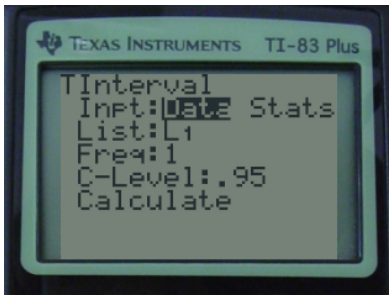
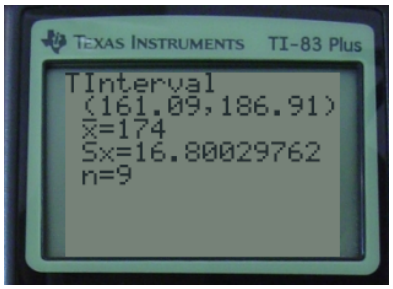
Tue, Apr 14 - Slide 300

1. Put the data in L1, and use tInterval. $n = 9$, $\overline{x} = 174$, and $s = 16.9$





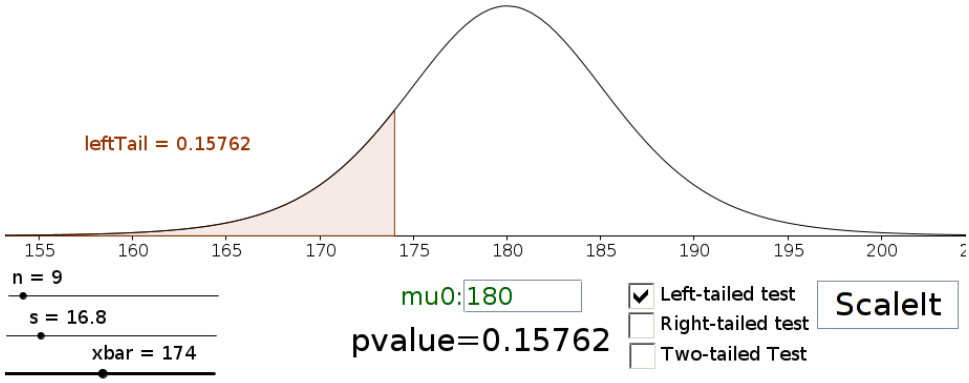
2. The 95% CI is $(161, 187) = 174 \pm 13$. We might think of this as a window estimate of Walter's true bowling ability.

3. $\mu = 180$ is inside the CI, so the data are reasonably consistent with that claim.

   He only averaged $\overline{x} = 174$, not 180. But given the small sample size and game-to-game variation, that is not surprising for a true 180 bowler.

Tue, Apr 14 - Slide 302

We will take a new perspective. Center a bell curve around the claimed population statistic, $\mu = 180$. Then see how **far out** in the tails the sample statistic $\overline{x} = 174$ is.



- He claimed to be a 180 bowler; he only bowled 174. Since $174 - 180 = -6$, we observed a six pin discrepancy.
- The picture shows there is a 15.8% chance that a true $\mu = 180$ bowler would have performed as poorly.
- The number .158 is called a **p-value**, a measure of how "**plausible**" the observed data are given the original claim.
- By most standards, .158 is not low enough to be **beyond reasonable doubt**. So we **fail to reject** the claim that $\mu = 180$.
- In other words, the 6 pin discrepancy is **not statistically significant**; we'll chalk it up to chance.

Tue, Apr 14 - Slide 303,304

The logic just described is called a **hypothesis test** (HT for short). Let's think of it in five parts. Not every problem will call for all five parts, but it's a helpful framework.

1. **null hypothesis** (H_0) - the original claim or baseline, a presumption

$$H_0 : \mu = 180$$

2. **alternative hypothesis** (H_1) - a directional suspicion we seek evidence for

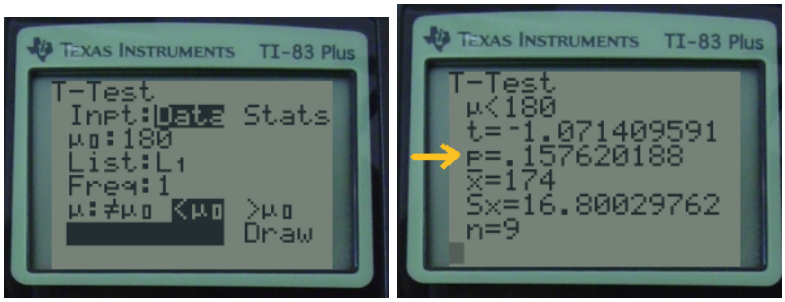
$$H_1 : \mu < 180$$

   This is called **"left-tailed"**, because we suspect he is really not as good as he claimed.

3. **effect size** - a measure of the discrepancy between claimed and observed statistics

$$174 - 180 = -6$$

4. **p-value** - mathematical probability (plausibility) of landing that far out in the distribution.
   Use your calculator: STAT-TESTS-tTest. Enter 180 for μ_0, and highlight the $<$ for the alternative.



   For all practical purposes, everything is **boiled down to a single number**, the **p-value** is 0.158. Your calculator refers to it as just "P".

5. **conclusion** - prescribe a threshold α (often .05)

   - if the p-value is less than α, **reject** H_0 and say the effect is **statistically significant**
   - otherwise, **fail to reject** H_0, and say the effect is **NOT statistically significant**

   In our case, $.158 > .05$, so we fail to reject H_0; the 6 pin difference is not statistically significant.

Tue, Apr 14 - Slide 306

Intuitively, all else equal, the p-value gets smaller if:

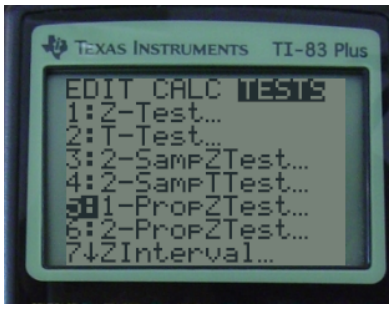
- the sample size increases (i.e. he was still averaging 174 after 30 games)
- the effect size (discrepancy) increases (i.e. he averaged only 164 pins)
- the standard deviation decreases (i.e. his scores were more consistent)

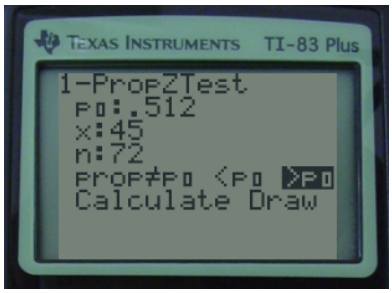
Tue, Apr 14 - Slide 308-312

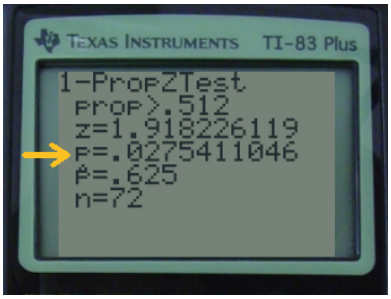
We've looked at a HT for a mean μ. Now let's test a hypothesis about a proportion p.

The natural rate is that 51.2% of babies are boys (note: the book uses 50%). Suppose (especially in historical contexts) that a couple would prefer to have a boy. There is some **"treatment"** available, (could be a superstition or some modern-day scientific pill or procedure) that **purports** to increase the chances of having a boy. Let's work through a formal HT.

1. $H_0 : p = .512$ (the baseline null hypothesis)
2. $H_1 : p > .512$ (alternative hypothesis is **"right-tailed"** because we suspect an increased p)
3. we observed $\hat{p} = \frac{45}{72} = .625$, so the "effect" is $.625 - .512 = .113$, or 11.3 percentage points
4. Use 1propZtest to get a p-value of .0275







5. Using $\alpha = .05$, we see that $.0275 < .05$, so by that standard, we would reject H_0 and say the effect is statistically significant. The treatment doesn't guarantee a boy, but there is strong evidence that the treatment increases the probability of a boy.

Note, that the p-value depends on the sample size as well as the effect size. With $\alpha = .05$,

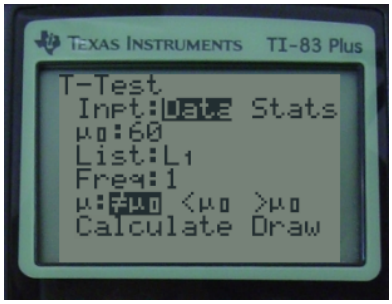
n	$\hat{p}$	p-value	stat.sig. ?
8	$\frac{5}{8} = 0.625$	.261	no
40	$\frac{25}{40} = 0.625$	.0764	no
72	$\frac{45}{72} = 0.625$	.0275	yes

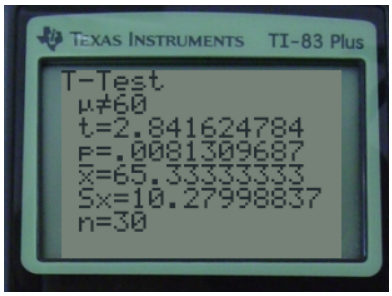
Tue, Apr 14 - Slide 315

Here is a sample of $n = 30$ times (in seconds) for students that tried to keep their eyes closed for a minute. Put them in L1.

54	63	65	65	52	80
69	63	68	73	63	80
77	70	71	88	53	69
77	64	80	69	61	49
62	54	55	45	62	59

1. $H_0 : \mu = 60$ (the benefit of the doubt says on average the internal clocks are accurate)
2. $H_1 : \mu \neq 60$ (a **"two-tailed"** alternative, to detect clocks that are too fast or too slow)
3. tTest gives $\overline{x} = 65.3$ (and $s = 10.28$), so the effect is $65.3 - 60 = 5.3$ seconds





4. tTest gives p-value .00813
5. This is a small p-value, certainly less than $\alpha = .05$, so we reject H_0, and say that there was "significant" error in the students' clocks.

Tue, Apr 14 - Slide 317-322

Although it is easy to mis-interpret, the HT procedure is ubiquitous in modern science. Do not get confused about what the p-value means. It tells you how **plausible** the observed data are if H_0 were true.

- p-value small ($\leq \alpha$): reject H_0, say the effect is statistically significant
- p-value not small ($> \alpha$): fail to reject H_0; the effect is not statistically significant

Make sure to recognize scientific notation, e.g. p-value $8.4E^{-}7 = 0.00000084$ would be extremely small, indicating statistical significance.

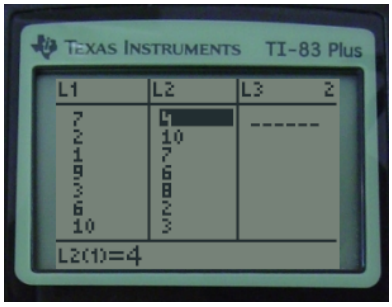
The significance level α is prescribed in the context of the research being done. We commonly use $\alpha = .05$ in class, an academic research journal might set $\alpha = .01$, and the FDA (food and drug administration) might use $\alpha = .001$. Go with $\alpha = .05$ unless I say otherwise.

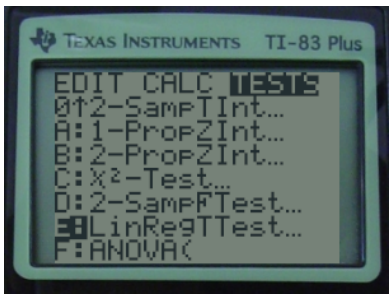
A p-value of $.05 = \frac{1}{20}$ means that if H_0 were true, then there would be a 1 in 20 chance of observing sample statistics this "far out" by purely dumb luck.

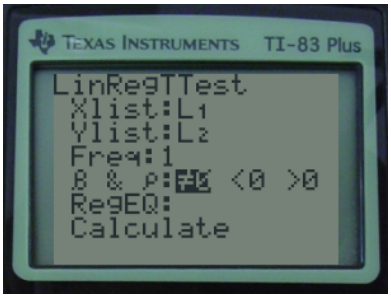
Tue, Apr 14 - Slide 327,328

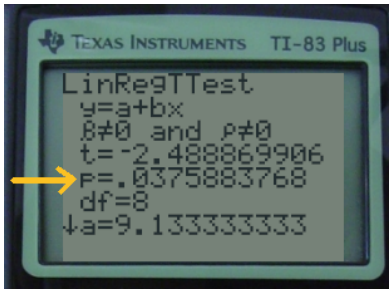
We have done HT for a mean μ and a proportion p. Now let's do a HT for a correlation ρ (Greek "rho"). Enter the given x and y values into L1 and L2.

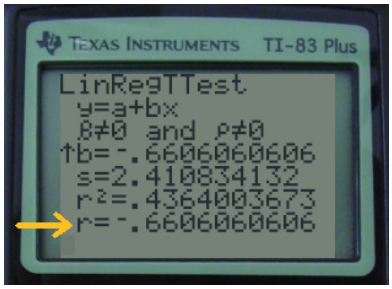
1. $H_0 : rho = 0$ (the variables are uncorrelated)
2. $H_1 : rho \neq 0$ (two-tailed to detect either positive or negative correlation)
3. **linregttest** gives the sample correlation is $r = -0.661$, so x and y appear to be negatively correlated. But with such a small sample size, is that correlation statistically significant ?











4. the p-value is .0376
5. $.0376 < .05$, so reject H_0 and say the observed correlation is statistically significant

Thu, Apr 16 - Slide 329,339

You have now seen the basic concepts of inference: confidence interval (CI) and hypothesis test (HT).

Now let us apply the same ideas in the context of comparing two populations to see if there is a significant **difference** between them. The null hypothesis is always the two statistics are equal.

Continue to shove the mathematics under the rug; it is much more important that you

- learn the jargon
- recognize which inference tool applies to the given statistic(s)
- know how to interpret confidence intervals and p-values

Just use your TI calculator as illustrated in the remaining examples. Here is slide 339, which summarizes TI functions for doing CI and HT inference. Of course, in practice use Excel or some other software to do essentially the same things.

stat	samples	distribution	T.I.
proportion	1	z (normal)	1-PropZtest 1-PropZint
proportion	2	z (normal)	2-PropZtest 2-PropZint
mean	1	t	tTest tInterval
mean	2	t	2-SampTtest 2-SampTint
st.dev.	2	F	2-SampFtest
correlation	(x, y)	t	LinRegTtest

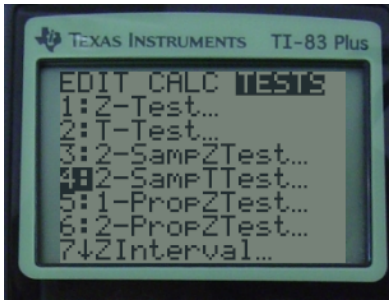
Thu, Apr 16 - Slide 331,332

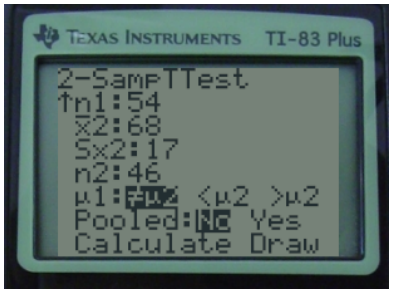
This example illustrates a **randomized controlled trial**, in which (unbeknownst to them):

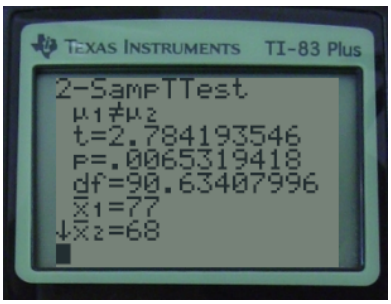
- some randomly chosen participants got the real **treatment**
- others got something else, a **control** (often a **placebo**)

So we have **two populations**, and we want to compare the results. First we'll do a HT.

1. $H_0 : \mu_T = \mu_C$ (there is no difference between treatment and control)
2. $H_1 : \mu_T \neq \mu_C$ (two-tailed, maybe the treatment hurts)
3. we observed the treatment group did $77 - 68 = 9$ points better on average
4. use **2sampTtest** to get a p-value of .00653

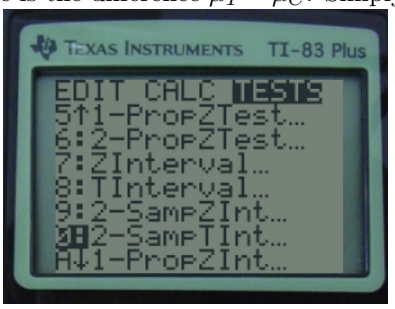


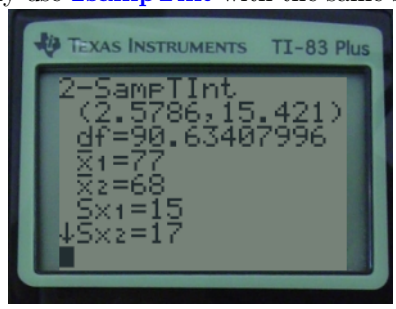




5. Since the p-value is small enough ($.00653 < .05$), reject H_0, and say that the treatment **significantly** boosts performance.

The researcher might also choose to report a confidence interval on the size of the **effect**, which in this case is the difference $\mu_T - \mu_C$. Simply use **2sampTint** with the same stats to get a 95% CI of $(2.58, 15.42)$.



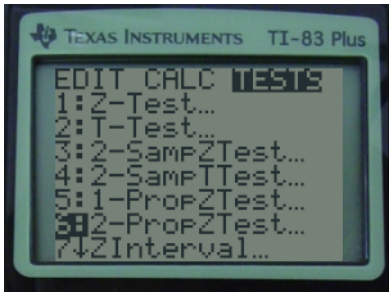


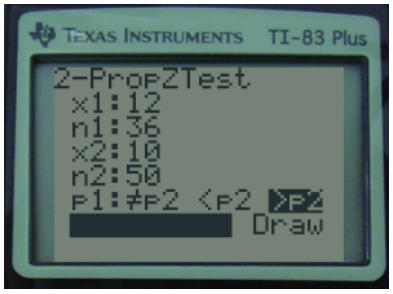
We might say that 5 hour energy boosts performance on this test by 9 ± 6.42 points.

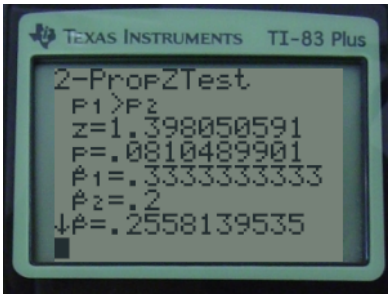
Thu, Apr 16 - Slide 333,334

Same idea, but this time we are comparing proportions, not means. We know **how many** of each group checked the box to indicate itchy ears, but not the degree of itchiness. Let's do a **one-tailed** HT, and then a CI.

1. $H_0 : p_T = p_C$ (null hypothesis is always that the stats are equal)
2. $H_1 : p_T > p_C$ (we would only care if the treatment "caused" itchy ears)
3. $\frac{12}{36} - \frac{10}{50} = .133$, so 13.3 percentage point difference
4. use **2propZtest** to get a p-value of .081

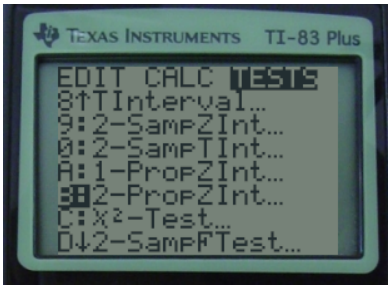


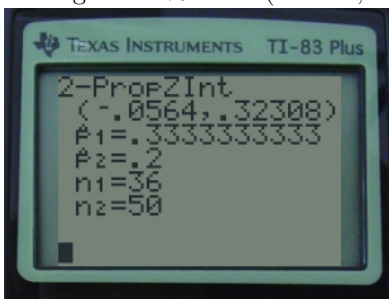




5. This p-value is not small enough ($.081 > .05$), so fail to reject H_0. The 13.3 percentage point difference, although large, is not statistically significant (primarily due to small sample size).

Use **2propZint** with the same stats to get a 95% CI of $(-.0564, .323) = .133 \pm .19$.





The CI is so wide, it contains both positive and negative numbers.

Thu, Apr 16 - Slide 336

By now you see that the HT and CI ideas can be applied to pretty much any statistic. But in the interest of time, we'll skip the **2sampFtest** for comparing two standard deviations.

However, without knowing any details, if somebody tells you that a p-value is .225, you automatically know that whatever effect they're talking about is **not** statistically significant ($.225 > .05$).

Thu, Apr 16 - Slide 340

For the remainder of the class, practice reading a problem and deciding what kind of inference is called for. Then get the CI or p-value with your calculator. Here are three questions to help you narrow it down:

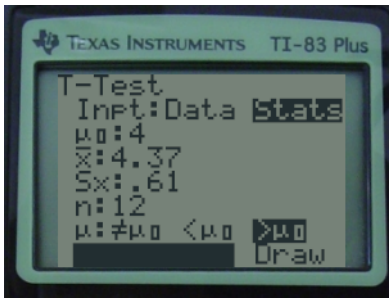
- 1 sample compared to a fixed number, 2 samples compared to each other, or (x, y) pairs
- mean, proportion, or correlation (remember we're skipping standard deviation)
- CI and/or HT

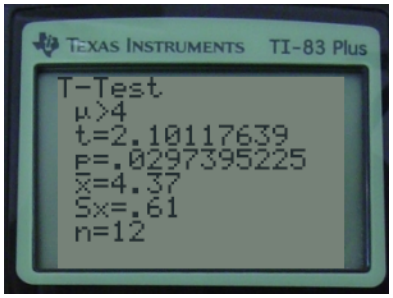
In a live setting, I might wax philisophical about how p-values are often cherry-picked, how we have a replication crisis in modern science, and how the term "significant" is used to manipulate an unsuspecting public. I'll refrain from that except to say **have a healthy degree of skepticism** when you read statistics. There's no telling what's going on in the kitchen.

Thu, Apr 16 - Slide 343

We have one sample, are computing a mean, and we'll do a HT. So use **tTest**.

1. $H_0 : \mu = 4$ (benefit of the doubt, meeting EPA limit)
2. $H_1 : \mu > 4$ (don't want to go over)
3. Observed $4.37 - 4 = .37$ units over. Is that statistically significant ?
4. **tTest** gives p-value .0297.





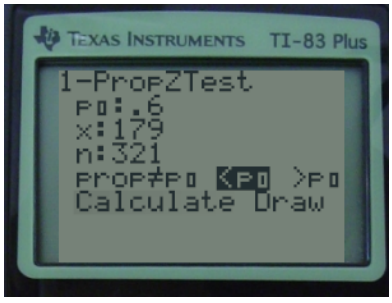
5. $.0297 < .05$, so reject H_0. The difference is "significant".

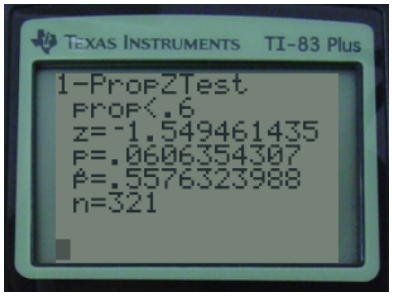
You can use **tInterval** with the same stats to get a 95% CI of $(3.98, 4.76) = 4.37 \pm .39$ units of radon.

Thu, Apr 16 - Slide 344

We have one sample, are computing a proportion, and we'll do a HT. So use **1propZtest**.

1. $H_0 : p = .6$ (benefit of the doubt, meeting the goal)
2. $H_1 : p < .6$ (don't want to go under)
3. Observed $\hat{p} = \frac{179}{321} = .5576$, which is 4.24 percentage points below the 60% goal. Is it "significantly" below the goal ?
4. **1propZtest** gives p-value .061.



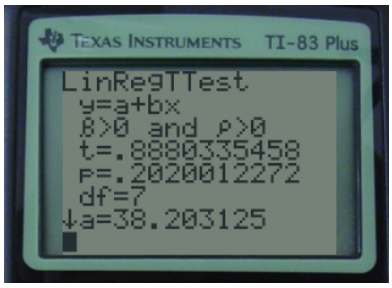


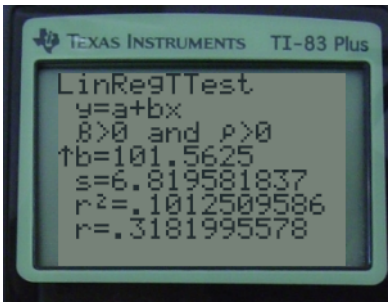
5. .061 > .05, so fail to reject H_0. Although the shortfall is not statistically "significant", the fact of the matter is that they did not meet their goal. If that persists in a bigger sample size, the p-value will drop below .05.

Thu, Apr 16 - Slide 345

We have x,y values, are computing a correlation, and we'll do a HT. So use **linregttest** (with data in L1,L2).

1. $H_0 : rho = 0$ (uncorrelated)
2. $H_1 : rho > 0$ (look for positive correlation)
3. Observed $r = .318$. Is that statistically significant ?
4. **linregttest** gives p-value .202.





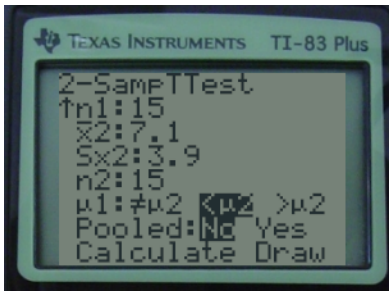
5. .202 > .05, so fail to reject H_0. The correlation is NOT "significant".

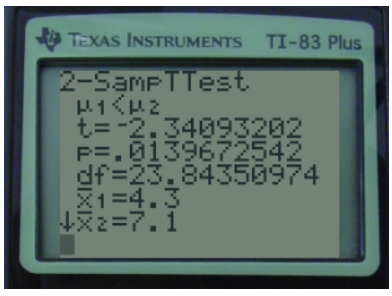
Thu, Apr 16 - Slide 347

You might want to watch `https://youtu.be/oMy1yjQ9HIs` about this problem.

We have two samples, are computing means, and we'll do a HT. So use **2sampttest**. Let's say the researcher wants a **one-tailed** test.

1. $H_0 : \mu_R = \mu_M$ (no difference)
2. $H_1 : \mu_R < \mu_M$ (looks longer at magic event)
3. $7.1 - 4.3 = 2.8$ seconds. Is that statistically significant ?
4. **2sampttest** gives p-value .014.



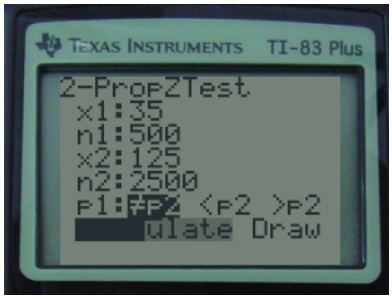


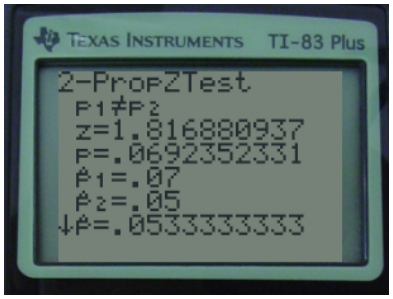
5. .014 < .05, so reject H_0. The babies look "significantly" longer at magical events

Thu, Apr 16 - Slide 348

We have two samples, are computing proportions, and we'll do a HT. So use **2propZtest**. Let's do a **two-tailed** test.

1. $H_0 : p_B = p_N$ (no difference)
2. $H_1 : p_B \neq p_N$ (brochures either encourage or discourage enrollment yield)
3. $.07 - .05 = .02$. Is the 2 percentage point difference statistically significant ?
4. Note that 7% of 500 is 35 out of 500 that got the brochure enrolled. Similarly $(.05)(2500) = 125$ of the non-brochure group enrolled.
   2propZtest gives p-value .069.





5. $.069 > .05$, so fail to reject H_0. The brochures can't be said to have a "significant" impact, but the p-value is fairly small, so we could justify continuing the experiment.

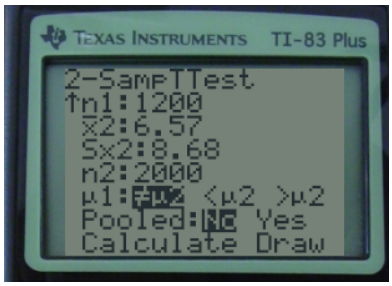
Tue, Apr 21 - Slide 360

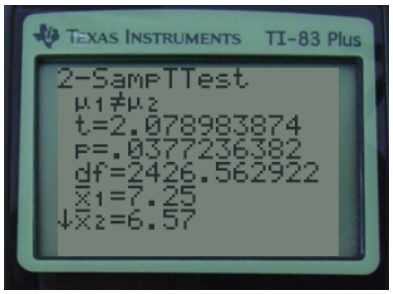
Let's put some numbers to (2.). A new late-stage cancer treatment drug is being compared to an old drug that will soon have its patent expire. We'll compare the life expectancies (in months). Suppose our research sets $\alpha = .01$ as the significance threshold.

	n	$\overline{x}$	s
new drug	1200	7.25	9.12
old drug	2000	6.57	8.68

We have two samples, are computing means, and we'll do a HT. So use **2sampttest**. Do a **two-tailed** test.

1. $H_0 : \mu_1 = \mu_2$ (no difference)
2. $H_1 : \mu_1 \neq \mu_2$ (one drug is "better")
3. $7.25 - 6.57 = .68$ months. Is that statistically significant ?
4. **2sampttest** gives p-value .0377.



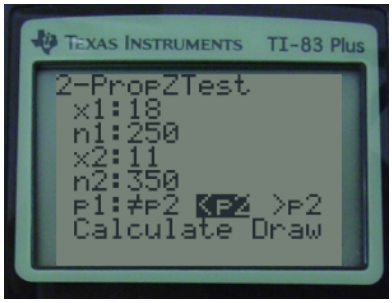


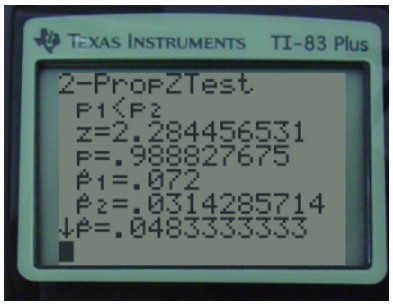
5. .0377 > .01 so fail to reject H_0. The new drug can't yet be said to be "significantly" better. But if the effect persists with a larger sample, the p-value will eventually drop below .01 and convince the researchers of the new drug's superiority.

Tue, Apr 21 - Slide 366

We have two samples, are computing proportions, and we'll do a HT. So use **2propZtest**.
Do a left-tailed (<) test because you suspect the signs reduce theft of petrified wood.

1. $H_0 : p_T = p_C$ (no difference)
2. $H_1 : p_T < p_C$ (signs reduce theft)
3. $\frac{18}{250} - \frac{11}{350} = .072 - .0314 = .0406$. Is that statistically significant ?
4. **2propZtest** gives p-value .989





5. This p-value is huge (.989 > .05), so fail to reject H_0.

This is a case where the treatment **backfired**. If anything, the signs seem to increase theft. Perhaps it gives hikers the idea, or makes them think that other people are stealing, so they should also before all the good pieces are gone.

- Redo the test as a two-tailed ($\neq$) test and you'll get a p-value of .022.
- Use **2propZint** to get a 95% CI of $(.0037, .0775) = .0406 \pm .0369$ for increase in theft.

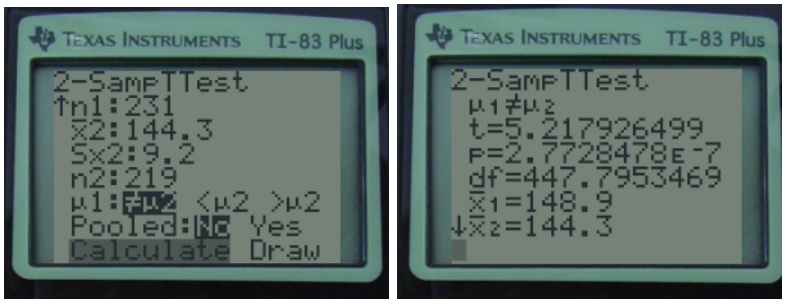
Tue, Apr 21 - Slide 368

Suppose an ultrasound technician recorded fetal heart rates as follows:

	n	$\overline{x}$	s
girls	231	148.9	9.5
boys	219	144.3	9.2

We have two samples, are computing means, and we'll do a HT. So use **2sampttest**. Do a **two-tailed** test.

1. $H_0 : \mu_G = \mu_B$ (no difference)
2. $H_1 : \mu_G \neq \mu_B$ (girls and boys have different heart rates before birth)
3. $148.9 - 144.3 = 4.6$ beats per minute. Is that statistically significant ?
4. **2sampttest** gives p-value $2.77E^{-}7 = 0.000000277$ (notice the scientific notation!)



5. This p-value is tee-tiny, so reject H_0. Girls have "significantly" higher fetal heart rates.

If we wanted a CI for the difference between girls' and boys' heart rates, use **2samptint** to get:
$(2.87, 6.33) = 4.6 \pm 1.73$ beats per minute.

Tue, Apr 21 - Slide 369

A certain smartwatch monitors sleep patterns. Here is a scatterplot of the time sleep began vs longitude within a time-zone.

1. What is the slope?
   Answer: Imperceptible, but the slope is $-.008896$. So as you go east, average bed times are earlier. In fact, a time zone is 15 degrees of arc, so on average, somebody on the eastern end goes to bed $(.008896)(15) = .13344$ hours, or about 8 minutes earlier than somebody on the western end.
2. What is the correlation ?
   Answer: $r = -\sqrt{.00145677} = -.0382$ (notice correlation is negative)
3. Is the correlation statistically significant at the $\alpha = .01$ level?
   Answer: yes, because the p-value is small ($.0031 < .01$)