Essential Math for Data Analysis Using Excel Online

Module 5, Lab 1: T-Test

Learning Objectives

- Compare the performance of Team X and Team Y on a KPI (average profit per team), using an independent samples *t*-test.

Description

Learners will use Excel to perform an inferential *t*-test to indicate whether any difference in two teams' performance is likely to continue in the future.

Data set

Mod5Lab1.csv

Overview

Imagine your company has hired two different freelance teams, Team X and Team Y, for a number of projects. For each project, the profit for your company is reported. In this lab, you'll calculate each team's mean profit and standard deviation. Then you'll use a *t*-test to determine whether the difference in each team's performance is statistically significant (and likely to continue in the future) or just due to random chance.

What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

Exercise 1: *T*-Values and *P*-Values

1. Open the data set in Excel. Here's what the data look like:

	A	B	C	D
1	projetid	team	profit.earned.for.company	
2	MF3688S	Team X	$1,590.56	
3	EQ7342V	Team X	$1,142.64	
4	PQ2146C	Team X	$989.90	
5	NR2174R	Team X	$1,030.48	
6	HF4826R	Team X	$818.83	
7	GV2691H	Team X	$1,150.00	
8	DG6685N	Team X	$1,026.49	
9	RS6664Z	Team X	$1,230.89	
10	SB4945B	Team X	$1,164.28	
11	CK6143K	Team X	$1,074.21	
12	SD4157X	Team Y	$289.01	
13	OG7194I	Team Y	$457.82	
14	XI3380I	Team Y	$887.38	
15	RR9444Q	Team Y	$471.54	
16	RM2073W	Team Y	$552.92	
17	XP7023E	Team Y	$1,313.18	
18	AL3024B	Team Y	$681.48	
19	ZX1975Y	Team Y	$548.42	

Team X's projects/profits are grouped at the top, and Team Y's are at the bottom.

2. In a new column, find the mean/average profit earned by Team X. Use Excel's AVERAGE function for each. Be careful to use the correct range of cells: Team X's projects run from C2 down to C11.

	A	B	C	D
1	projetid	team	profit.earnec	Team X avg
2	MF3688S	Team X	$1,590.56	$1,121.83
3	EQ7342V	Team X	$1,142.64	
4	PQ2146C	Team X	$989.90	
5	NR2174R	Team X	$1,030.48	
6	HF4826R	Team X	$818.83	
7	GV2691H	Team X	$1,150.00	
8	DG6685N	Team X	$1,026.49	
9	RS6664Z	Team X	$1,230.89	
10	SB4945B	Team X	$1,164.28	
11	CK6143K	Team X	$1,074.21	
12	SD4157X	Team Y	$289.01	
13	OG7194I	Team Y	$457.82	
14	XI3380I	Team Y	$887.38	
15	RR9444Q	Team Y	$471.54	
16	RM2073W	Team Y	$552.92	
17	XP7023E	Team Y	$1,313.18	
18	AL3024B	Team Y	$681.48	
19	ZX1975Y	Team Y	$548.42	

So Team X made a mean of $1,121.83 per project.

3. Now find the mean/average profit for Team Y. This time, the range of cells for Team Y runs from C12 to C19.

	A	B	C	D	E
1	projetid	team	profit.earnec	Team X avg	Team Y avg
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22
3	EQ7342V	Team X	$1,142.64		
4	PQ2146C	Team X	$989.90		
5	NR2174R	Team X	$1,030.48		
6	HF4826R	Team X	$818.83		
7	GV2691H	Team X	$1,150.00		
8	DG6685N	Team X	$1,026.49		
9	RS6664Z	Team X	$1,230.89		
10	SB4945B	Team X	$1,164.28		
11	CK6143K	Team X	$1,074.21		
12	SD4157X	Team Y	$289.01		
13	OG7194I	Team Y	$457.82		
14	XI3380I	Team Y	$887.38		
15	RR9444Q	Team Y	$471.54		
16	RM2073W	Team Y	$552.92		
17	XP7023E	Team Y	$1,313.18		
18	AL3024B	Team Y	$681.48		
19	ZX1975Y	Team Y	$548.42		

It looks like Team Y only had a mean profit of $650.22 per project. That's quite a bit lower than Team X's mean, but the question now is whether this difference is likely to hold up in the future. That's where the *t*-test comes in, but first you'll need a few more values to plug into the formula for finding the *t*-value.

As a quick reminder, here's what that formula looks like:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

It looks pretty horrifying, but you can use Excel to find all the values you need and crunch the numbers for you. Those two *x*-bars at the top of the fraction are the means of each data set, which you just found. The sigmas represent the standard deviation of each data set (which we'll find next), and the *n*-values are just the number of data points in each set. In the formula, imagine that everything with a subscript of 1 represents Team X, and everything with a subscript of 2 represents Team Y.

4. In a new column, find the standard deviation of Team X's profit values using Excel's STDEV function. The syntax is **=STDEV(first value:last value)**. You want the standard deviation of Team X's values only, so use the range C2:C11.

=STDEV(C2:C11)

	A	B	C	D	E	F
1	projetid	team	profit.earned	Team X avg	Team Y avg	Team X sd
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22	200.9829
3	EQ7342V	Team X	$1,142.64			
4	PO2146C	Team X	$989.90			

The standard deviation for Team X is about 200.9829. That's the value that you'll plug into the formula for $\hat{\sigma}_1$.

5. Now create another new column and grab the standard deviation of Team Y's profits using STDEV again. This time, the range of cells for Team Y runs from C12 to C19.

=STDEV(C12:C19)

	A	B	C	D	E	F	G
1	projetid	team	profit.earned	Team X avg	Team Y avg	Team X sd	Team Y sd
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22	200.9829	319.8181
3	EQ7342V	Team X	$1,142.64				
4	PO2146C	Team X	$989.90				

Team Y's standard deviation is 319.8181, so that's what you'll use for $\hat{\sigma}_2$ in the formula.

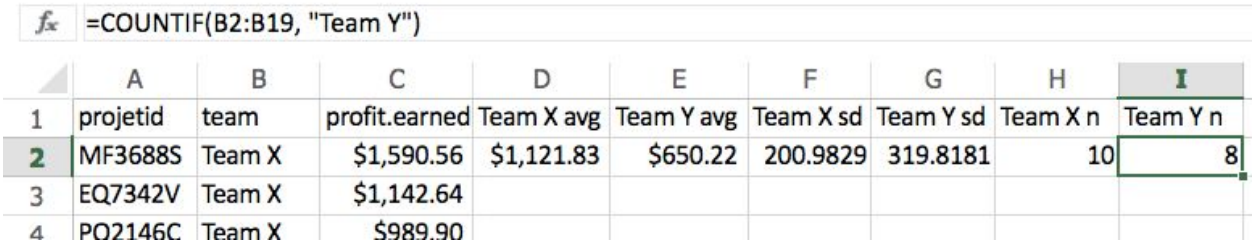
6. The only other values you need for the formula are n_1 and n_2. There aren't too many data points in this data set, so you can probably just count each team's number of values by hand (n_1 is the number of entries for Team X, and n_2 is the number of entries for Team Y).

   The safer way to do it, though, is to use the COUNTIF function, which is especially useful if you've got a huge data set. In a new column, use **=COUNTIF(B2:B19, "Team X")** to count the number of entries for Team X (and don't forget to put "Team X" in quotes, since it's text instead of numerical).

=COUNTIF(B2:B19, "Team X")

	A	B	C	D	E	F	G	H
1	projetid	team	profit.earned	Team X avg	Team Y avg	Team X sd	Team Y sd	Team X n
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22	200.9829	319.8181	10
3	EQ7342V	Team X	$1,142.64					
4	PO2146C	Team X	$989.90					

7. Use the same syntax (and another new column) to count the number of entries for Team Y.



	A	B	C	D	E	F	G	H	I
1	projetid	team	profit.earned	Team X avg	Team Y avg	Team X sd	Team Y sd	Team X n	Team Y n
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22	200.9829	319.8181	10	8
3	EQ7342V	Team X	$1,142.64						
4	PQ2146C	Team X	$989.90						

So Team X had 10 entries (that's n_1), and Team Y had 8 entries (n_2).

8. Now plug everything you know into the formula. The *x*-bar values up top are the mean/average values for each team, the sigmas are the standard deviations, and the *n*'s are the number of entries for each team.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\hat{\sigma}_1^2(n_1-1)+\hat{\sigma}_2^2(n_2-1)}{(n_1-1)+(n_2-1)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

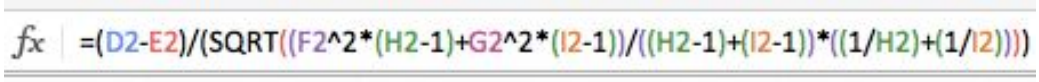
Here we go:

$$t = \frac{1121.83 - 650.22}{\sqrt{\left(\frac{200.9829^2(10-1)+319.8181^2(8-1)}{(10-1)+(8-1)}\right)\left(\frac{1}{10}+\frac{1}{8}\right)}}$$

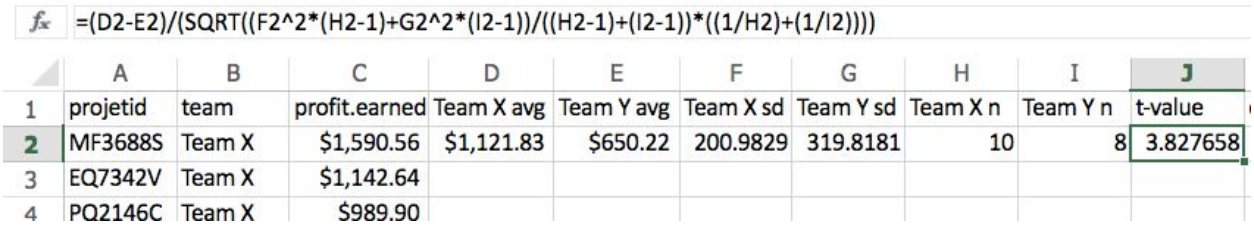
Yikes. You can either plug that into a calculator or set up a formula in Excel to calculate it. To do the latter, it might be easier to write the formula in terms of the cell numbers where all those values are located in your spreadsheet. Like so:

$$t = \frac{D2 - E2}{\sqrt{\left(\frac{(F2)^2(H2-1)+(G2)^2(I2-1)}{(H2-1)+(I2-1)}\right)\left(\frac{1}{H2}+\frac{1}{I2}\right)}}$$

9. In a new column for the *t*-value, plug all of that into an Excel formula… very, very carefully. Keep careful track of your parentheses. If you're using Excel 2016 on a desktop, the program will color-coordinate your parentheses for you.



If that's too brain-wrenching, it's also fine to break down each step of the formula into separate cells and combine them all at the end. Either way, here's the value you should get in the end:



	A	B	C	D	E	F	G	H	I	J
1	projetid	team	profit.earned	Team X avg	Team Y avg	Team X sd	Team Y sd	Team X n	Team Y n	t-value
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22	200.9829	319.8181	10	8	3.827658
3	EQ7342V	Team X	$1,142.64							
4	PQ2146C	Team X	$989.90							

The *t*-value is about 3.828. Translation: The difference between Team X and Team Y's average profits is 3.828 times *greater* than you would expect if the null hypothesis were true. In other words, the difference in the two teams' performance is 3.828 times greater than it would be by random chance… which means it's probably *not* random that Team X did so much better than Team Y. Team X really is performing better.

10. Now it's time to use that *t*-value to find a *p*-value, which is a probability. First, you'll need the degree of freedom (*df*). To get this value, add the number of entries for each team and subtract 2.

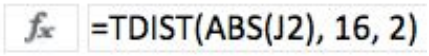
$df = n_1 + n_2 - 2$

Remember, Team X had 10 entries and Team Y had 8 entries.

$df = 10 + 8 - 2$
$df = 18 - 2$
$df = 16$

So the degree of freedom is 16.

11. Set up one final column for the *p*-value in your sprawling Excel spreadsheet. Use Excel's TDIST function with the syntax **=TDIST(ABS(*t-value*), *df*, 2)**. Use cell J2 for your *t*-value, and use 16 for *df*. (That 2 at the end of the TDIST syntax means you're using a 2-tailed distribution.)



	A	B	C	D	E	F	G	H	I	J	K
1	projetid	team	profit.earned	Team X avg	Team Y avg	Team X sd	Team Y sd	Team X n	Team Y n	t-value	p-value
2	MF3688S	Team X	$1,590.56	$1,121.83	$650.22	200.9829	319.8181	10	8	3.827658	0.001484
3	EQ7342V	Team X	$1,142.64								
4	PQ2146C	Team X	$989.90								

Finally, you've tracked down the *p*-value of 0.001484.

12. Okay, but what does that number mean, exactly? It's a probability, so if you convert 0.001484 to a percentage, it's 0.1484% (just move that decimal point two spots to the left). That's less than one percent. It means that if you assume both teams are equal in the long run, you would expect a discrepancy *this big* between Team X and Team Y's average profits to occur only 0.1484% of the time. That's a really low probability, so you can reject the null hypothesis.

    Translation: It would be nearly impossible to get a difference this big by random chance, so we can safely attribute the difference observed to a real gap in performance. In other words, Team X is better, and thus likely to continue to be better — at least in terms of average profits.