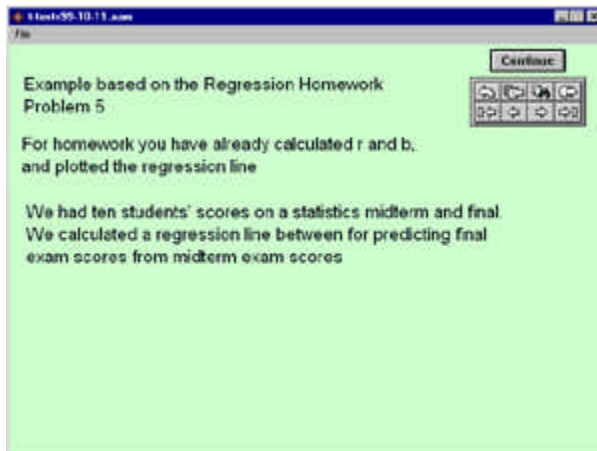


# t-test for b

Copyright 2000 Tom Malloy. All rights reserved.



## Regression

Recall, back some time ago, we used a descriptive statistic which allowed us to draw the best fit line through a scatter plot. We called this linear regression.

In regression every participant is measured on two variables, like height and weight, or GPA in college and ACT score at the end of high school. After you measure two variables of interest to you, you draw a scatterplot which represents each research participant as a point on the graph. Then you can use some rather complicated

formulas, especially if you include calculations for the Pearson  $r$ , to calculate the regression line. Recall that the regression line formula is the predicted score  $Y'$ , was equal to **a plus bx**, ( $Y' = a + bx$ ) where **a** is the intercept and **b** is the slope of the regression line. So **t for b**, then, is a t test to determine if the slope of the regression line, **b**, is significantly greater than zero.

## Example

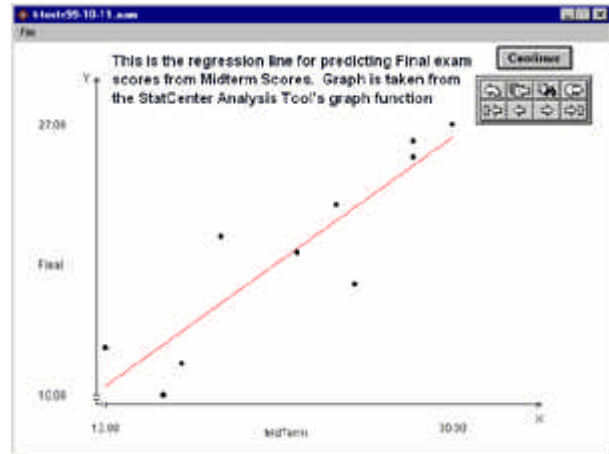
The example which we will use is an example which you may have already worked with on your practice homework for regression. In that example a teacher thought she could predict final exam scores from midterm exam scores. She also thought there would be a positive relationship between midterm and final scores so that the slope of the regression line should be positive. She had ten students who took a midterm and final. She used the regression formulas to predict the final scores from the midterm scores. In other words, she calculated a regression line between those two variables, X and Y. Let's call the Final Scores **Y** and the Midterm Scores **X**.

[Go To Menu Man](#)

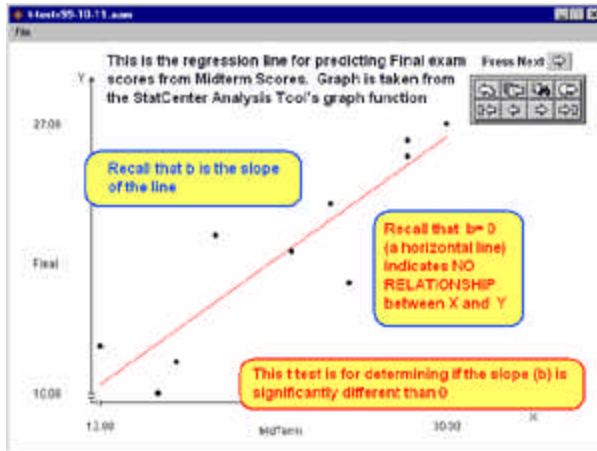
[Go to Home Map](#)

## Scatterplot

Here is how the data looks. This is actually just a screen capture from StatCenter's Analysis Tool which you've used several times by now. This tool has a graphing function and it can draw the scatter plot of appropriate data and then draw the regression line on the graph so that you don't have to do that. All I did was use the data that is already available in the Analysis Tool for the practice regression homework. Then I used Analysis Tool to graph it. Each dot is an individual participant with final exam scores on the vertical axis and their midterm exam scores on the horizontal. The best fit regression line is the red line shown here and it was calculated by the formulas that we've already considered in the lecture on regression.



**The Question.** In this lecture we are asking: Is  $b$ , the slope of that regression line, significantly different from zero or not?



## $b$ is the Slope of a Line

First, recall that  $b$  is the slope of the regression line. Second recall that if  $b$  equals zero then there is no slope. If  $b = 0$  then you would have a perfectly horizontal line and this indicates no relationship between the two variables  $X$  and  $Y$ . Think about it. If the line were horizontal, then no matter what value of  $X$  you put into the equation, you would get the same value of  $Y$  as a prediction.

So in the midterm and final scores case, if the red line was horizontal (obviously it's not horizontal) and it ran straight across the graph, that would indicate there was no relationship between midterm and final exam scores. Such an outcome would be quite counterintuitive from our knowledge of having taken a lot of exams.

**This t test then, is for determining if the slope,  $b$ , is significantly different from zero.**

**Note:** As a side note, **t for b** is completely redundant with **t for r**. If you took the exact same data that you have here for midterms and finals and calculated little  $r$ , the correlation coefficient, and then did t for  $r$ , you would get the same result as if you did the t for  $b$ . **The calculated t value for the two tests would be identical, except for rounding error.**

[Go To Menu Map](#)

## Directional Scientific Hypothesis

The scientific hypothesis is that there is a predictable positive relationship between midterm scores and final exam scores. This would be a pretty common sense scientific hypothesis.

This particular hypothesis is **directional**. We are saying there is a positive relationship and that we expect a positive slope. Another way of saying the same thing is that we expect **b** to be greater than zero.

Scientific Hypothesis: There is a positive relationship between Midterm Scores and Final Exam Scores

Is the Scientific Hypothesis Directional or Non-Directional?

Directional: We are saying there is a Positive relationship. That means we expect a positive slope (  $b > 0$  )

Remember that, in general, **b** could be less than zero. In that case the line would slope in the opposite direction; it would have high scores on the midterm predict low scores on the final. That would be a negative relationship. In this example though we are expecting a positive relationship between midterm and final exams. That is, we're expecting people with higher scores on the midterm exam will have higher final exam scores as well.

[Go To Menu Map](#)

$H_0: E(b) = 0$

$H_1: E(b) > 0$  (One-tailed)

$t = \frac{b - 0}{S_b}$

Notice that if  $H_0$  is true, then the top of the  $t$  formula will be zero. Anything divided into zero is zero, so that we expect  $t$  to equal zero

If  $H_0$  is true  $E(t) = 0$

## Null and Alternative Hypotheses

The skeptic would think that there's no relationship between mid term and final scores. Therefore the skeptic would expect that the slope of the regression line would be zero. So  $H_0$  is that we expect  $b$  to be equal to zero.

The scientist as we have said is expecting a positive slope. The alternative hypothesis,  $H_1$ , is that we expect a positive slope, the expected value of  $b$  is greater than zero.

Because the scientific hypothesis is directional  $H_1$  leads to a one-tailed test.

If the skeptic were presented with the scatterplot and regression line we looked at above, the skeptic would say that the scatterplot is completely random, it happened by chance alone, it is like shotgun pellets hitting the barn door. By luck alone there appears to be a positive relationship between midterms scores and final scores. So we do a  $t$  test for the significance of  $b$  to evaluate the PCH of Chance.

## Expected Value of t

On the bottom of the graphic you can see part of the formula for the t test. It is small and you don't even have to write it out yet, but the main idea I want you to comprehend here is that this **t** is like all other **t**'s should be zero if **F** is true.

The top of the **t for b** formula is going to be **b** times some standard deviation times some square root. It doesn't matter what the actual values will be, because if **b** equals 0, then the whole top must be equal to 0 (because anything times zero is zero). And so the whole value of **t** must be zero because anything divided into zero must be zero.

The null hypothesis predicts a value of zero for **t**.

[Go To Menu Map](#)

## The t for b Formula

t test for testing the significance of b

$$t = \frac{bS_x\sqrt{N}}{\sqrt{\frac{NS_y^2 - Nb^2S_x^2}{N - 2}}}$$

df = N - 2

**Statistical Results**

Variables:	MidTerm	Final
S <sup>2</sup>	34.9600	32.8900
s	5.9127	5.7350
s <sup>2</sup>	34.9644	32.8444

Standard Deviation:

Variable:	MidTerm	Final
S	5.9127	5.7350
s	6.2325	6.3452

Correlation Statistics:

Variable:	Final and MidTerm
Pearson r	0.9942

Regression Statistics:

Dependent (Y):	Final
Predictor (X):	MidTerm
Regression Line:	Y = 0.8673 X + 0.1804

df = N - 2

	X	Mid	Final	Y
	30	27		
	12	13		
	15	10		
	25	17		
	22	19		
	28	25		
	18	20		
	16	12		
	24	22		
	28	28		

S<sup>2</sup> = 34.96 32.89  
S = 5.9127  
b = 0.8673

**We need the Variance of Y (Final), Variance of X (Midterm), Standard Deviation of X (Midterm), and b (regression coefficient)**

**These statistics were taken from the StatCenter Analysis Tool**

**Can you substitute correctly into the formula?**

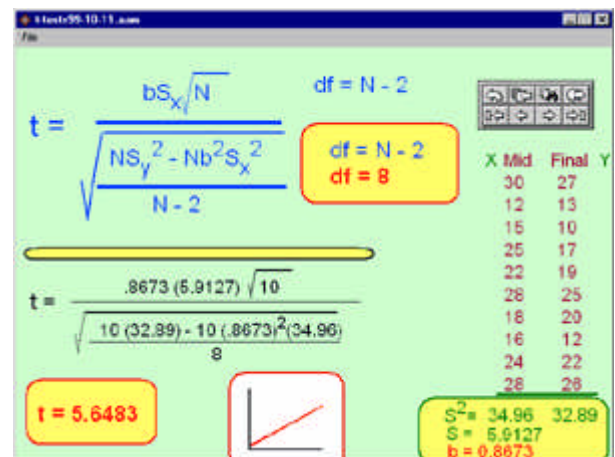
Here on the above graphics are the data and the entire formula for the t test. Also shown is the output of Analysis Tool from StatCenter. The Analysis Tool outputs are summarized below the data on the second graphic.

The main thing is to understand **X** is the midterm and **Y** is the final and that we have already done the summary statistics. As you write the t formula and the summary statistics down in your notes, please substitute the summary statistics into the t formula. Just make sure you can do that now because you will have to on homework and exams.

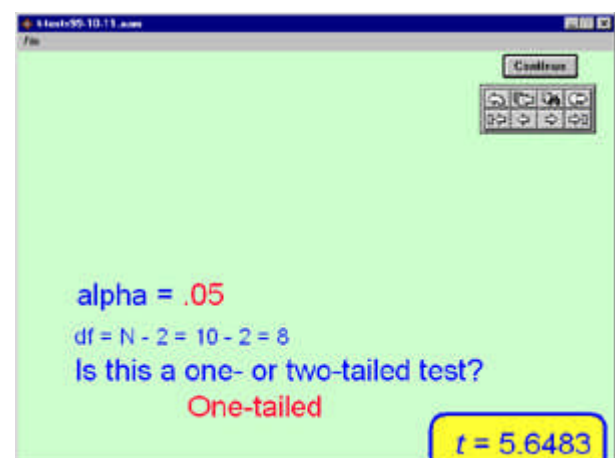
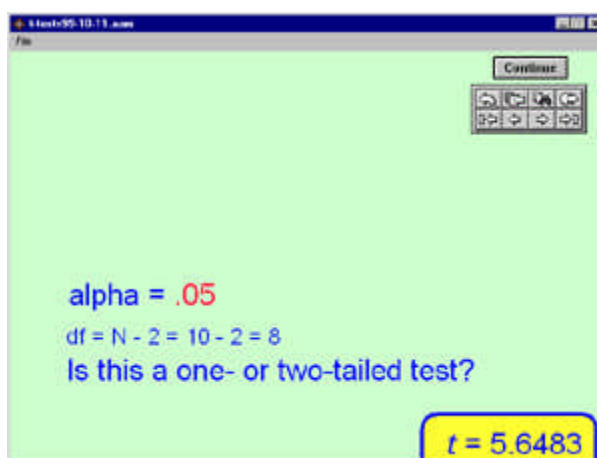
[Go To Menu Map](#)

## Calculated Value of t for b

Here's my substitution which you can use to check your work. Again, my recommendation is to substitute first and then look at my numbers. Once you have substituted the correct values into the formula, then it is just a matter of arithmetic. In this case, I skipped all the steps in the arithmetic. **The calculated t is equal to 5.6483.**

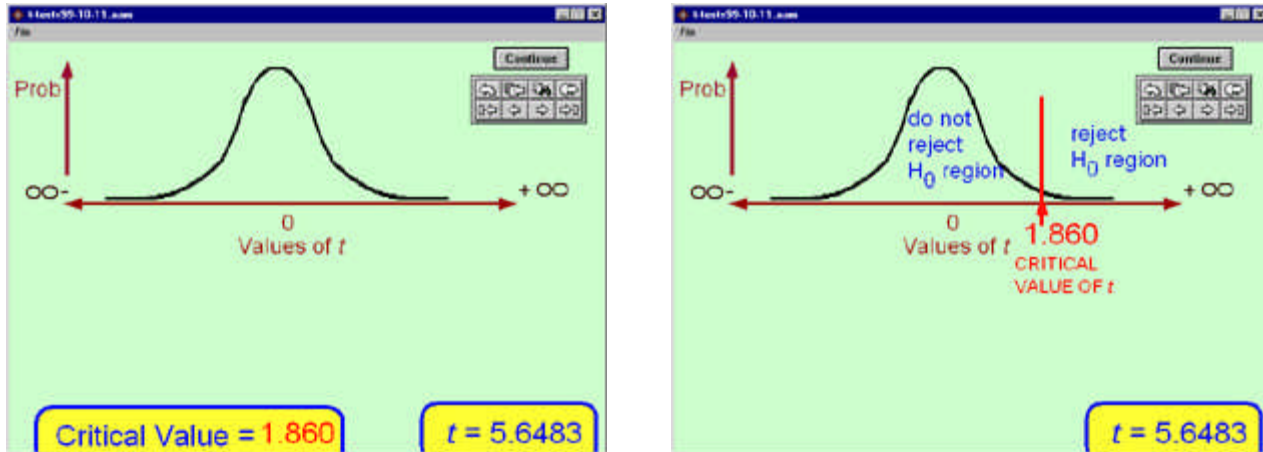


## Statistical Conclusion Validity

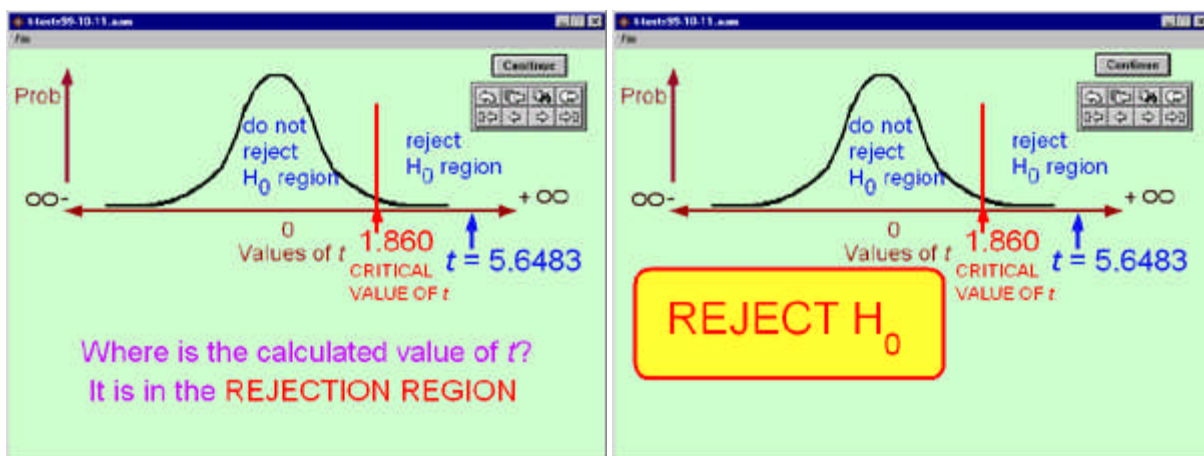


We choose alpha equal to .05. The degrees of freedom are equal to N minus 2. Therefore if we had 10 participants minus 2, the degrees of freedom are equal to 8. Is it one or two-tailed test? One-tailed. The t-test table tells us that for a one-tailed test with 8 degrees of freedom and alpha .05, **t critical is 1.860**.

## Sampling Distribution of t

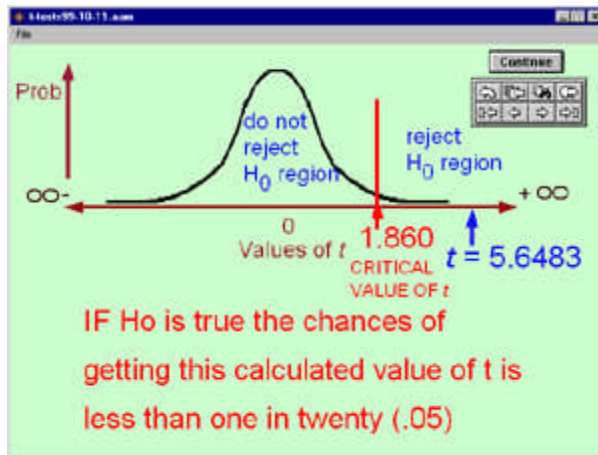


The above graphics show our sampling distribution of **t**. Notice that **t** is centered at zero, which is what H<sub>0</sub> is expecting. If **b** is zero, **t** ought to be zero. However due to chance, or some random error in our data, we wouldn't say **t** has to be exactly zero, but rather it should be near zero. Look at the sampling distribution shown here. If H<sub>0</sub> is true, the highest probabilities are for the values of **t** are very close to zero. The big bulge in the probability curve is right above zero and near zero. Out in the tails of the distribution the values of **t** are so far from zero that they have very small probabilities. Using our familiar logic then, we're going to place our **critical value** on the **t** number line far from zero and in the direction the scientist predicted. You can see that the value 1.86 is far out on the upper tail of the distribution and there's only a .05 probability of being at 1.86 or above. If H<sub>0</sub> is true, the probability of **t** being 1.86 or larger, is quite small, about a 1 in 20 chance.



Continuing with the same logic, we put our calculated value of  $t = 5.6483$  on the number line. The calculated value of **t** is out in the reject H<sub>0</sub> region, so we **reject the null hypothesis**.

[Go To Menu Map](#)



## The probability that H<sub>0</sub> is True

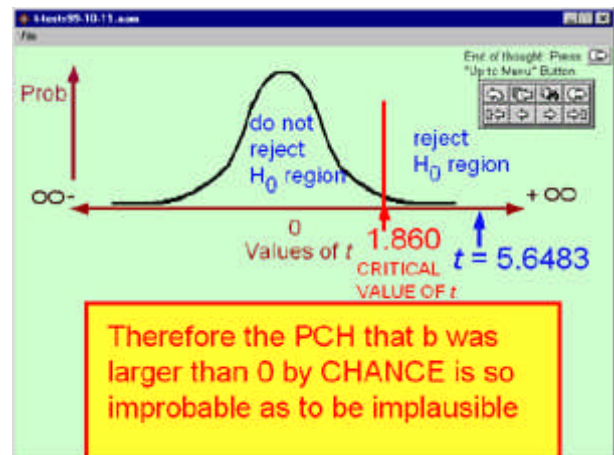
Just to repeat (because it's important to understand the rationale underlying this), if H<sub>0</sub> is true and  $t$  ought to be around zero, and if we've got just chance data, the probability we would get such chance data that would lead to a  $t$  this far from zero, is less than 1 in 20, or .05 or 5 %.

**Alpha is always the probability that we are wrong when we reject H<sub>0</sub>.**

## Plausibility of Chance

Back to the realm of science. Since H<sub>0</sub> has been tested in the realm of statistics and shown to be very improbable, (less than 1 chance in 20) then it seems that in the realm of science chance is pretty improbable, in fact chance is improbable enough to make it an implausible competing hypothesis.

[Go To Menu Map](#)



## The 4-Step Process

By now you should be familiar with the process that we follow in hypothesis testing. However, I will still round our our discussion of the t for b as a way of reviewing the big concepts involved. In step 1 we assume that the null hypothesis is true. Recall that the null hypothesis for this test is that  $b = 0$  therefore our calculated t will be equal to 0. We also assume that the data we collected is a normal population.

In step 2, we construct a random sample by collecting data on N number of participants. This random sampling helps assure that the data will represent the entire population.

Step 3 - We define the statistical formula for t. Here I have shown the general formula for t but it is analogous to the t for b formula.

**In Step 4, we determine the sampling distribution of t.** That means we can determine the probability of obtaining the value of t we calculated. Is it a highly likely or highly unlikely value if  $H_0$  is assumed to be true. Base on this probability we determine if we can rule out chance as a plausible competing hypothesis.

[Go To Menu Map](#)

