Practicing inference with two-sample t-tests (comparison of means) and regression

PAI 721

Prof. Hamersma

1. The Maxwell X Lab was enlisted to help a city increase payments for garbage pickup bills by experimenting with a more personalized bill. The study contained about 8,000 households. The control group paid on average $21.13 while those receiving a more personalized notice paid on average $25.59. They calculated the standard error of the difference to be $2.05.
2. What is the estimated effect of the intervention on average amount paid?

The difference in the means is the estimated effect: 25.59 – 21.13 = $4.46

1. Construct a 99% confidence interval for the effect of the intervention.

This sample is large, so I can use t\* = 2.58 (the same as z\* for 99% confidence)

The estimated effect is 4.46, and the (given) standard error of the difference is 2.05.

This means I have everything I need for the confidence interval:

4.46 +/- 2.58\*2.05 🡪 ( -0.829, 9.749)

1. Is the effect of the intervention statistically significant at the 99% level? How about the 95% level?

Statistically significant means statistically significantly different from zero – so my null hypothesis is an effect of zero (i.e. no effect).

Since zero is in the confidence interval in (b), I cannot reject the null of zero. This means the estimated effect is not statistically significant at the 99% level.

To check the 95% level, I could make a new confidence interval with t\* = 1.96 or do the formal test:

H0: μ1 – μ2 = 0

Ha: μ1 – μ2 ≠ 0

t = {(4.46) – 0} / 2.05 = 2.176 Since |2.176| > 1.96, I reject the null of no effect at the 95% confidence level, so the estimated effect \*is\* statistically significant at the 95% level.

1. Answer the following questions about each of the (fake) scenarios below.
   1. What is the estimated relationship between X and Y in a sentence?
   2. Construct a 95% confidence interval for the relationship.
   3. Is the relationship statistically significant at the 95% level? How about the 90% level?
2. Y is annual dental expenses in $, X is # times brushing teeth per week, n = 52

|  |  |  |
| --- | --- | --- |
|  | Estimate | Std error of estimate |
| Constant | 468.1 | 98.4 |
| Brushing teeth | -30.4 | 16.1 |

Each additional time brushing teeth per week is associated with a reduction of $30.40 in dental expenses for the year.

With n = 52, we have df = 50. For 95% confidence, this means t\* = 2.009

-30.4 +/- (2.009\*16.1) 🡪 (-62.745, 1.945)

t = -30.4/16.1 = 1.888

Since |1.888|<2.009, the relationship is NOT statistically significant at the 95% level. (You can also see that 0 is in the confidence interval, which would already be enough to conclude this.)

The 90% confidence choice of t\* would be t\* = 1.676. In this case, since |1.888|>1.676, we CAN reject the null hypothesis of no effect, i.e. the estimated effect is statistically significant at the 90% level.

ii. Y is city $ collected per day from parking tickets, X is the number of parking spots in the city, n = 1200

|  |  |  |
| --- | --- | --- |
|  | Estimate | Std error of estimate |
| Constant | 1.12 | 5.10 |
| Parking Spots | 4.57 | 3.14 |

An increase in one additional parking spot is associated with $4.57 in additional revenue from parking tickets.

Since n > 1000, I am safe to use t\* = 1.96 for 95% confidence (and t\* = 1.645 for 90% confidence)

4.57 +/- (1.96\*3.14) 🡪 (-1.584, 10.724)

t = 4.57/3.14 = 1.455

This t is below both critical values, so it is not statistically significant at 95% or 90% confidence.

iii. Y is a county’s % voting for President Trump, X is the county’s size (in 1000s), n = 2000 counties

|  |  |  |
| --- | --- | --- |
|  | Estimate | Std error of estimate |
| Constant | 85.25 | 8.9 |
| County size | -0.12 | 0.04 |

*Note, there is no way this should be linear -- just go with it!*

Each additional 1000 people in a county is associated with 0.12% fewer votes for Trump.

Since n > 1000, I am safe to use t\* = 1.96 for 95% confidence (and t\* = 1.645 for 90% confidence)

-0.12 +/- (1.96\*0.04) 🡪 (-0.198, -0.042)

t = -0.12/0.04 = - 3.000

This t is above both critical values in absolute value, so it is statistically significant at 95% and 90% confidence.

iv. Y is a person’s hourly wage, X is hours spent in a public job training program, n = 82

|  |  |  |
| --- | --- | --- |
|  | Estimate | Std error of estimate |
| Constant | 8.10 | 1.4 |
| Job training | 0.13 | 0.06 |

Each hour spent in job training is associated with a $0.13 higher hourly wage.

Since n = 82, we use t\* with 80 df. For 95% confidence, this is t\* = 1.990

0.13 + (1.990\*0.06) 🡪 (0.012, 0.249)

t = 0.13/0.06 = 2.167

This t is above the critical t\* of 1.990 for 95% confidence. This means it is also above 90% confidence (you can confirm by noting that t\* = 1.664 for 90% confidence with 80 df). So we can reject the null hypothesis of no relationship at either of the confidence levels; the estimate is statistically significant.