

Unit 3: Inference for Categorical and Numerical Data

3. Paired Data and the t-distribution (Chapter 4.2)

11/07/2016

Quiz 3 - Difference of Proportions and T-values

Recap from last time

1. You can use the Normal approximation for the difference of two proportions
2. The margin of error is not just the sum of the margin of errors for each proportion
3. If you think two proportions come from the same population, you can use a pooled estimate

Recap from last time

1. When our samples are too small, we shouldn't use the Normal distribution
2. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)
3. All of our statistical theory still holds, we are just plugging in different distributions

Key ideas

1. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values
2. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)
3. If you think two proportions come from the same population, you can use a pooled estimate

Key ideas

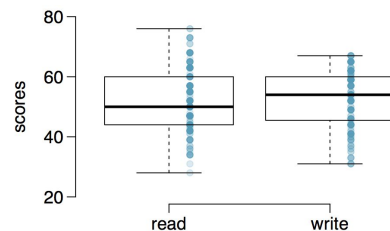
1. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values
2. The margin of error is not just the sum of the margin of errors for each proportion
3. If you think two proportions come from the same population, you can use a pooled estimate

An example of paired data



200 observations were randomly sampled from the HS&B survey. The same students took a reading and writing test, here are their scores.

Does there appear to be a difference between the average reading and writing test score?



An example of paired data



Are the reading and writing scores of each student independent of each other?

(a) Yes (b) No

	id	read	write
	1	70	57
	2	86	44
	3	141	63
	4	172	47
	⋮	⋮	⋮
	200	137	63

An example of paired data



Are the reading and writing scores of each student independent of each other?

(a) Yes (b) No

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

Analyzing paired data

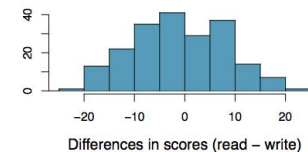
When two sets of data are not-independent, we call them *paired*.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

$$\text{diff} = \text{read} - \text{write}$$

Note: It's important that we always subtract using a consistent order.

	id	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
⋮	⋮	⋮	⋮	⋮
200	137	63	65	-2



Parameter and point estimate

Parameter of interest: Average difference between the reading and writing scores of all high school students.

$$\mu_{\text{diff}}$$

Point estimate: Average difference between the reading and writing scores of sampled high school students.

$$\bar{x}_{\text{diff}}$$

Setting up the Hypotheses

If there were no difference between scores on reading and writing exams, what difference would you expect on average?

0

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

H0: There is no difference between the average reading and writing score.

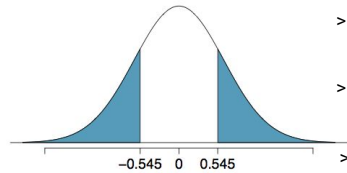
$$\mu_{\text{diff}} = 0$$

HA: There is a difference between the average reading and writing score.

$$\mu_{\text{diff}} \neq 0$$

Calculating the test-statistics and p-values

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these suggest a difference between the average scores on the two exams at $\alpha = 0.05$?



```
> t <- (-.545 - 0) / (8.887 / sqrt(200))  
= -.87  
  
> pt(-.87, df = 199)  
= .1927  
  
> p_val <- .1949 * 2  
= .3898
```

Since $p\text{-value} > 0.05$, fail to reject, the data do not provide convincing evidence of a difference between the average reading and writing scores.

Interpreting the p-value

Which of the following is the correct interpretation of the p-value?

- (a) Probability that the average scores on the two exams are equal.
- (b) Probability that the average scores on the two exams are different.
- (c) Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.
- (d) Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Interpreting the p-value

Which of the following is the correct interpretation of the p-value?

- (a) Probability that the average scores on the two exams are equal.
- (b) Probability that the average scores on the two exams are different.
- (c) Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.**
- (d) Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Hypothesis testing and Confidence Intervals

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

- (a) Yes
- (b) No
- (c) Cannot tell from the information given

Hypothesis testing and Confidence Intervals

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

- | | | |
|--|--|----------------------------------|
| (a) Yes | $-0.545 \pm 1.96 \frac{8.887}{\sqrt{200}}$ | $= -0.545 \pm 1.96 \times 0.628$ |
| (b) No | | $= -0.545 \pm 1.23$ |
| (c) Cannot tell from the information given | | $= (-1.775, 0.685)$ |

Key ideas

1. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values
2. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)
3. If you think two proportions come from the same population, you can use a pooled estimate