Unit 3: Inference for Categorical and Numerical Data

3. Paired Data and the t-distribution

(Chapter 4.2)

11/06/2017

Quiz 3 - Difference of Proportions and T-values

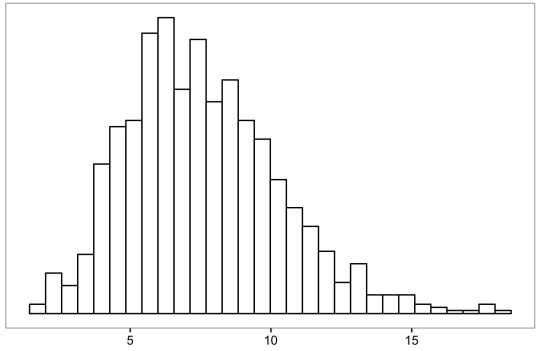
Recap from last time

1. When our samples are too small, we shouldn't use the Normal distribution
2. We should use the *t* distribution to make up for uncertainty in our sample statistics
3. All of our statistical theory still holds, we are just plugging in different distributions

Key ideas

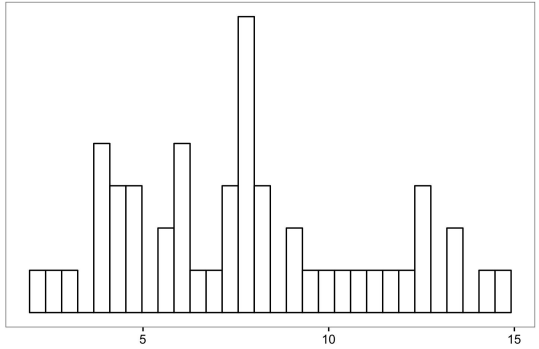
1. We use the t-distribution because of the difficulty of estimating standard deviation for small samples
2. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values
3. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)

A reminder about the Central Limit Theoreom



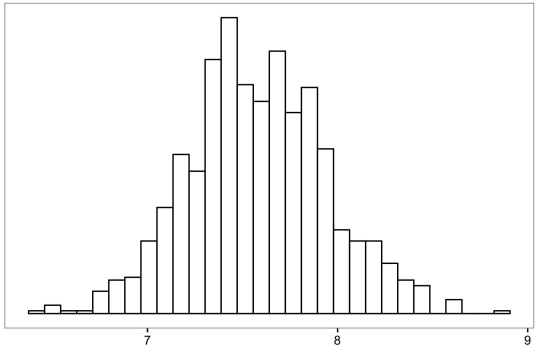
When I draw **independent samples** from the population, as sample size **approaches infinity,** the distribution of means approaches normality

But what is it's SD?



Take the mean,

Repeat many times…

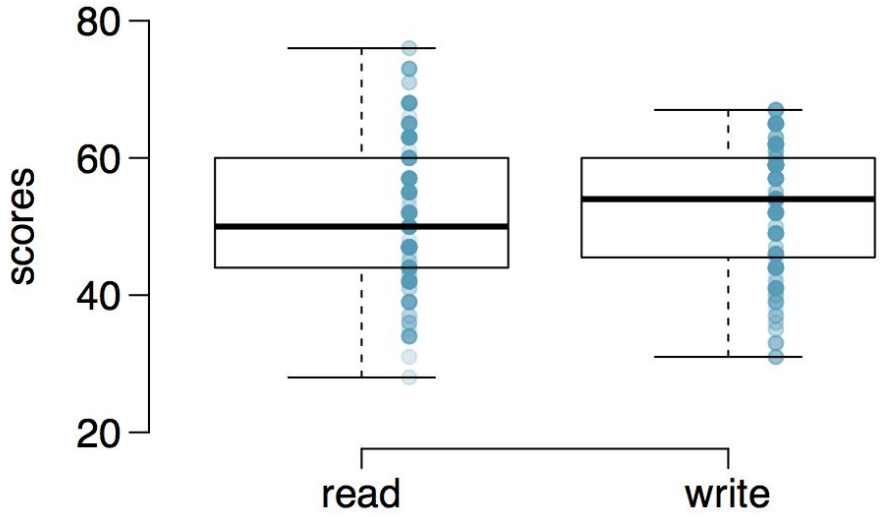


An example of paired data

HIGH SCHOOL & Beyond
providing a lifetime of insight

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

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

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

(a) Yes (b) No

An example of paired data

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

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

(a) Yes **(b) No**

Analyzing paired data

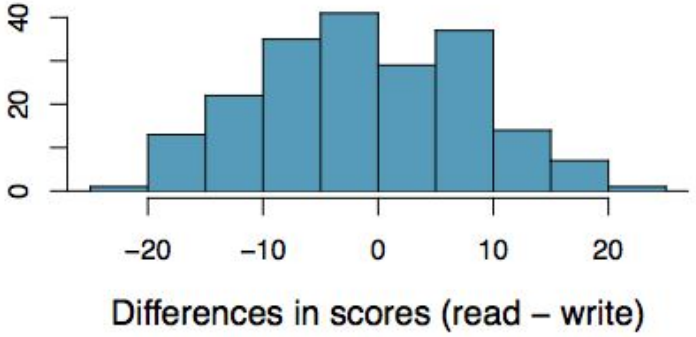
When two sets of data are not-independent, we call them ***paired***.

To analyze paired data, we first compute the difference between in outcomes of each pair of observations.

diff = read - 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



What counts as paired?

1. Verbal SAT and Math SAT from the same person
2. Spouse 1's height and Spouse 2's height
3. Parental anxiety score and child's anxiety score
4. Traffic flow at the same stop on Friday the 6th and Friday the 13th.
5. SAT scores at Harvard and Yale
6. Control group blood pressure and Treatment group blood pressure

Two sets of data are paired if each data point in the first set has one clear "partner" in the second data set.

Parameter and point estimate

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

$$\mu_{diff}$$

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

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

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.

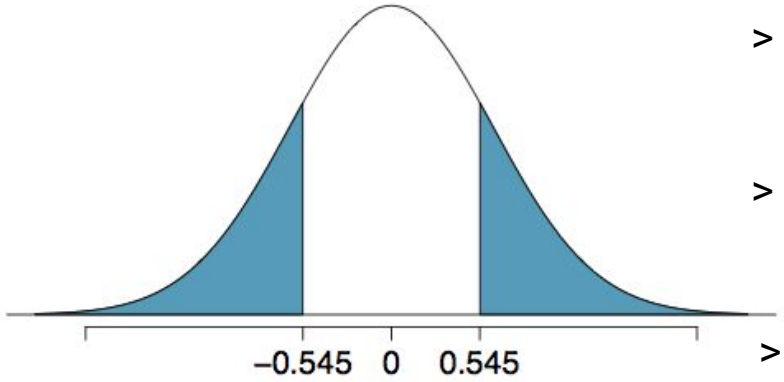
$$\mu_{diff} = 0$$

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

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

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-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

(b) No

(c) Cannot tell from the information given

$$
\begin{aligned}
-0.545 \pm 1.96\frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.96 \times 0.628 \\
&= -0.545 \pm 1.23 \\
&= (-1.775, 0.685)
\end{aligned}
$$

Key ideas

1. We use the t-distribution because of the difficulty of estimating standard deviation for small samples
2. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values
3. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)