ST 516: Foundations of Data Analytics

Inference for means in paired data

Moving beyond one sample of data

Last week we looked at methods of inference for a central value in a single population.

This week we extend those ideas to inference about the centers of more than one population.

This lecture looks at a special case where our current methods are appropriate.

Next lecture introduces a new method for means of two populations.

Motivating example

Sleuth case 2.2: The Schizophrenia case study

Researchers are interested in whether there are physiological differences between people affected by schizophrenia and those who aren't.

They recruit 15 identical twins, where one is affected by schizophrenia and other is not.

The measure the left hippocampus (an area of the brain) volume for all 30 people.

Questions of interest:

- Is there a difference in brain volume between people affected by schizophrenia and those unaffected?
- How big is the difference?

Paired data

Two samples are **paired**, if each observation in one sample, has a connection to one, and only one, observation in the other sample.

The Schizophrenia case study is an example of paired data.

A brain volume for a person in the *Unaffected* group is naturally paired with the brain volume for their twin in the *Affected* group.

A common approach to paired data is to analyse the differences between the members of each pair.

Analysis of paired data

Scientific questions:

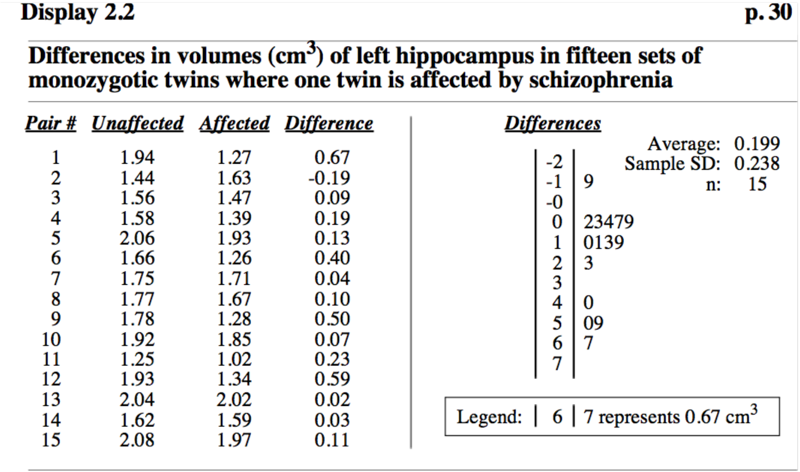
- Is there a difference in brain volume between people affected by schizophrenia and those unaffected?
- How big is the difference?

Let's consider the differences: volume from twin unaffected by Schizophrenia – volume from twin affected by Schizophrenia

We can consider these as a sample of differences from some population of differences, and ask questions about this population.

Statistical questions:

- Is the *mean difference* in volume different from zero? Equivalenty, is the difference in mean volumes different from zero?
- What's a likely range for the *mean difference* in volume?



Display 2.2 **p. 30**

Differences in volumes (cm^3) of left hippocampus in fifteen sets of monozygotic twins where one twin is affected by schizophrenia

Pair #	*Unaffected*	*Affected*	*Difference*
1	1.94	1.27	0.67
2	1.44	1.63	-0.19
3	1.56	1.47	0.09
4	1.58	1.39	0.19
5	2.06	1.93	0.13
6	1.66	1.26	0.40
7	1.75	1.71	0.04
8	1.77	1.67	0.10
9	1.78	1.28	0.50
10	1.92	1.85	0.07
11	1.25	1.02	0.23
12	1.93	1.34	0.59
13	2.04	2.02	0.02
14	1.62	1.59	0.03
15	2.08	1.97	0.11

Differences

Average:	0.199
Sample SD:	0.238
n:	15

```
| -2 |
| -1 | 9
| -0 |
|  0 | 23479
|  1 | 0139
|  2 | 3
|  3 |
|  4 | 0
|  5 | 09
|  6 | 7
|  7 |
```

Legend: | 6 | 7 represents 0.67 cm^3

Some notation

Let $i = 1, \ldots, n$ index the pairs, so for example, $i = 1$ refers to the 1st pair.

For the Schizophrenia case study, $n = 15$, there are 15 pairs.

Let,
y_i $i = 1, \ldots, n$ be the observations in one sample and
z_i, $i = 1, \ldots, n$ be the observations in the other sample.

The differences are $x_i = y_i - z_i$, $i = 1, \ldots, n$.

We have a sample of differences and our questions are about the population mean of the differences.

We can therefore apply the t-based procedures from last week to this sample of differences.

Example: Calculations

```r
library(Sleuth3)
diffs <- case0202$Unaffected - case0202$Affected
t.test(diffs)
```

```
##
##  One Sample t-test
##
## data:  diffs
## t = 3.2289, df = 14, p-value = 0.006062
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.0667041 0.3306292
## sample estimates:
## mean of x
## 0.1986667
```

```r
# alternative syntax
t.test(case0202$Unaffected, case0202$Affected, paired = TRUE)
```

Interpretation: just remember our inferences are now about the mean difference, Unaffected − Affected.

Example: Statistical Summary

There is convincing evidence the mean difference in brain volume between a person affected by schizophrenia and their twin unaffected by schizophrenia is not zero (paired t-test, p-value = 0.006, n = 15).

It is estimated the hippocampus in the twin affected by schizophrenia is 0.2 cm^3 smaller on average than in the twin unaffected by schizophrenia.

With 95% confidence the mean difference is between 0.07 cm^3 and 0.33 cm^3.

Why collect paired data?

Think about the variation in brain volume coming from two sources:

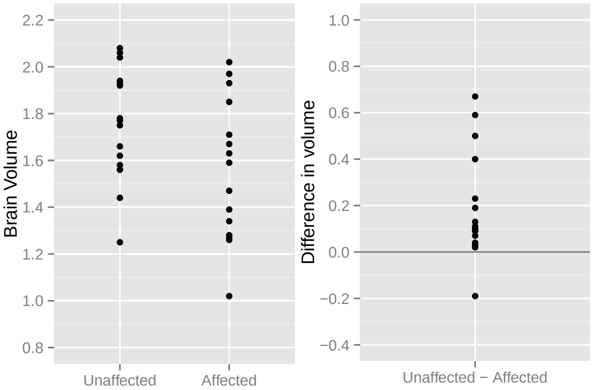
- variability attributable to the difference between being affected by schizophrenia or not, and
- variability attributable to other factors (e.g. environment, genetics, upbringing, . . .)

Twins would be expected to share a lot of the influences on brain volume. By taking differences within twins we are “subtracting” away the variation due to these variables. The variability left should only be attributable to the difference between being affected by schizophrenia or not.

Smaller variability leads to more precise estimates.

In general, differences within a pair will be less variable than differences between pairs. Estimates of differences in mean will often be more precise from paired data.

Why collect paired data?



Identifying paired data

Not all data that involves two groups are paired.

Look for something that links an observation in one group to an observation in the other.

Ask yourself: if I had one observation, could I easily (and uniquely) identify another observation in the other sample that it “pairs” with?

One giveaway of non-paired data, is if the two samples are of different sizes.

Another clue to paired data, is when only one sample of units is taken, but two measurements are made on each unit.

Paired or not?

We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock.

Paired or not?

We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock. **Not paired**

We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.

Paired or not?

We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock. **Not paired**

We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items. **Paired**

A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

Paired or not?

We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock. **Not paired**

We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items. **Paired**

A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.**Not paired**

To evaluate the general fitness level of OSU students, we randomly sample 50 students and get two measurements: their heart rate after 10 minutes of jump rope, and their heart rate again 2 minutes later.

Paired or not?

We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days for Intel's stock and another random sample of 50 days for Southwest's stock. **Not paired**

We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items. **Paired**

A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.**Not paired**

To evaluate the general fitness level of OSU students, we randomly sample 50 students and get two measurements: their heart rate after 10 minutes of jump rope, and their heart rate again 2 minutes later. **Paired**

Coming up next: two sample t-test

We'll see a new method for questions about means when the two samples are not paired.