# Chapter Twenty One: Paired Data

A teacher is testing out a new program to help high school students learn math. A random sample of 40 students was selected. The teacher wants to determine if the new program helps the students perform better on exams. Consider the following two scenarios:

1. The students were randomly divided into two groups, each of size 20. One group of students were enrolled in the new program, and the other group was not. All students took the same exam.
2. All 40 students were enrolled in the new program. Each student completed the same math exam before entering the program and another (very similar) exam after completing the program.

What is different about the two studies?

How would we proceed to analyze each of these studies?

For the first scenario:

- we would compare the mean exam scores of the two groups (those who were enrolled in the program and those that were not)
- This is a difference in means situation Ch. 20

For the second scenario:

- we would compare the mean change in scores of the two exams for each student. (exam2 - exam1 score for each student).
- This is a paired data situation Ch. 21

## Paired Data

Sometimes we collect data from two groups that are dependent on one another. We refer to this as paired data. For example:

- Two measurements are taken from the same subject
- measurements from two different subjects taken at the same time or location.

The general data structure of paired data is as follows:

| Pair | Group 1 | Group 2 | Differences |
|---|---|---|---|
| 1 | $y_{11}$ | $y_{21}$ | $d_1 = y_{11} - y_{21}$ |
| 2 | $y_{12}$ | $y_{22}$ | $d_2 = y_{12} - y_{22}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $y_{1n}$ | $y_{2n}$ | $d_n = y_{1n} - y_{2n}$ |
| Means | $\bar{y}_1$ | $\bar{y}_2$ | $\bar{d}$ |
| SDs | $s_1$ | $s_2$ | $s_d$ |

We define $\bar{d}$ and $s_d$ as follows

$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \bar{y}_1 - \bar{y}_2$ $\quad s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}} \neq s_1 - s_2$

Now let's look at a few examples of scenarios of experiments and determine whether it is paired data:

1. We would like to determine if students taking an ACT prep course will score better than students not taking the course. A random sample of 25 students was chosen who took the course and a random sample of another 25 students was chosen who did not take the course. At the end of the prep course, both groups were given the ACT. is there paired data, two independent sample data, or one sample data?

   two independent groups sample data

2. We would like to determine if students can improve their ACT score by taking a prep course. A random sample of 25 students was chosen. They first took the ACT test. Then they spent 6 weeks taking the prep course. At the end of the 6 weeks, they took the ACT test again. Is there paired data, two independent sample data, or one sample data?

   paired data

3. We would like to determine ACT scores for students that take a prep course. A random sample of 50 students was selected. All 50 students took the prep course. At the end of the prep course, students took the ACT test. Is there paired data, two independent sample data, or one sample data?

   one sample data

### Notation for Paired Data

$\mu_D$: mean difference between two values in populations
$n$: # of differences in sample (# of pairs)
$\bar{d}$: mean difference between two values in a sample
$s_D$: standard deviation of differences between two values in the sample

## Inference for Paired Data

Like the last two chapters we will consider two types of inference: Confidence Intervals and Hypothesis tests. The process for both of these types of inferences are very similar to previous chapters.

### Assumptions for CIs and Hypothesis Tests

The following assumptions need to be met in order to construct a confidence interval or perform a hypothesis test for paired data.

i. **Randomization Condition:** samples must be randomly selected

ii. **10% Condition:** sample size is less than 10% of population.

iii. **Nearly Normal Condition:**
1) population distribution is already Normal
2) sample size is sufficiently large.
   - ↳ Symmetric (but not normal): $n = 10$ or larger
   - ↳ Skewed: $n = 50$ or larger
   - ↳ very skewed: $n = 100$ or larger

## Confidence Interval

If the assumptiosn are met, then the confidence interval for paired data can be computed using the formulat below. Not how similar it is for CIs for one sample mean:

**Paired Data**

$$\bar{d} \pm t^* \frac{s_D}{\sqrt{n}}$$

**One Sample Mean**

$$\bar{y} \pm t^* \frac{s}{\sqrt{n}}$$

In both cases, $t^*$ has a $t$ distribution with $n-1$ degrees of freedom.

## Hypothesis Test Procedure

**Step 1: Hypotheses**

- **Null Hypothesis**: population mean difference is 0

  $H_0: \mu_D = 0$

- **Alternative Hypothesis**: population mean difference is diff. than 0

  1) $H_A: \mu_D < 0$
  2) $H_A: \mu_D > 0$
  3) $H_A: \mu_D \neq 0$

**Step 2: Assumptions** See previous pages.

**Step 3: Test Statistic**

Calculate the test statistic below (or find with JMP). Note the similarity to the Chapter 18 one saple mean formula.

**Paired Data**:

$$t = \frac{\bar{d}}{s_D/\sqrt{n}}$$

**One Sample Mean**:

$$t = \frac{\bar{y}}{s/\sqrt{n}}$$

both follow t distribution w/ n-1 df

**Step 4: Find P-value** The p-value is found using a t-distribtion with n-1 degrees of freedom. Remember, the way the p-value is found depends on the alternative hypothesis.

- $H_A: \mu_D < 0$ → p-value is area less than t.
- $H_A: \mu_D > 0$ → p-value is area greater than t
- $H_A: \mu_D \neq 0$ → p-value is the area less than -|t| plus area greater than |t|

**Step 5: List your Decision**

| P-value | Evidence (against Ho) |
|---|---|
| Greater than .10 | Little to no evidence |
| Between .05 and .10 | Weak evidence |
| Between .01 and .05 | Moderate Evidence |
| Less than .01 | Strong evidence |

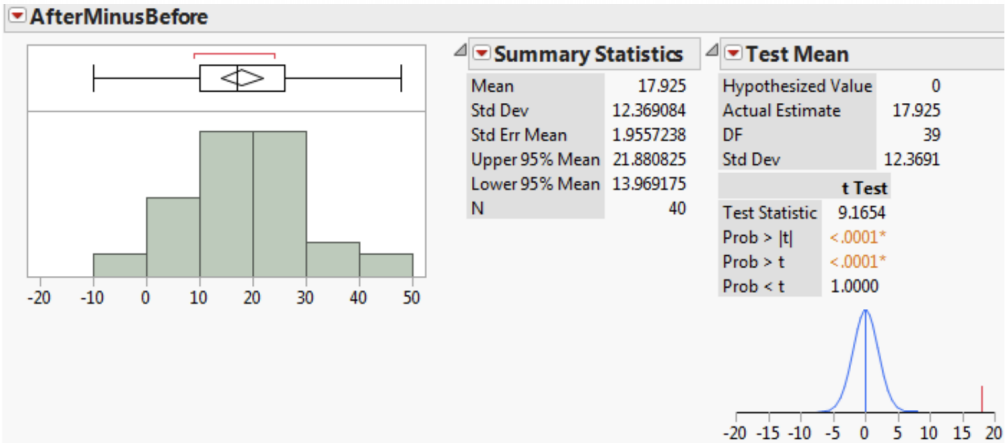
**Step 6: Conclusion**

Make a statement about $\mu_D$ (the population difference in means with pairing) given the information from the hypothesis test.

Make sure to include:

- parameter
- context
- whether or not there is evidence against the null.

A teacher is testing out a new program to help high school students learn math. A random sample of 40 students was selected. The teacher wants to determine if the new program helps the students perform better on exams. Each student completed the same math exam before entering the program and another (very similar) exam after completing the program. Construct a 95% confidence interval and conduct a hypothesis test to see if there is a difference in score before and after completing the program, and if so how much.




Check Conditions:
1) students randomly selected ✓
2) probably more than 10*40 students ✓
3) distribution is roughly symmetric, sample size suffices.

CI: from JMP
(13.97, 21.88)

We are 95% confident the true mean difference in scores after and before the program falls between (13.97, 21.88) points

Step 1: $H_0: \mu_D = 0$
$H_a: \mu_D \neq 0$

Step 2: See above

Step 3: 9.1654

Step 4: p-value = .0001
single prob>|t| = .0001
from JMP

Step 5: strong evidence against the null

Step 6:
We have strong evidence to suggest that the true mean difference in scores before and after the program is not equal to 0.