SECTION 4: MATCHED PAIRS DESIGNS

In this section, you will again compare two sets of quantitative observations but with one key difference in how the data were collected. You will see advantages to collecting data with a "paired" design, and then you will investigate how to make the corresponding changes in the appropriate analyses.

Investigation 4.8: Speed It Up

Student researchers (Coutin & Heffler, 2021) wanted to know whether listening to up-tempo music causes college students to tend to type faster. To collect their data (number of words typed correctly in one minute), the students planned to use the 60-second Easy-Text typing test (TypingTest.com). They recruited 34 college students from groups they were associated with on campus (e.g., athletic teams, musical groups). For the up-tempo music they selected *Overture to Candide* performed by the London Symphony Orchestra.

Let $\mu_{nomusic}$ represent the population mean typing speed without the music and μ_{nusic} the population mean typing seed with the music.

(a) State the student researchers' null and alternative hypotheses in symbols and in words.

Design

(b) Describe a *completely randomized design* for conducting this study. Identify the experimental units and variables of interest. Classify the variables as explanatory and response, as well as quantitative or categorical.

Design:

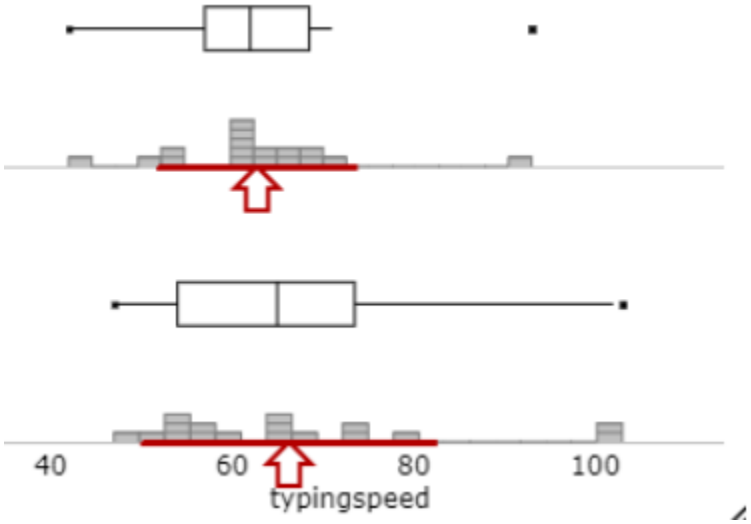
Experimental units:

Explanatory variable: Type:

Response variable: Type:

(c) Identify some precautions you would take in carrying out this design.

Consider the following results for the number of words per minute by each participant.



(d) Do you see much of a difference between the two distributions? What is the observed difference in the sample means? Do you think this difference in sample means will be statistically significant? Explain.

(e) Is a two-sample *t*-test likely valid for these data? Explain.

The one-sided p-value from a two-sample *t*-test is 0.2299, failing to provide convincing evidence that the average typing speed is faster with (this) music than the average typing speed without music. But maybe there is a genuine difference and our sample sizes are just too small to detect it.

(f) Suppose everything else remained the same but the sample sizes had been 34 in each group. How would that impact the p-value? By a lot or by a little? Would the difference now be statistically significant?

Increasing the sample size helps reduce the "random chance" variation in our statistic, but does not reduce the person-to-person variation in typing speeds. When the person-to-person variation is large, it will still be difficult for us to detect the underlying treatment difference even if one exists. If the actual difference in typing speeds is 4 words per minute on average, we would need sample sizes on the order of 100 participants, in each group, to have at least 80% power.

(g) Sketch curves illustrating this power calculation.

But there is another way we can improve the design of this study for detecting differences in typing speeds without using any additional people!

(h) How could you modify the experimental design to give you a better chance of detecting a difference between the typing speeds between the two conditions if one exists? Be sure to explain why you believe this new design will be advantageous in detecting a difference between the two conditions.

(i) How will randomness be used in this new study design and why is that important?

(j) In your new study design, what response variable will you measure on each individual? Will it be quantitative or categorical?

Definition: In a ***paired design***, rather than splitting the observational units into two distinct groups, they are paired in a way where we expect the observations within a pair to be more similar to each other than to observations in other pairs. This can explain some variation in the response variable.
For example, we can have each person take the typing test both with and without music. This allows us to compare the two typing speeds for each individual to each other which should be very similar apart from the music, and to account for variation in typing speeds across individuals.

(k) Should the participants be given an identical typing test both times? Explain.

(l) The students originally wanted to compare music students to athletes. Explain how you could create a paired design using information on whether the participant was a music student or an athlete. Do you think this design will be as effective as the above design? Explain.

Practice Problem 4.8A

Suppose that a baseball manager wants to study whether a player can run from second base to home plate more quickly by taking a wide angle around third base or a narrow angle. Forty players are available to use as subjects in an experiment.

(a) Suggest a completely randomized design and a paired design for this research question and explain why a paired design is likely to be more effective.

(b) Suppose the players arrive for the study at different times. The manager decides to pair the first two arrivers and have them each do a different angle. The manager continues pairing the next two players as they arrive, etc. Is this a paired design? Is the paired design likely to be more effective than the completely randomized design in this case?

Practice Problem 4.8B

For each of the following research study designs, indicate whether the data collection plan will result in two independent samples (completely randomized design) or "dependent" samples (matched-pairs design).

(a) A farmer wants to see whether referring to cows by name increases their milk production. He selects half of his cows at random, gives them names, and frequently calls them by name. The other half of his cows he does not call by name. Then he measured the milk production of each cow over one week.

(b) A farmer wants to know whether hand-milking or machine-milking tends to produce more milk from cows. He examines records of how much milk the cows have produced in the past, and order them from most to least productive. For the top two milk producers, randomly assign one to hand-milking and the other to machine-milking. Do the same for the next two and the next two and so on.

(c) You wonder whether students at your school tend to drive newer cars than faculty at your school. You take a random sample of 20 students and a random sample of 20 faculty members, and ask each person how old their car is.

(d) To investigate whether knee surgery is effective, you randomly assign half of the subjects to receive knee surgery and the other half to undergo a "placebo" operation.

(e) To investigate the effectiveness of an online language study program, participants were assigned to enroll in a six-week summer session, after which their language skills were assessed, and then to spend six-weeks using an online program (Duolingo), after which their language skills were assessed.

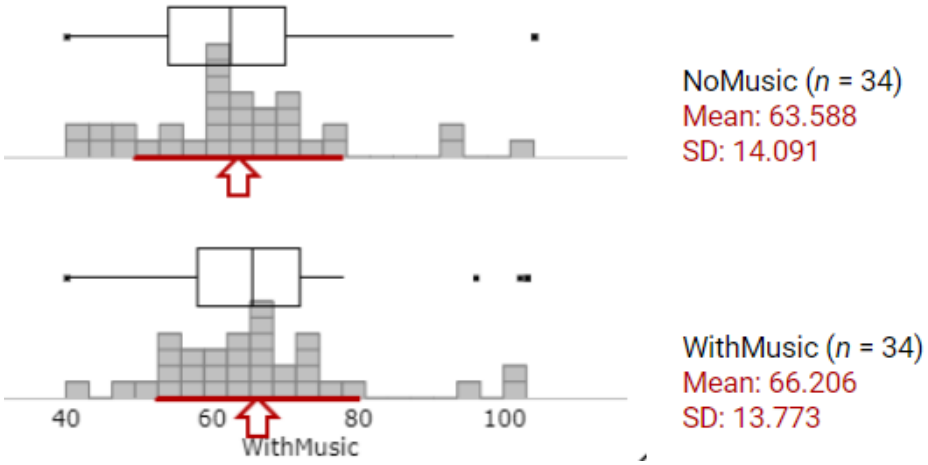
Investigation 4.9: Speed It Up (cont.)

Recall that student researchers wanted to compare the mean typing speed with and without up-tempo music.

H_0: $\mu_{\text{no music}} - \mu_{\text{music}} = 0$ (no difference in the long run average speed)

H_a: $\mu_{\text{no music}} - \mu_{\text{music}} < 0$ (on average typing speed is faster with up-tempo music)

Below are the results from the students' matched pairs design.

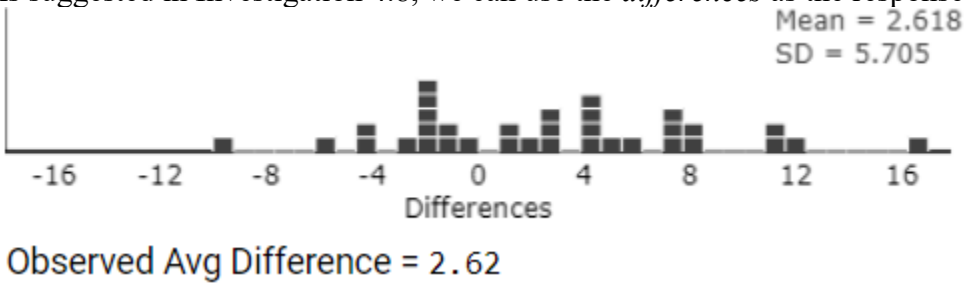


(a) Compare these results to the results in Investigation 4.8. What has changed? Will these changes impact the p-value? If so, how?

(b) Explain why we *can't* do a randomization test or a two-sample *t*-test with these data.

Key Idea: When the data are *paired* (e.g., repeat observations on the same observational unit) we should not treat the two samples as independent. This ignores the information that two measurements were taken for each observational unit (we couldn't mix up the values in the second column without altering the information in the data).

As suggested in Investigation 4.8, we can use the *differences* as the response variable.



(c) Summarize what you learn about the distribution of differences.

(d) How does the mean of the differences ($\bar{x}_{diff}$) compare to the difference in means ($\bar{x}_{music} - \bar{x}_{nomusic}$)?

(e) How does the standard deviation of the differences compare to the standard deviation of the original typing speeds? What does this tell you about the effectiveness of the pairing in this context?

(f) Define an appropriate parameter for investigating whether comparing typing speeds with and without music. State a null and an alternative hypothesis about this parameter.

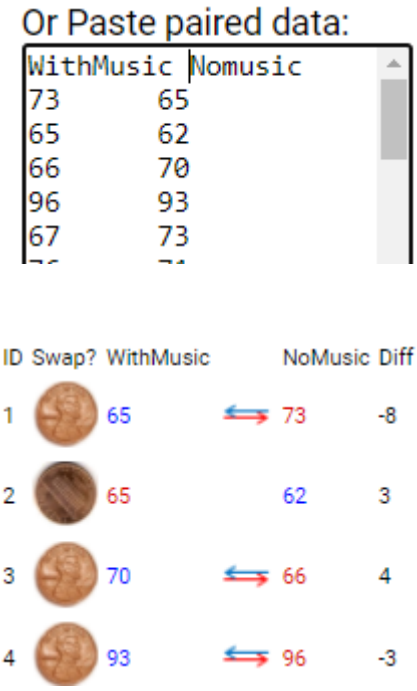
It actually doesn't matter whether we use the "difference in means" or the "mean difference" as our statistic/parameter. What is important is how we estimate the chance variation in that statistic, assuming the null hypothesis is true.

Simulation

(g) Outline (pseudo-code) how you could use a coin to simulate a randomization test for paired data to compare the two sets of measurements to assess how unusual it is for the average difference in typing speeds to be at least this extreme just by chance. Keep in mind that you want the simulation to mimic the randomization process used in the study design, assuming the presence/absence of music does not affect typing speed.

Copy and paste the original raw data (TypingMusic.txt) into the Matched Pairs Randomization applet:

- View the data window, with one column for the speeds with music and a second column for the speeds without music (each row is one person). You can also include an initial column of identifiers (e.g., student IDs or initials).The dotplots should then show both sets of data, connecting the paired observations, and their differences.
- Check the **Randomize** box and press **Randomize**.
  For each pair, the applet will virtually "flip a coin" and if it lands heads, the two observations for that person will change positions. The new dotplots and the new set of differences for these rearranged values will be displayed. The mean of these differences will appear in the bottom dotplot.
- Uncheck **Animate**.
- Press **Randomize** four more times to get a sense of the variability in the results from repetition to repetition.
- Change the number of randomizations from 1 to 995 (for a total of 1000) and press **Randomize**.



(h) Explain what distribution is being displayed in the bottom dotplot (what order of subtraction did the applet use?). What do you notice about the high outliers (fastest typers) in each condition?

(i) Where is the distribution of the average differences centered? Why should you expect that?

(j) How surprising does our observed value for the mean difference appear to be, under the simulation's assumption that presence/absence of up-tempo music does not affect typing speed?

(k) Use the applet to determine the proportion of simulated Average Differences that are more extreme than what we observed, and report the empirical p-value. [*Hint*: Be sure to consider the alternative hypothesis when deciding what to consider as "more extreme."] What conclusion will you come to based on this p-value? Can you draw a cause-and-effect conclusion? For what population?

Mathematical Model

(l) Does the randomization distribution appear that it would be well modeled by a normal distribution? If you change the **Statistic** from Avg Difference to ***t*-statistic**, do the standardized statistics appear well-modeled by a *t*-distribution?

Definitions: A paired *t*-test standardizes the mean of the *differences* from a matched-pairs design.

$$t = (\bar{x}_d - 0)/(s_d / \sqrt{n_d}),$$

where $\bar{x}_d$ is the sample mean of differences and s_d is the sample standard deviation of the differences. The test statistic above assumes the hypothesized difference is zero, but this can be changed.

Technical conditions: When the distribution of differences is normally distributed or the sample size is large (e.g., $n \geq 30$ pairs of observations), this *t*-statistic is well modeled by a *t*-distribution with $n - 1$ degrees of freedom.

A paired *t*-confidence interval for μ_d has the form $\bar{x}_d \pm t^*_{n-1}(s_d / \sqrt{n_d})$

Note: These are a special case of the *one-sample t-procedures* that can be applied to a single sample of quantitative data (see Investigation 2.5). In this case, variable of interest is the difference in the quantitative response for each observational unit pair.

(m) Calculate and interpret the value of this test statistic, by hand, based on the summary statistics.

(n) Using the **Overlay *t* distribution** in the applet, how does the p-value compare to the empirical p-value from the simulated paired-randomization test? Do the *t*-procedures appear to be valid for these data?

(o) Use the check box to display the **95% CI for average difference**. Report and interpret this interval in context. Is the confidence interval consistent with the p-value? What additional information is provided by the confidence interval?

(p) Verify your results using software.

Technology Detour – Paired *t*-tests

In R: you can use the t.test command as before but specify the data are paired, e.g.,

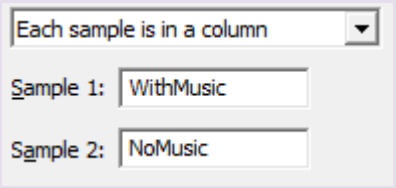
```
>  t.test(WithMusic, NoMusic, alternative="greater", conf.level =
+  .95, paired=TRUE)
```

OR with stacked data

```
>  t.test(speed~condition, alt="greater", paired=TRUE)
```

In Minitab

- Choose **Stat > Basic Statistics > Paired t**.
- Specify the two columns of data in the "Sample 1" and "Sample 2 boxes.
- Under **Options** specify the direction of the alternative.
- Press **OK** twice.



In JMP

- Choose **Analyze > Specialized Modeling > Matched Pairs**
- Specify the two columns in the **Y, Paired Response** box and press **OK**..

The output will include the test statistic (t-Ratio), p-values, and confidence interval endpoints.

Discussion: To compare two groups on a quantitative variable, a more powerful study design than randomly assigning individuals to two groups, if possible, is to pair the individuals and measure both responses in each pair (See Example 4.4). This pairing accounts for the variability from individual to individual and allows for a more direct comparison between the two conditions. For example, if two measurements are taken on each individual, there should not be any other systematic differences between the measurements other than the treatment effect and random chance. Randomizing will still be important in determining the *order* of the two treatments, thereby eliminating order as a potential confounding variable. By accounting for the variability in individuals, this should increase the power of the test of significance, making it easier to detect a difference between the two conditions if one really exists. To analyze such data, perform a matched-pairs randomization test or a (one sample) *t*-test on the *differences*.

Study Conclusions

A paired experiment comparing typing speeds with and without up-tempo classical music (Overture to Candide performed by the London Symphony Orchestra) found participants were significantly faster on average with the music. A paired *t*-test on the mean difference give a one-sided p-value of 0.006, similar to the simulation-based p-value using "random swapping" of the speeds to represent no difference between the two treatment conditions. We are 95% confident that participants like those in this study (e.g., college students, music students or athletes willing to help out a friend) type, on average, 0.63 to 4.61 more words per minute when listening to the up-tempo music. This evidence is much stronger than when we only compared the participants on their first tests, because of both effectively doubling the sample size and also reducing the among of "unexplained variation" in the response variable. The person-to-person variation in typing speeds was around 15 wpm, compared to the person-to-person variation in difference in typing speeds which was around 5 wpm. A further analysis could compare the average improvement with music between the music students and the athletes.

Practice Problem 4.9A

(a) Use statistical software to determine the power of detecting a difference of 5 wpm in a two-sample *t*-test if the sample standard deviations are 15 wpm. (You can assume a 5% level of significance and a one-sided alternative, as well as a total sample size of 34.)
(a) Repeat (a) assuming the sample standard deviations are 5 wpm.
(c) Use statistical software to determine the power of detecting a difference of 5 wpm in a one-sample paired *t*-test assuming the sample standard deviation of the differences is 5 wpm.
(d) How does the power of the paired *t*-test compare to the power of the two-sample *t*-test? (Cite appropriate evidence.)

Practice Problem 4.9B

Scientists have long been interested in whether there are physiological indicators of diseases such as schizophrenia. In a 1990 study by Suddath et. al., reported in Ramsey and Schafer (2002), researchers used magnetic resonance imaging to measure the volumes of various regions of the brain for a sample of 15 monozygotic twins, where one twin was affected by schizophrenia and other not ("unaffected"). The twins were found in a search through the United States and Canada, the ages ranged from 25 to 44 years, with 8 male and 7 female pairs. The data (in cubic centimeters) for the left hippocampus region of the brain are in hippocampus.txt. The primary research question is whether the data provide evidence of a difference in hippocampus volumes between those affected by schizophrenia and those unaffected.
(a) Calculate the difference in hippocampus volumes for each pair of twins (unaffected – affected).
(b) Calculate and interpret a 95% confidence interval for the mean volume difference using the paired *t*-interval. Also comment on the validity of this procedure.
(c) Based on this confidence interval, is there statistically significant evidence that the mean difference in left hippocampus volumes is different from zero? Explain.