Comparing Two Dependent Means

10/24

Previously

- Independent sample t-tests
- Pooled variance
- Effect sizes: Cohen’s D and distribution overlap
- Assumptions and dealing with violations
    - Normality: Wilcoxon rank sum test
    - Homogeneity of variance: Welch’s t-test
    - Independence?

Dealing with independence violations

The independent samples t-tests (student's and Welch's) assume that the responses in one group are uncorrelated with the responses in the other group.

That independence assumption is sometimes violated, as in the following common research situations:

- Longitudinal data
- Paired samples
- Paired measures

Note that these are violations between groups, but not within groups. Ideally, you want to avoid dependency within groups

Research Designs with Dependencies

- In <u>longitudinal</u> research, the same people provide responses to the same measure on two occasions.
- In <u>paired-sample</u> research, the individuals in the two groups are different, but they are related and their responses are assumed to be correlated. Examples would be responses by children and their parents, members of couples, twins, etc.
- In <u>paired-measures</u> research, the same people provide responses to two different measures that assess closely related constructs. This resembles longitudinal research, but data collection occurs at one time.
- All of these are instances of repeated measures designs.

Research Designs with Dependencies

- The advantage of repeated measures designs is that, compared to an independent groups design of the same size, the repeated measures design is more powerful.
- Two groups are more alike than in simple randomization
- The correlated sampling units will have less variability on "nuisance variables" because those are either the same over time (longitudinal) or over measures (paired measures), or very similar over people (paired samples). Nuisance variables -- anything that isn't relevant to the study.

t-test for Repeated Measures

- You can convert two dependent measures into a single difference score, $D = X_1 - X_2$.
- It is then appropriate to conduct a one-sample t-test on the difference score, with a null mean of 0.

$$t_{df=N-1} = \frac{\bar{\Delta} - \mu}{\frac{\hat{\sigma}_\Delta}{\sqrt{N}}}$$

$$H_0 : \bar{\Delta} = \mu \qquad H_0 : \bar{\Delta} = 0$$

$$H_1 : \bar{\Delta} \neq \mu \qquad H_1 : \bar{\Delta} \neq 0$$

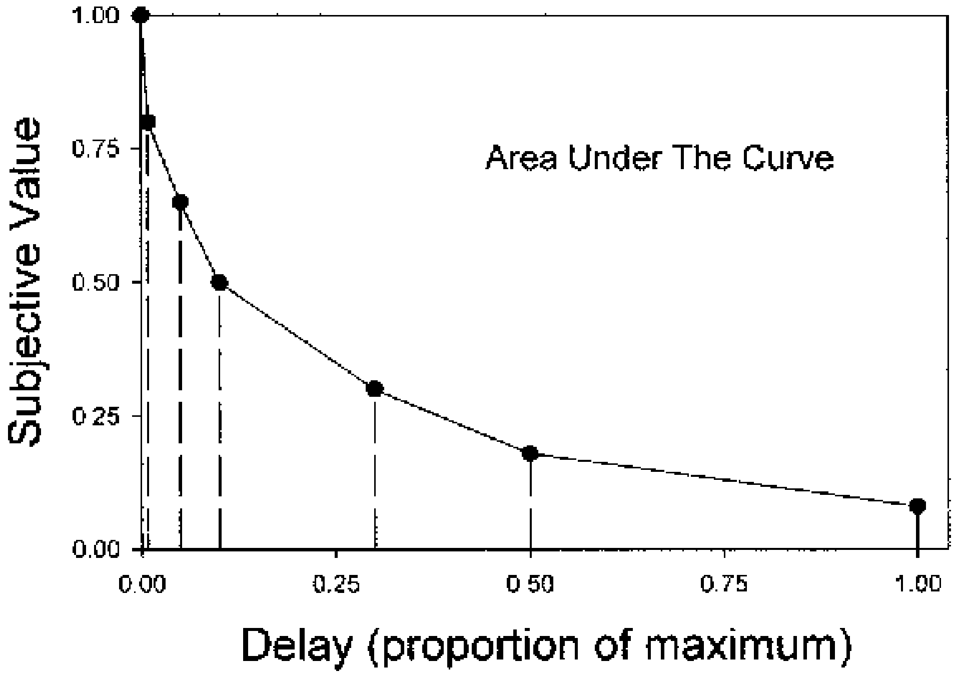
Example

- People value delayed outcomes less, a phenomenon known as delay discounting.
- Outcomes are discounted more the longer they are delayed.
- Both rewards(gains) and losses are subject to delay discounting.
- In this study, we examined how people discount delayed losses that are preceded by immediate gains.

Learn more about the study: https://doi.org/10.1002/jeab.929

- Each participant completed three conditions.
- In the <u>loss-only</u> condition, participants discounted simple delayed losses. (Pay $1,000)
- In the <u>net-loss</u> condition, participants discounted combination outcomes where immediate gains are followed by delayed losses, resulting in a net gain. (Get $750 now, pay $1,000 later)
- In the <u>net-gain</u> condition, participants discounted similar combination outcomes that resulted in a net gain. (Get $1,250 now, pay $1,000 later)
- n = 76

- Assessing the subjective values of delayed outcomes at different delays allows us to calculate the Area under the Curve (AuC), a measure of degree of discounting.
- AuC is constrained between 0 and 1. High values indicate shallower discounting (more self-control).



Our Hypothesis

- We’re interested in if people discount combination outcomes that result in net gains differently from simple delayed losses.
- Because each participant completes all three conditions, we use paired-samples t-test.

H_0: There is no difference in the degree of discounting between the loss-only condition and the net-gain condition.

H_1: People discount more shallowly in one of the conditions.

Sampling Distribution

t-distribution requires two parameters, the mean and the standard distribution.

According to our null hypothesis, the mean of our sampling distribution is 0.

The standard deviation of our sampling distribution is the standard error of difference scores. We can find this by:

- calculating difference scores
- calculating the standard deviation of difference scores
- dividing the standard deviation by the square root of the number of pairs.

```
discounting = read.csv("discounting_data.csv")
difference = discounting$loss_only - discounting$net_gain
```

Calculate the mean of difference scores:

```
m_delta = mean(difference) # -.14
```

And the standard deviation:

```
s_delta = sd(difference) # .28
```

We can calculate the standard error by dividing the standard deviation by the square root of the number of pairs or, in the case of repeated measures, the number of subjects (n).

```
se_delta = s_delta/sqrt(nrow(discounting)) # .03
```

Test Statistic

$$t_{df=N-1} = \frac{\bar{\Delta} - \mu}{\frac{\hat{\sigma}_\Delta}{\sqrt{N}}}$$

N here refers to the number of pairs, not total sample size

$$t_{df=N-1} = \frac{-.14 - 0}{.03} = -4.4$$

Is a paired samples t-test any different from a one-sample t-test on difference scores?

```r
t_stat = m_delta/se_delta # -4.4
```

You may use the same procedure for a one-sample t-test to find the p-value. It estimates the probability of finding this test statistic or more extreme.

```r
pt(t_stat, df = 76-1, lower.tail = T) # 1.72e-05
```

```r
pt(t_stat, df = 76-1, lower.tail = T)*2 # 3.45e-05
```

Your Turn!

Do people discount combination outcomes that result in net losses differently from simple delayed losses?

Do people discount combination outcomes that result in net gains differently from those that result in net losses?

t-test Functions

You can also use the t.test function for paired samples t-tests.

```
t.test(x = discounting$loss_only, y = discounting$net_gain, paired = TRUE)
```

```
t.test(x = discounting$loss_only, y = discounting$net_loss, paired = TRUE)
```

```
t.test(x = discounting$net_loss, y = discounting$net_gain, paired = TRUE)
```

Cohen’s D

Calculating a standardized effect size for a paired samples t-test (and research design that includes nesting or dependency) is slightly complicated, because there are two levels at which you can describe results.

The first level is the within-subject level, and this communicates effect size in the unit of differences (of units).

$$d = \frac{\bar{\Delta}}{\hat{\sigma_{\Delta}}} = \frac{-.14}{.28} = -.51$$

The interpretation is that, on average, variability within a single participant is about .51 standard deviations of differences.

Cohen’s D Functions

```
lsr::cohensD(x = discounting$loss_only, y = discounting$net_gain, method = "paired") # .51
```

The second level is the between-conditions variance, which is in the units of your original outcome and communicates how the means of the two conditions differ.

For that, you can use the Cohen's d calculated for independent samples t-tests.

```
lsr::cohensD(x = discounting$loss_only, y = discounting$net_gain, method = ”pooled") # .62
```

Interpreting Cohen’s D

Unlike hypothesis testing, there are no standards for Cohen’s d.

Generally, a Cohen’s d of .2 is considered “small”, .5 is considered “moderate”, and .8 is considered “large”.

Historically, the within-subject version is encouraged because it mirrors the hypothesis test.

Some argue the between-conditions version is actually better because the paired-design is used to reduce noise by adjusting our calculation of the standard error. But that shouldn't make our effect bigger, just easier to detect. The other argument is that using the same formula (the between-conditions version) allows us to compare effect sizes across many different designs, which are all trying to capture the same effect.

Cohen’s D from t

This can be calculated from t-statistics, allowing you to calculate standardized effect sizes from manuscripts even when the authors did not provide them.

One sample or within-subjects for paired

$$d = \frac{t}{\sqrt{N-1}}$$

Independent sample

$$d = \frac{2t}{\sqrt{N_1 + N_2 - 1}}$$

What are we looking for?

$$\frac{\color{blue}{Signal}}{\color{red}{Noise}}$$

Make Your Study More Powerful

- Improving the signal-to-noise ratio:
  - Study larger effects! (not always possible)
  - Increase your sample size! (easier but still not always feasible)
  - Use within-subject designs
  - Replications

ASSUMPTIONS!

- Independence between pairs
- Normality

Is the discounting data normal?

Violating Normality

We can use the paired-samples version of Wilcoxon test when the assumption of normality is violated.

```
wilcox.test(discounting$loss_only, discounting$net_gain, paired = T) # p<.01
```

The Wilcoxon test converts data into ranks.

Are you satisfied with this solution?

- What if each measure is not independent from each other?
  - Later measures might be biased
  - Counterbalancing
  - Sometimes unavoidable (longitudinal research)

- Homogeneity of variance
  - Different standard deviations within pairs

- What if our statistical tests are not independent?
  - Exploratory studies
  - We’ll try a different test if the first one is not significant…

NEWS | 14 October 2024 | Clarification 17 October 2024 | Clarification 23 October 2024

‘Doing good science is hard’: retraction of high-profile reproducibility study prompts soul-searching

A paper by some of the biggest names in scientific integrity is retracted for issues including misstatements about the research plan.

Next time

Data visualization

(Exam 2 review)

Exam 2…