Comparing two dependent means

Previously...

t-tests galore!

Thoughts on signs

Thoughts on conservatism (Welch vs. Student, t vs. normal, etc.)

Previously...

- independent samples t-tests
- pooled variance
- effect sizes -- Cohen's D and distribution overlap
- assumptions and dealing with violations
    - normality -- Wilcoxon Rank Sum
    - homogeneity of variance -- Welch's *t*-test
    - independence -- ??

Today

The independent samples t-test (Student's or Welch's) assumes 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. For now, we'll deal with the problem of violating independence between two groups, but these techniques won't help you if there is dependency within a group

Research Designs with Dependencies

In **longitudinal research**, the same people provide responses to the same measure on two occasions (the individuals in the two groups are the same).

In **paired-sample 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 **paired-measures 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.

The Pros

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.

Dealing with Dependencies

Each of these repeated measures problems can be viewed as a transformation of the original two measures into a single measure: a difference score. This reduces the analysis to a **one-sample *t*-test** on the difference score, with null mean = 0.

Dealing with Dependencies

If the repeated measures are X_1 and X_2, then their difference is $D = X_1 - X_2$. This new measure has a mean and standard deviation, like any other single measure, making it appropriate for a one-sample *t*-test.

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

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

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

Example

Human-wildlife conflict in urban areas endangers wildlife species. One species under threat is the *Larus argentatus* or herring gull, which is considered a nuisance owing to food-snatching and other behaviors. A recent study examined whether herring gull behavior is influenced by human behavior cues and whether this could be used to reduce human-gull conflict.

In this study, experimenters visited coastal towns in the UK and found locations with multiple gulls. They placed a bag of potato chips (250 g) in front of them and measured how long it took gulls to peck at the food.

Looking At: the experimenter directed their gaze towards the eye(s) of the gull and turned their head, if necessary, to follow its approach path until the gull completed the trial by pecking at the food bag.

Looking Away: the experimenter turned their head and eyes approximately 60° (randomly left or right) away from the gull and maintained this position until they heard the gull peck at the food bag.

"We adopted a repeated measures design... We randomly assigned individuals to receive Looking At or Looking Away first, and trial order was counterbalanced across individuals. Second trials commenced 180 s after the completion of the first trial to allow normal behaviour to resume."

```
gulls = read.delim(here("data/gulls/pairs.txt"))
gulls
```

```
##    GullID  At Away
## 1   FAL01 210   35
## 2   FAL03 300   80
## 3   FAL04   6    3
## 4   PEN03  18   21
## 5   W120M  47   13
## 6    W019  25    4
## 7   PNZ01   4   13
## 8   PNZ02   9    8
## 9   STI01 300   18
## 10  STI02 300    6
## 11   W186  11    8
## 12  STI03   4    3
## 13  STI04   4    6
## 14  HEL02  12   19
## 15  NEW01 300    6
## 16  NEW02  63   16
## 17  NEW03 300  166
## 18  PER01  24   66
## 19  TRU01 300  167
```

Hypothesis testing

Use a paired-samples *t*-test because we have the same gulls in both conditions.

H_0: There is no difference in how long it takes gulls to approach food between conditions.

H_1: Gulls take longer to approach food in one of the conditions.

Sampling distribution

t-distribution requires two parameters, a mean and a standard deviation.

The mean of our sampling distribution comes from our null hypothesis, so

$$\mu = 0$$

Our standard deviation of our sampling distribution is the **standard error of difference scores**. This can be found by

1. calculating difference scores
2. calculating the standard deviation of the difference scores, and
3. dividing the standard deviation by the square root of the number of *pairs* in your study.

```
difference = gulls$At - gulls$Away
```

```
## [1] 175 220   3  -3  34  21
```

We can take the mean of this new variable:

```
(m_delta = mean(difference))
```

```
## [1] 83.10526
```

And we can calculate the standard deviation

```
(s_delta = sd(difference))
```

```
## [1] 115.8485
```

Std Error: divide the standard deviation by the square root of the number of *pairs* or, in the case of repeated measures, the number of *subjects*.

```
(se_delta = s_delta/sqrt(nrow(gulls)))
```

```
## [1] 26.57747
```

$$\frac{\hat{\sigma}_{\Delta}}{\sqrt{N}} = 26.58$$

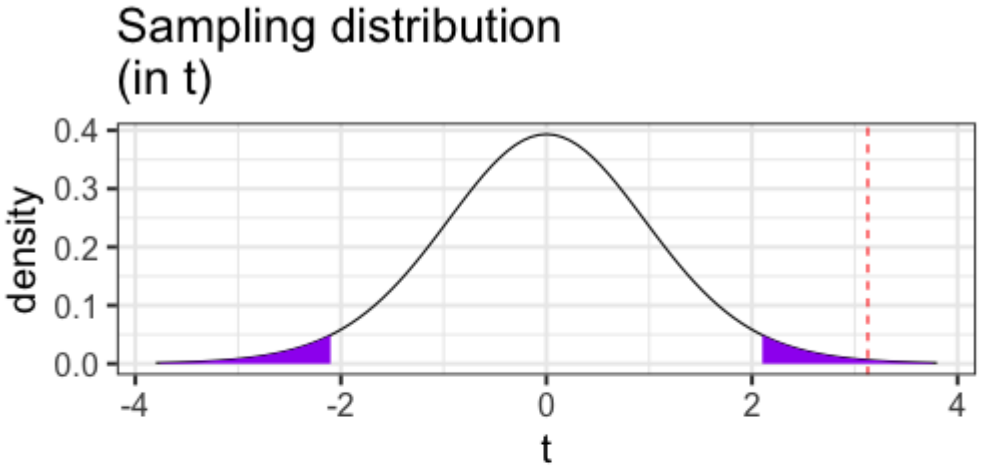
Test statistic

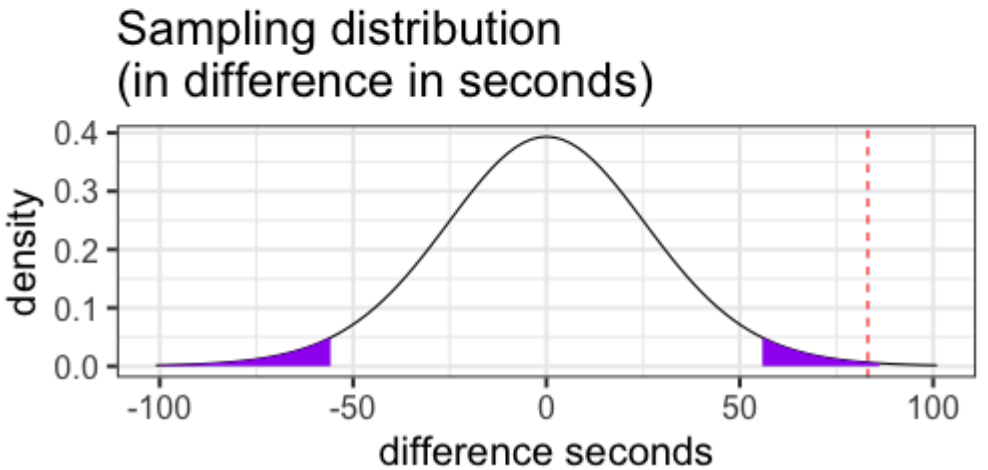
$$t_{df=N-1} = \frac{\bar{\Delta} - \mu}{\frac{\hat{\sigma}_{\Delta}}{\sqrt{N}}}$$

In this case, N refers to the number of pairs, not the total sample size.

$$t_{df=N-1} = \frac{83.11 - 0}{26.58} = 3.13$$

Note: A paired-samples *t*-test is *exactly* the same as a one-sample *t*-test on the difference scores.


Sampling distribution
(in t)
density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
t


Sampling distribution
(in difference in seconds)
density
0.4
0.3
0.2
0.1
0.0
-100
-50
0
50
100
difference seconds

Get the *p*-value

Calculate the area above the absolute value of the test statistic and multiply that by two -- this estimates the probability of finding this test statistic or more extreme.

```
(t_statistic = m_delta/se_delta)
```

```
## [1] 3.126906
```

```
pt(t_statistic, df = 19-1, lower.tail = F)
```

```
## [1] 0.002912942
```

```
pt(t_statistic, df = 19-1, lower.tail = F)*2
```

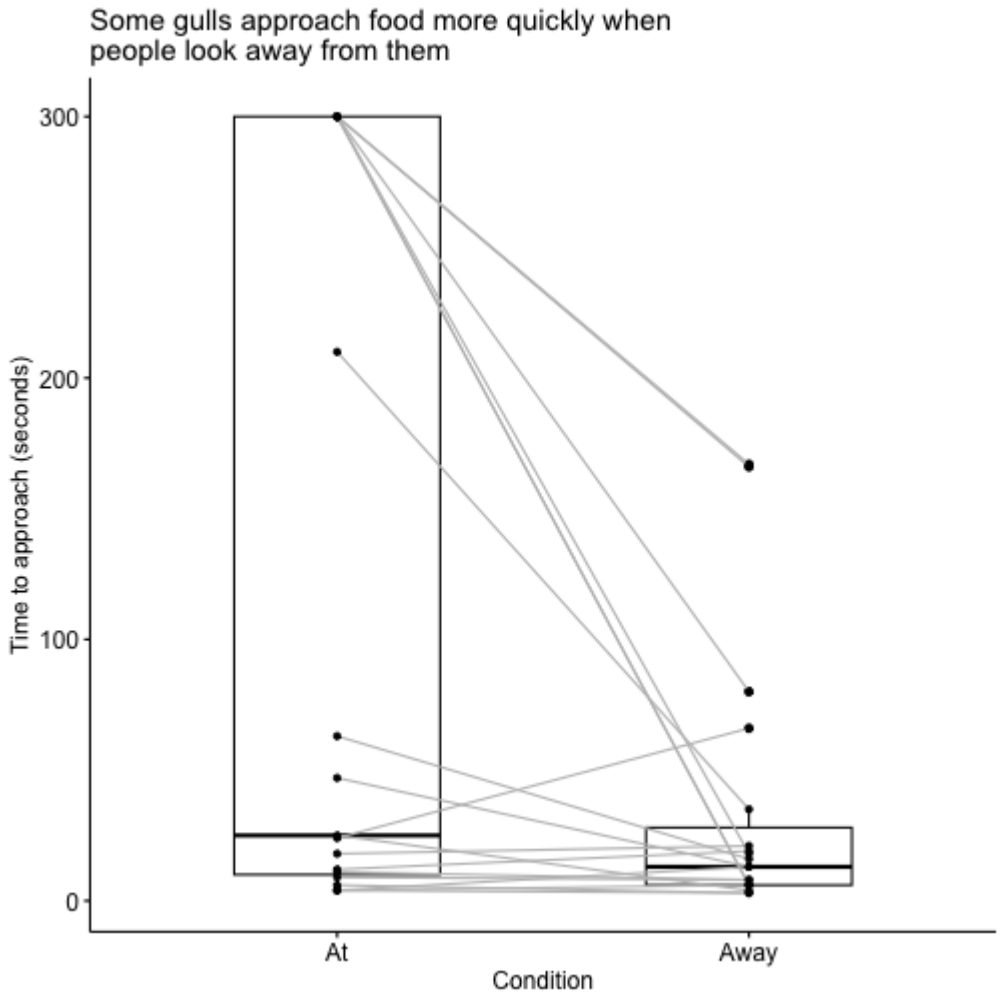
```
## [1] 0.005825884
```

t-test functions

```
t.test(x = gulls$At, y = gulls$Away,
       paired = TRUE)
```

```
##
##     Paired t-test
##
## data:  gulls$At and gulls$Away
## t = 3.1269, df = 18, p-value = 0.005826
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   27.26807 138.94246
## sample estimates:
## mean difference
##        83.10526
```

```
ggpubr::ggpaired(data = gulls, cond1 = "At", cond2 = "Away", line.color
                 ylab = "Time to approach (seconds)",
                 title = "Some gulls approach food more quickly when \np
```



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** (or within-pair, or within-gull) level, and this communicates effect size in the unit of differences (of units).

$$d = \frac{\bar{\Delta}}{\hat{\sigma_{\Delta}}} = \frac{83.11}{115.85} = 0.72$$

The interpretation is that, on average, variability within a single gull is about .72 standard deviations of differences.

Cohen's D

```
lsr::cohensD(x = gulls$At, y = gulls$Away, method = "paired")
```

```
## [1] 0.7173615
```

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 = gulls$At, y = gulls$Away, method = "pooled")
```

```
## [1] 0.8137369
```

Which one should you use?

There are no standards for effect sizes. When Cohen (1988) developed his formula, he never bothered to precisely define σ. Interpretations have varied, but no single method for within-subjects designs has been identified.

Most often, textbooks will argue for the within-pairs version, because this mirrors the hypothesis test.

Some argue the between-conditions version is better because the paired-design is used to reduce noise by adjusting our calculation of the standard error. The other argument is that using the same formula (betwn-conditions version) lets us to compare effect sizes across many different designs, which are all trying to capture the same effect.

Westfall, J. (2016). Five different "Cohen's d" statistics for within-subject designs

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

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

Assumptions

- Independence (between pairs)
- Normality

Note: These are the same assumptions as a one-sample *t*-test.

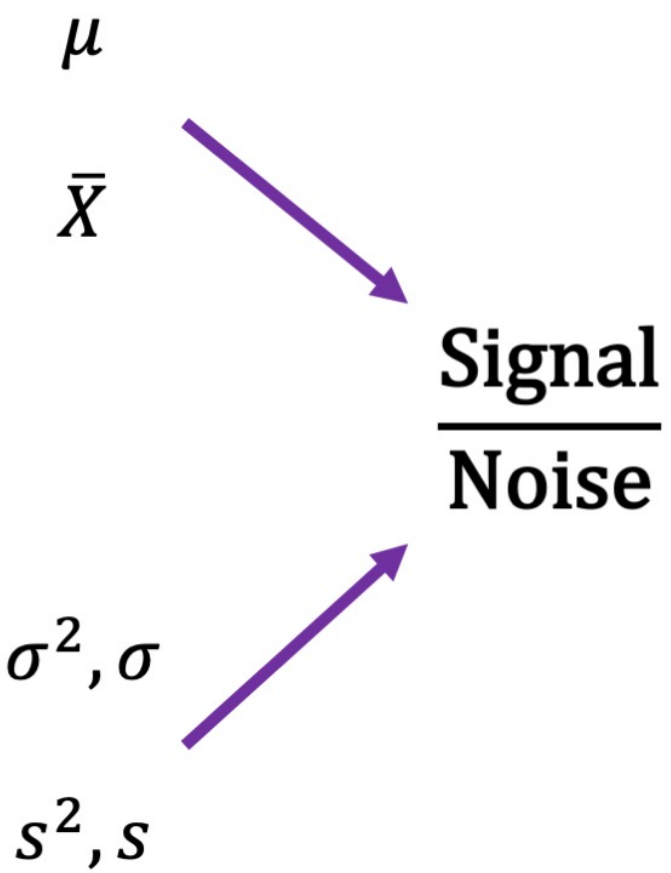
Violating normality

For two independent samples, we can use the Wilcoxon Sum Rank test if we have severe violations of normality. There is a paired-samples version as well.

```
wilcox.test(gulls$At, gulls$Away, paired = T)
```

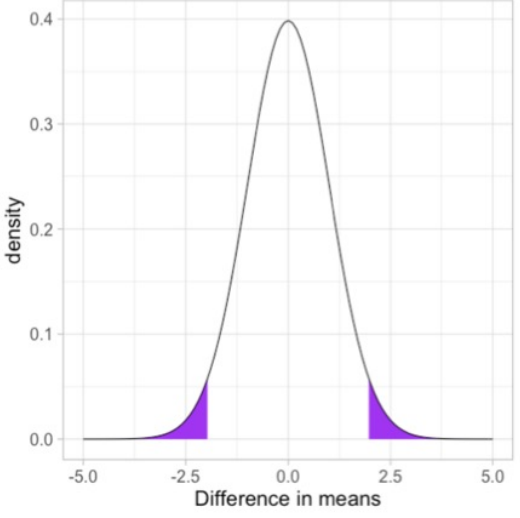
```
##
##     Wilcoxon signed rank test with continuity correction
##
## data:  gulls$At and gulls$Away
## V = 156, p-value = 0.01485
## alternative hypothesis: true location shift is not equal to 0
```

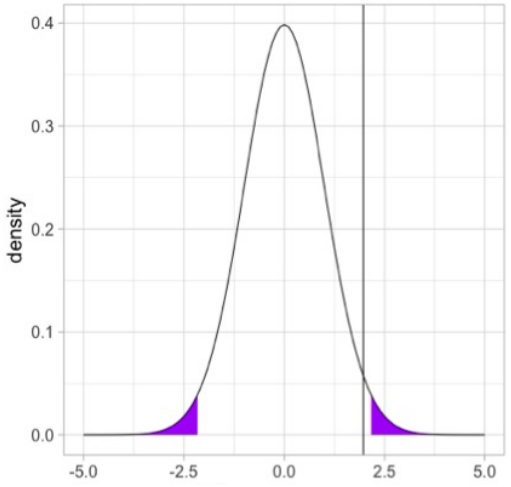
$$\frac{\text{Signal}}{\text{Noise}}$$


μ
$\bar{X}$
Signal
Noise
σ^2, σ
s^2, s

$$\frac{\text{Signal}}{\text{Noise}}$$

How much noise should we expect?





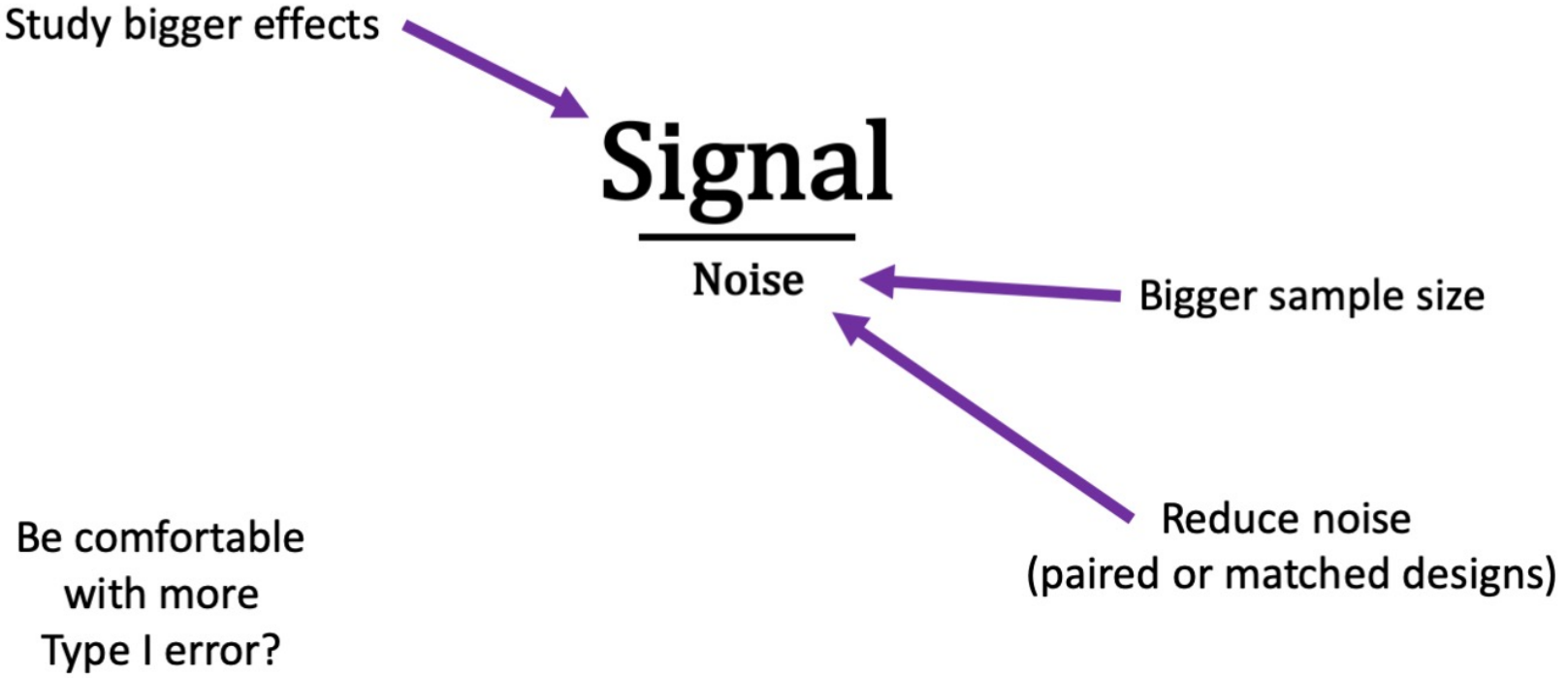
How likely is this?

$$\frac{\text{Signal}}{\text{Noise}}$$

How can I make my study more powerful?

$$\frac{\text{Signal}}{\text{Noise}}$$

How can I make my study more powerful?



Because we're using probability theory, we have to assume independence between trials.

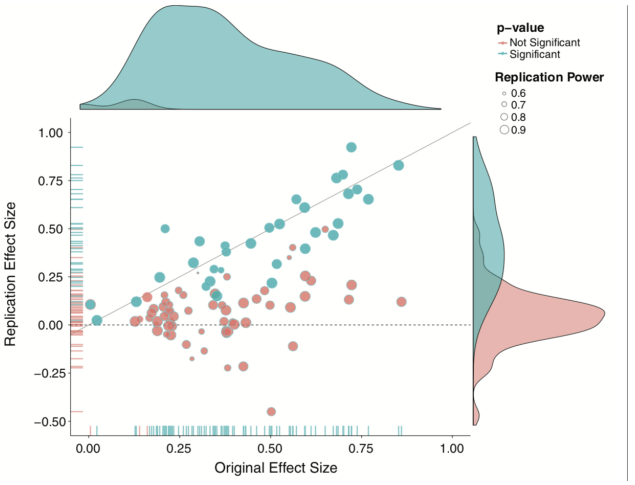
- independent rolls of the die
- independent draws from a deck of cards
- independent... tests of hypotheses?

What happens when our tests are not independent of one another?

- For example, what happens if I only add a covariate to a model if the original covariate-less model is not significant?

- Similarly, what happens if I run multiple tests on one dataset?

The issue of independence between tests has long been ignored by scientists. It was only when we were confronted with impossible-to believe results (listening to "When I'm Sixty-Four" by The Beatles makes you younger) that we started to doubt our methods. + a lot of fraud.

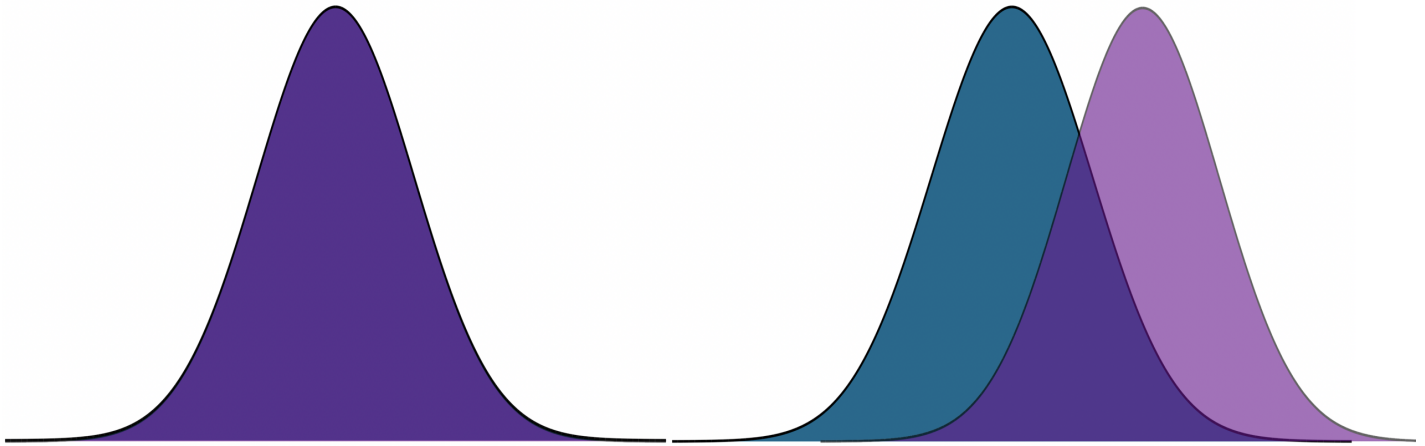


Since then (2011), there's been greater push to revisit "known" effects in psychology and related fields, with sobering results.

The good news is that things seem to be changing, rapidly.

Independence isn't the only assumption we have to make to conduct hypothesis tests.

We also make assumptions about the underlying populations.



Normality

Homogeneity of variance

Sometimes we fail to meet those assumptions, in which case we may have to change course.

Picture credit: Cameron Kay

How big of a violation of the assumption is too much?

- It depends.

How do you know your operation X measures your construct A well?

- You don't.

What should alpha be?

- You tell me.

Coming up...

- Written Exam 2 11/2 (ommmgggggggg!!!!!)
- Oral Exam 2 11/6-11/8