Introduction to Longitudinal Data Analysis

Andrew Zieffler

Longitudinal Data

- Repeated measurements of the same individual(s) over time

Primary Goals: (1) Characterize the change in response over time; (2) understand the factors that influence this change

- Cross-sectional studies do NOT allow the study of within-individual change over time

Example: Suppose a researcher was interested in studying the change in achievement scores over time. In a cross-sectional design, the researcher might obtain achievement data from *two separate groups of students*, say 4th-graders and 6th-graders. By comparing the means (e.g., two-sample *t*-test) she can say that the average achievement between 4th- and 6th-graders is different. However, it is impossible to get a reasonable estimate of the ACTUAL effect of age on achievement because the are many potential confounders with the different cohorts which distort this effect.

- In longitudinal studies individuals act as their own control group and the effect of age on achievement is estimated free of the between-individual variation.

Clustered Data

- Because it is collected from the SAME individual(s), longitudinal data is **clustered**.
- Measurements within a cluster (e.g., from the same person) are generally more similar (typically positively correlated) than measurements from different clusters.

DANGER: The within-cluster correlation of measurements has implications for the analysis. The assumption of independence will be violated.

- Clustered data (a.k.a.: hierarchical data; nested data) describes data structures common throughout numerous research domains, not all of which are longitudinal data.
- Longitudinal data is also **temporal**; the first measurement necessarily precedes the second measurement, etc.

Clustered Non-Longitudinal Data

- Clustered data commonly arises when intact GROUPS are randomly assigned to conditions.

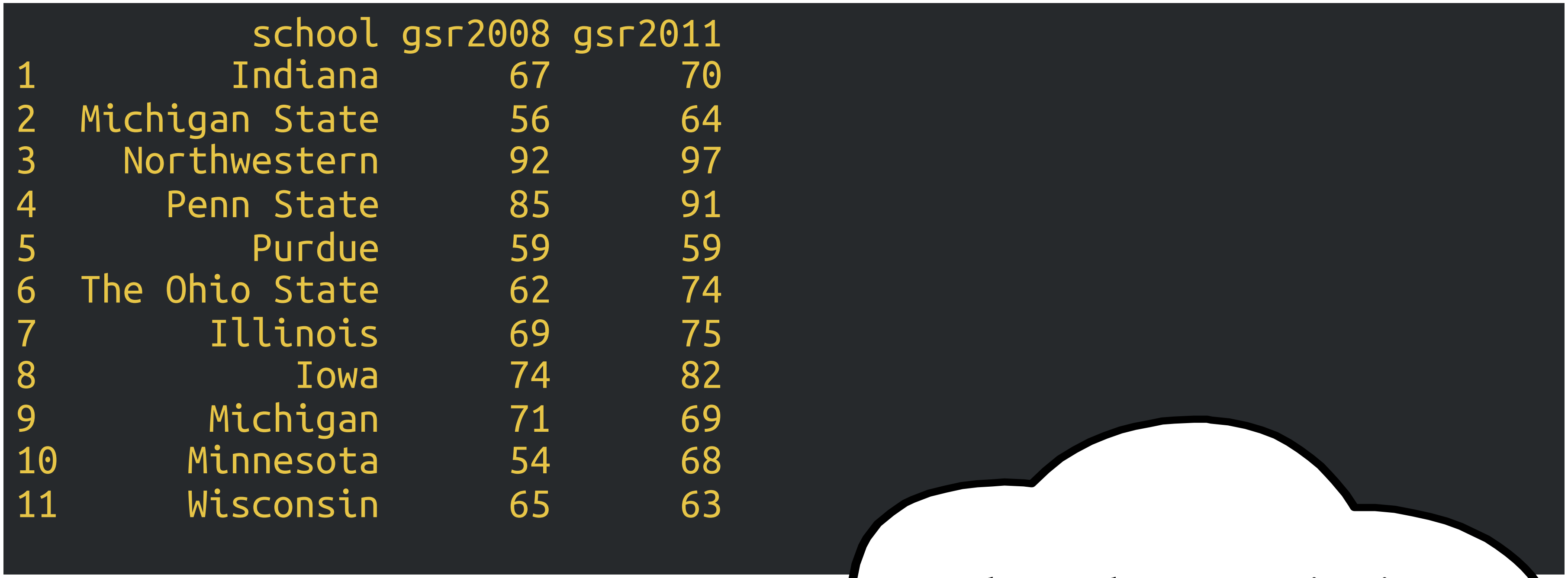
Example: A researcher assigns whole classrooms of students to a condition (group-randomized trial). Student-level data is then collected for analysis to compare the conditions.

- In this example, the classroom is the cluster and the measurements of the individual students will typically be more similar (correlated) than measurements between clusters.
- This is not longitudinal data since the measurements within the cluster do not have a temporal order (all students data within the classroom are collected simultaneously).
- Natural clusters: Families, neighborhoods, schools, classrooms, hospitals, offices

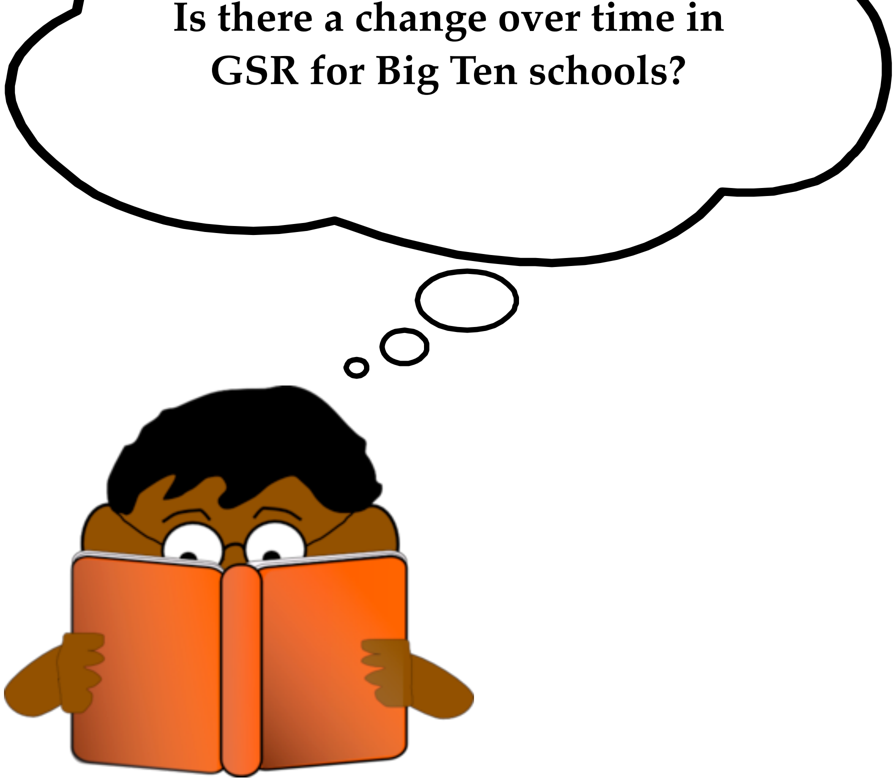
Analysis of Clustered Data

- The correlation of measurements within clusters violates the assumption of independence.
- Statistical models used to analyze clustered data must account for this correlation. (Statistical models used to analyze longitudinal data must also account for the temporal nature of the data, and are a special case of the methods used to analyze clustered data.)

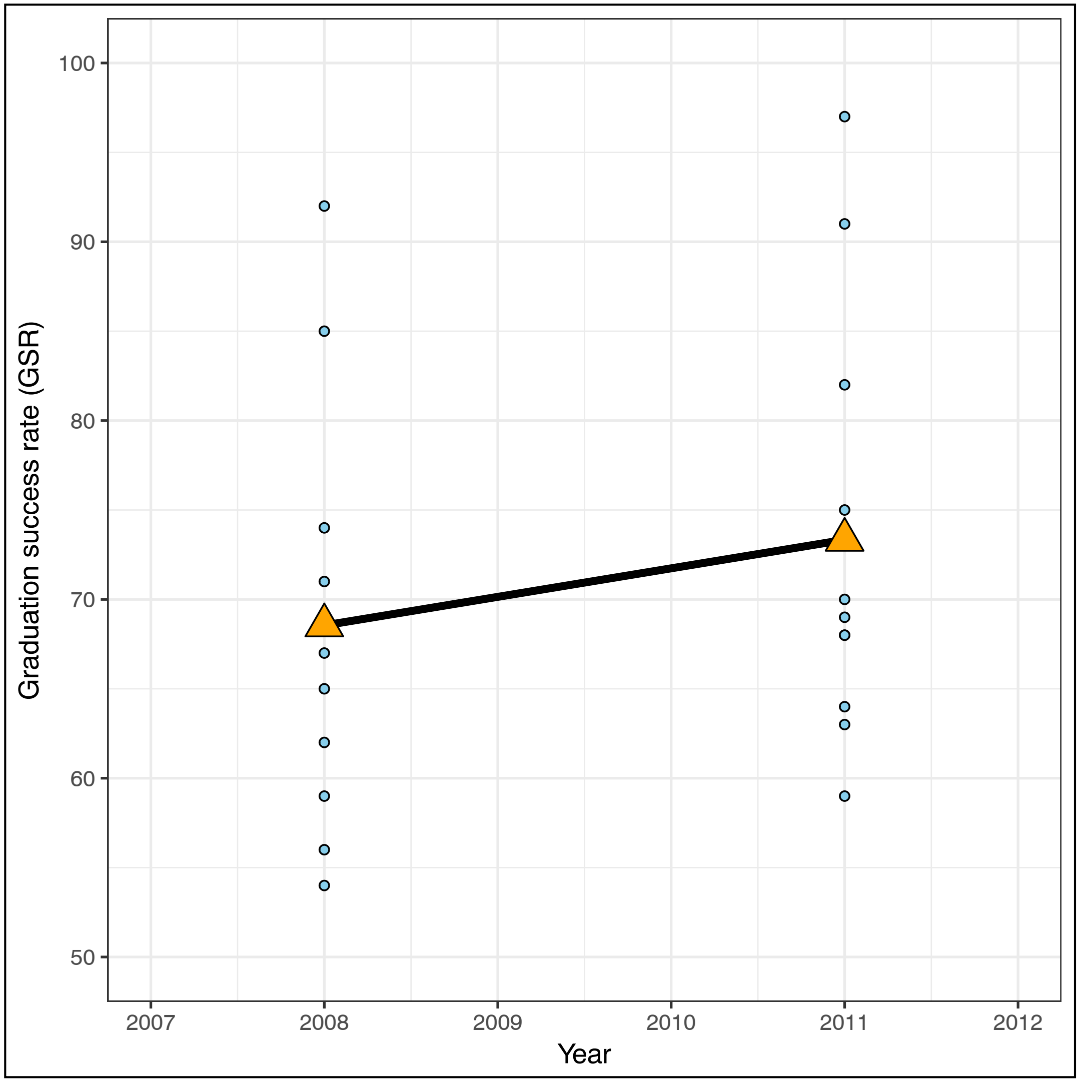
Example: A researcher wants to examine whether graduation success rate (GSR) has changed over time for football teams in the NCAA Big Ten conference. She has GSR data for the Big Ten teams from the 2008 and 2011 (based on the 2002 and 2005 freshmen cohorts).



```
              school gsr2008 gsr2011
1            Indiana      67      70
2     Michigan State      56      64
3       Northwestern      92      97
4         Penn State      85      91
5             Purdue      59      59
6     The Ohio State      62      74
7           Illinois      69      75
8               Iowa      74      82
9           Michigan      71      69
10         Minnesota      54      68
11         Wisconsin      65      63
```



Year	*N*	*M*	*SD*
2008	11	68.5	11.7
2011	11	73.8	11.9



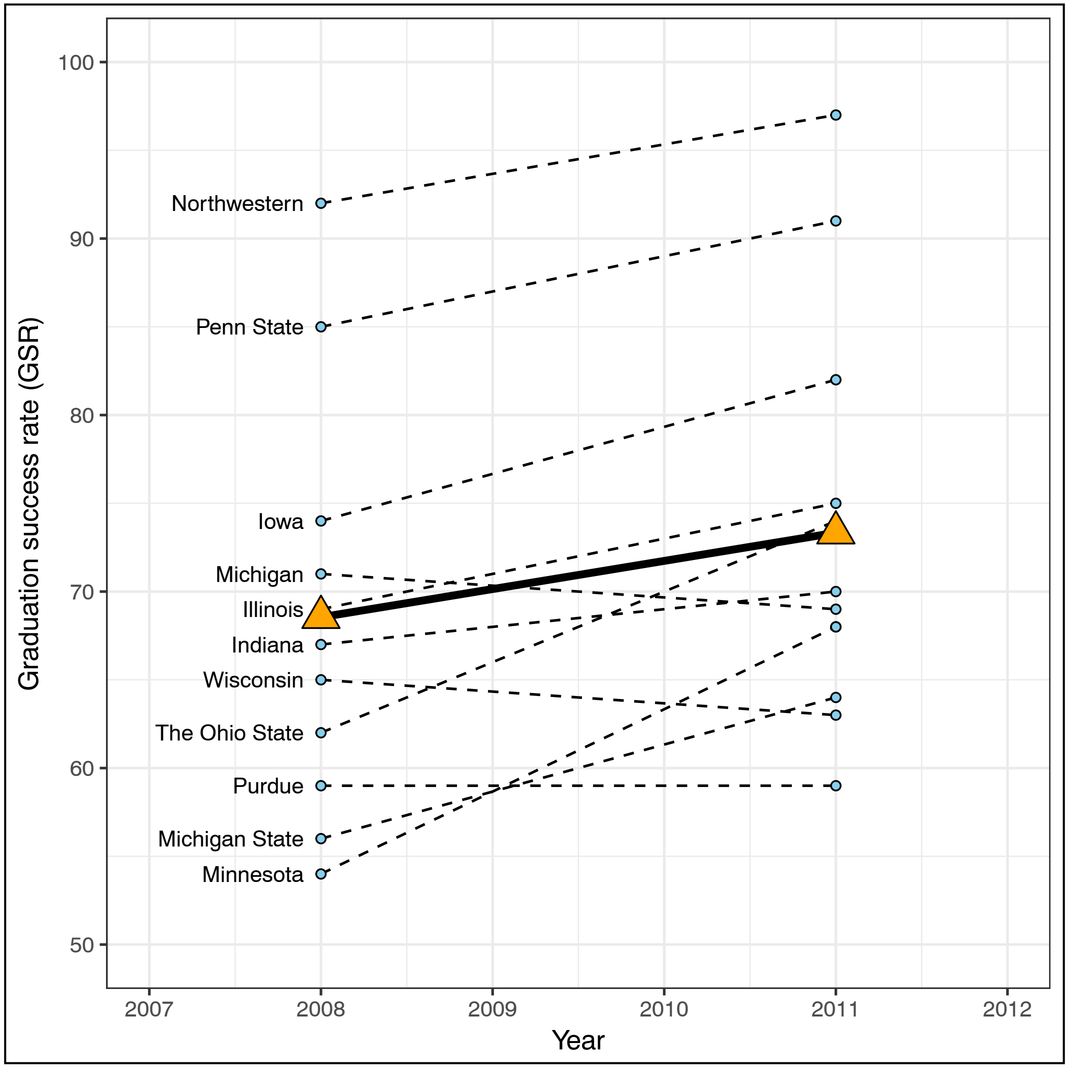
The summaries and plot suggest that: (1) the average GSR increased from 2008 to 2011; and (2) the variation in GSR for both years is approximately equal.

An **inappropriate statistical analysis** of these data would be to test the hypothesis of no mean differences using an independent-samples *t*-test or using linear regression.

```
# Results of the independent-samples t-test

t = -1.0488, df = 20, p-value = 0.3068
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.759966   5.214511


# Results of linear regression

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   68.545      3.555  19.281 2.17e-14 ***
yeargsr2011    5.273      5.028   1.049    0.307
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.79 on 20 degrees of freedom
Multiple R-squared:  0.05213, Adjusted R-squared:  0.004736
F-statistic:   1.1 on 1 and 20 DF,  p-value: 0.3068
```

Both analyses suggest there is not a statistically significant effect of time on GSR ($p = 0.307$). This suggests that the difference we observed in the sample data is only due to chance.

```
# Correlation between 2008 GSR and 2011 GSR


          gsr2008   gsr2011
gsr2008 1.0000000 0.9007227
gsr2011 0.9007227 1.0000000
```



Since the 2008 and 2011 data come from the same set of schools, we can compute the correlation between the measurements. The correlations between 2008 and 2011 GSR is positive and strong ($r = 0.90$). In order to meet the assumption of independence the correlation between measurements needs to be zero (in the population).

This non-independence needs to be accounted for in the analysis. To do this, we link the clustered measurements. In the two-group situation, we can use a **dependent-samples *t*-test** (i.e., the paired samples *t*-test).

```
# Results of the dependent-samples t-test

t = -3.3276, df = 10, p-value = 0.007648
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.803286 -1.742168
```

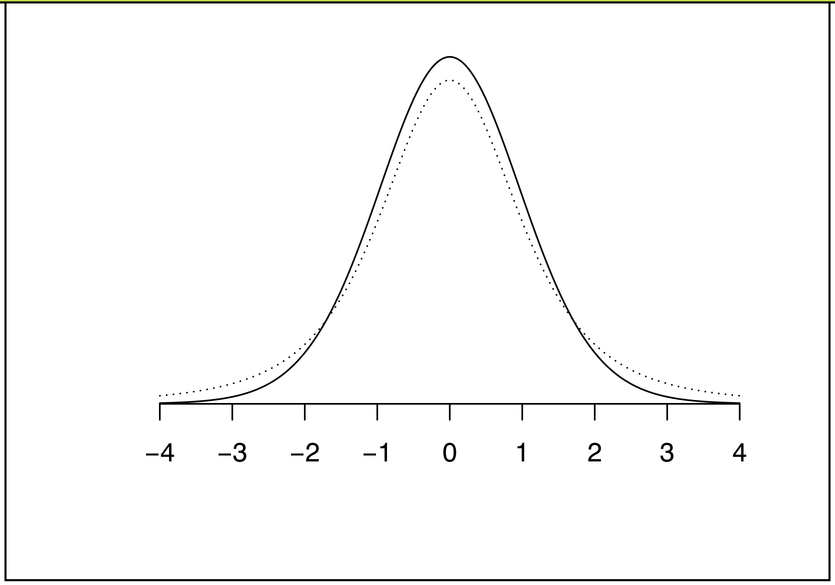
This analysis, which correctly accounts for the correlation in measurements, suggests there is a statistically significant effect of time on GSR ($p = 0.008$). This suggests that the difference we observed in the sample data is not only due to chance.

$$t = \frac{\text{Mean Difference}}{SE_{\text{Mean Difference}}}$$

t-Test	ΔM	*SE*	*t*	*df*	*p*
Independent samples	5.27	5.03	–1.05	20	0.307
Dependent samples	5.27	1.58	–3.33	10	0.008

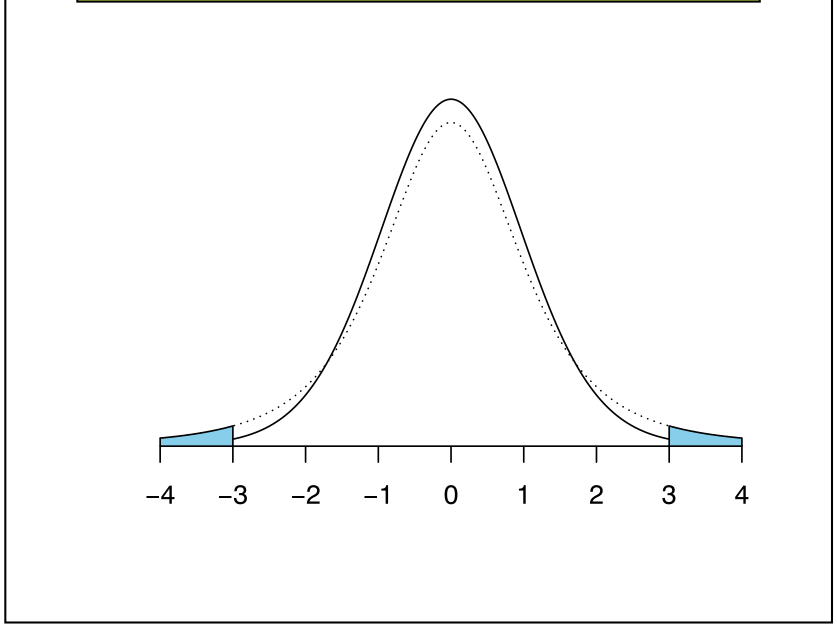
Remember the *t*-distribution is defined by its *df*.

In the figure, the solid line represents the t(20) distribution. The dotted line corresponds to the *t*(10) distribution.

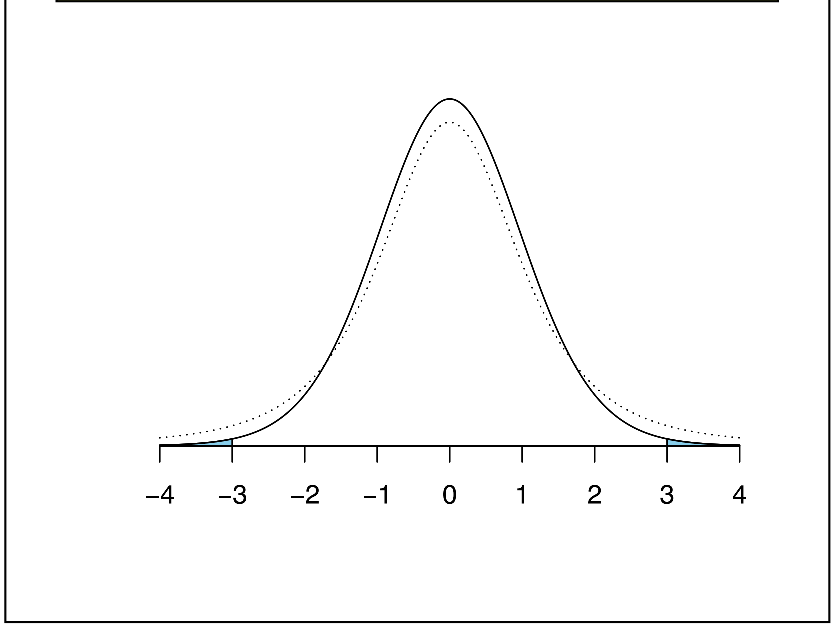


The *p*-value corresponds to the area as extreme or more extreme than the observed *t*-value. Consider a *t*-value of 3.

two-tailed *p*-value for a *t*-value of 3 in the *t*(10) distribution.

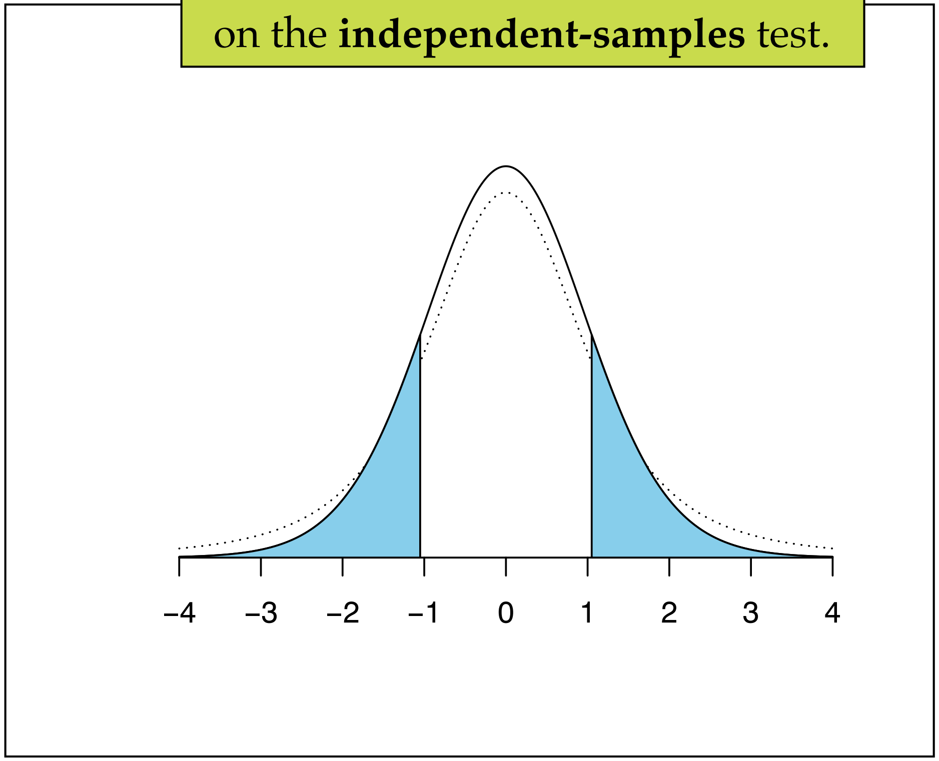


two-tailed *p*-value for a *t*-value of 3 in the *t*(20) distribution.

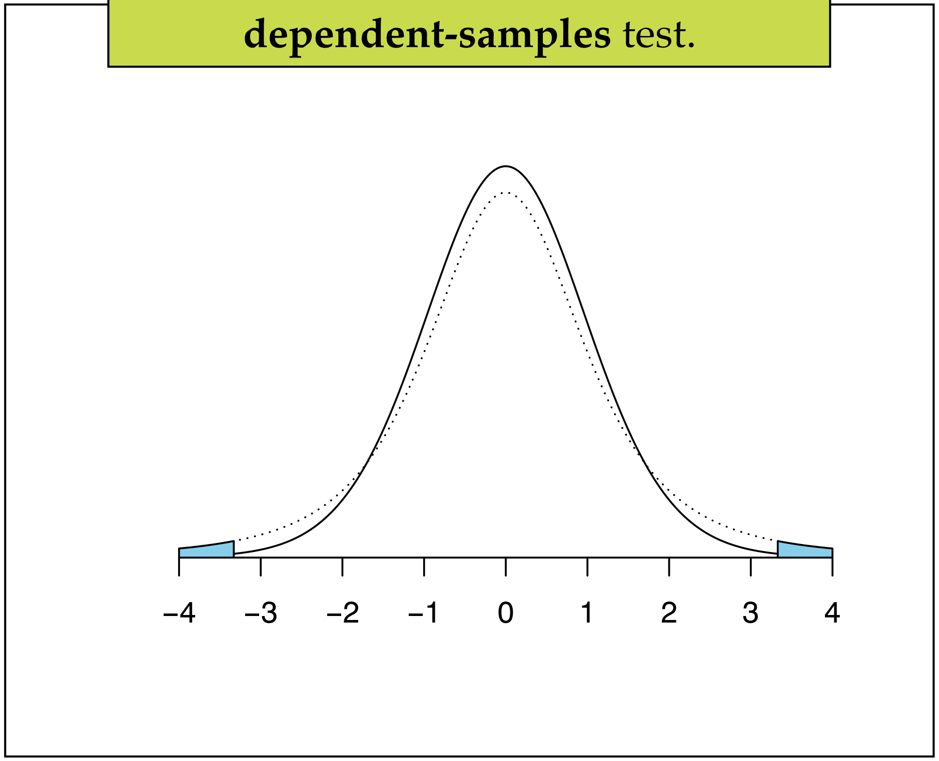


If the *t*-value were the same, the *p*-value would be smaller in the *t*(20) distribution than the *t*(10) distribution….larger *N* = more power.

The bigger difference is in the *SE* which affects the size of the *t*-value.



two-tailed *p*-value for a *t*-value of 1.05 in the *t*(20) distribution based on the **independent-samples** test.



two-tailed *p*-value for a *t*-value of 3.33 in the *t*(10) distribution based on the **dependent-samples** test.

Using the analysis that accounts for the correlation in measurements not only appropriately models the data, but also typically yields more statistical power!

References and Source Material

- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. New York: Wiley.
- Kohli, N. (2016). *Introduction to longitudinal data analysis (course notes)*. Minneapolis, MN: Author.
- National Collegiate Athletic Association. (2016, December 12). *Graduation success rate.* Retrieved July 21, 2017, from http://www.ncaa.org/about/resources/research/graduation-success-rate

- Graphics used from Open Clip Art Library (https://openclipart.org/)