EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

March 19, 2020

Last time

- Repeated measures ANOVA
- ANCOVA

Today

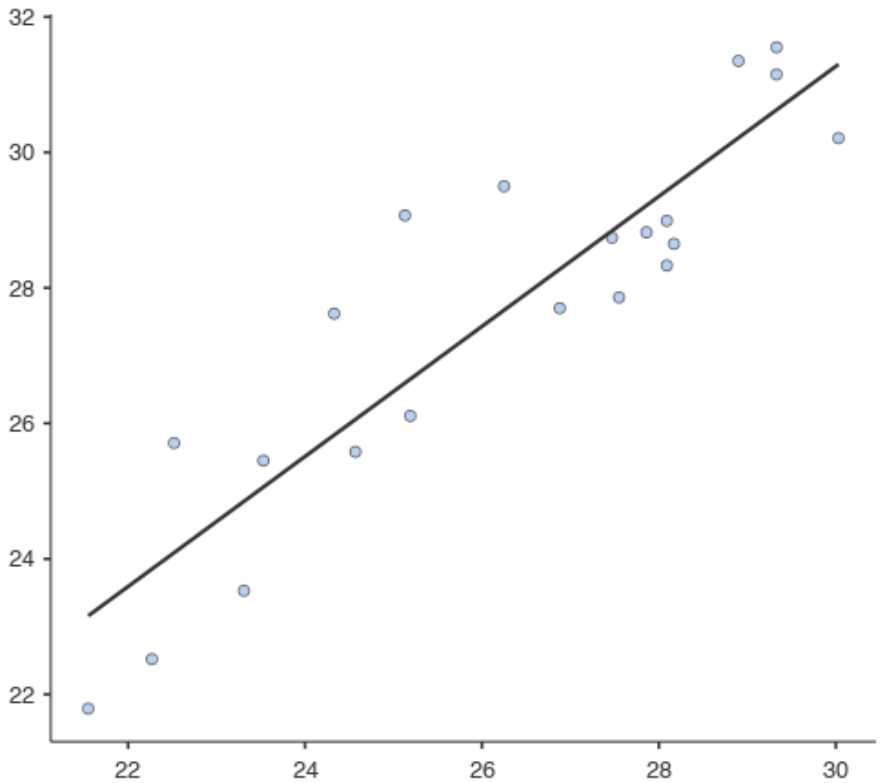
- ANCOVA review
- Case study: Motl et al. (2015)
- Starting nonparametric methods: chi-squared tests

Analysis of Covariance (ANCOVA)

- ANOVA relates a *continuous response* of interest to a set of *categorical* explanatory variables.
- Analysis of Covariance (ANCOVA) extends the ANOVA framework to allow control for *continuous* explanatory variables as well.
- This is *NOT* the same thing as regression. In particular, ANCOVA does *not* allow you to estimate the *effect* of a continuous explanatory variable on a continuous response; it only *removes* the variation explained by the continuous explanatory variable, thus:
  - reducing residual error.
  - allowing better estimates of the categorical marginal and interaction effects of interest.
- In an ANCOVA, the continuous explanatory variable is *never* of interest. It is merely a *nuisance* variable to be eliminated.

Analysis of Covariance (ANCOVA) rationale

- Let Y_i be the response of interest for sample unit i. Let X_i be the covariate (continuous explanatory variable) for sample unit i
- First, find the "best fitting" line through the points (X_i, Y_i):



Analysis of Covariance (ANCOVA) rationale

- There are many ways to define "best fitting," but here we take the classical definition; i.e. the *ordinary least squares* (OLS) fitted line obtained by *minimizing the sum of the squared errors*.
- That is, if we write

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

for some random error $\varepsilon \sim N(0, \sigma^2)$, we can find numbers $\widehat{\beta}_0$ for β_0 and $\widehat{\beta}_1$ for β_1 that minimize

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- This is a simple calculus exercise and yields the OLS estimators:

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}, \qquad \widehat{\beta}_1 = \frac{S_{XY}}{S_X^2}$$

Analysis of Covariance (ANCOVA) rationale

- Now, with the "best fitting" (OLS regression) line estimated, we can plug in the OLS estimators and rearrange the equation:

$$\begin{aligned} Y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \varepsilon_i \\ &= \bar{Y} - \widehat{\beta}_1 \bar{X} + \widehat{\beta}_1 X_i + \varepsilon_i \\ &= \bar{Y} + \widehat{\beta}_1 (X_i - \bar{X}) + \varepsilon_i \end{aligned}$$

  Thus,

$$Y_i - \widehat{\beta}_1 (X_i - \bar{X}) = \bar{Y} + \varepsilon_i$$

- Denote the lefthand side of this equation by

$$Y_i^{adj} := Y_i - \widehat{\beta}_1 (X_i - \bar{X})$$

  This is our response of interest, Y, adjusted for the effect of the covariate X.

Analysis of Covariance (ANCOVA) rationale

- So, we now have a *transformed* version of Y that we can fit ANOVA models to. For example, if W is some categorical explanatory factor of interest for Y, we can now estimate the ANOVA model:

$$Y^{adj} = \mu + \tau_W + \delta$$

- This will give us information about the *effect of W on Y adjusted for the effect of X*.
- The classic (and most common) application: estimating the effect of some intervention Y *adjusting for baseline X* over groups of W.
- Note: we can adjust for *multiple covariates* by using the same "best fit" adjustment procedure for each covariate.

RM-ANOVA vs. ANCOVA

- Suppose we have a pre-test and post-test measurement on 21 people subjected to one of three experimental treatments (a *nested* design).
- Performing a RM-ANOVA, we could address the question of whether or not the average change in pre and post-test measurement differs among the three experimental groups.
- Or, treating the pre-test measurement as a nuisance variable, we can perform an ANCOVA to address the question of whether or not the average post-test measurement, adjusted for baseline differences in pre-test measurements, differs among the three experimental groups.
- ANCOVA quantifies *differences of post-test means* between groups (adjusted for baseline); RM-ANOVA quantifies *change from pre-test to post-test* between groups.

Assumptions of ANCOVA

- The usual ANOVA assumptions (independence, homoskedasticity, normality of residuals)
- Relationship between response and covariate is *linear*. **(check plausibility with scatterplots)**
- All regression slopes between the covariate and the response are *equal* across each level of the explanatory factor(s). **(check plausibility with improper ANCOVA and group-wise scatterplots)**
- In an RM-ANCOVA framework, the regression slopes are also *equal* over each repeated measurement (virtually *never* satisfied in practice).
- Independence of the covariate and the other explanatory factors (often suspect).

RM-ANCOVA Example (three repeated measures, covariate does *not* adjust for baseline of response variable)

- Recall RM-ANOVA example from last time: we assess 24 students on their confidence in their math abilities after participating in two weekend workshops. 8 students have not taken a math course in the last 5 years, 8 have taken a course within the past 5 years but not within the last year, and 8 have taken a course in the last year.
- Suppose we actually give the students a short math test *before* they take any of the workshops. Then their individual *math ability* measured by this test is likely correlated with their baseline *confidence*. Since this could explain important variation, we could try to fit an RM-ANCOVA to assess the effect of the workshops on *confidence*, while controlling for their baseline math *ability*.

RM-ANCOVA Example

Here is the RM-ANOVA output *without* controlling for baseline ability.

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Time	51.162	2	25.581	35.031	<.001
Time ✻ Group	18.475	4	4.619	6.325	<.001
Residual	30.670	42	0.730		

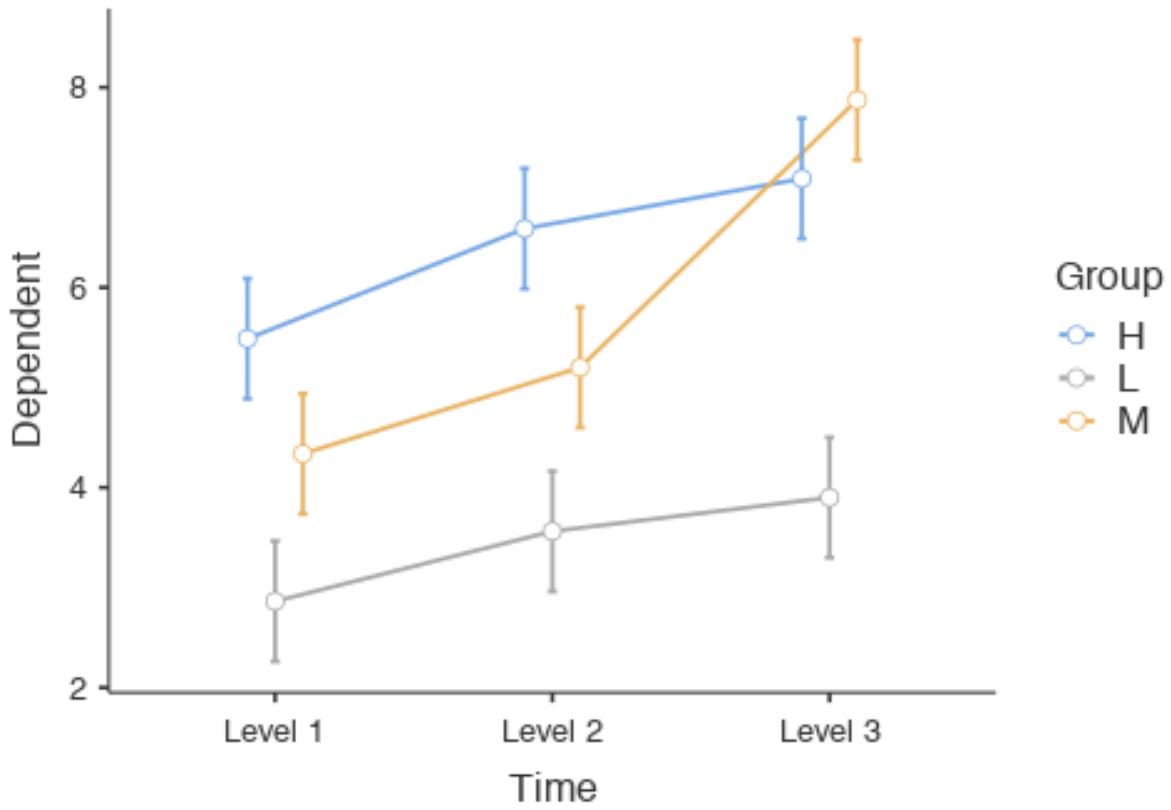
Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	116.797	2	58.398	81.826	<.001
Residual	14.987	21	0.714		

RM-ANCOVA Example

- Interaction plot (time vs. group) of RM-ANOVA *without* controlling for baseline ability.



RM-ANCOVA Example

RM-ANCOVA output, controlling for baseline ability (very different results!).

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Time	6.442	2	3.221	4.906	0.013
Time ✻ Group	3.594	4	0.899	1.369	0.264
Time ✻ Baseline.Ability	4.552	2	2.276	3.466	0.042
Time ✻ Group ✻ Baseline.Ability	3.538	4	0.885	1.347	0.271
Residual	23.636	36	0.657		

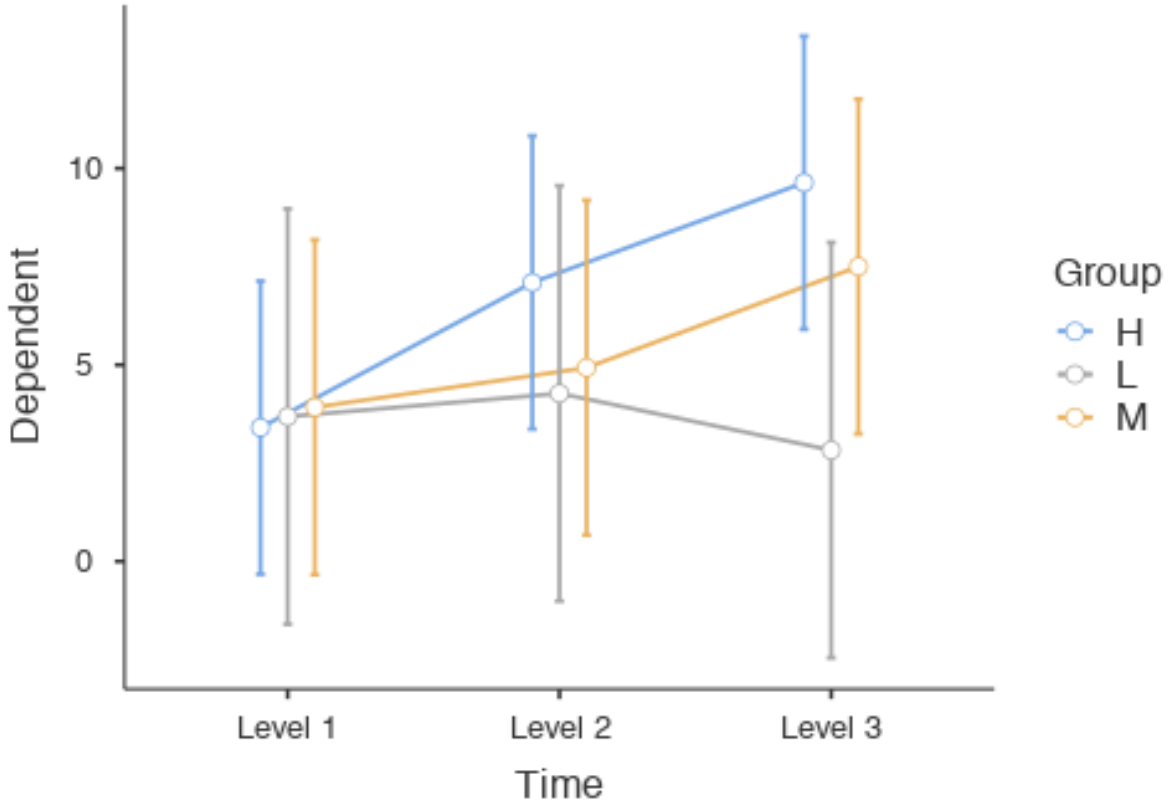
Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	1.997	2	0.999	1.283	0.301
Baseline.Ability	3.752e−4	1	3.752e−4	4.820e−4	0.983
Group ✻ Baseline.Ability	0.893	2	0.447	0.574	0.573
Residual	14.013	18	0.778		

RM-ANCOVA Example

- Interaction plot (time vs. group) of RM-ANCOVA, controlling for baseline ability.
- No significant interaction effect present.

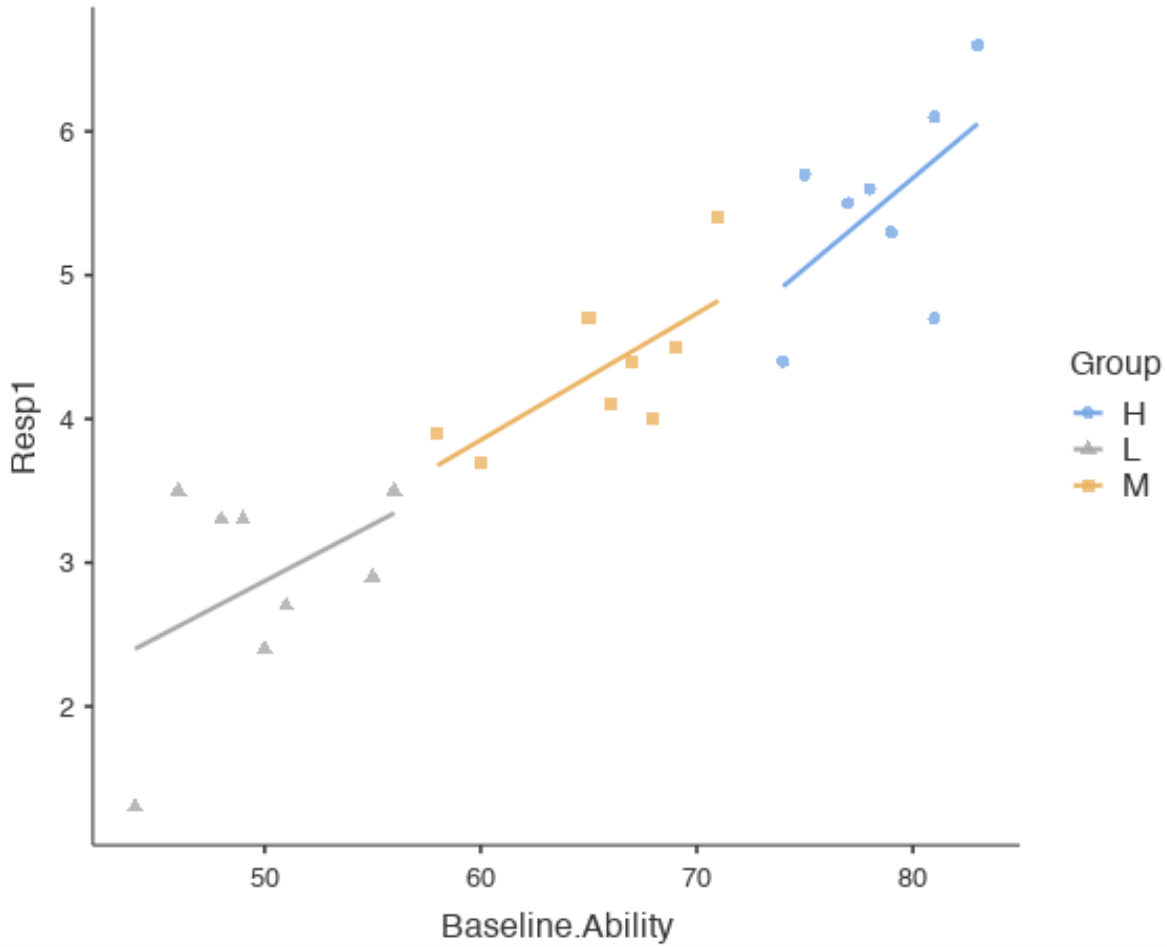


RM-ANCOVA Example

Which analysis is better?

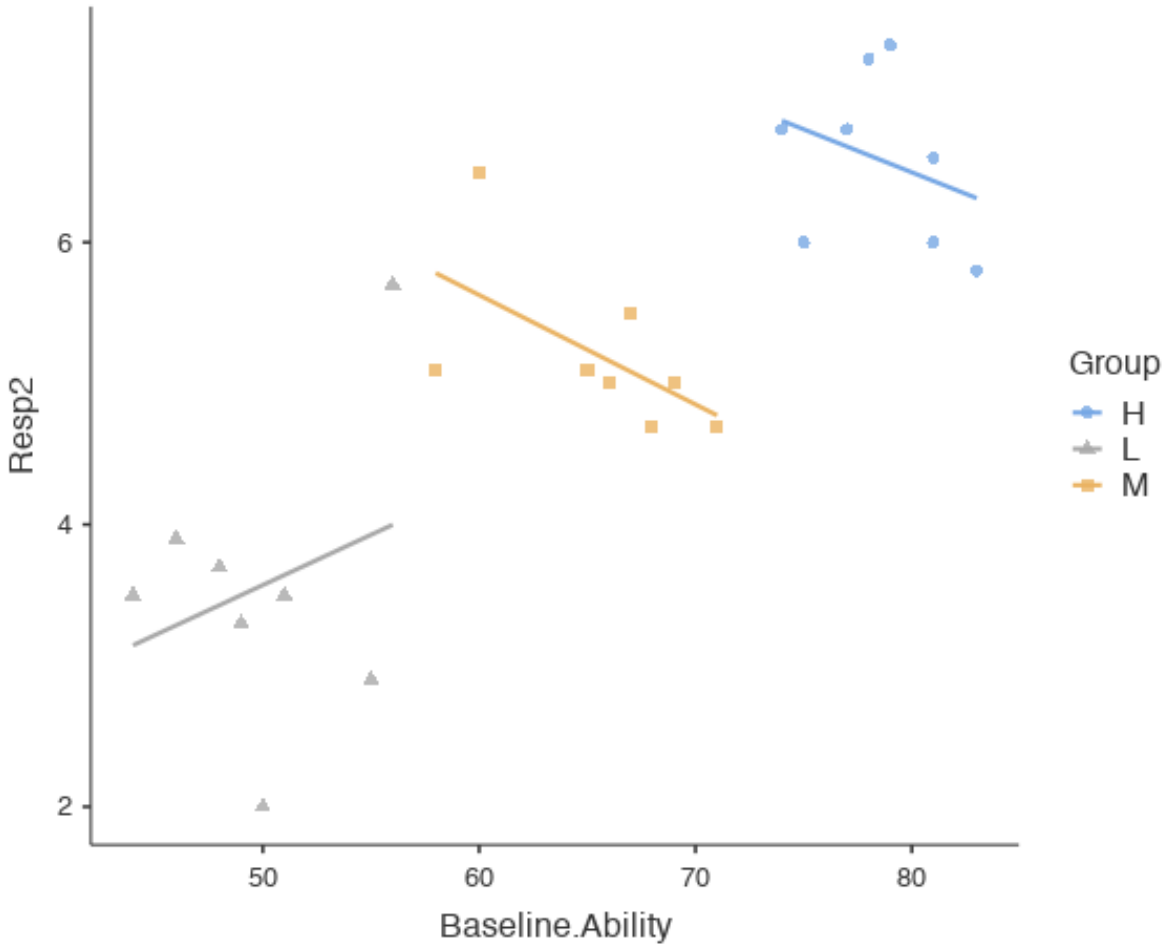
- On the one hand, the ANCOVA attempts to control for more variation in the response of interest.
- On the other hand, we have not checked the ANCOVA assumptions.
- One major assumption is that the regression "best fit" lines should be the same across all groups.
- But now, since we have repeated measures, we need to make sure that these "best fit" lines are the same across all groups *and across all time points*.

RM-ANCOVA Example



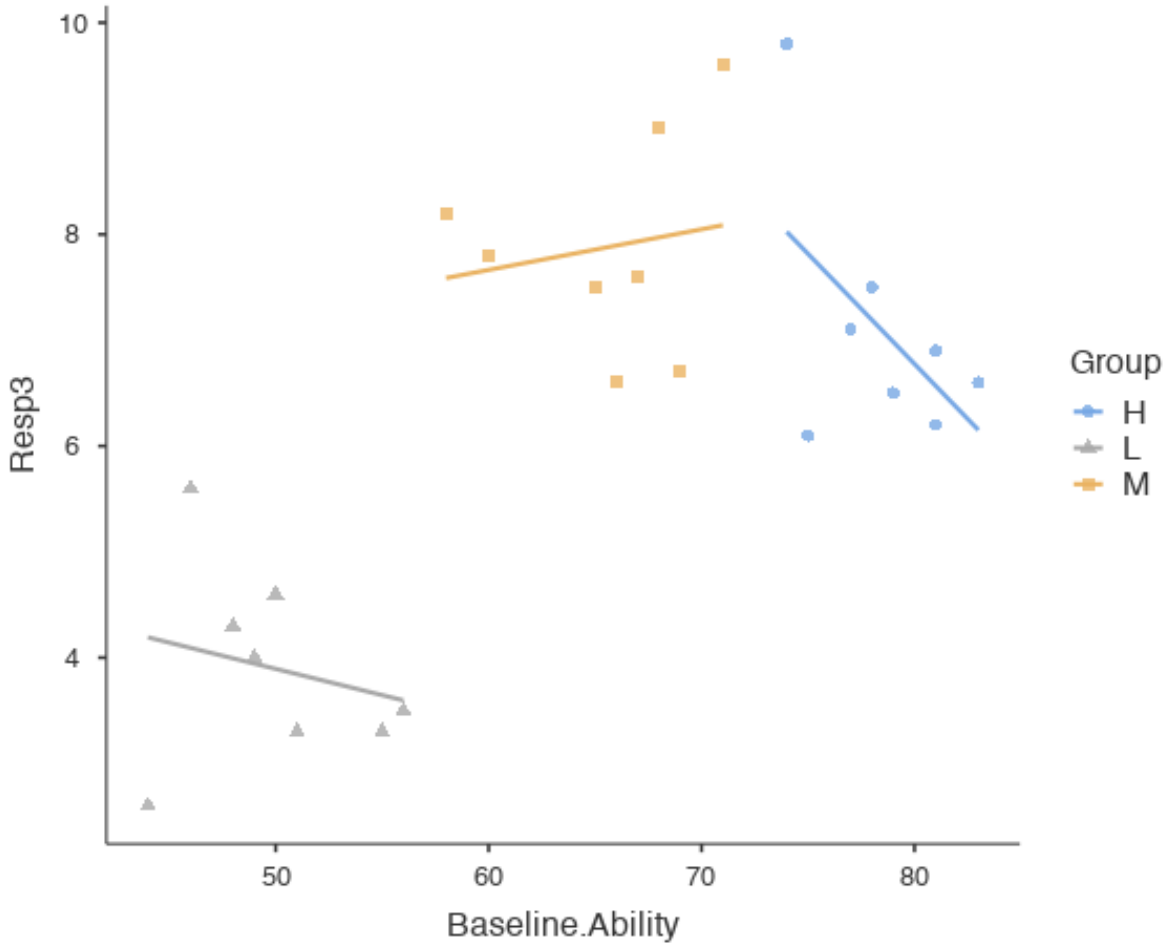
- No major evidence of heterogeneity of regression slopes at time 1.

RM-ANCOVA Example



- Definite evidence of heterogeneity of regression slopes at time 2.

RM-ANCOVA Example



- Definite evidence of heterogeneity of regression slopes at time 3.

RM-ANCOVA Example

Which analysis is better?

- Thus, the RM-ANCOVA should **not** be trusted.
- This type of violation of homogeneity of regression slopes is an extremely common occurence when trying to run an RM-ANCOVA.
- Because of this, it is *rarely* advisable to perform RM-ANCOVA.

Case Study: Motl *et al.* (2015)

- Case study posted on webpage.

Nonparametric procedures

- Analyzing categorical response data:
  - Contingency tables
  - Chi-squared tests
  - Fisher's exact test
  - Sign and McNemar tests
- Analyzing continuous *or* categorical response data:
  - Mann-Whitney and Wilcoxon tests
  - Kruskal-Wallis nonparametric one-way ANOVA
  - Friedman test
  - Permutation and randomization tests

Nonparametric Statistics

- So far, all statistical tests that we have considered (e.g. t-tests, F-tests, ANOVAs) have been examples of *parametric* procedures. A parametric test is one that makes a *distributional* assumption about the data in some way.
  - t-tests assume data are *normally distributed* (or use CLT).
  - F-tests assume data are *normally distributed* (or use CLT).
  - Traditional ANOVAs (and regressions) assume that residuals are *normally distributed*.
- In contrast, a *nonparametric procedure* does *not* make any distributional assumptions about the data.
- Next few weeks, we will become familiar with some common and extremely useful nonparametric procedures.

When to use nonparametric procedures

Consider using nonparametric procedures in the following contexts:

- When you are worried about severe violations of assumptions of robust parametric procedures (e.g. t-tests, ordinary ANOVA).
- When you are worried about mild violations of assumptions of sensitive parametric procedures (e.g. RM-ANOVA).
- When you have only nominal or ordinal response data (e.g. rankings of preferences, or *some* would argue Likert responses).
- When you have outliers in your data (robust procedures may also be available).
- When you have too little data to reasonably check assumptions of parametric procedures (remember: checking those assumptions requires enough data in *all* groups in order to have enough power to detect violations).

Chi-squared Tests

- Chi-squared (χ^2) tests are often used to test for the presence of relationships between categorical variables. For example:
  - Testing if multiple (two or more) categorical variables are independent.
  - Testing goodness-of-fit for a categorical variable; i.e. if a categorical variable follows a certain distribution.
  - Testing homogeneity of a categorical variable over all the levels of another categorical variable.
- People often refer to these three tests like they are different from each other, but all these tests are simple chi-squared tests that use *the exact same math/procedure*.
- More simply, **chi-squared tests always ask if a set of categorical variables all follow the same probability distribution.**

Chi-squared Tests rationale

- Suppose we observe the value of some categorical random variable, X, for n sample units (people).
- Suppose X has L distinct factor levels, encoded as $1, 2, \ldots, L$, and let O_i denote the number of people for whom we observe $X = i$. Our data of interest is thus a series of observed counts:

factor levels	$X = 1$	$X = 2$	$\ldots$	$X = L$
observed counts	O_1	O_2	$\ldots$	O_L

- Now suppose we hypothesize that X should follow a certain probability distribution. That is, we hypothesize

$$H_0 : \Pr(X = 1) = p_1, \ \Pr(X = 2) = p_2, \ldots, \Pr(X = L) = p_L,$$

for some *fixed numbers* $p_1, p_2, \ldots, p_L$.

Chi-squared Tests rationale

- To test this hypothesis, we can calculate the *expected counts* for each event $X = i$ for our sample of size n by using the proposed null distribution:

factor levels	$X = 1$	$X = 2$	$\ldots$	$X = L$
observed counts	O_1	O_2	$\ldots$	O_L
expected counts	$n \cdot p_1$	$n \cdot p_2$	$\ldots$	$n \cdot p_L$

- Now we can calculate a test statistic that quantifies the discrepancy between our observed and expected counts:

$$T = \sum_{i=1}^{L} \frac{(O_i - n \cdot p_i)^2}{n \cdot p_i}$$

Chi-squared Tests rationale

- *Under the null hypothesis*, this test statistic will *asymptotically* follow a chi-squared distribution on $L-1$ degrees of freedom:

$$T = \sum_{i=1}^{L} \frac{(O_i - n \cdot p_i)^2}{n \cdot p_i} \sim \chi^2(L-1) \text{ as } n \to \infty$$

- Thus, we can calculate (approximate) p-values by calculating probabilities under the χ^2 curve, a known probabiliity distribution.
- This procedure can be immediately generalized to test if a set of two or more categorical variables all follow the same probability distribution.

Chi-squared Tests: assumptions

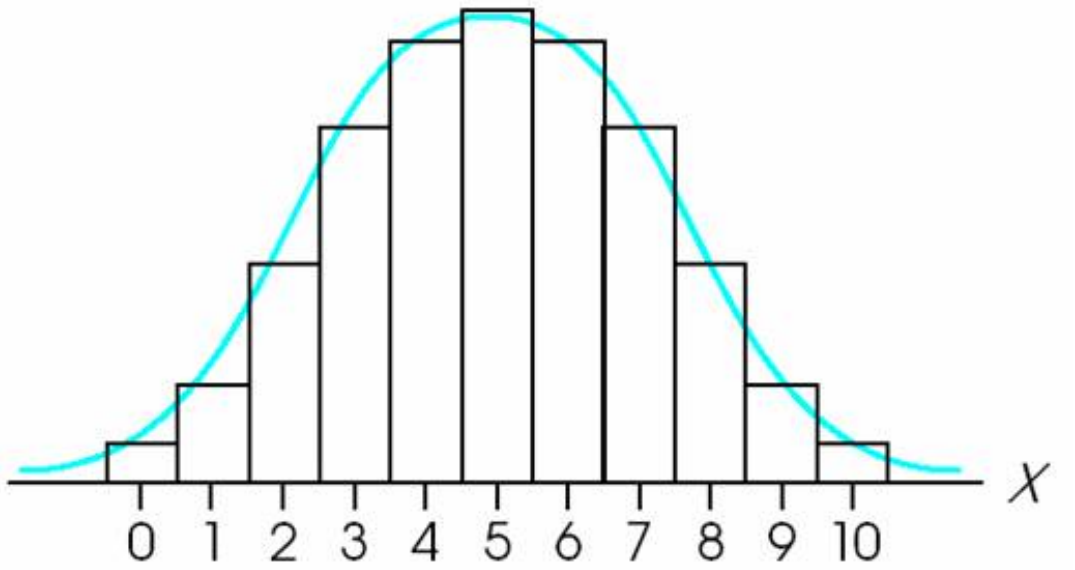
- The variable(s) being tested are categorical (nominal or ordinal)
- Test data arise from a *simple random sample*; i.e. every sample unit (person) in the study population had an equal chance of being observed.
- Observations are all independent (won't work for paired data)
- Sample size is sufficient for the asymptotics to kick in (much like the CLT)
- Expected cell counts must be large enough. Most often cited advice: at least 5 in a 2×2 table, and at least 5 in 80% of cells in larger tables, with no zero expected cell counts.

Chi-squared Tests: common errors to avoid

- You can only perform chi-squared tests on **count** data. In particular, you should *not* perform chi-squared tests:
  - on percentage or proportion data.
  - on continuous data that has been arbitrarily discretized.
  - when your total sample size is less than about 30.
- χ^2 tests are often referred to as *nonparametric*, but they still rely on parametric asymptotics to make sense.

Chi-squared Tests: common errors to avoid

- You should *always* perform a continuity correction (usually, of the Yates' variety) because a χ^2 test approximates a categorical distribution by a continuous one. Easy option to check-off in Jamovi.



- Notice: if you calculated the area under the curve between, say, 0 and 3, this would not exactly agree with if you had calculated the total area of the *histogram* from 0 to 3. A continuity correction adjusts (mostly) for this discrepancy.

Chi-squared Tests: example 1

- Suppose we sampled 96 people at UBC about their job preferences. We classified job type according to the following six categories for the survey respondents: academic, commercial (for profit), commerical (not for profit), industrial, government, and other. We observe the following data:

job types	acad.	com.(FP)	com.(NFP)	ind.	gov.	other
obs. counts	23	29	15	12	10	7

- We would like to test the hypothesis that our sample population shows *no preference* among the six job types. Thus, our particular null hypothesis here is:

$$H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$$

Chi-squared Tests: example 1

In Jamovi, enter the data as follows:

Observed or Expected	Factor levels	Total counts
Observed	1	23
Observed	2	29
Observed	3	15
Observed	4	12
Observed	5	10
Observed	6	7
Expected	1	16
Expected	2	16
Expected	3	16
Expected	4	16
Expected	5	16
Expected	6	16

- Note that the expected counts are calculated as $n \cdot p_i$ for each cell. For our particular null hypothesis of choice here, these all become:

$$96 \cdot \frac{1}{6} = 16$$

Chi-squared Tests: example 1

In Jamovi:

- Click on the 'Frequencies' tab, then select the 'Independent Samples X^2 test of association' option.
- Assign 'Observed or Expected' to the 'Rows' option; this identifies the variables to be tested.
- Assign 'Factor levels' to the 'Columns' option; this identifies the different factor levels of the categorical variables being tested.
- Assign 'Total counts' to the 'Counts' option; this identifies the observed and expected counts.
- Under the 'Statistics' menu, select the 'χ^2 continuity correction' option.

Chi-squared Tests: example 1

Contingency Tables

	Factor levels						
Observed or Expected	1	2	3	4	5	6	Total
Observed	23	29	15	12	10	7	96
Expected	16	16	16	16	16	16	96
Total	39	45	31	28	26	23	192

χ^2 Tests

	Value	df	p
χ^2 continuity correction	10.522	5	0.062

- Here, we find weak evidence against the null hypothesis. Thus we find that the data are inconsistent with the hypothesis that our target population has no preference among the six job types.
- The data are displayed in a *contingency table*.

Chi-squared Tests: example 2

- Suppose we conduct a survey asking students about their current satisfaction with their academic program. People respond on a 5-point Likert scale. In total, we survey 63 undergraduates and 61 Master's students. We would like to see if there is evidence that undergrads and Master's students report different levels of job satisfaction according to our survey.
- Thus, we would like to test the hypothesis that the distribution of our categorical Likert response is the same across both academic levels. Our particular null hypothesis here is:

$$H_0 : p_i(U) = p_i(M) \text{ for all } 1 \leqslant i \leqslant 5$$

Chi-squared Tests: example 2

Contingency Tables

	Degree		
LikertResp	U	M	Total
1	9	7	16
2	13	9	22
3	16	10	26
4	17	20	37
5	8	15	23
Total	63	61	124

χ^2 Tests

	Value	df	p
χ^2 continuity correction	4.705	4	0.319

- Here, we find no evidence against the null hypothesis. Thus we find that the data are consistent with the hypothesis that undergrads and Master's students do not report different levels of academic satisfaction.

Chi-squared Tests: example 3

- Now suppose we survey an additional 73 PhD students about their current satisfaction with their degree program.
- Again, we would like to test the hypothesis that the distribution of our categorical Likert response is the same across all program types (now there are three program types). Our particular null hypothesis here is:

$$H_0 : p_i(U) = p_i(M) = p_i(P) \text{ for all } 1 \leqslant i \leqslant 5$$

Chi-squared Tests: example 3

Contingency Tables

	Degree.Program			
Likert	U	M	P	Total
1	9	7	11	27
2	13	9	15	37
3	16	10	29	55
4	17	20	10	47
5	8	15	8	31
Total	63	61	73	197

χ^2 Tests

	Value	df	p
χ^2 continuity correction	17.712	8	0.023

- Here, we find weak evidence against the null hypothesis. Note that we do *not* require equal sample sizes among groups.

Chi-squared Tests: example 4

- Suppose we would like to test if patient survival is independent of the type of drug used in treatment. Here, we have data on 31 patients taking Drug A, and 59 patients taking Drug B. After 5 years, we have the following counts:

	Drug A	Drug B
Death	14	22
Survival	17	37

- Here, testing independence amounts to testing if the likelihood of survival (or death) is the same for both drugs. Thus, our particular null hypothesis is:

$$H_0 : \Pr(Survival \mid DrugA) = \Pr(Survival \mid DrugB)$$

Chi-squared Tests: example 4

Contingency Tables

Survival		Treatment: Drug A	Treatment: Drug B	Total
No	Observed	14	22	36
	Expected	12.400	23.600	
Yes	Observed	17	37	54
	Expected	18.600	35.400	
Total	Observed	31	59	90
	Expected	31	59	

χ^2 Tests

	Value	df	p
χ^2 continuity correction	0.248	1	0.618
Fisher's exact test	1.380		0.503

- Here, we find no evidence against the null hypothesis. Note the Fisher's exact test statistic at the bottom that seems to be corroborating the result of the χ^2 test. More on this next time.