EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

Two-Way ANOVA, in practice

Two-way fixed effects ANOVA (full) model:

$$Y = \mu + \tau_A + \tau_B + \tau_{A\times B} + \varepsilon$$

- # of obs. in each category can be different. If all the same, then the design is said to be "balanced".
- Balanced analyses have higher power and are more robust to unequal variances across categories (i.e. heteroskedasticity). They are also robust to moderate departures from normality; i.e. skewness not a big problem, but multiple modes or outliers can be.
- The interaction term, $\tau_{A\times B}$, is often of the greatest interest.
- However, need lots of data to detect meaningful interaction effects.

Two-Way ANOVA: partitioning the variance

Return to the two-way fixed effects ANOVA (full) model:

$$Y = \mu + \tau_A + \tau_B + \tau_{A\times B} + \varepsilon$$

- Recall that we worked out mathematically how a one-way ANOVA model *partitions* the observed variance in our *response variable* into two pieces:
  - (1) variance explained by the (average) differences between the explanatory (categorical) variable,
  - (2) variance leftover (attributable to within-group/individual differences).
- An analagous kind of partitioning happens when we work with a more complicated ANOVA model....

Extreme examples to clarify partitions of variance: Ex. 1

- Suppose we have these sample data on white blood cell concentration over two categorical variables: patient biological sex, X, and presence of a tumour, Z:

	$X = M$	$X = F$
$Z = No$	2.0, 2.5, 2.3	1.9, 2.3, 2.6
$Z = Yes$	1.5, 1.6, 1.1	1.6, 1.7, 0.9

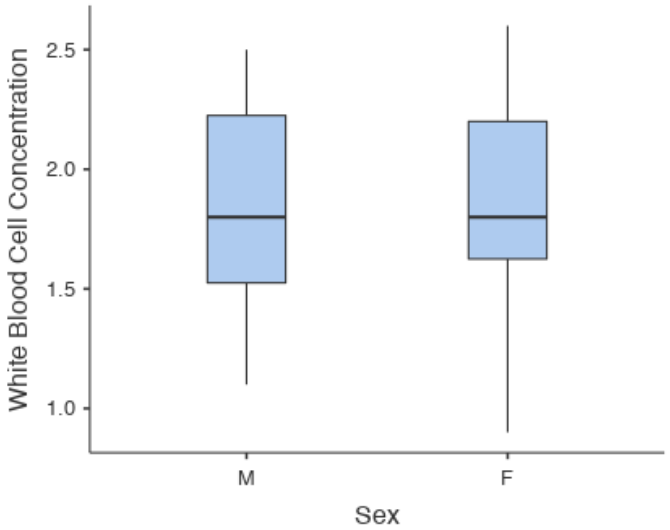
- Then:

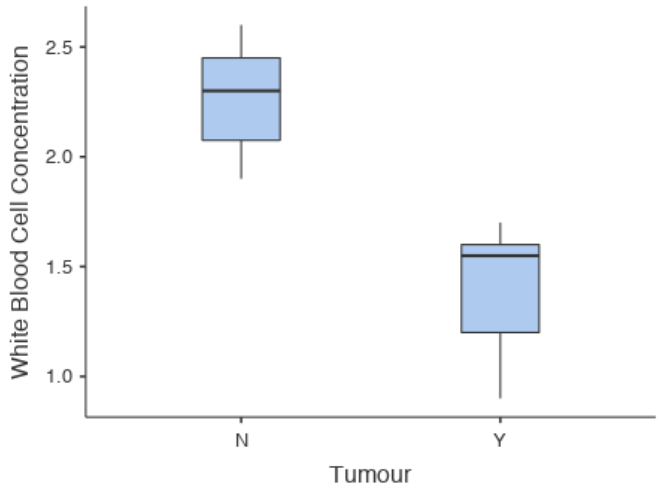
$$\overline{X_M} = 1.83,\ \overline{X_F} = 1.83,\ \overline{Z_N} = 2.27,\ \overline{Z_Y} = 1.40$$

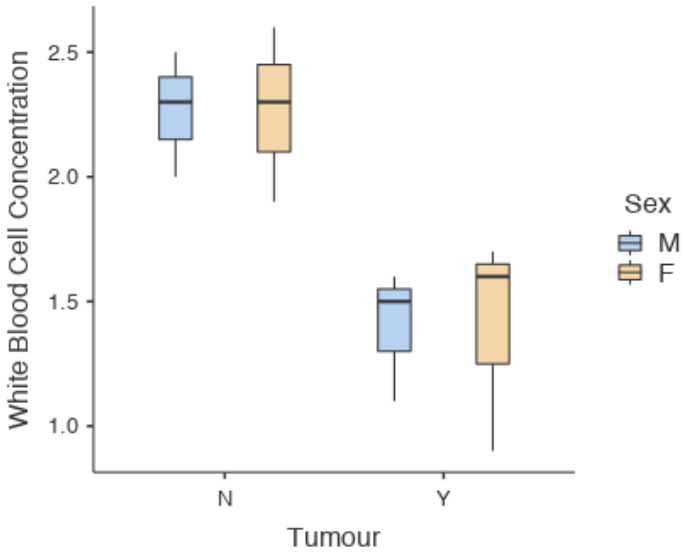
- And

$$\overline{X_M Z_N} = 2.27,\ \overline{X_M Z_Y} = 1.40,\ \overline{X_F Z_N} = 2.27,\ \overline{X_F Z_Y} = 1.40$$

Extreme examples to clarify partitions of variance: Ex. 1







Extreme examples to clarify partitions of variance: Ex. 1

ANOVA

	Sum of Squares	df	Mean Square	F	p
Sex	0.000	1	0.000	4.152e−30	1.000
Tumour	2.253	1	2.253	20.179	0.002
Sex ✻ Tumour	0.000	1	0.000	4.312e−30	1.000
Residuals	0.893	8	0.112		

- No variation explained by averaging over sex
- Clear variation explained by averaging over tumour presence
- No additional variation explained by averaging over sex$\times$tumour factor levels
- Leftover (residual) variation present from individual observations within each fixed factor level

Extreme examples to clarify partitions of variance: Ex. 2

- Now response is measure of blood pressure; experimental design assigns people to no drug, Drug A, Drug B, or both.

	Drug A No	Drug A Yes
Drug B No	2.0, 2.5, 2.3	1.6, 1.7, 0.9
Drug B Yes	1.5, 1.6, 1.1	1.9, 2.3, 2.6

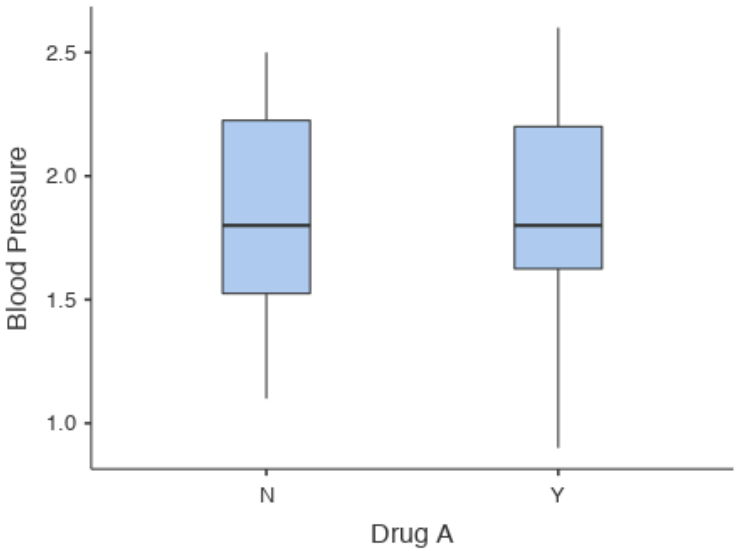
- Then:

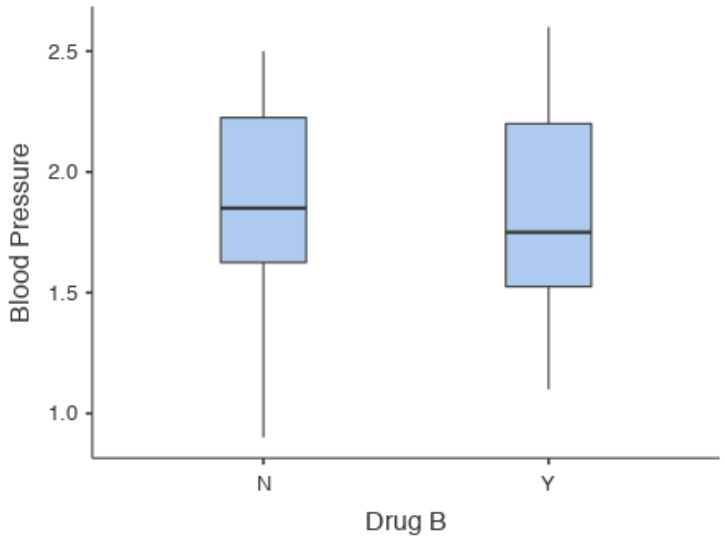
$$\overline{A_N} = 1.83,\ \overline{A_Y} = 1.83,\ \overline{B_N} = 1.83,\ \overline{B_Y} = 1.83$$

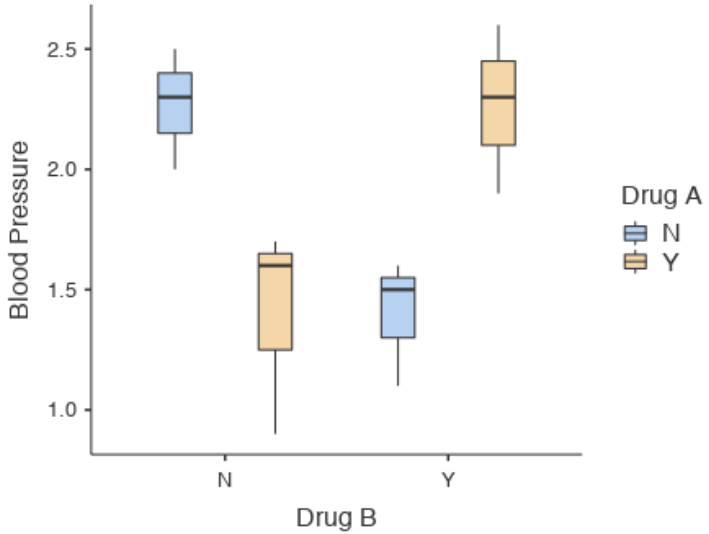
- And

$$\overline{A_N B_N} = 2.27,\ \overline{A_N B_Y} = 1.40,\ \overline{A_Y B_N} = 1.40,\ \overline{A_Y B_Y} = 2.27$$

Extreme examples to clarify partitions of variance: Ex. 2







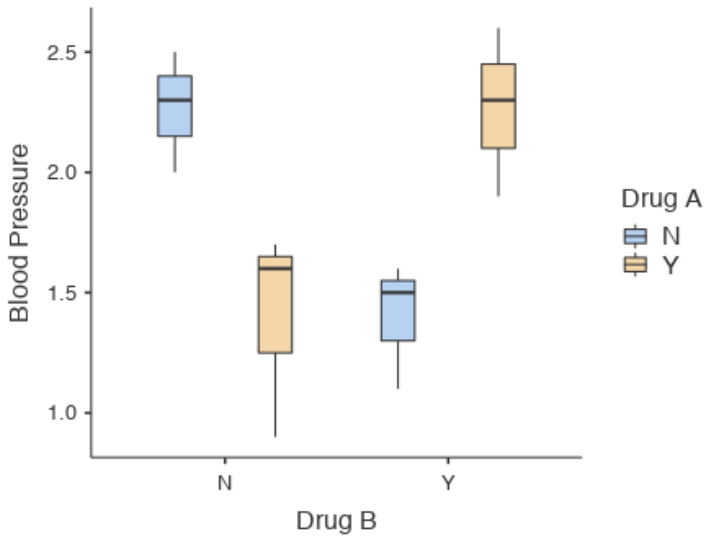
Extreme examples to clarify partitions of variance: Ex. 2

ANOVA

	Sum of Squares	df	Mean Square	F	p
Drug A	0.000	1	0.000	4.760e−30	1.000
Drug B	0.000	1	0.000	3.140e−30	1.000
Drug A ✻ Drug B	2.253	1	2.253	20.179	0.002
Residuals	0.893	8	0.112		

- No variation explained by taking Drug A and ignoring Drug B (marginal effect)
- No variation explained by taking Drug B and ignoring Drug A (marginal effect)
- Clear variation explained by considering *both* Drug A and Drug B simultaneously
- Leftover (residual) variation present from individual observations within each fixed factor level

Extreme examples to clarify partitions of variance: Ex. 2



- Those who took Drug A *only* saw blood pressure go down.
- Those who took Drug B *only* saw blood pressure go down.
- But those who took *both* drugs (or neither) have high blood pressure; drugs seem to be interacting to negate effects of treatment.

Extreme examples to clarify partitions of variance: Ex. 3

- Now suppose we have these sample data on Y over two categorical variables X and Z with 2 factor levels each:

	$X = A$	$X = B$
$Z = 1$	2.0, 2.5, 2.3	-1.2, 4.9, 3.1
$Z = 2$	0.7, 3.0, 3.1	1.9, 2.3, 2.6

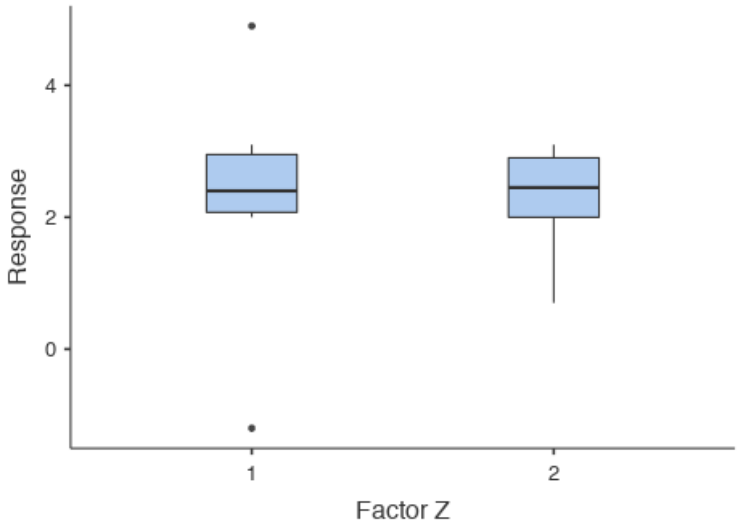
- Then:

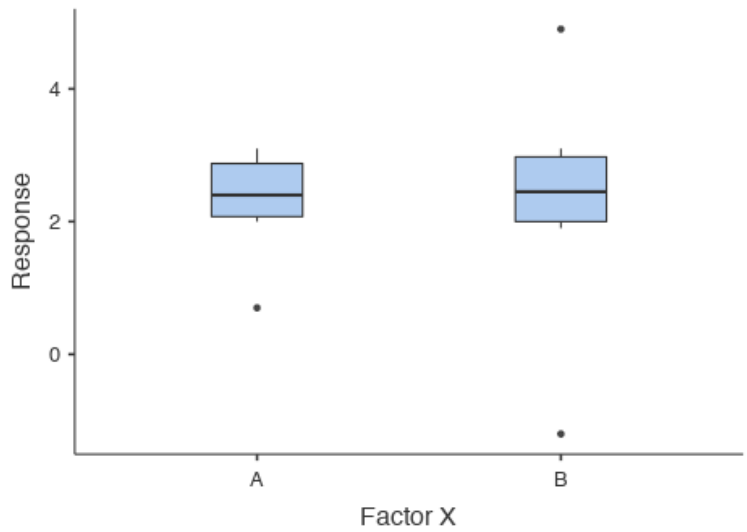
$$\overline{X_A} = 2.27,\ \overline{X_B} = 2.27,\ \overline{Z_1} = 2.27,\ \overline{Z_2} = 2.27$$

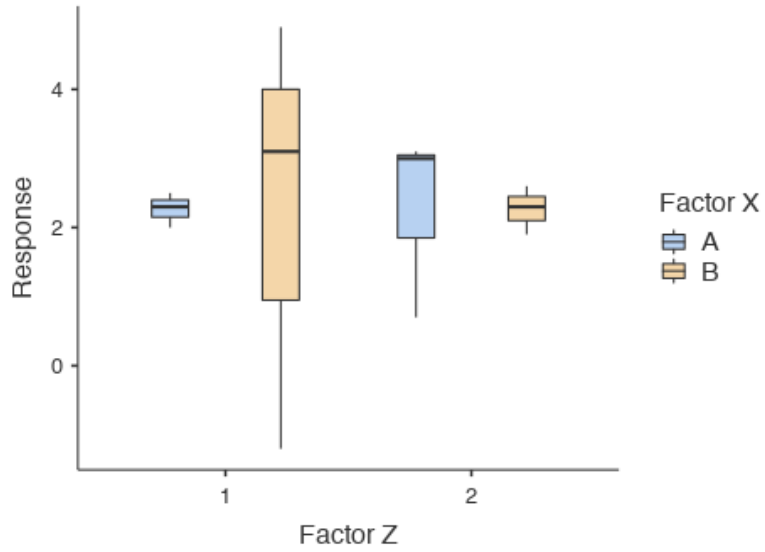
- And

$$\overline{X_A Z_1} = 2.27,\ \overline{X_A Z_2} = 2.27,\ \overline{X_B Z_1} = 2.27,\ \overline{X_B Z_2} = 2.27$$

Extreme examples to clarify partitions of variance: Ex. 3







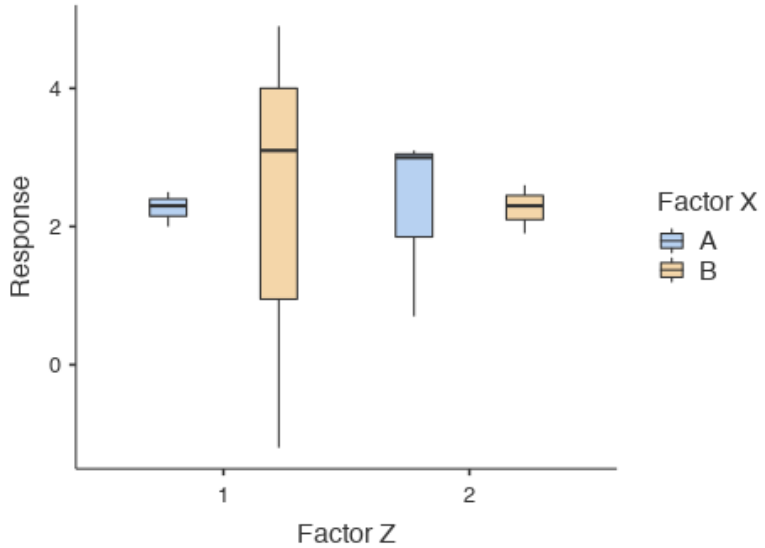
Extreme examples to clarify partitions of variance: Ex. 3

ANOVA

	Sum of Squares	df	Mean Square	F	p
Factor X	0.000	1	0.000	1.664e−32	1.000
Factor Z	0.000	1	0.000	9.359e−33	1.000
Factor X ✻ Factor Z	0.000	1	0.000	1.664e−32	1.000
Residuals	23.707	8	2.963		

- No variation explained by averaging over X factor levels
- No variation explained by averaging over Z factor levels
- No variation explained by averaging over $X \times Z$ factor levels
- All variation is residual variation from individual observations within each fixed factor level

Extreme examples to clarify partitions of variance: Ex. 3



Notice: obviously there are differences between the $X \times Z$ groups, but not *average* differences.

- ANOVAs are only able to detect *average differences* between groups.
- But there are many ways groups can be different, e.g. different variance, skewness, kurtosis, etc.
- This is why it is always important to **look at your data** (notice: ANOVA assumptions are violated); don't just rely on statistical tests of hypotheses.

Generic *n*-way ANOVAs

Nothing special about two factors; can write models with as many explanatory factors as we like.

- For example, three-way fixed effects ANOVA (full) model:

$$Y = \mu + \tau_A + \tau_B + \tau_{A\times B} + \tau_C + \tau_{A\times C} + \tau_{B\times C} + \tau_{A\times B\times C} + \varepsilon$$

- Or, for example, a four-way ANOVA with two pairwise interactions:

$$Y = \mu + \tau_A + \tau_B + \tau_C + \tau_D + \tau_{A\times C} + \tau_{B\times D} + \varepsilon$$

- Theoretically, the possibilities are endless.

Generic *n*-way ANOVAs

- However, in practice, the more complicated your model:
  (1) the more data you need to detect effects
  (2) the better experimental control you need to make sure you are isolating the effects of interest
  (3) the harder it is to diagnose your model and check your assumptions (need lots more data!)
  (4) the easier it is to fool yourself into thinking "complicated answer" means the same thing as "right answer"
- In particular, if you need about 10 people per treatment to have reasonable ability (i.e., power) to detect a meaningful average difference in treatments, then you are going to need about *twice* as many people in those groups if you want to detect a meaningful *interaction* between the treatments and a second variable.

Exchangeability of sample units

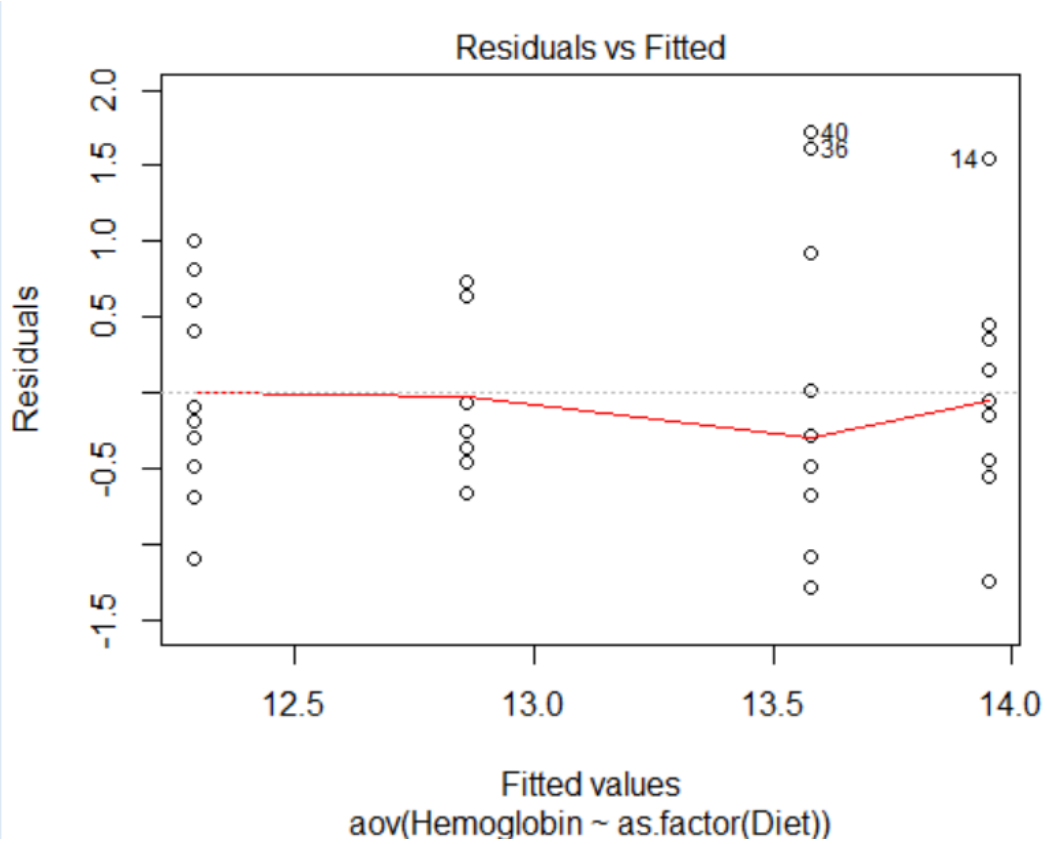
General Rule: *Never* discretize a continuum (unless you have clinical justification for doing so).

- This is an incredibly common thing in the applied literature, and it is a bad idea at least 99% of the time.
- The main reasons for doing this:
  - ANOVAs can only handle discrete predictors (i.e., grouping explanatory variables), and people only know how to perform ANOVAs (i.e., ignorance)
  - Discretizing a continuum is a form of "p-hacking"/effect-hacking/result-hacking.

Exchangeability of sample units

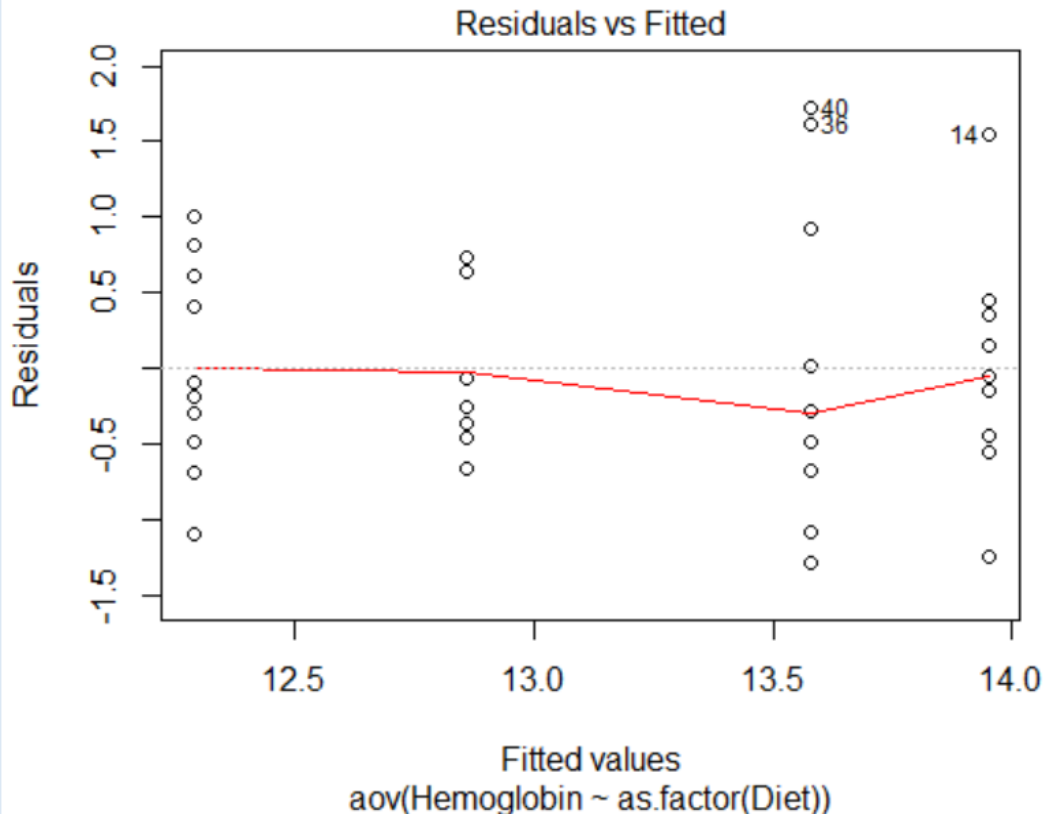
Recall the HW4 dataset from the study testing the effects of four different diet regiments on hemoglobin levels in adult pateints (10 per group, randomly assigned) with iron deficiency anemia after 6 months.

- Conditional on assigned *Diet* group, how interchagneable/exchangeable are the individual sample units?



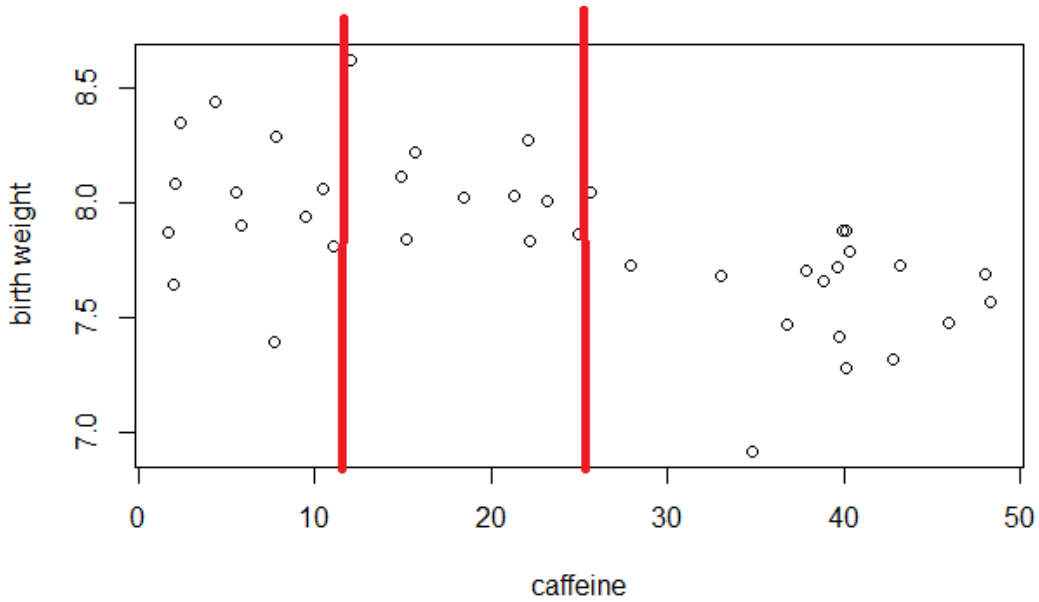
Exchangeability of sample units

- Random sampling ensures exchangeability (on average, given large enough sample sizes)
- Study design: Important to ensure reasonable balance of sample units on possible confounders (e.g., biological sex, age, health history, regular exercise, etc.) when recruting study subjects



Exchangeability of sample units

- Do NOT discretize an observed continuum because:
  - Will rely on arbitrary cutoffs/thresholds
  - Implicitly assumes that all sample units in one category are *exchangeable* (often unjustified/untenable)
  - Observations close to thresholds can create artificial effects that appear as (spurious) associations/effects in a formal analysis.



Types of sums of squares

- For ANOVAs with at least two factors, there is more than one way to decompose a total sum of squares and to define a reasonable F-test on marginal and interaction effects.
- We will *not* get into the math behind this.
- **Generally, always use the Type 3 sum of squares** (R and SPSS default to this, so basically don't worry about it).
  - Type 1: some nice mathematical properties, but order of variables matters (bad)
  - Type 2: more powerful when interactions not present (unlikely to be relevant in practice)
  - Type 3: good mathematical properties
  - Type 4: preferable in certain types of experimental designs

Measures of effect size

- Always important to report effect sizes for any comparison, statistical test, or model. For example:
  - Observed difference in two sample means (t-test)
  - Observed ratio of two sample variances or standard deviations (F-test)
  - Sums of squares or mean squares (main effects, interactions) in ANOVA
  - Regression coefficients in regressions
- Also need to report an estimate of *sample variability* (e.g. standard error, confidence interval, residual sum of squares)
- Cannot properly assess the meaning of an experiment/study without *at least*:
  - Observed effect size
  - Estimate of variability
  - Sample size(s)

Measures of effect size

- Many statistics have been developed to try to communicate these three pieces of information (i.e., observed effect size, sample variation, sample size) in a single number.
- Unfortunately, this has the effect of obscuring easily understandable statistics (e.g. means, standard deviations, counts) into obtuse derivative quantities (e.g. η^2, ω^2).
- Worst of all, since these derivative quantities are not immediately interpretable, people have developed rules of thumb for interpretation that now take the place of critical thought.
- Extremely common in social science literature (some health and natural sciences as well)

Cohen's d

- Cohen's d is the ordinary *standardized effect size* for the average difference between two groups:

$$d = \frac{\overline{X_1} - \overline{X_2}}{s},$$

  where s is an estimate of the overall standard deviation of the two groups.

- This is not an inherently bad statistic; if scales are different, standardizing can aide interpretation.

- However, Cohen's rules of thumb has become virtual gospel among applied practitioners. He advises:
  - $d \approx 0.2$ means small effect size
  - $d \approx 0.5$ means medium effect size
  - $d \approx 0.8$ means large effect size

Cohen's *d*

- **NEVER interpret your data this way.**
- First of all, it communicates nothing about *sample size*.
- Secondly, the "small, medium, large" advice of Cohen only makes sense when *all your data are normally distributed in each subgroup with equal variances* (i.e., when ANOVA assumptions hold). Even mild deviations from normality completely destroy these rules of thumb.
- Cohen's *d* is commonly reported with *t*-tests and post hoc pairwise comparisons from an ANOVA.
- Fine to report Cohen's *d*, but **avoid** the "small, medium, large" trichotomizing of effect sizes. You should be interested in clinically relevant/believable effect sizes unique to your particular research domain. And just like with *p*-values, *discretizing a continuum to make decisions is guaranteed to lead to trouble*.

Eta-squared, η^2

- η^2 is a measure of how much variation is explained by one factor (or one interaction) in an ANOVA:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

- Again, this is not an inherently bad statistic; we have been informally calculating it everytime we look at an ANOVA table.
- However, the "proportion of total variance explained" interpretation only holds when:
  - group sizes are all equal (i.e. balanced ANOVAs)
  - there are no repeated measures (will study these ANOVAs soon)

Eta-squared, η^2

- η^2 can be useful for heuristics, but it can also hide a lot of important info:
  - Again, it communicates nothing about *sample size*.
  - It can hide the fact that your data don't explain much variation at all.
  - Again, the (tenuous) interpretation breaks down for non-normal data.
- η^2 is always a *biased* estimator of the true variance explained.
- **Note:** η^2 for ANOVAs is the direct analogue of R^2 for regression models.
- Again, there are ill-advised rules of thumb for interpretation ($0.01 \approx$ small, $0.06 \approx$ medium, $0.14 \approx$ large): **NEVER use these.**

Partial eta-squared, $\eta^2_{partial}$

- $\eta^2_{partial}$ is a measure of how much variation is explained by one factor (or one interaction) relative to the residual variation:

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

- This is a bit more obscure (i.e. less intuitively interpretable) of a statistic.
- This is no longer "proportion of total variance explained" in any sense.
- This is a comparison of effect variance to residual variance.
  - Works when group sizes are not all equal (i.e. unbalanced ANOVAs)
  - Works with repeated measures (will study these ANOVAs soon)

Partial eta-squared, $\eta^2_{partial}$

- $\eta^2_{partial}$ hides and obscures a lot of important info:
  - Again, it communicates nothing about *sample size*.
  - Again, it can hide the fact that your data don't explain much variation at all.
  - Again, the (tenuous) interpretation breaks down for non-normal data.
  - It will *automatically increase* as you add more terms to your ANOVA model, since the leftover variation, SS_{error}, will automatically go down.
- $\eta^2_{partial}$ is again always a *biased* estimator of the true variance explained.
- **Note:** $\eta^2_{partial}$ for ANOVAs is analogous to $R^2_{partial}$ for regression models.
- Again, there are ill-advised rules of thumb for interpretation ($0.01 \approx$ small, $0.06 \approx$ medium, $0.14 \approx$ large): **NEVER use these.**

Omega-squared, ω^2

- ω^2 is a measure of how much variation is explained by one factor (or one interaction) relative to the total and residual variation:

$$\omega^2 = \frac{SS_{effect} - df_{effect} \cdot MS_{error}}{SS_{total} + MS_{error}}$$

- This is a *lot* more obscure of a statistic.
- It tries to again mimic the "variance explained by the effect of interest" paradigm.
- This is a comparison of effect variation to total and residual variation.

Omega-squared, ω^2

- ω^2 hides and obscures a lot of important info:
  - Again, it communicates nothing about *sample size*.
  - Again, it can hide the fact that your data don't explain much variation at all.
  - Again, the (obscure) interpretation breaks down for non-normal data.
- ω^2 is again always a *biased* estimator of the true variance explained, although not as badly biased as η^2 or $\eta^2_{partial}$.
- **Note:** ω^2 for ANOVAs is analogous to $R^2_{adjusted}$ for regression models.
- Again, there are ill-advised rules of thumb for interpretation (0.01 $\approx$ small, 0.06 $\approx$ medium, 0.14 $\approx$ large): **NEVER use these.**

If a journal requires you to report these statistics, then fine. But do NOT rely on them for your own purposes.