Section 2

Completely Randomized Designs

The slides for this class are adapted from multiple sources:

- The main outline comes from *ANOVA and Mixed Models: A Short Introduction Using R*, by Lukas Meier (https://stat.ethz.ch/~meier/teaching/anova/index.html)
- *Introduction to Design and Analysis of Experiments*, by George W. Cobb
- Notes from prior semesters created by William Christensen and Dennis Tolley

Learning Outcomes

This section and the associated R examples and assignments achieve the following course expected learning outcomes:

- **Data Import:** Create datasets in R from space-, comma-, tab-delimited files
- **Summary Statistics:** Compute summary statistics from R datasets
- **Create Graphics:** Create graphics in R for exploratory data analysis and communicating results
- **Analyze Data:** Analyze data from 'Treatment-Control' or 'A/B' experiments using professional statistical software
- **Randomized Design:** Analyze data from completely randomized designs using professional statistical software
- **Variability:** Understand the concept of variability in data and the attempt to identify sources of that variability
- **Writing Statistical Models:** Practice writing statistical models
- **Sample Size:** Demonstrate the impact of increasing the number of replicates on confidence intervals and hypothesis tests

2. Completely Randomized Designs

2.1 One-Way Analysis of Variance

2.2 Checking Model Assumptions

2.3 Nonparametric Approaches

2.4 Power or "What Sample Size Do I Need?"

2.5 Adjusting for Covariates

2.6 Appendix

2.1 One-Way Analysis of Variance (ANOVA)

Research Question: Does Orville brand microwave popcorn taste better than the generic brand?

- We purchase 4 bags of Orville popcorn and 4 bags of generic popcorn.
- We assume all bags of popcorn are essentially the same. Any potential differences can be summed up as being due to the brand.
- We have only 2 brands: Orville and generic
- This is called a "completely random design" or **CRD design structure**
- **Response variable**: flavor score (quantitative)
- **Explanatory variable**: brand (categorical)
- The **explanatory variable** is a **factor** with 2 **levels**.
- We wish to compare the mean flavor score for each brand to see if the brand affects flavor score. Similarly, is there a difference in the flavor score means for each brand?

2.1 One-Way Analysis of Variance (ANOVA)

Former STAT 230 Student Project

Research Question: Are all stain removers equally effective?

- 48 fabric samples were stained using various methods.
- 4 stain removers were considered: Tide, Oxiclean, Shout, and Miss Mouth's Messy Eater
- We assume all stained fabric samples are essentially the same.
- Each of the 4 stain removers was assigned at random to treat 12 of the 48 fabric samples.
- This is called a "completely random design" or **CRD design structure**
- **Response variable**: cleaning effectiveness score (quantitative)
- **Explanatory variable**: stain remover (categorical)
- The **explanatory variable** is a **factor** with 4 **levels**.
- We wish to compare the mean effectiveness for each stain remover to see if the stain remover affects the average cleaning effectiveness. Similarly, is there a difference in the mean cleaning effectiveness score for each stain remover?

2.1 One-Way Analysis of Variance (ANOVA)

Typical of ANOVA:

- **Response variable** is *quantitative*
- **Explanatory variable** is *categorical*
- Assume the experimental units are **homogenous** (if they are not, use a different analysis, **blocking**)
- We wish to **compare multiple means**
  - When we have only one explanatory variable, it is considered a **one-way treatment structure**

R code for random assignment of 4 treatments to 48 subjects

If we wish to randomize a total of 48 fabric samples (experimental units) to the four treatments (Tide, Oxiclean, Shout, Miss Mouth's Messy Eater), labeled "Tide," "Oxi," "Shout," "Mouth" using a balanced design with 12 experimental units per treatment, we can use the following R code:

```{r}
treat.ord <- rep(c("Tide", "Oxi", "Shout", "Mouth"), each = 12)
treat.ord ## display the ordered treatments
```

```
 [1] "Tide"  "Tide"  "Tide"  "Tide"  "Tide"  "Tide"  "Tide"  "Tide"  "Tide"
"Tide"  "Tide"  "Tide"  "Oxi"
[14] "Oxi"   "Oxi"   "Oxi"   "Oxi"   "Oxi"   "Oxi"   "Oxi"   "Oxi"   "Oxi"   "Oxi"
"Oxi"   "Shout" "Shout"
[27] "Shout" "Shout" "Shout" "Shout" "Shout" "Shout" "Shout" "Shout" "Shout"
"Shout" "Mouth" "Mouth" "Mouth"
[40] "Mouth" "Mouth" "Mouth" "Mouth" "Mouth" "Mouth" "Mouth" "Mouth" "Mouth"
```

```{r}
sample(treat.ord) ## randomize
```

```
 [1] "Tide"  "Shout" "Tide"  "Mouth" "Oxi"   "Tide"  "Oxi"   "Tide"  "Tide"  "Oxi"
"Mouth" "Tide"  "Oxi"
[14] "Mouth" "Tide"  "Shout" "Shout" "Shout" "Mouth" "Tide"  "Mouth" "Shout"
"Mouth" "Tide"  "Mouth" "Shout"
[27] "Mouth" "Shout" "Shout" "Tide"  "Shout" "Oxi"   "Oxi"   "Tide"  "Oxi"
"Shout" "Mouth" "Oxi"   "Mouth"
[40] "Oxi"   "Shout" "Mouth" "Oxi"   "Tide"  "Oxi"   "Shout" "Mouth" "Oxi"
```

This means that the first fabric sample will be treated by *Tide*, the second by *Shout*, and so on.

Means Model

Population means model for our data:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where

- μ_i is the expected value of treatment group I
- $\epsilon_{ij} \sim i.i.d.\ N(0, \sigma^2)$ random errors/residuals
- σ^2 is the variance of the Normal distribution

Y_{ij} is a random variable that represents our data. Our model suggests that our data come from a system where the outcomes are independent and identically distributed (Normal) with potentially a different mean for each treatment (μ_i) and the same variance (σ^2)for all treatments.

This is referred to as a **means model**. It is also referred to as the "**cell means model**."

Means Model Example: Popcorn Flavor by Brand

Population model for our data:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Brand	Flavor
orville	36
orville	28
orville	63
orville	81
generic	47
generic	42
generic	67
generic	85

Summary Statistics

Brand	N	Mean	Standard Deviation
generic	4	60.25	19.72
orville	4	52.00	24.45

Means Model Example: Popcorn Flavor by Brand

Decomposition of our data in the context of our model:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Observed Values

Brand	Flavor
orville	36
orville	28
orville	63
orville	81
generic	47
generic	42
generic	67
generic	85

=

Factor Means

Brand	Mean Flavor
orville	52.00
generic	60.25

+

Error

Error
-16
-24
11
29
-13.25
-18.25
6.75
24.75

This gives us an estimate of our error (or residual).

Example: Plant Growth Data

Consider the following data from a plant growth study that measured the dry weight of plants under three conditions (ctrl, trt1 and trt2). There were 10 observations for each group.
<Section2_PlantGrowth.R Part 1>

weight	group
4.17	ctrl
5.58	ctrl
5.18	ctrl
6.11	ctrl
...	

Effects Model

Population **means model** for our data:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

We can rewrite it as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where

- μ is the overall mean of the data (sometimes referred to as an *intercept*)
- α_i is the **effect** of treatment i on the response
- $\epsilon_{ij} \sim i.i.d.\ N(0, \sigma^2)$ random errors/residuals

This is called the **effects model** because it explicitly identifies the treatment effect, α_i, as a component of the model.

Note: $\mu_i = \mu + \alpha_i$, which also means that $\alpha_i = \mu_i - \mu$

Effects Model Example: Popcorn Flavor by Brand

Decomposition of our data in the context of our model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Observed Values

Brand	Flavor
orville	36
orville	28
orville	63
orville	81
generic	47
generic	42
generic	67
generic	85

Overall Mean

= 56.125 +

Factor Effects

Brand	Mean Flavor
orville	-4.125
generic	4.125

+

Error

Error
-16
-24
11
29
-13.25
-18.25
6.75
24.75

Factor Means

Brand	Mean Flavor
orville	52.00
generic	60.25

This gives us an estimate of our error (or residual).

Knowledge Check

Does Generic "King Cola" taste just as good as Coke?

Decomposition of our data in the context of our model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Observed Values

Brand	Flavor
Coke	32
Coke	
Coke	54
Coke	66
King	
King	40
King	77
King	80

Overall Mean

= 52.375 +

Factor Effects

Brand	Mean Flavor
Coke	-7.875
King	

+

Error

Error
-18.5
9.5
21.5
-16.25
-20.25
19.75

This gives us an estimate of our error (or residual).

Can you solve for the highlighted values in the model?

Knowledge Check

Does Generic "King Cola" taste just as good as Coke?

Decomposition of our data in the context of our model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Observed Values

Brand	Flavor
Coke	32
Coke	26
Coke	54
Coke	66
King	44
King	40
King	77
King	80

Overall Mean

= 52.375 +

Factor Effects

Brand	Mean Flavor
Coke	-7.875
King	7.875

+

Error

Error
-12.5
-18.5
9.5
21.5
-16.25
-20.25
16.75
19.75

This gives us an estimate of our error (or residual).

Conceptual understanding of the differences in the models

Means Model: $Y_{ij} = \mu_i + \epsilon_{ij}$

Effects Model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Hypothesis Testing for Means Model: For the means model we can compare means as we have before with our t-tests.

H_0: μ_i are all equal

H_A: At least one μ_i differs from the others

Hypothesis Testing for Effects Model: For the effects model we can test whether or not the factor "has an effect" on the mean response.

H_0: $\alpha_i = 0, \forall i$

H_A: At least one $\alpha_i \neq 0$

In a way we are testing the same thing, but from a different perspective.

Constraints that allow the effects model to work

The effects model has too many parameters to estimate. So, a constraint must be put in place to remove one of the parameters. Here are three constraints, along with the R implementations:

Name	Side Constraint	Meaning of μ	R
weighted sum-to-zero	$\sum_{i=1}^{g} n_i \alpha_i = 0$	$\mu = \frac{1}{N} \sum_{i=1}^{g} n_i \mu_i$	-
sum-to-zero	$\sum_{i=1}^{g} \alpha_i = 0$	$\mu = \frac{1}{g} \sum_{i=1}^{g} \mu_i$	`contr.sum`
reference group	$\alpha_1 = 0$	$\mu = \mu_1$	`contr.treatment`

TABLE 2.1: Different side constraints.

Constraints that allow the effects model to work

Which constraint do I use?

- They all produce the same sums of squares, F statistics, and p-values.
- The *reference group* constraint treats the first group like it is different than the others. So, if you have a control group vs. multiple treatment groups you can treat the control group as the "reference" group. Hence, $\hat{\mu}$ is the estimated mean of the control group. All other effects, $\hat{\alpha}_i$, then indicate how much each treatment group mean differs from that reference group mean.
- The sum-to-zero constraint considers all the effects as deviations from the overall mean response. So, the magnitude of each effect, $\hat{\alpha}_i$, is how much its group mean differs from the overall mean.

For the popcorn example:

Reference Group Constraint

$\hat{\mu} = 60.25, \hat{\alpha}_1 = 0, \hat{\alpha}_2 = -8.25$
Mean flavor for *generic* is 60.25
Mean flavor for *orville* is 60.25 - 8.25 = 52

Sum-to-Zero Constraint

$\hat{\mu} = 56.125, \hat{\alpha}_1 = 4.125, \hat{\alpha}_1 + \hat{\alpha}_2 = 0$
Therefore, $\hat{\alpha}_2 = -4.125$
Mean flavor for *generic* is 56.125 + 4.125 = 60.25
Mean flavor for *orville* is 56.125 - 4.125 = 52

Summation Notation

- This notation is used to **simplify** repetitive mathematical equations.
- Summation notation is the **foundation** of such formulas as **standard deviation** and **statistical models**.
- We want to understand

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- It is like building a macro or a function in computer coding.
- The capital Greek letter sigma Σ is the symbol for summation.

Summation Notation

- Assume we have the following 5 test scores:

$$88, 76, 79, 77, 82$$

- We can generically denote the 5 test scores as

$$X_1, X_2, X_3, X_4, X_5$$

- Meaning,

$$X_1 = 88, X_2 = 76, X_3 = 79, X_4 77, X_5 = 82$$

(note: we might use upper-case X-values to represent random variables and lower-case x-values to represent actual data that has already been observed)

Summation Notation

- Then, using summation notation,

$\sum_{i=1}^{5} X_i$ is the same as $X_1 + X_2 + X_3 + X_4 + X_5$
where i indexes the different observations.

- Or, more generically, we can say $\sum_{i=1}^{n} X_i$ where n represents the total sample size regardless of the size.

Summation Rules

- The distributive property states that

$$cX_1 + cX_2 + \cdots + cX_n = \mathrm{c}(X_1 + X_2 + \cdots + X_n)$$

Where c is some constant.

If $c = \frac{1}{n}$ then $\frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{(X_1 + X_2 + \cdots + X_n)}{n} = \bar{X}$

Using summation notation, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Multiplying a constant c by each observation in our data with a mean of $\bar{X}$ will make the mean of our new set of values $c\bar{X}$.

Summation Rules

- Adding or subtracting a constant from a sum

$$\sum_{i=1}^{n}(X_i + c) = (X_1 + X_2 + \cdots + X_n) + (c + c + \cdots + c) = \sum_{i=1}^{n} X_i + nc$$

$$\sum_{i=1}^{n}(X_i - c) = (X_1 + X_2 + \cdots + X_n) - (c + c + \cdots + c) = \sum_{i=1}^{n} X_i - nc$$

- What does that imply?
  - Adding a constant to each observation in our data with a mean of $\bar{X}$ will make the mean of the new set of values $\bar{X} + c$
  - Subtracting a constant from each observation in our data with a mean of $\bar{X}$ will make the mean of the new values $\bar{X} - c$

Summation Rules

Example of adding, subtracting, multiplying, and dividing a constant from scores (n) in a distribution.

n	Distribution	Distribution + 3	Distribution - 2	Distribution x 2	Distribution/2
1	5	8	3	10	2.5
2	8	11	6	16	4
3	6	9	4	12	3
4	2	5	0	4	1
5	4	7	2	8	2
Σ	25	40 or 25+15	15 or 25-10	50 or 25x2	12.5 or 25/2
$\bar{x}$	5	8 or 5+3	3 or 5-2	10 or 5x2	2.5 or 5/2

Sum of Deviations

- A deviation is an observation minus the mean.

  Deviation: $X_i - \bar{X}$

- The deviation can be positive or negative.
- The sum of all deviations is zero.

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = \sum_{i=1}^{n} X_i - n\bar{X} = \sum_{i=1}^{n} X_i - n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0$$

That is why we square deviations before adding them up. Otherwise, they would always add up to zero.

Sum of Squared Deviations

n	Distribution	Deviations	Squared Deviations
1	5	5-5 = 0	0
2	8	8-5 = 3	9
3	6	6-5 = 1	1
4	2	2-5 = -3	9
5	4	4-5 = -1	1
Σ	25	0	20
$\bar{x}$	5	0	4

This information is used to calculate standard deviations and variances.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Other important information:

- The mean of differences is equal to the difference in the means
- The mean of several means is equal to the grand mean of the individual observations that make up the several means

Using Multiple Subscripts

- We use subscripts to index our observations. For example: $X_1 + X_2 + \cdots + X_n$ can be used to represent one variable with n observations.
- We can use multiple subscripts to index multiple variables and observations.
- Assume we had three categorical variables Age (young, middle, old), Height (short, medium, tall), and Weight (under, average, over), and we have 10 subjects being measured.
- X_{ijkl} can be used to represent observation l for the i^{th} level of Age, the j^{th} level of Height, and the k^{th} level of Weight.
- So, X_{2315} represents the 5^{th} subject. That subject is middle-aged, tall, and under-weight.

Using Multiple Subscripts with dot notation

- Again, assume we had three categorical variables Age (young, middle, old), Height (short, medium, tall), and Weight (under, average, over), and we have 10 subjects being measured.
- We can use dot notation to represent sums and means.
  - $X_{\cdot}$ is a sum
  - $\bar{X}_{\cdot}$ is a mean
- Because the second subscript in X_{ijkl} represents Height, we can use $\bar{X}_{1\cdot 2\cdot}$ to represent the mean of the young-aged average-weight individuals. This is found by averaging over all the Heights and all the subjects who are at level 1 for our first variable, Age, and level 2 for our third variable Weight. The numbers represent specific levels. The dots represent summing or averaging over all levels of the respective variables.

Summation Notation: In Review

We use the following notation:

$$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij} \qquad \text{sum of group } i$$

$$y_{\cdot\cdot} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} \qquad \text{sum of all observations}$$

$$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \qquad \text{mean of group } i$$

$$\bar{y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} \qquad \text{overall (or total) mean}$$

Least Squares Estimation

We follow the ***least squares criterion***. We want to choose parameter estimates that minimize the sum of the squared deviations of our observations from the model estimates and predictions.

$$\widehat{\mu}, \widehat{\alpha}_i = \text{argmin}_{\mu,\, \alpha_i} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(y_{ij} - \mu - \alpha_i\right)^2.$$

Or equivalently, when working directly with the μ_i's, we get

$$\widehat{\mu}_i = \text{argmin}_{\mu_i} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(y_{ij} - \mu_i\right)^2.$$

Doing so, we would call $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\mu}_i$ ***least squares estimates***.

Least Squares Estimation

We can show that the estimate of the population group mean is the sample group mean:

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

Similarly, the estimate of the overall mean is the overall sample mean:

$$\hat{\mu} = \bar{y}_{\cdot\cdot}$$

and the estimated effect is the sample mean for the group minus the overall sample mean:

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$$

Using the weighted sum-to-zero constraint.

If we use the reference constraint, then the estimated effect is the sample mean for the group in question minus the sample mean for the reference group.

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{1\cdot}$$

Calculation of the estimates for the means do not change with different constraints as do the estimates for the effects.

< Section 2 Plant Growth.QMD (Part 2)>

Least Squares Estimation

Means Model: $y_{ij} = \mu_i + \epsilon_{ij}$

Residuals are the difference between our observed values y_{ij} and the cell mean $\hat{\mu}_i$

$$r_{ij} = y_{ij} - \hat{\mu}_i$$

Least Squares Estimation

The sum of the squared residuals (error) is

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (r_{ij})^2$$

The estimate of the error variance, $\hat{\sigma}^2$, also called **mean squared error**, MSE is

$$\hat{\sigma}^2 = MSE = \frac{1}{N-g} SSE$$

For a sample,

$$s_i^2 = \frac{1}{n_i - 1} SSE$$

Tests

We would now like to conduct a hypothesis test to simultaneously compare all treatment means.

H_0: μ_i are all equal

H_A: At least one μ_i differs from the others

This test assesses whether or not the differences among the treatments (signal) is significantly larger than the variability among the individuals within the groups (noise).

The first step will be to calculate all the variability in the data and partition it according to the different model components.

Tests

We would now like to conduct a hypothesis test to simultaneously compare all treatment means.

H_0: μ_i are all equal

H_A: At least one μ_i differs from the others

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2}_{SS_T} = \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_{\text{Trt}}} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}_{SS_E},$$

where

$$\begin{aligned} SS_T &= \text{total sum of squares} \\ SS_{\text{Trt}} &= \text{treatment sum of squares (between groups)} \\ SS_E &= \text{error sum of squares (within groups)} \end{aligned}$$

Hence, we have

$$SS_T = SS_{\text{Trt}} + SS_E.$$

Finally, we discuss what degrees of freedom are and how to calculate them.

- From Intro STATS:
  - One Sample T-test: df = n-1
  - Two Sample T-test: df = n-2
  - One-way ANOVA:

Source	df
Treatment	t-1
Error	n-t
Total (corrected)	n-1

  - RCBD:

Source	df
Block	b-1
Treatment	t-1
Error	(b-1)(t-1)
Total (corrected)	n-1

  - Two-way ANOVA:

Source	df
Factor A	a-1
Factor B	b-1
AB Interaction	(a-1)(b-1)
Error	n-ab
Total (corrected)	n-1

Where do degrees of freedom come from?

- Each observation provides one unit of information about chance error
- This information is distributed across your model
- Each unit in a model component provides information unless its value is determined by repetition or other constraints
- We say the units that are not constrained or repeated are "free" to vary. That is why we use the term "degrees of freedom."
- Degrees of freedom (df) are the number of free numbers (in a model component)

Observations					Grand Mean (Benchmark)					Method Effect					Residuals			
5.4	5.2	6.1	4.8	=	5	5	5	5	+	.375	.375	.375	.375	+	.025	-.175	.725	-.575
5.0	4.8	3.9	4.0		5	5	5	5		-.575	-.575	-.575	-.575		.575	.375	-.525	-.425
4.8	5.4	4.9	5.7		5	5	5	5		.2	.2	.2	.2		-.4	.2	-.3	.5

Where do degrees of freedom come from?

Observations					Grand Mean (Benchmark)					Method Effect					Residuals			
5.4	5.2	6.1	4.8	=	5	5	5	5	+	.375	.375	.375	.375	+	.025	-.175	.725	-.575
5.0	4.8	3.9	4.0		5	5	5	5		-.575	-.575	-.575	-.575		.575	.375	-.525	-.425
4.8	5.4	4.9	5.7		5	5	5	5		.2	.2	.2	.2		-.4	.2	-.3	.5

Observations					Grand Mean (Benchmark)					Method Effect					Residuals			
5.4	5.2	6.1	4.8	=	5	R	R	R	+	.375	R	R	R	+	.025	-.175	.725	S
5.0	4.8	3.9	4.0		R	R	R	R		-.575	R	R	R		.575	.375	-.525	S
4.8	5.4	4.9	5.7		R	R	R	R		S	R	R	R		-.4	.2	-.3	S

$df = 12$ $df = 1$ $df = 2$ $df = 9$

Each column adds to zero

Each row adds to zero

R = repeat, S = sum

Tests

We can summarize the partitioning of the variance in an ANOVA table.

Source	df	Sum of Squares (SS)	Mean Squares (MS)	F-ratio	P-value
Treatment	g-1	SSTrt	MSTrt = SSTrt/g-1	$\frac{MSTrt}{MSE}$	
Error (Residual)	N-g	SSE	MSE = SSE/N-g		
Total	N-1	SSTot			

If the means are similar, F-ratio will be close to 1. If the means are dissimilar, the F-ratio will be significantly larger than 1.

Tests

Consider the following ANOVA table. Use what you have learned to fill in the blanks:

Source	df	Sum of Squares (SS)	Mean Squares (MS)	F-ratio	P-value
Treatment		100			
Error (Residual)		160			
Total	20				

Note: We have 5 treatment levels.

Tests

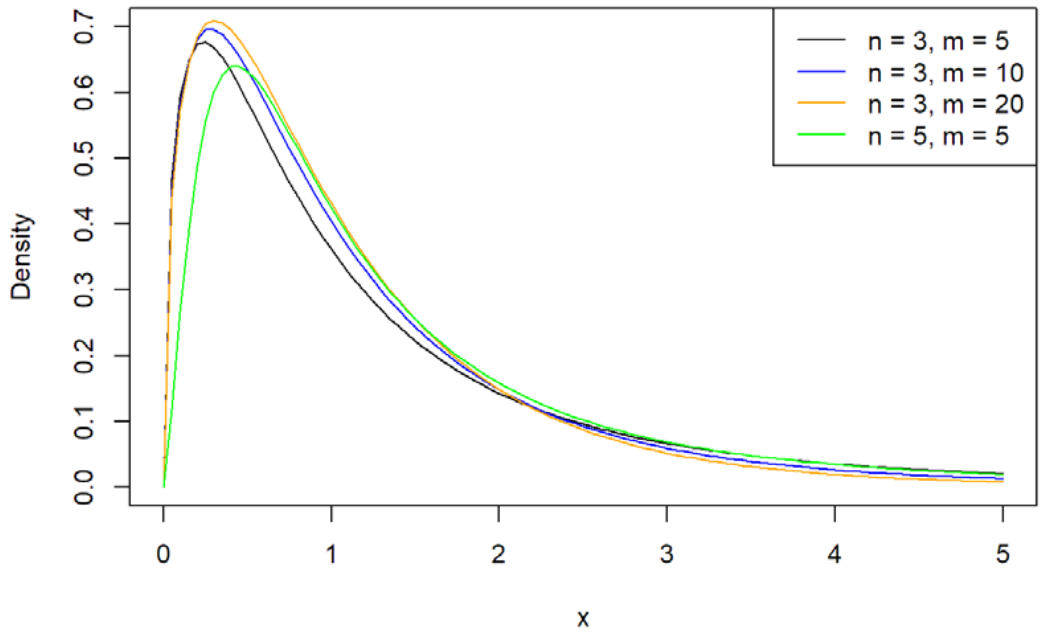
Consider the following ANOVA table. Use what you have learned to fill in the blanks:

Source	df	Sum of Squares (SS)	Mean Squares (MS)	F-ratio	P-value
Treatment	4	100	25	2.5	
Error (Residual)	16	160	10		
Total	20	260			

Note: We have 5 treatment levels.

Tests

The F-distribution is used to calculate the p-value and critical ranges.



The F-distribution is indexed by three values.

- df associated with the numerator of the F-ratio
- df associated with the denominator of the F-ratio
- The value of the test statistic (when calculating a p-value)

Tests

For the plant data example, we have the following output:

```
             Df Sum Sq Mean Sq F value Pr(>F)
group         2  3.766  1.8832   4.846 0.0159 *
Residuals    27 10.492  0.3886
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	df	Sum of Squares (SS)	Mean Squares (MS)	F-ratio	P-value
group	2	3.766	1.8832	4.846	0.0159
Residuals	27	10.492	0.3886		
Total	29	14.258			

Why is the total *df* = 29 instead of 30?

There are 30 observations. Why not 30 *df*?

- corrected vs. uncorrected *df*

We can also calculate the p-value using the following R function:

```
pf(4.846, 2, 27, lower.tail = FALSE)
[1] 0.01591099
```

< Section 2 Plant Growth.QMD (Part 3)>

Estimates and Confidence Intervals

We can compute means or effects:

Means Model: This provides us with estimates, tests, and confidence intervals for the group **means**.

```
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
groupctrl    5.0320     0.1971   25.53   <2e-16 ***
grouptrt1    4.6610     0.1971   23.64   <2e-16 ***
grouptrt2    5.5260     0.1971   28.03   <2e-16 ***

 2.5 %   97.5 %
groupctrl 4.627526 5.436474
grouptrt1 4.256526 5.065474
grouptrt2 5.121526 5.930474
```

Estimates and Confidence Intervals

Effects Model: This provides us with estimates, tests, and confidence intervals for the group **effects**.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0730     0.1138  44.573   <2e-16 ***
group1       -0.0410     0.1610  -0.255   0.8009
group2       -0.4120     0.1610  -2.560   0.0164 *

                2.5 %      97.5 %
(Intercept)  4.8394768  5.30652317
group1      -0.3712516  0.28925164
group2      -0.7422516 -0.08174836
```

Means vs. EMMeans

- EMMeans (Expected Marginal Means) are equally-weighted means, regardless of the sample size in each group
- Means, our basic arithmetic average, are actually weighted means that differ with differing sample sizes for each group

- Extreme example: 2-factors and we want to compare the means for the levels of Factor A.

Table of Means

Factor A	Factor B: 1	2	3
1	7	3	2
2	7	3	2

Means vs. EMMeans

Example: 2-factors

Factor B

Factor A	1	2	3
1	7	3	2
2	7	3	2

Table of **Means**

Factor B

Factor A	1	2	3
1	8	2	2
2	2	2	8

Table of **Sample Sizes**

Calculate Means (weighted)

Level A1: $\frac{(8x7)+(2x3)+(2x2)}{12} = 5.5$

Level A2: $\frac{(2x7)+(2x3)+(8x2)}{12} = 3.0$

There appears to be a difference in the level means for Factor A

Calculate EMMeans

Level A1: $\frac{7+3+2}{3} = 4.0$

Level A2: $\frac{7+3+2}{3} = 4.0$

There appears to be NO difference in the level means for Factor A

When to use EMMeans vs. Means

Because EMMeans are equally weighted, **use EMMeans** when the populations sampled from have approximately equal sizes, or when designing an experiment and intending for the sample sizes to be equal.

Use Means (weighted) when the populations sampled from have drastically different sizes and you want to emphasize those different sizes.

Example 1: We wish to compare the average feeling of belonging for students across the different classes (Freshmen, Sophomores, Juniors, Seniors). Which means method should we use?

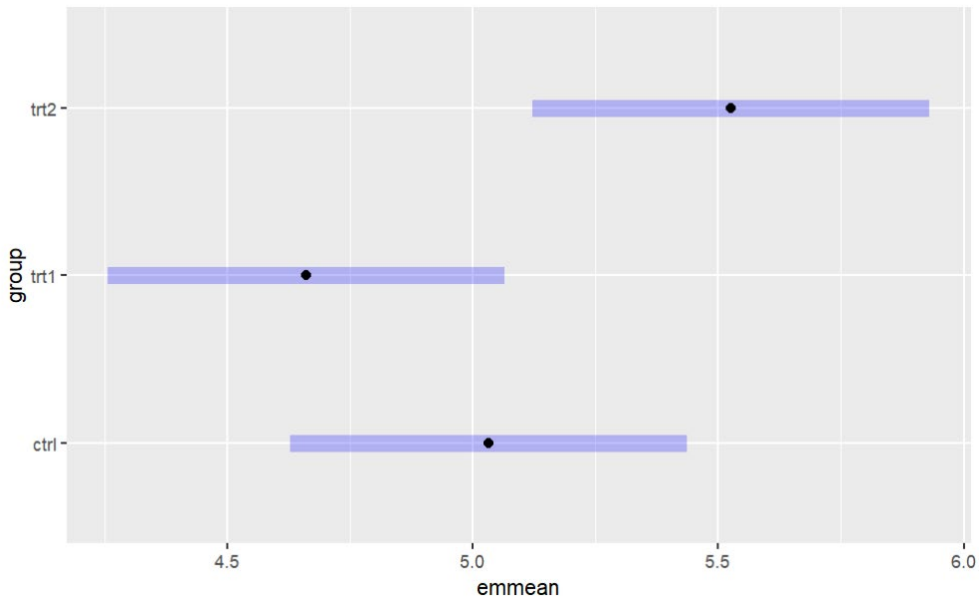
Example 2: We have designed an experiment to compare different types of paper on the distance a paper airplane can fly. Which means method should we use?

Estimates and Confidence Intervals

We can compute estimated marginal means (EMMeans):

EMMeans: This provides us with estimates, tests, and confidence intervals for the group **emmeans**. We can calculate these from either a means model or effects model.

group	emmean	SE	df	lower.CL	upper.CL
ctrl	5.03	0.197	27	4.63	5.44
trt1	4.66	0.197	27	4.26	5.07
trt2	5.53	0.197	27	5.12	5.93



< Section 2 Plant Growth.QMD (Part 4)>

Checking Model Assumptions

Previously, we stated $\epsilon_{ij} \sim i.i.d.\ N(0, \sigma^2)$

What does ***i.i.d.*** stand for?

ANOVA Assumptions:

- The errors are independent
- The errors are normally distributed
- The error variance is constant
- The errors have mean zero

Checking Model Assumptions

ANOVA Assumptions:

- The errors are independent
  - Use critical thinking to determine if the responses might be correlated in some way
- The errors are normally distributed
  - Check a Normal Q-Q plot for the points to follow a 45^o line
- The error variance is constant
  - Plot of residuals vs. fitted values should have even spread
- The errors have mean zero
  - Plot of residuals vs. fitted values should be centered at zero

< Section 2 Plant Growth.QMD (Part 5)>

Checking Model Assumptions

What do we do if the assumptions are not met?

1. Transformation: If the data are not normally distributed, we can try to transform them to *make* them normally distributed.
2. Nonparametrics: Use distribution-free tools to analyze the data
3. Generalized Linear Model (GLM): Model the data according to the distribution that more-appropriately fits the data.

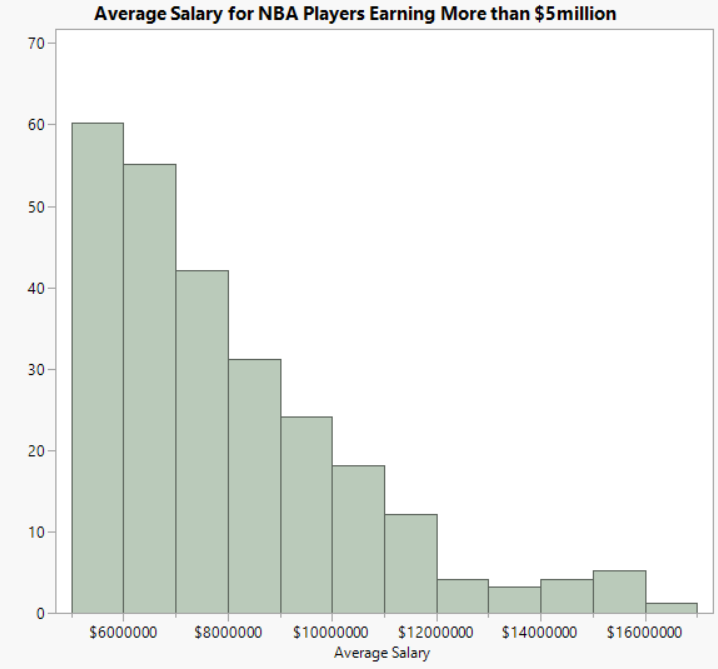
We will discuss the first two here. The third is covered in a different statistics class.

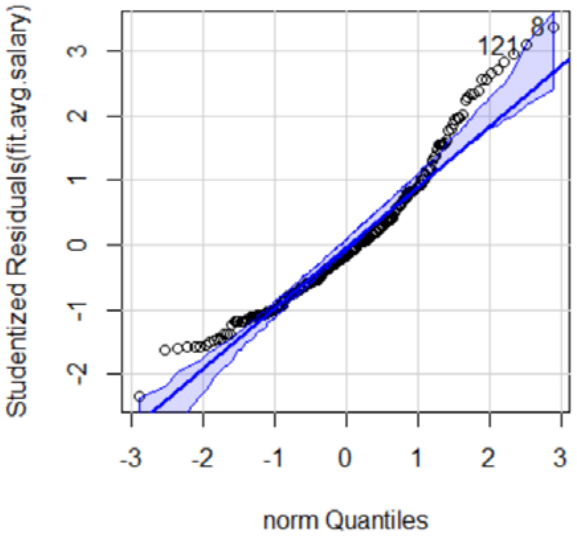
Checking Model Assumptions

Consider the following example where our assumption of normality is not met.

Ex. NBA salaries for players earning more than $5million

The data are right-skewed



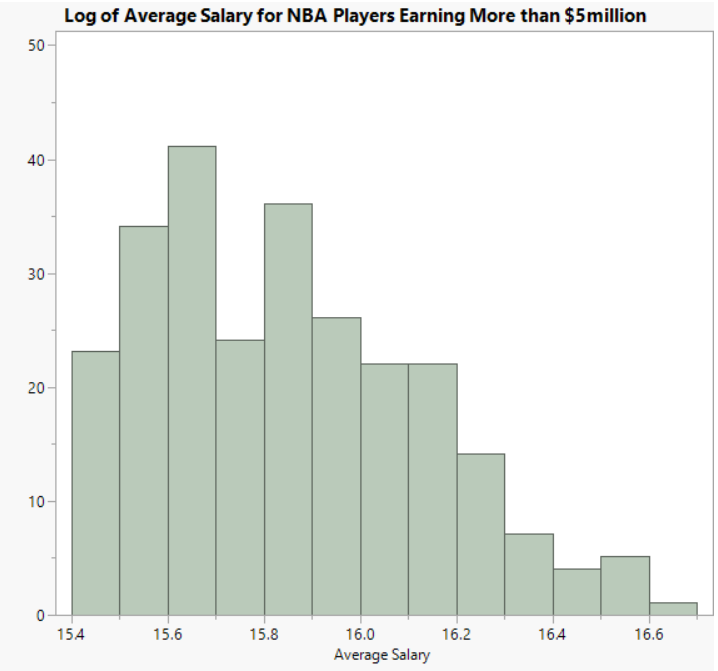


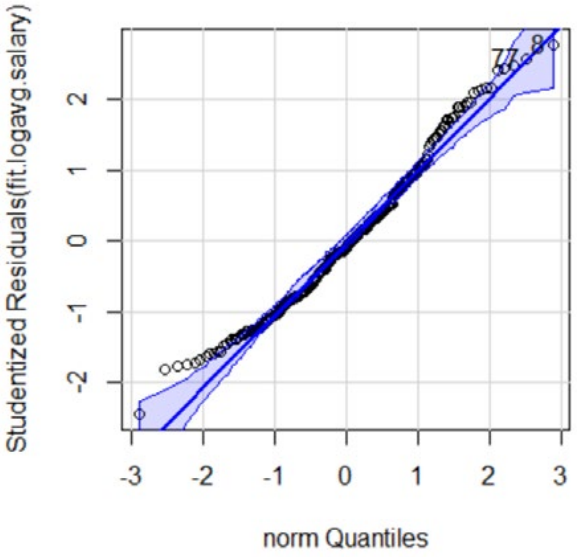
Transformations

When assumptions are not met, we might consider transforming the data so that the assumptions *are* met.

With right-skewed data we can apply a log transformation:

$log(Y_{ij}) = \mu' + \alpha'_i + \epsilon'_{ij}$ instead of $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$





Not perfect, but better

Transformations

All the results are on the log scale, log salaries rather than salaries. To properly interpret the data, we must first back-transform the results.

Although $E(Y_{ij}) = \mu + \alpha_i$,

We can't simply back-transform $E[log(Y_{ij})] \neq e^{\mu\prime+\alpha\prime_i}$ to solve for $\mu + \alpha_i$

But, we can work with medians rather than means and come up with reasonable results.

$$median[log(Y_{ij})] = \mu' + \alpha'_i$$

Which means that

$$median(Y_{ij}) = e^{\mu\prime+\alpha\prime_i}$$

That allows us to interpret our results in terms of medians rather than means. But, it is close. ☺

Nonparametric Statistics Approaches

The second solution when assumptions are not met is to use a nonparametric test.

Nonparametric tests do not rely on the data coming from a specific distribution.

The only assumptions that need to be met are

- The responses need to be independent
- The responses cannot be of ordinal type

How do these tests work?

- Signs
- Ranks
- Permutations

Nonparametric Statistics Approaches

Here are three nonparametric tests with their parametric equivalents:

Parametric Test	Nonparametric Version
Independent Samples T-test	Mann-Whitney U Test
Paired T-test	Wilcoxon Signed Rank Test
One-way ANOVA	Kruskal-Wallis Test

Mann-Whitney U Test

- Replaces independent two-sample t-test
- Assume all observations are independent of each other
- Variable type must not be nominal

H_0: The distributions of both populations are identical

H_A: The distributions of both populations are not identical

We are testing if the median of one population is larger than the median of the other population

1. Put all observations together in one group in order of size
2. Rank each observation, using 0.5 for ties
3. Sum the ranks for each population. We expect them to be the same under the null hypothesis. The further apart they are, the more we favor the alternative hypothesis.
4. Use a computer to calculate the p-value for this test

Wilcoxon Signed Rank Test

- Replaces one-sample t-test or paired t-test
- Variable type must not be nominal

H_0: The median (or median of the differences) equals zero

H_A: The median (or median of the differences) is not equal to zero

1. If paired data then calculate differences in each pair. If one-sample, skip this step.
2. Calculate the absolute value of each observation (or difference)
3. Assign ranks to each value
4. Apply a sign to each rank, + if it was initially positive and – if it was initially negative
5. The test statistic is the sum of the signed ranks
6. Use a computer to calculate the p-value for this test

Kruskal-Wallis Test

- Replaces one-way ANOVA F-test
- Variable type must not be nominal

H_0: The distributions (medians) of all populations are identical

H_A: The distributions (medians) of all populations are not identical

Steps:

1. Put all observations together in one group in order of size
2. Rank each observation, using averages of ranks for ties
3. The test statistic is calculated using this formula →

4. Use a computer to calculate the p-value for this test

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \text{ where:}$$

- N is the total number of observations across all groups
- g is the number of groups
- n_i is the number of observations in group i
- r_{ij} is the rank (among all observations) of observation j from group i
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
- $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

Power

Power is how likely we are to detect actual differences.

Ex.: A study was conducted to determine if the "turn right after stopping at a red light" traffic rule would increase the number of car accidents. A study was conducted, and data were collected. They failed to reject the null hypothesis. Why?

- The null hypothesis is true
- Type 2 error

If the power is high, it is more likely that the null hypothesis was true.

Power

Power is how likely we are to detect actual differences.

Ex.: A study was conducted to determine if the "turn right after stopping at a red light" traffic rule would increase the number of car accidents. A study was conducted, and data were collected. They failed to reject the null hypothesis. Why?

It turns out, the sample size was not large enough.

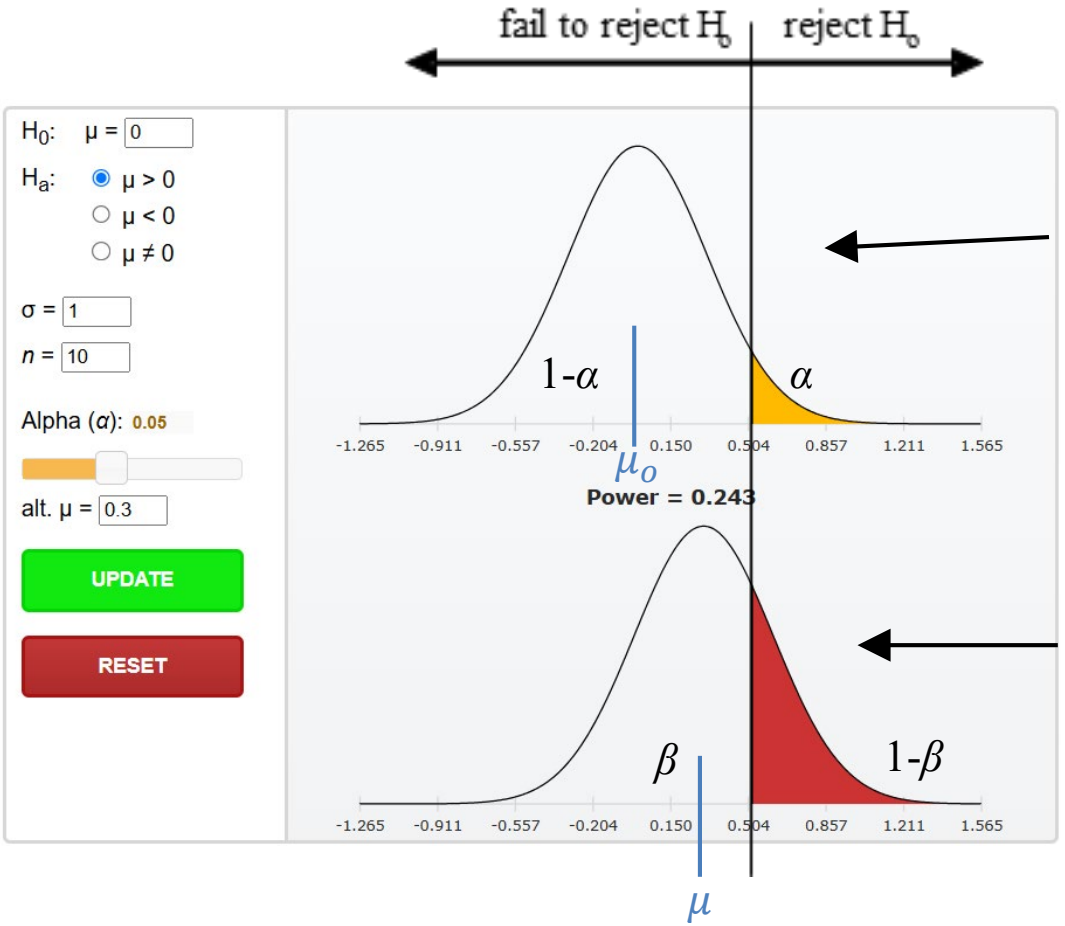
So, power was too small.

Government agencies require power to be high (.80+) for tests that involve approvals of regulated materials (USDA).

Thought Question: If the result of the test is "Reject H_o" why is power not an issue?

What is probability of correctly rejecting H_0?

Power!



Assumes H_o is true

This represents the situation where H_A: $\mu > \mu_0$

Assumes H_A is true

- Note that as μ moves further to the right the power $(1 - \beta)$ gets larger. Why is that?

https://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html

How Do We Calculate Power?

Power depends on the following items:

1. The design of the experiment
2. Significance level α
3. Parameter setting under the alternative hypothesis (including error variance σ^2
4. Sample size n

Including power, that gives us 5 values in the power calculation. If we know 4 of those values, we can solve for the 5th. Typically, we either solve for power or the sample size (not the other three values).

Calculating power for 2-sample t-tests and one-way ANOVA

Calculating power for 2-sample t-tests and ANOVA can be much more complicated. We will discuss how to do so in R rather than with the formula.

In general, power is the probability of correctly rejecting H_o

For example,

$$P(F > f^*|H_A)$$

Where F is essentially our test statistic, and f^* is the cutoff in our F-sampling distribution that separates "reject H_o" from fail to "reject H_o"

Calculating power for 2-sample t-tests and one-way ANOVA

Consider a situation where we are testing

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_5$$
$$H_A: At\ least\ one\ \mu_i\ not\ equal\ to\ the\ others$$

To calculate an approximation of power, you must make some assumptions. We need to assume values for

- Which statistical model do we plan to use
- Our population means (or at least the difference between two of the means that must be observed to reject H_o
- The value of our error variance, σ^2
- The sample size for each group

Example:

Let's say that for H_A we assume $\mu_1 = 57, \mu_2 = 63, \mu_3 = 60, \mu_4 = 60, \mu_5 = 60$, $\sigma^2 = 7$, $n_i = 4$ for each group

Calculating power for one-way ANOVA

R Script:

```
mu     <- c(57, 63, 60, 60, 60)
sigma2 <- 7
power.anova.test(groups = length(mu), n = 4, between.var = var(mu),
                 within.var = sigma2)
```

Output:

```
##      Balanced one-way analysis of variance power calculation
##
##           groups = 5
##                n = 4
##      between.var = 4.5
##       within.var = 7
##        sig.level = 0.05
##            power = 0.578
##
## NOTE: n is number in each group
```

The scenario we described will have power of approximately 58%

Calculating n for one-way ANOVA to achieve 80% power

R Script:

```
power.anova.test(groups = length(mu), between.var = var(mu),
                 within.var = sigma2, power = 0.8)
```

Output:

```
##      Balanced one-way analysis of variance power calculation
##
##           groups = 5
##                n = 5.676
##     between.var = 4.5
##      within.var = 7
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

The scenario we described to achieve 80% power will require samples of size 6 or more

Adjusting for Covariates

- A covariate is like a lurking variable. It is a variable that may have an impact on the results, but it was not of primary interest in our study. So, we include it in the model.

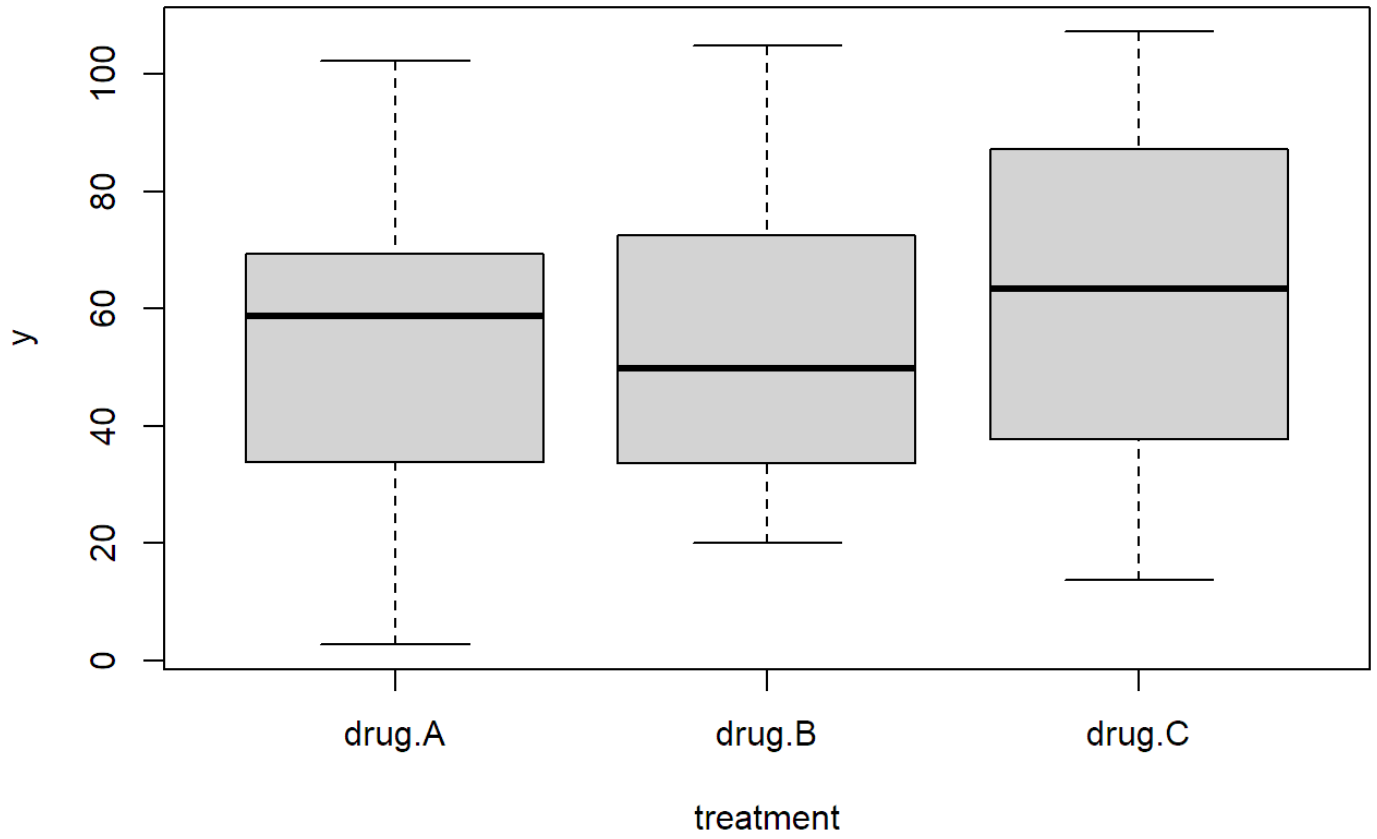
- Example: We conduct a weight loss study. Of primary interest is to see if our 3 diets have an impact on how much weight a person is able to lose.
-
  - Response variable: weight lost
  - Explanatory variable: diet (A, B, C)

Adjusting for Covariates

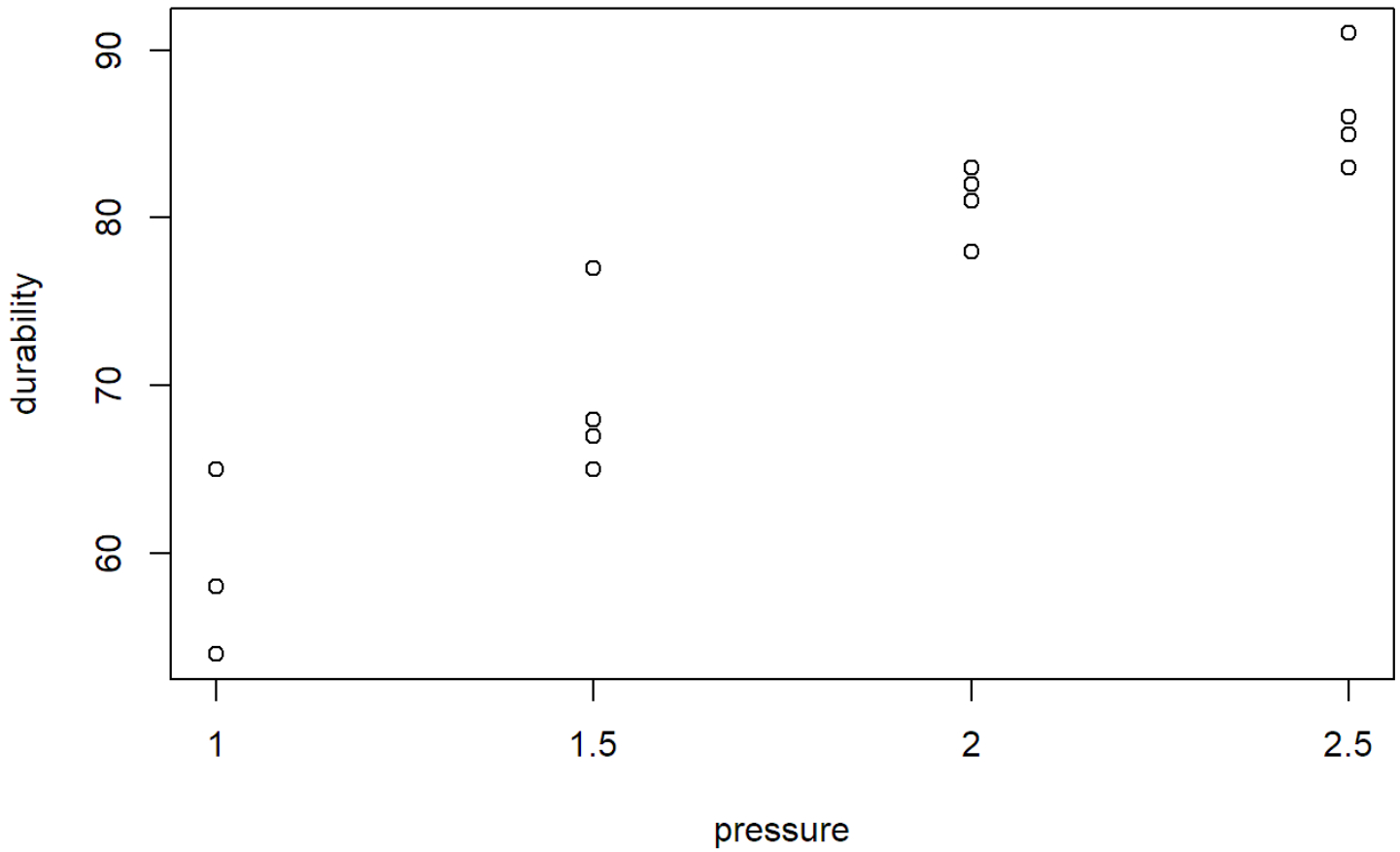
- We know that a person's initial weight when they enter the study will have an impact on how much weight they can lose. A 300lbs individual can lose 150lbs. But a 140lbs individual cannot lose 150lbs.

- By including the person's initial weight in the model as a covariate, we can account for the variability caused by that variable, removing it from our random error, making our tests more precise.

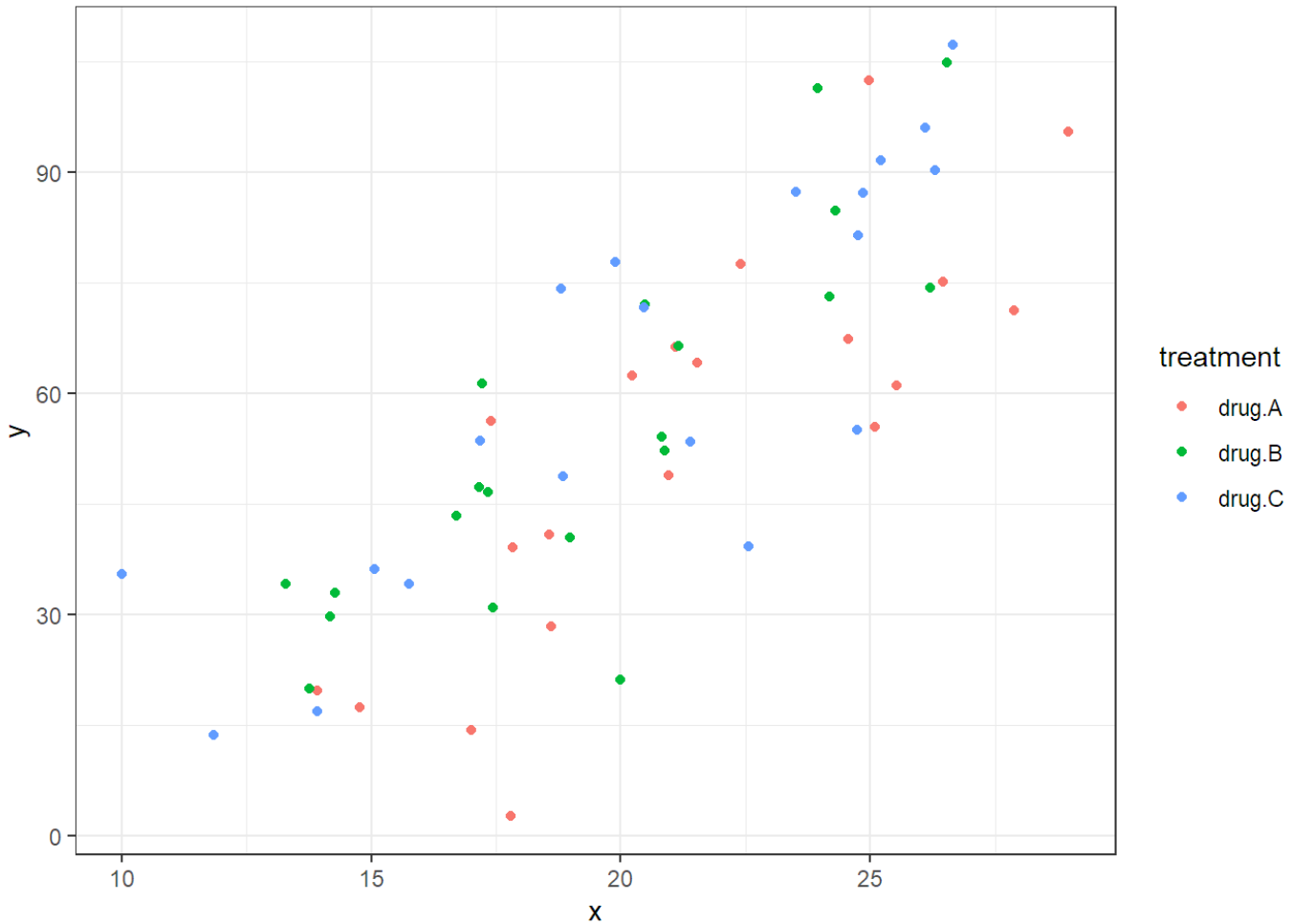
When we observe a quantitative response variable and categorical explanatory variable, we use ANOVA



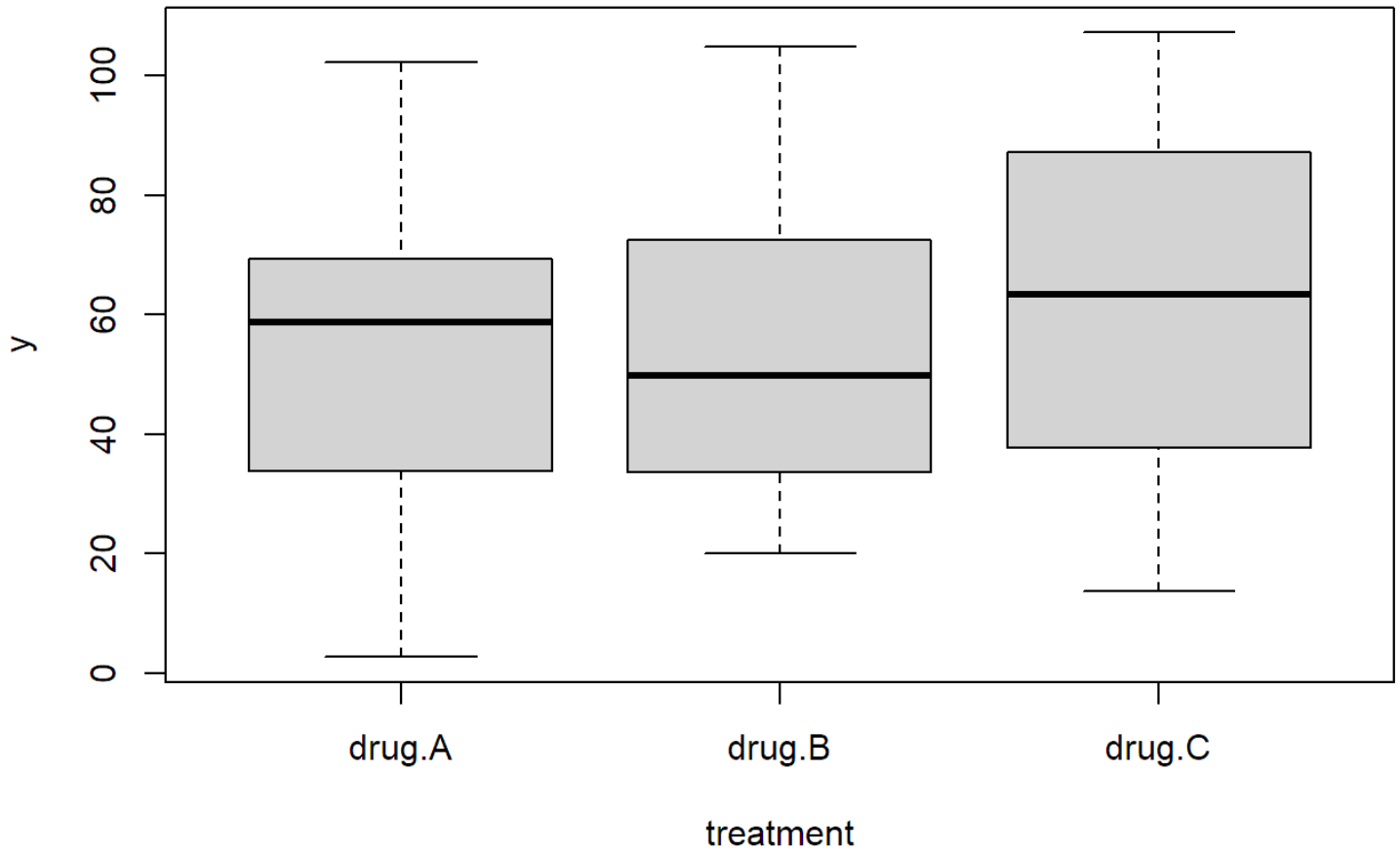
When we observe a quantitative response variable and quantitative explanatory variable, we use regression



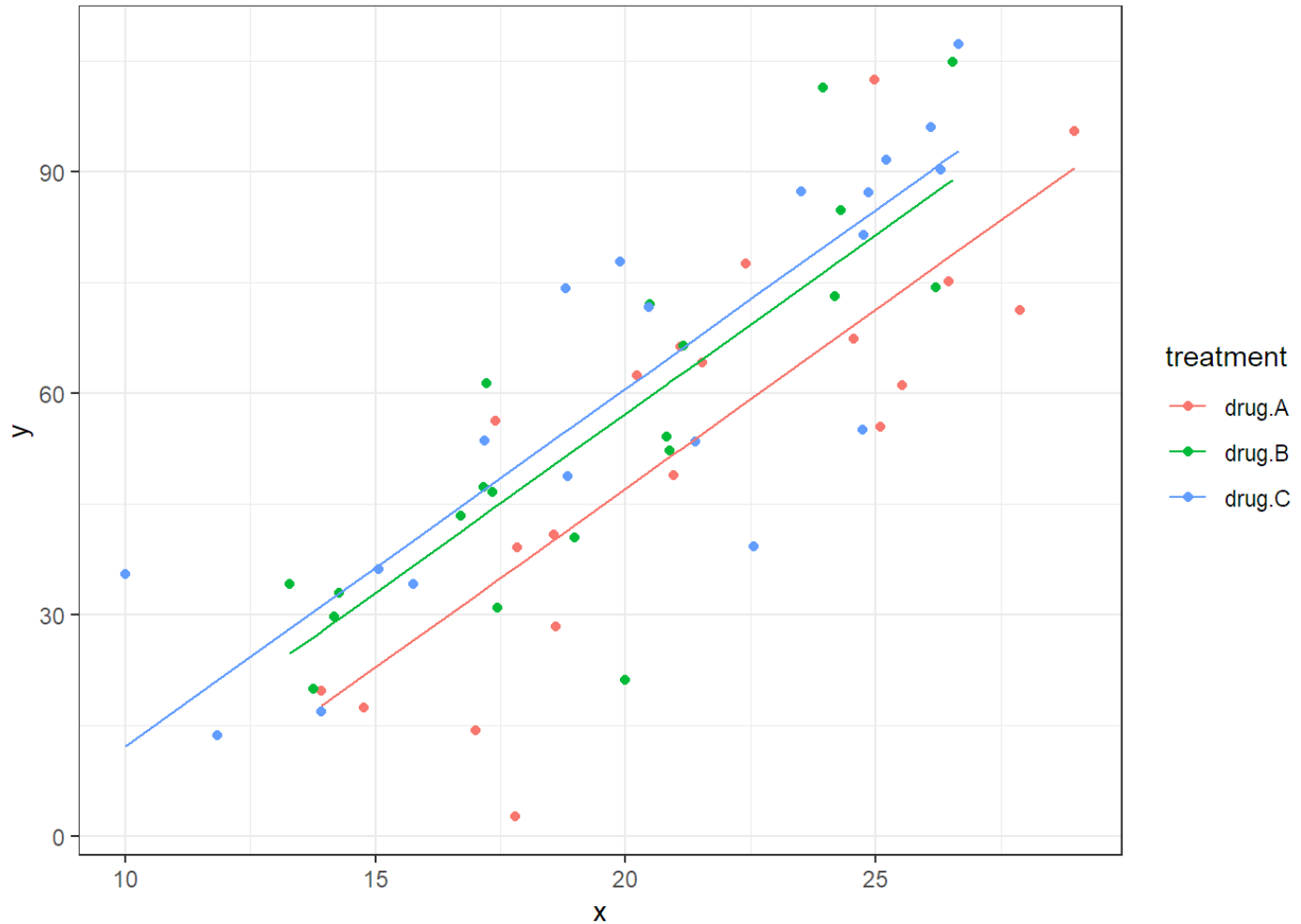
What do we do when we have both categorical and quantitative explanatory variables?



ANOVA cannot detect these drug differences



Using both categorical and quantitative explanatory variables
Rather than means we assess vertical distances among lines



Analysis of Covariance (ANCOVA)

- The quantitative variable is called a "covariate"
- Covariate helps to clarify the relationship between the response variable and the other explanatory variables.
- Using covariates is similar to blocking.
- Effects Model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- Effects Model with covariate: $Y_{ij} = \mu + \alpha_i + \beta \cdot x_{ij} + \epsilon_{ij}$
- Note that $\mu + \alpha_i$ now characterizes multiple intercepts and β characterizes a common slope for multiple regression lines.

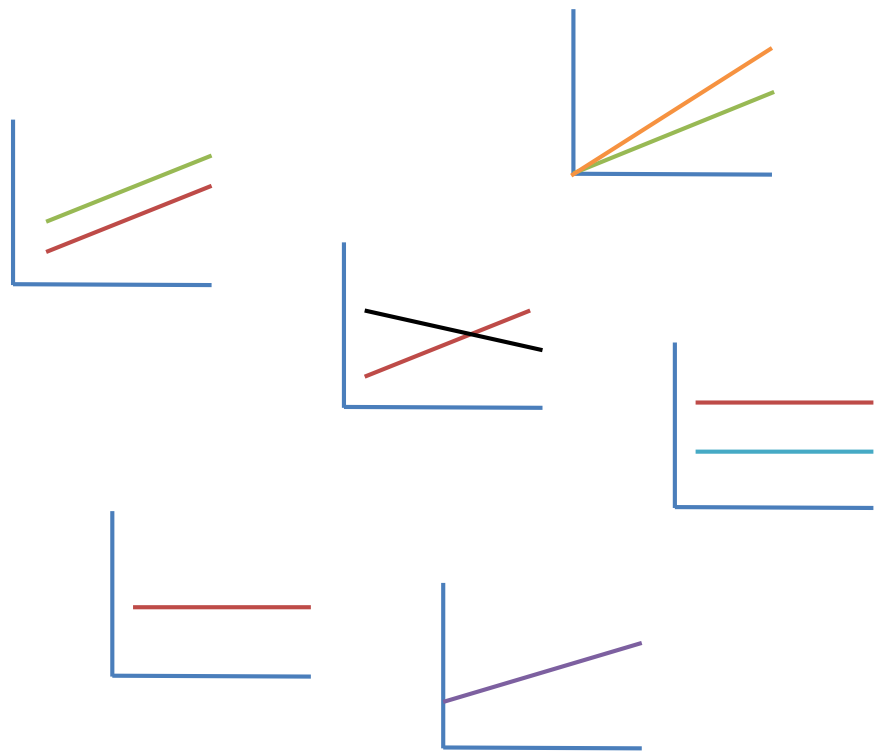
- <Section2_ANCOVA.R>

Analysis of Covariance (ANCOVA)

Note that α represents one intercept while α_i represents the potential for multiple intercepts. Additionally, β represents one slope while β_i represents the potential for multiple slopes.

What do each of these models represent? How many slopes and how many intercepts? Can you match the models with the appropriate graphs?

- $Y_{ij} = \mu + \alpha + \gamma + \beta \cdot x_{ij} + \epsilon_{ij}$
- $Y_{ij} = \mu + \alpha_i + \beta \cdot x_{ij} + \epsilon_{ij}$
- $Y_{ij} = \mu + \alpha_i + \beta + \epsilon_{ij}$
- $Y_{ij} = \mu + \alpha_i + \beta_i \cdot x_{ij} + \epsilon_{ij}$
- $Y_{ij} = \mu + \beta \cdot x_{ij} + \epsilon_{ij}$
- $Y_{ij} = \beta \cdot x_{ij} + \epsilon_{ij}$
- $Y_{ij} = \mu + \theta_i + \pi \cdot x_{ij} + \epsilon_{ij}$
- $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- $Y_{ij} = \mu + \alpha + \epsilon_{ij}$



Summary

Section 2 looks at the completely randomized design (CRD).

Things to know:

- Identify CRD
- Identify one-way treatment structure
- Know how to check model assumptions
- Understand nonparametric approaches and know how to perform them in R
- Understand and know how to calculate power and sample size
- Understand covariates and know how to analyze ANCOVA models in R