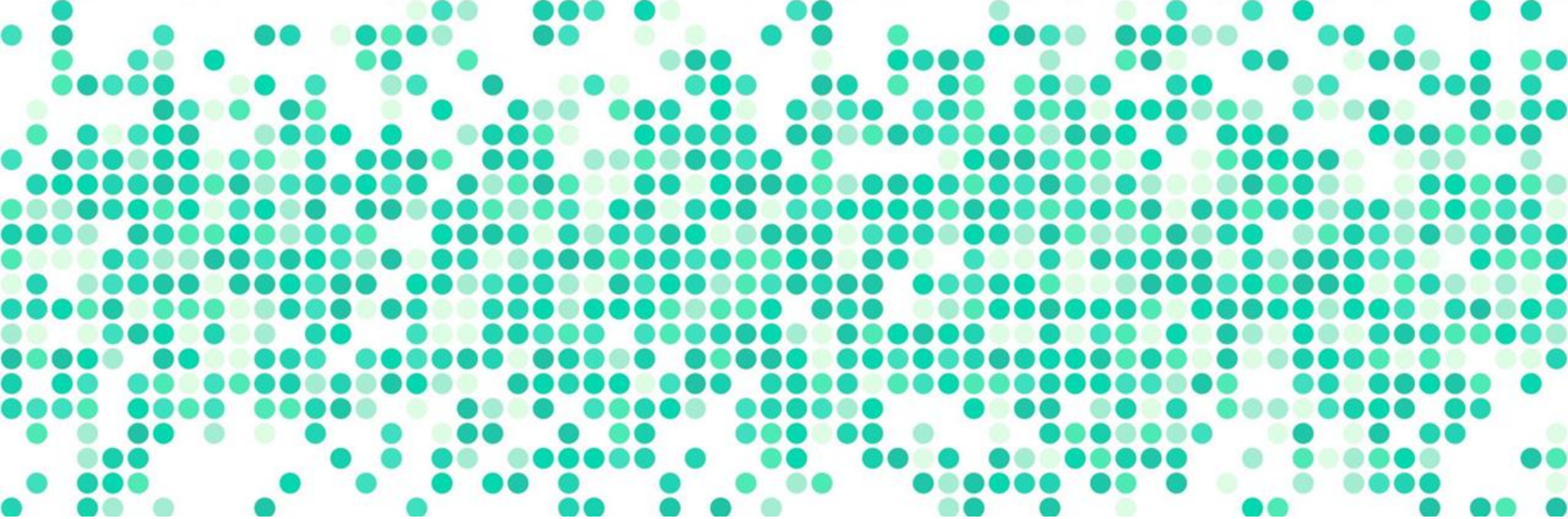


# DATA ANALYSIS

Week 13: Additional predictors

# logistics

- three more content weeks
- final opt-out deadline for problem sets is next Monday
- midterm 2 scores will be released latest by the end of the week with current percentages

| | | |
|---|---|---|
| 13 | T: April 15, 2025 | W13: Additional Predictors |
| 13 | Th: April 17, 2025 | W13 continued… |
| 13 | Su: April 20, 2025 | **Week 13 Quiz due** |
| 14 | M: April 21, 2025 | **PS6 due / PS4 / Opt-out Deadline 3** |
| 14 | T: April 22, 2025 | W14: Repeated Measures |
| 14 | Th: April 24, 2025 | W14 continued… |
| 14 | Su: April 27, 2025 | **Week 14 Quiz due** |
| 15 | M: April 28, 2025 | **PS5+ PS6 revision due** |
| 15 | T: April 29, 2025 | W15: Miscellaneous Data |
| 15 | Th: May 1, 2025 | W15 continued… |
| 16 | M: May 5, 2025 | **PS7 due** |
| 16 | T: May 6, 2025 | W16: Last Class / Final Exam review |
| 17 | Th: May 15, 2025 | **PS7 revision + Computational Exam Due by 1.30 pm** |
| 17 | Th: May 15, 2025 | **Conceptual Exam (1.30-3 pm, VAC South)** |

# logistics

| Letter grade | Points |
|---|---|
| A | 95 - 100+ |
| A- | 90 - 94.99 |
| B+ | 87 - 89.99 |
| B | 83 - 86.99 |
| B- | 80 - 82.99 |
| C+ | 77 - 79.99 |
| C | 73 - 77.99 |
| C- | 70 - 72.99 |
| D | 60 - 69.99 |
| F | fewer than 60% |

| | | |
|---|---|---|
| 13 | T: April 15, 2025 | W13: Additional Predictors |
| 13 | Th: April 17, 2025 | W13 continued… |
| 13 | Su: April 20, 2025 | **Week 13 Quiz due** |
| 14 | M: April 21, 2025 | **PS6 due / PS4 / Opt-out Deadline 3** |
| 14 | T: April 22, 2025 | W14: Repeated Measures |
| 14 | Th: April 24, 2025 | W14 continued… |
| 14 | Su: April 27, 2025 | **Week 14 Quiz due** |
| 15 | M: April 28, 2025 | **PS5+ PS6 revision due** |
| 15 | T: April 29, 2025 | W15: Miscellaneous Data |
| 15 | Th: May 1, 2025 | W15 continued… |
| 16 | M: May 5, 2025 | **PS7 due** |
| 16 | T: May 6, 2025 | W16: Last Class / Final Exam review |
| 17 | Th: May 15, 2025 | **PS7 revision + Computational Exam Due by 1.30 pm** |
| 17 | Th: May 15, 2025 | **Conceptual Exam (1.30-3 pm, VAC South)** |

# logistics

| Letter grade | Points |
|---|---|
| A | 95 - 100+ |
| A- | 90 - 94.99 |
| B+ | 87 - 89.99 |
| B | 83 - 86.99 |
| B- | 80 - 82.99 |
| C+ | 77 - 79.99 |
| C | 73 - 77.99 |
| C- | 70 - 72.99 |
| D | 60 - 69.99 |
| F | fewer than 60% |

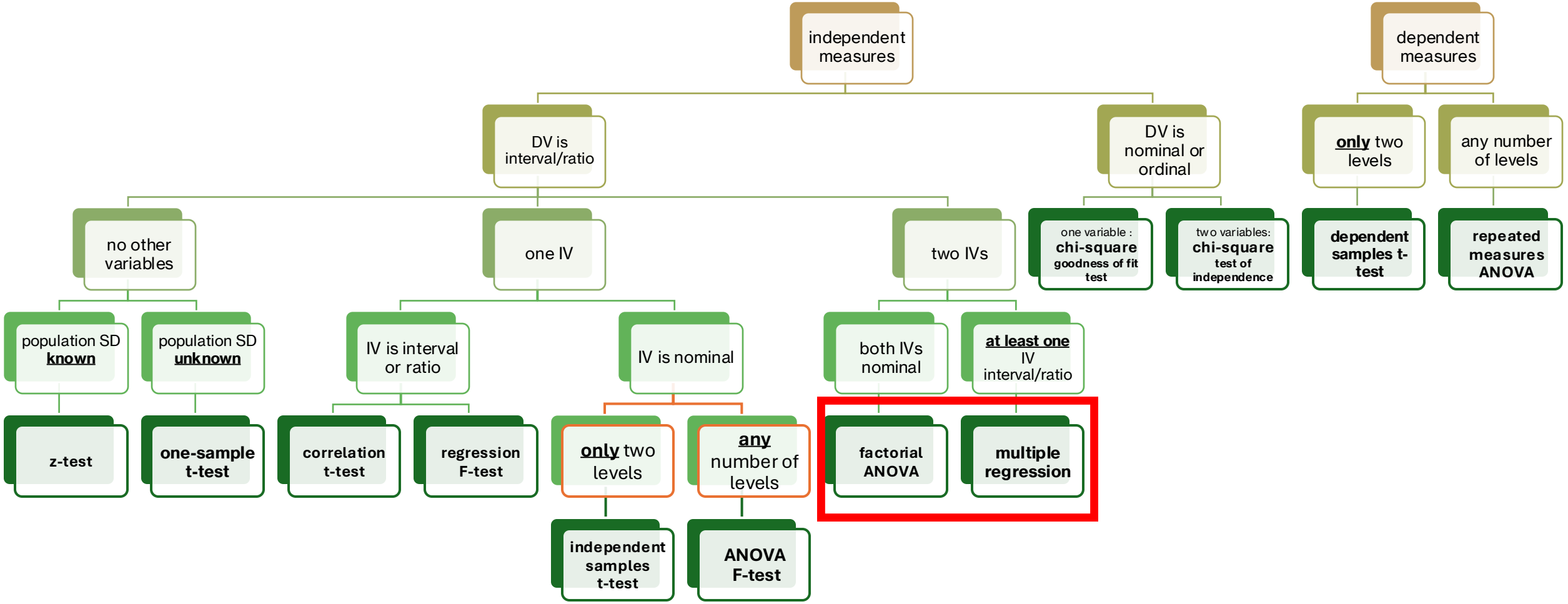
1. Complete class surveys (2 points) : There will be 3 surveys during the semester to gather your reflections and suggestions to improve the course. You will be able to earn 1 point for completing the first survey, and 0.5 for the other two surveys. With the exception of the pre-class survey, the two other surveys will be anonymous.

2. Win Conceptual Czar (1 point): To incentivize timely preparation and encourage you to master the class content, you will have the opportunity to submit multiple-choice and/or true/false questions based on the course content covered each week. Submitting questions for 8 of the 11 content weeks will earn you this extra credit point. Additionally, if your question is selected to be on any of the exams, you will earn an additional extra credit point.

3. Win Analysis Ace (1 point): To incentivize timely preparation and encourage you to master data analysis, the two students who score the highest on the *computational* exams throughout the semester will earn 1 extra credit point each.

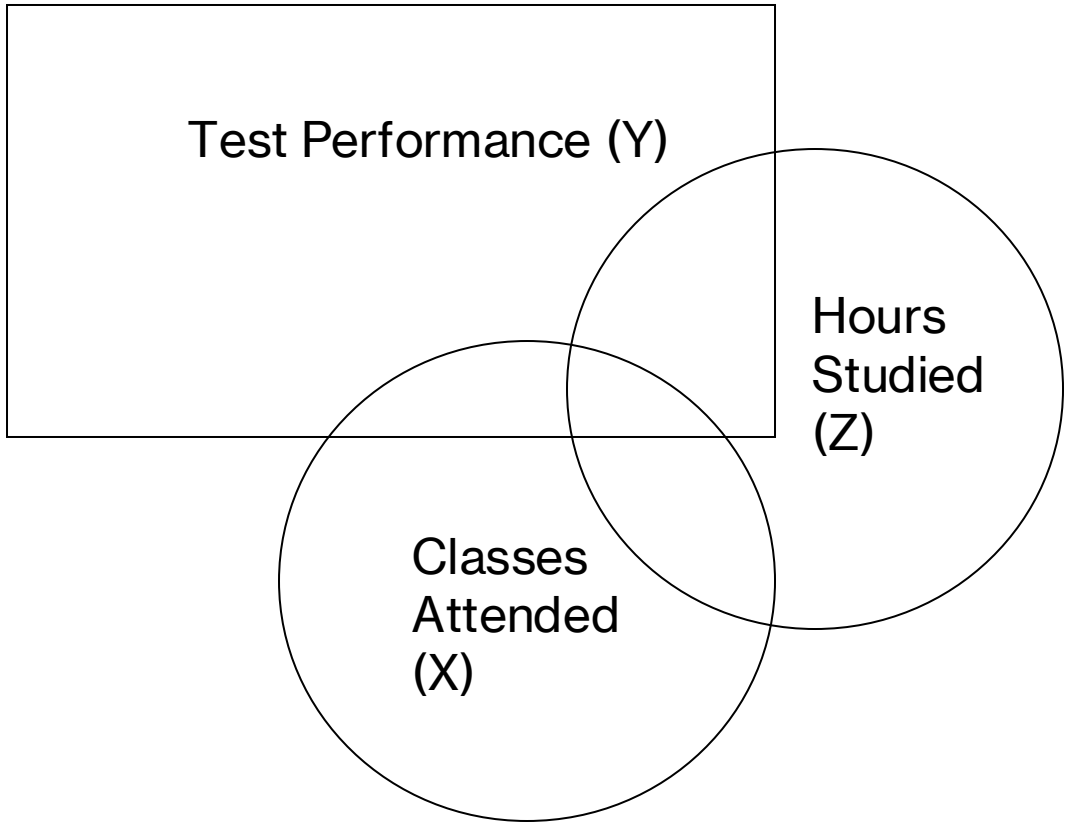
4. Win Memer of the Semester (1 point) : Each week, you will have the opportunity to submit a meme via Canvas, that reflects your experience with the course content of that week. Memes should be *original*, i.e., they should be course-specific and something you have created yourself and not simply found on the internet, although you are allowed to use common images/tropes from popular memes as a starting point. All memes will be gathered and sent to the class anonymously at the end of the semester for a survey, and the student(s) with the average highest score and the best scoring meme will both receive 1 additional point. **Note**: A student can only receive a maximum of 1 point through this mechanism, even if the same student has the highest average score in the context *and* the best scoring meme.

# hypothesis testing flowchart



# additional predictors = complex models

- often, outcomes/dependent variables depend on not just one IV, but several IVs
- in such situations, modeling the variation in our dependent variable using only one variable leads to an impoverished model: we could do better by examining multiple variables
- data = model + error
  - *one IV*: $Y = a + bX + error$
  - *multiple IVs*: $Y = a + b_1X_1 + b_2X_2 + ... + error$



# complex models: data types

- for a one DV and one IV situation, we saw how the data could come in different forms
- when more than one IV is involved, several permutations and combinations are possible
  - one DV ~ interval/ratio $IV_1$ + interval/ratio $IV_2$
  - one DV ~ interval/ratio $IV_1$ + nominal $IV_2$
  - one DV ~ nominal $IV_1$ + interval/ratio $IV_2$
  - one DV ~ nominal $IV_1$ + nominal $IV_2$
- no fear...general linear models are here!

| | **one independent variable** | | |
|---|---|---|---|
| **dependent variable** | nominal | ordinal | interval/ ratio |
| nominal | | | |
| ordinal | | | |
| interval/ratio | F / anova | | t / F |

# the tooth growth dataset

- this in-built R dataset contains the “length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid”
- think about the design of this experiment
  - dependent variable?
  - independent variable(s) and their levels?
  - broad research question?

# factorial designs

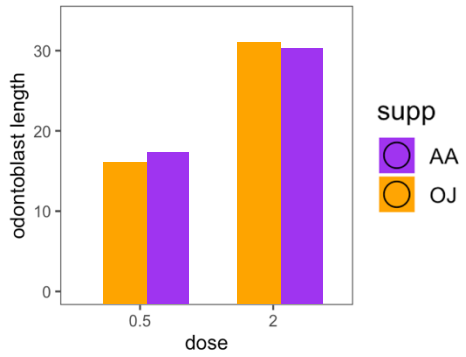
- factorial designs refer to situations where more than one independent variable or "factor" is manipulated in the same experiment (nominal IVs)
- common terminology
  - 2 x 2 factorial design, i.e., two independent variables (number of x's + 1), and each of them had 2 levels
  - 3 x 2 factorial design, i.e., 2 independent variables, one of them had 3 levels, and another had 2 levels
  - 3 x 5 x 4 x 6 factorial design, i.e., you are crazy
- what about our tooth decay design?
  - technically a 3 (dose: 0.5/1/2) x 2 (delivery: OJ, AA) design
  - we will examine a subset of this data that is 2 x 2
  - PS 6 has a problems with a 3 x 2 design! (arousal x task difficulty)

# tooth growth dataset: visualization

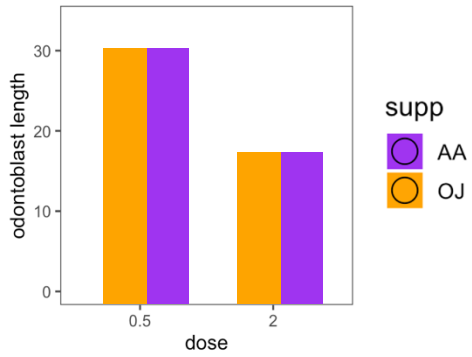
- let's try to visualize the pattern of **tooth growth** as a function of **dose** and **supplements**
  - **dose**: 0.5 mg and 2 mg
  - **supplements**: OJ and AA
- sketch a possible bar graph of tooth growth based on the research question: is tooth growth impacted by dosage and delivery method of vitamin C?
  - **dose** on x axis
  - **tooth growth** on y axis
  - **supplement** by color

# tooth growth dataset: visualization

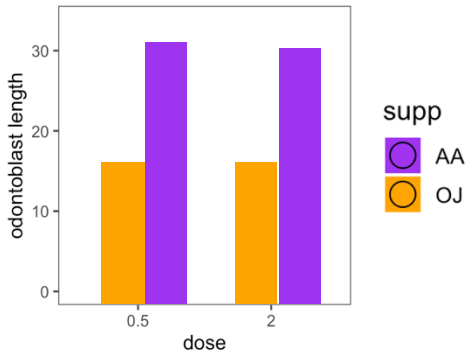
dose matters
supplement does not matter



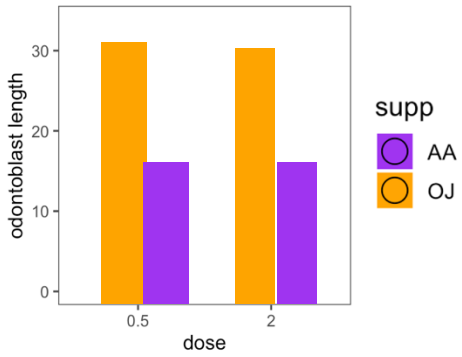

OR



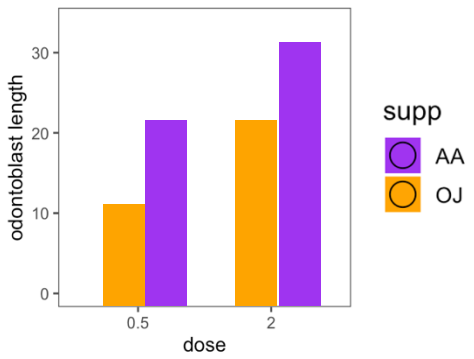

dose does not matter
supplement matters



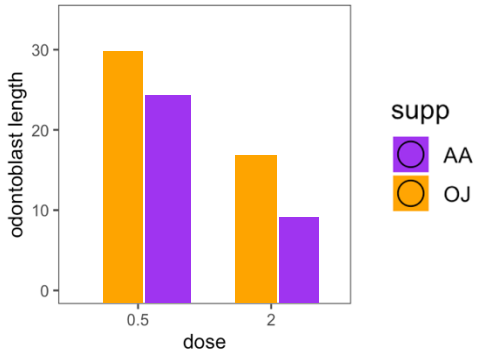

OR



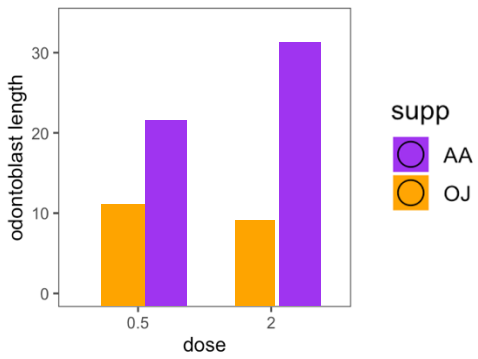

dose matters
supplement matters
dose and supplement
do not influence each other



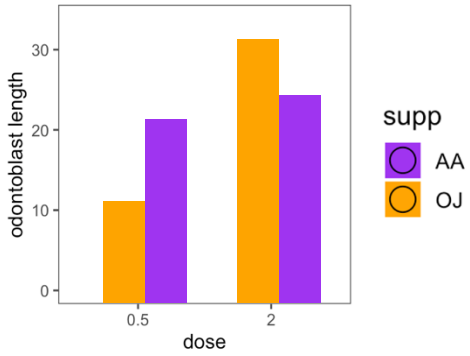

OR




dose matters
supplement matters
dose and supplement
influence each other

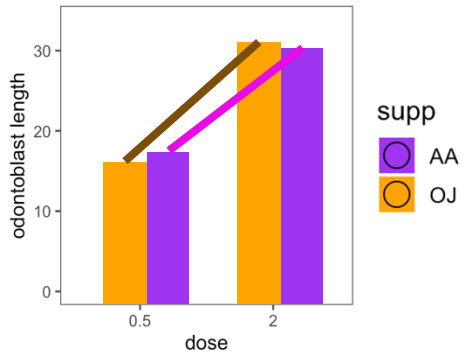



OR

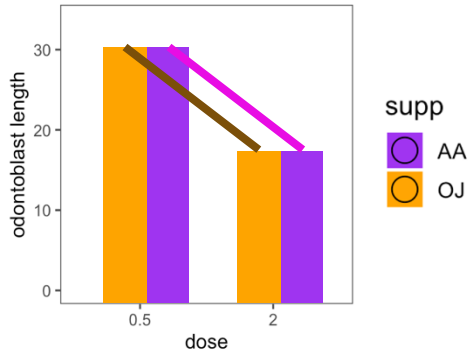


# tooth growth dataset: visualization

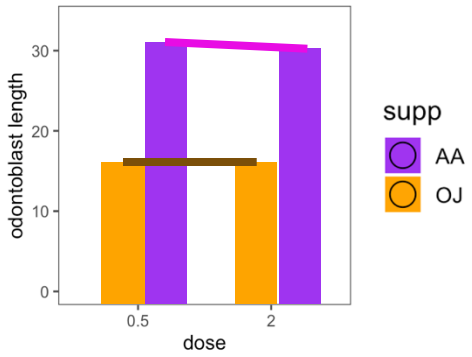
dose matters
supplement does not matter



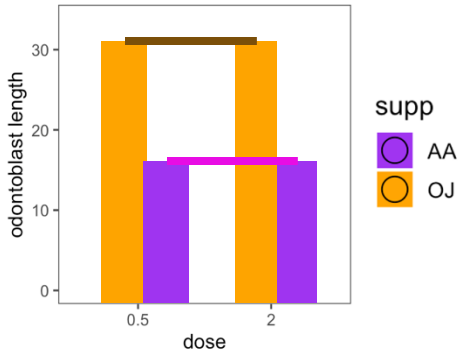

OR



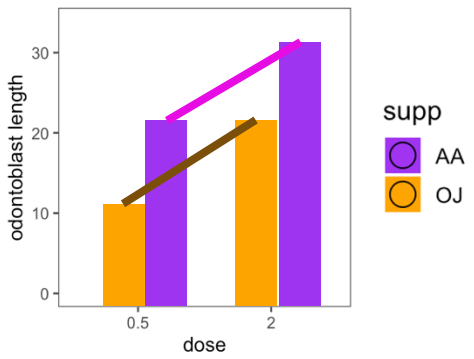

dose does not matter
supplement matters



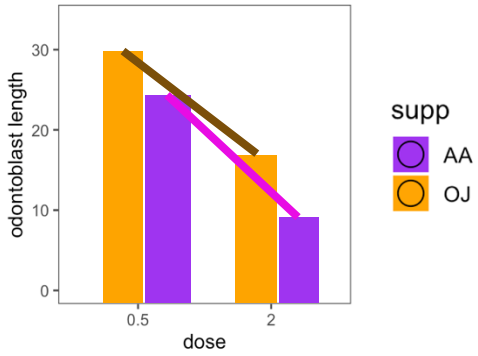

OR



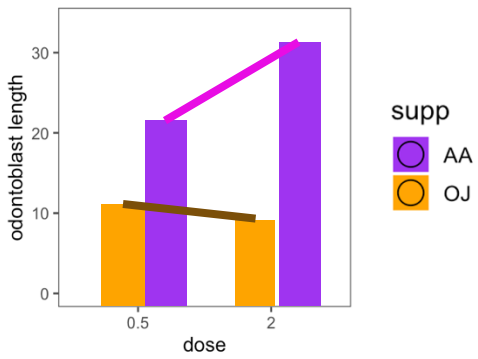

dose matters
supplement matters
dose and supplement
do not influence each other



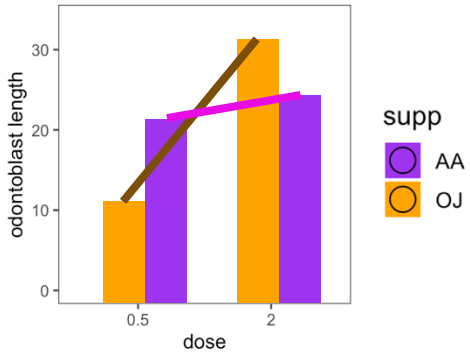

OR




dose matters
supplement matters
dose and supplement
influence each other




OR



# W13 Activity 1: compute the means

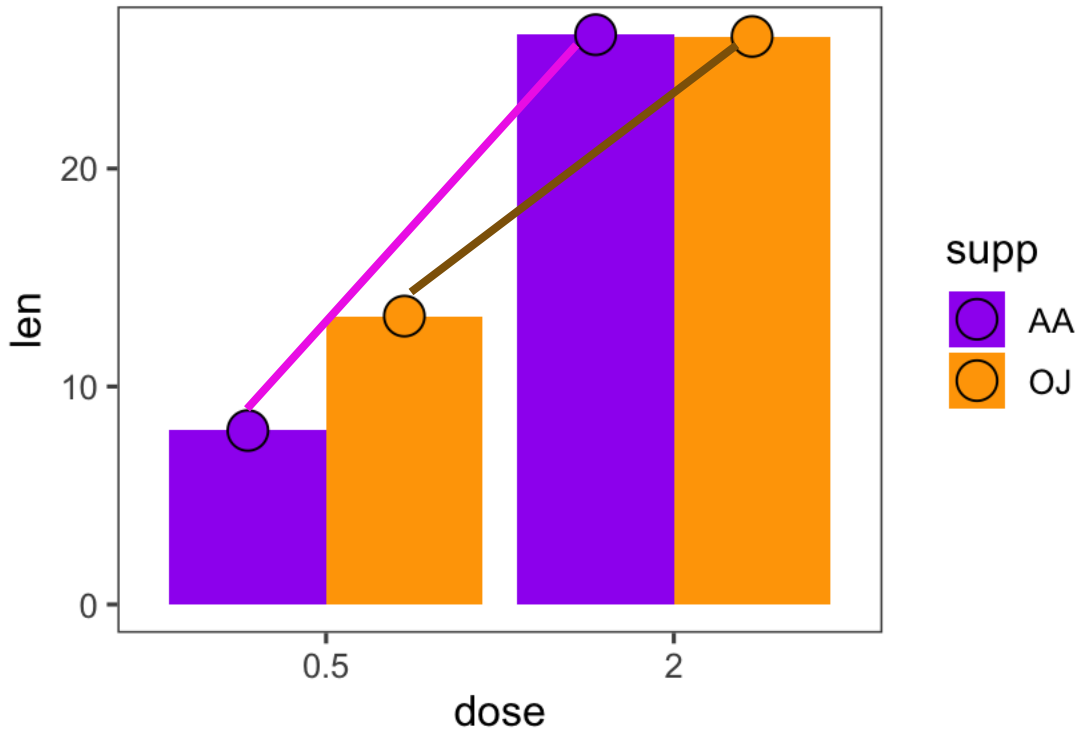
- use the tooth growth dataset
- compute all means
- plot the graph

| supplement | dose=0.5 | dose=2 |
|---|---|---|
| AA | 7.98 | 26.14 |
| OJ | 13.23 | 26.06 |

| | |
|---|---|
| AA_overall | 17.06 |
| OJ_overall | 19.645 |
| | |
| dose_0.5 | 10.605 |
| dose_2 | 26.1 |

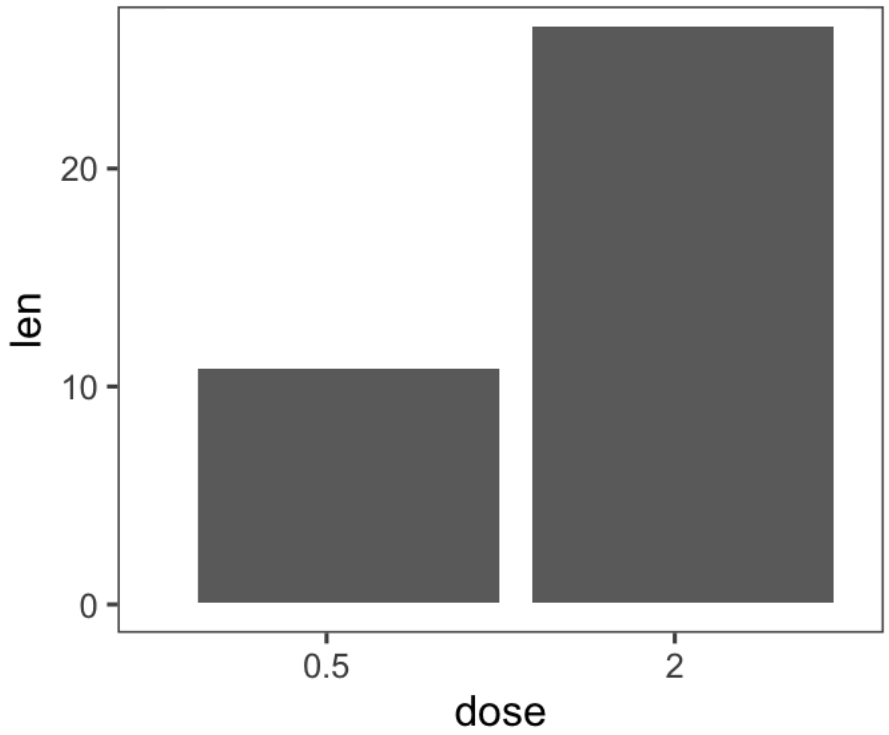
# tooth growth dataset: actual pattern

- **dose** matters (0.5 mg << 2 mg)
- **supplement** matters (OJ > AA slightly)
- **dose** and **supplement** influence each other
  - at 0.5 mg, delivery method matters
  - at 2 mg, delivery method stops mattering



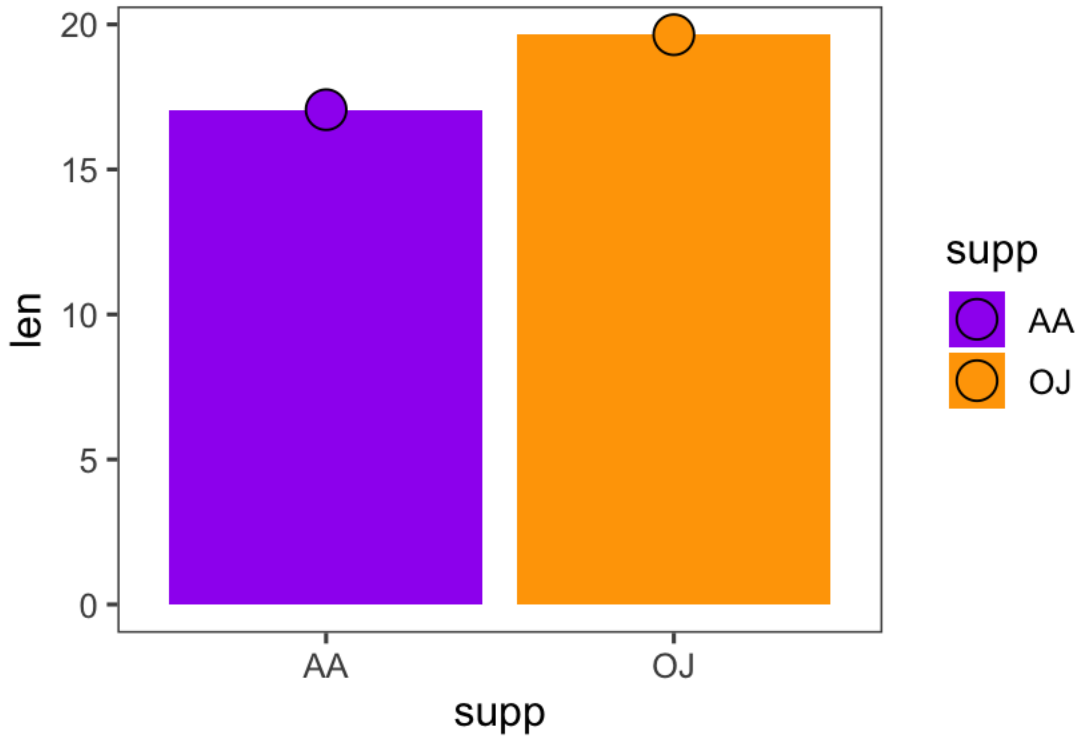
# tooth growth dataset: main effects

- **dose** matters (0.5 mg << 2 mg)
  - MAIN effect: the "overall" effect of dose (ignoring delivery method), i.e., difference in tooth growth for 0.5 mg vs. 2 mg
  - $M_{0.5mg} - M_{2mg}$



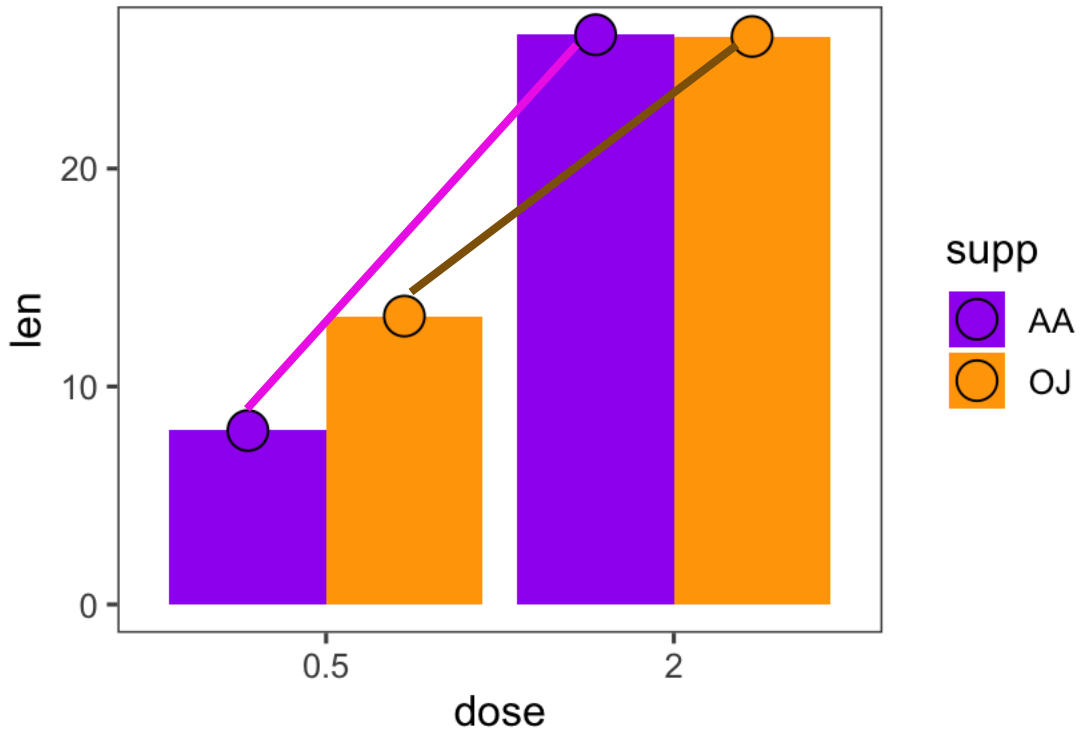
# tooth growth dataset: main effects

- **supplement** matters (OJ > AA)
  - MAIN effect: the “overall” effect of supplement (ignoring dose), i.e., difference in tooth growth for OJ vs. AA
  - $M_{OJ} - M_{AA}$



# tooth growth dataset: interactions

- **dose** and **supplement** influence each other
  - INTERACTION effect: the difference between differences
    - $OJ_{0.5mg} - OJ_{2mg}$ vs. $AA_{0.5mg} - AA_{2mg}$
- what would the plot look like if there was NO interaction?
  - parallel lines!



# main effects and interactions

- main effects represent the “overall” effect of one independent variable when ignoring the influence of other variables
- interactions represent the full relationship between multiple independent variables
- when interactions are present in the model, **main effects need to be qualified**, i.e., you cannot truly understand the influence of that variable in isolation

# main effects and interactions

| supplement | dose=0.5 | dose=2 |
|---|---|---|
| AA | 7.98 | 26.14 |
| OJ | 13.23 | 26.06 |

**difference**

$AA_{0.5mg} - AA_{2mg} = -18.16$

$OJ_{0.5mg} - OJ_{2mg} = -12.83$

**difference of differences = interaction**

$(AA_{0.5mg} - AA_{2mg}) - (OJ_{0.5mg} - OJ_{2mg}) = -5.33$

| | |
|---|---|
| AA_overall | 17.06 |
| OJ_overall | 19.645 |
| | |
| dose_0.5 | 10.605 |
| dose_2 | 26.1 |

main effect of **supplement**

$M_{OJ} - M_{AA} = 2.585$

main effect of **dose**

$M_{0.5mg} - M_{2mg} = 15.495$

# W13 Activity 2 (Canvas)

# practice question #1

- For a two-factor experiment with 2 levels of factor A and 3 levels of factor B and n = 10 subjects in each treatment condition, how many participants are in each level of factor B?
  - 10
  - 20
  - 30
  - 60

# practice question #2

- A two-factor research study is used to evaluate the effectiveness of a new blood-pressure medication. In this two-factor study, Factor A is medication versus no medication and factor B is male versus female. The medicine is expected to reduce blood pressure for both males and females, but it is expected to have a much greater effect for males. What pattern of results should be obtained if the medication works as predicted?
  - significant main effect for factor A (medication).
  - a significant interaction.
  - a significant main effect for factor A and a significant interaction.
  - none of the above.

# practice question #3

- In a line graph showing the results from a two-factor experiment, the levels of factor A (A1 and A2) are presented on the X-axis and separate lines are used to display the means for B1 and B2. If the points on the line for B1 are consistently 10 points lower than the corresponding point on the line for B2, what pattern of results is indicated?
  - an indication of an overall A-effect
  - an indication of an overall B-effect
  - an indication of a significant interaction
  - no claims can be made

# practice question #4

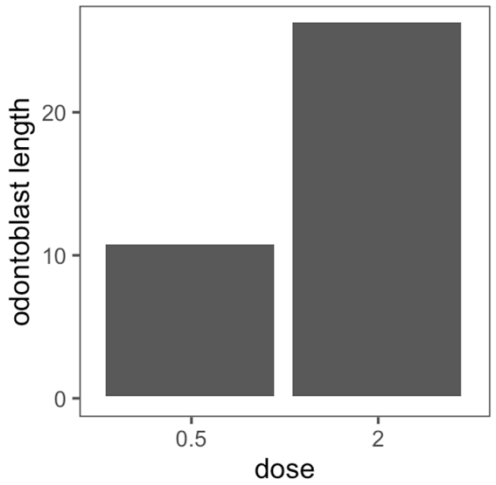
- In a line graph showing the results from a two-factor experiment, the levels of factor A (A1 and A2) are presented on the X-axis and separate lines are used to display the means for B1 and B2. If the points on the line for B1 are consistently **at least** 10 points lower than the corresponding point on the line for B2, what pattern of results is indicated?
  - an indication of an overall A-effect
  - an indication of an overall B-effect
  - an indication of a significant interaction
  - no claims can be made

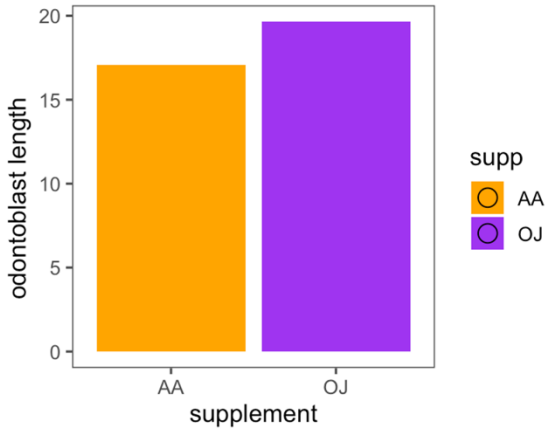
# overall strategy

- we want to explain the variation in tooth growth as a function of dose and supplement
- if we had no information about dose or supplement, what would our model be?
- if we had information about dose but not supplement, what would our model be?
- if we had information about supplement but not dose, what would our model be?

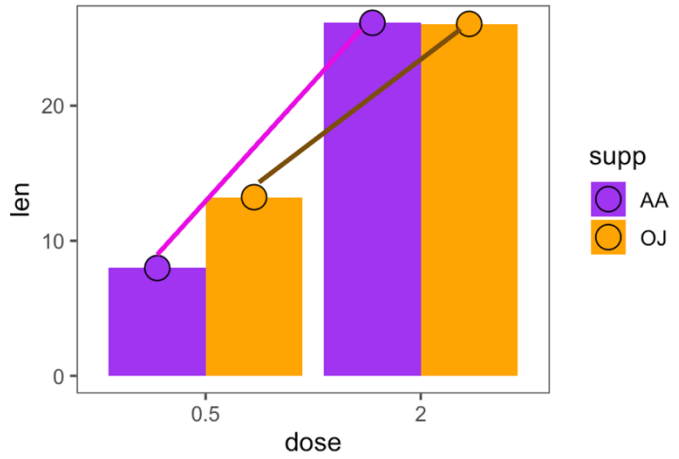
# building a factorial model

- we can start with three simple models
- grand mean model : toothGrowth ~ **grand mean**
- main effect 1: toothGrowth ~ **dose**
  - model = dose means
  - obtain $SS_{dose\_model} = SS_{total} - SS_{Y-\hat{Y}_{dose\_model}}$
- main effect 2: toothGrowth ~ **supp**
  - model = supplement means
  - obtain $SS_{supp\_model} = SS_{total} - SS_{Y-\hat{Y}_{supp\_model}}$
- next, we fit our more complex model
  - interaction model: toothGrowth ~ **dose + supp + (dose)(supp)**



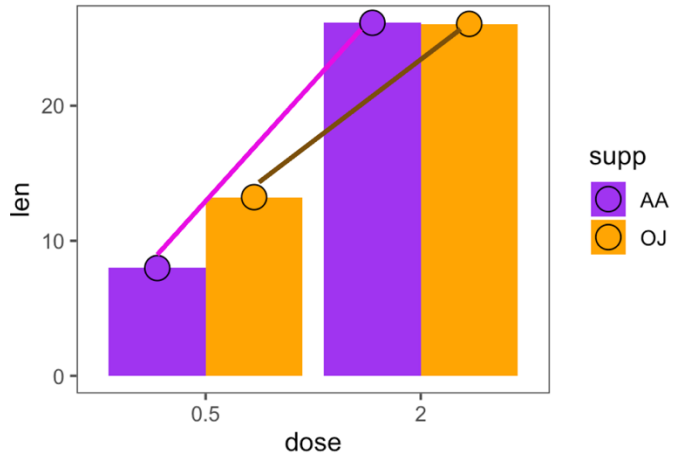



# activity: build the models




- build the **grand mean** model
  - obtain $SS_{total}$
- build the **dose** model using dose means
  - obtain $SS_{dose_{model}}$
- build the **supplement** model using supplement means
  - obtain $SS_{supp_{model}}$

# activity: build the models




- build the **grand mean** model
  - obtain $SS_{total} = 3056.29975$
- build the **dose** model using dose means
  - obtain $SS_{dose_{model}} = 2400.95025$
- build the **supplement** model using supplement means
  - obtain $SS_{supp_{model}} = 66.82225$

| | |
|---|---|
| **SStotal** | 3056.29975 |

| | **SS** |
|---|---|
| **supplement_model** | 66.82225 |
| **dose_model** | 2400.95025 |

# next time

- more on factorial ANOVA
- multiple regression

Here are the to-do's for this week:

- Submit Week 13 Quiz
- Submit Problem Set 6
- Submit any lingering questions here!
- Extra credit opportunities:
  - Submit Exra Credit Questions
  - Submit Optional Meme Submission

## Before Tuesday

- Nothing to do, take a breather!

## Before Thursday

- Watch: Factorial ANOVA (30 minute video).
  - Practice Data
  - Solution Sheet

## After Thursday

- See Apply section.