Data Analytics
CS301
An Overview of Anova Models

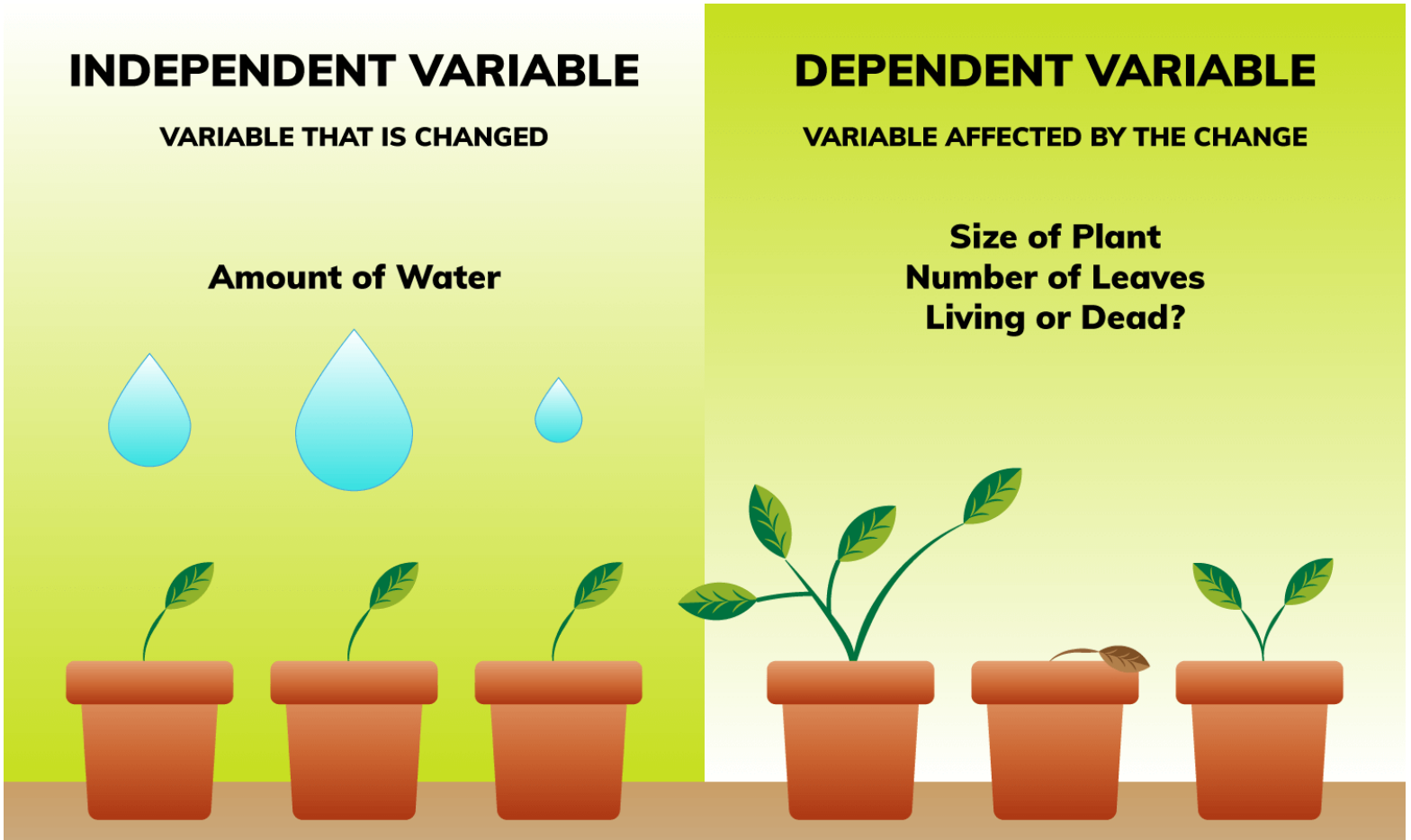
Week 8
Spring 2023
Oliver BONHAM-CARTER

Analysis of Variance (Anova)

- A statistical test used to analyze the difference between the means of more than two groups.
- In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.
- **Quick example:**
  - **Independent variable is social media use, and you assign groups to low, medium, and high levels of social media use.**
  - **Find out if there is a difference in hours of sleep per night**

Independent, Dependent Variables?

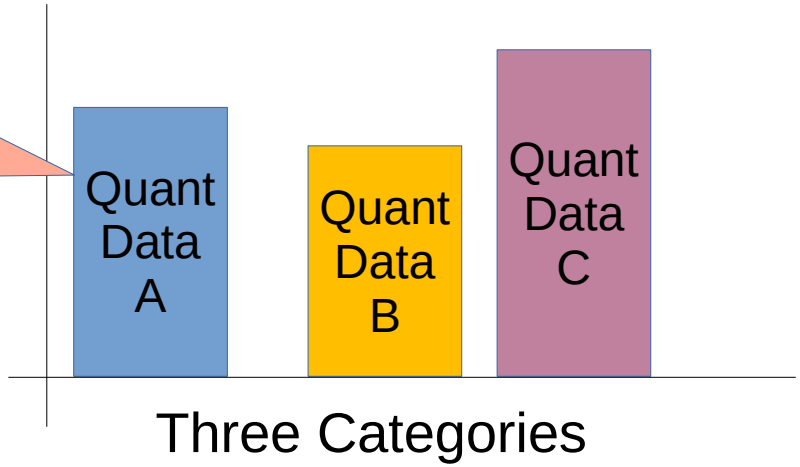
- **Independent** is seemingly random or unpredictable
- **Dependent** is not random; behavior depends on independent variable.



When to Use ANOVA?

- The independent variable should have at least three levels (i.e. at least three different groups or categories)
- ANOVA tells you if the *dependent* variable changes according to the level of the *independent* variable

Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data.



ANOVA: Different from t-Test

- **The Student's t-test is used to compare the means between *two* groups**
- **ANOVA is used to compare the means among *three or more* groups**
- **One way ANOVA: is a hypothesis test in which only one categorical variable or single factor is considered**
- **Makes comparisons between of means of three or more samples**
- **Each test study difference in means and the spread of the distributions (i.e., variance) across groups**
- **The statistical mechanisms of each test calculate statistical significance differently**

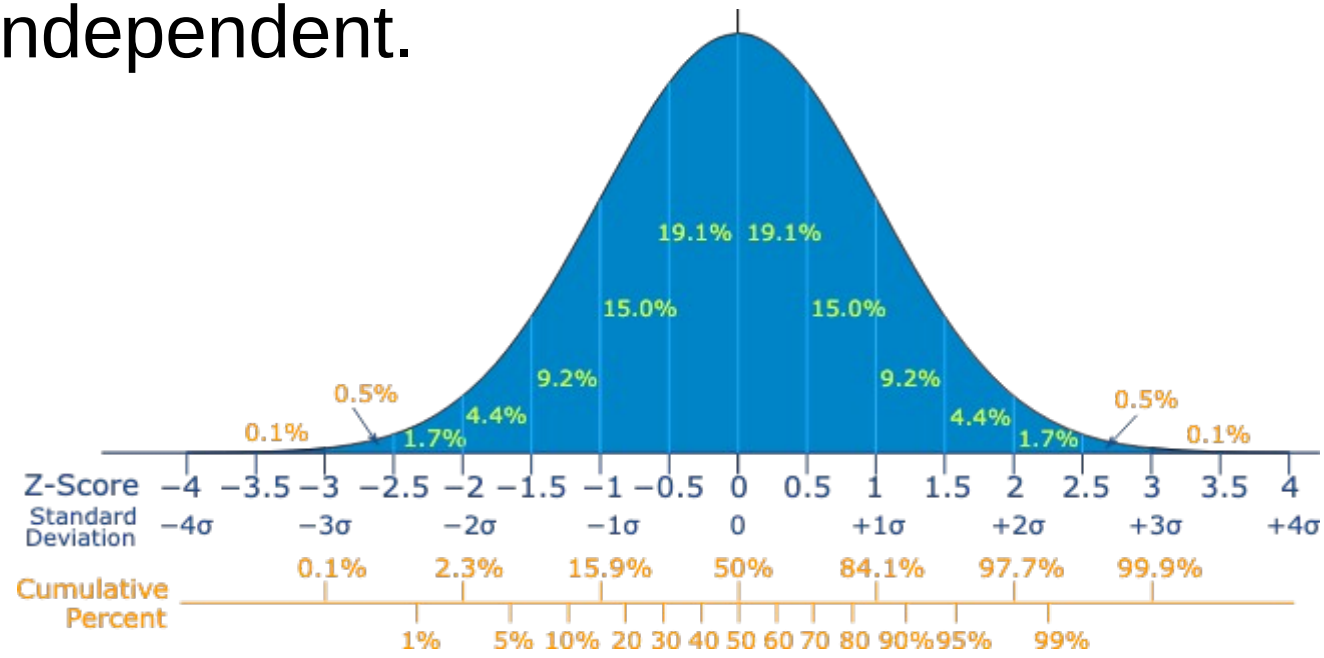
Avoiding Type-1 Errors

- When there are more than two means, then we can use the t-test to compare the means
- When conducting multiple t-tests, we run into more **type-1** errors from the test
- **Using ANOVA reduces type I error rates**

- **A Type I error is to reject the null hypothesis when we should have accepted it**
- Erroneous conclusions that results are statistically significant when, actually, they are not significant

Assumptions of ANOVA

- The responses for each factor level have a normal population distribution
- These distributions have the same variance
- The data are independent.



Testing Hypotheses in Anova

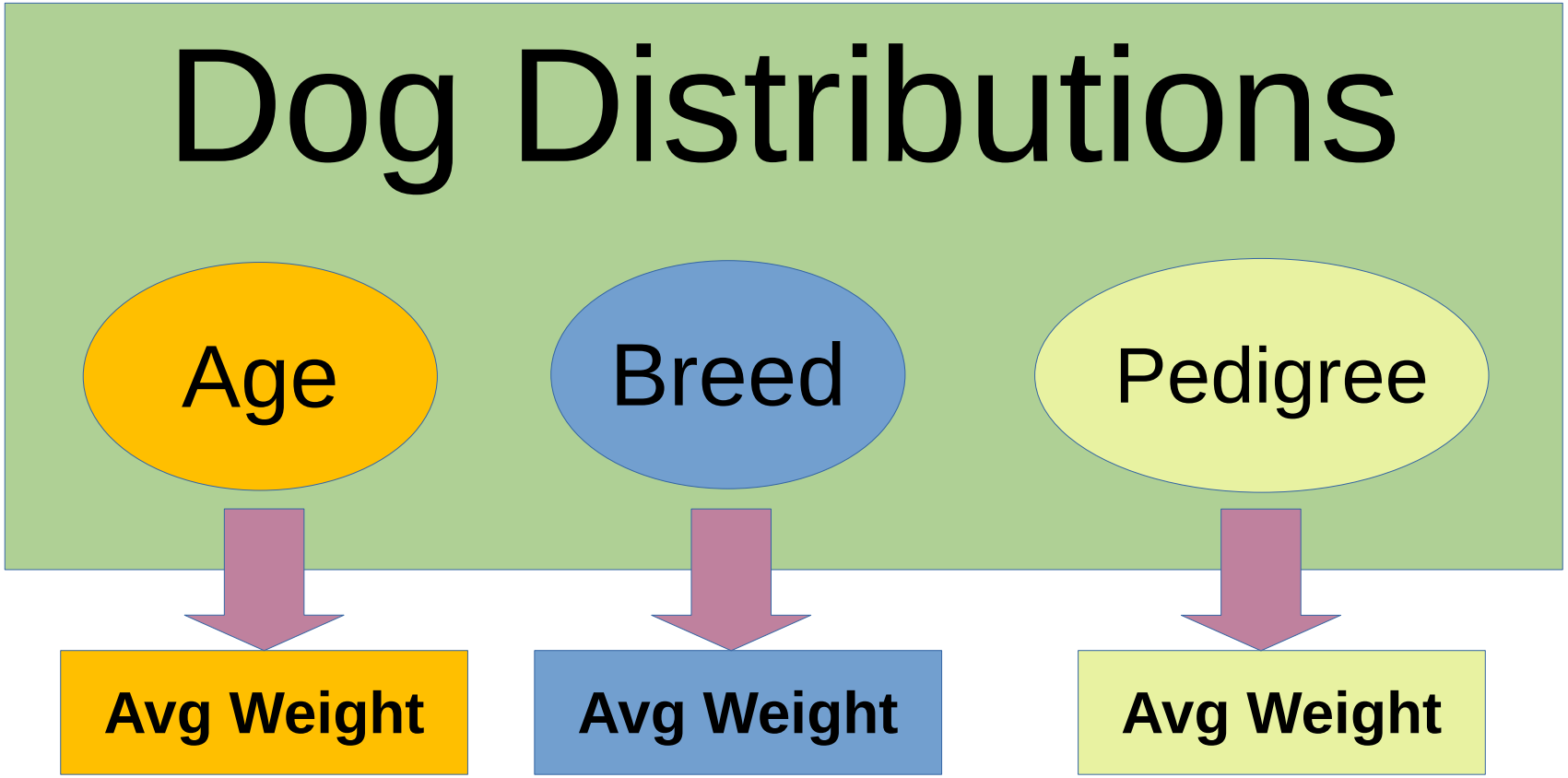
- Null hypothesis (H0) of the ANOVA: *there is no difference between means*
  - Written as, **H0: μ1 == μ2 == μ3**
- Alternative hypothesis (Ha): *the means are different from one another*
  - Written as, **Ha: μ1 != μ2 != μ3**
- For,
  - **H0** = the null hypothesis,
  - **Ha** = the alternative hypothesis,
  - **μ1** = the mean of population 1
  - **μ2** = the mean of population 2
  - **μ3** = the mean of population 3

Groups of Dogs

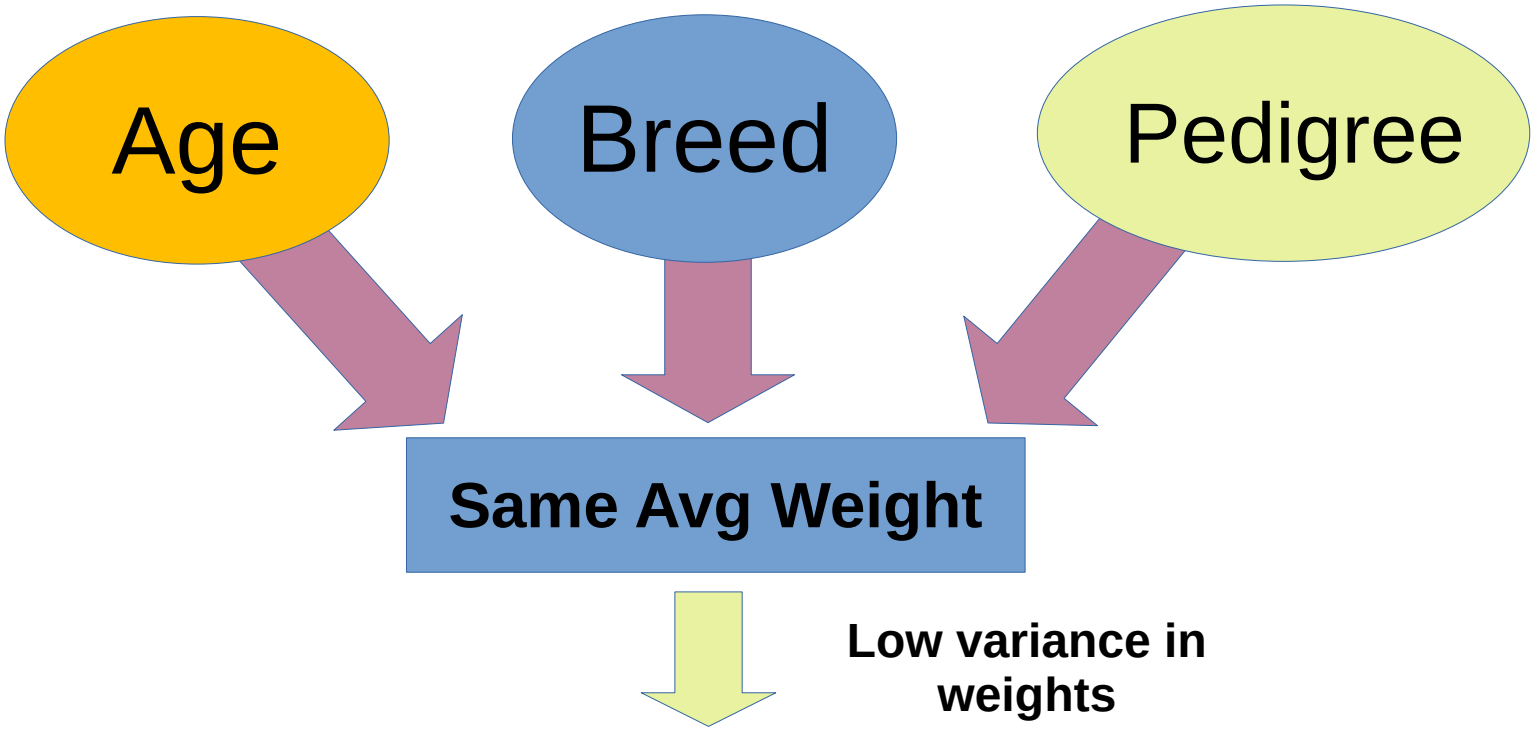
The Dog Show

Prediction by Group Type

- We want to ***predict the weight of a dog based on a certain set of characteristics*** for each dog (age, breed, and pedigree)



Explain Dog Weights According to Groups



If groups have the same mean, then it isn't reasonable to conclude that the groups are, in fact, separate in any meaningful way

Similar Means From Low Variances

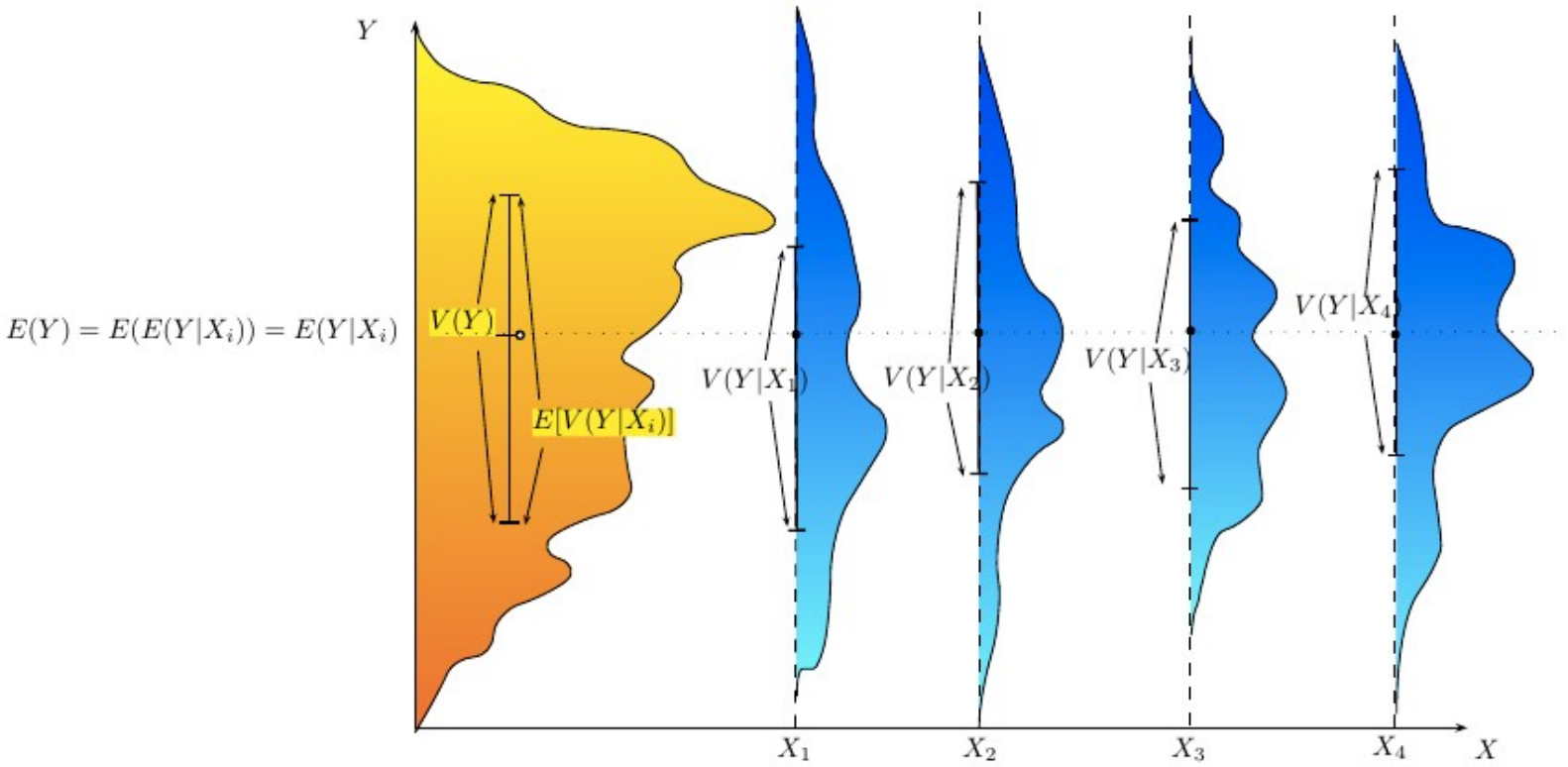
- Separate dogs into distributions having low variance of dog weights (a heterogeneous group) …
- The weight-means between groups would be distinct
- However, if weight-means were similar between groups, then the groups would be similar.

We create two groups:
are-cute *and* ***have floppy ears.***

The mean-weights would be similar from each group.

All these Beagles are ***cute*** and have ***floppy ears***!

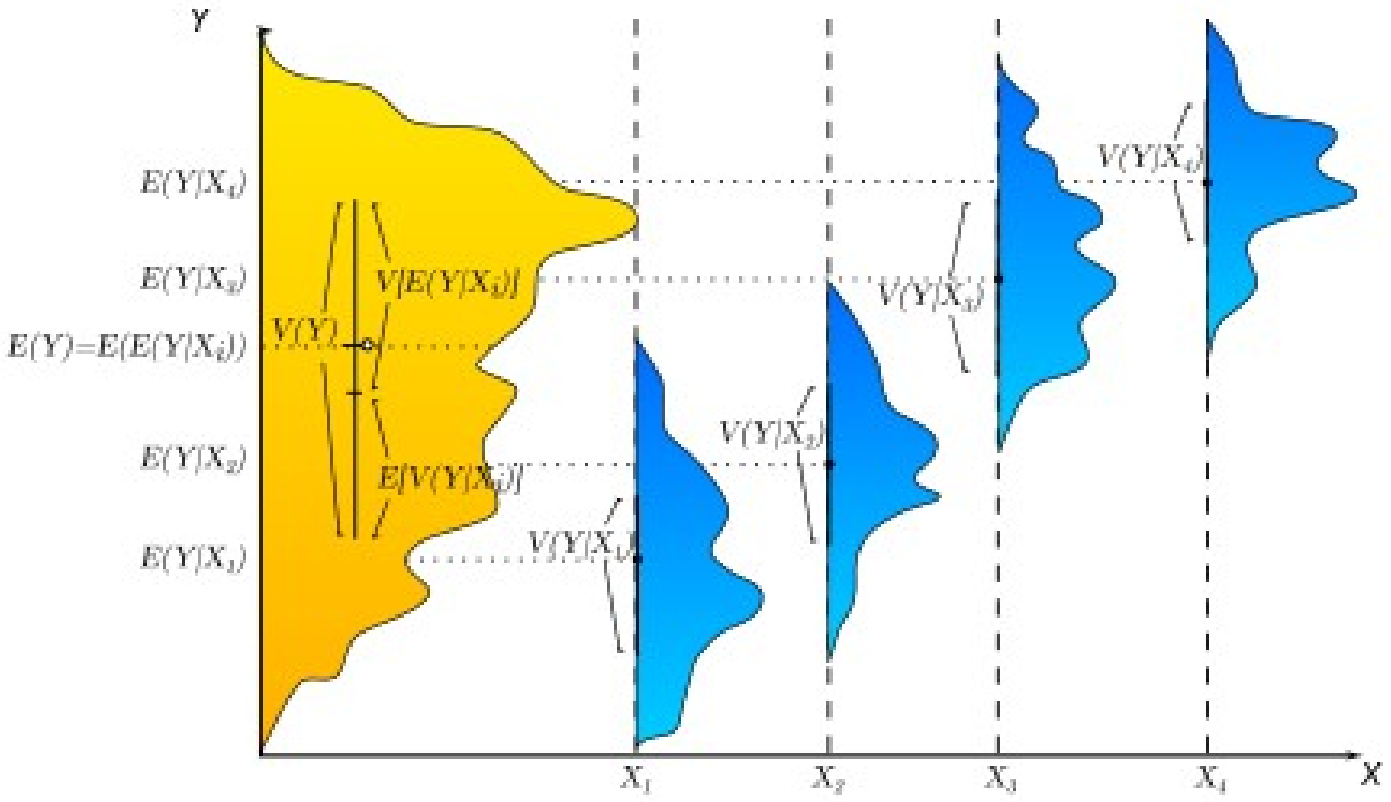
Comparison of Variances
No Fit



- Orange: mean-weights of all dogs
- Blues: All mean-weights by types of groups
- *Groups do not explain variation in distributions*

https://en.wikipedia.org/wiki/Analysis_of_variance

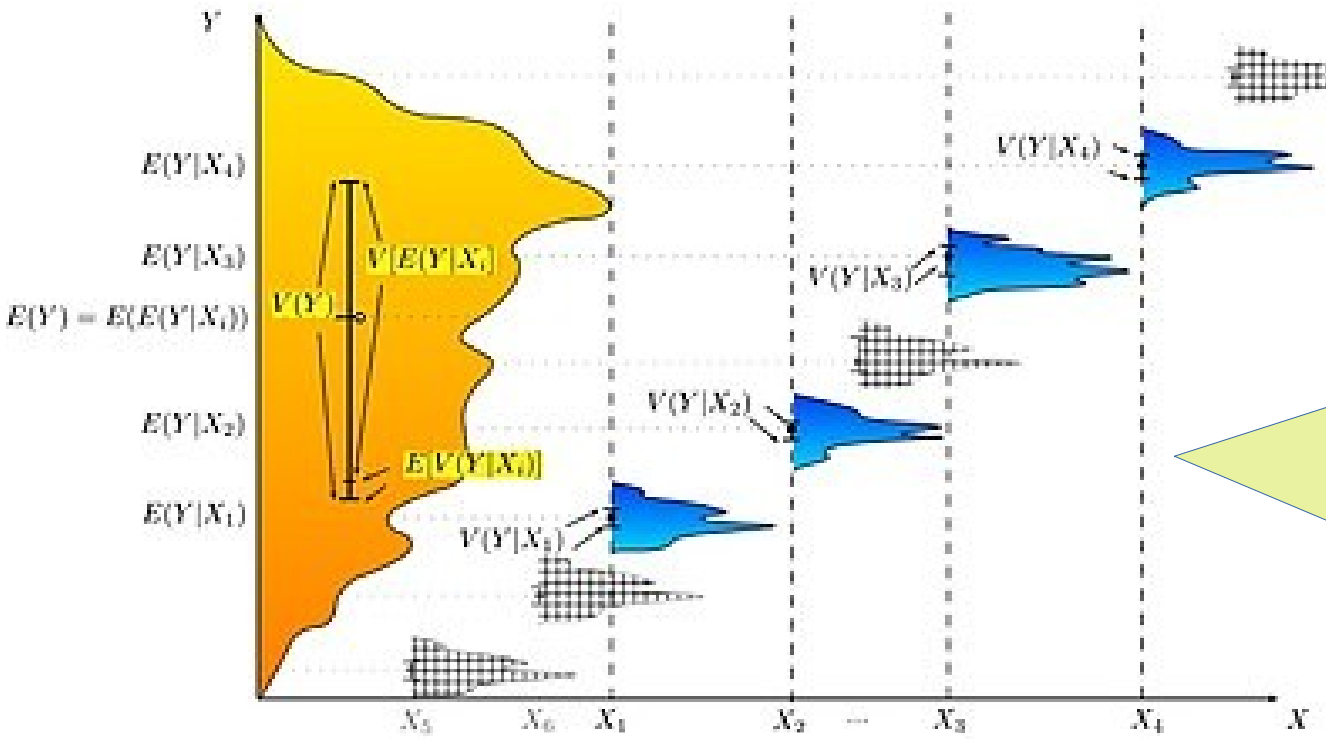
Comparison of Variances Better/Fair Fit



- Orange: mean-weights of all dogs
- Blues: All mean-weights by types of groups
- *Groups are **becoming distinguishable** by mean-weights*

https://en.wikipedia.org/wiki/Analysis_of_variance

Comparison of Variances Better/Fair Fit



The heaviest dogs are likely to be bigger breeds, while the lighter breeds tend to be smaller dogs.

- Orange: mean-weights of all dogs
- Blues: All mean-weights by types of groups
- *Groups **are distinguishable** by mean-weights*

https://en.wikipedia.org/wiki/Analysis_of_variance

Group-making: Lighter Dogs

Dogs groups to similar types of weight

Group-making: Heavy Dogs

Dogs groups to similar types of weight

So What Then?

ANOVA provides a statistical test of equality of weight-means across the groups of dogs.

An ANOVA generalizes the t-test beyond two means.

Let's Hit The Code

Follow the code from the Tutorial and address some questions. URL on next slide.

Activity 05: Complete a Tutorial

ANOVA in R | A Complete Step-by-Step Guide with Examples

Published on March 6, 2020 by Rebecca Bevans. Revised on November 17, 2022.

ANOVA is a statistical test for estimating how a quantitative dependent variable changes according to the levels of one or more categorical independent variables. ANOVA tests whether there is a difference in means of the groups at each level of the independent variable.

https://www.scribbr.com/statistics/anova-in-r/

Repository for tutorial

Click me

Dataset for tutorial

Sample dataset for ANOVA

https://classroom.github.com/a/XasMhHU_