Data Science
CS301
An Overview of Anova Models

Week 7-8
Fall 2024
Oliver BONHAM-CARTER

Are you here today?!

https://forms.gle/iaY7zBmxj8KvsDMa8

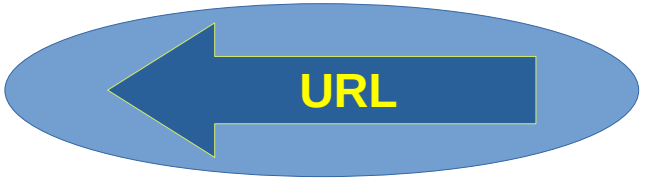
A New Test ...

- **A statistical test to compare the *means* of multiple groups**
- **Allows us to determine the existence of significant differences between them.**
- **Commonly used in research studies to analyze the effects of different variables on an outcome (i.e., useful in scientific research fields.**

Analysis of Variance (Anova)

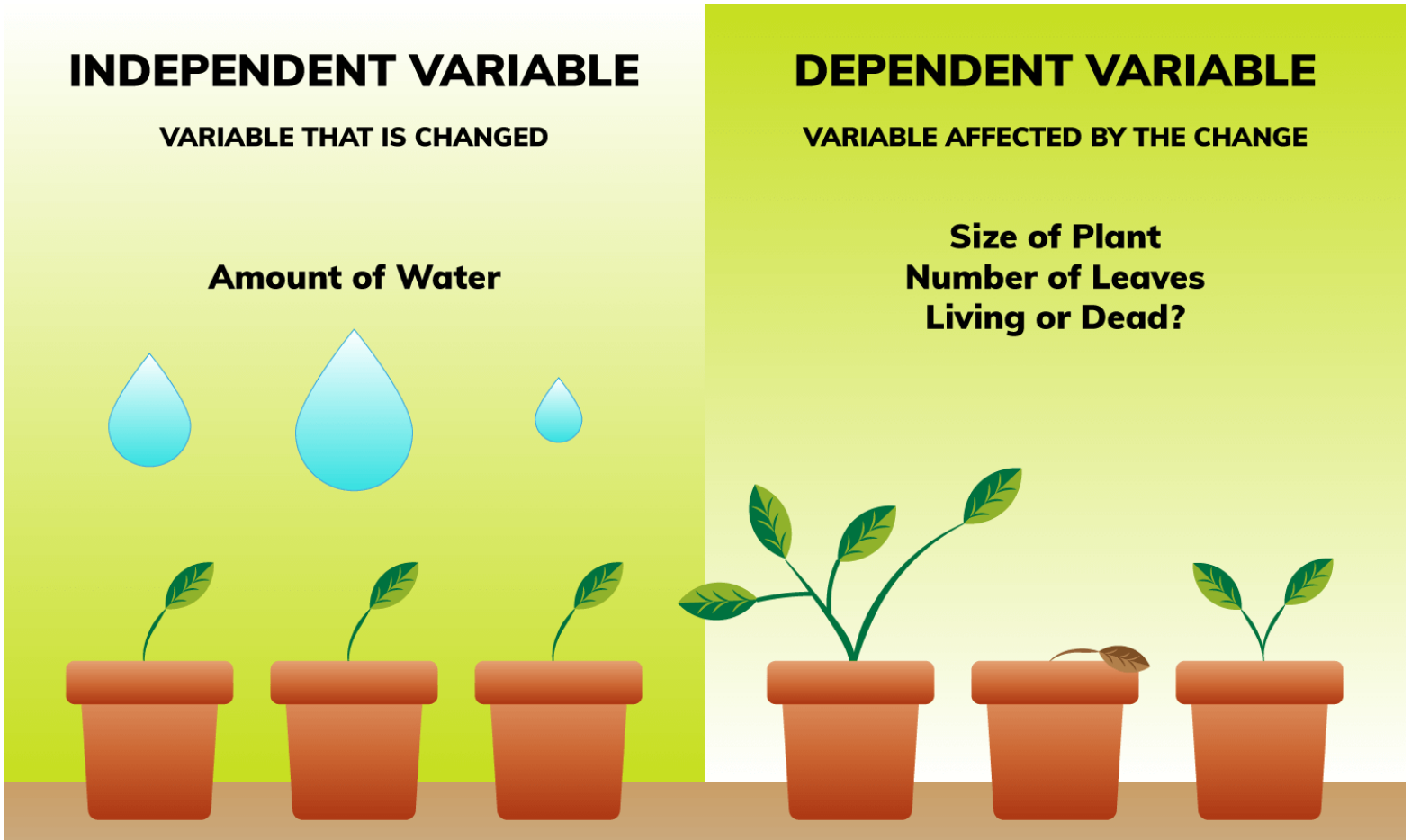
In its simplest form, ANOVA provides a statistical test of whether two (or more) population means are equal, and therefore generalizes the t-test beyond two means.

- **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**
  - **Find out more! CLICK HERE**



Independent, Dependent Variables?

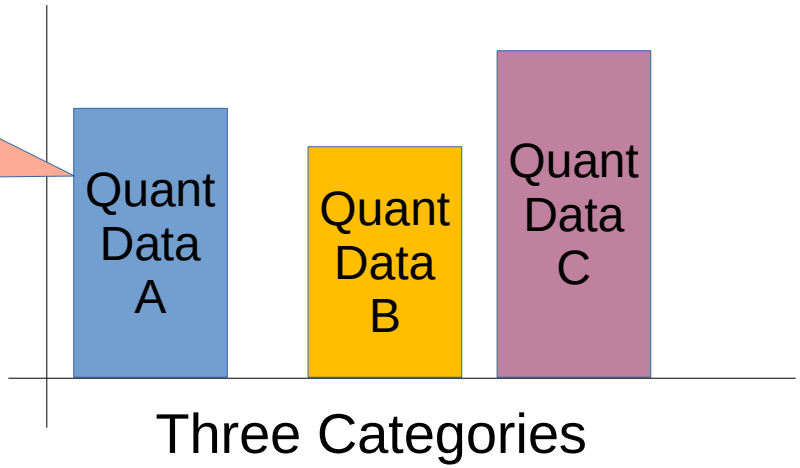
- **Independent** is seemingly random or unpredictable
- **Dependent** is not random; behavior depends on independent variable.



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 (levels)**
- **ANOVA is used to compare the means among *three or more* groups (levels)**
- **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 studies differences in means and the spread of 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 could use the **t-test** to compare the means
- But! 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

Did You Say, *Errors*?

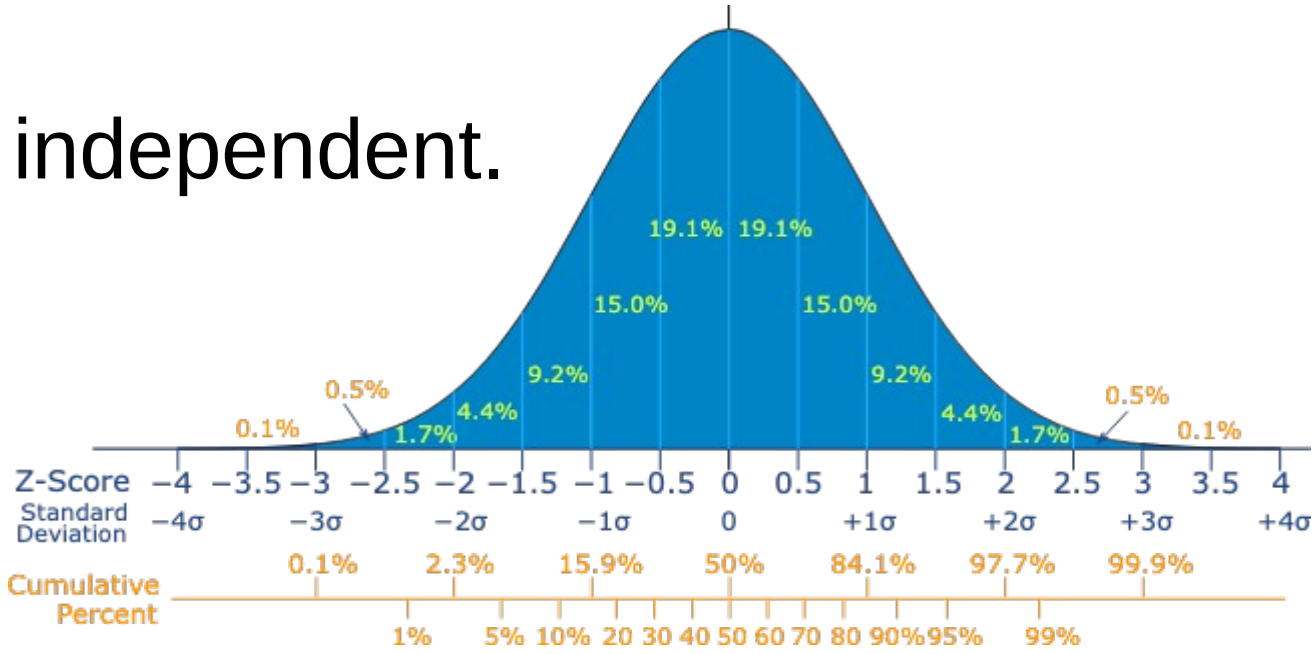
ALLEGHENY COLLEGE

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

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 mean of at least one group is different from the others*
  - Written as, **Ha: μ1 != μ2 != μ3**

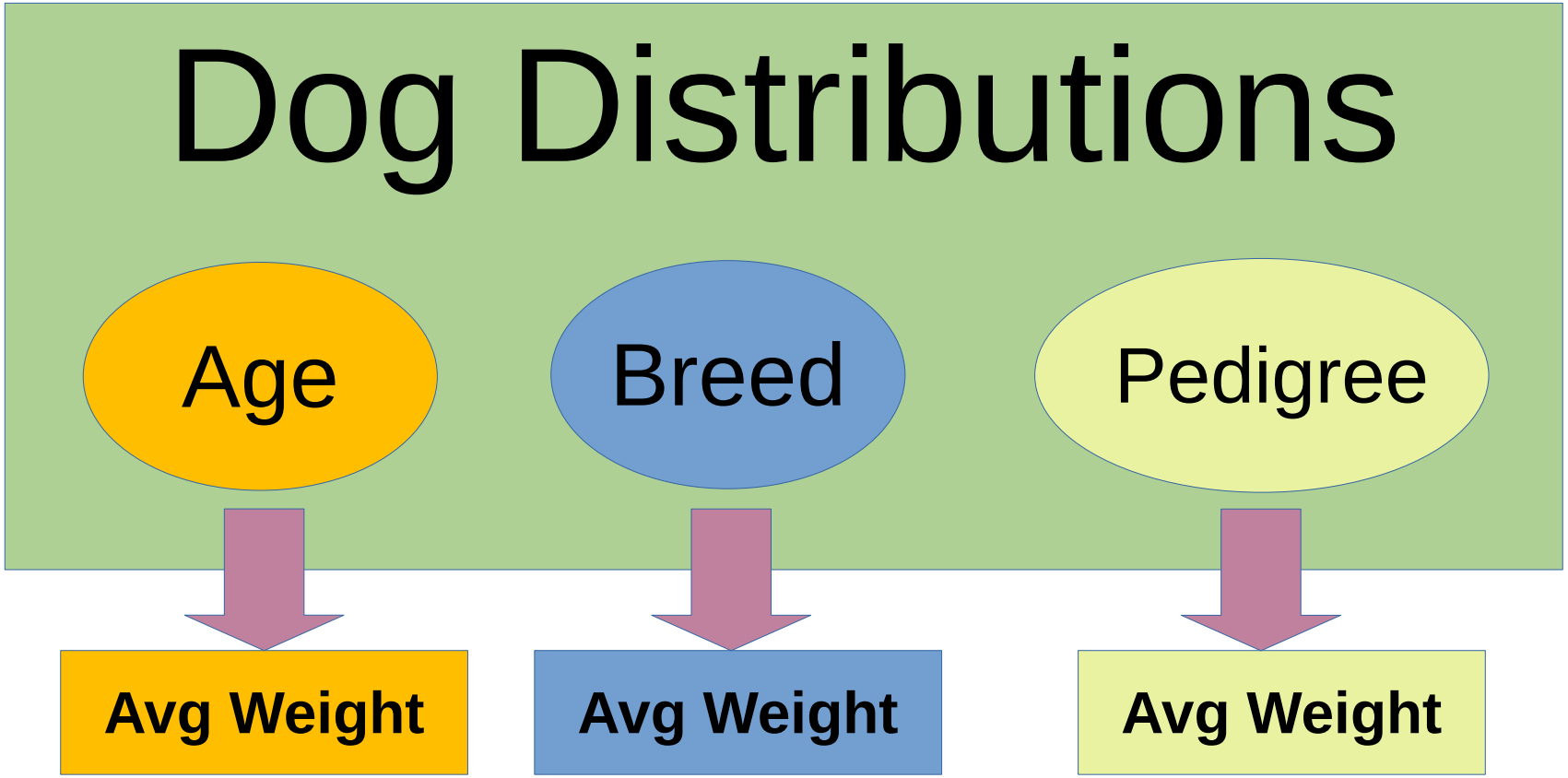
- For,
  - **H0** = the null hypothesis,
  - **Ha** = the alternative hypothesis,
  - **μ1** = the mean of population 1
  - **μ2** = the mean of population 2
  - **μ3** = the mean of population 3

Groups of Dogs

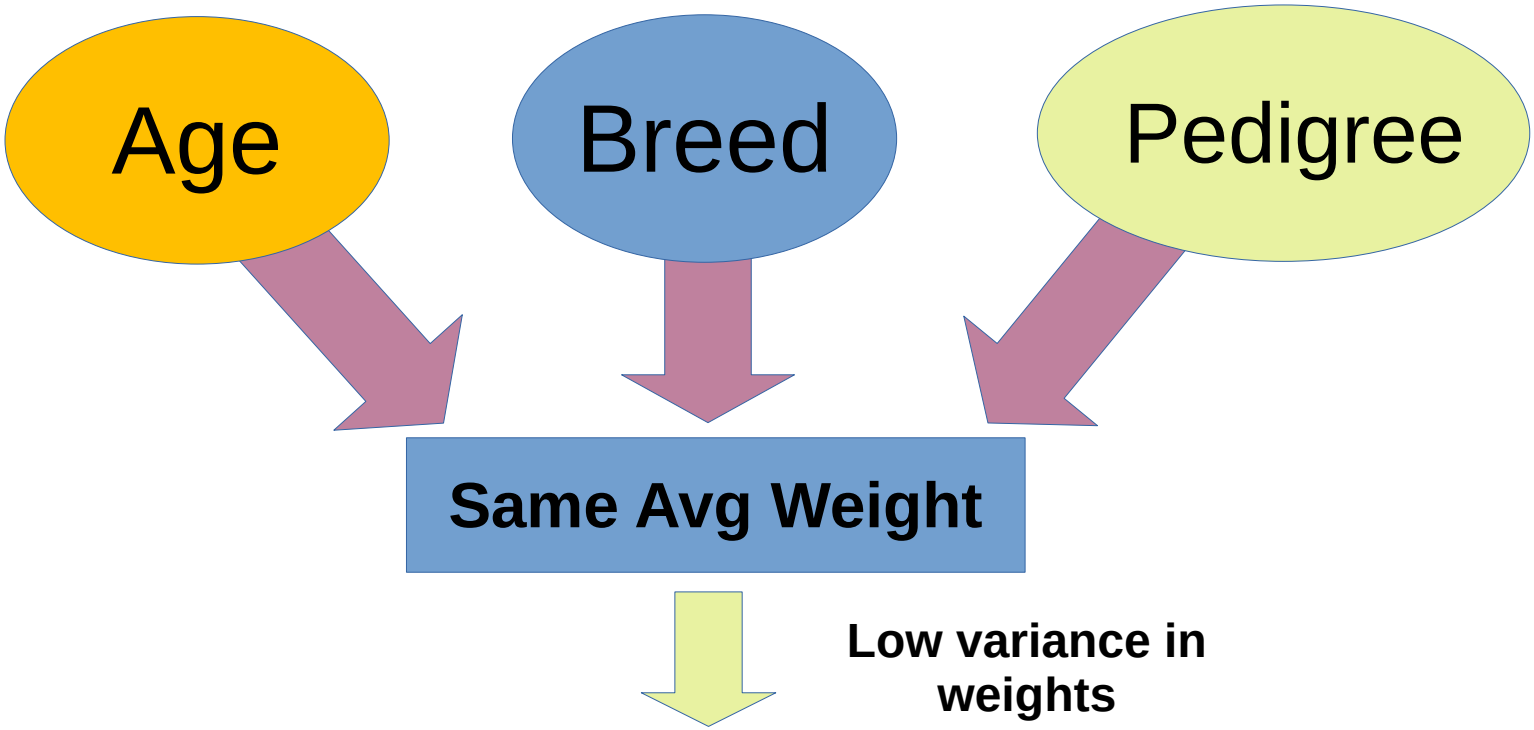
The Dog Show

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

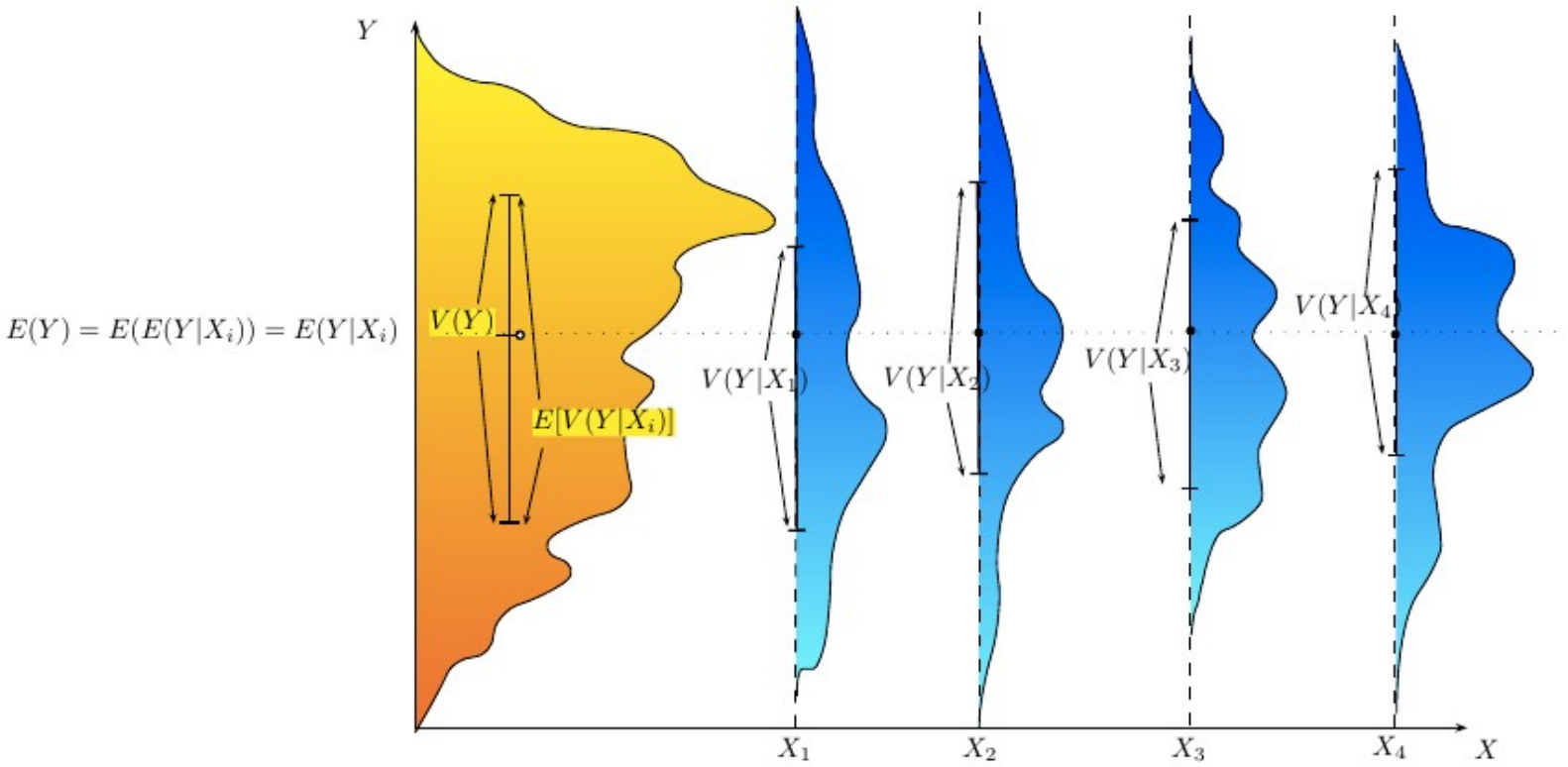
- Separate dogs into distributions having low variance of dog weights (a *heterogeneous* group) …
- The weight-means between groups should be distinct
- If weight-means were similar between groups, then the groups would be similar.

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

The mean-weights would
be similar from each group.

All these Beagles are (both)
cute and
have ***floppy ears***!

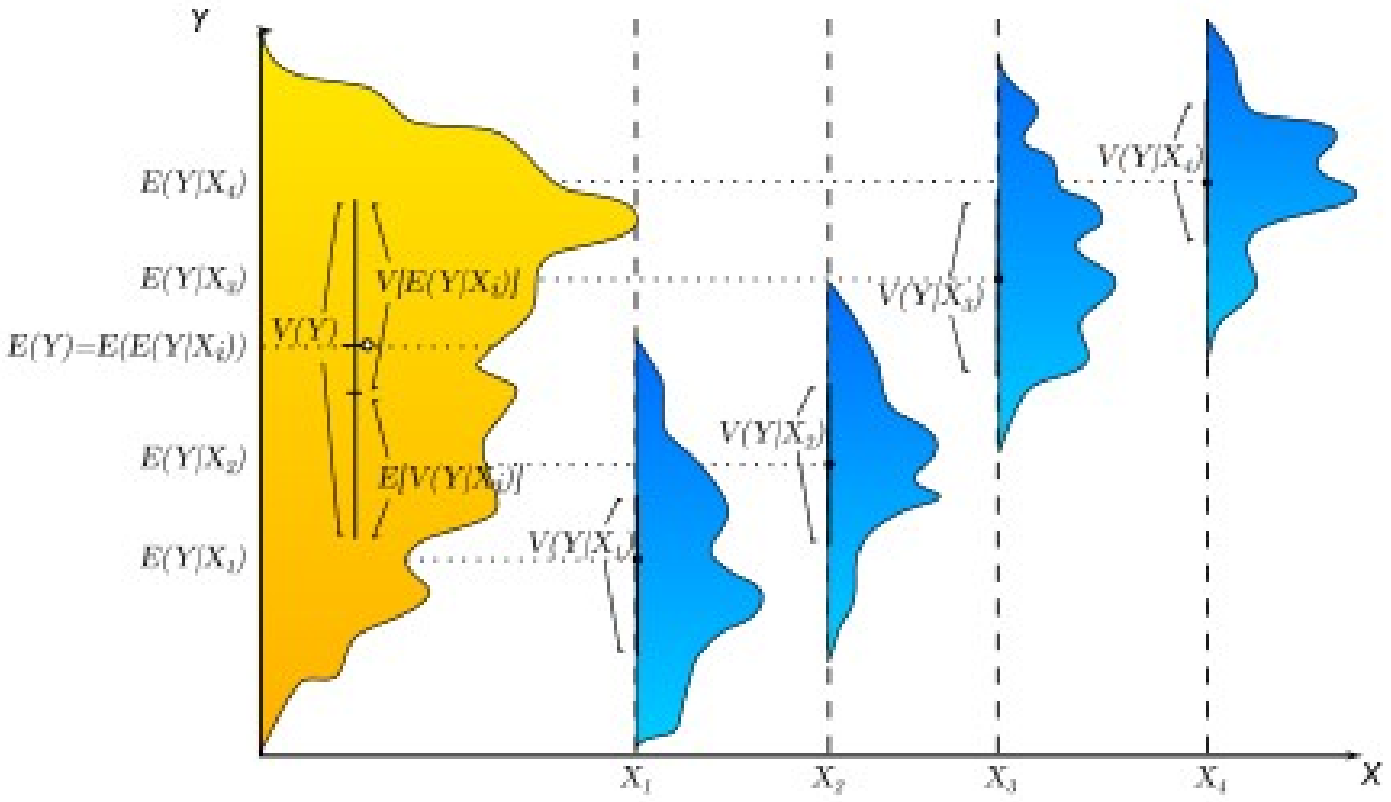
Comparison of Variances
No Fit



- 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

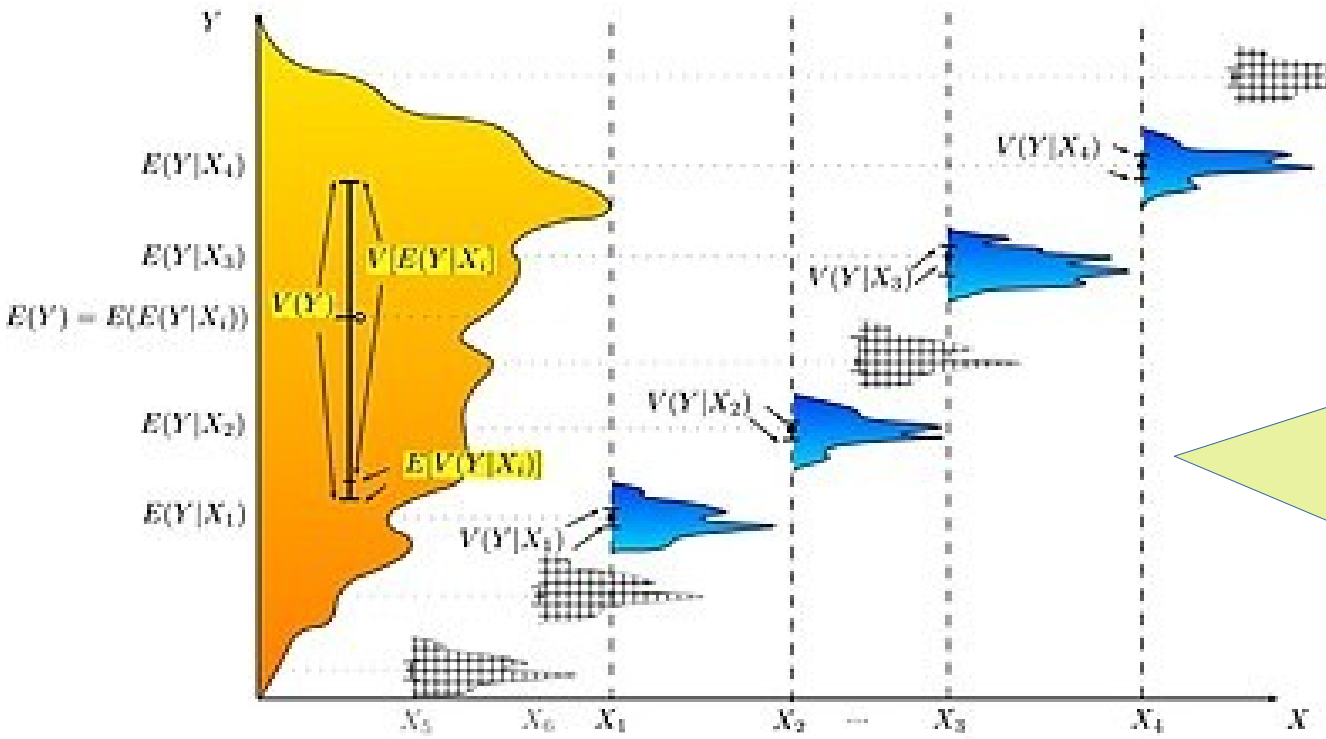
Comparison of Variances Better/Fair Fit



- 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

Dog groups to similar types of weight

Group-making: Heavy Dogs

Dog 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.

Let’s Look at Some Code!

```
# clear out the variables
  rm(list = ls())
# clear out all plots from previous
work.
  graphics.off()
# clear the console
  cat("\014")

# install libraries
if(!require('tidyverse')) {
  install.packages('tidyverse')
  library('tidyverse')
}
```

Create Random Data to Test

```
# Set seed for reproducibility
set.seed(123)

# Generate synthetic data
group1 <- rnorm(30, mean = 5, sd = 1)
group2 <- rnorm(30, mean = 7, sd = 1)
group3 <- rnorm(30, mean = 6, sd = 1)
```

Randomly generated and normal distribution

```
> group1
 [1] 4.439524 4.769823 6.558708 5.070508 5.129288 6.715065 5.460916 3.734939 4.313147
[10] 4.554338 6.224082 5.359814 5.400771 5.110683 4.444159 6.786913 5.497850 3.033383
[19] 5.701356 4.527209 3.932176 4.782025 3.973996 4.271109 4.374961 3.313307 5.837787
[28] 5.153373 3.861863 6.253815
```

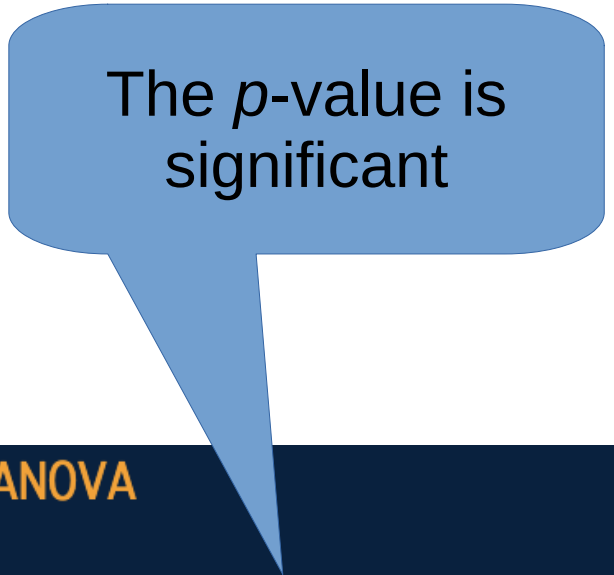
Groups Into Data Frame

```
# Set seed for reproducibility
set.seed(123)

# Generate synthetic data
group1 <- rnorm(30, mean = 5, sd = 1)
group2 <- rnorm(30, mean = 7, sd = 1)
group3 <- rnorm(30, mean = 6, sd = 1)

# Place into a data frame
data <- data.frame(
  value = c(group1, group2, group3),
  group = factor(rep(c("Group 1", "Group 2", "Group 3"),
each = 30))
)
```

Run the Test, Visualize to Understand Results

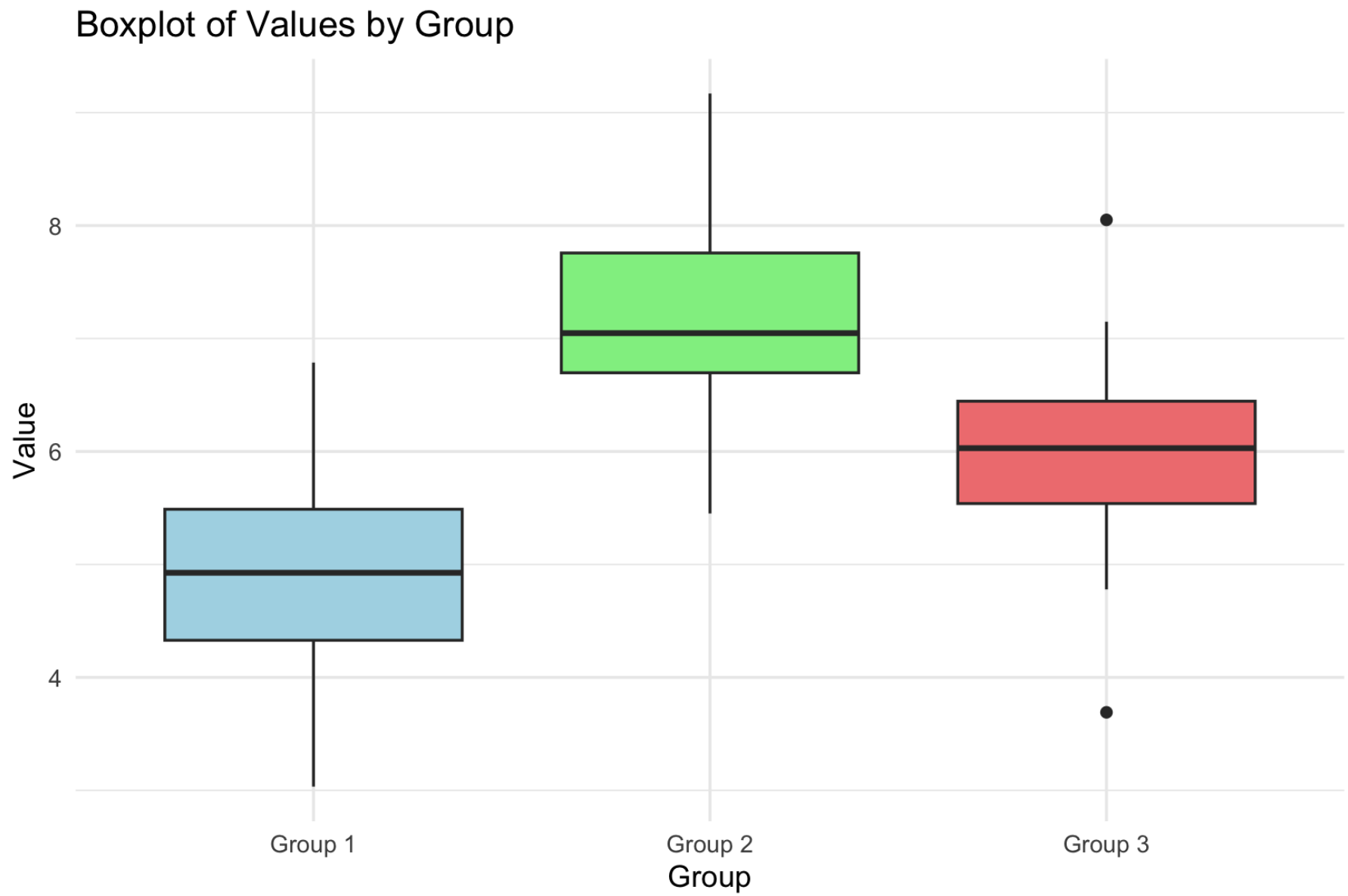
```
# Perform ANOVA
anova_result <- aov(value ~ group, data = data)

# Display the summary of the ANOVA
summary(anova_result)

# Visualize means by boxplot
ggplot(data, aes(x = group, y = value)) +
  geom_boxplot(fill = c("lightblue", "lightgreen",
"lightcoral")) +
  theme_minimal() +
  labs(title = "Boxplot of Values by Group",
       x = "Group",
       y = "Value")
```

Check the Significance

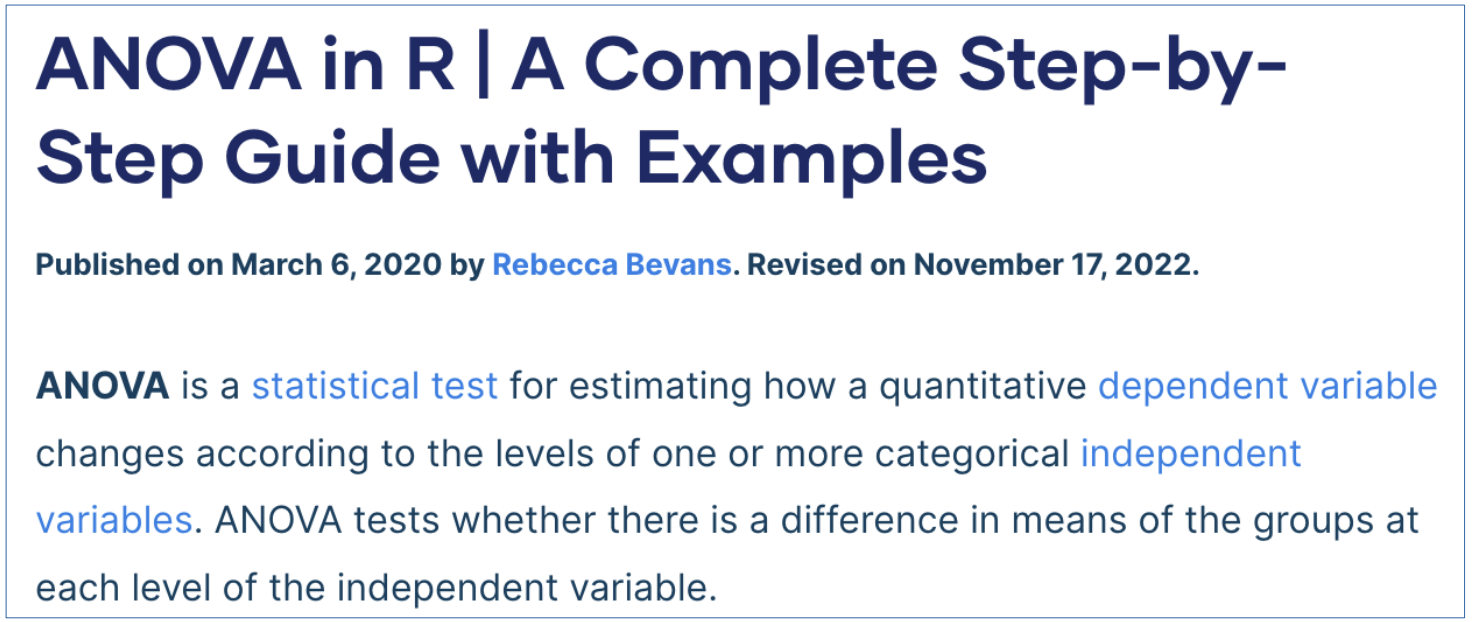



```
> # Display the summary of the ANOVA
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
group         2  74.32   37.16   46.14 2.19e-14 ***
Residuals    87  70.08    0.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Visualize Difference Between Group Means



Anyone For Completing 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/