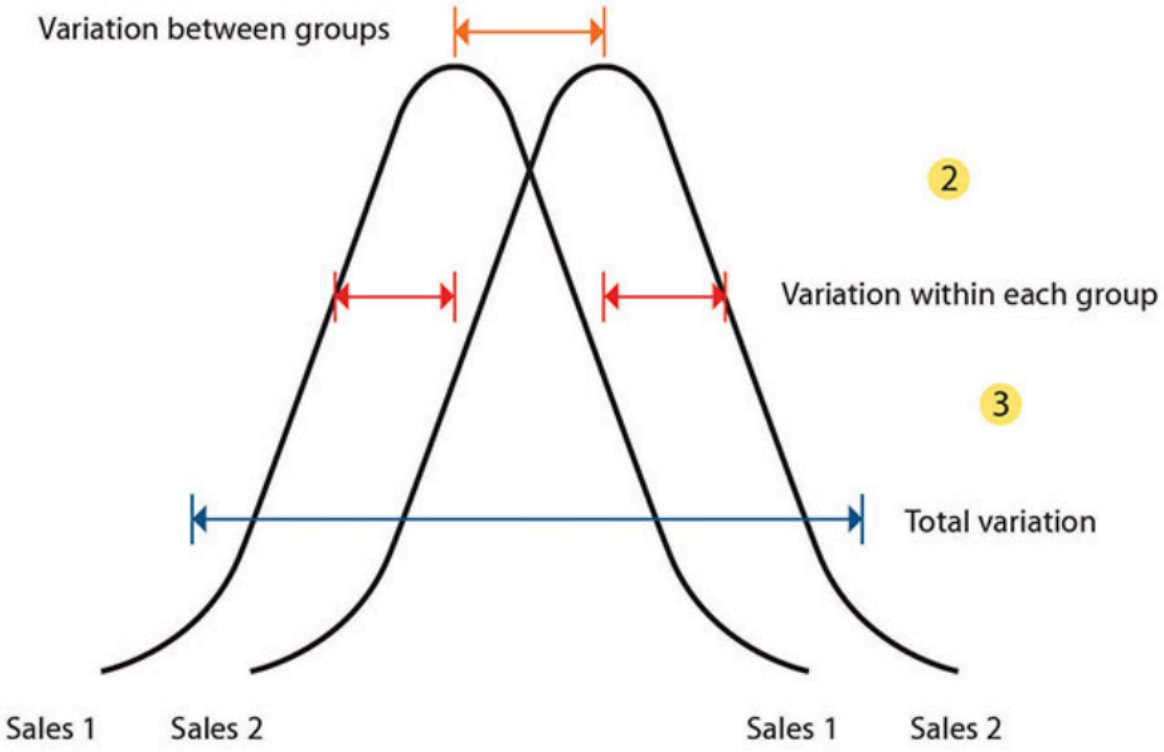

ANOVA



Sales 1	Sales 2
150	170
155	162
157	177
145	192
130	184
170	169
165	155

Two sample groups of sales data

ANOVA: Analysis of Variance is a *variability ratio*

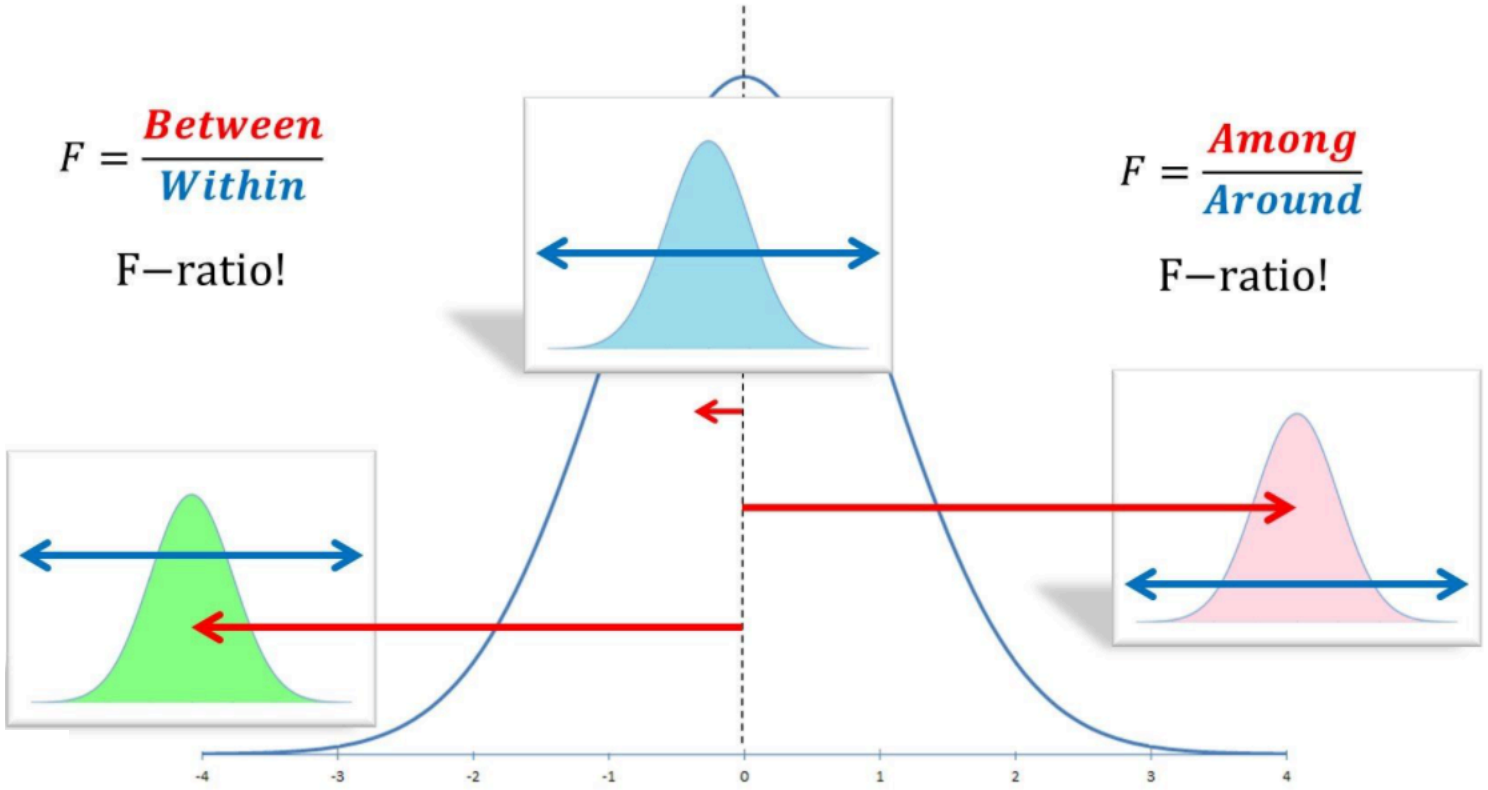
$$Variance\ Between + Variance\ Within = Total\ Variance$$

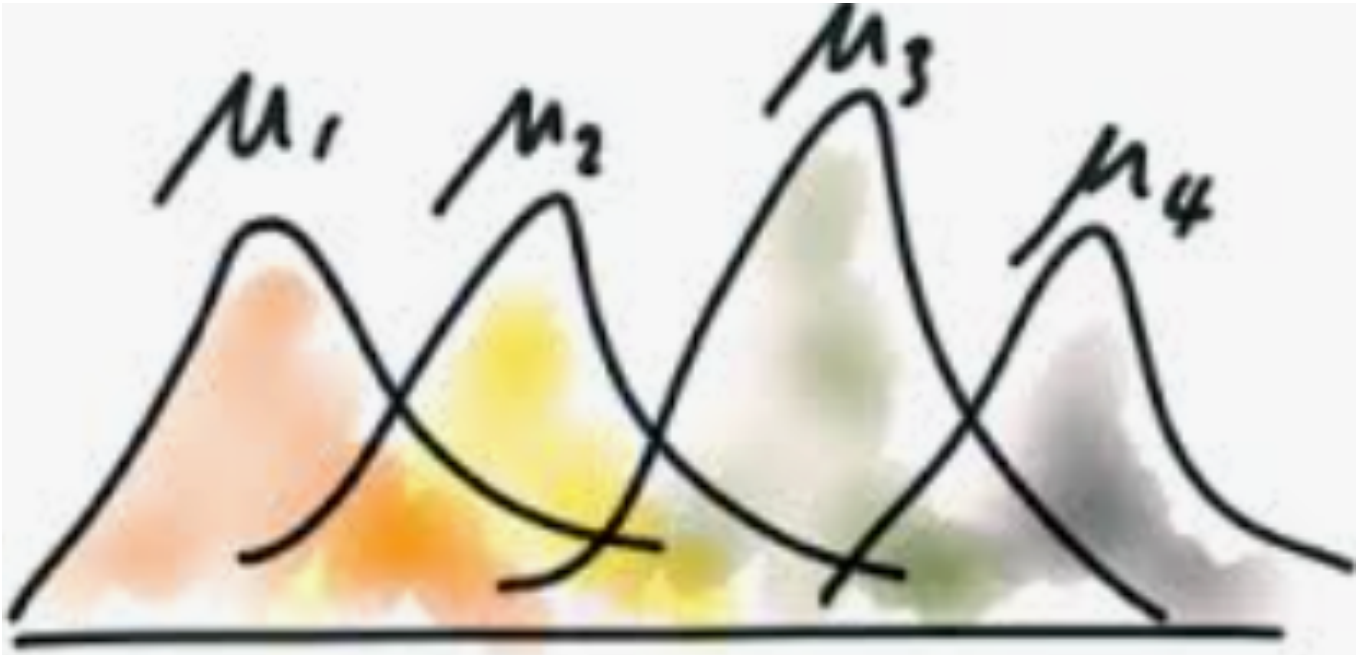
$$F = \frac{\textbf{\textit{Between}}}{\textbf{\textit{Within}}}$$

F−ratio!

$$F = \frac{\textbf{\textit{Among}}}{\textbf{\textit{Around}}}$$

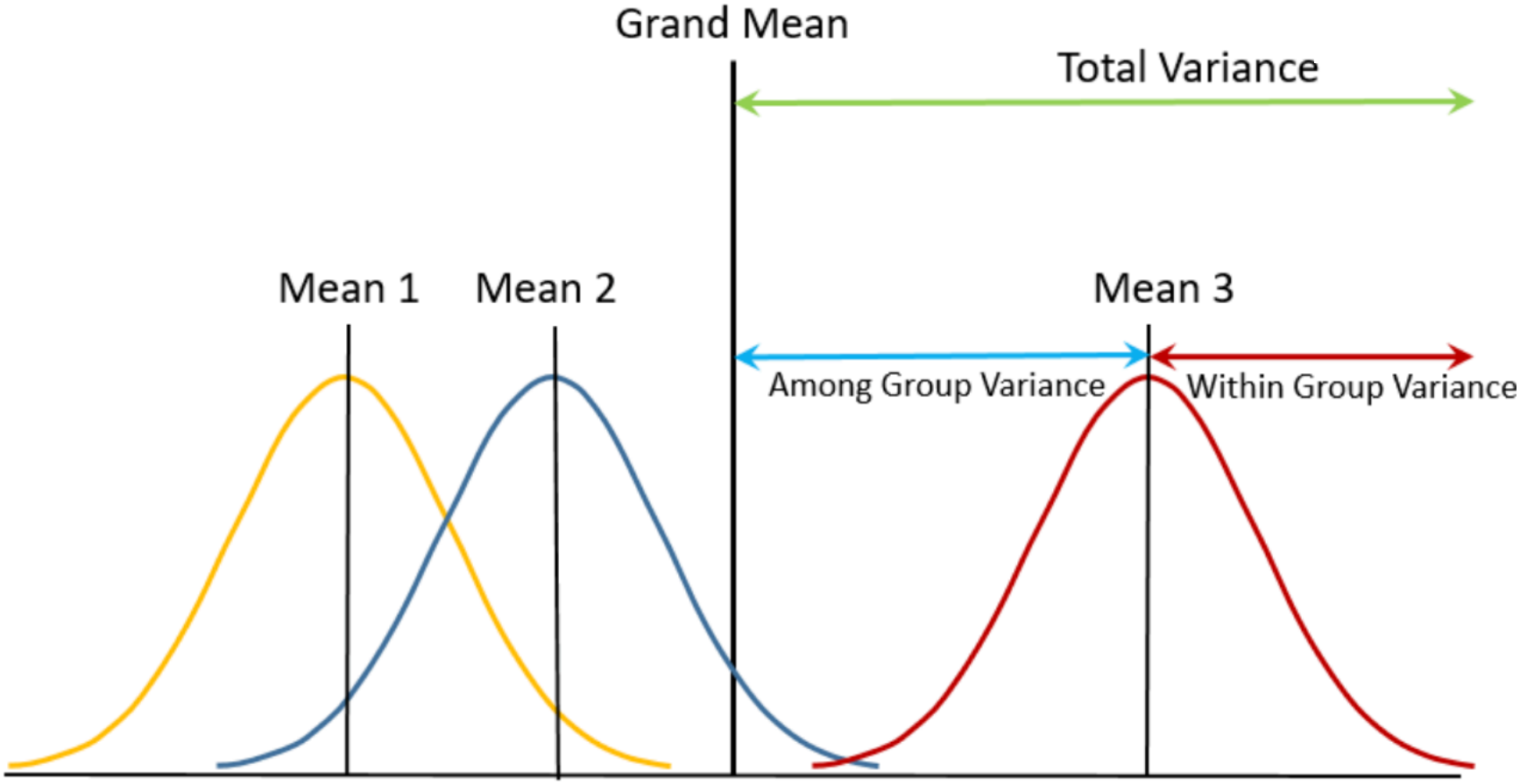
F−ratio!





ANOVA

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 \; ?$$


Grand Mean
Total Variance
Mean 1
Mean 2
Mean 3
Among Group Variance
Within Group Variance

Analysis of Variance

- Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample

- ANOVA was developed by statistician and evolutionary biologist Ronald Fisher

- The observed variance in a particular variable is partitioned into components attributable to different sources of variation

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

Analysis of Variance

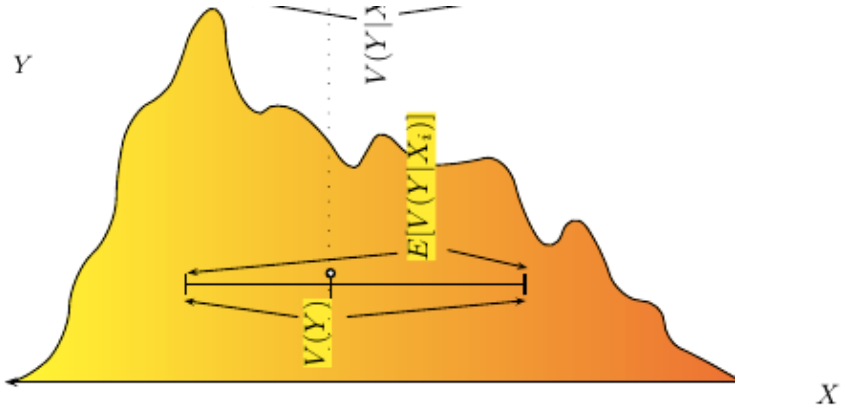
- ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the t-test to more than two groups
- ANOVA is useful for comparing (testing) three or more group means for statistical significance
- It is conceptually similar to multiple two-sample t-tests, but is more conservative, resulting in fewer type I errors and is therefore suited to a wide range of practical problems

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

Analysis of Variance

Example

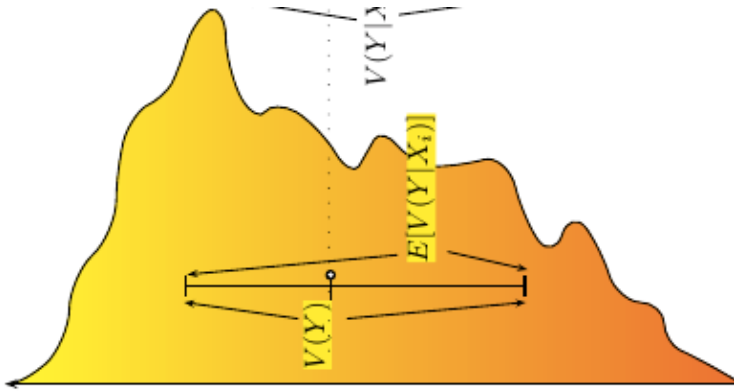
- Used as an exploratory tool to explain observations
- Example: A dog show: limited to dogs that are adult, pure-bred, and exemplary
- A histogram of dog weights from a show is shown in the yellow-orange distribution shown in the illustration



$$E(Y) = E(E(Y|X_i)) = E(Y|X_i)$$

Analysis of Variance

- We want to predict the weight of a dog based on a certain set of characteristics of each dog.
- One way to do that is to explain the distribution of weights by dividing the dog population into groups based on those characteristics
- A successful grouping will split dogs such that
  - each group has a low variance of dog weights (assuming the group is relatively homogeneous)
  - the mean of each group is distinct (if two groups have the same mean, then it isn't reasonable to conclude that the groups are, in fact, separate in any meaningful way)



$$E(Y) = E(E(Y|X_i)) = E(Y|X_i)$$

Analysis of Variance

- Groups are identified as X1, X2, X3, X4
- The dogs are divided according to two binary groupings: young vs old, and short-haired vs long-haired
  - Group X1 is young, short-haired dogs
  - Group X2 is young, long-haired dogs
  - Group X3 is old, short-haired dogs
  - Group X4 is old, long-haired dogs
- The distributions of dog weight within each of the groups (shown in blue) has a large variance, and the means are very similar across groups. Grouping dogs by these characteristics does not produce an effective way to explain the variation in dog weights: knowing which group a dog is in doesn't allow us to predict its weight much better than simply knowing the dog is in a dog show
- Thus, this grouping fails to explain the variation in the overall distribution (yellow-orange).

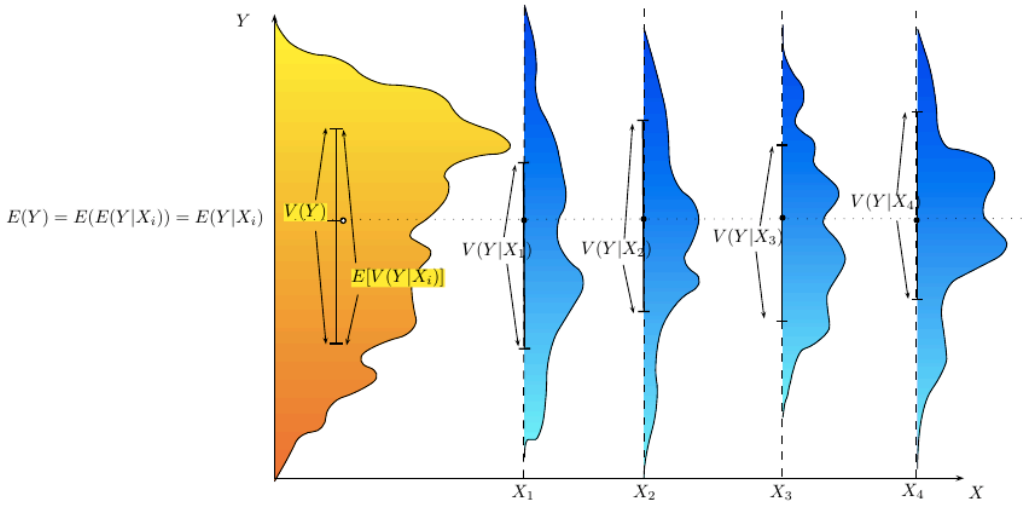


Figure 2: ANOVA : No fit

Analysis of Variance

- An attempt to explain the weight distribution by grouping dogs as pet vs working breed and less athletic vs more athletic would probably be somewhat more successful (fair fit).
- The heaviest show dogs are likely to be big, strong, working breeds, while breeds kept as pets tend to be smaller and thus lighter.
- As shown by the second illustration, the distributions have variances that are considerably smaller than in the first case, and the means are more distinguishable. However, the significant overlap of distributions, for example, means that we cannot distinguish X1 and X2 reliably.
- Grouping dogs according to a coin flip might produce distributions that look similar.

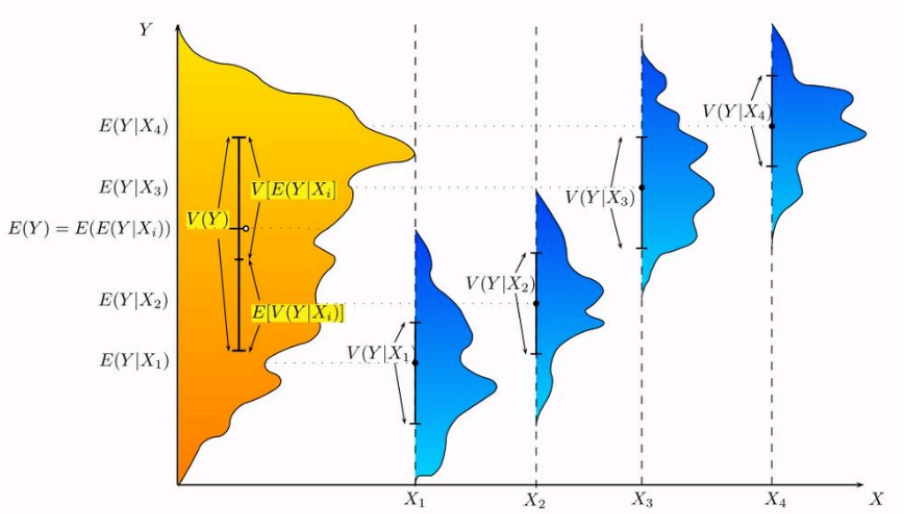


Figure 1: ANOVA : Fair fit

Analysis of Variance

- An attempt to explain weight by breed is likely to produce a very good fit. All Chihuahuas are light and all St Bernards are heavy.
- The difference in weights between Setters and Pointers does not justify separate breeds.
- The analysis of variance provides the formal tools to justify these intuitive judgments.
- The method has some advantages over correlation: not all of the data must be numeric and one result of the method is a judgment in the confidence in an explanatory relationship.

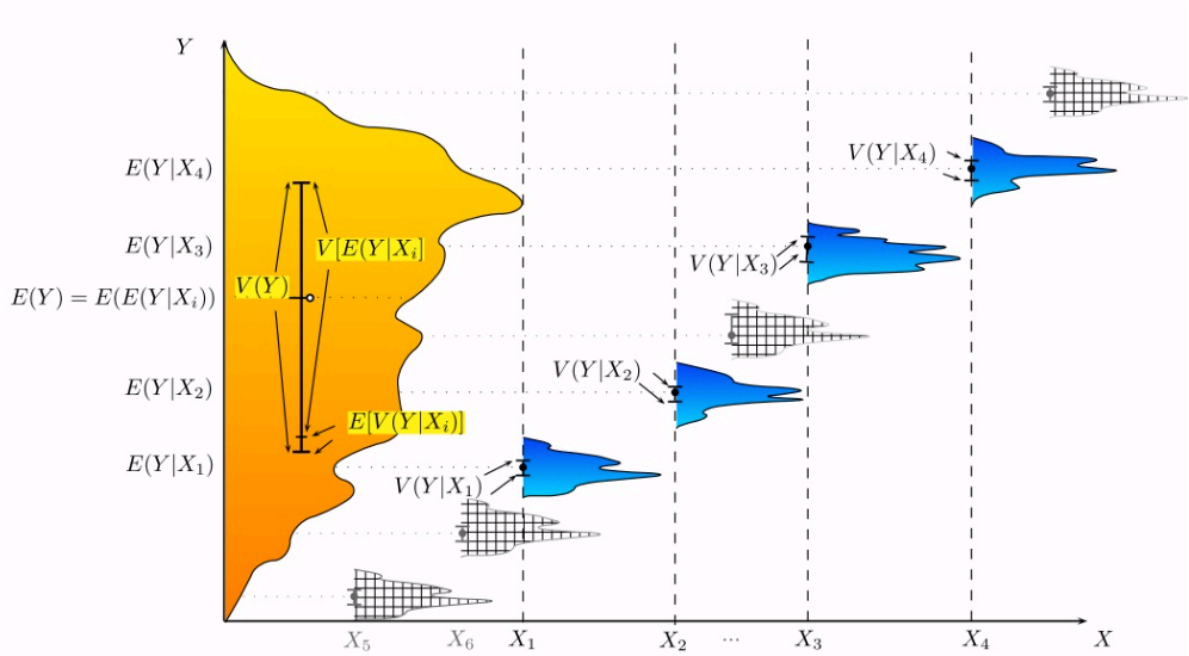


Figure 3: ANOVA : very good fit

One-way ANOVA

There are many situations where you need to compare the mean between multiple groups. For instance, the marketing department wants to know if three teams have the same sales performance.

- Team: 3 level factor: A, B, and C
- Sale: A measure of performance

The ANOVA test can tell if the three groups have similar performances.

To clarify if the data comes from the same population, you can perform a **one-way analysis of variance** (one-way ANOVA hereafter). This test, like any other statistical tests, gives evidence whether the H0 hypothesis can be accepted or rejected.

One-way ANOVA

Hypothesis in one-way ANOVA test:

- H0: The means between groups are identical
- H3: At least, the mean of one group is different

In other words, the H0 hypothesis implies that there is not enough evidence to prove the mean of the group (factor) are different from another.

This test is similar to the t-test, although ANOVA test is recommended in situation with more than 2 groups. Except that, the t-test and ANOVA provide similar results.

Assumptions

We assume that each factor is randomly sampled, independent and comes from a normally distributed population with unknown but equal variances.

Interpret ANOVA test

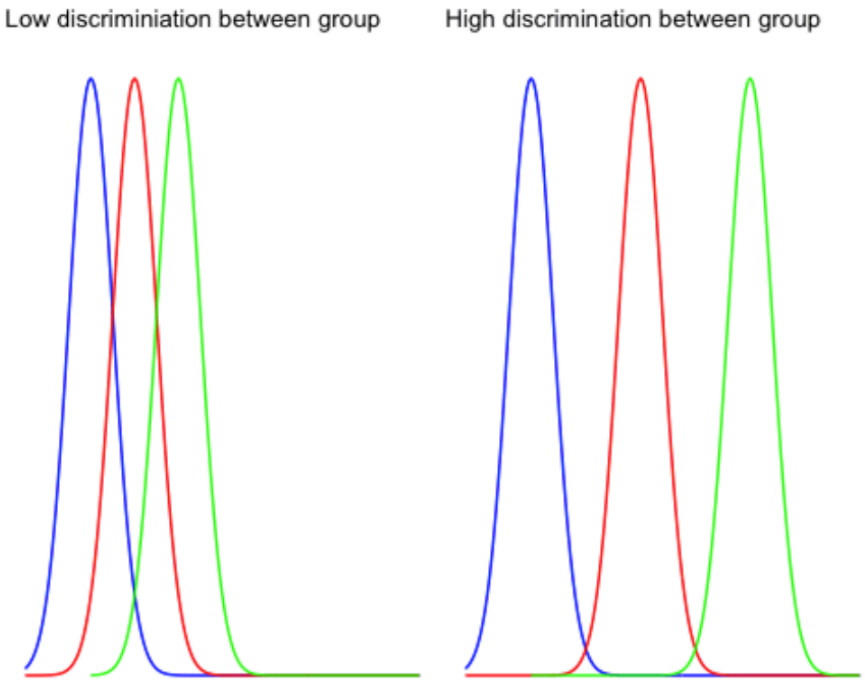
The F-statistic is used to test if the data are from significantly different populations, i.e., different sample means.

To compute the F-statistic, you need to divide the **between-group variability** over the **within-group variability**.

The **between-group** variability reflects the differences between the groups inside all of the population. Look at the two graphs below to understand the concept of between-group variance.

The left graph shows very little variation between the three group, and it is very likely that the three means tends to the **overall** mean (i.e., mean for the three groups).

The right graph plots three distributions far apart, and none of them overlap. There is a high chance the difference between the total mean and the groups mean will be large.

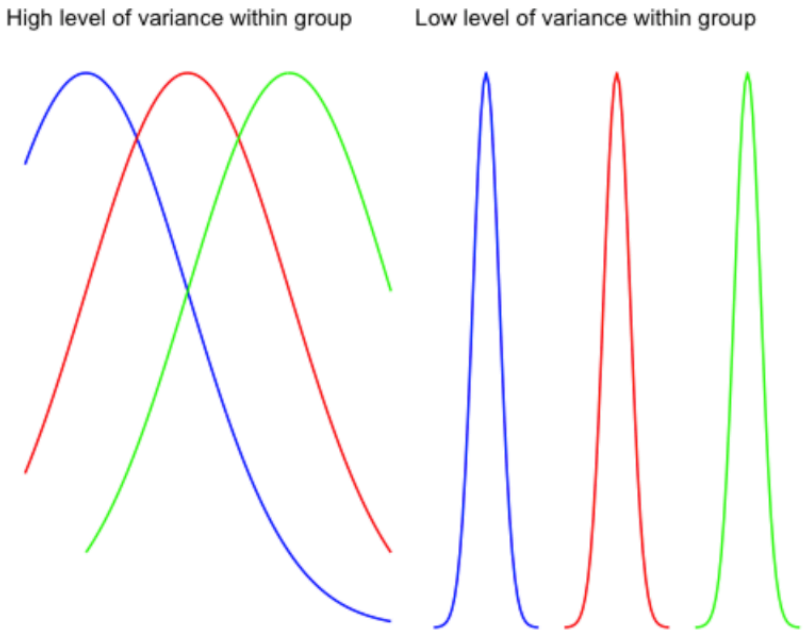


The **within group** variability considers the difference between the groups. The variation comes from the individual observations; some points might be totally different than the group means. The **within group** variability picks up this effect and refer to the sampling error.

To understand visually the concept of within group variability, look at the graph below.

The left part plots the distribution of three different groups. You increased the spread of each sample and it is clear the individual variance is large. The F-test will decrease, meaning you tend to accept the null hypothesis

The right part shows exactly the same samples (identical mean) but with lower variability. It leads to an increase of the F-test and tends in favor of the alternative hypothesis.



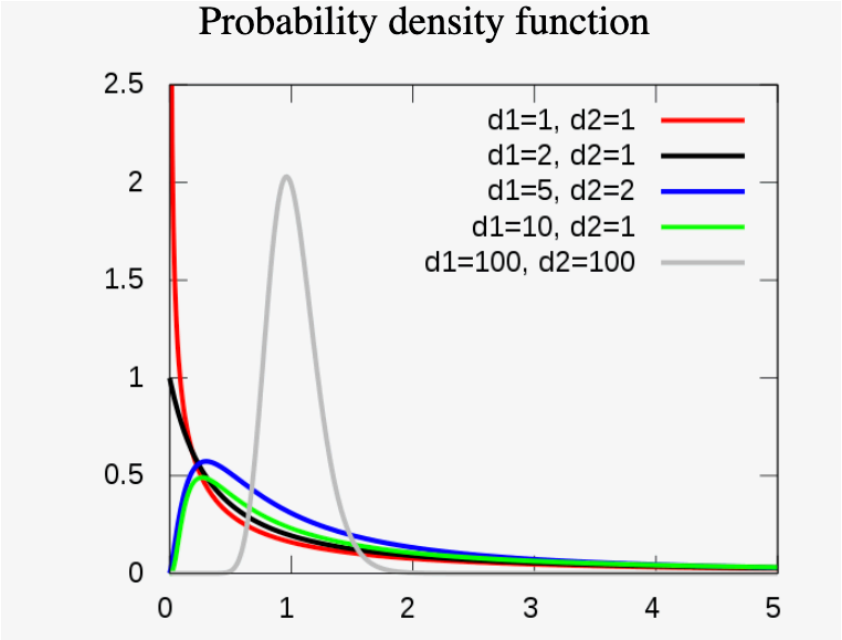
The F Test (Fisher-Snedecor)

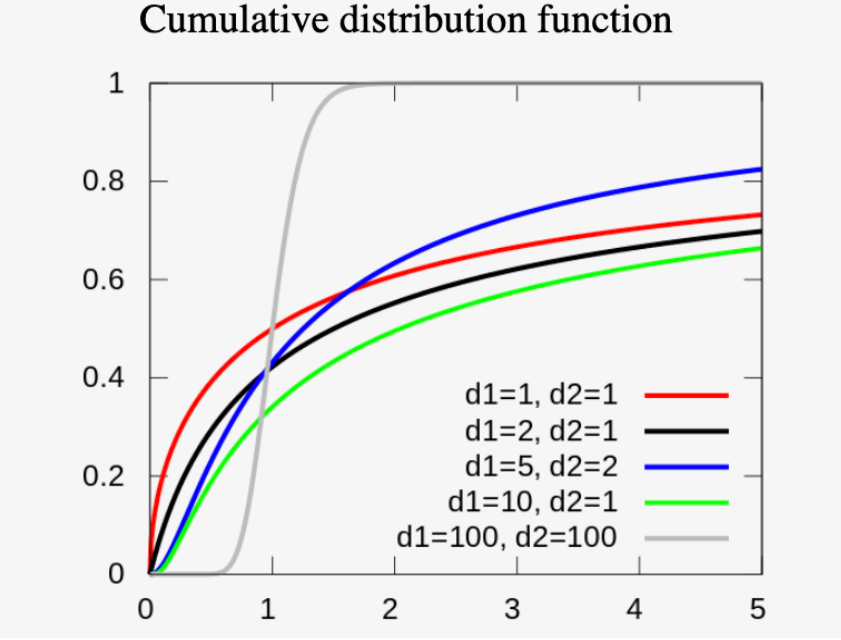
- An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.
- **It is most often used when comparing statistical models** that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.
- Exact "F-tests" mainly arise when the models have been fitted to the data using least squares.

https://en.wikipedia.org/wiki/F-test

The F Distribution

It is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA), e.g., F-test



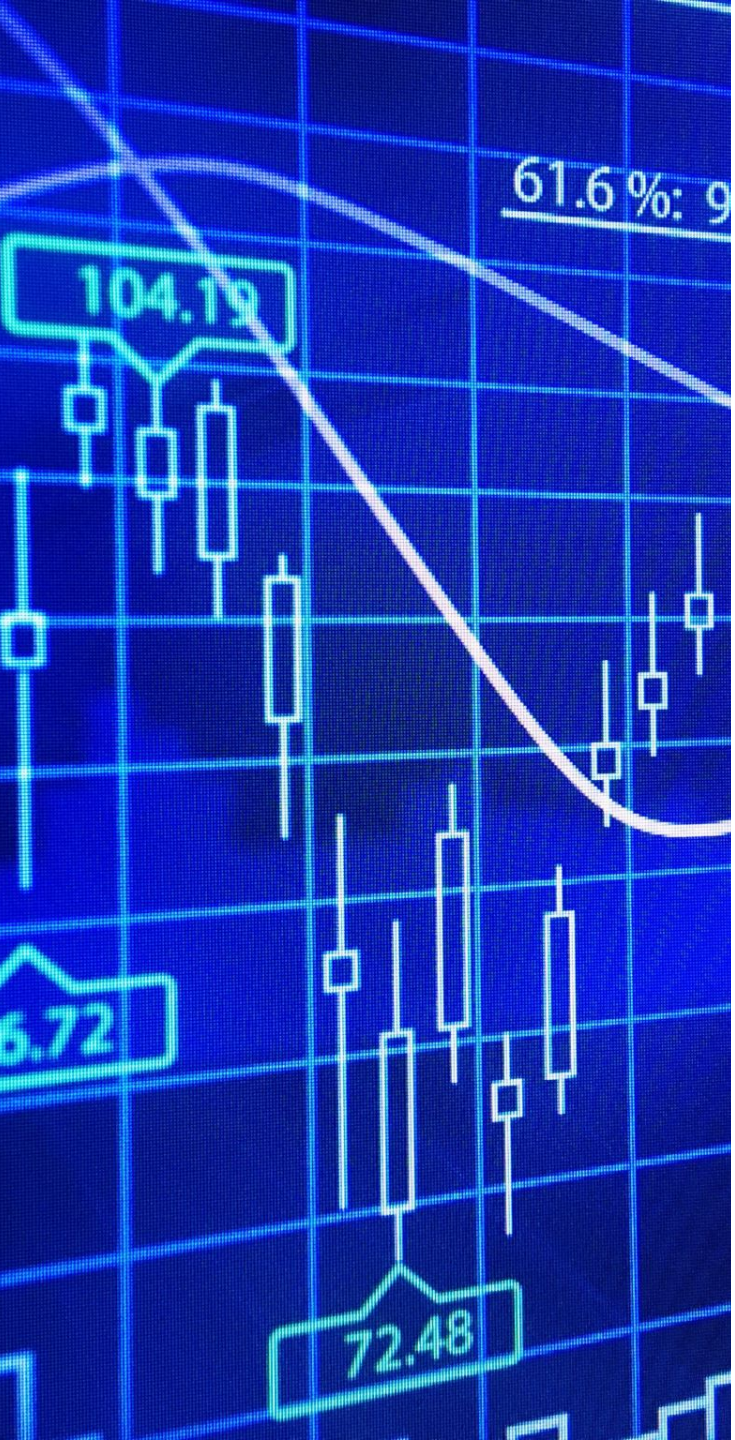


https://en.wikipedia.org/wiki/F-distribution

Uses of the F-Test

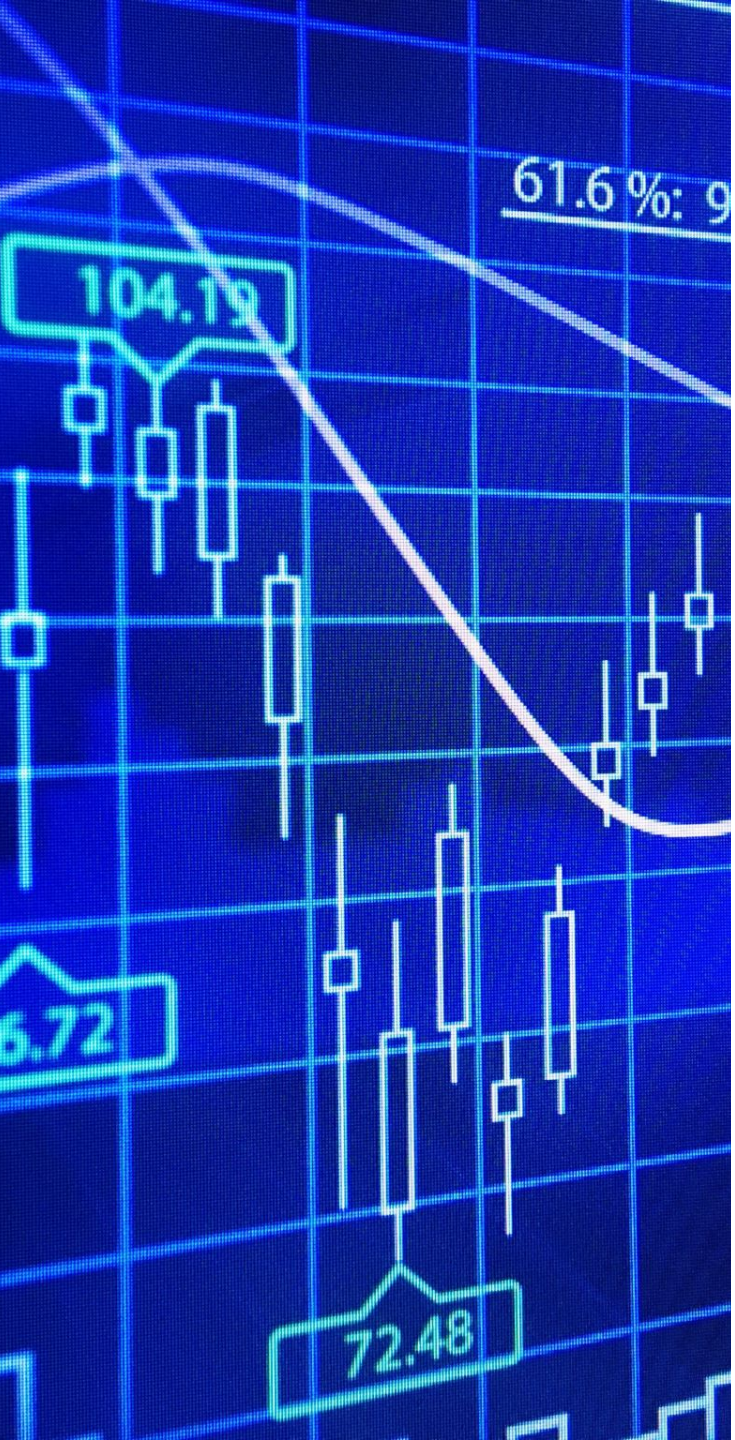
- The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal. This is perhaps the best-known F-test and plays an important role in the analysis of variance (ANOVA).
- The hypothesis that a proposed regression model fits the data well. See Lack-of-fit sum of squares.
- The hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested within each other.

https://en.wikipedia.org/wiki/F-test



Analysis of Variance in Python

- Download Python script anova1.py
- https://datasciencechalktalk.com/2019/09/04/one-way-analysis-of-variance-anova-with-python/



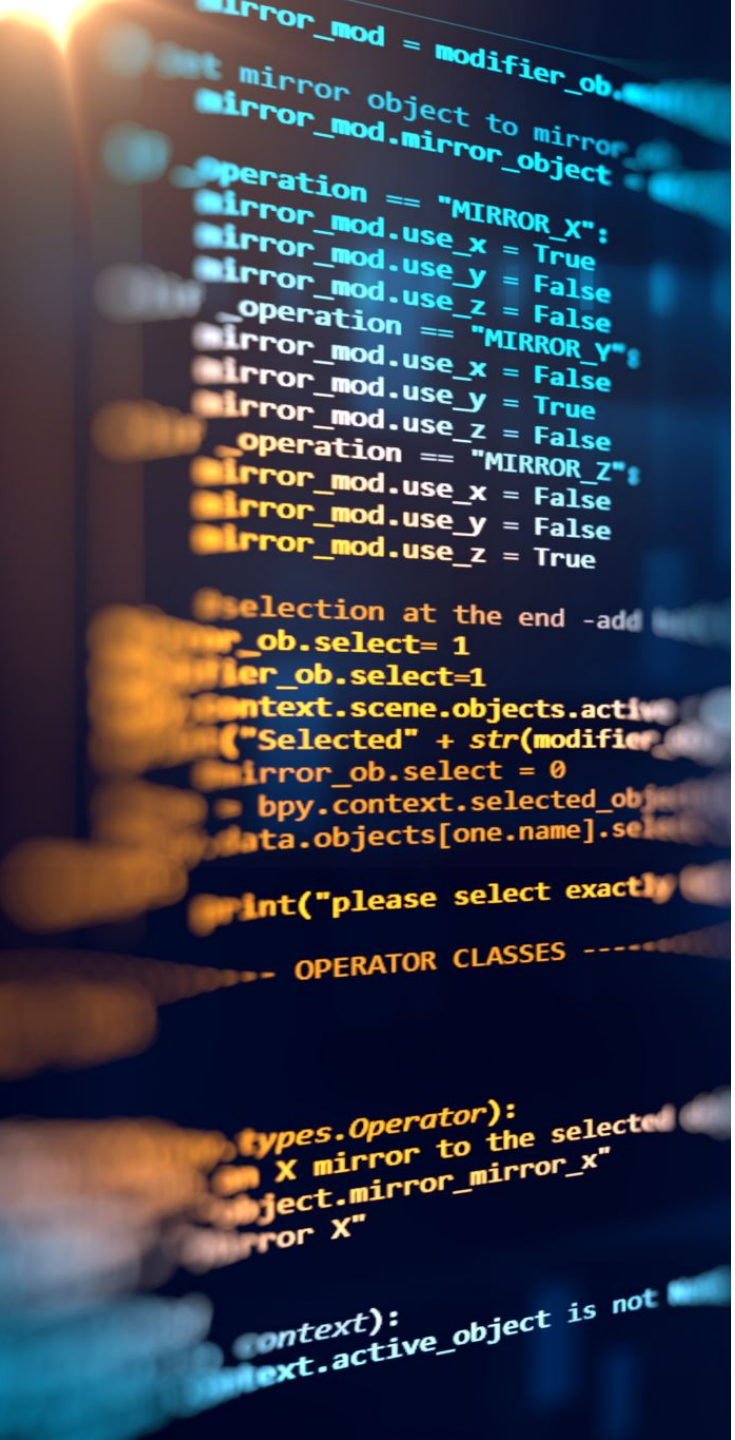
Analysis of Variance in Python

- Download Python script anova2.py
- https://www.reneshbedre.com/blog/anova.html

17 Normality Tests with Python

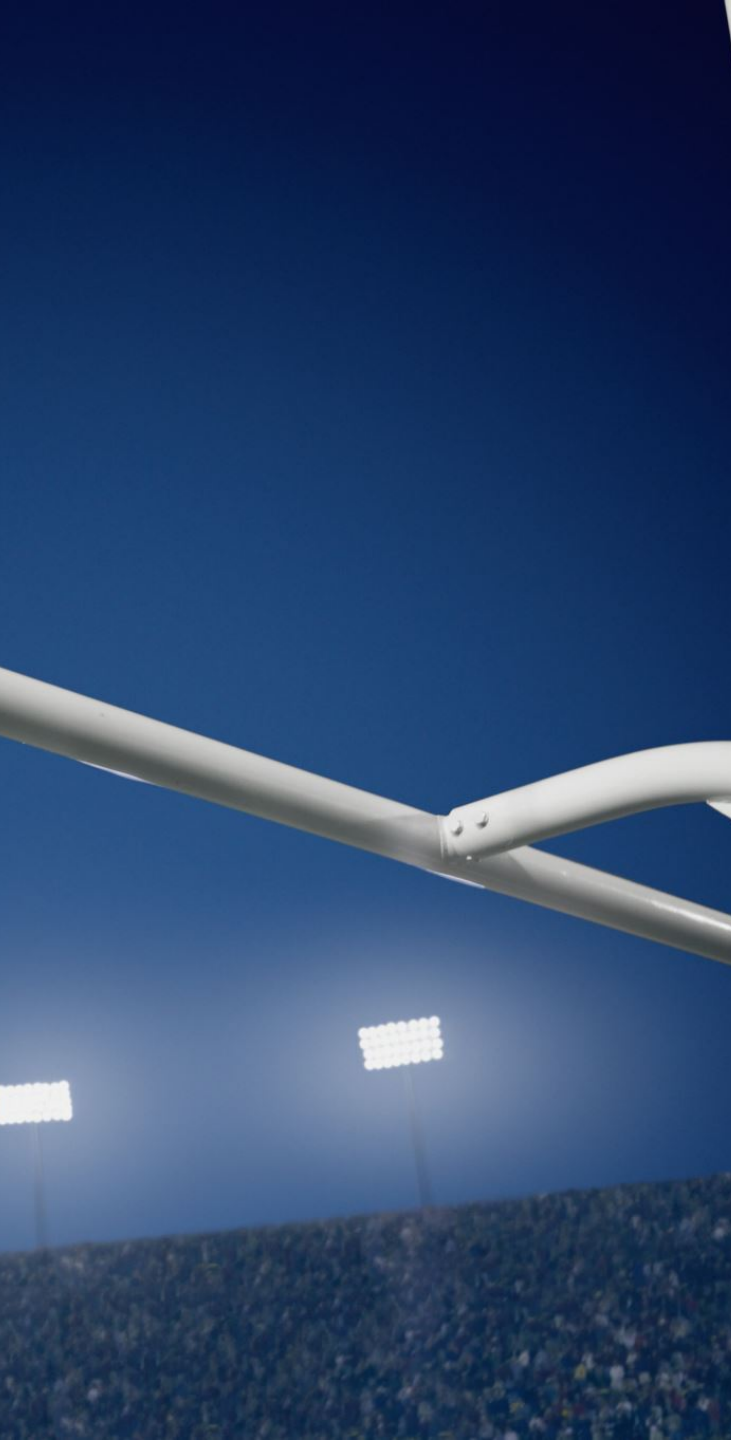
Download Python script
normality-tests.py

https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/



Other ANOVA examples in Python

- https://www.pythonfordatascience.org/anova-python/
- https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/
- https://medium.com/@rrfd/f-tests-and-anovas-examples-with-the-iris-dataset-fe7caa3e21d0



Hands-on Exercises in R

- https://data-flair.training/blogs/hypothesis-testing-in-r/

ANOVA Examples in R

Example One way ANOVA Test

You will use the poison dataset to implement the one-way ANOVA test. The dataset contains 48 rows and 3 variables:

- Time: Survival time of the animal
- poison: Type of poison used: factor level: 1,2 and 3
- treat: Type of treatment used: factor level: 1,2 and 3

Before you start to compute the ANOVA test, you need to prepare the data as follow:

- Step 1: Import the data
- Step 2: Remove unnecessary variable
- Step 3: Convert the variable poison as ordered level

```
library(dplyr)
PATH <- "https://raw.githubusercontent.com/guru99-edu/R-Programming/master/poi
sons.csv"
df <- read.csv(PATH) %>%
select(-X) %>%
mutate(poison = factor(poison, ordered = TRUE))
glimpse(df)
```

ANOVA examples in R

Our objective is to test the following assumption:

- H0: There is no difference in survival time average between group
- H3: The survival time average is different for at least one group.

In other words, you want to know if there is a statistical difference between the mean of the survival time according to the type of poison given to the Guinea pig.

You will proceed as follow:

- Step 1: Check the format of the variable poison
- Step 2: Print the summary statistic: count, mean and standard deviation
- Step 3: Plot a box plot
- Step 4: Compute the one-way ANOVA test
- Step 5: Run a pairwise t-test

ANOVA examples in R

Step 1) You can check the level of the poison with the following code. You should see three character values because you convert them in factor with the mutate verb.

```
levels(df$poison)
```

Output:

```
## [1] "1" "2" "3"
```

Step 2) You compute the mean and standard deviation.

```
df % > %
        group_by(poison) % > %
        summarise(
                count_poison = n(),
                mean_time = mean(time, na.rm = TRUE),
                sd_time = sd(time, na.rm = TRUE)
        )
```

Output:

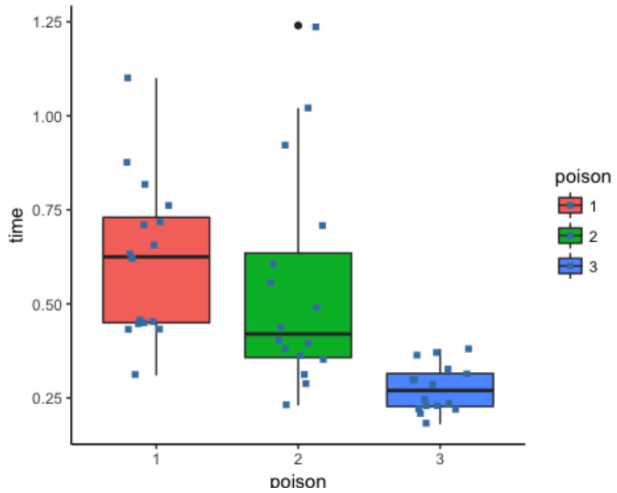
```
##
# A tibble: 3 x 4
##   poison count_poison mean_time    sd_time
##    <ord>        <int>     <dbl>      <dbl>
## 1      1           16  0.617500 0.20942779
## 2      2           16  0.544375 0.28936641
## 3      3           16  0.276250 0.06227627
```

ANOVA examples in R

Step 3) In step three, you can graphically check if there is a difference between the distribution. Note that you include the jittered dot.

```
ggplot(df, aes(x = poison, y = time, fill = poison)) +
    geom_boxplot() +
    geom_jitter(shape = 15,
        color = "steelblue",
        position = position_jitter(0.21)) +
    theme_classic()
```

Output:



ANOVA examples in R

Step 4) You can run the one-way ANOVA test with the command aov. The basic syntax for an ANOVA test is:

```
aov(formula, data)
Arguments:
- formula: The equation you want to estimate
- data: The dataset used
```

The syntax of the formula is:

```
y ~ X1+ X2+...+Xn # X1 +  X2 +... refers to the independent variables
y ~ . # use all the remaining variables as independent variables
```

You can answer our question: Is there any difference in the survival time between the Guinea pig, knowing the type of poison given.

Note that, it is advised to store the model and use the function summary() to get a better print of the results.

```
anova_one_way <- aov(time~poison, data = df)
summary(anova_one_way)
```

Code Explanation

- aov(time ~ poison, data = df): Run the ANOVA test with the following formula
- summary(anova_one_way): Print the summary of the test

ANOVA examples in R

Output:

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## poison       2  1.033  0.5165   11.79 7.66e-05 ***
## Residuals   45  1.972  0.0438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is lower than the usual threshold of 0.05. You are confident to say there is a statistical difference between the groups, indicated by the "*".

ANOVA examples in R

Pairwise comparison

The one-way ANOVA test does not inform which group has a different mean. Instead, you can perform a Tukey test with the function TukeyHSD().

```
TukeyHSD(anova_one_way)
```

Output:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ poison, data = df)

$poison
         diff        lwr         upr     p adj
2-1 -0.073125 -0.2525046  0.10625464 0.5881654
3-1 -0.341250 -0.5206296 -0.16187036 0.0000971
3-2 -0.268125 -0.4475046 -0.08874536 0.0020924
```

Upper bound of the difference in mean between group

Adjusted p-value when there are multiple groups

Lower bound of the difference in mean between group

ANOVA Examples in R

Two-way ANOVA

A two-way ANOVA test adds another group variable to the formula. It is identical to the one-way ANOVA test, though the formula changes slightly:

$$y=x1+x2$$

with is a quantitative variable and and are categorical variables.

Hypothesis in two-way ANOVA test:

- H0: The means are equal for both variables (i.e., factor variable)
- H3: The means are different for both variables

You add treat variable to our model. This variable indicates the treatment given to the Guinea pig. You are interested to see if there is a statistical dependence between the poison and treatment given to the Guinea pig.

ANOVA Examples in R

```
anova_two_way <- aov(time~poison + treat, data = df)
summary(anova_two_way)
```

Output:

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## poison       2 1.0330  0.5165   20.64 5.7e-07 ***
## treat        3 0.9212  0.3071   12.27 6.7e-06 ***
## Residuals   42 1.0509  0.0250
## ---
```

You can conclude that both poison and treat are statistically different from 0. You can reject the NULL hypothesis and confirm that changing the treatment or the poison impact the time of survival.

ANOVA Summary in R

Summary

We can summarize the test in the table below:

Test	code	hypothesis	p-value
One way ANOVA	`aov(y ~ X, data = df)`	H3: Average is different for at least one group	0.05
Pairwise	`TukeyHSD(ANOVA summary)`		0.05
Two way ANOVA	`aov(y ~ X1 + X2, data = df)`	H3: Average is different for both group	0.05