USC Department Of Translational Genomics
Keck Medicine of USC

# TRGN 527: Applied Data Science and Bioinformatics

## UNIT IV. Un-supervised Statistical Tests, Multiple Testing

## Week 8 - Lecture 1

**Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S.** | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

**David W. Craig, Ph.D.** | **Professor and Vice Chair**

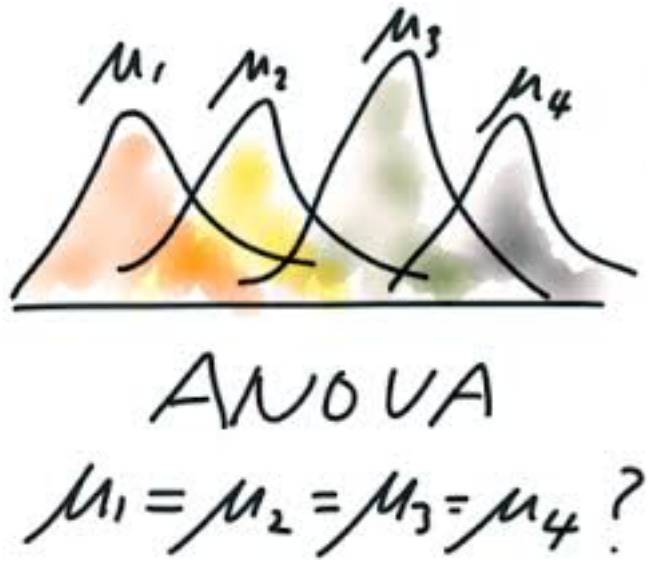
Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

# Topics

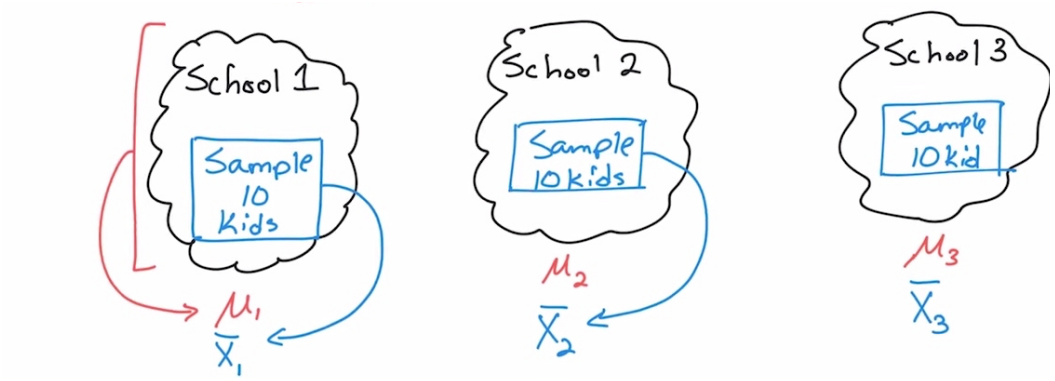
- Analysis of Variance (ANOVA)



# ANOVA

- Last classes we explained how two-sample t-tests can be used to evaluate hypotheses regarding the difference between the means of two groups.
  - We typically use this approach to investigate the relationship between a binary categorical (factor) variable.
    - It specifies the two groups, and a numerical variable, which is regarded as the response variable.

- Now we will discuss about "Analysis of Variance" (ANOVA)
  - Models that generalize the t-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories.




NOTE: A hypothesis test that is used to compare the means of **two populations** is called **t-test**.
A statistical technique that is used to compare the means of **more than two populations** is known as Analysis of Variance or **ANOVA**

# ANOVA

- As mentioned before:
  - The **categorical variable** is called the factor and is typically considered as the explanatory variable
  - The **numerical variable**, whose means across different groups are compared, is regarded as the response variable.
- Now we will focus first on ANOVA models with only one factor.
  - These models are known as one-way ANOVA.
- Also we will briefly discuss two-way ANOVA models that include two factors
  - Two categorical explanatory variables in the analysis.

# ANOVA

- Analysis of Variance (ANOVA)
  - Dataset:
    - We will use a dataset from the MASS package
      - It is the Cushings data
    - Background:
      - Cushing's syndrome is a hormone disorder associated with high level of cortisol secreted by the adrenal gland.
      - The Cushings data set includes 27 observations (n = 27).
      - For each individual in the sample, the urinary excretion rates of two steroid metabolites are recorded.
        - Urinary excretion rate (mg/24 hr) of Tetrahydrocortisone and
        - Urinary excretion rate (mg/24 hr) of Pregnanetriol.
        - The Type variable in the data set shows the underlying type of syndrome, which can be one of four categories:
          - adenoma (a)
          - bilateral hyperplasia (b)
          - carcinoma (c)
          - unknown (u).

# ANOVA

- Load the data in R:

```
23 # Load Data in R
24 ```{R}
25
26 library(MASS)
27 head(Cushings)
28 Cushings
29
30 ```
```

| | Tetrahydrocortisone | Pregnanetriol | Type |
|---|---|---|---|
| a1 | 3.1 | 11.70 | a |
| a2 | 3.0 | 1.30 | a |
| a3 | 1.9 | 0.10 | a |
| a4 | 3.8 | 0.04 | a |
| a5 | 4.1 | 1.10 | a |
| a6 | 1.9 | 0.40 | a |
| b1 | 8.3 | 1.00 | b |
| b2 | 3.8 | 0.20 | b |
| b3 | 3.9 | 0.60 | b |
| b4 | 7.8 | 1.20 | b |
| b5 | 9.1 | 0.60 | b |
| b6 | 15.4 | 3.60 | b |
| b7 | 7.7 | 1.60 | b |
| b8 | 6.5 | 0.40 | b |
| b9 | 5.7 | 0.40 | b |
| b10 | 13.6 | 1.60 | b |
| c1 | 10.2 | 6.40 | c |
| c2 | 9.2 | 7.90 | c |
| c3 | 9.6 | 3.10 | c |
| c4 | 53.8 | 2.50 | c |
| c5 | 15.8 | 7.60 | c |
| u1 | 5.1 | 0.40 | u |
| u2 | 12.9 | 5.00 | u |
| u3 | 13.0 | 0.80 | u |
| u4 | 2.6 | 0.10 | u |
| u5 | 30.0 | 0.10 | u |
| u6 | 20.5 | 0.80 | u |

For 27 individuals, the urinary excretion rates of two steroid metabolites and the underlying type of syndrome are recorded

Outlier with Tetrahydrocortisone = 53.8

We should further investigate this observation and remove it only if we are convinced that it was recorded by mistake and we cannot recover the correct values

# ANOVA

- **The objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone.**
  - We denote by:
    - Y the urinary excretion rate of Tetrahydrocortisone
    - X the Type variable, where X = 1 for Type=a, X = 2 for Type=b, X = 3 for Type=c, and X = 4 for Type=u.

- Also, our objective could be defined as investigating whether the mean of the response variable Y differs for different values (levels) of the factor X.

# ANOVA

- Denote the individual observations as:
  - yij : the urinary excretion rate of Tetrahydrocortisone of the
  - Jth: individual in group i.
- The total number of observations is n=27, and the number of observations in each group is n1 = 6, n2 = 10, n3 = 5, and n4 = 6.
- The overall (for all groups) observed sample mean for the response variable is $\bar{y}$ = 10.46.

```
32 # Summarizing data
33 ```{R}
34 summary(Cushings)
35 ```
```

```
 Tetrahydrocortisone Pregnanetriol    Type
 Min.   : 1.90       Min.   : 0.040   a: 6
 1st Qu.: 3.85       1st Qu.: 0.400   b:10
 Median : 7.80       Median : 1.000   c: 5
 Mean   :10.46       Mean   : 2.242   u: 6
 3rd Qu.:12.95       3rd Qu.: 2.800
 Max.   :53.80       Max.   :11.700
```

# ANOVA

- We also find the group specific means, which are
  - $\bar{y}1 = 3.0$
  - $\bar{y}2 = 8.2$
  - $\bar{y}3 = 19.7$
  - $\bar{y}4 = 14.0$

```
# Summarizing data
```{R}
Cuchings_db <- Cushings %>%
select(Tetrahydrocortisone, Pregnanetriol, Type) %>%
group_by(Type) %>%
summarise(Tetrahydrocortisone = mean(Tetrahydrocortisone), Pregnanetriol = mean(Pregnanetriol))

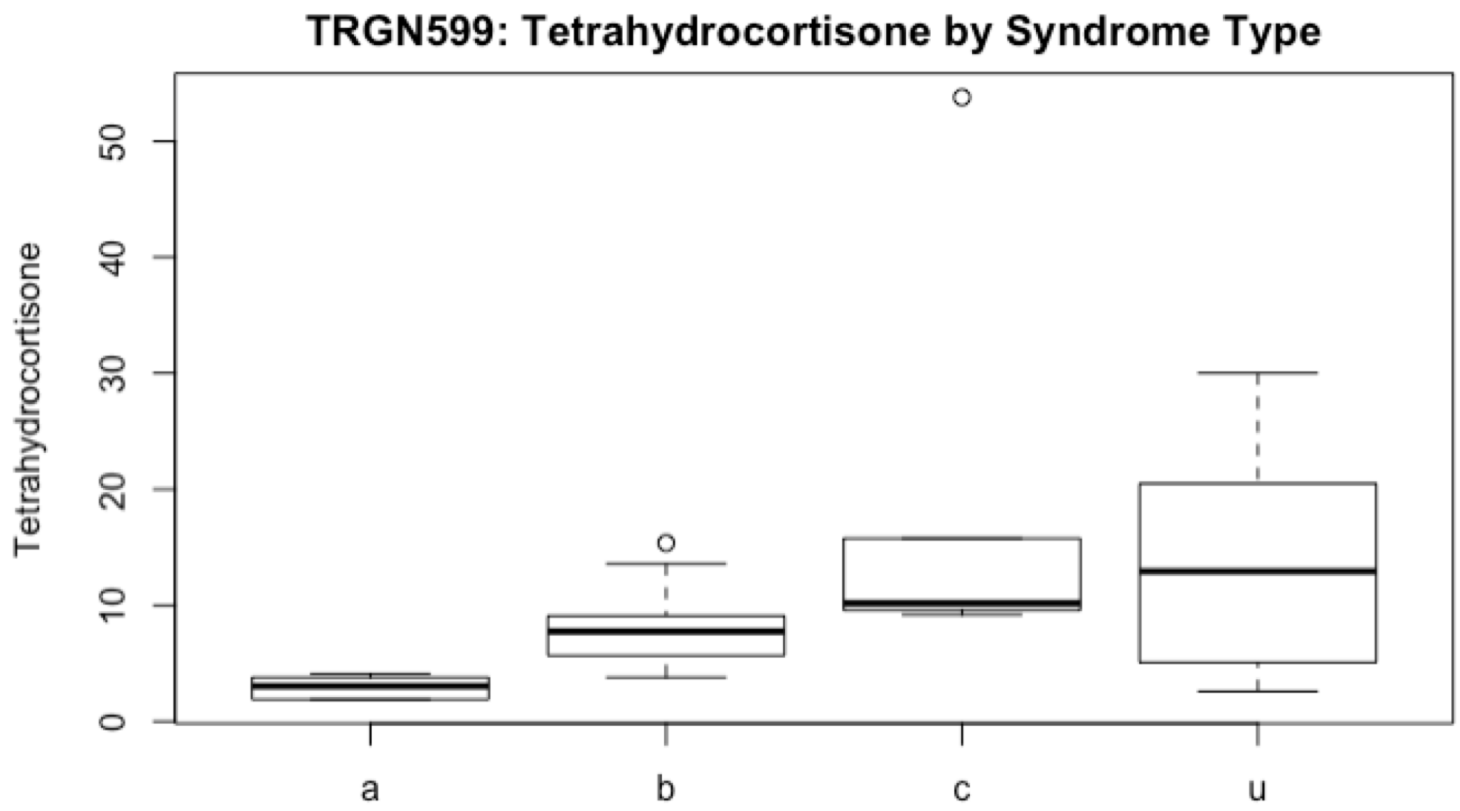
Cuchings_db
```
```

| Type <fctr> | Tetrahydrocortisone <dbl> | Pregnanetriol <dbl> |
|---|---|---|
| a | 2.966667 | 2.44 |
| b | 8.180000 | 1.12 |
| c | 19.720000 | 5.50 |
| u | 14.016667 | 1.20 |

# ANOVA

- Plot the data

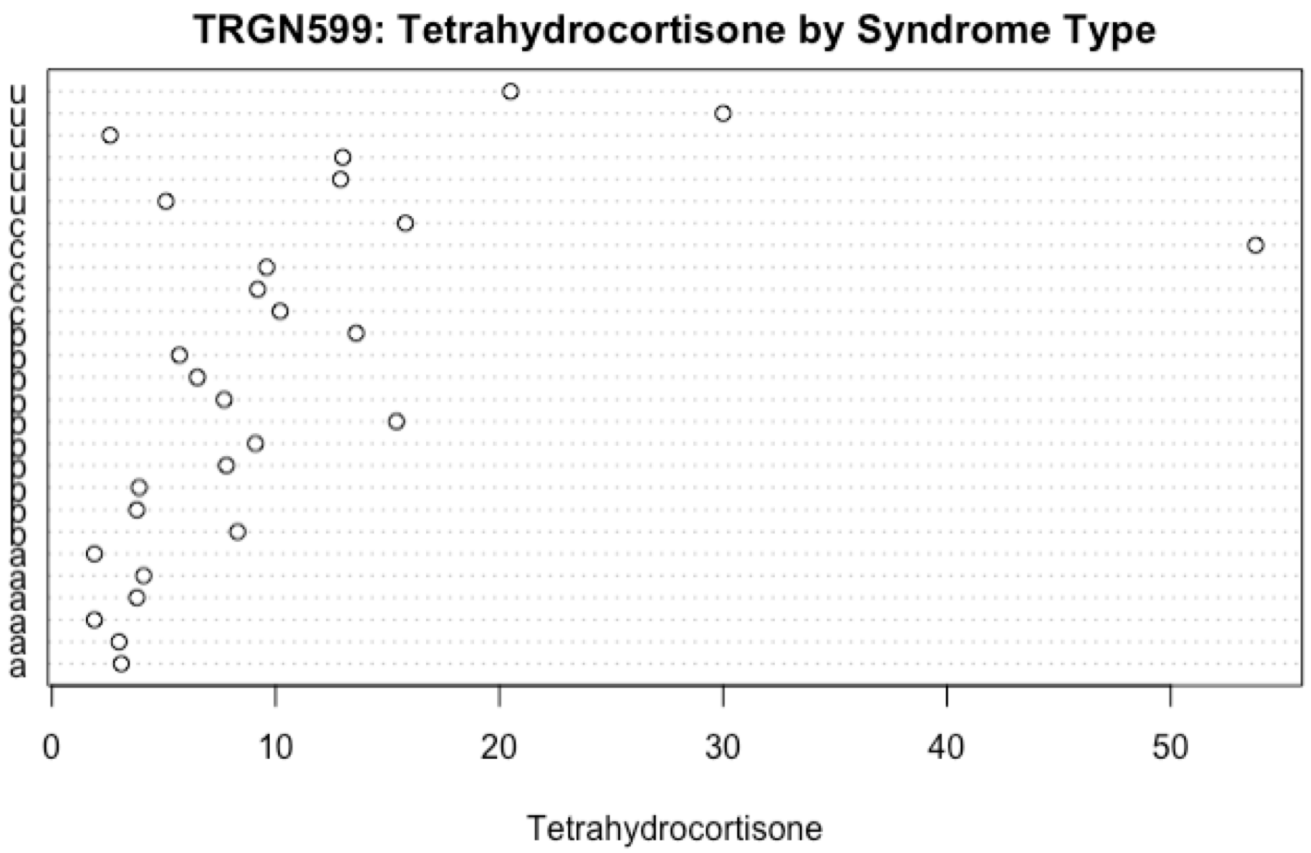
```
46 # Plotting data
47 ```{R}
48 plot(Cushings$Type, Cushings$Tetrahydrocortisone, main="TRGN599: Tetrahydrocortisone by Syndrome Type", xlab="Type", ylab="Tetrahydrocortisone")
49 ```
```



# ANOVA

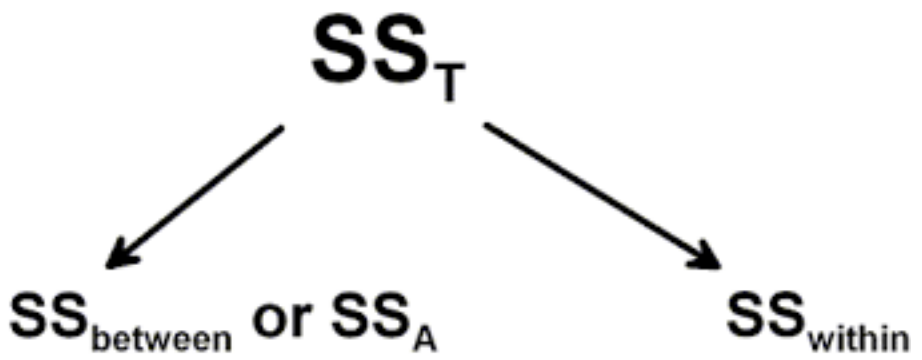
- Plot the data where observations are represented by points.

```
51 # Plotting data
52 ```{R}
53
54 dotchart(as.numeric(Cushings$Tetrahydrocortisone), Cushings$Type, main="TRGN599: Tetrahydrocortisone by Syndrome Type", xlab="Tetrahydrocortisone")
55
56 ```
```



# ANOVA

- Overall:
  - Across the four groups (between groups), there appears to be considerable variation in the group means.
  - Likewise, within groups, there are different degrees of variation of the observations from their specific mean.
  - Both sources of variation contribute to the total variation of the observations around the overall mean.



# ANOVA

- To find SSB:
  - We first find the squared difference between each group mean and the overall mean.
  - In order to account for varying sample sizes, the squared distance is then multiplied by the number of observations in that group, ni.
  - Therefore, groups with more observations are weighted more heavily.
  - The sum of these squared and weighted differences is the between-groups variation.

In general, the **between-groups variation** is denoted as $SS_B$ and calculated by

$$SS_B = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2, \qquad (1)$$

where $k$ is the number of groups (here, 4).

# ANOVA

- To find SSW:
  - We first calculate the sum of squared deviations of each observation (i.e., the point) from the group mean (i.e., the short horizontal line) for each group separately.
  - Then we add the results over all groups.

The **within-groups variation** is denoted as $SS_W$ and calculated by

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

# ANOVA

- To find SS:
  - We find the sum of squared distances of each observation to the overall average (i.e., the dashed line).
  - It seems intuitive and can be shown that the total variation SS is equal to the sum of the between-groups variation SSB and the within-groups variation SSW:

We measure the **total variation** in $Y$ by

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 .$$

# ANOVA

- In other words:
  - The total variation can be attributed partly to the variation within groups and partly to the variation between groups.
  - SSB is interpreted as the part of total variation SS that is associated with (and can be explained by) the factor variable X (e.g., syndrome type).
  - SSW is regarded as the unexplained part of total variation and is regarded as random.

$$SS = SS_B + SS_W$$

- In our example, if Tetrahydrocortisone does not depend on the type of syndrome
  - We expect the group-specific averages to be the same.
  - On the other hand, if there is a substantial difference in Tetrahydrocortisone depending on the type of syndrome, then we would expect the variation between groups to be large.
  - We examine the amount of between-groups variation relative to the variation within groups (which occurs randomly).

# ANOVA

- Let us denote the overall population mean of Y as μ and group-specific population means as μ1,...,μ4.

- Then we can express the null hypothesis of no difference in means between the groups as:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu.$$

- The alternative hypothesis HA is that at least one of the group means μi is different from the overall mean μ.

# ANOVA

- The process of evaluating hypotheses regarding the group means of multiple populations is called the Analysis of Variance (ANOVA).
- Since we are only considering one factor only, this method is specifically called one- way ANOVA.
- The test statistic for examining the null hypothesis is called F -statistic
  - More specifically, ANOVA F -statistic and is defined as:

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

# ANOVA

- Where:
  - n is the total sample size
  - k is the number of groups
- The numerator is called the **mean square for groups**
- The denominator is called the **mean square error** (MSE)

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

# ANOVA

- Note that the test statistic is based on:
  - Comparing the variation between groups (which is explained by the factor) and
  - The variation within groups (which is unexplained and random).

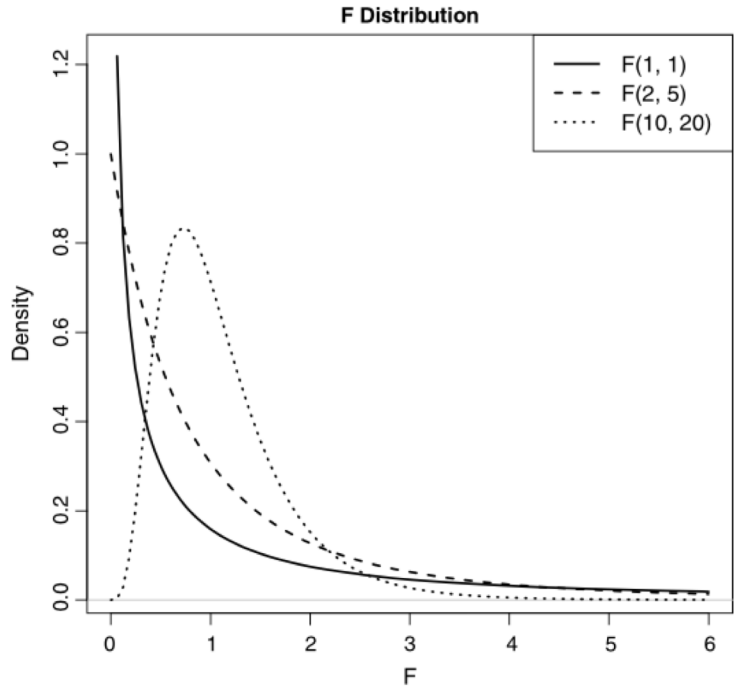
- When the group means are substantially different, and their variation is relatively large compared to the random variations within groups, the value of the F statistic becomes large.

- We denote the observed value of the F -statistic as f . If the null hypothesis is true, then the test statistic F has an F -distribution.

$$F = \frac{MS_{group}}{MS_{error}}$$

# ANOVA

- The F-distribution, which is a continuous probability distribution, is very important for hypothesis testing.
- It is specified by two parameters, df1 and df2, and is denoted as F(df1,df2).
- We refer to df1 and df2 as the *numerator degrees of freedom* and *denominator degrees of freedom*, respectively.
- Both parameters must be positive.
- Figure below shows the pdf of F-distribution for different values of df1 and df2.

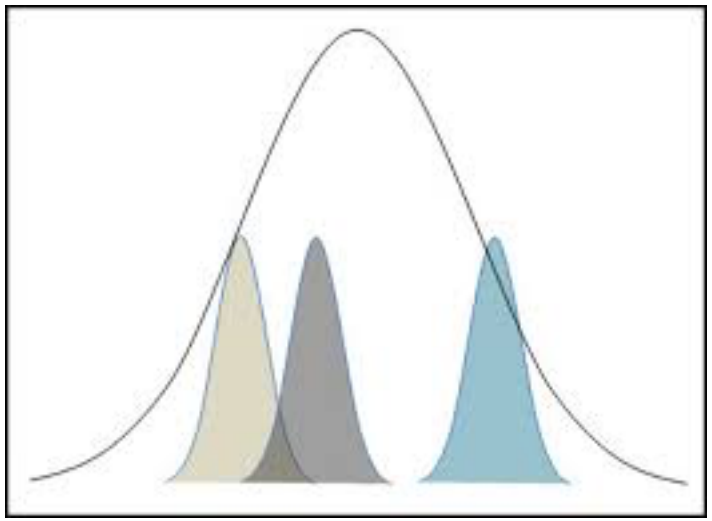



Comparing the plots of the probability density function for an F -distribution with various degrees of freedom.
The *solid line* represents the pdf of F (1, 1), the *dashed line* represents the pdf of F (2, 5), and the *dotted line* represents the pdf of F (10, 20)

df is the the number of independent values or quantities which can be assigned to a statistical distribution.

# ANOVA

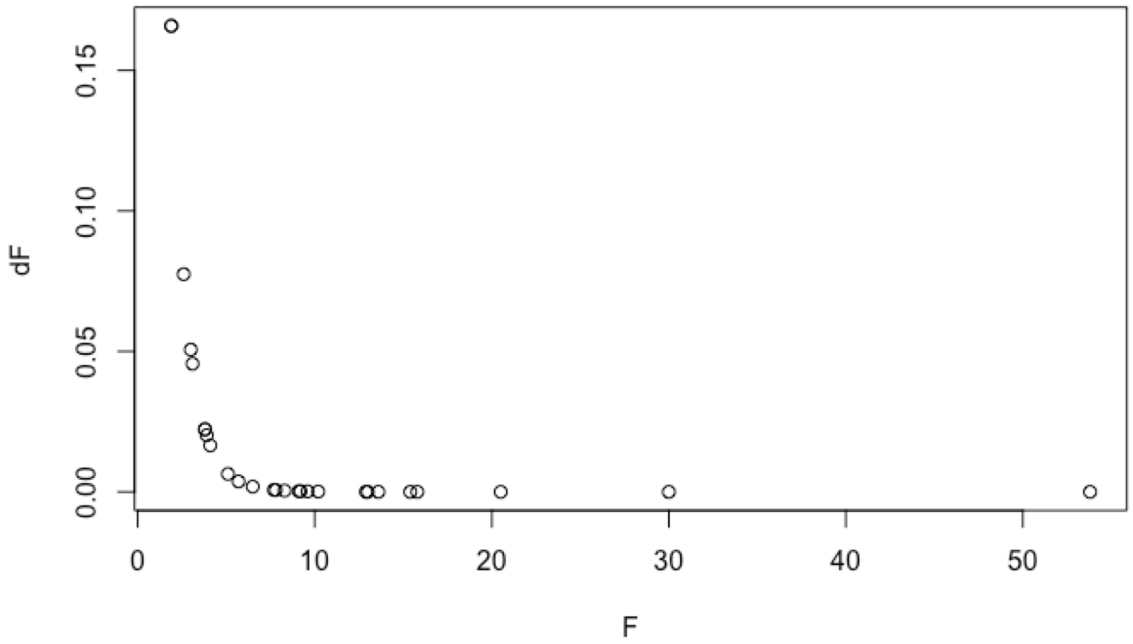
- For the one-way ANOVA, the F-statistic has F(df1 = k − 1, df2 = n − k) distribution under the null hypothesis (i.e., assuming that the null hypothesis is true).
- Here, df1 = k − 1, which is the number of groups minus 1, is called the numerator degrees of freedom, and df2 = n − k, which is the sample size minus the number of groups, is called the denominator degrees of freedom.
- The underlying assumption here is that the observations in each group have a normal distribution.
- Additionally, the underlying assumption of the ANOVA method discussed here is that all groups have the same population variance, σ2, which is unknown.



# ANOVA

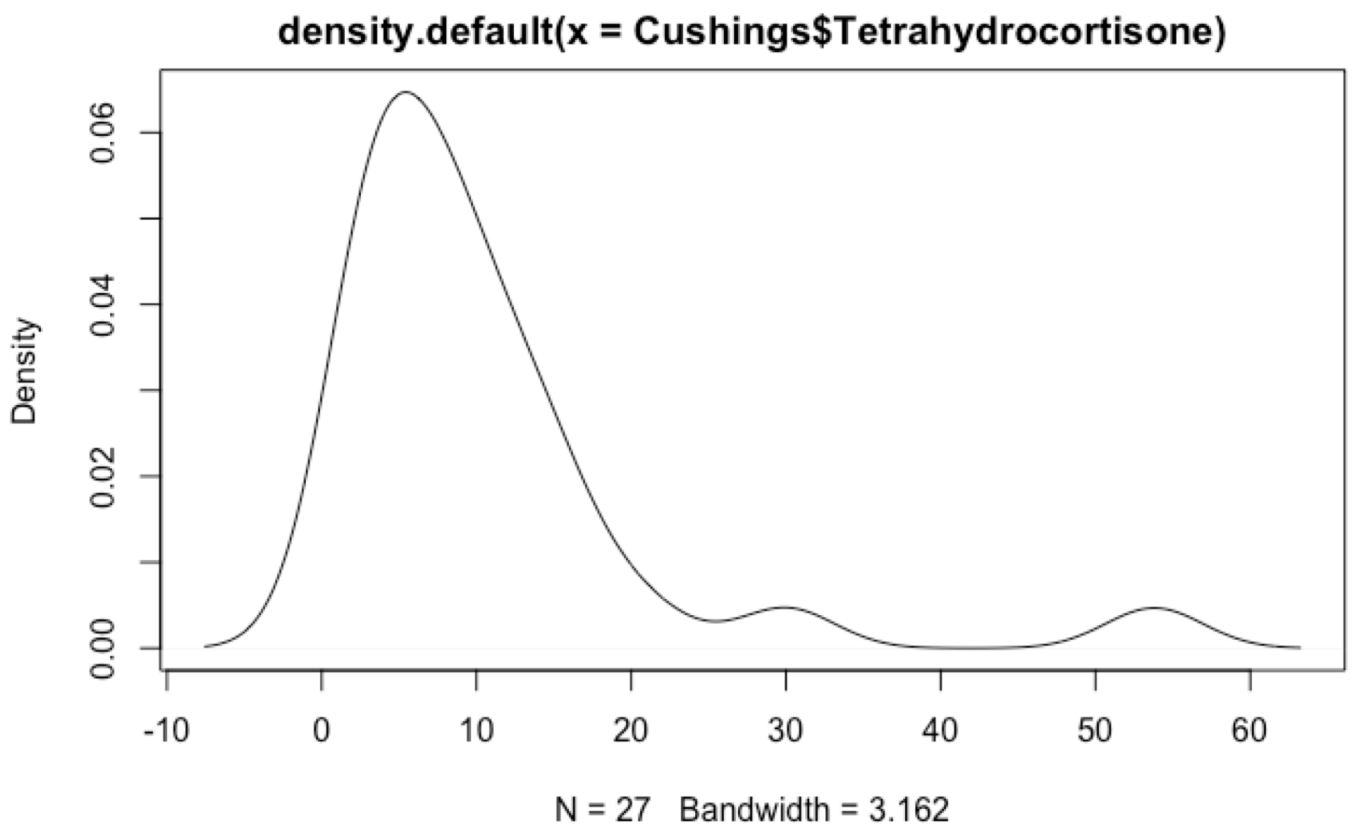
- For the above example, the degrees of freedom parameters are df1 = 4 − 1 = 3 and df2 = 27 − 4 = 23.
- Try plotting the F(3,23) distribution
- Using R: Set the Numerator degrees of freedom to 3 and the Denominator degrees of freedom to 23:
  - df(x, df1, df2)

```
119 ```{R}
120 F <- Cushings$Tetrahydrocortisone
121 dF <- df(x=F, df1=3, df2=23)
122 plot(F, dF)
123 ```
```



# ANOVA

```
131 ```{R}
132 d <- density(Cushings$Tetrahydrocortisone) # returns the density data
133 plot(d) # plots the results
134 ```
```



# ANOVA

- We can use the function aov() to perform ANOVA directly.
- For this, we specify the response and factor variables using the same formula notation we used for the t -test: response ~ factor.
- Here, the response variable is Tetrahydrocortisone, and the factor is Type.
- The output of ANOVA is assigned to the object TRGN599_anova_TT.

````
61 # We can use the function aov() to perform ANOVA directly.
62 # Here the response variable is Tetrahydrocortisone, and the factor is Type:
63 ```{R}
64
65 library(MASS)
66 data(Cushings)
67 TRGN599_anova_TT <- aov(Tetrahydrocortisone ~ Type, data = Cushings)
68
69 ```
````

# ANOVA Table

- We can create the ANOVA table by applying the summary() function to this object:

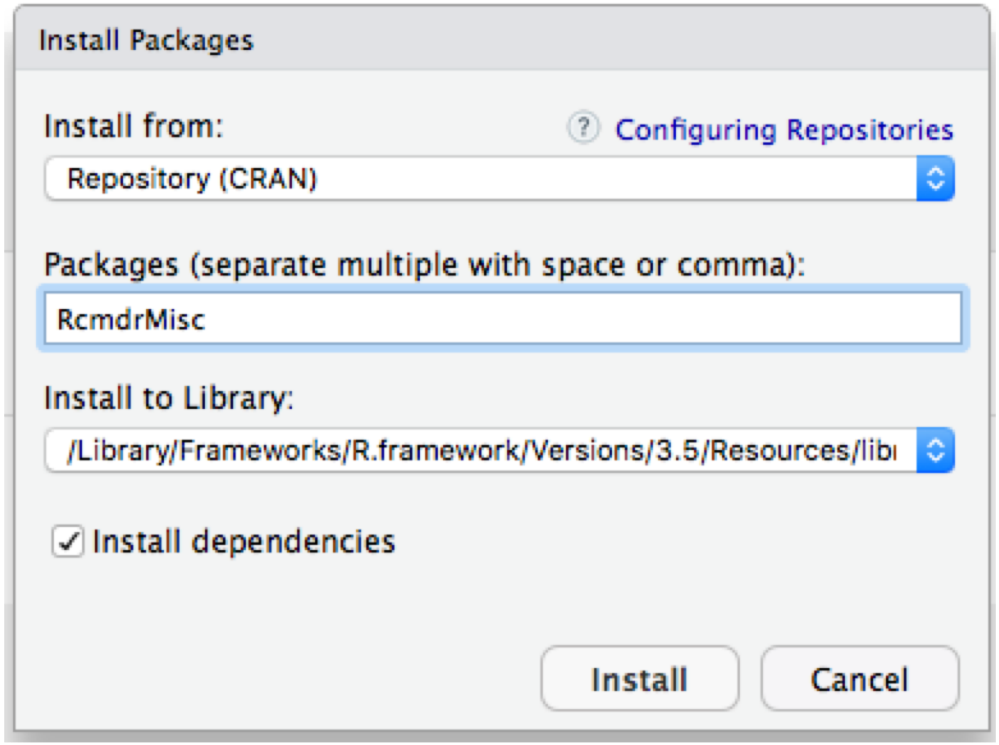
````
71 # Summarizing
72 ```{R}
73
74 summary(TRGN599_anova_TT)
75
76 ```
````

```
            Df Sum Sq Mean Sq F value Pr(>F)
Type         3  893.5  297.84   3.226 0.0412 *
Residuals   23 2123.6   92.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table resulting from the hypothesis test regarding the mean Tetrahydrocortisone of various syndrome types. Specifically, the null hypothesis is that there is no difference in the group means.
The first row corresponds to the factor, and the second to the residuals

# ANOVA

- Using the Package ‘RcmdrMisc’:



# ANOVA Table

```
# Summarizing
```{R}

summary(TRGN599_anova_TT)

```
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Type         3  893.5  297.84   3.226 0.0412 *
Residuals   23 2123.6   92.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

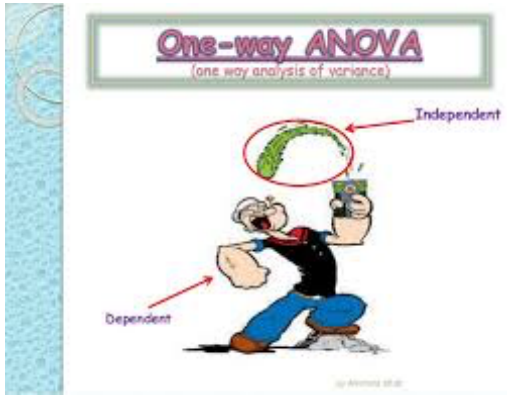
```{R}
install.packages("RcmdrMisc")
library(RcmdrMisc)
numSummary(Cushings$Tetrahydrocortisone, groups=Cushings$Type, statistics = c("mean", "sd"))

```
```

```
       mean         sd data:n
a  2.966667  0.9244818      6
b  8.180000  3.7891072     10
c 19.720000 19.2388149      5
u 14.016667 10.0958242      6
```
```

# ANOVA

- Interpretation:
  - The results of the analysis of variance are presented as a table called the *ANOVA table*.
  - The first row of this table is for the group variable (Type) and shows the explained part of the total variation (i.e., between groups).
  - The last row (Residuals) shows the unexplained part (i.e., random variations within groups) of the total variation in the data.
  - The first column shows the degrees of freedom (Df), which are k − 1 = 3 and n − k = 23, respectively.
  - The values of the second column, labeled Sum Sq, are the between-groups and within-groups variations: SSB = 893.5 and SSW = 2123.6.
  - The observed value of F -statistic is f = 3.2 given under the column labeled F value.
  - The resulting p-value is then 0.04.
  - Below the ANOVA table, R-Commander provides the group-specific means, the group-specific standard deviations, and the number of observations in each group, ni.



# ANOVA

- Assumptions:
  - When the group means are very different from each other, the between-groups variation *SS*B is high.
    - As a result, the F -statistic is large.
  - Therefore, large values of the F -statistic are considered as extreme if the null hypothesis is true.
  - Therefore, large values of F provide strong evidence against the null hypothesis.
  - To find the observed significance level pobs, we find the probability of values as or more extreme than the observed value of the test statistic, f .

# ANOVA

- In case you have the F-statistic from an experiment:
    - We can use the F -distribution to obtain the corresponding p-value
        - by calculating the upper tail probability of f .

```
# After we obtain f , the observed value of the F -statistic, we can use the F -distribution to obtain the corresponding p-value by calculating the upper tail probability of f.
# The upper tail probability of 3.2 is obtained as follows:
# Note that we need to set the lower.tail to “FALSE” to obtain the upper tail probability.
```{R}
pf(3.2, df1 = 3, df2 = 23, lower.tail = FALSE)
```

[1] 0.04226148
```

# ANOVA

- Two-Way ANOVA
  - Consider the study by Bailey (1953) to investigate the inheritance of maternal influences on the growth of the rat.
  - In this study, rat litters were separated from their natural mothers, and they were nurtured by foster mothers.
  - Mothers and litters can have four different genotypes: A, B, I, and J.

```
136 # Printing Genotype database
137 ```{R}
138 genotype
139 ```
```

| | Litter | Mother | Wt |
|---|---|---|---|
| 1 | A | A | 61.5 |
| 2 | A | A | 68.2 |
| 3 | A | A | 64.0 |
| 4 | A | A | 65.0 |
| 5 | A | A | 59.7 |
| 6 | A | B | 55.0 |
| 7 | A | B | 42.0 |
| 8 | A | B | 60.2 |
| 9 | A | I | 52.5 |
| 10 | A | I | 61.8 |
| 11 | A | I | 49.5 |
| 12 | A | I | 52.7 |
| 13 | A | J | 42.0 |
| 14 | A | J | 54.0 |
| 15 | A | J | 61.0 |
| 16 | A | J | 48.2 |
| 17 | A | J | 39.6 |
| 18 | B | A | 60.3 |
| 19 | B | A | 51.7 |
| 20 | B | A | 49.3 |
| 21 | B | A | 48.0 |
| 22 | B | B | 50.8 |
| 23 | B | B | 64.7 |
| 24 | B | B | 61.7 |
| 25 | B | B | 64.0 |
| 26 | B | B | 62.0 |
| 27 | B | I | 56.5 |
| 28 | B | I | 59.0 |
| 29 | B | I | 47.2 |
| 30 | B | I | 53.0 |
| 31 | B | J | 51.3 |
| 32 | B | J | 40.5 |
| 33 | I | A | 37.0 |
| 34 | I | A | 36.3 |
| 35 | I | A | 68.0 |
| 36 | I | B | 56.3 |
| 37 | I | B | 69.8 |
| 38 | I | B | 67.0 |
| 39 | I | I | 39.7 |
| 40 | I | I | 46.0 |
| 41 | I | I | 61.3 |
| 42 | I | I | 55.3 |
| 43 | I | I | 55.7 |
| 44 | I | J | 50.0 |
| 45 | I | J | 43.8 |
| 46 | I | J | 54.5 |
| 47 | J | A | 59.0 |
| 48 | J | A | 57.4 |
| 49 | J | A | 54.0 |
| 50 | J | A | 47.0 |
| 51 | J | B | 59.5 |
| 52 | J | B | 52.8 |
| 53 | J | B | 56.0 |
| 54 | J | I | 45.2 |
| 55 | J | I | 57.0 |
| 56 | J | I | 61.4 |
| 57 | J | J | 44.8 |
| 58 | J | J | 51.5 |
| 59 | J | J | 53.0 |
| 60 | J | J | 42.0 |
| 61 | J | J | 54.0 |

# ANOVA

- In R:
  - Load the genotype data set from the MASS library.
  - Suppose that we want to investigate whether weight gain (Wt) of the litter (in grams) at age 28 days is related to foster mother's genotype (Mother).
    - Rat litters were separated from their natural mothers at birth and given to foster mothers.
  - For this, we could use the **one-way ANOVA** procedure to compare the means of weight gain across different groups (genotypes).
  - That is, we regard Wt as the response variable and Mother as the factor.
    - For this example, however, we might want to take into account the **genotype of rat litters (Litter) as well**.
    - The litter's genotype is itself a factor, and even though it is not the main factor in this study, it should be included in the analysis since we believe that it could influence the relationship between the main factor, mother's genotype, and the response variable, weight gain.

# ANOVA

- An ANOVA with two factors is called a two-way ANOVA.
  - In general, we can have a multi-way ANOVA by including multiple factors.
- In many two-way ANOVA procedures, one of the two factors is the main explanatory variable of interest.
- The other factor is included since it is believed to be important in the study of the relationship between the main factor and the response variable.
- This is the case for the “rat genotype” example.
- In this example, we are mainly interested in the variation of weight gain across different genotypes of mothers.
- However, we need to account for possible weight gain variation due to the genotype of the litters.

# ANOVA

- By including both factors Mother and Litter, we are dividing the total variation, SS, into three sources:
  - (1) variation explained by the mother's genotype, SSM
  - (2) variation explained by the litter's genotype, SSL
  - (3) the random variation, SSE , of weight gain not explained by either mother's genotype or litter's genotype.
  - Note that we have switched our notation from SSW to SSE .

$$SS = SS_M + SS_L + SS_E.$$

# ANOVA

- This type of two-way ANOVA is commonly used for experiments with a randomized block design.
- For these experiments, the treatment variable is the factor whose effect on the response variable is of main interest.
- The categorical variable used for blocking is the factor which is believed to be important, but its relationship with the response variable is not the focus of the experiment.

````
84
85 # For two-way ANOVA, we use also use aov().
86 # but the right side of the formula includes both factors.
87
88 ```{R}
89 library(MASS)
90 data(genotype)
91 TRGN599_anova_ML <- aov(Wt ~ Mother + Litter, data = genotype)
92 ```
````

# ANOVA

```
94 # Summarizing
95 ```{R}
96
97 summary(TRGN599_anova_ML)
98
99 ```
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
Mother        3    772  257.20   4.254 0.00905 **
Litter        3     64   21.21   0.351 0.78870
Residuals    54   3265   60.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# ANOVA

```
# To include the interaction of the two factors in the ANOVA model, we use "*" instead of "+".
```{R}

library(MASS)
data(genotype)
TRGN599_anova_MLi <- aov(Wt ~ Mother * Litter, data = genotype)

```


# Summarizing
```{R}

summary(TRGN599_anova_MLi)

```
```

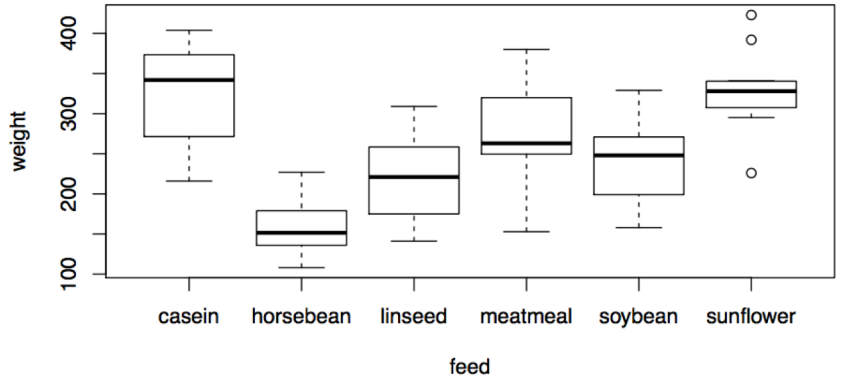
```
             Df Sum Sq Mean Sq F value  Pr(>F)
Mother        3  771.6  257.20   4.742 0.00587 **
Litter        3   63.6   21.21   0.391 0.76000
Mother:Litter 9  824.1   91.56   1.688 0.12005
Residuals    45 2440.8   54.24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

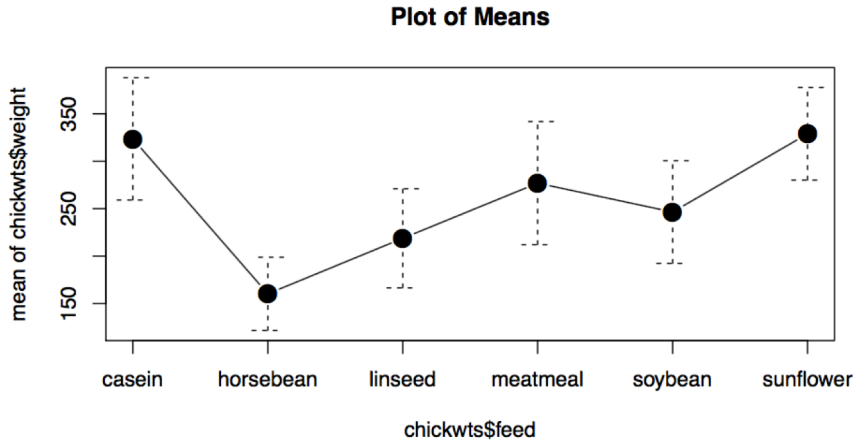
# ANOVA

- The interpretation of sum of squares, degrees of freedom, F -statistic, and p-value is similar to one-way ANOVA.
- In this example, SSM = 771.6, SSL = 63.63, and SSM×L = 824.07.
- Based on these results, only the relationship between mother's genotype and weight gain is statistically significant at 0.05 level (pobs = 0.006).
- The interaction effect (shown as Litter: Mother) and the main effect of litter's genotype are not statistically significant at 0.05 level.

# ANOVA - Interpretation

- EXAMPLE 1:
  - We would like to investigate the effectiveness of various feed supplements (feed) on the growth rate (weight) of chickens.
  - The following plot of means visualize the difference between feed types.
    - Comment on your findings and appropriateness of your assumptions.



# ANOVA - Interpretation

- EXAMPLES:
  - We would like to investigate the effectiveness of various feed supplements (feed) on the growth rate (weight) of chickens.
  - ANOVA examine the effectiveness of feed supplements.
    - Comment on your findings and appropriateness of your assumptions.

```
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER:
The observed value of *F*-statistic is $f = 15.37$, and the corresponding *p*-value is quite small.
Therefore, we can reject the null hypothesis and conclude that based on this experiment,
various feed supplements have quite different effects on the growth rate and
the relationship between the two variables (feed type and weight) is statistically significant.

# ANOVA - Interpretation

- EXAMPLE 2:
  - We are reviewing a study comparing ascorbic acid (one form of vitamin C) content between two different cultivars (c39 and c52) of cabbage.
  - In this data set, the two different cultivars were planted on three different dates, denoted as d16, d20, or d21.
  - The variable Data is a factor that specifies the planting date for each cabbage.
  - The results of the two-way ANOVA to evaluate the relationship between the vitamin C content and cultivars while controlling for the effect of planting dates are the following.

```
Anova Table (Type II tests)

Response: VitC
           Sum Sq Df F value    Pr(>F)
Cult       2496.2  1 54.1095 1.089e-09 ***
Date        909.3  2  9.8555 0.0002245 ***
Cult:Date   144.3  2  1.5640 0.2186275
Residuals  2491.1 54
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# ANOVA

- EXAMPLE 2:

  - Interpret the results of the two-way ANOVA to evaluate the relationship between the vitamin C content and cultivars while controlling for the effect of planting dates are the following.

ANSWER:

Using a two-way ANOVA, where Cult and Date are the two factors, we find that Cult and date are significantly associated with VitC (*p*-value = 1.089×10−09 & 0.0002).