Chapter 24: Analysis of Variance

In Chapters 19 and 20, we examined inferential methods for comparing the means of two populations. We will now study **analysis of variance** (**ANOVA**), which gives us a method of comparing the means of three or more populations.

One-way ANOVA is the simplest kind of ANOVA. It is used when we have one categorical explanatory variable (called a **factor**) which defines three or more populations and a quantitative response variable on which we are comparing the populations. (groups)

Example: Suppose we want to compare the mean battery life of four brands of laptop batteries. The explanatory variable would record the brand of the batteries and the response variable would record the amount of time that the batteries last.

Suppose we have k populations whose means (with respect to the response variable) are $\mu_1, \mu_2, \ldots, \mu_k$ and whose standard deviations are $\sigma_1, \sigma_2, \ldots, \sigma_k$. An ANOVA requires that we assume that the populations all have the same standard deviation, that is, $\sigma_1 = \sigma_2 = \cdots = \sigma_k$. (Denoted σ).

For an ANOVA, the null hypothesis is that the means are all equal to each other, that is,

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

To test the null hypothesis, we analyze the variation in the sample data.

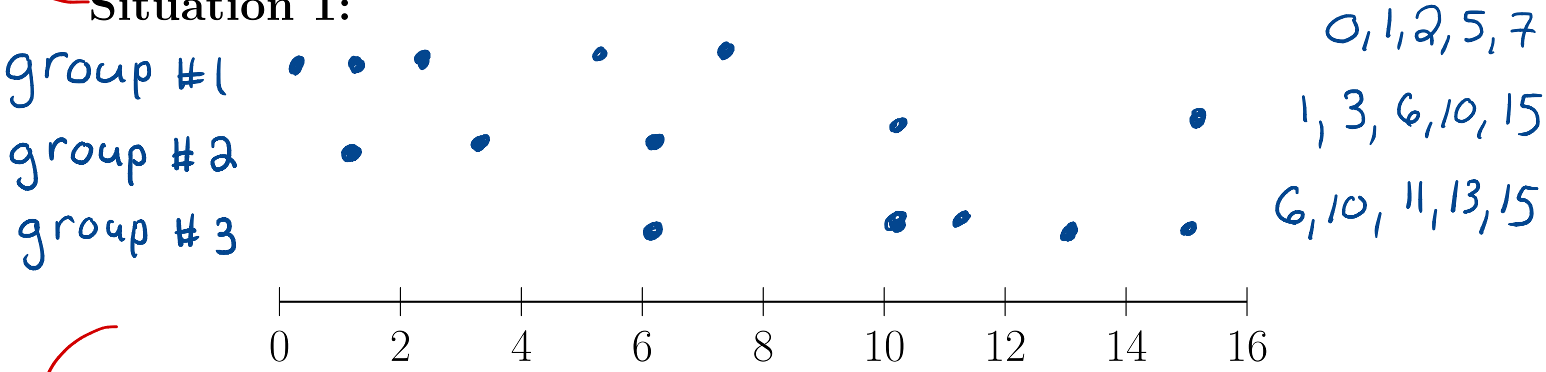
Variance measures the spread of data. ANOVA is concerned with two types of variability:

- Variability **within** groups: variance within groups is the "average" of the variances of the groups.
- Variability **between** groups: variance between groups is the "variance" of the groups means.

natural sampling variability could account for differences in means

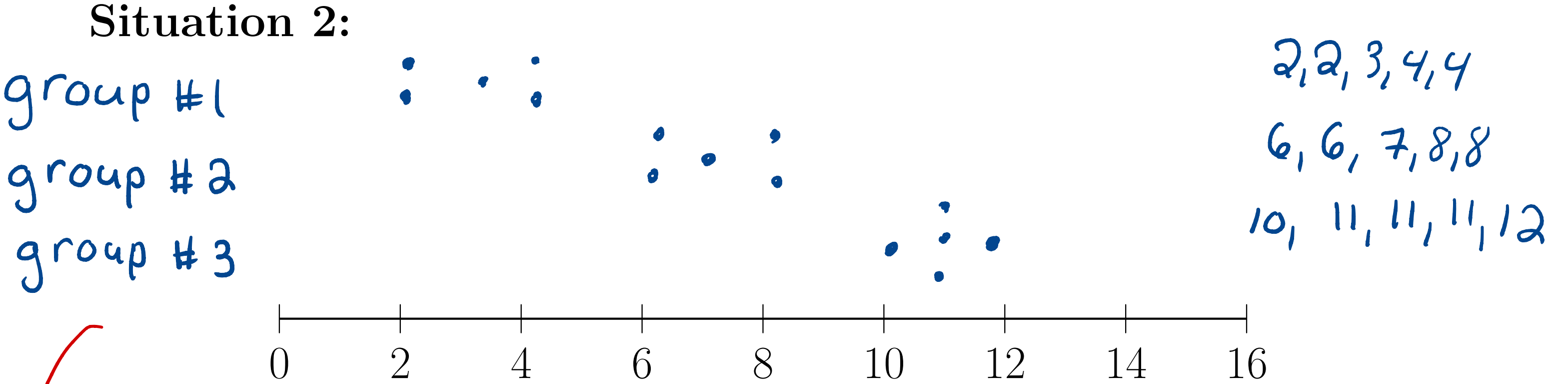
Example: Let's consider two situations:

Situation 1:



could have occured from pops. with same mean

Situation 2:



differences between means large relative to variability within each sample

In both cases, $\bar{y}_1 = 3$, $\bar{y}_2 = 7$, and $\bar{y}_3 = 11$:

- the variance between groups is the same.
- the variance within the groups is smaller in situation 2.

To find evidence against the null hypothesis, ANOVA uses the test statistic

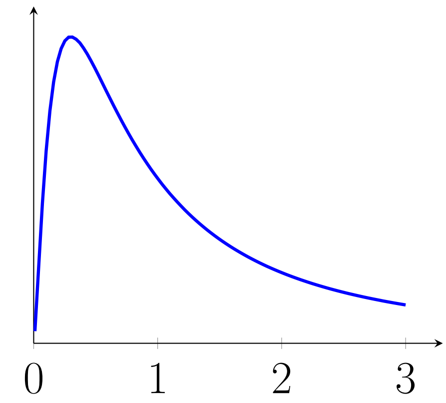
$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

The larger the F-score, the more evidence we have against H_0. The F-score increases when:

- the variance between groups is large
- the variance within groups is small

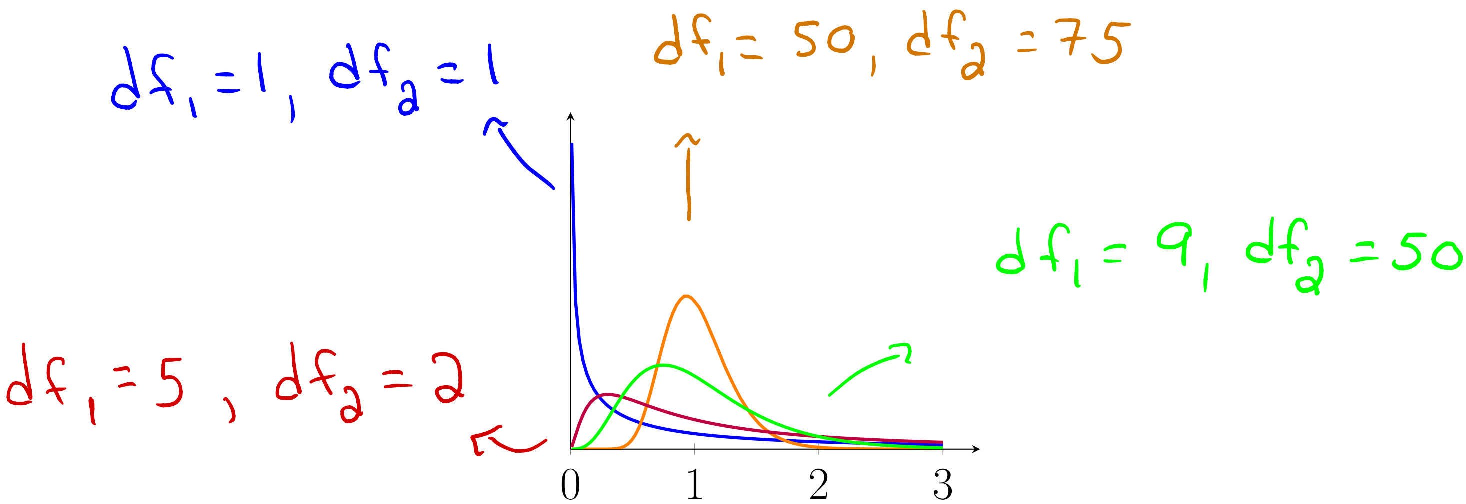
The F-Distribution

ANOVA procedures rely on a collection of models called the F-models. An F-distribution has an associated F-curve



which has the following properties:

- The total area under an F-curve is 1.
- An F-curve starts at 0 on the horizontal axis and extends indefinitely to the right, getting very close to, but never touching the horizontal axis.
- An F-curve is right-skewed.
- There are infinitely many F-distributions/F-curves. Each one is parameterized by **two** degrees of freedom:
  - degrees of freedom for the numerator, denoted df_1.
  - degrees of freedom for the denominator, denoted df_2.



How do we Measure the Variance Between and Within Groups?

Sample 1	Sample 2	Sample 3
8	14	10
7	16	12
9	12	16
13	17	15
10	11	12

$k = 3, \quad N = 15$

$n_1 = 5, \; n_2 = 5, \; n_3 = 5$

$\bar{y}_1 = 9.4, \; \bar{y}_2 = 14, \; \bar{y}_3 = 13$

$\bar{y} = 12.13$

$y_{12} = 7, \quad y_{25} = 11, \quad y_{34} = 15$

Let k be the number of populations/groups.
Let n_i be the number of subjects sampled in group i.
Let $\bar{y}_i$ be the sample mean of group .
Let s_i be the sample standard deviation of group i.
Let $\bar{y}$ be the mean of all observations sampled (grand mean).
Let y_{ij} be the j^{th} observation in the i^{th} group.
Let N be the total number of observations in all groups.
$(N = n_1 + n_2 + \ldots n_k)$

Treatment Sum of Squares SS_T:

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

$$= \underbrace{(9.4 - 12.13)^2 + (9.4 - 12.13)^2 + \cdots + (9.4 - 12.13)^2}_{n_1 \text{ times}}$$

$$+ \underbrace{(14 - 12.13)^2 + \cdots + (14 - 12.13)^2}_{n_2 \text{ times}}$$

$$+ \underbrace{(13 - 12.13)^2 + \cdots + (13 - 12.13)^2}_{n_3 \text{ times}}$$

$$= 5(9.4 - 12.13)^2 + 5(14 - 12.13)^2 + 5(13 - 12.13)^2$$

Error Sum of Squares SS_E:

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

$= (8 - 9.4)^2 + (7 - 9.4)^2 + \cdots + (10 - 9.4)^2$

$+ (14 - 14)^2 + \cdots + (11 - 14)^2$

$+ (10 - 13)^2 + \cdots + (12 - 13)^2$

$y_{ij} - \bar{y}_i$ is the residual of y_{ij}

Total Sum of Squares SS_{Total}:

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

$= (8 - 12.13)^2 + (7 - 12.13)^2 + \cdots + (10 - 12.13)^2$

$+ (14 - 12.13)^2 + \cdots + (11 - 12.13)^2$

$+ (10 - 12.13)^2 + \cdots + (12 - 12.13)^2$

One-way ANOVA Identity:

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

$$SS_{\text{Total}}=SS_T+SS_E$$

The variance between the groups is measured by the **Treatment Mean Square** (or the **Between Mean Square**), denoted MS_T, and the variance within the groups is measured by the **Error Mean Square** (or the **Within Mean Square**), denoted MS_E. These are computed by dividing SS_T and SS_E by their respective degrees of freedom:

$$MS_T=\frac{SS_T}{k-1} \qquad\qquad MS_E=\frac{SS_E}{N-k}$$

ANOVA Table:

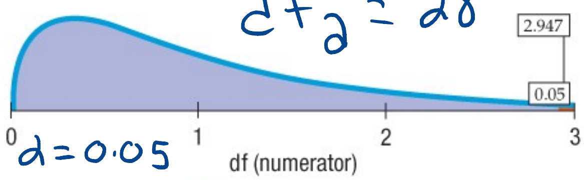
Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Treatment (Between)	$df_1=k-1$	SS_T	$MS_T=\dfrac{SS_T}{k-1}$	$\dfrac{MS_T}{MS_E}$	From Software
Error (Within)	$df_2=N-k$	SS_E	$MS_E=\dfrac{SS_E}{N-k}$		
Total	df_1+df_2 $=N-1$	SS_{Total} $=SS_T+SS_E$			

To use the F-table to find P-values: → always upper-tailed tests

- Go to the table for the given α level.
- Locate the entry corresponding to the appropriate numbers of degrees of freedom.
- If your test statistic F_0 is larger than that entry, your P-value is less than α.

$df_1 = 3$

$df_2 = 28$



$\alpha = 0.05$

Ex: for $F_0 = 6.82$,

$6.82 > 2.947$

$\Rightarrow$ P-value < 0.05

df (denominator)

	1	2	3	4	5	6	7
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313

for $F_0 = 1.38$

$1.38 < 2.947$

$\Rightarrow$ P-value > 0.05

<u>Note:</u> Recall: we are assuming that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ and we denote this common variance by σ^2 :

- MS_E is the **pooled estimate of the variance** σ^2, so we write it as s_p^2. $s_p = \sqrt{MS_E}$
- When we assume H_0 is true, MS_T is also an estimate for σ^2.

Thus, if H_0 is true, then $F_0 = \dfrac{MS_T}{MS_E}$ should be close to 1. If H_A is true, then MS_T should be larger and $F_0 > 1$.

Example: Complete the following ANOVA table for an experiment involving five groups with eight subjects in each group:

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Treatment (Between)	4	160	$\frac{160}{4} = 40$	$\frac{40}{5} = 8$	0.0001
Error (Within)	35	175	$\frac{175}{35} = 5$		
Total	39	335			

$5 \times 8 = 40$ observations

Example: Complete the following ANOVA table:

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Treatment (Between)	3	2.124	0.708	0.75	0.535
Error (Within)	20	18.88	0.944		
Total	23				

$$F = \frac{MS_T}{MS_E} \Rightarrow MS_E = \frac{MS_T}{F} = \frac{0.708}{0.75} = 0.944$$

$$MS_E = \frac{SS_E}{df_2} \Rightarrow SS_E = df_2 \, MS_E = 20(0.944) = 18.88$$

$$MS_T = \frac{SS_T}{df_1} \Rightarrow df_1 = \frac{SS_T}{MS_T} = \frac{2.124}{0.708} = 3$$

Assumptions and Conditions:

a) **Independence Assumption:**

   i) **Independent Responses Assumption:** the data sampled should come from independently responding individuals.

   ii) **Independent Groups Assumption:** the k samples, one from each population, must be independent of each other.

   iii) **Randomization Condition:** for each sample, the data should be drawn independently, using random selection from the population or come from a completely randomized experiment.

b) **Equal Variance Assumption:** The populations all have the same variance (and so also the same standard deviation).

   To check this assumption, we can check:

   i) that the largest of the sample standard deviation is less than twice the smallest of the sample standard deviations:

$$\frac{\text{largest } s}{\text{smallest } s} < 2$$

   ii) **Similar Spreads Condition:** Look at side-by-side boxplots of the groups to see whether they have the same spread.

c) **Normal Population Assumption:** The k populations are all Normally distributed. To check this, we check:

   Nearly Normal Condition: Look at the boxplots for each group to check for skewness and outliers. Examine a histogram or Normal probability plot of all of the residuals together.

   all $y_{ij} - \bar{y}_i$

A One-way ANOVA F-test has five steps:

1. **Assumptions/Conditions:**
   - The k populations all have the same standard deviation, but the value is unknown.
   - We have k random samples, one from each of the k populations.
   - Within each sample, we have independent responses from the individuals.
   - The k samples are independent of each other.
   - The k populations are all Normally distributed.

2. **Hypotheses:**

$$H_0 : \quad \mu_1 = \mu_2 = \ldots = \mu_k$$

at least 1 μ_i is different ←

$$H_A : \quad \text{the } \mu_i \text{ are not all equal}$$

3. **Test Statistic:**

$$F_0 = \frac{MS_T}{MS_E} = \frac{\dfrac{SS_T}{k-1}}{\dfrac{SS_E}{N-k}}$$

   Assuming H_0 is true, F_0 follows an F-distribution with $df_1 = k - 1$ and $df_2 = N - k$, where k is the number of samples/populations and N is the total number of observations.

4. ***P*-value:** The P-value $= P(F > F_0)$ can be computed using software or the F-table, where $df_1 = k - 1$ and $df_2 = N - k$.

5. **Conclusion:** Given a significance level α,
   - if P - value $\leq \alpha$, we reject H_0 at level α
   - if P- value $> \alpha$, we do not reject H_0 at level α

Example: Musical Preference and Reckless Behaviour

Group	1	2	3	4
Musical Preference	Acoustic/Pop	Mainstream Rock	Hard Rock	Heavy Metal
	2	3	3	4
	3	2	4	3
	4	1	3	4
	1	2	1	3
	3	3	2	3
	3	4	1	3
	3	3	4	3
	3	2	2	3
	2	4	2	2
	2	4	2	4
	1	4	3	4
	3	4	3	5
	2	2	4	4
	2	3	3	5
	2	2	3	3
	3	2	2	4
	2	2	3	5
	2	3	4	4
	3	1	2	2
	4	3	4	3
$\bar{y}$	2.5	2.7	2.75	3.55
s	0.827	0.979	0.967	0.887

smallest largest

Researchers wanted to determine whether adolescents who preferred certain types of music are more likely to engage in reckless behaviours, such as speeding. Independently chosen random samples (all of size 20) were taken from four groups of adolescents with different musical preferences: (1) acoustic/pop, (2) mainstream rock, (3) hard rock, (4) heavy metal.

Each adolescent was asked how many times they had driven over 130 kph in the last year. Their responses are given in the table above.

a) Complete the Following ANOVA table: $N = 20 \times 4 = 80$

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Treatment (Between)	3	12.85	4.283	5.09	0.0029
Error (Within)	76	63.9	0.841		
Total	79	76.75			

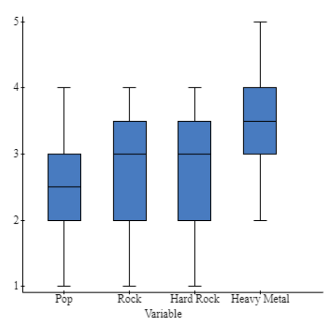
b) Carry out a one-way ANOVA F-test to determine if the mean number of times driving over 130 kph varies with musical preference. Use $\alpha = 0.01$.

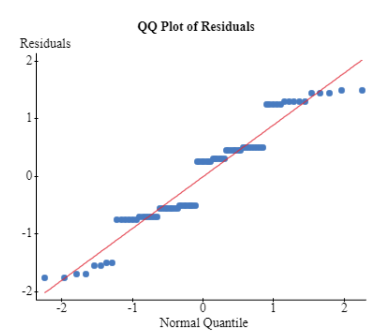
1. **Assumptions/Conditions:**

- assume populations have same standard deviation

$$\frac{\text{largest } s}{\text{smallest } s} = \frac{0.979}{0.827} = 1.18 < 2$$

- data gathered randomly
- assume independence within and between samples
- assume all populations are approximately normally distributed.





2. **Hypotheses:**

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

H_A: the μ_i are not all equal

3. **Test Statistic:**

$F_0 = 5.09$

4. ***P*-value:**

P-value = 0.0029

5. **Conclusion:**

Since P-value < 0.01, we reject H_0 at the 0.01 significance level, that is, there is enough statistical evidence to conclude that the mean number of times driving over 130 kph is the not the same for all four musical preference groups.