Review: We know how to test the null hypothesis that one population mean is equal (or greater than or equal, or less than or equal) to a real value and how to test the null hypothesis that population mean 1 is equal (or greater than or equal, or less than or equal to population mean 2. The analysis of variance tests the null hypothesis that three or more population means are equal in what is called an omnibus test. It does this by comparing the variance of the sample means to the variance inside the samples using an F test, which we have already used to compare two variances.

A. Introduction
ANOVA is a statistical procedure for determining whether three or more sample means were drawn from populations with equal means. In everyday language, ANOVA tests the null hypothesis that the population means (estimated by the sample means) are all equal. If this null hypothesis is rejected, then we conclude that the population means are not all equal. A more precise formulation of the null and alternative hypotheses for comparing k means is:

\[ H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \]
\[ H_1 : \text{at least one pair of means is different or not equal} \]
\[ \mu_i \neq \mu_j, \quad i, j = 1, \ldots, k; \quad i \neq j \]

Because it tests for differences between multiple pairs of means in one test, it is called an omnibus test.

Motivation and need for ANOVA: You may ask, “Why don’t we just test if the means are equal with t-tests? After all, we know about t-tests.” The problem with t-tests on even three sample means is that the t-tests are not independent because you use each sample mean more than once. Consider a simple example.

Example 1: A chemical engineer must decide between 3 catalysts in terms of reaction completion times. The chemical engineer takes three samples, each with 25 completion times. If she tries to make inferences about the population means using t-tests on the three sample means, she will use each sample mean twice: \( \bar{x}_1 \) vs. \( \bar{x}_2 \), \( \bar{x}_1 \) vs. \( \bar{x}_3 \), and \( \bar{x}_2 \) vs. \( \bar{x}_3 \). Because she used each sample mean in two t-tests the t-tests are not independent and, therefore she can not know the probability of a Type I error. In this case a Type I error is rejecting one of the three null hypotheses when it is true. Clearly, the situation becomes worse when we compare more than three means. Can you think of a way to remedy this situation?

Problem: Inflation of Type I error with multiple t-tests. One remedy is to use ANOVA. The null hypothesis of ANOVA assumes that all means are equal. This is equivalent to stating that all samples were taken from the same population. The ANOVA test eliminates the problem of multiple t-tests on the same sample
means by testing all the means at once to see if any of the means are different. If we reject the null hypothesis, all we know is that at least one pair of means is different. We cannot tell which pair of means is different until we run a multiple comparison test which is “follow-up” test we run only when we reject the null hypothesis in ANOVA.

**Completely Randomized Designs (also called One-Way ANOVA or a single-factor ANOVA)**

**Example 2:** A mining engineer wishes to compare five strains of bacteria with respect to mean copper yield (lb/ton).

**Example 3:** A chemical engineer assesses the effect of four pressure settings on product yield (grams/liter).

<table>
<thead>
<tr>
<th>Sample Batch</th>
<th>Low (1)</th>
<th>Moderate (2)</th>
<th>Strong (3)</th>
<th>High (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>30</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>29</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>30</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>30</td>
<td>33</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>28</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>29</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>28</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>30</td>
<td>32</td>
<td>30</td>
</tr>
</tbody>
</table>

B. **Assumptions of single-factor ANOVA:**

1. The dependent variable is **normally distributed.**
2. Each population has the **same or equal variance** (homogeneity of variance). The factor has equal variances at all levels of the factor.
3. A non-mathematical assumption is that the samples **represent independent random samples.**

Note that, taken together, the assumptions cause the null hypothesis to be that the samples were randomly selected from a single population which is \( N(\mu, \sigma^2) \)

C. **The hypothesis-testing procedure for ANOVA.**

1. We formulate a null hypothesis and an appropriate alternative **hypothesis.** Consider the single-factor ANOVA model for a single factor with 4 levels, such as the four different pressure settings in Example B. Each population mean may be represented as:
\[
\mu_j = \mu + B_j, \quad \text{where} \quad \mu = \frac{\sum_{j=1}^{4} \mu_j}{4}
\]
is the population mean of the product yield at the four product settings and \( B_j \) is the effect of pressure setting \( j \). Note that \( B_j \) may be positive if a pressure setting increases product yield over the mean setting or negative if a product setting decreases product yield. Under the null hypothesis, \( B_j = 0 \) for each \( j \).

The single-factor ANOVA null hypothesis for our example, \( H_0 \): \( \mu_1 = \mu_2 = \mu_3 = \mu_4 \), includes the assumption that the variances are equal; **so, the null hypothesis is equivalent to assuming that all samples were drawn from a one population distributed as \( N(\mu, \sigma^2) \)**. Under this assumption we know that the standard error of the sample means is \( \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \). Consequently, we know how the sample means should be distributed if the samples are drawn from a single population or identical populations.

2. **We specify the probability of Type I error**, \( P(\text{Type I error}) = \alpha \). \( P(\text{Type I error}) \) is usually set at 0.05 or 0.01.

3. **Based on the sampling distribution of an appropriate statistic, we construct a criterion for testing the null hypothesis against the given alternative.**

   The F-statistic is used to compare the variability displayed by the \( k \) sample means (the among-sample treatment variance) to the variability displayed within the \( k \) samples (the within-sample error variance). The form of the test or observed F-statistic is

   \[
   F = \frac{\text{Mean Square for treatments}}{\text{Mean Square for error}} = \frac{MS_{\text{treatment}}}{MS_{\text{error}}} = \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{n-k}^2} = F_{k-1, n-k, \alpha}.
   \]

   The distribution of the F-statistic is generated from the null hypothesis which assumes that all samples were taken from equivalent populations. Note that the numerator of the F statistic estimates the variance of the sample mean \( \sigma_{\bar{X}}^2 \). The denominator estimates the variance of X inside the samples divided by the sample size: \( \frac{\sigma^2}{n} \). So, finally, the numerator and the denominator estimate the variance of the mean \( \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \) using two different sample statistics. Therefore, if the null hypothesis is true, the F value should vary randomly around the value 1. If the null hypothesis is NOT TRUE, the numerator will not be a good estimate of the variance of the mean and will be larger than 1. So, if the "variability among the sample means" is greater than the "variability within the samples", then F will tend to be larger than the critical value and the null hypothesis is rejected. In that
case we conclude that there is at least one pair of population means that are not equal.

Note: if we compare only 2 means with the $F$ statistic, the values of the $F$ and $t$ are related by the equation $F = t^2$.

**Criterion or decision rule:** For the one-factor ANOVA, the degrees of freedom for the numerator of the $F$ statistic $v_1 = k - 1$ and the degrees of freedom for the denominator $v_2 = n - k$, where $n$ is the sum of all sample sizes. The F-test is always a one-sided test. We reject the null hypothesis only if the obtained $F$ value is too large. This means that the differences between the sample means are too large for all the sample means to represent the single mean of one population.

4. **We calculate from the data the value of the (obtained) $F$ statistic.**

**ANOVA Table.** An ANOVA table is used to calculate an F statistic. Here is a summary of relevant notation.

- $k =$ number of groups (samples from distinct populations)
- $n_i =$ sample size of $i$th group
- $N =$ sum of the $k$ sample sizes (total number of observations)
- $\bar{x}_i =$ mean of $i$th sample
- $\bar{x} =$ grand mean or mean of all observations
- $s_i =$ standard deviation of $i$th sample

$$v_1 = df_{treatment} = k-1 \quad = \text{degrees of freedom among sample means}$$
$$v_2 = df_{error} = n-k \quad = \text{degrees of freedom within samples}$$

$$SS_{treatment} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \quad \text{(treatment sum of squares among sample means)}$$

$$SS_{error} = \sum_{i=1}^{k} (n_i-1)s_i^2 \quad \text{(error sum of squares within samples)}$$

If you already have the sample variance, the following formula is much easier to use.

$$SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = SS_{treatment} + SS_{error}$$

$$MS_{treatment} = SS_{treatment} / (k-1) = \text{mean square among sample means}$$
$$MS_{error} = SS_{error} / (n-k) = \text{mean square within samples}$$

$$F = MS_{among} / MS_{within}$$

The p-value is the probability of an $F$ as large or larger than the one obtained.

5. **We decide to reject the null hypothesis or we fail to reject it.** Decision:

We reject or FTR the null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ at the specified $\alpha$ level by using the decision rule.
**Example** from the (now nonfunctioning) site: [https://onlinecourses.science.psu.edu/stat500/node/63](https://onlinecourses.science.psu.edu/stat500/node/63)

Twenty young pigs were randomly assigned to 4 groups, each of which is characterized by a different diet. The table gives each pig's weight after 10 months on the diet.

<table>
<thead>
<tr>
<th>Group</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60.80</td>
</tr>
<tr>
<td>1</td>
<td>57.10</td>
</tr>
<tr>
<td>1</td>
<td>65.00</td>
</tr>
<tr>
<td>1</td>
<td>58.70</td>
</tr>
<tr>
<td>1</td>
<td>61.80</td>
</tr>
<tr>
<td>2</td>
<td>68.30</td>
</tr>
<tr>
<td>2</td>
<td>67.70</td>
</tr>
<tr>
<td>2</td>
<td>74.00</td>
</tr>
<tr>
<td>2</td>
<td>66.30</td>
</tr>
<tr>
<td>2</td>
<td>69.90</td>
</tr>
<tr>
<td>3</td>
<td>102.60</td>
</tr>
<tr>
<td>3</td>
<td>102.20</td>
</tr>
<tr>
<td>3</td>
<td>100.50</td>
</tr>
<tr>
<td>3</td>
<td>97.50</td>
</tr>
<tr>
<td>3</td>
<td>98.90</td>
</tr>
<tr>
<td>4</td>
<td>87.90</td>
</tr>
<tr>
<td>4</td>
<td>84.70</td>
</tr>
<tr>
<td>4</td>
<td>83.20</td>
</tr>
<tr>
<td>4</td>
<td>85.80</td>
</tr>
<tr>
<td>4</td>
<td>90.30</td>
</tr>
</tbody>
</table>

**Hypothesis testing for this experiment.**

1. \( H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \)
2. \( H_1: \mu_i \neq \mu_j, i \neq j, i, j = 1,2,3,4 \)
3. \( F_{k-1, N-k, \alpha} = F_{3,16, 0.05} = 3.25 \). We reject \( H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \) when \( F_{\text{obtained}} > 3.25 \)
4. We calculate the \( F_{\text{obtained}} \).
Hypothesis testing: ANOVA: More than Two Means

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of freedom</th>
<th>Mean Square</th>
<th>( F_{obtained} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>SS_{treatment}</td>
<td>k-1</td>
<td>SS_{treatment}/(k-1)</td>
<td>MS_{treatment} / MS_{error}</td>
</tr>
<tr>
<td>Error</td>
<td>SS_{error}</td>
<td>N-k</td>
<td>SS_{error} /(N-k)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SS_{total}</td>
<td>N-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( k \) = number of groups (samples from distinct populations)
\( n_i \) = sample size of \( i \)th group
\( N \) = sum of the \( k \) sample sizes (total number of observations)
\( \bar{x}_i \) = mean of \( i \)th sample
\( \bar{x} \) = grand mean or mean of all observations
\( s_i \) = standard deviation of \( i \)th sample

\( v_1 = df_{treatment} = k-1 \) = degrees of freedom among sample means
\( v_2 = df_{error} = N - k \) = degrees of freedom within samples

\[ SS_{treatment} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \] (treatment sum of squares among sample means)

\[ SS_{error} = \sum_{i=1}^{k} (n_i-1)s_i^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \] (error sum of squares within samples)

\[ SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = SS_{treatment} + SS_{error} \]

\[ MS_{treatment} = SS_{treatment} / (k-1) \] mean square among or between sample means
\[ MS_{error} = SS_{error} / (N-k) \] mean square within samples

\[ F = MS_{treatment} / MS_{error} \]

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>60.6800</td>
<td>3.02771</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>69.2400</td>
<td>2.95770</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>100.3400</td>
<td>2.16402</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>86.3800</td>
<td>2.78155</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>79.1600</td>
<td>15.93495</td>
</tr>
</tbody>
</table>
D. Multiple comparisons

Which pairs of means are different? We use ANOVA to avoid inflating the Type I error rate by running t-tests on the different pairs of means. So, we cannot use regular t-tests to find out which set of means is different.

Bonferroni method of following up rejection of the Null Hypothesis in ANOVA

The Bonferroni obtained statistic is \[ \frac{|\bar{x}_i - \bar{x}_j|}{\sqrt{MSS_{error} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} \text{, } i \neq j \text{, } i, j = 1, 2, ..., k \text{.} \] The critical t is obtained by using a two-tailed t test at the level \( \alpha' = \frac{2\alpha}{k(k-1)} \). So \( \frac{\alpha'}{2} = \frac{\alpha}{k(k-1)} \) for the two-tailed test with the degrees of freedom=n-k, the degrees of freedom for the MS_{error}.

If we calculate the Bonferroni statistic for each pair of means, we find that each of the 6 possible pairs is different at the 0.05 level. Notice that the Bonferroni process protects the Type I error rate by reducing the Type I error rate for each comparison.

Calculating the 6 Bonferroni comparisons is in the homework assignment.
There are many approaches to multiple comparison tests, many of them very tailored to specific situations. The Bonferroni method is based on adjusted probabilities. If you, the experimenter, have \( k \) means to compare, then there \( \binom{k}{2} = \frac{k(k-1)}{2} \) possible two-sample t-tests. So, the Bonferroni method uses \( \frac{\alpha}{k(k-1)} \) instead of \( \frac{\alpha}{2} \) for the two-sided t-tests.

- **Example 3**: effects of four pressure settings on product yield
  1. **Hypotheses**: The null hypothesis for comparing \( k = 4 \) means is:
     
     \[ H_0 : \mu_1 = \mu_2 = \ldots = \mu_4 \]

     \[ H_1 : \text{at least one pair of means is different, } \mu_i \neq \mu_j, \quad i, j = 1, \ldots, k; \quad i \neq j \]

  2. **P(Type I error) and Test Statistic**:
     The engineer chose \( \alpha = 0.05 \).

     **Assumptions for the F statistic**: The engineer had evidence that larger samples at various pressures displayed symmetric histograms; so, the assumption of normality seemed appropriate. Note: The sample standard deviations may vary by a factor of 2 without producing invalid results; so, given the standard deviations in a table for step 4, the assumption of homogeneity of variances seems reasonable. The engineer, therefore, chose the F statistic.

     **Decision rule**: The F for \( v_1 = 4 - 1 = 3 \) and \( v_2 = 40 - 4 = 36 \) degrees of freedom for \( \alpha = 0.05 \) is 2.86. The engineer’s decision rule is to reject the null hypothesis if the obtained F > 2.86.

     (Note that I found the number 2.86 in a table which includes \( v_2 = 36 \). Using a table which present \( v_2 \) in increments of 10, the closest you could come to the correct criterion F is by finding the critical F value for \( v_1 = 3 \) and \( v_2 = 30 \). (Why not 40?) This value is 2.92. Note that this will be conservative in the sense that...
an obtained F of 2.87 will not be rejected for degrees of freedom 3 and 30, but will be (correctly) rejected for degrees of freedom 3 and 36.)

4. Calculate F value.
Note: I displayed only the first three decimal places in this example. I actually stored the standard deviations in the memory of my calculator at greater precision. Therefore, if you work through this example, you may differ from my answers beginning in the hundredths place.
Recall that the formula for the sample standard deviation is:

\[ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^{n} X_i^2 - n \overline{X}^2 \right)} \]

<table>
<thead>
<tr>
<th>Pressure</th>
<th>Mean Yield</th>
<th>N</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>28.400</td>
<td>10</td>
<td>2.2706</td>
</tr>
<tr>
<td>2.00</td>
<td>29.800</td>
<td>10</td>
<td>1.3984</td>
</tr>
<tr>
<td>3.00</td>
<td>30.700</td>
<td>10</td>
<td>2.2632</td>
</tr>
<tr>
<td>4.00</td>
<td>29.200</td>
<td>10</td>
<td>2.0976</td>
</tr>
<tr>
<td>Total</td>
<td>29.525</td>
<td>40</td>
<td>2.1362</td>
</tr>
</tbody>
</table>

\[ SS_{treatment} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{X})^2 = 10[(28.4 - 29.525)^2 + (29.8 - 29.525)^2 + (30.7 - 29.525)^2 + (29.2 - 29.525)^2] = 28.275 \]

\[ SS_{error} = \sum_{i=1}^{k} (n_i-1)s_i^2 = 9[2.2706^2 + 1.3984^2 + 2.2632^2 + 2.0976^2] = 149.700 \]

\[ MS_{treatment} = SS_{treatment} / (k-1) = 28.275/(4-1) = 9.425 \]

\[ MS_{error} = SS_{error} / (N-k) = 149.700/(40-4) = 4.158 \]

\[ F = MS_{among} / MS_{within} = 9.425/4.158 = 2.267 \]

ANOVA Table for Example 3

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>28.275</td>
<td>3</td>
<td>9.425</td>
<td>2.267</td>
</tr>
<tr>
<td>Error</td>
<td>149.700</td>
<td>36</td>
<td>4.158</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>177.975</td>
<td>39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Decision: We FTR the null hypothesis \( H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \) at \( \alpha=0.05 \) because the obtained F 2.267 is less than 2.86. Note that computer programs give you a p-value. Using the p-value, you reject the null hypothesis if the obtained p-value is less than or equal to 0.05. In our example the obtained p-value is greater than 0.05; so, we FTR the null hypothesis.
Because we FTR the null hypothesis, we do not execute any multiple comparison tests. We use multiple comparisons between sample means to analyze/explain why the null hypothesis was rejected. If the null hypothesis was not rejected, there is nothing to explain.

Example: A study is made of the effect of the curing temperature (\(^\circ C\)) on the compressive strength (MPa) of concrete.

<table>
<thead>
<tr>
<th>Temp</th>
<th>Compressive strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>0(^\circ)C</td>
<td>31, 30, 31, 30, 32, 30, 31, 30, 32, 30, 32</td>
</tr>
<tr>
<td>10(^\circ)C</td>
<td>30, 29, 31, 32, 31</td>
</tr>
<tr>
<td>20(^\circ)C</td>
<td>36, 35, 37, 35, 37</td>
</tr>
<tr>
<td>30(^\circ)C</td>
<td>38, 37, 38, 36, 38</td>
</tr>
</tbody>
</table>

1. Entering the data in R:
   ```r
   > cmpStrength=c(31, 30, 31, 30, 32, 30, 31, 30, 32, 30, 36, 35, 37, 35, 37, 38, 37, 38, 36, 38)
   > temp=c(rep("0",5),rep("10",5),rep("20",5),rep("30",5))
   > concrete=data.frame(cmpStrength,temp)
   ```

2. Boxplots (exploratory)
   ```r
   > plot(cmpStrength~temp,data=concrete)
   ```
3. Analysis of variance results

```r
> results <- aov(cmpStrength ~ temp, data = concrete)
> summary(results)

        Df Sum Sq Mean Sq F value   Pr(>F)
temp         3  185.0   61.67   64.91 3.56e-09 ***
Residuals   16   15.2    0.95
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Pairwise comparisons

Because the ANOVA rejected the null hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, we need to find out which pairs of means are different. We will use two difference pairwise comparisons: Bonferroni and TukeyHSD.

```r
> pairwise.t.test(cmpStrength, temp, p.adjust = "bonferroni")

Pairwise comparisons using t tests with pooled SD

data:  cmpStrength and temp
```

---

Note: The box plot in the image is not described in the text; it likely visualizes the distribution of `cmpStrength` grouped by `temp`.
P value adjustment method: bonferroni

> TukeyHSD(results, conf.level = 0.95)
Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = cmpStrength ~ temp, data = concrete)

$temp
diff      lwr      upr     p adj
10-0  -0.2  -1.9636511 1.563651  0.9877518
20-0   5.2   3.4363489 6.963651  0.0000015
30-0   6.6   4.8363489 8.363651  0.0000001
20-10  5.4   3.6363489 7.163651  0.0000009
30-10  6.8   5.0363489 8.563651  0.0000000
30-20  1.4  -0.3636511 3.163651  0.1466777