Northeastern University

ALY-6015 – Intermediate Analytics

Module 2 – Chi-square and ANOVA

Jean-Sebastien Provost
j.provost@northeastern.edu

January 14th, 2026

Northeastern University – Vancouver Campus

Learning Objectives

- Test a distribution for goodness of fit, using chi-square.

Testing for Goodness of Fit 1

- The **chi-square goodness-of-fit test** is used to test if a frequency distribution fits a specific given expected frequency distribution.
- Consider the following table of the observed occurrences of arrests for four kinds of crimes:

Larceny thefts	Property crimes	Drug use	Driving under the influence
38	50	28	44

- A total of 160 arrests are listed in this table. If arrests for each of the four types of crime happened with the same frequency, the table would look like this:

Larceny thefts	Property crimes	Drug use	Driving under the influence
40	40	40	40

Testing for Goodness of Fit 2

	Larceny thefts	Property crimes	Drug use	Driving under the influence
Observed	38	50	28	44
Expected	40	40	40	40

The observed frequencies will always differ from the expected frequencies due to sampling error.
Our hypotheses are:

- H_0: there is no difference in the number of arrests for each type of crime.
- H_1: there is a difference in the number of arrests for each type of crime.

Testing for Goodness of Fit 3

- To perform this hypothesis test, we use the following chi-square test statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- Degrees of freedom: Number of categories - 1.

- Test ≥ 0

- The larger the test statistic, the worse the fit, meaning this is a right-tailed test.

- In order to perform this test, the data must come from a random sample, and the expected frequency of each category must be at least 5.

Arrests for Crimes 1

Let's test our hypotheses at $\alpha = 0.05$. Remember, the hypotheses were.

- H_0: there is no difference in the number of arrests for each type of crime.
- H_1: there is a difference in the number of arrests for each type of crime.

- 4 categories
- 3 degrees of freedom
- The critical value for a right-tailed test at $\alpha = 0.05$ is 7.815

	Larceny thefts	Property crimes	Drug use	Driving under the influence
Observed	38	50	28	44
Expected	40	40	40	40

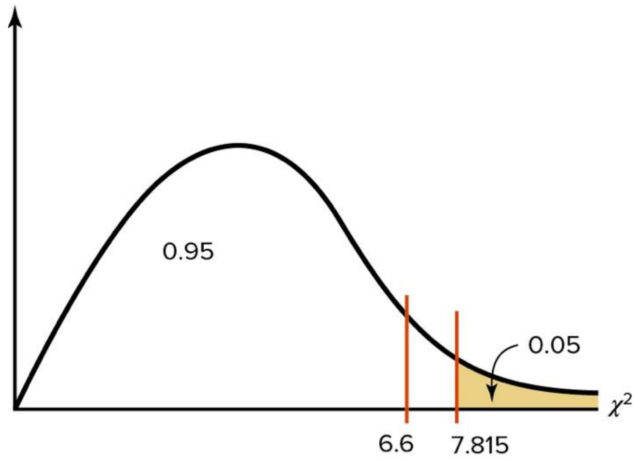
$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(38-40)^2}{40} + \frac{(50-40)^2}{40} + \frac{(28-40)^2}{40} + \frac{(44-40)^2}{40}$$

$$= 0.1 + 2.5 + 3.6 + 0.4$$

$$= 6.6$$

Arrests for Crimes 2

- 6.6 is not in the region of rejection so we should not reject the null hypothesis.



- Alternatively, we can see that 6.6 falls between 6.251 and 7.815 in the d.f. = 3 row of table A-6, so the *P*-value is between 0.05 and 0.10. So we would not reject the null hypothesis because the *P*-value is greater than the level of significance.
- We conclude that there is not enough evidence to reject the claim that there is no difference in the number of arrests for each type of crime.

Test of Normality 1

We can use the chi-square goodness-of-fit test to test for normality. We use the hypotheses.

- H_0: The variable is normally distributed.
- H_1: The variable is not normally distributed.

Next, we make a frequency table for our data values, as if we were making a histogram. For example:

Boundaries	Frequency
89.5–104.5	24
104.5–119.5	62
119.5–134.5	72
134.5–149.5	26
149.5–164.5	12
164.5–179.5	4
	Total = 200

- Next what we do is find the mean and standard deviation of the data. That way we can use the z scores of the midpoints of each class to find expected frequencies for those classes.

Test of Normality 2

- To find the mean and standard deviation of the variable, use the formula for grouped data.

Boundaries	f	X_m	$f \cdot X_m$	$f \cdot X^2_m$
89.5–104.5	24	97	2,328	225,816
104.5–119.5	62	112	6,944	777,728
119.5–134.5	72	127	9,144	1,161,288
134.5–149.5	26	142	3,692	524,264
149.5–164.5	12	157	1,884	295,788
164.5–179.5	4	172	688	118,336
	200		24,680	3,103,220

$$\bar{X} = \frac{\sum f \cdot X_m}{\sum f} = \frac{24680}{200} = 123.4$$

$$s = \sqrt{\frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}} = \sqrt{\frac{200(3,103,220) - 24,680^2}{200(199)}} = \sqrt{290} = 17.03$$

- Now we find the probability that a normal random variable with mean 123.4 and standard deviation 17.03 would fall within each class, using table A-4.

Test of Normality 3

Class	$P(X \text{ in class})$	Expected
89.5–104.5	0.1335	
104.5–119.5	0.2755	
119.5–134.5	0.3332	
134.5–149.5	0.1948	
149.5–164.5	0.0550	
164.5–179.5	0.0080	

- We used $P(X < 104.5)$ for the first probability and $P(X > 164.5)$ for the last probability.
- There were 200 total observations, so now multiply each probability by 200 to get the expected frequencies.
- N.B.: Probability for X in class is a z-test:
  - Z = upper bound – X_bar / s (ex.: Z = (104.5 – 123.4) / 17.03 = -1.11 => P = 0.1335
  - Look at the probability associated to the z value

Test of Normality 3

Class	$P(X$ in class)	Expected
89.5–104.5	0.1335	
104.5–119.5	0.2755	
119.5–134.5	0.3332	
134.5–149.5	0.1948	
149.5–164.5	0.0550	
164.5–179.5	0.0080	

- Z = upper bound – X_bar / s (ex.: Z = (104.5 – 123.4) / 17.03 = -1.11
  - Z(<-1.11) => P = 0.1335
- Z = (119.5 – 123.4) / 17.03 = -0.23
  - -1.11 < z < -0.23 => 0.4090 - 0.1335 = 0.2755

And so on…

Test of Normality 3

Class	$P(X \text{ in class})$	Expected
89.5–104.5	0.1335	26.7
104.5–119.5	0.2755	55.1
119.5–134.5	0.3332	66.64
134.5–149.5	0.1948	38.96
149.5–164.5	0.0550	11.0
164.5–179.5	0.0080	1.6

- We used $P(X < 104.5)$ for the first probability and $P(X > 164.5)$ for the last probability.
- There were 200 total observations,
    - Expected frequencies = probability x 200
- The last class does not have an expected frequency of at least 5, so we will merge it with the class above.

Test of Normality 5

Class	Observed	Expected
89.5–104.5	24	26.7
104.5–119.5	62	55.1
119.5–134.5	72	66.64
134.5–149.5	26	38.96
149.5–179.5	16	12.6

- 5 categories
- 4 degrees of freedom
  - We lose an additional degree of freedom for every parameter we estimate. We estimated the mean and the standard deviation, so we have
    - 5 − 3 = 2 degrees of freedom.
- CV at α = 0.05 = 5.991

$$\chi^2 = \frac{(24-26.7)^2}{26.7} + \frac{(62-55.1)^2}{55.1} + \frac{(72-66.64)^2}{66.64} + \frac{(26-38.96)^2}{38.96} + \frac{(16-12.6)^2}{12.6}$$
$$= 6.797$$

- 6.797 > 5.991, so we reject the null hypothesis and conclude that the variable is not normally distributed.

Learning Objectives

- Test two variables for independence, using chi-square.
- Test proportions for homogeneity, using chi-square.

Postoperative Procedure 1

- A new postoperative procedure is administered to a number of patients at a large hospital. The researcher wants to know if the doctors feel differently from the nurses about the procedure, or do they basically feel the same way?
- A random sample of doctors and nurses results in the following data.

Group	Prefer new procedure	Prefer old procedure	No preference
Nurses	100	80	20
Doctors	50	120	30

To answer our research question, we test the hypotheses.

- H_0: The opinion is independent of the profession.
- H_1: The opinion is dependent of the profession.

Note that the question is not about whether one group prefers the procedure and the other doesn't, but whether the proportion of nurses with each opinion is the same as the proportion of doctors with that opinion.

Contingency Tables

- Suppose you take a random sample from a population,
  - Measure the value of two different random variables from each member of the sample.
  - Assume the two variables can only take a few different values.

- Make a table, where the rows correspond to the possible values of one variable, called the **row variable**, and the columns correspond to the possible values of the other variable, called the **column variable**.

- Each cell of the table contains a **cell value**, the number of members of the sample who had the value of the row variable corresponding to that row, and the value of the column variable corresponding to that column.

- For example, a 5 × 3 contingency table has 5 rows and 3 columns. The cell value $C_{2,1}$ is the value in the second row and first column.

Testing for Independence

- If we have observations of the values of two variables recorded in a contingency table, we can use the **chi-square independent test** to determine if the variables are independent or dependent.

- We use the exact same test statistic as with the goodness-of-fit test:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- The number of degrees of freedom is the product

  $(\text{number of rows minus } 1)(\text{number of columns minus } 1)$.

- For chi-square independence test
  - the data must come from a random sample,
  - Expected value in each cell must be 5 or more.
  - If the expected value in a cell is less than 5, combine categories.

- Therefore, the expected value of each cell of the contingency table is the value that would occur if the two variables used to generate the table were independent. They can be calculated using the formula.

$$\text{Expected value} = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}}$$

Postoperative Procedure 2

$$\text{d.f.} = (\text{rows} - 1)(\text{columns} - 1) = (2 - 1)(3 - 1) = 2$$

- The critical value for d.f. = 2 and $\alpha = 0.05$ is 5.991.
- To perform the hypothesis test, we must calculate the sum of each row and column.

Group	Prefer new procedure	Prefer old procedure	No preference	Total
Nurses	100	80	20	200
Doctors	50	120	30	200
Total	150	200	50	400

- We then calculate the expected values of each cell by dividing the row and column sums by the grand total.

$$E_{1,1} = \frac{(200)(150)}{400} = 75 \quad E_{1,2} = \frac{(200)(200)}{400} = 100 \quad E_{1,3} = \frac{(200)(50)}{400} = 25$$

$$E_{2,1} = \frac{(200)(150)}{400} = 75 \quad E_{2,2} = \frac{(200)(200)}{400} = 100 \quad E_{2,3} = \frac{(200)(50)}{400} = 25$$

Postoperative Procedure 3

- The contingency table, with the expected values in parentheses, is shown below.

Group	Prefer new procedure	Prefer old procedure	No preference
Nurses	100 (75)	80 (100)	20 (25)
Doctors	50 (75)	120 (100)	30 (25)

- If the group and the opinion are independent:
  - Then the proportion of each group with a given opinion is the same as the total proportion of that group in the sample.
  - For example, 150 out of 400 people prefer the new procedure. There are 200 nurses. So, if the null hypotheses were true, (150/400)(200) = 75 of the nurses would prefer the new procedure.
- The test statistic is

$$\chi^2 = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \frac{(20-25)^2}{25} + \frac{(50-75)^2}{75} + \frac{(120-100)^2}{100} + \frac{(30-25)^2}{25} = 26.667$$

- We reject the null hypothesis: The opinion is dependent of the profession

Test for Homogeneity of Proportions 1

- We can also use the chi-square test and a contingency table to perform the **test for homogeneity of proportions**.
- It claims that different populations have the same proportion of subjects with a certain attitude or characteristic.
- We take a sample from each of several different populations, measure how many members of each population do and do not have a given characteristic and use these as our observed values.
- For example, 100 people in each of four different income groups are asked if they are “very happy”. The results are in the following contingency table.

Household income	Less than $30,000	$30,000—$74,999	$75,000—$99,999	$100,000 or more	Total
Yes	24	33	38	49	144
No	76	67	62	51	256
	100	100	100	100	400

Test for Homogeneity of Proportions 2

If p_1, p_2, p_3, and p_4, are the proportions of each income group who answered "Yes" then our hypotheses are

- $H_0 : p_1 = p_2 = p_3 = p_4$.
- H_1 : At least one of the proportions is different.

If the null hypothesis is true, then the cell values could be calculated using the formula

$$\text{Expected value} = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}}$$

The resulting expected values are written in parentheses in each cell.

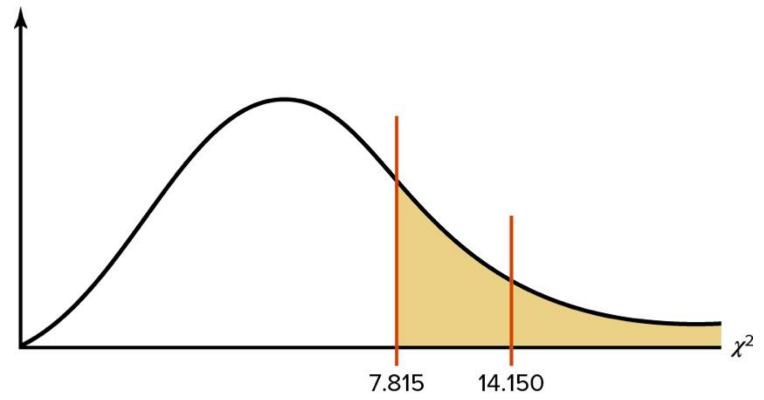
Household income	Less than \$30,000	\$30,000—\$74,999	\$75,000—\$99,999	\$100,000 or more	Total
Yes	24 (36)	33 (36)	38 (36)	49 (36)	144
No	76 (64)	67 (64)	62 (64)	51 (64)	256
	100	100	100	100	400

Test for Homogeneity of Proportions 3

- The table has 2 rows and 4 columns, so d.f. = $(1)(3)$ = 3. If we do the test at α = 0.05, our critical value is 7.815.
- The test statistic is

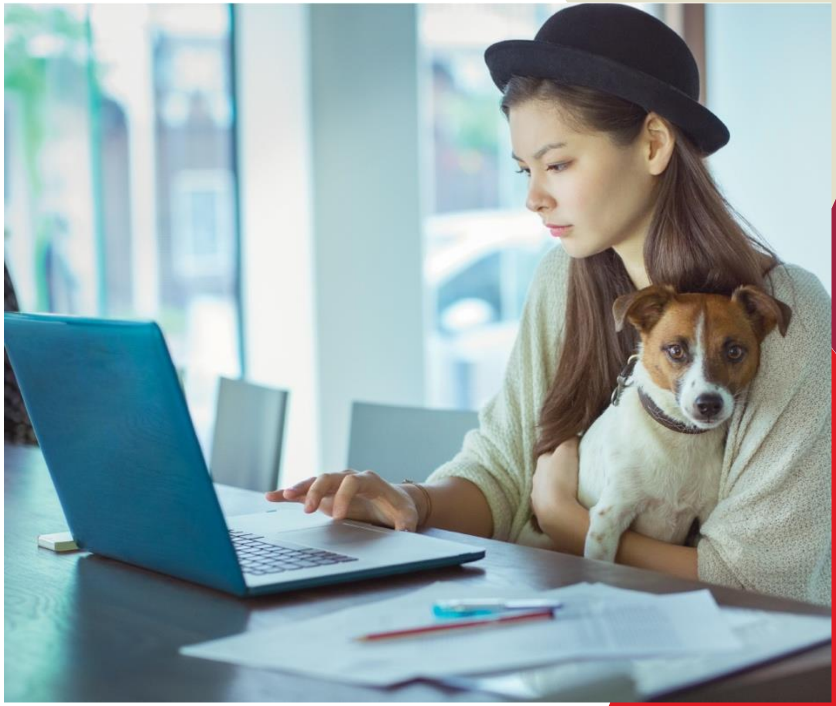
$$\chi^2 = \frac{(24-36)^2}{36} + \frac{(33-36)^2}{36} + \frac{(38-36)^2}{36} + \frac{(49-36)^2}{36} + \frac{(76-64)^2}{64} + \frac{(67-64)^2}{64} + \frac{(62-64)^2}{64} + \frac{(51-64)^2}{64} = 14.150$$

- This is in the critical region, so we reject the null hypothesis and conclude that the proportions of each income group that is “very happy” are not all equal.



Learning Objectives

- Use the one-way ANOVA technique to determine if there is a significant difference among three or more means.



ANOVA

- **Analysis of variance**, or **ANOVA**:
  - *F*-test is used to test a hypothesis about the means of three or more populations
  - **One-way analysis of variance** test is used to test the equality of three or more means using sample variances.
    - One-way analysis because the different populations are distinguished by a single variable, called a **factor**.

- We can do multiple t-tests but can be:
  - Time consuming
  - Increase the chance of making a type I error by doing more tests

- Suppose we do 10 *t*-tests at α=0.05.

$$(0.95)^{10} = 0.599$$

In other words, there is a 40% chance of making a type 1 error.

Between-Group Variance

- Even though we are comparing the means of the populations, we use two variances to calculate our test statistic.

- The **between-group variance**:
  - This is the variance of the means of the populations.

- The **grand mean** of the data values in all the samples is

$$\bar{X}_{\mathrm{GM}} = \frac{\sum X}{N}$$

- The between-group variance is then

$$s_B^2 = \frac{\sum n_i \left(\bar{X}_i - \bar{X}_{\mathrm{GM}}\right)^2}{k-1}$$

where there are k groups, n_i is the size of the ith sample, and $\bar{X}_i$ is the mean of the ith sample.

Within-Group Variance

The **within-group variance** is the weighted average of the variances s_i^2 of each of the samples.

$$s_W^2 = \frac{\sum (n_i - 1) s_i^2}{\sum (n_i - 1)}$$

The numerator of this expression is called the **sum of squares within groups (SS_W)** or the **sum of squares for the error**.
The F-test statistic is then

$$F = \frac{s_B^2}{s_W^2}$$

Degrees of freedom of numerator: $k - 1$
Degrees of freedom of denominator: $N - k$
where N is the sum of the sample sizes of the groups
and k is the number of groups
We do a right-sided test with the hypotheses:

- H_0 : All of the population means are the same.
- H_1 : At least one population mean is different from the others.

ANOVA Summary Table

The results of a one-way ANOVA can be summarized using an **ANOVA summary table**.

TABLE 12–1 Analysis of Variance Summary Table

Source	Sum of squares	d.f.	Mean square	*F*
Between groups	SS_B	$k - 1$	MS_B	
Within groups (error) Total	SS_W	$N - k$	MS_W	

- Each mean square: the sum of squares / degrees of freedom.
- Test statistic: Between-groups mean square / within-groups mean square.

In order to perform this test:

- All of the populations must be normally or approximately normally distributed.
- All of the samples are independent.
- All of the populations have the same variance.
- There is exactly one simple random sample per population.

Miles Per Gallon 1

The following table shows the fuel economy for 12 cars in three different categories: small cars, sedans, and luxury cars.

Small	**Sedans**	**Luxury**
36	43	29
44	35	25
34	30	24
35	29	
	40	

Test the claim that all three categories of car have the same mean fuel economy at $\alpha = 0.05$.
Our hypotheses are

- $H_0 : \mu_1 = \mu_2 = \mu_3$ (claim).
- H_1 : At least one mean is different from the others.

Miles Per Gallon 2

- There are 3 categories and 12 data values, so N = 12 and k = 3.
- d.f.N = $k - 1 = 2$ and d.f.D. = $N - k = 9$.
- The CV for a right-tailed F-test at $\alpha = 0.05$ is 4.26.
- The means and variances for the three samples are:

$$\text{Small cars}: \bar{X} = 37.25 \quad s^2 = 20.917$$

$$\text{Sedans}: \bar{X} = 35.4 \quad s^2 = 37.3$$

$$\text{Luxury cars}: \bar{X} = 26 \quad s^2 = 7$$

- The grand mean is

$$\bar{X}_{\text{GM}} = \frac{\sum X}{N} = \frac{404}{12} = 33.667$$

- The between-groups variance is

$$s_B^2 = \frac{\sum n_i \left(\bar{X}_i - \bar{X}_{\text{GM}}\right)^2}{k-1} = \frac{4(37.25-33.667)^2 + 5(35.4-33.667)^2 + 3(26-33.667)^2}{3-1} = 121.359$$

- The within-groups variance is

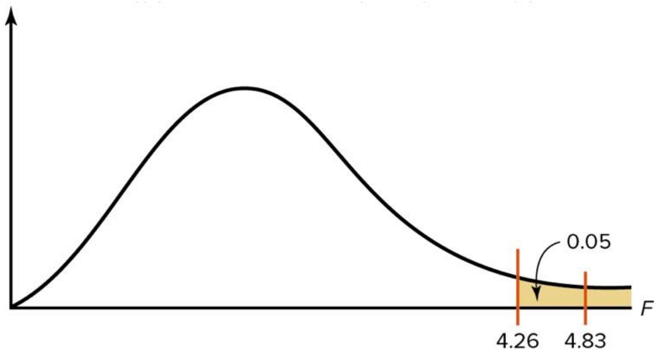
$$s_W^2 = \frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)} = \frac{(4-1)(20.917) + (5-1)(37.3) + (3-1)(7)}{(4-1)+(5-1)+(3-1)} = 25.106$$

Miles Per Gallon 3

- The F statistic is

$$F = \frac{s_B^2}{s_W^2} = \frac{121.359}{25.106} = 4.83$$

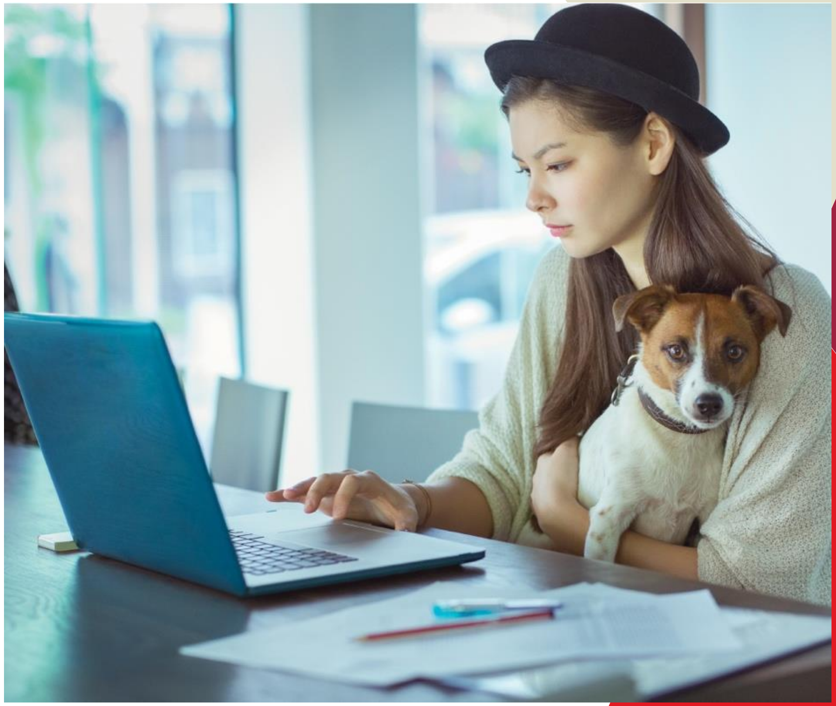
- 4.83 > 4.26, so we reject the null hypothesis.



- We conclude that at least one mean is different from the others.
    - Note that we do not have evidence about which means are different; that would require doing other tests, such at the Tukey or Scheffé test.

Learning Objectives

- Determine which means differ using:
  - Scheffé test
  - Tukey test
  - Bonferroni test



Which Mean is Different?

- The most commonly used methods are the **Scheffé test** and the **Tukey test**.

- Both tests compare the means of the various samples one pair at a time.
  - The Scheffé test uses an F test.
  - The Tukey test uses a test statistic q which has a different distribution.

- The Tukey test is more powerful, but they can only be used if all of the samples have the same sample size.

- A general rule of thumb is to use the Tukey test if all the sample sizes are the same, and to use the Scheffé test otherwise.

The Scheffé Test 1

- Suppose we want to test which of three means is different. We must do three comparisons using the Scheffé test:

$$\bar{X}_1 \text{ versus } \bar{X}_2 \quad \bar{X}_1 \text{ versus } \bar{X}_3 \quad \bar{X}_2 \text{ versus } \bar{X}_3$$

- To compare two means $\bar{X}_i$ and $\bar{X}_j$, compute the test statistic.

$$F_S = \frac{\left(\bar{X}_i - \bar{X}_j\right)^2}{s_W^{\ 2}\left[\left(1/n_i\right)+\left(1/n_j\right)\right]}$$

where n_i and n_j are the sample sizes and $s_W^{\ 2}$ is the within-group variance.

- We compare this to the critical value.

$$F' = (k-1)(\text{C.V.})$$

where k is the total number of groups and C.V. is the critical value used in the ANOVA F test.

- Verify if Fs is different than F’

The Tukey Test 1

- The Tukey test can only be used if each sample has the same size n.

- To compare two means $\bar{X}_i$ and $\bar{X}_j$, compute the test statistic.

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{s_W^2 / n}}$$

where $s_W{}^2$ is the within-group variance.

- We compare the absolute value of q to a critical value from Table A-9.
- This critical value is based on:
  - The k: the number of groups
  - The v: the degrees of freedom for the within-group variances
    - which is $N - k$.

The Bonferroni Test 1

- The **Bonferroni test** compares each pair of means using a *t*-test with $N - k$ degrees of freedom.
- It compensates for the increase in the significance level due to the number of tests being done by using a smaller value of α to find the critical value.
- First, pick a level of significance.
  - Divide α by the number of different comparisons being done:

Alpha / number of tests

- Alternatively, leave α as it is, but when calculating the *P*-value of each *t* statistic, multiply it by

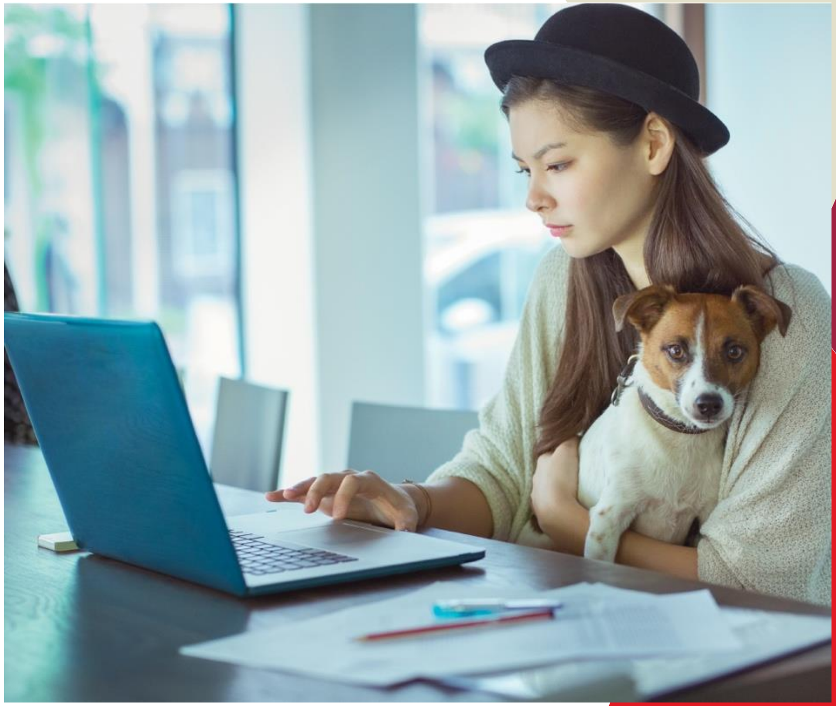
$${}_kC_2.$$

- The test statistic for two means $\bar{X}_i$ and $\bar{X}_j$ is

$$t = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{s_W^{\,2}\left[(1/n_i) + (1/n_j)\right]}}$$

Learning Objectives

- Use the two-way ANOVA technique to determine if there is a significant difference in the main effects or interaction.



Two-Way ANOVA

- **Two-way ANOVA** compare populations that differ due to two independent variables or factors.

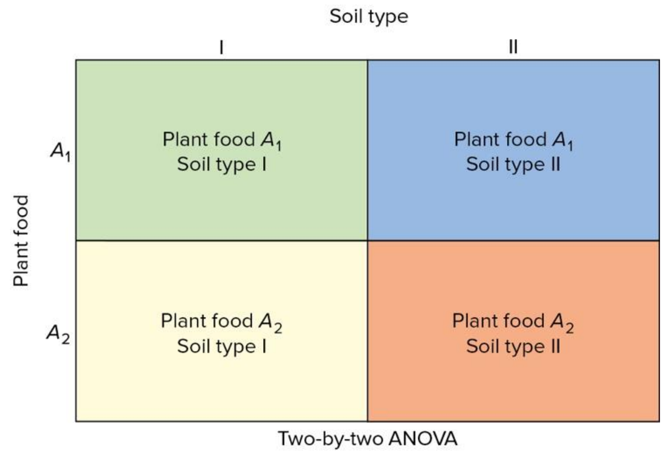
- Performing a two-way ANOVA is very complicated, and our treatment will be only a brief introduction.

- We are interested in the effect two different independent random variables have on the value of a single dependent variable.

- For example,
  - Test the effect of:
  - 2 different types of plant food
  - 2 different types of soil
  - On the growth of certain plants.

Treatment Groups and Effects

- Suppose we call the two types of plant food A_1 and A_2, and the two different types of soil I and II. Then we would have four **treatment groups**, one for each combination of soil and plant food.

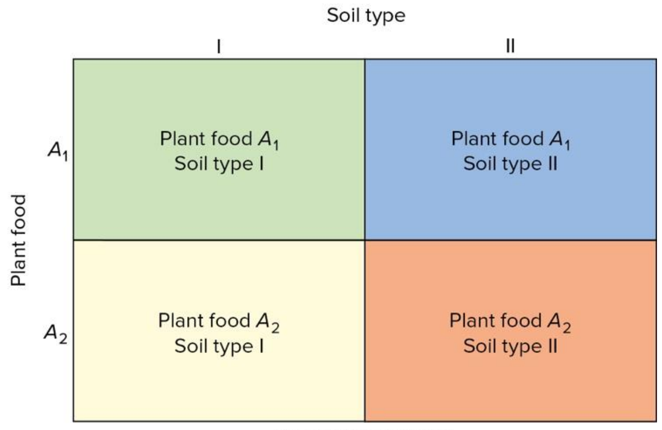


- **Main effect**: When the plant food is changed and the soil type is the same, or vice versa

- **Interaction effect**: A change in the response variable caused by changing both factors.
  - Difference between using plant food A_1 and soil type I and using plant food A_2 and soil type II

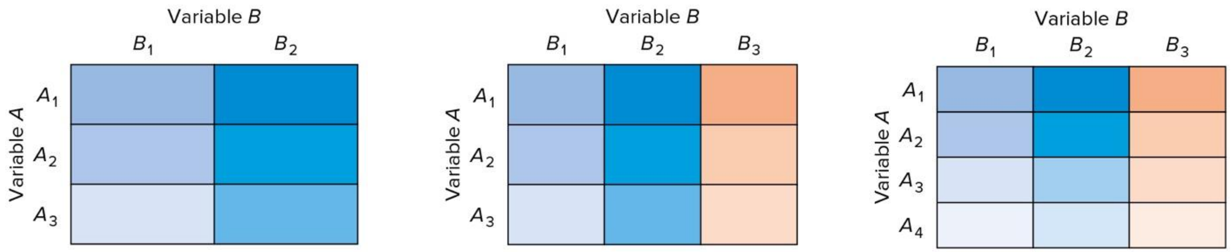
Two-way ANOVA Designs

- Our experimental design is a 2 × 2 ANOVA because each factor has two different possible values.



Two-by-two ANOVA

- If the factors have more than 2 different values, the result is a different design, with more treatment groups.



(a) 3 × 2 design **(b)** 3 × 3 design **(c)** 4 × 3 design

Hypotheses

In a two-way ANOVA there are three sets of hypotheses. One for each factor, and one for the interaction effect.
The hypotheses for plant food are:

- H_0: There is no difference in means of heights of plants grown using different foods.
- H_1: There is a difference in means of heights of plants grown using different foods.

The hypotheses for soil are:

- H_0: There is no difference in means of heights of plants grown using different soils.
- H_1: There is a difference in means of heights of plants grown using different soils.

The hypotheses regarding the plant food – soil type interaction effect are:

- H_0: There is no interaction effect between type of plant food and type of soil.
- H_1: There is an interaction effect between type of plant food and type of soil.

Two-way ANOVA Summary Table 1

- Just as with one-way ANOVA, the sums of squares, degrees of freedom, means squares, and *F* statistics used can be put in a summary table.

TABLE 12–5 ANOVA Summary Table

Source	Sum of squares	d.f.	Mean square	F
Row factor A	SS_A	$a-1$	MS_A	F_A
Column factor B	SS_B	$b-1$	MS_B	F_B
Interaction $A \times B$	$SS_{A\times B}$	$(a-1)(b-1)$	$MS_{A\times B}$	$F_{A\times B}$
Within (error)	SS_W	$ab(n-1)$	MS_W	
Total	SS_{Total}	$N-1$		

- The process for calculating the sums of squares is very complicated

- The mean square for a specific source is calculated using the formula.

$$MS = \frac{SS}{d.f.}$$

- The *F* statistic for one of the factors or the interaction effect is calculated using the formula.

$$F = \frac{MS}{MS_W}$$

Questions