ANOVA: Conditions and F-test

Learning Objective: In a given context, carry out the inferential method for comparing groups and draw the appropriate conclusions.

Step 2: Checking Conditions and Finding the Test Statistic

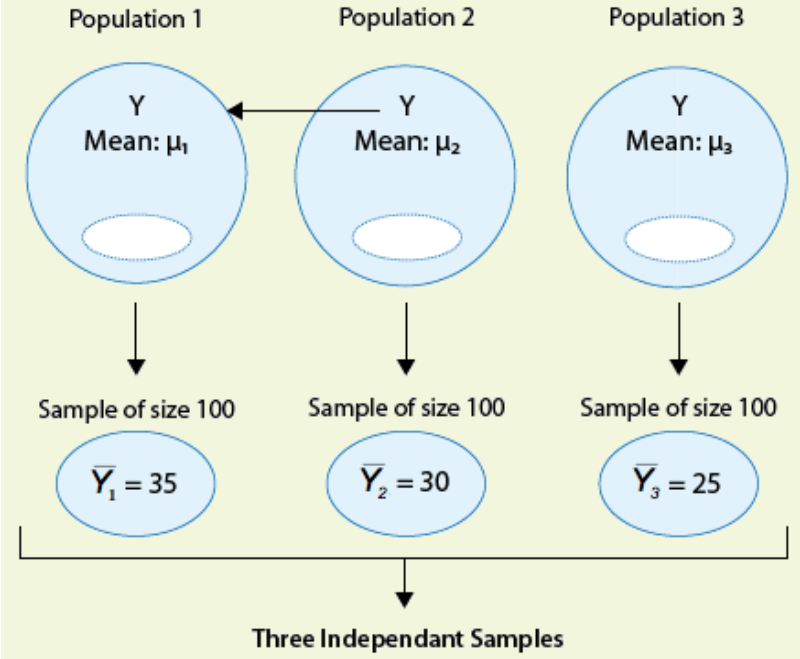
The test statistic of the ANOVA F-test, called the F statistic, has the form

$$F = \frac{\text{VARIATION AMONG SAMPLE MEANS}}{\text{VARIATION WITHIN GROUPS}}$$

As we mentioned earlier, it has a different structure from all the test statistics we've looked at so far; however, it is similar in that it is still a measure of the evidence against H_0. The larger F is (which happens when the denominator, the variation within groups, is small relative to the numerator, the variation among the sample means), the more evidence we have against H_0.

Did I Get This?

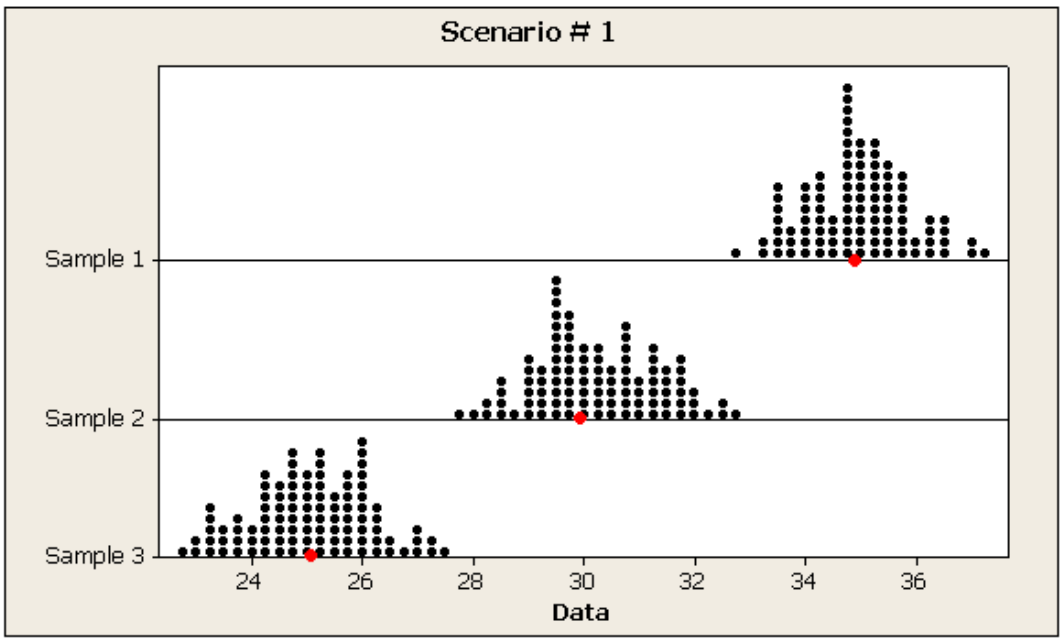
Consider the following generic situation:

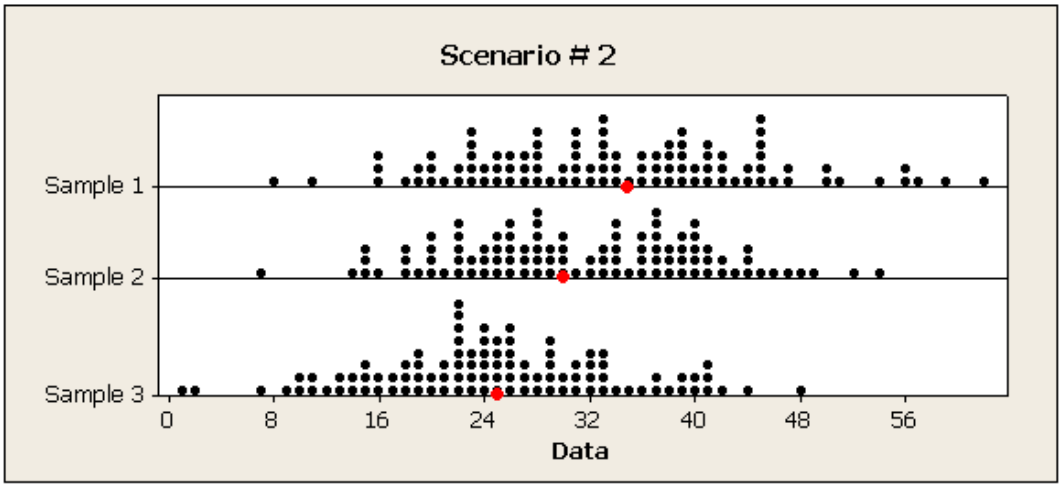


where we're testing:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : not\ \ all\ \ the\ \mu's\ \ are\ \ equal$$

The following are two possible scenarios of the data (note in both scenarios the sample means are 25, 30, and 35).





Did I Get This

1/1 point (graded)

Consider Scenario 1 (compared with Scenario 2):

Is the within-group variation in Scenario 1 relatively small or relatively large as compared to Scenario 2?

- (•) small ✔
- ( ) large

Answer

Correct: Indeed the within-group variation in Scenario 2 is larger than in Scenario 1.

Submit

Did I Get This

1/1 point (graded)

Therefore, is the F statistic for Scenario 1 as compared to Scenario 2 relatively small or relatively large?

Recall that the F statistic has the form:

$$\frac{\text{Variation among sample means}}{\text{Variation within groups}}$$

- small
- large ✔

Answer
Correct:
Indeed, the within-group variation appears in the denominator of the F statistic, and therefore if the within-group variation is relatively small (as it is for Scenario 1), the F statistic will be relatively large.

Submit

Did I Get This

1/1 point (graded)
Does this indicate that H_o will probably be rejected or probably not be rejected?

- be rejected ✔
- not be rejected

Answer
Correct:
Indeed, the larger the F statistic (which happens when the variation within groups is relatively small) the more evidence we have against H_o.

Submit

Did I Get This

1/1 point (graded)
Finally, what will we conclude about the population means?

- they may be equal
- they are not all equal ✔

Answer
Correct:
Indeed, when H_o is rejected, we accept the alternative claim, which in the ANOVA F-test says that the means are not all equal.

Submit

Did I Get This

1/1 point (graded)

Now consider Scenario 2 (compared to Scenario 1):

Is the within-group variation in Scenario 2 relatively small or relatively large as compared to Scenario 1?

- small
- large ✔

Answer

Correct: Indeed, the within-group variation in Scenario 1 is smaller than in Scenario 2.

Submit

Did I Get This

1/1 point (graded)

Therefore, is the F statistic for Scenario 2 as compared to Scenario 1 relatively small or relatively large?

- small ✔
- large

Answer

Correct:
Indeed, the within-group variation appears in the denominator of the F statistic, and therefore if the within-group variation is relatively large (as it is for Scenario 2), the F statistic will be relatively small.

Submit

Did I Get This

1/1 point (graded)

Does this indicate that H_o will probably be rejected or probably not be rejected?

- be rejected
- not be rejected ✔

Answer
Correct:
Indeed, the smaller the F statistic (which happens when the variation within groups is relatively large) the less evidence we have against H_o.

Submit

Did I Get This

1/1 point (graded)
Finally, what will we conclude about the population means?

- may be equal ✔
- are not all equal

Answer
Correct:
Indeed, when we cannot reject H_o, we're essentially saying that it is possible that the means are equal. Recall, however, that not rejecting H_o does not mean that we accept it. In hypothesis testing, we never accept H_o.

Submit

Scenario: Age by Psychological Test Score

Suppose that we would like to compare four populations (for example, four races/ethnicities or four age groups) with respect to a certain psychological test score. More specifically we would like to test:

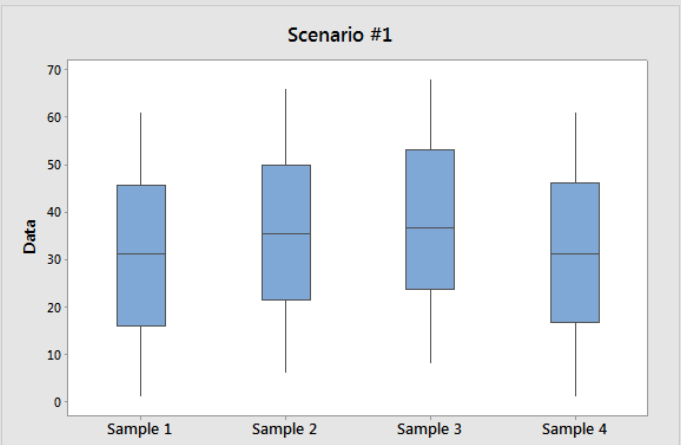
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a : *Not all the* μ*'s are equal*

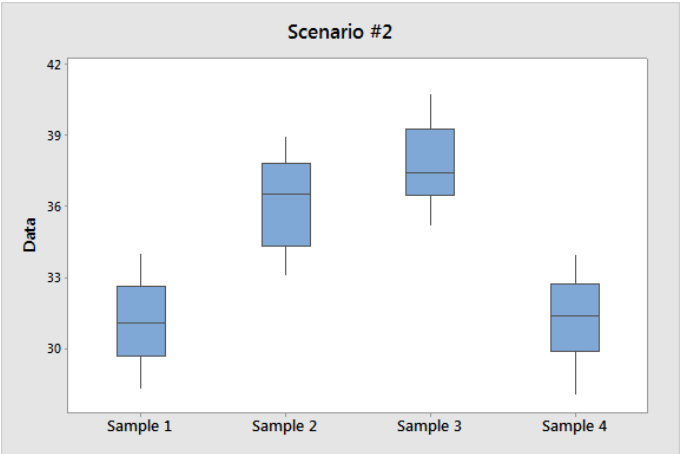
Where: μ_1 is the mean test score in population 1 μ_2 is the mean test score in population 2 μ_3 is the mean test score in population 3 μ_4 is the mean test score in population 4

We take a random sample from each population and use these four independent samples in order to carry out the test.

The following are two possible scenarios for the data:



	Score Averages
Sample 1	$\bar{x}_1 = 30.9$
Sample 2	$\bar{x}_2 = 35.6$
Sample 3	$\bar{x}_3 = 38.0$
Sample 4	$\bar{x}_4 = 31.2$



	Score Averages
Sample 1	$\bar{x}_1 = 31.1$
Sample 2	$\bar{x}_2 = 36.2$
Sample 3	$\bar{x}_3 = 37.7$
Sample 4	$\bar{x}_4 = 31.2$

Note that in both scenarios, the score averages of the four samples are very similar.

Learn By Doing

1/1 point (graded)

In which scenario do the data have a larger within-group variability?

- (●) Scenario 1. ✔
- ( ) Scenario 2.
- ( ) Both scenarios have the same within-group variability.

Answer

Correct:
In scenario 1 we see more spread within each of the four samples (IQR roughly 30 and full range roughly 60) compared to scenario 2 (IQR roughly 3 and the full range is roughly 6).

Submit

Learn By Doing

1/1 point (graded)
In which scenario do the data have a larger F test statistic?

- Scenario 1.
- Scenario 2. ✔
- Both scenarios have the same F test statistic.

Answer
Correct:
In both scenarios the numerator of the F test statistic is roughly equal (since the sample means in both scenarios are very similar). Since in scenario 2 we have smaller within group variability and therefore the denominator of the F test statistic is smaller, the F test statistic is larger.

Submit

Learn By Doing

1/1 point (graded)
In which scenario are we more likely to reject H_0 and conclude that the four population score means are not all equal?

- Scenario 1.
- Scenario 2. ✔

Answer
Correct:
Scenario 2 has smaller within-group variability and therefore a larger F test statistic. The larger the F test statistic, the more evidence the data provide against the null hypothesis.

Submit

Comments

1. The focus here is for you to understand the idea behind this test statistic. We are not going to go into any of the details about how the two variations are measured. This will be included in an extension module to this course in the future. We will rely on software output to obtain the F-statistic.

2. This test is called the ANOVA F-test. So far, we have explained the ANOVA part of the name. Based on the previous tests we introduced, it should not be surprising that the "F-test" part comes from the fact that the null distribution of the test statistic, under which the p-values are calculated, is called an F-distribution. We will say very little about the F-distribution in this course, which will essentially be limited to this comment and the next one.

3. It is fairly straightforward to decide if a z-statistic is large. Even without tables, we should realize by now that a z-statistic of 0.8 is not especially large, whereas a z-statistic of 2.5 is large. In the case of the t-statistic, it is less straightforward, because there is a different t-distribution for every sample size n (and degrees of freedom n - 1). However, the fact that a t-distribution with a large number of degrees of freedom is very close to the Z (standard normal) distribution can help to assess the magnitude of the t-test statistic.

   When the size of the F-statistic must be assessed, the task is even more complicated, because there is a different F-distribution for every combination of the number of groups we are comparing and the total sample size. We will nevertheless say that for most situations, an F-statistic greater than 4 would be considered rather large, but tables or software are needed to get a truly accurate assessment.

Example

Analysis of Variance results:
Data stored in separate columns.
Column means

Column	n	Mean	Std. Error
Business	35	7.3142858	0.48984894
English	35	11.771428	0.35286513
Mathematics	35	13.2	0.3639189
Psychology	35	14.028571	0.52096504

ANOVA table

Source	df	SS	MS	F-Stat	P-value
Treatments	3	939.85	313.28333	46.600895	<0.0001
Error	136	914.2857	6.722689		
Total	139	1854.1357			

Here is the statistics software output for the ANOVA F-test. In particular, note that the F-statistic is 46.60 which is very large, indicating that the data provide a lot of evidence against H_o . (we can also see that the p-value is so small that it is essentially 0, which supports that conclusion as well).

Let's move on to talk about the conditions under which we can safely use the ANOVA F-test, where the first two conditions are very similar to ones we've seen before, but there is a new third condition. It is safe to use the ANOVA procedure when the following conditions hold:

1. The samples drawn from each of the populations we're comparing are independent.

2. The response variable varies normally within each of the populations we're comparing. As you already know, in practice this is done by looking at the histograms of the samples and making sure that there is no evidence of extreme departure from normality in the form of extreme skewness and outliers. Another possibility is to look at side-by-side boxplots of the data, and add histograms if a more detailed view is necessary. For large sample sizes, we don't really need to worry about normality, although it is always a good idea to look at the data.

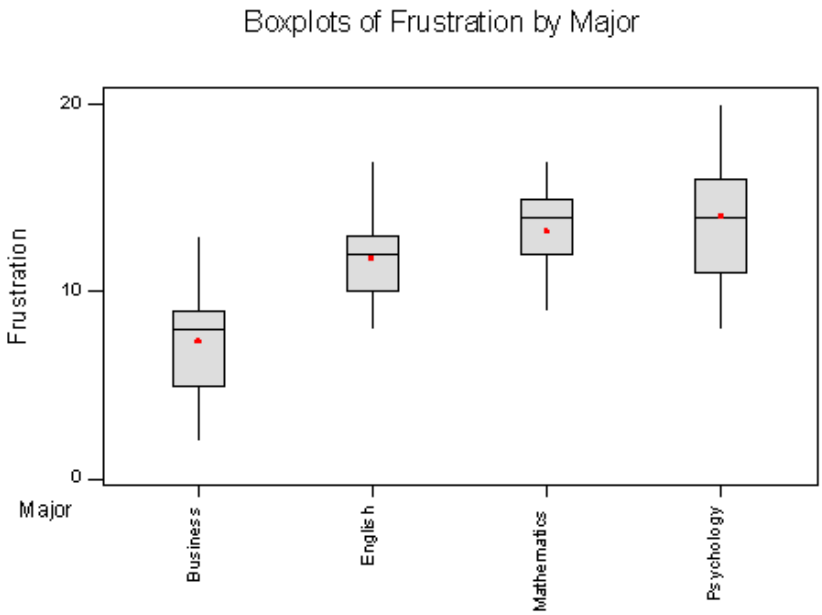
3. The populations all have the same standard deviation. The best we can do to check this condition is to find the **sample** standard deviations of our samples and check whether they are "close." A common rule of thumb is to check whether the ratio between the largest sample standard deviation and the smallest is less than 2. If that's the case, this condition is considered to be satisfied.

Example

In our example all the conditions are satisfied:

1. All the samples were chosen randomly, and are therefore independent.

2. The sample sizes are large enough (n = 35) that we really don't have to worry about the normality; however, let's look at the data using side-by-side boxplots, just to get a sense of it:



You'll recognize this plot as Scenario 2 from earlier. The data suggest that the frustration level of the business students is generally lower than students from the other three majors. The ANOVA F-test will tell us whether these differences are significant.

3. In order to use the rule of thumb, we need to get the sample standard deviations of our samples. Here is the output from statistics software:

Summary statistics:

Column	n	Mean	Std. Err.	Std. Dev.	Min	Q1	Median	Q3	Max
Business	35	7.3142858	0.48984894	**2.8979855**	2	5	8	9	13
English	35	11.771428	0.35286513	**2.0875783**	8	10	12	13	17
Mathematics	35	13.2	0.3639189	**2.1529734**	9	12	14	15	17
Psychology	35	14.028571	0.52096504	**3.0820706**	8	11	14	16	20

The rule of thumb is satisfied since 3.082 / 2.088 < 2.

Scenario: College Credits by Class Standing

In each of the following 3 questions, you'll find two designs for comparing number of credits taken by freshmen vs. sophomores vs. juniors vs. seniors. In each case, one of the designs should not be handled with ANOVA. Your task is to identify which of the two it is.

Did I Get This

1/1 point (graded)

(i) Survey a random sample of 40 seniors and get them to report the number of credits they took in each of their four years, then compare the four sample mean numbers of credits.

(ii) Survey random samples of 40 seniors, 40 juniors, 40 sophomores, and 40 freshmen, and get them to report the number of credits they are taking.

Which of the above two designs should **NOT** be handled with ANOVA?

- (●) design (i) ✔
- ( ) design (ii)

Answer

Correct:

Indeed, this design should **not** be handled with ANOVA since the samples are not independent (all four samples have the same 40 students).

Submit

Did I Get This

1/1 point (graded)

(i) Survey random samples of 5 seniors, 5 juniors, 5 sophomores, and 5 freshmen, and get them to report the number of credits they are taking. Part-time students are included, so there are some extreme low outliers in the data.

(ii) Survey random samples of 5 seniors, 5 juniors, 5 sophomores, and 5 freshmen, and get them to report the number of credits they are taking. Only full-time students are being considered, and the distributions do not display any skewness or outliers.

Which of the above two designs should **NOT** be handled with ANOVA?

- (●) design (i) ✔
- ( ) design (ii)

Answer

Correct:
Indeed, this design should **not** be handled with ANOVA. The sample sizes are quite low (all of size 5), and there is a violation of the normality assumption in the form of extreme outliers.

Submit

Did I Get This

1/1 point (graded)
(i) Survey random samples of 40 seniors, 40 juniors, 40 sophomores, and 40 freshmen, and get them to report the number of credits they are taking. Sample standard deviations are 3.1, 3.4, 1.9, and 2.7.

(ii) Survey random samples of 20 seniors, 20 juniors, 20 sophomores, and 20 freshmen, and get them to report the number of credits they are taking. Sample standard deviations are 3.1, 5.4, 1.9, and 2.7.

Which of the above two designs should **NOT** be handled with ANOVA?

- design (i)
- design (ii) ✔

Answer
Correct:
Indeed, the largest among the sample standard deviations (5.4) is more than twice as large as the smallest one (1.9). Since this rule of thumb is not satisfied, we cannot assume that condition (iii) of equal population standard deviations is satisfied, and cannot therefore use ANOVA.

Submit