S18

**Mathematics**
**Support Centre**

# **Title:** Calculating a Manual F-test (ANOVA)

**Target:** On completion of this worksheet you should be able to identify key assumptions and conduct a manual F-test for normal distribution

**Key Information to start:**
An F-test is a statistical test that is used on the variances of the data of at least 2 samples (i.e. there are two or more independent variables). The outcome of the test will tell us whether the means are statistically different of our independent variables.

**Assumptions**
1) The higher variance is referenced as 1, or the numerator to use a right hand tail which is easier to calculate
2) If degrees of freedom are not listed in the table take the next highest degree, it avoids a type 1 error. Also if it's part way, you can take an estimate of the average between the two values in the table
3) Both samples are normally distributed
4) The samples are independent to each other

**F-ratio**
F-ratio $= \dfrac{\sigma_1^2}{\sigma_2^2}$
where $\sigma_1^2$ is the variance of the first sample and $\sigma_2^2$ is the variance of the second sample.

**Using the F-table**
Degrees of freedom are from our numerator using $v_1 = n_1 - 1$ for sample$_1$, and our denominator using $v_2 = n_2 - 1$ for sample$_2$. We would then require an alpha value to be able to use the table. $v_1$ is degrees of freedom in the numerator $v_2$ is degrees of freedom in the denominator and $\alpha$ is the upper tail probability which we usually use $5\%$ unless stated otherwise.

**Accepting/ rejecting hypothesis**
You should accept the null hypothesis $H_0$ when the F-ratio is smaller than your F-critical meaning the variances are equal or fairly similar. The means are statistically equal. If your F-ratio is greater than your F-critical then we should reject the null hypothesis $H_0$ and accept the alternative hypothesis, $H_1$, meaning our variances are statistically different therefore the means are not statistically equal.

**Example 1**
Kelly was looking at the cars in a car park. On one day she went in and counted the cars and noted what their colours were. She went back in on another day and counted the colours again. On the first day, she found $6$ black cars, $3$ red cars, $8$ blue cars and $9$ white cars, while on the other day she found $14$ black cars, $8$ red cars, $13$ blue cars, and $17$ white cars. She wanted to test if the variances for the two days were statistically different.

**Answer**
Firstly we find the mean. For day $1$ we get $6.5$ cars and day $2$ we get $13$. Using the variance formula $\dfrac{\Sigma(X-\bar{X})^2}{(n-1)}$ we get,
Day 1 variance
$$= \frac{(6-6.5)^2+(3-6.5)^2+(8-6.5)^2+(9-6.5)^2}{4-1}$$
$$= 7$$
Day 2 variance
$$= \frac{(14-13)^2+(8-13)^2+(13-13)^2+(17-13)^2}{4-1}$$
$$= 14$$
Our variance for day $1$ is $7$ and variance for day $2$ is $14$. Let day $2$ represent our numerator and day $1$ be our denominator because $14 > 7$, our F-ratio is
F-ratio $= \dfrac{14}{7} = 2$
Looking up our F-critical, our degrees of freedom for our numerator is $4-1=3$ and our denominator is $4-1=3$. This is because there are $4$ colours of cars that Kelly was looking at so $n=4$. Our alpha is not stated so we take it to be $5\%$ therefore our critical value is $9.28$ according to the table.
$2 < 9.28$ therefore we do no reject our hypothesis, meaning our variance for day $1$ is statistically equal to the variance of day $2$, suggesting our means are therefore statistically the same.


P.T.O.

**F-Test in Regression**

An F-test is used in regression analysis to determinewhether the model created is significantly different to that of a model with no coefficents. In other words, it determines whether your model is worth using. Our Sum of squares is calculated for regression as $SSR = \sum_{j=1}^{N}(\hat{y}_j - \bar{y})^2$,

for error as $SSE = \sum_{j=1}^{N}(y_j - \hat{y})^2$

and for total as $SST = \sum_{j=1}^{N}(y_j - \bar{y}_j)^2$

where $\hat{y}$ is the mean of all observations in the sample

$\bar{y}$ is the mean of all observations

and $y$ is the different values in the sample. Our regression degrees of freedom would be $k$ where $k$ is the number of independent variables. Our residual/ error degrees of freedom is $N - k - 1$ where $N$ is the number of observations. Our total degrees of freedom is $N - 1$. Our mean square is calculated as $MSR = SSR \div K$ and $MSE = SSE \div (N - k - 1)$. To find our F-ratio we would need to do $F = MSR/MSE$. To find the $p$-value we would use the table and our F-ratio given in our created table like before. If our F-ratio is greater than the F-crits in the tables, then it is significant and $p$ is very close to $0$ that we could write $0.00...$. If the F-ratio is smaller than the F-crit, it would be insignificant and $p$ value would be $> c$ where $c$ is the highest percentage we reject at. Our table would look like the following:

| $Source$ | $SS$ | $DF$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| $Regression$ | $SSR$ | $K$ | $MSR = \frac{SSR}{k}$ | $F = \frac{MSR}{MSE}$ | $p$ |
| $Error$ | $SSE$ | $N - k - 1$ | $MSE = \frac{SSE}{(N-k-1)}$ | | |
| $Total$ | $SST$ | $N - 1$ | | | |

**Example on regression**

| $Source$ | $SS$ | $DF$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| $Regression$ | 36464 | 1 | 36464 | $Q4$ | $Q5$ |
| $Error$ | 17173 | $Q2$ | $Q3$ | | |
| $Total$ | $Q1$ | 48 | | | |

Find the values $Q1, Q2, Q3, Q4, Q5$

**Answer**

To find $Q1$ we know that regression add error will give us the total, so, $Total = 36464 + 17173 = 53637$ So $Q1 = 53637$.

For $Q2$ using the formula of $N - k - 1$ where $N - 1$ is $48$ and $k$ is 1 so $N - 1 - k = 48 - 1 = 47$ So $Q2 = 47$

For $Q3$ we use the formula of $MSE = \dfrac{SSE}{(N - k - 1)}$ therefore $MSE = \dfrac{17173}{47} = 365$ So $Q3 = 365$

For $Q4$ we use the formula again $F = \dfrac{36464}{17173} = 99.8$ So $Q4 = 99.8$

Finally for $Q5$ we would need to look at our F distribution table to find our F-crit. Our numerator degrees of freedom is $1$ as its our regression and denominator is $47$ which is our error. Looking at the column of $1$ and row of $47$ or near $47$ if the value is not in the table, our value at $5\%$ is $4.06$ if we go to a lower percentage, (bottom of the group is $0.1\%$) we get $12.50$. It is clear that our F-ratio is far greater than our crit therefore we accept and our $p$-value is $< 0.001$ which can be written as either $< 0.001$ or $0.00...$ to show how small it is and how close it is to 0. So our table would look like this:

| $Source$ | $SS$ | $DF$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| $Regression$ | 36464 | 1 | 36464 | 99.8 | 0.00.. |
| $Error$ | 17173 | 47 | 365 | | |
| $Total$ | 53637 | 48 | | | |

**Exercises**

1. The management are testing out a new method to enhancing the company's performance. They wanted to see if the productivity as a whole has improved and want to make sure that the numbers have generally improved and not a conicendence. They evaluated $16$ performances of the old method and $16$ performances of the new method. The old had a productivity of per hour: $6, 8, 15, 11, 14, 9, 12, 7, 13, 12, 9, 11, 13, 8, 10, 15$. The new had a productivity of per hour: $18, 12, 15, 13, 17, 14, 11, 17, 15, 14, 16, 19, 13, 15,\ 10, 18$. Use a F-test to show the variances and means are similar and can exclude any outside influences.

2. A meteorologist recorded the amount of rainfall per day in mm on April and July. They wanted to see if the data was related in their variance. The recording of April was: $5.1, 0, 2.3, 1.6, 1.7, 3.2, 2.6, 4.6, 0, 0, 0, 2.1, 3.5, 0, 4.1, 1.2, 0.3, 2.4, 0, 0.5, 2.7, 6.1, 0.2, 0.4, 1.5, 0, 0,$ $5.2, 3.4, 1.3$.
   The recording of July was: $0, 0, 0, 1.3, 0, 3.1, 0, 0, 1.2, 1.5, 0.5, 0, 0.6, 1.3, 0, 0,$ $0.3, 0.9, 2.3, 0, 0, 2.6, 2.1, 1.5, 0, 0, 1.3, 1.7, 0, 0, 0$.
   Use the F-test to see if their variances are related.

3. A bank investor was looking into another company to invest. It is using a company that they work well with to be the comparison for the new company. The company they already work with has patrnership deals rated out of 10 by the investors as: $8, 6, 7, 5, 9, 6, 7, 7, 4, 8$. The new company they are looking into has some patnership deals that are rated: $3, 5, 9, 2, 9$. They want to see if their means and variances are the same, because if they are, then the bank investors believe that they will be as successful as the other company and would be very wise to invest into. Use F-test to see if the variances and means are the same and say whether it is worth for the bank investors to invest in.

4. Find $Q1, Q2, Q3, Q4, Q5, Q6$

| $Source$ | $SS$ | $DF$ | $MS$ | $F$ | $p$ |
|---|---|---|---|---|---|
| $Regression$ | 1450 | 1 | $Q3$ | $Q5$ | $Q6$ |
| $Error$ | $Q1$ | 8 | $Q4$ | | |
| $Total$ | 2056 | $Q2$ | | | |

**Answers**

1. The variances are $7.895833$ and $6.829167$, the F-value is $1.156193$ and critical is $2.4034$. Therefore do not reject null hypothesis, meaning there was no influences and the new method has a higher performance rate, and should therefore use the new method.

2. April variance was $3.4$ and July was $0.85$. The value is higher than the crit therefore reject null hypothesis.

3. Old company variance $= 2.233333$, new company variance $= 10.8$. F-value is $4.835821$, F-crit is $4.7725$. Reject null hypothesis as F-ratio $>$ F-crit. The bank investors should not invest with this company just yet.

4. $Q1 = 606, Q2 = 9, Q3 = 1450, Q4 = 75.8, Q5 = 19.14, Q6 => 0.1\%$