Answer on Question #83498 – Math – Statistics and Probability

Question

A study was done to quantify the effect of cigarette smoking on standard measures of lung function in patients with idiopathic pulmonary fibrosis across different age groups. Among the measurements taken were percent predicted residual volumes. The results by smoking history were as follows

Age group	Never	Former	Current
10-15	35	62	95
16-20	120	73	107
20-25	90	60	63
26-30	109	77	134
30-35	82	52	140
36-10	40	115	103
Above 40	68	82	158

Using the above data can we conclude that there is a difference among population residual means at 5% level of significance in terms of smoking history?

Solution

State the null and alternative hypotheses. The null hypothesis for an ANOVA always assumes the population means are equal.
H_0: The residual means are statistically equal in terms of smoking history.

Since the null hypothesis assumes all the means are equal, we could reject the null hypothesis if only mean is not equal. Thus, the alternative hypothesis is:
H_a: At least one residual mean is not statistically equal.

First find mean for each sample:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

For sample «Never» X_N we have $\bar{X}_N = \frac{35+120+90+109+82+40+68}{7} = 77.7143.$

For sample «Former» X_F: $\bar{X}_F = \frac{62+73+60+77+52+155+82}{7} = 74.4286.$

For sample «Current» X_C: $\bar{X}_C = \frac{95+107+63+134+140+103+158}{7} = 114.2857.$

So σ is unknown, we use is the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$. or

For sample X_N we have $s_N^2 = 1046.238$, sample X_F: $s_F^2 = 429.619$, sample X_C $s_C^2 = 1023.905$.

Then find total Sum of Squares (SST)

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{s}(X_{ij} - \bar{\bar{X}})^2,$$

where *r* is the number of rows in the our table, c is the number of columns, $\bar{\bar{X}}$ is the grand mean. Using the data in the table above we may find the grand mean:

$$\bar{\bar{X}} = \frac{\sum X_{ij}}{N} = \frac{35 + 120 + 90 + 109 + 82 + 40 + 68 + 62 + 73 + \ldots + 103 + 158}{21} = 88.8095$$

$$SST = (35 - 88.8095)^2 + (120 - 88.8095)^2 + (90 - 88.8095)^2 \ldots + (103 - 88.8095)^2 = 21851.2381.$$

Then find Treatment Sum of Squares (SSTR):

$$SSTR = \sum r_j(\bar{X}_j - \bar{\bar{X}})^2,$$

where *rj* is the number of rows in the *j*-th treatment and *X j* is the mean of the *j*-th treatment. Use the data $\bar{X}_N = 77.7143, \bar{X}_F = 74.4286, \bar{X}_C = 114.2857, \bar{\bar{X}} = 88.8095$:

$$SSTR = 7 \cdot (77.7143 - 88.8095)^2 + 7 \cdot (74.4286 - 88.8095)^2 + 7 \cdot (114.2857 - 88.8095)^2 = 6852.6667.$$

Find Error Sum of Squares (SSE):

$$SSE = \sum\sum(X_{ij} - \bar{X}_j)^2$$

$$SSE = (35 - 77.7143)^2 + (120 - 77.7143)^2 + (90 - 77.7413)^2 \ldots + (103 - 114.2857)^2 = 14998.5714$$

Note that $SST = SSTR + SSE$. In our case $21851.2381 = 6852.6667 + 14998.5714$

The next step in an ANOVA is to compute the “average” sources of variation in the data using SST, SSTR, and SSE.

Total Mean Squares (MST):

$$MST = \frac{SST}{N - 1},$$

where N is the total number of observations.

$$MST = \frac{21851.2381}{21 - 1} = 1092.5619$$

Mean Square Treatment (MSTR):

$$MSTR = \frac{SSTR}{c - 1},$$

where c is the number of columns in the data table.

$$MSTR = \frac{6852.6667}{3 - 1} = 3426.3334$$

Mean Square Error (MSE):

$$MSE = \frac{SSE}{N - c}.$$

$$MSE = \frac{14998.5714}{21-3} = 833.2539$$

The test statistic may now be calculated. For a one-way ANOVA the test statistic is equal:

$$F = \frac{MSTR}{MSE} = \frac{3426.3334}{833.2539} = 4.11$$

Find the critical value from an F distribution with 5% level of significance. FCV has df1 and df2 degrees of freedom, where df1 is the numerator (MSTR) degrees of freedom equal to $c-1$ and df2 is the denominator (MSE) degrees of freedom equal to $N-c$.
In our case

$$df1 = c-1 = 3-1 = 2, df2 = N-c = 21-3 = 18.$$

We need to find $F_{2,18}^{CV}$ corresponding to $\alpha = 0.05$. Using the F tables we determine $F_{2,18}^{CV} = 3.55$. We reject the null hypothesis if: F (observed value) > F^{CV} (critical value). In our case $4.11 > 3.55$. So we reject the null hypothesis. And we can conclude that there is a difference among population residual means at 5% level of significance in terms of smoking history

Answer: Yes**,** we can conclude that there is a difference among population residual means at 5% level of significance in terms of smoking history.