STAT 3202: Practice 08

Spring 2019, OSU

Exercise 1

A study compared 15 students who intended to major in engineering with 15 students who intended to major in language and literature. Given in the accompanying table are the means and standard deviations of the scores on the verbal and mathematics portion of the SAT for the two groups of students:

	Verbal	Math
Engineering	$\bar{y} = 446$, $s = 42$	$\bar{y} = 548$, $s = 57$
Language/literature	$\bar{y} = 534$, $s = 45$	$\bar{y} = 517$, $s = 52$

Do three things:

- Perform the ANOVA overall F-test to determine whether there is sufficient evidence to claim a difference in the mean math SAT scores between high school students who intend to major in engineering and and those who intend to major in language/literature. Report the p-value of this test. (You should find the p-value using `R`.)
- Perform a two-sample t-test to determine whether there is sufficient evidence to claim a difference in the mean verbal SAT scores for high school students who intend to major in engineering and language/literature. Again, report the p-value of this test. (You should find the p-value using `R`.)
- Compare the p-values of the two tests.

Solution

To perform the F-test, we find the MSE based on the s_i^2 values:

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{14 \cdot 57^2 + 14 \cdot 52^2}{28} = 2976.5$$

In order to complete the ANOVA table, we want to calculate MST which is given by:

$$\frac{1}{K-1} \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2$$

We first calculate:

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \frac{15(548) + 15(517)}{30} = 532.5$$

And then:

$$\text{MST} = \frac{1}{2-1} \cdot \left(15 \cdot (548 - 532.5)^2 + 15 \cdot (517 - 532.5)^2\right) = 7207.5$$

So then our test statistic (which can also be found by completing the ANOVA table) is:

$$F_{\text{obs}} = \frac{\text{MST}}{\text{MSE}} = \frac{7207.5}{2976.5} = 2.421$$

We compare this to a $F_{1,28}$ distribution:

```
1 - pf(2.421, df1 = 1, df2 = 28)
```

```
## [1] 0.1309504
```

Thus, the p-value is: $\boxed{0.131}$

To perform the two sample t-test, we calculate the t-statistic:

$$t_{\text{obs}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The calculation for s_p^2 is identical to the MSE calculated for the F-test, so:

$$s_p^2 = 2976.5$$

So then

$$t_{\text{obs}} = \frac{548 - 517}{\sqrt{2976.5\left(\frac{1}{15} + \frac{1}{15}\right)}} = 1.556$$

```
2 * (1 - pt(1.556, df = 28))
```

```
## [1] 0.1309401
```

Thus, the p-value is: $\boxed{0.131}$

The p-values for the two tests are $\boxed{\text{identical}}$, up to rounding errors.

Exercise 2

In a comparison of the strengths of concrete produced by four experimental mixes, three specimens were prepared from each type of mix. Each of the 12 specimens was subjected to increasingly compressive loads until breakdown. The accompanying table gives the compressive loads, in tons per square inch, attained at breakdown.

Mix A	Mix B	Mix C	Mix D
2.30	2.20	2.15	2.25
2.20	2.10	2.15	2.15
2.25	2.20	2.20	2.25

Do the data provide evidence at the $\alpha = 0.05$ level that at least one of the concretes differs in average strength from the others? You may use R to complete this exercise. Your answer should include:

- The null and alternative hypotheses you are testing.
- A completed ANOVA table – i.e., a table formatted like the one below, with values in all starred entries:

Source	df	SS	MS	F-test statistic	p-value
Treatment	**	**	**	**	**
Error	**	**	**		
Total	**	**			

- A one sentence conclusion in words.

Solution

We are testing:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D \text{ vs. } H_A : \mu_i \neq \mu_j \text{ for some } i, j$$

where μ_i is the mean compressive load, in tons per square inch, attained at breakdown for Mix i.

We can enter the data into `R`:

```r
concrete_data = data.frame(
  mix = factor(rep(c("A", "B", "C", "D"), each = 3)),
  breakdownload = c(2.30, 2.20, 2.25, 2.20, 2.10, 2.20,
                    2.15, 2.15, 2.20, 2.25, 2.15, 2.25))
```

We then run `aov()` and then `summary()` to obtain the ANOVA table.

```r
summary(aov(breakdownload ~ mix, data = concrete_data))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## mix          3  0.015  0.0050       2  0.193
## Residuals    8  0.020  0.0025
```

Our ANOVA table is thus:

Source	df	SS	MS	F-test statistic	p-value
Treatment	3	0.015	0.0050	2	0.193
Error	8	0.020	0.0025		
Total	11	0.035			

And our conclusion is:

At the $\alpha = 0.05$ level, we fail to reject the null hypothesis, and conclude that there is not statistical support at that significance level to conclude there are any differences in concrete strength.

Exercise 3

Water samples were taken at four different locations in a river to determine whether the quantity of dissolved oxygen, a measure of water pollution, differed from one location to another. Locations 1 and 2 were selected above an industrial plant, one near the shore and the other in midstream; location 3 was adjacent to the industrial water discharge for the plant; and location 4 was slightly downriver in midstream. Five water specimens were randomly selected at each location, but one specimen, from location 4, was lost in the laboratory. The sample sizes, means, and variances are shown in the accompanying table (the greater the pollution, the lower will be the dissolved oxygen readings). Do the data provide sufficient evidence to indicate a difference in mean dissolved oxygen content for the four locations? Provide a completed ANOVA table, including p-value, for this situation.

Location	Sample Size	Sample Mean	Sample Variance
1	5	6.08	0.022
2	5	6.44	0.013
3	5	4.78	0.097
4	4	6.025	0.029

Solution

The SSE is:

$$\text{SSE} = (n_1-1)S_1^2+(n_2-1)S_2^2+(n_3-1)S_3^2+(n_4-1)S_4^2 = 4(0.022)+4(0.013)+4(0.097)+3(0.029) = 0.615$$

To calculate the SST, we first calculate the grand me:

$$\bar{y} = \frac{n_1\bar{y}_1+n_2\bar{y}_2+n_3\bar{y}_3+n_4\bar{y}_4}{n_1+n_2+n_3+n_4} = \frac{5\cdot 6.08+5\cdot 6.44+5\cdot 4.78+4\cdot 6.025}{5+5+5+4} = 5.82$$

Then the SST is:

$$\begin{aligned}\text{SST} &= n_1(\bar{y}_1-\bar{y})^2+n_2(\bar{y}_2-\bar{y})^2+n_3(\bar{y}_3-\bar{y})^2+n_4(\bar{y}_4-\bar{y})^2\\ &= 5(6.08-5.82)^2+5(6.44-5.82)^2+5(4.78-5.82)^2+4(6.025-5.82)^2\\ &= 7.836\end{aligned}$$

We can now complete the ANOVA table:

Source	df	SS	MS	F-test statistic	p-value
Treatment	3	7.836	2.612	63.7	nearly 0
Error	15	0.615	0.041		
Total	18	8.451			

The treatment degrees of freedom are $K-1=4-1=3$; the error degrees of freedom are $n-K=5+5+5+4=19$. Both entries in the Total row are the sum of the entries above them. The MS column is the SS column divided by the df column. The F-test statistic is the ratio of the MST to the MSE. The p-value is found by comparing the test statistic 63.7 to an $F_{3,15}$. (You should verify this results using `R`.)

Example 4

Returning to the data in Exercise 3, suppose researchers want to know whether the locations just above the industrial plant are less polluted than the location downriver. That is, they want to test whether the dissolved oxygen at locations 1 and 2 is higher than at location 4. Write down the appropriate hypotheses to test. Find the appropriate test statistic. What are your results at $\alpha = 0.05$?

Solution

We want to test:

$$H_0: \frac{\mu_1+\mu_2}{2}-\mu_4 = 0 \text{ vs. } H_A: \frac{\mu_1+\mu_2}{2}-\mu_4 > 0$$

Our estimator for $\frac{\mu_1+\mu_2}{2} - \mu_4$ is given by:

$$\frac{\bar{y}_1 + \bar{y}_2}{2} - \bar{y}_4$$

Given this data, our estimate is:

$$\frac{6.08 + 6.44}{2} - 6.025 = 0.235$$

The variance of our estimator is:

$$\begin{aligned}
\text{Var}\left[\frac{\bar{y}_1 + \bar{y}_2}{2} - \bar{y}_4\right] &= \frac{1}{4}\text{Var}[\bar{y}_1] + \frac{1}{4}\text{Var}[\bar{y}_2] + \text{Var}[\bar{y}_4], \text{ since the groups are independent} \\
&= \sigma^2\left(\frac{1}{4n_1} + \frac{1}{4n_2} + \frac{1}{n_4}\right) \\
&= \sigma^2\left(\frac{1}{20} + \frac{1}{20} + \frac{1}{4}\right)
\end{aligned}$$

We estimate the variance using the MSE, and take the square root to get the standard error (the estimate of the standard deviation):

$$\text{SE}\left[\frac{\bar{y}_1 + \bar{y}_2}{2} - \bar{y}_4\right] = \sqrt{\text{MSE}\left(\frac{1}{20} + \frac{1}{20} + \frac{1}{4}\right)} = 0.1198$$

Then the observed test statistic is then:

$$\frac{0.235}{0.1198} = 1.96$$

We compare to a $t_{n-K} = t_{15}$:

```r
1 - pt(1.96, df = 15)
```

```
## [1] 0.03442208
```

The p-value is 0.03, so at $\alpha = 0.05$, we reject the null and conclude that there is evidence of their being more pollution downriver from the plant compared with just above the plant.

Exercise 5

Three skin cleansing agents were used on nine people. For each person, a patch of skin were exposed to a contaminant and afterward cleansed by using one of the three cleansing agents, assigned at random. After 8 hours, the residual contaminant was measured, with the following results:

- SST = 1.18
- SSE = 3.02

Each cleansing agent was used three times. Test the hypothesis that there are no differences among the treatment means, using $\alpha = 0.05$.

Solution

We can complete the ANOVA Table:

Source	df	SS	MS	F-test statistic	p-value
Treatment	2	1.18	0.59	1.18	0.369
Error	6	3.02	0.50		
Total	8	4.20			

We fail to reject the null, and conclude that this data does not give us evidence of any differences among the cleansing agents.