STAT 3202: Practice 09

Spring 2019, OSU

Exercise 1

An experiment was designed to test the effect of nematodes (microscopic worms) on plant growth. A botanist prepares 16 identical planting pots and then introduces different numbers of nematodes into the pots. Four pots get 0 nematodes, four get 1,000, four get 5,000, and four get 10,000. A tomato seedling is transplanted into each pot. Seedlings are assigned to pots (i.e., to treatments) using a completely randomized design. Below are the data on the increase in height of the seedlings in centimeters 16 days after planting.

Number of Nematodes			
0	1,000	5,000	10,000
10.8	11.1	5.4	5.8
9.1	11.1	4.6	5.3
13.5	8.2	7.4	3.2
9.2	11.3	5.0	7.5

In R, perform the ANOVA overall F-test, at $\alpha = 0.01$. Then use the `TukeyHSD` function to determine what differences there are among the treatment groups. What levels of nematodes produce different seedling growth, at family-wise error rate (FWER) 0.01? Summarize in a sentence or two what this tells us about the treatment levels.

Solution

We read the data into R.

```
nema <- data.frame(NumNematodes=factor(c(0,0,0,0,1000,1000,1000,1000,5000,5000,5000,5000,
 10000,10000,10000,10000)),SeedlingGrowth=c(10.8,9.1,13.5,9.2,11.1,11.1,8.2,11.3,5.4,4.6,
 7.4,5.0,5.8,5.3,3.2,7.5))
summary(nema)
```

```
 NumNematodes SeedlingGrowth
 0    :4      Min.   : 3.200
 1000 :4      1st Qu.: 5.375
 5000 :4      Median : 7.850
 10000:4      Mean   : 8.031
              3rd Qu.:10.875
              Max.   :13.500
```

We perform the ANOVA overall F-test:

```
aov.fit <- aov(SeedlingGrowth~NumNematodes, data=nema)
summary(aov.fit)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
NumNematodes  3 100.65   33.55   12.08 0.000616 ***
Residuals    12  33.33    2.78
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find a p-value of 0.0006, which is less than 0.01, and which suggests that there are differences in mean seedling growth among the nematode treatment groups.

Next, we use the Tukey method for pairwise comparisons:

```r
TukeyHSD(aov.fit, conf.level=0.99)
```

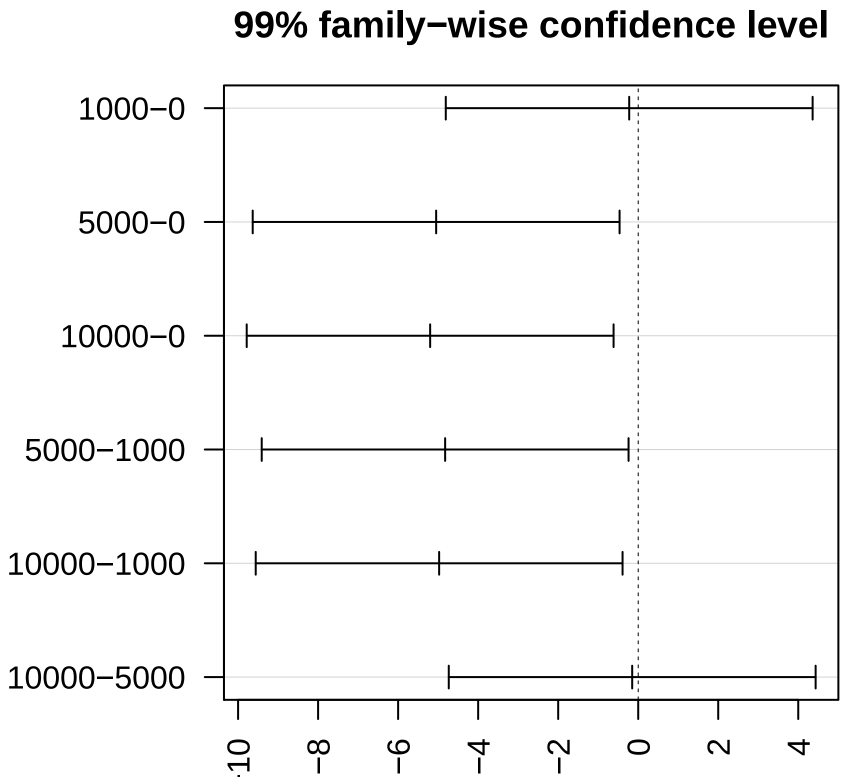
```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = SeedlingGrowth ~ NumNematodes, data = nema)

$NumNematodes
             diff       lwr        upr     p adj
1000-0     -0.225 -4.809287  4.3592874 0.9973921
5000-0     -5.050 -9.634287 -0.4657126 0.0050470
10000-0    -5.200 -9.784287 -0.6157126 0.0040599
5000-1000  -4.825 -9.409287 -0.2407126 0.0070131
10000-1000 -4.975 -9.559287 -0.3907126 0.0056301
10000-5000 -0.150 -4.734287  4.4342874 0.9992199
```

Note that you can ask R to plot the results, which can make it easier to see what is happening.

```r
par(mar=c(2, 7, 2, 2))
plot(TukeyHSD(aov.fit, conf.level=0.99), las=2)
```



To summarize the differences identified, we compare the *adjusted p-values* to our desired α level directly – i.e., the p-values have already been adjusted for multiple comparisons. (The p-values actually don't change when you change the confidence level – try it! – the confidence level only affects the intervals.) So here, since `5000-0`, `10000-0`, `5000-1000`, `10000-1000` all have p-values less than 0.01, we declare those to be significant at FWER 0.01. Our data suggest that the 5000 and 10,000 nematode treatments both reduce seedling growth vs. the 0 nematode treatment, and they both reduce seedling growth vs. the 1,000 nematode treatment. The

0 and 1,000 groups are indistinguishable, and the 5,000 and 10,000 groups are indistinguishable.

Exercise 2

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).

Investigate whether locations 1, 2, and 3 are different from location 4. That is, if μ_i is the mean dissolved oxygen at location i, build confidence intervals on $\mu_1 - \mu_4$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$. Construct the three confidence intervals using a *Bonferroni correction* to ensure that the *family-wise* confidence level is 95%.

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

Note that we found in the previous practice problem set that the MSE was 0.041.

To use a Bonferroni correction to get an overall FWER of 0.05, we want to construct each confidence interval to have level 0.05/3=0.0167, so $\frac{\alpha}{2} = 0.0083$ in each interval, and so $t_{\frac{\alpha}{2},n-K} = t_{0.0083,15} = 2.70$.

So then our three intervals, with family-wise confidence 95%, are:

$$\mu_1 - \mu_4 \in 6.08 - 6.025 \pm 2.70\sqrt{0.041\left(\frac{1}{5} + \frac{1}{4}\right)} = \boxed{(-0.312, 0.422)}$$

$$\mu_2 - \mu_4 \in 6.44 - 6.025 \pm 2.70\sqrt{0.041\left(\frac{1}{5} + \frac{1}{4}\right)} = \boxed{(0.048, 0.782)}$$

and

$$\mu_3 - \mu_4 \in 4.78 - 6.025 \pm 2.70\sqrt{0.041\left(\frac{1}{5} + \frac{1}{4}\right)} = \boxed{(-1.612, -0.878)}.$$

Exercise 3

Do you hate Mondays? Researchers in Germany have provided another reason for you: They concluded that the risk of heart attack on a Monday for a working person may be as much as 50% greater than on any other day. The researchers kept track of heart attacks and coronary arrests over a period of 5 years among 330,000 people who lived near Augsberg, Germany. In an attempt to verify the researcher's claim, 200 working people who had recently had heart attacks were surveyed. The day on which their heart attacks occurred appear in the following table.

Sunday	Monday	Tuesday	Wednesday	Thurdsay	Friday	Saturday
24	36	27	26	32	26	29

Do these data present sufficient evidence to indicate that there is a difference in the percentages of heart attacks that occur on different days of the week? Test using $\alpha = 0.05$.

Solution

We let p_1 be the probability of a heart attack on Sunday, p_2 on Monday, and so on.

We want to test:

$$H_0 : p_1 = p_2 = \cdots = p_7 = \frac{1}{7}$$

vs.

$$H_A : \text{ Some } p_i \neq \frac{1}{7}.$$

We begin by calculating the expected number of heart attacks (out of 200 total) on each day assuming H_0 holds. For each day, we expect:

$$\frac{1}{7}(200) = 28.571$$

heart attacks. We note that these expected cell counts are much larger than 5, so the Chi-Square Test will be reliable.

We construct the test statistic, writing X_i as the observed number of heart attacks, with numbering matching the p_i.

$$\begin{aligned}
X^2 &= \sum_{i=1}^{7} \frac{(X_i - E[X_i])^2}{E[X_i]} \\
&= \frac{(24 - 28.571)^2}{28.571} + \frac{(36 - 28.571)^2}{28.571} + \frac{(27 - 28.571)^2}{28.571} + \frac{(26 - 28.571)^2}{28.571} \\
&\quad + \frac{(32 - 28.571)^2}{28.571} + \frac{(26 - 28.571)^2}{28.571} + \frac{(29 - 28.571)^2}{28.571} \\
&= 0.7313024 + 1.93168 + 0.08638273 + 0.2313549 \\
&\quad + 0.4115376 + 0.2313549 + 0.006441532 \\
&= 3.630054
\end{aligned}$$

Under H_0, $X^2 \sim \chi^2_{\text{df}}$, where here $\text{df} = k - 1 = 7 - 1$, since we have seven cells and only one constraint (that all cell probabilities sum to 1). Our p-value is thus:

```
1-pchisq(3.630054, df=6)
```

```
## [1] 0.7265938
```

With a p-value of 0.73, we fail to reject H_0 at level $\alpha = 0.05$. Our data do not show a significant deviation from the hypothesis that the heart attack rate is the same across days of the week.

In R, we can use `chisq.test`.

```
chisq.test(x=c(24, 36, 27, 26, 32, 26, 29))
```

```
##
##  Chi-squared test for given probabilities
##
```

```
## data:  c(24, 36, 27, 26, 32, 26, 29)
## X-squared = 3.63, df = 6, p-value = 0.7266
```

```
## explicitly specifying the model -- the default is equal cell probabilities
chisq.test(x=c(24, 36, 27, 26, 32, 26, 29),
 p=rep(1/7, 7))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(24, 36, 27, 26, 32, 26, 29)
## X-squared = 3.63, df = 6, p-value = 0.7266
```

Exercise 4

Is the chance of getting a cold influenced by the number of social contacts a person has? A study by Sheldon Cohen, a psychology professor at Carnegie Melon University, seems to show that the more social relationships a person has, the *less susceptible* the person is to colds. A group of 276 healthy men and women were grouped according to their number of relationships (such as parent, friend, church member, and neighbor). They were then exposed to a virus that causes colds. An adaptation of the results is given in the following table.

	Number of Relationships		
	3 or fewer	4 or 5	6 or more
Cold	49	43	34
No cold	31	57	62
Total	80	100	96

Do the data present sufficient evidence to indicate that susceptibility to colds is affected by the number of relationships that people have? Test at the 5% level of significance.

Solution

Let p_i be the probability that an individual in group i developed a cold, where Group 1 is has 3 or fewer relationships; Group 2 has 4 or 5; and Group 3 has 6 or more.

We want to test:

$$\boxed{H_0 : p_1 = p_2 = p_3 \text{ vs. } H_A : \text{ At least one } p_i \neq p_j.}$$

(It's also fine to write that hypothesis out in words / talk about row-column independence / etc.)

We can find the expected numbers in each cell in various ways, but one way is to use the formula $\frac{r_i c_j}{N}$. The row sums are $r_1 = 126$ and $r_2 = 150$, and the total sample size is $N = 276$.

	Number of Relationships		
	3 or fewer	4 or 5	6 or more
(Expected) Cold	$\frac{126 \cdot 80}{276} = 36.52$	$\frac{126 \cdot 100}{276} = 45.65$	$\frac{126 \cdot 96}{276} = 43.83$
(Expected) No cold	$\frac{150 \cdot 80}{276} = 43.48$	$\frac{150 \cdot 100}{276} = 54.35$	$\frac{150 \cdot 96}{276} = 52.17$

Then our test statistic is:

$$X^2 = \frac{(49 - 36.52)^2}{36.52} + \frac{(43 - 45.65)^2}{45.65} + \frac{(34 - 43.83)^2}{43.83} + \frac{(31 - 43.48)^2}{43.48} + \frac{(57 - 54.35)^2}{54.35} + \frac{(62 - 52.17)^2}{52.17} = \boxed{12.19}$$

We compare to a χ^2_2, since the degrees of freedom are $\boxed{(r-1)(c-1)=2}$, and get the p-value of $\boxed{0.002}$. Thus, we $\boxed{\text{reject}}$ H_0 and conclude there are differences in susceptibility to cold among these three groups.

We can check our work in R:

```
chisq.test(matrix(c(49, 43, 34, 31, 57, 62), byrow=TRUE, nrow=2))
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix(c(49, 43, 34, 31, 57, 62), byrow = TRUE, nrow = 2)
## X-squared = 12.182, df = 2, p-value = 0.002263
```

Exercise 5

Returning to the risk of developing a cold data in the previous exercise, calculate two confidence intervals for $p_1 - p_2$ and $p_1 - p_3$, where p_i is the probability of getting a cold in Group i, where Group 1 has 3 or fewer relationships; Group 2 has 4 or 5; and Group 3 has 6 or more. Use a Bonferroni correction to calculate these confidence intervals so as to maintain a *family-wise confidence level* of 95%.

Solution

For a family-wise confidence level of 95%, we want to divide the $\alpha = 0.05$ among two intervals, so each interval has $\alpha = 0.025$ (and has level 97.5%). The critical value in each interval should be: $z_{0.025/2} = z_{0.0125} = 2.24$. Thus, the intervals are:

$$\begin{aligned} p_1 - p_2 &\in \frac{49}{80} - \frac{43}{100} \pm 2.24\sqrt{\frac{\frac{49}{80}(1-\frac{49}{80})}{80} + \frac{\frac{43}{100}(1-\frac{43}{100})}{100}} = \boxed{(0.018, 0.347)} \\ p_1 - p_3 &\in \frac{49}{80} - \frac{34}{96} \pm 2.24\sqrt{\frac{\frac{49}{80}(1-\frac{49}{80})}{80} + \frac{\frac{34}{96}(1-\frac{34}{96})}{96}} = \boxed{(0.095, 0.422)} \end{aligned}$$

Thus, with family-wise confidence level 95%, we find that p_1 is bigger than both p_2 and p_3.