JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Statistics for laboratory scientists II

Solutions for the homework problems for lecture 6

1. Here is the R code for creating the observed data:

```
x <- c(159, 190, 204, 206, 222, 223)
y <- c(370, 376, 418, 488, 490, 503, 512, 532, 587, 605, 637)
```

   a. We can use the function `var.test()` to test whether the two treatments have the same underlying population variance.

   ```
   var.test(x,y) # P-value = 0.0099; 95% CI for var ratio = (0.017, 0.48)
   ```

   I would conclude that the variability of response to the two *is* different.

   b. To obtain a 95% confidence interval for the ratio of the two underlying proportions, we take the square-root of the 95% CI for the ratio of the population variances.

   ```
   result <- var.test(x,y)
   sqrt( result$conf.int )      # 95% CI = (0.13, 0.70)
   ```

   You might be interested in the ratio of the SD under treatment B to the SD under treatment A (i.e., the reciprocal of that considered above) instead:

   ```
   result <- var.test(y,x)
   sqrt( result$conf.int )      # 95% CI = (1.4, 7.6)
   ```

2. We first download the data file and read it into R. (The file is comma-delimited.)

```
mydata <- read.csv("data_hw07-1.csv")
```

The resulting object, `mydata` has two columns, `"diet"` and `"gain"`. Unfortunately, the `"diet"` column is not made a *factor*, and so the function `aov()` for performing the analysis of variance will not work correctly. Thus we need to do the following.

```
is.factor(mydata$diet)                    # Darn! It's not a factor
mydata$diet <- as.factor(mydata$diet)
is.factor(mydata$diet)                    # Now it is.
```

   a. To get the ANOVA table and the p-value for the test of whether the average weight gain is the same for the three diets, we do the following.

```
out <- aov(gain ~ diet, data = mydata)    # perform the ANOVA
summary(out)                              # get table and p-value
```

The ANOVA table we obtain is as follows:

Source	**SS**	**df**	**MS**
Between	36	2	18
Within	210	9	23.3
Total	246	11	

Because the $MS_{between}$ is less than the MS_{within}, we're clearly not going to reject the null hypothesis. We get an F statistic of 0.77 and a P-value of 0.49.

b. Since the P-value is ~50%, we *fail to reject* the null hypothesis, and conclude that **the data are insufficient to conclude that the average weight gains on these three diets are different**.

3. Download and read in the data using something like the following:

```
mydata <- read.csv("data_hw07-2.csv")
```

We can calculate the sample means, sample SDs and sample sizes for each group using the following:

```
tapply(mydata$Length, mydata$Group, mean)
tapply(mydata$Length, mydata$Group, sd)
tapply(mydata$Length, mydata$Group, length)
```

a. We can make a dotplot of the data using the following:

```
stripchart(mydata$Length ~ mydata$Group, method="jitter")
```

b. To get the ANOVA table and the p-value for the test of whether the average weight gain is the same for the three diets, we do the following.

```
out <- aov(Length~Group, data=mydata)     # perform the ANOVA
summary(out)                              # get table and p-value
```

The ANOVA table we obtain is as follows:

Source	SS	df	MS
Between	871.4	4	217.9
Within	3588.5	60	59.8
Total	4459.9	64	

We get an F statistic of 3.64 and a P-value of 0.01.

c. Since the P-value is quite small, we conclude that there are differences in the average lengths of daffodils in the different areas.