**NATIONAL UNIVERSITY OF SINGAPORE**
**Department of Statistics and Applied Probability**

**2018/19 Semester 2** **ST2137 Computer Aided Data Analysis** **Tutorial 7**

1. The retailing manager of a supermarket chain wants to determine whether product location has any effect on the sale of pet toys. Three different aisle locations are considered: front, middle, and rear. A random sample of 18 stores is selected with 6 stores randomly assigned to each aisle location. The size of the display area and price of the product are constant for all stores. At the end of a one-month trial period, the sales volumes (in thousands of dollars) of the product in each store were recorded in the file "locate.txt" (free format with space as delimiter).
   (a) At the 5% level of significance, is there any evidence of a significant difference in average sales among the various aisle locations. Use SAS.
   (b) If appropriate, which aisle locations appear to differ significantly in average sales? Use SAS.
   (c) At the 5% level of significance, is there any evidence of a significant difference in the variation in sales among the various aisle locations? Use SAS.
   (d) Suppose that at the outset of the study the manager suspected the sales of the front aisle location to be different from the other two locations. By testing on a set of appropriate contrasts, clarify the manager's doubt.
   (e) Repeat parts (a) to (d) using R.
   (f) Repeat parts (a) to (d) using SPSS.

2. Refer to Question 1.
   (a) Normality assumption is one of the assumptions made in testing whether there are differences in average sales among the various aisle locations. Check the normality assumption using SAS.
   (b) Plot the residuals against the predicted values using SAS.
   (c) Repeat parts (a) and (b) using R.
   (d) Repeat parts (a) and (b) using SPSS.

3. An operations manager in a company that manufactures electronic audio equipment is inspecting a new type of battery. A batch of 20 batteries is randomly assigned to four groups (so that there are 5 batteries per group). Each group of batteries is then subjected to a particular pressure level — low, normal, high, or very high. The batteries are simultaneously tested under these levels and the times to failure (in hours) are recorded in the data file "batfail.txt" (free format with space as delimiter).
   The operations manager, by experience, knows that such data come from populations that are not normally distributed. He wants to use a non-parametric procedure for purposes of data analysis.
   (a) At the 5% level of significance, analyze the data to determine whether there is evidence of a significant difference in the four pressure levels with respect to median batteries life. Use SAS. (Computation of exact p-value is not required.)
   (b) Repeat part (a) using R.
   (c) Repeat part (a) using SPSS.

4. The area of a unit square ($[0,1] \times [0,1]$) is 1. The area of a circle (centred at $(0.5, 0.5)$ with radius 0.5) enclosed by the unit square is $\pi(1/2)^2$. We generate a large number of uniformly distributed random points in the unit square. The proportion of these random points that fall in the circle is $\pi/4$. Write an R program to do a simulation to estimate $\pi$ based on the above discussion.

**Answers to selected questions**

1. (a) Test $H_0$: $\mu_1 = \mu_2 = \mu_3$ against $H_1$: $\mu_i \neq \mu_j$ for some $i \neq j$. $F_{obs} = 13.03 > F_{0.05} = 3.68$ (or $p$-value = 0.0005 < 0.05). Reject $H_0$. SAS code: p9.16. R code: p9.19-9.20. SPSS: p9.22.
   (b) $LSD = 1.6001$. $|\bar{X}_F - \bar{X}_M| = 3.8 > LSD$; $|\bar{X}_F - \bar{X}_R| = 2.3333 > LSD$; $|\bar{X}_M - \bar{X}_R| = 1.4667 < LSD$. $\mu_F \neq \mu_M$, $\mu_F \neq \mu_R$, $\mu_M = \mu_R$. SAS code: p9.27. R code: p9.29-9.30. SPSS: p9.31.
   (c) Test $H_0$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ against $H_1$: $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$. Levene's test: $F_{obs} = 1.86$, $p$-value = 0.1891 (SAS is based on squared deviation from the mean), Levene's test: $F_{obs} = 2.46$, $p$-value = 0.1191 (R and SPSS are based on absolute deviation from the mean) Bartlett's test: $\chi^2_{obs} = 4.305$, $p$-value = 0.1162. Do not reject $H_0$. SAS code: p9.42. R code: p9.44-9.45. SPSS: p9.46.
   (d) $C_1 = \mu_2 - \mu_3$ and $C_2 = 2\mu_1 - \mu_2 - \mu_3$. Test $H_{01}$: $C_1 = 0$ against $H_{11}$: $C_1 \neq 0$ and $H_{02}$: $C_2 = 0$ against $H_{12}$: $C_2 \neq 0$. $F_{C_1} = 3.82$ (or $p$-value = 0.0696) Reject $H_{01}$. $F_{C_2} = 22.25$ (or $p$-value = 0.0003) Reject $H_{02}$. SAS code: p9.35. R code: p9.37-9.38. SPSS: p9.39.
2. (a) and (b) Test $H_0$: data from a normal distribution against $H_0$: data not from a normal distribution. KS-test: $D_{obs} = 0.1147$ with $p$-value > 0.150 ($p$-value = 0.9719 from R). Do not reject $H_0$. SAS code: p9.47-9.50.
   (c) R code: p9.51-9.52.
   (d) SPSS: p9.54-56.
3. Test $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1$: $\mu_i \neq \mu_j$ for some $i \neq j$. $\chi^2_{Obs} = 11.91 > \chi^2_{0.05}(3) = 7.81$ (or $p$-value = 0.0077). Reject $H_0$. SAS code: p9.61. R code: p9.63. SPSS: p9.64.
4. Part of R code:

```
n <- 10000;
x <- runif(n); y <- runif(n)
   # Create a vector w which contains the information if (x,y) lies in the circle (i.e. the
   distance of (x,y) from the centre (0.5, 0.5) is less than the radius.)
4*(sum(w)/n) # why multiplied by 4?
```

   Refer to p10.15.