NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

2019/20 Semester 1 **ST2137 Computer Aided Data Analysis** **Tutorial 4**

Three data sets ("wip.txt", "testscores.txt" and "furniture.txt") have been uploaded to the course website in the IVLE. All the three data files are free format files with header, and space as the delimiter.

- There are two variables in the "wip.txt" data set. They are "time" and "plant".
- There are three variables in the "testscores.txt" data set. They are "A", "B" and "gender".
- There is one variable, "days", in the "furniture.txt" data set.

1. In many manufacturing processes the term" work-in-process" (often abbreviated WIP) is used. In a book manufacturing plant the WIP represents the time it takes for sheets from a press to be folded, gathered, sewn, tipped on end sheets, and bound. The data set "wip.txt" represents samples of 20 books at each of two production plants and the processing time (defined as the time in days from when the books came off the press to when they were packed in cartoons) for these jobs.
   (a) For each of the two plants, using SAS to compute the descriptive statistics: the mean, median, first quartile, third quartile, minimum, maximum, the range, interquartile range, variance and standard deviation.
   (b) Draw the histogram and the box plot for the processing time for each of the two plants using SAS.
   (c) Based on (a) and (b), are there any differences between the processing times of the two plants? Explain.
   (d) Repeat (a) and (b) using R.
   (e) Repeat (a) and (b) using SPSS.

2. The director of a training program for a large insurance company wanted to know if the two tests, Test A and Test B, are correlated. Thirty trainees sat for both tests. The test scores are recorded in the file "testscores.txt".
   (a) Draw the scatter plot for the two test scores for all the trainees with a different symbol for different gender.
   (b) Draw separate scatter plots for the two test scores, one for male group and one for female group.
   (c) Based on the plots in (a) and (b), is there any relation between the two test scores? Explain.
   (d) Repeat (a) and (b) using R.
   (e) Repeat (a) and (b) using SPSS.

3. One of the major measures of the quality of service provided by any organization is the speed with which the organization responds to customer complaints. During a recent year a company got 50 complaints. The file "furniture.txt" consists of the number of days between the receipt of the complaint and the resolution of the complaint.
   (a) Use SAS to find the 20% trimmed mean and 20% Winsorized mean.
   (b) Use SPSS to find the Huber's, Tukey's, and Hampel's M-estimators for the location.
   (c) Comment on the robust estimators in (a) and (b). What can you say about the "central" location of the distribution?

(d) Use SAS to find the following three robust estimates of the scale parameter, σ. They are the interquartile range (IQR), the median absolute difference (MAD), and the Gini's mean difference.
(e) Use R to find the interquartile range and the median absolute difference (MAD).
(f) What can you say about the "variability" of the distribution?

Partial answers/hints to selected questions

1. (a) SAS: proc univariate; class plant; var time;
Plant 1: mean = 9.382, median = 8.515, Q_1 = 7.395, Q_3 = 11.170, min = 4.42, max = 21.62, range = 17.20, IQR = 3.775, var = 15.9812, s.d. = 3.9977.
Plant 2: mean = 11.3535, median = 11.96, Q_1 = 7.71, Q_3 = 13.98, min = 2.33, max = 27.75, range = 23.42, IQR = 6.27, var = 26.2774, s.d. = 5.1262.
(b) SAS: proc univariate; class plant; var time; histogram time/midpoints=1 to 25 by 2; proc boxplot; plot time*plant;
(d) R: plant.a <- time[plant==1]; summary(plant.a); hist(plant.a,col="grey", breaks=seq(0,27,3)); boxplot(time~plant)
(e) SPSS: "Analyze" → "Descriptive Statistics" → "Explore…" → "time" to "Dependent List" and "plant" to "Factor List" → "Plots…" → "Histogram" → "Continue"
2. (a) SAS: proc sgplot; scatter x=A y=B/ group=gender;
(b) SAS: proc sgplot; by gender; scatter x=A y=B;
(d) R: Create data.frame "t4q2"
Plot with different symbols: Refer to p5.64-p5.67
plot(t4q2$A, t4q2$B, type="n"); points(t4q2[gender=="F", c("A", "B")]); points(t4q2[gender=="M", c("A", "B")], pch=2, col=2)
Separate plots: Refer to p5.73 to p5.75
t4q2.m <- t4q2[gender=="M", c("A","B")]; plot(t4q2$A, t4q2.m$B)
(e) SPSS: Refer to p5.73 to p5.75 for one plot with different symbols.
"Graphs" → "Legacy Dialogs " → "Scatter Plot…" → "Simple Scatter" → "Define" → "B" to y-axis and "A" to x-axis → "gender" to "Set Markers by"
Refer to p5.76 to p5.78 for separate plots. "gender" to "Columns".
3. (a) proc univariate trimmed=0.2 winsorized=0.2; var days; (Refer to p6.19-p6.21)
20% trimmed mean = 30.7, 20% Winsorized mean = 34.62
(b) "Analyze"→ "Descriptive Statistics" → "Explore…". → "Statistics…" → "M-estimators" → "Continue" → "OK" (Refer to p6.23)
Huber = 29.7741, Tukey = 24.1952, Hampel = 27.4688
(d) proc univariate robustscale; var days; (Refer to p6.22)
The estimates of σ are 29.65 (from IQR = 40), 37.68 (from Gini = 42.514), and 22.98 (from MAD = 15.5).
(e) IQR(days) = 38.25, mad(days) = 22.98 (Refer to p6.22)