STATS 2MB3 - Statistical Methods and Applications

Sang Woo Park

January 13, 2017

Course Outline

- Website: https://ms.mcmaster.ca/ bprotas/MATH3Q03/
- Textbook: *Numerical Mathematics*
- Supplemental references: *Approximation Theory and Approximation Practice*
- Five assignments

Contents

1 Introduction

1.1 Probability and statistics

Definition 1.1 (Probability). *A collection of concepts and methos useful to:*

1. *understand and quantify uncertainty (i.e. variability of randomness)*
2. *model uncertainty (e.g. discrete and continuous distributions)*

Definition 1.2 (Statistics). *A collection of analytical and graphical methods useful to*

1. *describe, picture, and summarize a data set (descriptive statistics)*
2. *draw conclusion about a* population *on the basis of observing a portion of it, i.e. a* sample *(inferential statistics).*
3. *verify and refute hypotheses made about a population on the basis of observing a sample (test of statistical hypothesis).*
4. *develop prediction equations from experimental data in the presence of uncertainty (model building, regression model).*

Remark. The statistical methods rely heavily of probability.

1.2 Descriptive statistics

Definition 1.3 (Stem-and-leaf plot). *Stem-and-leaf plots is a graphical method that is useful when summarizing numerical data. A* stem *and* leaf *are associated with every data point.*

Remark. Number of stems in a stem-and-leaf plot is approximately equal to $\sqrt{n}$ where n is the number of points in a given data set (i.e. the sample size).

Example 1.2.1 (Shower flow rate data).

- Do a stem-and-leaf plot of the data.
- Typical observations are 7 l/min, 7.5 l/min, etc.
- It is highly concentrated in the lower side of the scale and spaced out in the upper of the scale.
- The data shows assymetry with a high concentration in the lower side of the scale and spaced out on the larger values. This is called *positive asymmetry* or *positive skewness*.
- Flow rate of 18.9 l/min appears to be unusually far away from the rest of the data. We can consider it to be an *outlier*.

```r
stem(shower_flow)

##
##   The decimal point is at the |
##
##    2 | 23
##    3 | 2344567789
##    4 | 01356889
##    5 | 00001114455666789
##    6 | 0000122223344456667789999
##    7 | 00012233455555668
##    8 | 02233448
##    9 | 012233335666788
##   10 | 2344455688
##   11 | 2335999
##   12 | 37
##   13 | 8
##   14 | 36
##   15 | 0035
##   16 |
##   17 |
##   18 | 9
```

Definition 1.4 (Dot plots)**.** *Dot plots is a useful tool to describe data with repeated observations. Each data point is represented by a* dot*, and the dots are stacked.*

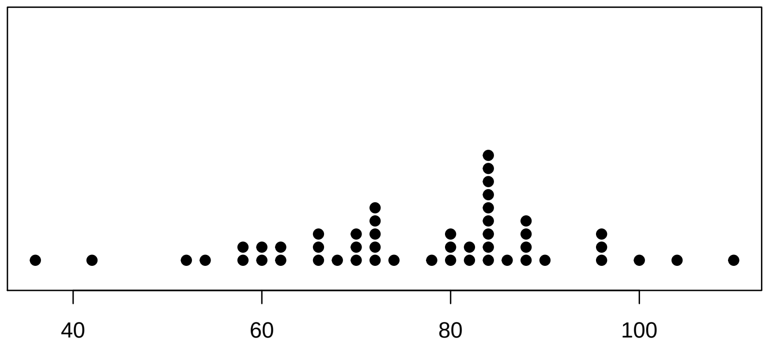
Example 1.2.2 (Pulse rate data)**.** Pulse rate data contains the following:

- n: 50 biomed students
- pulse rate: number of heart beats over 30 seconds multiplied by 2.

With this data, we wish to answer the following questions:

1. Do a dot plot of the data.
2. What can you say about the distribution of the data based on the plot?
3. Are there outliers in the data?

```r
pulse_rate <- scan("pulse_rate.txt")
## http://stackoverflow.com/questions/15244938/how-to-draw-a-stacked-dotplot-in-r
stripchart(pulse_rate, method = "stack", pch = 19, offset = 0.5, at = 0.15)
```



- The data show some negative skewness.
- Observation 42 seems a bit low. Perhaps it's a mild outlier

Definition 1.5 (Frequency table)**.** *Frequency table is a tabular method of visualizing data.*

1. *n: sample size*
2. *Represent the observations by* $x_1, x_2, \ldots, x_n$
3. *Identify the smallest observation,* $x_{(1)}$*, and the largest observation,* $x_{(2)}$*.*
4. *Divide the range of the data into non-overlapping subintervals of equal length.*
5.

Example 1.2.3. Going back to the flow rate data, $n = 129$. Since $\sqrt{n} = 11.69$, we take $k = 12$. Also, we have $x_{(1)} = 2.2$ and $x_{(129)} = 18.9$. Now, we can find the length of each subinterval:

$$L = \frac{x_{(n)} - x_{(1)}}{k} = 1.396 \approx 1.4$$

We can make some observations based on the table (see table 1):

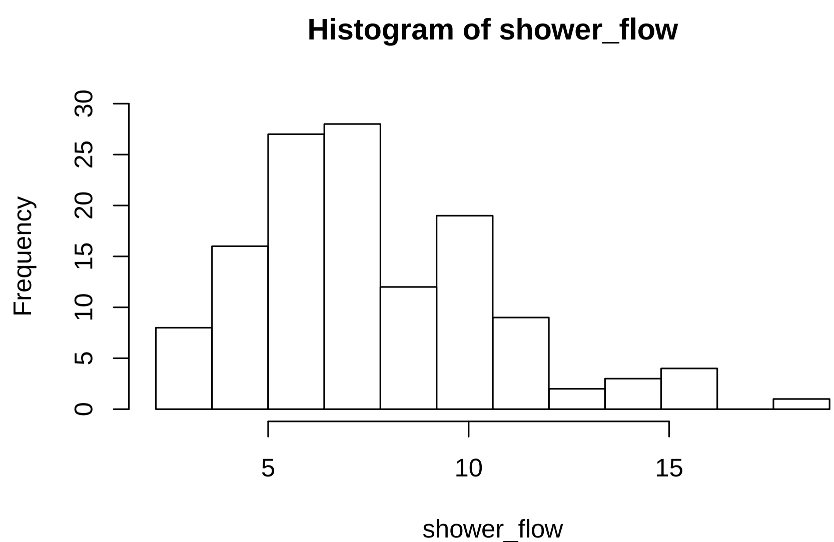
- The proportion of flow rate less than $6.4l/min$ is $0.0542+0.1007+0.2171 = 0.3720$ or 37.2%.
- The proportion of shower flow rates between $5.0l/min$ and $10l/min$ (exclusive) is $0.2171 + 0.2326 + 0.0853 + 0.1550 = 0.69$.

Table 1: Shower flow frequency table

Class	Frequency	Relative frequency
$2.2 \leq x \leq 3.6$	7	0.0542
$3.6 \leq x \leq 5.0$	13	0.1007
$5.0 \leq x \leq 6.4$	28	0.2171
$6.4 \leq x \leq 7.8$	30	0.2326
$7.8 \leq x \leq 9.2$	11	0.0853
$9.2 \leq x \leq 10.6$	20	0.1550
$10.6 \leq x \leq 12.0$	10	0.0775
$12.0 \leq x \leq 13.4$	2	0.0155
$13.4 \leq x \leq 14.8$	3	0.0233
$14.8 \leq x \leq 16.2$	4	0.0310
$16.2 \leq x \leq 17.6$	0	0.0000
$17.6 \leq x \leq 19.0$	1	0.0078

Definition 1.6 (Histogram)**.** *Plot of bars for the frequencies or relative frequencies.*

```r
hist(shower_flow, breaks = seq(2.2, 19, by = 1.4), ylim = c(0, 30))
```



- The histogram converys similar information as the stem-and-leaf plot
- Both plots show right skewness
- Some unusually large observations are noted

- Summarizing data through measures: measures of location or centrality

Definition 1.7. *Given a data $(x_1, x_2, \ldots, x_n)$ with sample size n,* Sample mean or average *is given by*

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Definition 1.8. Sample median*, $\widetilde{x}$, of a data is the middle observation when we order the data. If n is odd, median is the $(n+1)/2$th ordered observation. If n is even, median is given by taking the average of the $(n/2)$th and $(n/2+1)$th ordered observations.*

Definition 1.9 (The trimmed mean)**.** *Trimmed mean of a data ($\bar{x}_{\text{tr}(k)}$) is given by taking the mean of the data after removing the k% smallest observations and the k% largest observations.*

Example 1.2.4. Compute the sample mean, median, and trimmed mean for the shower flow data.

Sample mean. We have $n = 129$. Then, the sample mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i} x_i = 7.7085$$

Sample median. $n = 129$ is odd so we have $\frac{n+1}{2} = 65$. x_i is the 65th observation.

10% trimmed mean. 10% of 129 $= 0.1 \times 129 = 12.9 \approx 13$. We leave out 13 smallest and 13 largest observations. Then, sample size becomes $129 - 2 \times 13 = 103$. Then, we have

$$\bar{x}_{\text{tr}(10)} = \frac{1}{103} \sum_{i} x_i = 7.449$$

- The sample mean, median, and trimmed mean are measures of centrality. They represent typical key points in the centre of the data

- The mean is sensitive to extreme values (large or small), whereas the median is not

- The sample trimmed means typically fall between the sample mean and the sample median

- The sample mean or average is by far the most widely used, followed by the sample median.

- Since the shower flow rate is positively skewed, $\bar{x} > x$.

Definition 1.10 (Sample variance)**.** *sample variance measures variability or dispersion. Sample variance is given by*

$$\begin{aligned}\sigma^2 &= \frac{(x_1-\bar{x}_1)^2+(x_2-\bar{x}_2)^2+\cdots+(x_n-\bar{x}_n)^2}{n-1}\\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2\end{aligned}$$

Note that we can calculate variance more easily through the following formula:

$$\begin{aligned}\sigma^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &= \frac{\sum x_i^2 - n\bar{x}}{n-1}\end{aligned}$$

Definition 1.11 (Sample standard deviation)**.** *Sample standard deviation is obtained by taking a square root of variance:*

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

Theorem 1.1.

$$\sum_{i=1}^{n}(x_i-\bar{x})=0$$