STATS 2MB3 - Statistical Methods and Applications

Sang Woo Park

January 6, 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 | 1234456789
##    4 | 01356889
##    5 | 0000011145666789
##    6 | 000122223344456667789999
##    7 | 00012233455556678
##    8 | 02233448
##    9 | 012233335666788
##   10 | 2344556889
##   11 | 2335999
##   12 | 37
##   13 | 8
##   14 | 03
##   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.*

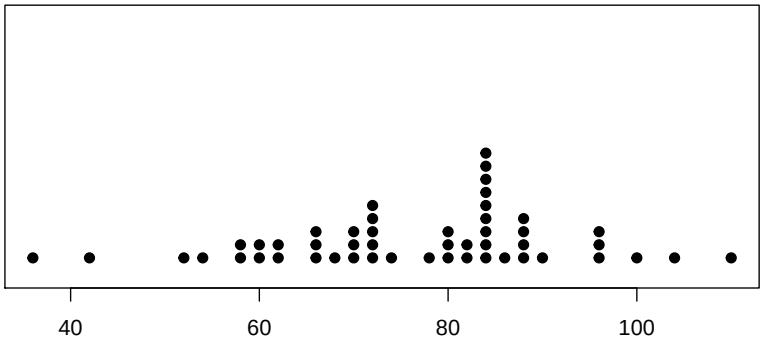
Example 1.2.2 (Pulse rate data)**.** Pulse rate data contains the following:

- 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.*