Class Jan-14-2016

Dajiang Liu

Review of Last Class

- What is a variable, and a case?
- How many sampling strategies are there? What are they?
- How many types of variables?
  - What is numerical random variable?
    - Can you give an example?
  - What is categorical random variable?
    - Can you give an example?

Today Class

- A walk-through of R
- Download the class dataset
  - https://www.openintro.org/stat/textbook.php?stat_book=os
- Download R at CRAN
  - https://cran.r-project.org/

- A nice and cute R book:
  - https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Stent Example in Practice

- Rscript-01-15-2015.R

Scatter Plot

- Email50 dataset
  - Explore relations
- Cars dataset:
  - Explore the relation between price and weight

Mean Value

- Mean value measures the average for a variable
  - $\bar{X} = \frac{(X_1 + X_2 + \cdots + X_N)}{N}$
- Use mean() function in R

- Practice:
  - Can you obtain the mean value for the car price and weight?

Histogram

- Bin the data and count the number in each bin:

Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	25-30	···	55-60	60-65
Count	19	12	6	2	3	5	···	0	1

Table 1.20: The counts for the binned `num_char` data.

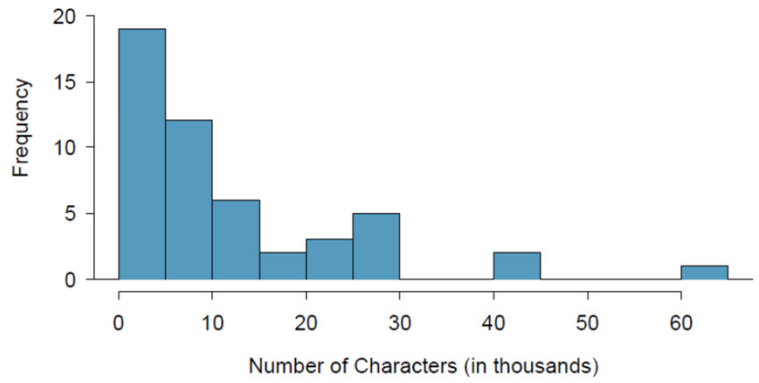
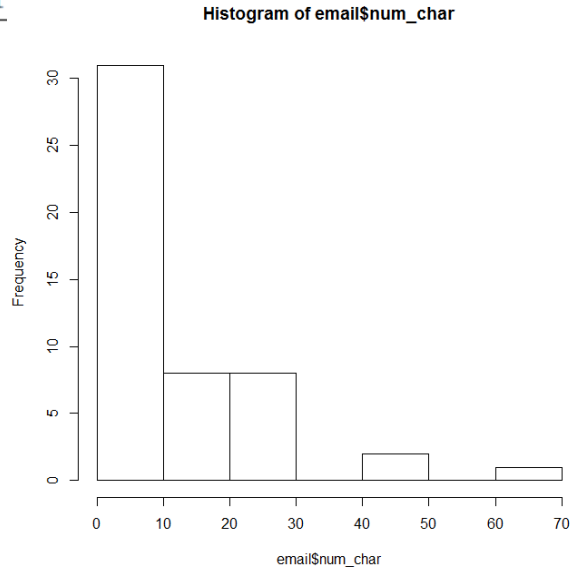


Figure 1.21: A histogram of `num_char`. This distribution is very strongly skewed to the right.



Unimodal, Bimodal and Multimodal

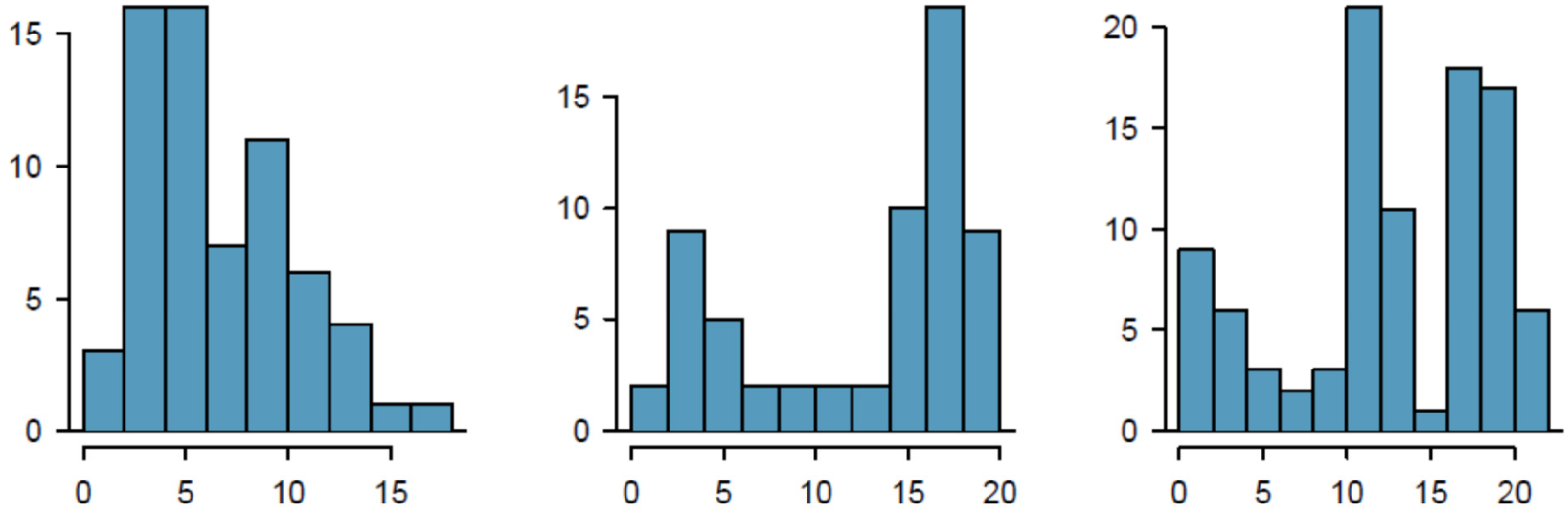


Figure 1.22: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

Mean and Variance

- Variance measures the deviance of each case from the mean
- Mean and variance are not the only description for a distribution

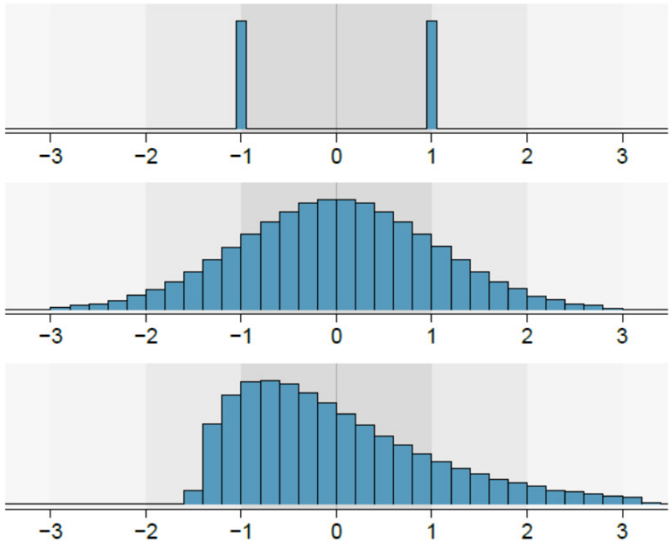


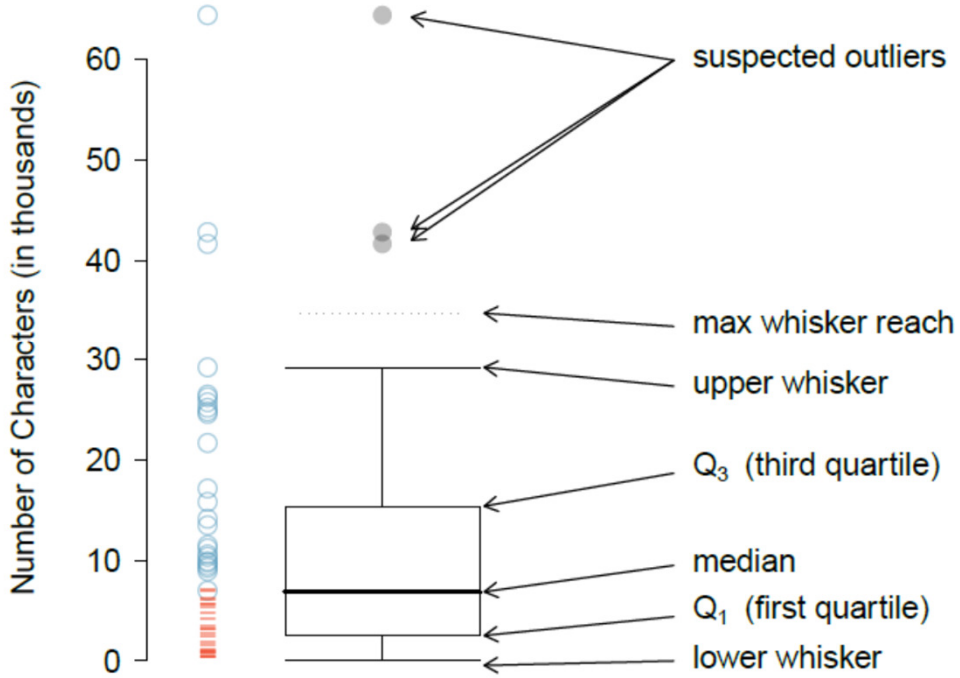
Figure 1.24: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

Median and Quartiles

- Median is the number in the middle
  - If an even number of cases, the median in the average of the two values in the middle

- First and third quartile
  - 25^{th} and 75^{th} percentile in the data
  - Interquartile range
    - Q3-Q1
  - Whisker
    - Upper whisker:
      - Q3+1.5*IQR
    - Lower whisker
      - Q1-1.5*IQR

Box Plot


Number of Characters (in thousands)
0
10
20
30
40
50
60
suspected outliers
max whisker reach
upper whisker
Q_3 (third quartile)
median
Q_1 (first quartile)
lower whisker

R Examples

- Email50 data
  - Num_char
  - What is the IQR, median, mean, upper and lower whisker

Categorical Data

- Contingency table:

		number			
		none	small	big	Total
spam	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

- Barplot
  - Segmented barplot

Homework

- Exercise: 1.1,1.2,1.5,1.6
  - Due Jan-19-2016 in class

- Self study chapters 1-6 Rintro book
- https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf