Columbia University
Instructor: Carsten Chong

STAT GU4001
Spring 2022

Sheet 1: Data Collection and Descriptive Statistics

Exercise 1

[D., p. 50, Exercise 79]
(a)

$$\bar{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{n\bar{x}_n + x_{n+1}}{n+1} \tag{1}$$

(b) In a first step, observe that

$$ns_{n+1}^2 = \sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 = (n-1)s_n^2 + x_{n+1}^2 + n\bar{x}_n^2 - (n+1)\bar{x}_{n+1}^2. \tag{2}$$

Now combine with (a) to get

$$ns_{n+1}^2 = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1}^2 + \bar{x}_n^2 - 2x_{n+1}\bar{x}_n) = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2.$$

(c) The mean is $\bar{x}_{16} = 200.5/16 = 12.53$. Then solving (b) for s_{n+1}^2 gives

$$s_{n+1}^2 = 14/15 * (0.512)^2 + 1/16 * (11.8 - 12.58)^2 = 0.24.$$

The standard deviation is 0.53.

Exercise 2

In the 2017 New York City Housing and Vacancy Survey, randomly selected households were asked about their borough, their monthly rent, their monthly income, the number of bedrooms in their unit, the presence of person elevators, among many other variables.

(a) Indicate which of the listed variables are numerical and which are categorical. If they are numerical, are they continuous or discrete? If they are categorical, list the corresponding categories (= possible values).

(b) Describe one procedure that would have made sure that the households were, at least, approximately, selected according to a simple random sampling.

(c) For the following three alternative procedures, explain why they would have very likely resulted in biased samples:

(1) Ask randomly selected Columbia students and faculty.

(2) Ask randomly selected NYC residents on Facebook and Instagram.

(3) Ask randomly selected people at randomly selected NYC supermarkets.

8

Solution:

Only give point if both rent and income are correct

(a) Borough: categorical ✓ (Bronx, Brooklyn, Manhattan, Queens, Staten Island); ✓ monthly rent, monthly income: Numerical, ✓ continuous; ✓ number of bedrooms: Numerical, ✓ discrete; ✓ presence of person elevator: categorical ✓ (yes, no). ✓

(b) Use recent census data (the actual survey was based on the 2010 census address list).

(c) (1) In 2017, many Columbia students and faculty resided in Morningside Heights or neighboring areas, which are likely not representative for the whole city.

(2) There are many NYC residents that do not use Facebook or Instagram. These residents may have a different response pattern than residents using Facebook or Instagram.

(3) This is closer to a simple random sampling, but still: There are people who go more often to supermarkets than others and the former would have been selected with a higher chance, thus resulting in a bias. ✓

1

only grade (3)

Exercise 3

[D, Section 1.3, Exercise 34ab]

(a) Sample mean of U is

$$\frac{1}{11}(6+5+11+33+4+5+80+18+35+17+23) = 21.55.$$

Sample mean of F is

$$\frac{1}{15}(4+14+11+9+9+8+4+20+5+8.9+21+9.2+3+2+0.3) = 8.56.$$

The sample mean of U is larger.

(b) Sorted data of U:

$$[4, 5, 5, 6, 11, 17, 18, 23, 33, 35, 80].$$

Sorted data of F:

$$[0.3, 2, 3, 4, 4, 5, 8, 8.9, 9, 9, 9.2, 11, 14, 20, 21].$$

From this, the sample medians are 17 and 8.9, respectively. The sample median of U is larger. Mean and median are so different in the U sample because of the outlier 80.0.

Exercise 4

[D, Section 1.2, Exercise 19]

No point if there is no calculation (see underlined parts)

2

(a) From this frequency distribution, the proportion of wafers that contained at least one particle is (100-1)/100 = .99, or 99%.

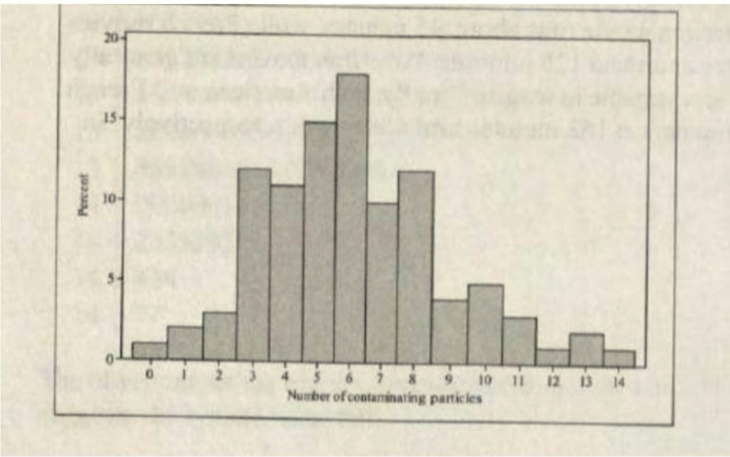
Note that it is much easier to subtract 1(which is the number of wafers that contain 0 particles) from 100 than it would be to add all the frequencies for 1,2,3, ... particles.

2

In a similar fashion, the proportion containing at least 5 particles is (100 - 1-2-3-12-11)/100 = 71/100 = .71, or, 71%.

(b) The proportion containing between 5 and 10 particles is (15+ 18+ 10+ 12+4+5)/100 = 64/100 = .64, or 64%. The proportion that contain strictly between 5 and 10 (meaning strictly more than 5 and strictly less than 10) is (18+ 10+12+4)/100 = 44/100 = .44, or 44%.

(c) The histogram is almost symmetric and unimodal; however, the distribution has a few smaller modes and has a very slight positive skew.



Exercise 5

[D, Section 1.4, Exercise 44]

(a) Range is 2.30.

(b) Variance $s^2 = \sum(x_i - \bar{x})^2/(n-1) = \frac{(180.5-181.41)^2+...+(180.5-181.41)^2}{11} = 0.48.$

(c) Standard deviation is $\sqrt{0.48} = 0.69$.

(d) $s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = (394913.57 - 12 * 181.41^2)/11 = 0.48.$

Exercise 6

4

No point without calculation

[D, Section 1.4, Exercise 51]

(a) $s^2 = \sum(x_i - \bar{x})^2/(n-1) = 1264.77$, $s = \sqrt{1264.77} = 35.56.$

(b) By Proposition 1.2, $s^2 = \frac{1264.77}{60^2} = 0.35$, $s = \sqrt{0.35} = 0.59.$

Exercise 7

Draw a boxplot for the discoveries data that we discussed in class. To do so, first determine the lower quartile, the median, the upper quartile, the LOB, the UOB, the set of outliers, the smallest data point that is not an outlier and the largest data point that is not an outlier.

Sorted data: 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8 9 10 12

Median: 3 ✓

Lower quartile: 2 } ✓

Upper quartile: 4

IQR: 2

LOB: 2 - 1.5 * 2 = -1 } ✓

UOB: 4 + 1.5 * 2 = 7

Outliers: 8, 9, 10, 12 ✓

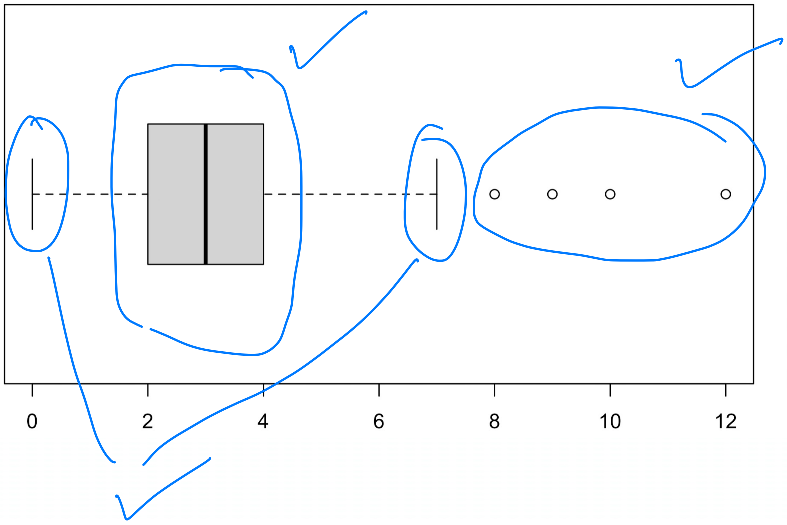
The smallest data point that is not an outlier: 0 } ✓

The largest data point that is not an outlier: 7

}: Only give point if both are correct

Give partial credit: e.g., if lower quartile is wrong, give points if everything else is done correctly by using this wrong value of Q_1

8



Exercise 8

For each landmass exceeding 10,000 square miles, the following table lists its name along with the logarithm of its area in thousands of square miles. You can use that $\widetilde{x} = 3.71$, $Q_1 = 2.99$ and $Q_3 = 5.21$.

(a) Construct a histogram with class intervals $]2, 3],]3, 4], \ldots,]9, 10]$ and relative frequency on the vertical axis. Is the histogram unimodal, bimodal, or multimodal?

(b) Determine the LOB, the UOB, the set of outliers, the smallest data point that is not an outlier and the largest data point that is not an outlier.

(c) Based on this information, draw a boxplot of the data.

```
##          Africa      Antarctica            Asia       Australia
##        9.350624        8.612503        9.740262        7.995644
##    Axel Heiberg          Baffin           Banks          Borneo
##        2.772589        5.214936        3.135494        5.634790
##         Britain         Celebes           Celon            Cuba
##        4.430817        4.290459        3.218876        3.761200
##           Devon       Ellesmere          Europe       Greenland
##        3.044522        4.406719        8.228177        6.733402
##          Hainan      Hispaniola        Hokkaido          Honshu
##        2.564949        3.401197        3.401197        4.488636
##         Iceland         Ireland            Java          Kyushu
##        3.688879        3.496508        3.891820        2.639057
##           Luzon      Madagascar       Melville        Mindanao
##        3.737670        5.424950        2.772589        3.583519
##        Moluccas     New Britain      New Guinea New Zealand (N)
##        3.367296        2.708050        5.723585        3.784190
##  New Zealand (S)     Newfoundland  North America   Novaya Zemlya
##        4.060443        3.761200        9.147401        3.465736
##  Prince of Wales        Sakhalin   South America     Southampton
##        2.564949        3.367296        8.823942        2.772589
##     Spitsbergen         Sumatra          Taiwan        Tasmania
##        2.708050        5.209486        2.639057        3.258097
## Tierra del Fuego           Timor       Vancouver        Victoria
##        2.944439        2.564949        2.484907        4.406719
```

Solution:

(a) See Figure 2. The histogram is bimodal. ✓

5

- Subtract 1 point if one column is wrong
- No points if >1 column is wrong

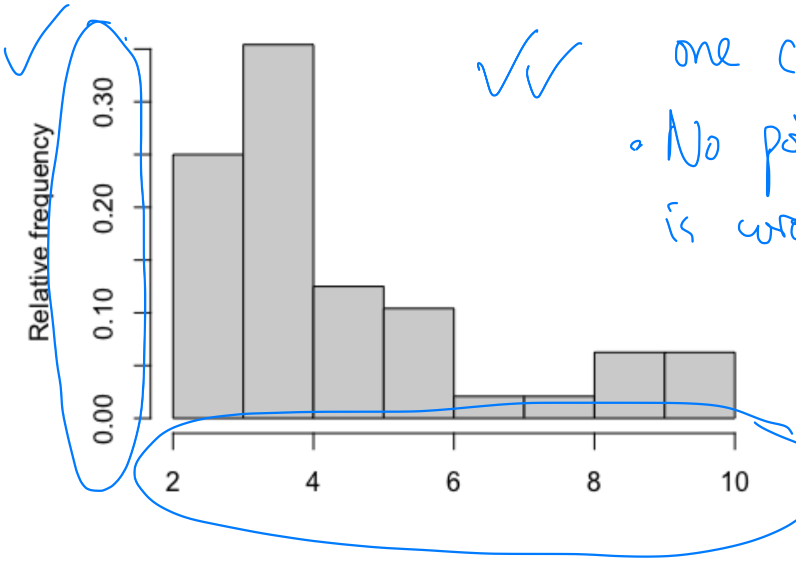


Figure 2: Histogram

(b) We have $Q_1 = 2.99$, $\tilde{x} = 3.71$, $Q_3 = 5.21$, $LOB = 2.99 \quad 1.5 - (5.21 \times 2.99) = -0.34$, $UOB = 5.21 + 1.5 \times (5.21 - 2.99) = 8.54$. Moreover, the smallest data point that is not an outlier is 2.485(Vancouver), the largest data point that is not an outlier is 8.23 (Europe) and the set of outliers consists of the values corresponding to Antartica, Africa, Asia, North America and South America.

(c) See Figure 3

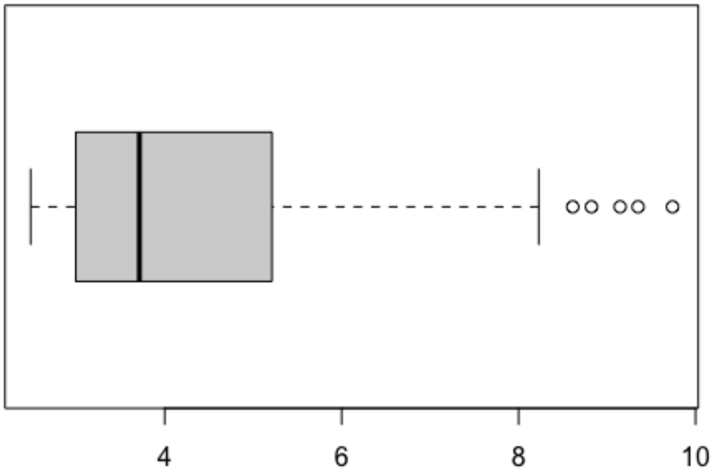


Figure 3: Boxplot

Total : 30 points