大數據統計分析與預測 第二章與第三章作業

created by 黃彥鈞**(Weber Huang)**

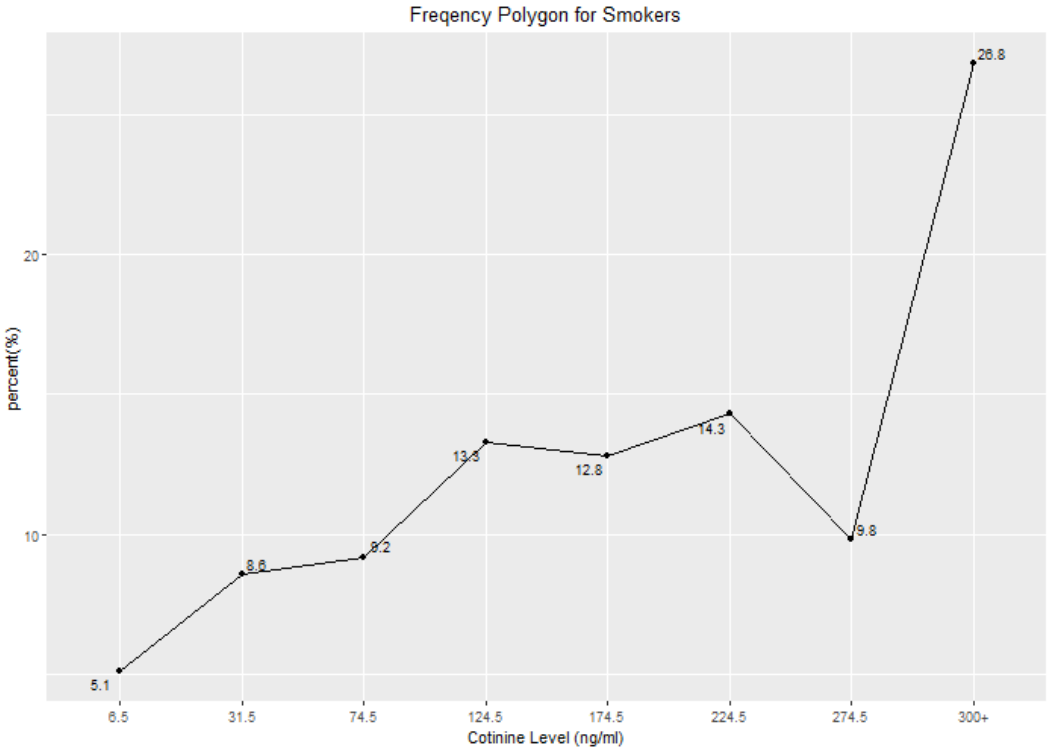
第二章

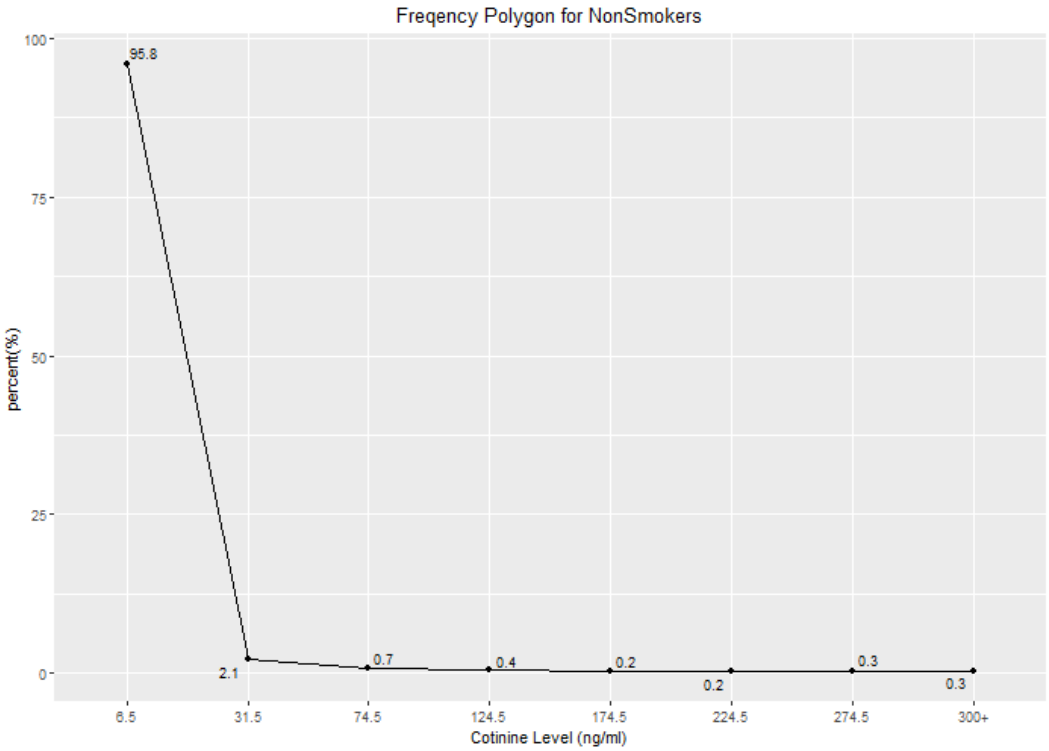
- 2-5. The bar-chart is usually being used to display nominal and ordinal observations.

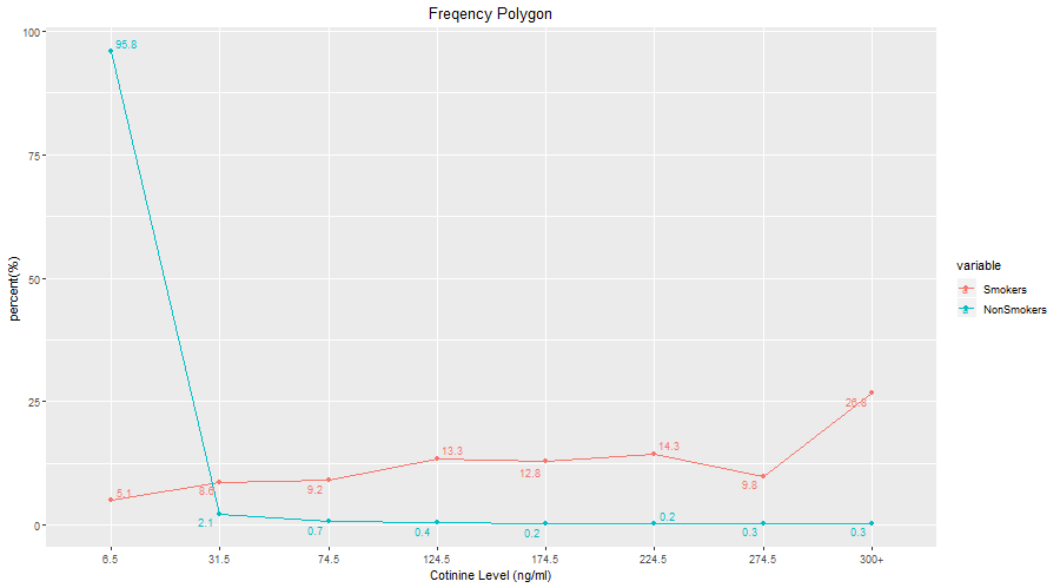
  The histogram is usually being used to display discrete or continuous observations.
- 2-6. The percentiles are what we are used to describe the shape of a data set. For instance, 40th and 60th percentiles represent an equal distance from the midpoint.
- 2-7.
  - (a) Discrete data. Since the suicide is based on people, the minimum unit of people is one person.
  - (b) Continuous data. The concentration is being calculated by the quality of water and lead, and it is common that the quality data are in the float type, so we cannot find out an minimum unit of the data.
  - (c) Continuous data. Since the question does not give us the unit of the measurement time (day, hour, or minute), the survival time of each patient may be differ from each other. The time value might be both continuous or discrete.
  - (d) Discrete data. The number of miscarriages cannot be split into any other smaller unit.
- 2-9. I agree with the statement. What we recognize the "most often" is to find of the **mode** of the data set. In this data set, we can see that the mode is lied in 16-30 minutes.
- 2-13.
  - (a) It is not fair to make a comparison in each interval of smokers and non-smokers via absolute frequencies, since the total number of these two groups are different. We have to calculate the relative frequencies from absolute frequencies first.
  - (b) Relative frequencies table

Cotinine level (ng/ml)	Smokers (%)	Non-smokers (%)
0-13	5.1	95.8
14-49	8.6	2.1
50-99	9.2	0.7
100-149	13.3	0.4
150-199	12.8	0.2
200-249	14.3	0.2
250-299	9.8	0.3
300+	26.8	0.3

- (c)

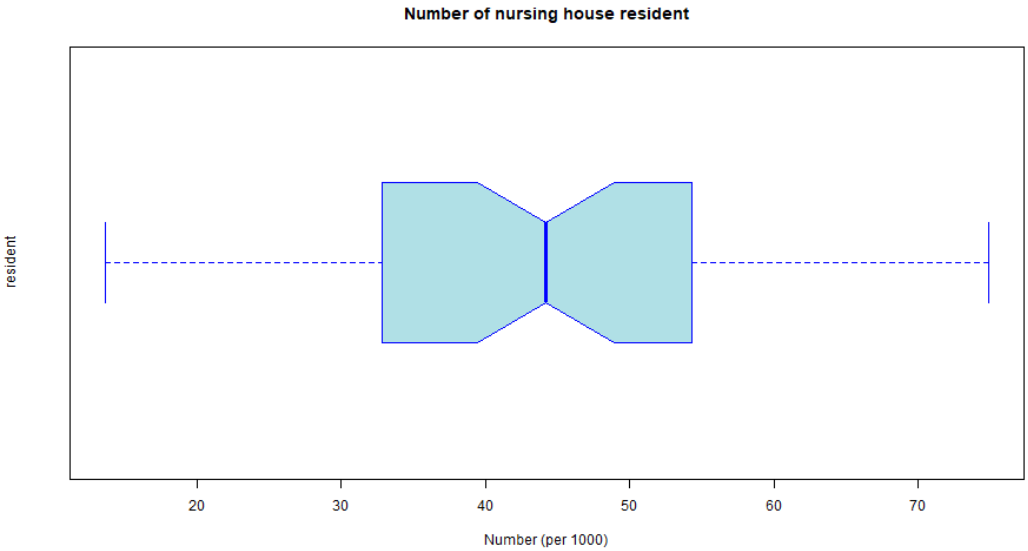

Freqency Polygon for Smokers
percent(%)
20
10
5.1
8.6
9.2
13.3
12.8
14.3
9.8
26.8
6.5
31.5
74.5
124.5
174.5
224.5
274.5
300+
Cotinine Level (ng/ml)


Freqency Polygon for NonSmokers
percent(%)
100
75
50
25
0
95.8
2.1
0.7
0.4
0.2
0.2
0.3
0.3
6.5
31.5
74.5
124.5
174.5
224.5
274.5
300+
Cotinine Level (ng/ml)



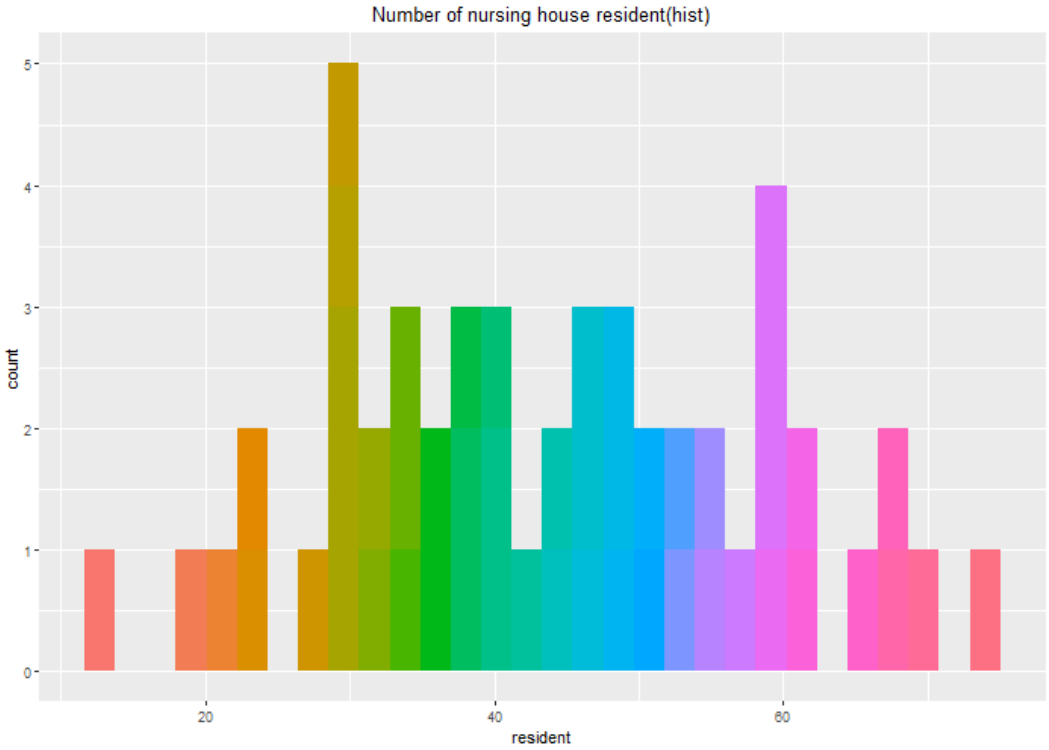
- (d) According to the frequency polygon, the data of smokers is tend to be smoother than the data of non-smokers.
  - In the data of smokers, The frequency is **gradually rising** by the cotinine level ascending, but we can find out that in 250-299 of cotinine level, the frequency suddenly drop to under 10%.
  - On the other hand, the frequency of non-Smokers, is perfectly **right-skewed**, most of the people is in 0-13 of cotinine level.
- (e) We have to notice if there is any interference factor in the study. For example, how can we classify those non-smoker who are highly exposure in secondhand smoke environment?

- 2-18.
  - (a) Hawaii has the smallest number of nursing house residents per 1000 population 65+ age. On the other hand, South Dakota has the largest number of nursing house residents per 1000 population 65+ age.
    There are many factor may cause such as huge difference among those state, like population, age proportion of each state, social welfare, etc.
  - (b)



  - (c) According to the box-plot, the observation of the data is almost symmetric. Only 2 states that the number is larger than 70 and also 2 states the number is smaller than 20. owing to the boxplot, it seems that there isn't any consideration about outliers

- (d)



In the histogram, we can clearly find out that which number is the mode and the frequency of data, while the boxplot tell us the five significant feature of the dataset and detect the outliers. It is hard to say which one is more informative. The type of chart aid chosen depends on the type of data collected, rough analysis of data trends, and project goals.

第三章

- 3-2.
  - mean : Often being used when the data set is normal distribution.
  - median : In the situation that we want to know about data set that is skewed distribution. Unlike mean, median is believed to be robust. for example, national salary analysis.
  - mode : In the circumstance that we want to find out which data occur most. For example, A store sells several kinds of flavor of ice cream, and the store keeper want to know which one is the most popular.
- 3-5. It lets us to say that for any number k that is greater than or equal to 1, at least of :

$$[1-(\frac{1}{k})^2]$$

measurements in the set of data lie within k standard deviations of their mean.

The Empirical rule have to use under some condition, the data should be symmetric and unimodal.

- 3-6.
  - (a)
    - (i) mean : 25.91
    - (ii) Median : 24.00
    - (iii) Mode : 12, 24
    - (iv) Range : 95.9
    - (v) interquartile range : 32

        - (vi) standard deviation : 27.37
    - (b)

      ```
      x <- c(0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96)
      sum(x-mean(x))
      ```

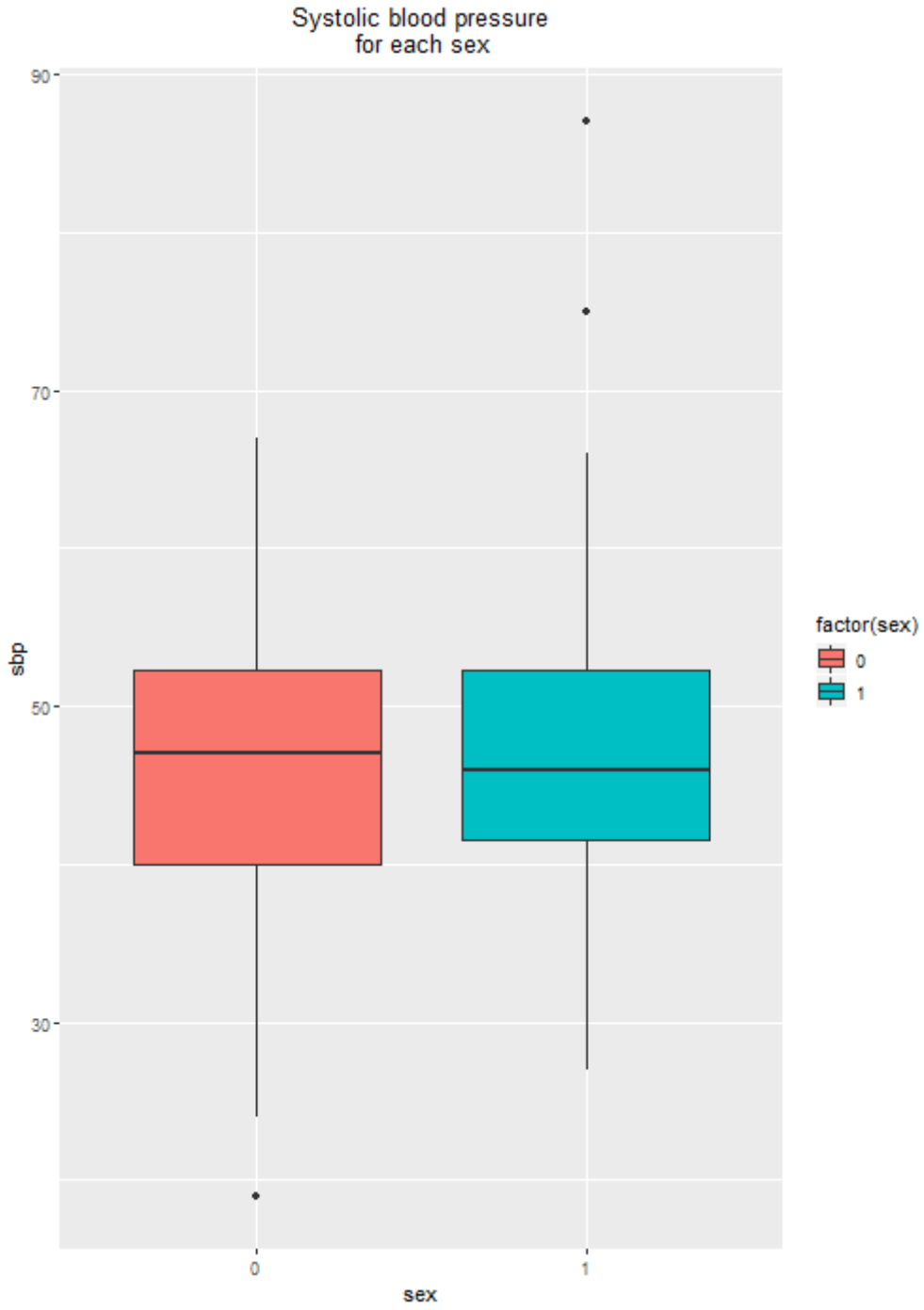
      The answer is `-2.131628e-14` , which is `-0.00000000000002131628`. It is highly approach to 0

- 3-9.
    - (a) Without any step, according to two graphs, Europe's data tend to be concentrate and small, and there isn't any outliers in it. Africa's boxplot shows that it has the largest median.

      I consider Europe has the smallest mean of infant mortality rates, and Africa has the largest median, and Europe has the smallest standard deviation.
    - (b) I will expect mean and median of Africa's mortality rates is approximately equal since seeing the histogram we can say that Africa's data are approximately normal distribution. While Asia's data are total right-skewed, most of the data is concentrated on the left side and there are some outliers on the right side, so the mean of Asia's data may larger than the median.
- 3-11.
    - (a)
        - mean : 87.94
        - median : 86.00
        - standard deviation : 16.00
        - range : 103
        - IQR : 22
    - (b) *Chebyshev's* Inequality tell us that no matter what, for any number that is >= 1, at least of : $[1-(\frac{1}{k})^2]$ measurements in the set of data lie within k standard deviations of their mean.

      For instance, when k=2, at least 75% (or greater) of the value lie within 2 standard deviations of the mean, that is, in this case we can say that there is at least 75% of values is in the range of 55.93-119.95
    - (c) At least 75.0% (347 data) in 2 standard deviations, 88.9% (411 data) in 3 standard deviations, 75% in 2 standard deviations.
    - (d) Since the zinc data isn't well normal distribution. using empirical rule won't be better than Chebyshev's Inequality
- 3-15.

- (a)



We can find out in males' data (number 1), The data seems more concentrated than females' data, but there are some outliers in males' data.

- (b)

	male	female
mean	47.86	46.46
standard deviation	11.81	11.15

According to the data table on the top, we can see that males has a larger mean of sbp 47.86, and also has larger standard deviation 11.81

- (c)

- Coefficient of Variation = (sd/mean)·100%
  `(sd(boy$sbp)/mean(boy$sbp))*100`
  `(sd(girl$sbp)/mean(girl$sbp))*100`

  male's cv = 24.67

  female's cv = 23.99

  Male's coefficient of Variation is slightly larger than female's, so the variability of male is slightly larger female.