Summarizing data

- Summary statistics for central location.
  - Sample mean (樣本平均): average; often denoted by $\bar{X}$.
  - Sample median (樣本中位數): the middle number or the average of the two middle numbers for the sorted data.
  - Sample median is less sensitive to extreme values in the data than the sample mean.
- Example 1. Consider the sample $(2, 7, 3)$. Find the sample mean and the sample median.
  - Sol. The sample mean is
  $$\frac{1}{3}(2 + 7 + 3) = 4.$$
  The sorted data are 2, 3, 7, so the middle number is 3, which is the sample median.
  - Software sol. To find the sample mean and median for the sample $(2, 7, 3)$ using R, at the R prompt, enter
  ```
  x <- c(2,7,3); mean(x); median(x)
  ```
  Then R returns the sample mean and sample median.
- For nominal data, we use mode to describe the central location instead of using sample mean/median.

  Example 2. Students may get to school by various means. Below is a summary table of transportation tools for a class of 30 students, where the coding rule for different tools is as follows: 1 for bus, 2 for feet, 3 for motorcycle and 4 for other.

transportation tool	1	2	3	4
count	10	8	7	5

  The mode for the sample is 1.
- Summary statistics for dispersion. Let $(X_1, \ldots, X_n)$ be a sample with sample mean $\bar{X}$.
  - Mean deviation:
  $$\frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|.$$
  - Sample variance (樣本變異數):
  $$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right).$$

- Sample standard deviation (樣本標準差): $\sqrt{\text{sample variance}}$.
- Range: maximum − minimum

- Example 3. Consider the sample $(2, 7, 3)$. Find the mean deviation, the sample variance and sample standard deviation.
  - Sol. From Example 1, the sample mean for the sample $(2, 7, 3)$ is 4, so the mean deviation is
  $$\frac{1}{3}\left(|2-4|+|7-4|+|3-4|\right) = 2,$$
  the sample variance is
  $$\frac{1}{3-1}\left((2-4)^2+(7-4)^2+(3-4)^2\right) = \frac{1}{3-1}\left(2^2+7^2+3^2-3\times 4^2\right) = 7$$
  and the sample standard deviation is $\sqrt{7} \approx 2.645751$.
  - To find the sample variance and sample standard deviation for the sample $(2, 7, 3)$ using R, at the R prompt, enter

    ```
    x <- c(2,7,3); var(x); sd(x)
    ```

    Then R returns the sample variance and sample standard deviation.

- Chebyshev's Theorem. For a sample $(X_1, \ldots, X_n)$ with sample mean $\bar{X}$ and sample standard deviation S,
$$\frac{1}{n}\left(\text{ number of } X_i\text{'s such that } |X_i - \bar{X}| \le kS\ \right) \ge 1 - \frac{1}{k^2}.$$

- Example 4. Suppose that we have a sample of 1000 exam scores, where the sample mean and sample standard deviation are 75 and 2 respectively. At least what percent of the scores are between 70 and 80?
  - Sol. Note that $(80-75)/2 = 2.5$ and $(70-75)/2 = -2.5$, so the range $75 \pm (2.5)(2)$ is the range from 70 to 80. Take $k = 2.5$ and apply Chebyshev's Theorem, then at least $1 - 1/(2.5)^2 = 84\%$ of the scores are within the range $75 \pm (2.5)(2)$, so at least 84% of the scores are between 70 and 80.

- Example 5. Suppose that we have a sample of 1000 exam scores, where the sample mean and sample standard deviation are 75 and 2 respectively. Find a range that covers at least 80% of the scores.
  - Sol. Solving $1 - 1/k^2 = 0.8$ gives $k = \sqrt{5}$. By Chebyshev's Theorem, at least 80% of the scores are in the range from $75 - 2\sqrt{5} \approx 70.52786$ to $75 + 2\sqrt{5} \approx 79.47214$.

- Histogram construction for a sample $(X_1, \ldots, X_n)$ based on Scott's rule (Scott (1979)). Determine k: the number of classes. Choose k to be the smallest number such that
$$k \ge \frac{\text{range of the sample}}{(24\sqrt{\pi})^{1/3} \cdot S \cdot n^{-1/3}} \approx \frac{\text{range of the sample}}{3.5 \cdot S \cdot n^{-1/3}},$$
where S is the sample standard deviation.

- Example 6. For a sample of size 999 with minimum 15546, maximum 35925 and sample standard deviation 3314.289, determine the number of classes for drawing a histogram using Scott's rule.

  Sol. Choosing the smallest k such that

$$\begin{aligned} k &\geq \frac{35925 - 15546}{(24\sqrt{\pi})^{1/3} \times 3314.289 \times (999)^{-1/3}} \\ &\approx \frac{35925 - 15546}{3.5 \times 3314.289 \times (999)^{-1/3}} \approx 17.6 \end{aligned}$$

  gives $k = 18$.

- Drawing a histogram using R. Suppose that the sample has been generated and stored in a vector x in R by running

```
x <- qnorm(seq(0.001, 1-0.001, 0.001))*20000/6
```

  Below are the R codes for drawing a histogram for x based on Scott's rule.

```
hist(x, nclass="scott")
```

- For a histogram that shows a shape with a unique peak (the mode), we can tell from the histogram

  1. the central location of the data,
  2. the range for most of the data (for example the range for the middle 50% of the data), and
  3. whether the shape is symmetric about the peak.

  If the shape of the histogram is essentially symmetric about the peak, then the mode and the median for the binned data are approximately the same. It is natural to use the peak location as the central location of the data. If the histogram is essentially asymmetric, then the mode and the median are not the same.

- For a histogram that shows more than one peak, we can still tell where most of the data are located from the histogram.

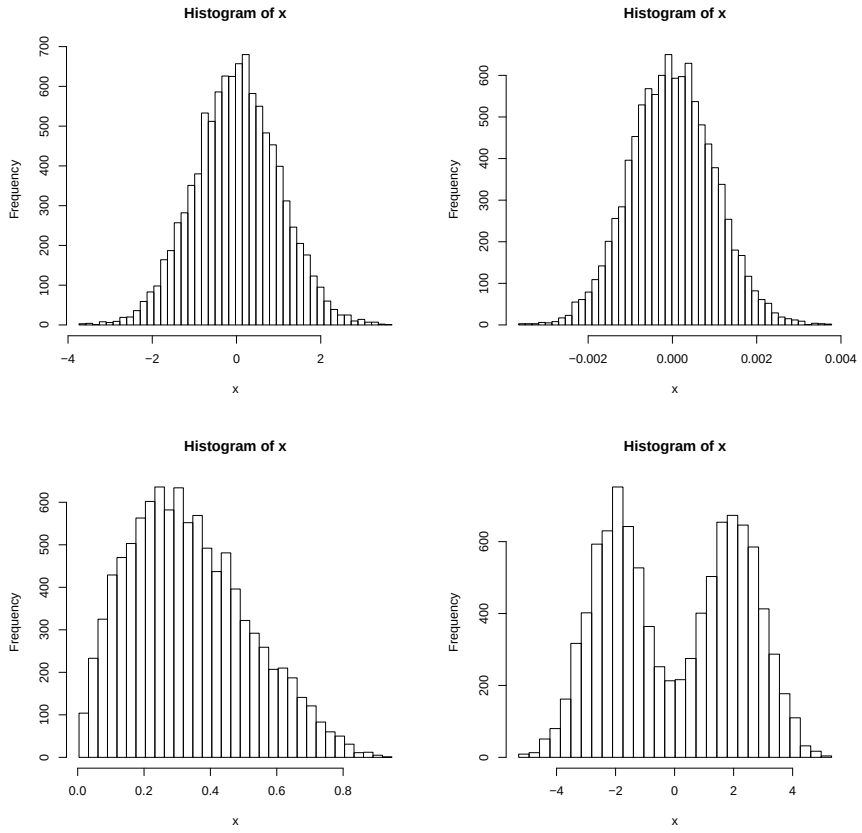
- Try to determine the central location(s) and the range for most of the data for each of the following histogram.

  – Left-upper histogram. Mode and Median: 0. At least 50% of the data are between -1.5 and 1.5. All data are between -4 and 4.

  – Right-upper histogram. Mode and Median: 0. At least 50% of the data are between -0.0015 and 0.0015. All data are between -0.004 and 0.004.

  – Left-bottom histogram. Mode: 0.286. Median > 0.286. At least 50% of the data are between 0.2 and 0.6. All data are between 0 and 1.

  – Right-bottom histogram. Most data are near -2 or 2. At least 25% of the data are between -3 and -1 and at least another 25% of the data are between 1 and 3. All data are between -6 and 6.



References

[1] D. W. Scott, *On optimal and data-based histograms*, Biometrika, 66 (1979), pp. 605–610.