MATH 2411 T1B Tutorial 1

HOU Zhen

zhouah@connect.ust.hk

February 10, 2026

Descriptive Statistics

Types of data/variables

- **Categorical variables**
  - Nominal: cannot be ranked.
    Examples: dog species, UG majors, payment method (cash / Octopus).
  - Ordinal: can be ranked.
    Examples: typhoon signals No. 1/3/8, 1–5 star ratings.
- **Numerical variables**
  - Discrete: counts.
    Examples: number of courses you take, number of siblings.
  - Continuous: measurements.
    Examples: height, weight, time, temperature.

R: categorical data as factors

- In R, categorical variables are declared as **factors** and possible values are **levels**.
- `as.factor()` converts a numeric/character vector into a factor, i.e. it treats the values as labels (categories) instead of numbers for arithmetic.

```
> x = c(1, 2, 3, 4, 5)
> x
[1] 1 2 3 4 5
> y = as.factor(x)
> y
[1] 1 2 3 4 5
Levels: 1 2 3 4 5
```

R: histogram and frequency table

```
> x =
c(970, 612, 1201, 1003, 666, 1088, 744, 898, 964, 1135,
  983, 1016, 1029, 1058, 1085, 1122, 1022, 623, 1197, 883)

> hist(x, 7)

> res = hist(x, 7)
> res$breaks
[1] 600 700 800 900 1000 1100 1200 1300
> res$counts
[1] 3 1 2 3 7 3 1
```

R: line chart with `plot()`

- `plot()` is a generic plotting function in R.
- For time-series style data, we often pass an x-vector (time) and a y-vector (measurements), and set `type = "l"` for a line chart.

```
> temp <- c(18, 19, 21, 20, 22, 24, 23)
> days <- 1:length(temp)
> plot(
+   days, temp, type = "l",
+   xlab = "Day", ylab = "Temperature (°C)",
+   main = "Daily temperature over a week"
+ )
```

R: help and dollar sign

- `help()` gives documentation of a function.
- The $ sign extracts a variable from a list/data frame.

```
> help(hist)
> help("hist")

> res = hist(x, 7)
> res$breaks
> res$counts
```

Center

Definition (sample mean). For data $x_1, x_2, \ldots, x_n$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Definition (sample median). Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the ordered data. Then

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{if } n \text{ is even.} \end{cases}$$

Definition (trimmed mean). The $\alpha\%$ trimmed mean $\bar{x}_{\mathrm{tr}(\alpha)}$ is computed by trimming away the largest $\alpha\%$ and the smallest $\alpha\%$ of the ordered sample values, and then taking the sample mean of the remaining data.

R: sample mean and median

```
> x =
c(970, 612, 1201, 1003, 666, 1088, 744, 898, 964, 1135,
  983, 1016, 1029, 1058, 1085, 1122, 1022, 623, 1197, 883)

> mean(x)
[1] 964.95
> median(x)
[1] 1009.5
```

Sample Variance

Definition (sample variance). For data $x_1, x_2, \ldots, x_n$, the sample variance is given by

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Definition (sample standard deviation). For data $x_1, x_2, \ldots, x_n$, the sample standard deviation is given by

$$s_{n-1} = \sqrt{s_{n-1}^2}.$$

Exercise

Prove that $\bar{x} = \operatorname{argmin}_a \sum_{i=1}^{n} (x_i - a)^2$.

R: sample variance and standard deviation

```
> var(x)
[1] 31790.26

> sd(x)
[1] 178.2982
```

Quartiles

In general, given a set of data, we **sort** the data from the **smallest** to the **largest** and mark the median. We then divide the data into four quarters.

- The 0.25 quantile is the 1st quartile Q_1.
- The 0.50 quantile is the median Q_2.
- The 0.75 quantile is the 3rd quartile Q_3.

Range, IQR and Outliers

Range. The range of a dataset is defined as the difference between the maximum data value and the minimum data value:

$$\text{range} = \text{max} - \text{min}\,.$$

Inter-Quartile Range (IQR). The inter-quartile range of a dataset is defined as the difference between the third quartile and the first quartile:

$$\text{IQR} = Q_3 - Q_1.$$

Outlier. A data point in a dataset is called an outlier if

$$\text{Observation} \geq Q_3 + 1.5 \times \text{IQR}$$

or

$$\text{Observation} \leq Q_1 - 1.5 \times \text{IQR}.$$

Remark: outliers are data that are significantly larger or smaller than the majority of the data.

R: boxplot

- `boxplot()` creates a boxplot using quartiles, range and outliers.
- Example dataset from the lecture.

```
> x =
c(12, 13, 14, 15, 16, 17, 18, 18, 19, 20,
  20, 21, 22, 22, 23, 24, 25, 26, 26, 27,
  28, 29, 29, 30, 31, 32, 33, 34, 35, 42, 47)

> boxplot(x,
          main = "Boxplot of example data",
          ylab = "Value")
```