STA2002: Probability and Statistics II

Introduction and Preliminary

Fangda Song and Ka Wai Tseng

School of Data Science, CUHK(SZ)

September, 2025

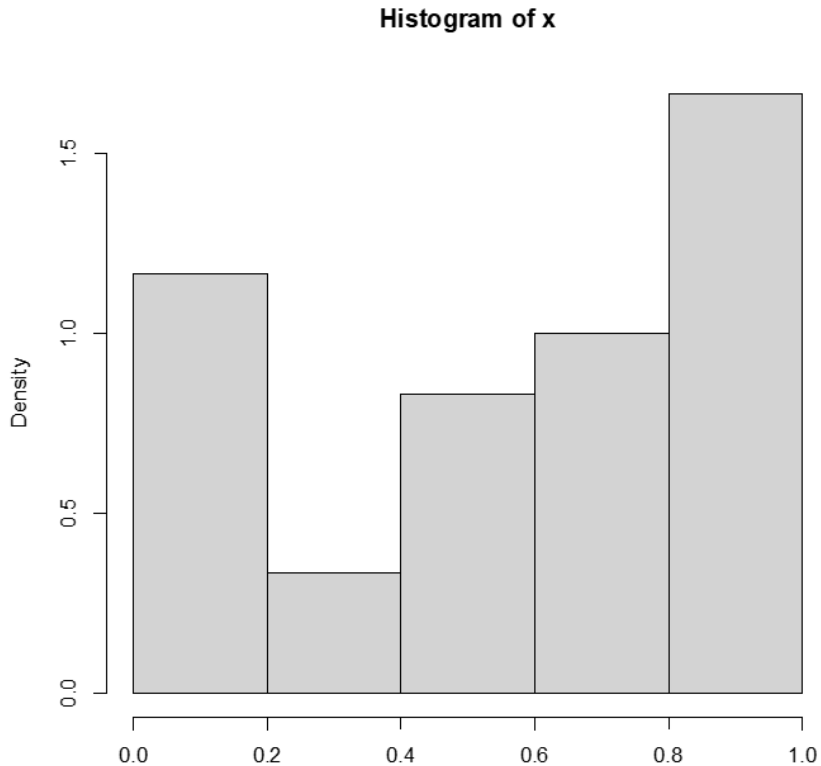
Introduction

In this lecture, we will learn.

- Data visualization
- Explanatory data analysis
- Maximum likelihood estimation (MLE)
- Suggested reading: Chapter 6.1, 6.2 and 6.4 in the book.

Histogram

```
0.97742091 0.99841003 0.48011321
0.08661486 0.58601176 0.61739409
0.65261468 0.38614097 0.87670884
0.73602044 0.16408397 0.02296603
0.43263809 0.64313709 0.31603898
0.90661059 0.67659943 0.70756372
0.92746488 0.81213553 0.95981676
0.84393371 0.45332443 0.13658223
0.03843230 0.83956954 0.03093576
0.50768522 0.81352515 0.11346368
```

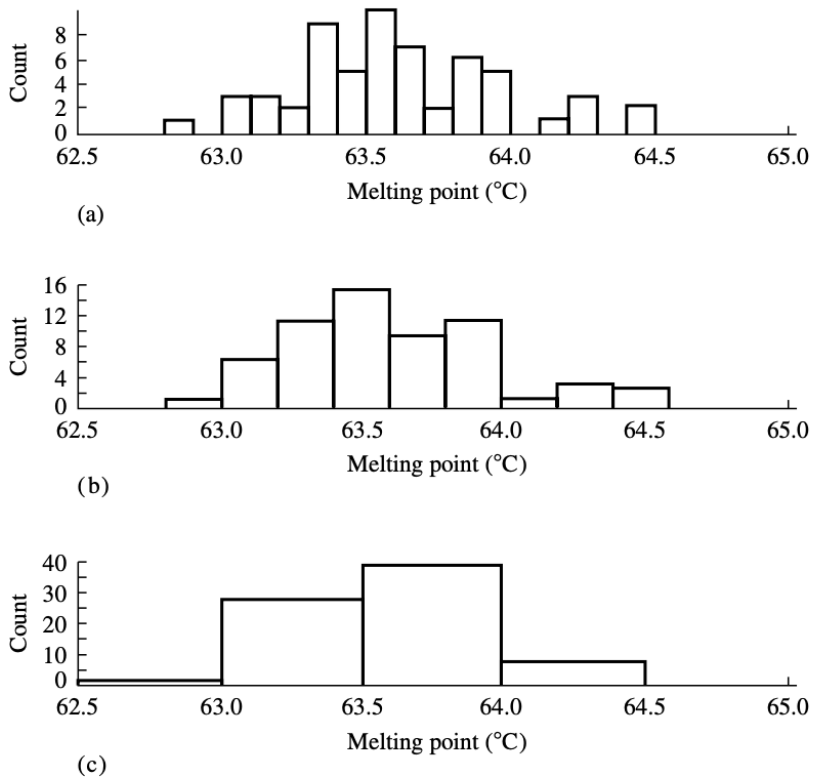


- Visualize continuous observations by grouping them into several classes (bins).
- The **height**: the **relative frequency** or **density** of data points
- The **area**: proportional to the number of observations

Histogram

The choice of bin size (or number of bins) is important.

- **Too small:** Creates artificial peaks and valleys, overfitting the data.
- **Too large:** Hides important details and patterns in the data.



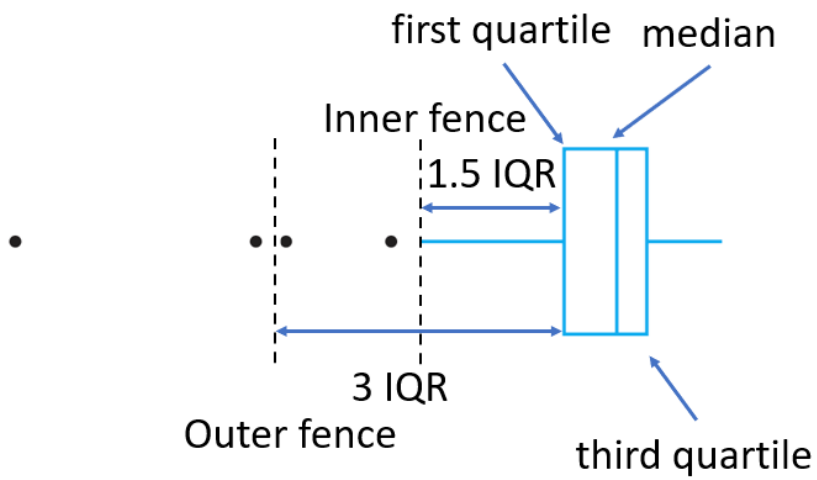
Histograms of melting points of beeswax with bin width equal to (a) 0.1 (b) 0.2 and (c) 0.5, respectively

Most statistical software (e.g., R, Python) can automatically select a reasonable number of bins based on some rules.

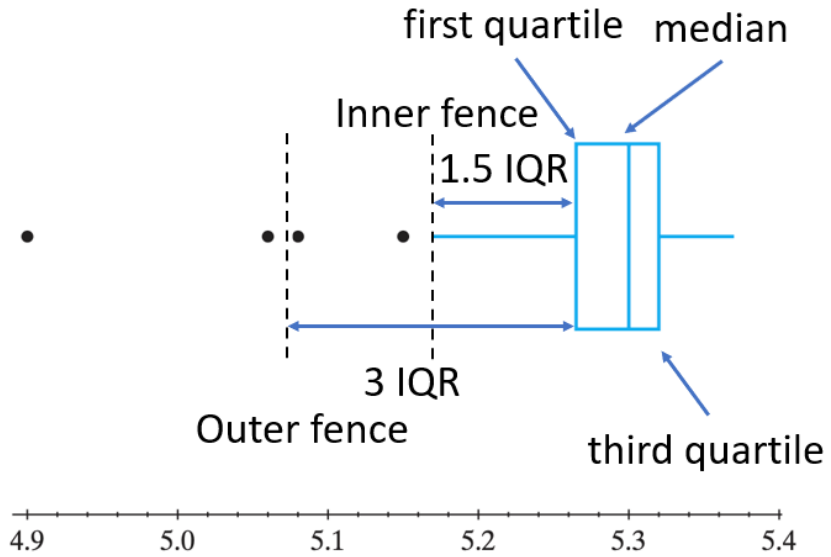
Boxplot (Box-and-Whisker Diagram)

A standardized way of displaying the distribution of data based on a five-number summary

- Minimum, first quartile (Q1), median, third quartile (Q3), maximum
- Interquartile range (IQR): Q3 − Q1
- Inner (Outer) fence: the left or right of the box at a distance of 1.5 × IQR (3 × IQR)
- Lower (Upper) whisker is the larger (smaller) one of lower (upper) inner fence and minimum (maximum)
- Suspected outlier: between the inner and outer fences
- Outlier: beyond the outer fence



Boxplot interpretation



- The **box** shows the middle 50% of the data.
- The **whiskers** show the range of "normal" data points, extending to the points within 1.5×IQR of the box.
- The two dots beyond the outer fence are identified as **outliers**.
- The distribution appears slightly **left-skewed** as the upper whisker is shorter than the lower one.

Boxplot enables us to understand the **variance** and **skewness** of the data and identify **outliers**.

Explanatory data analysis

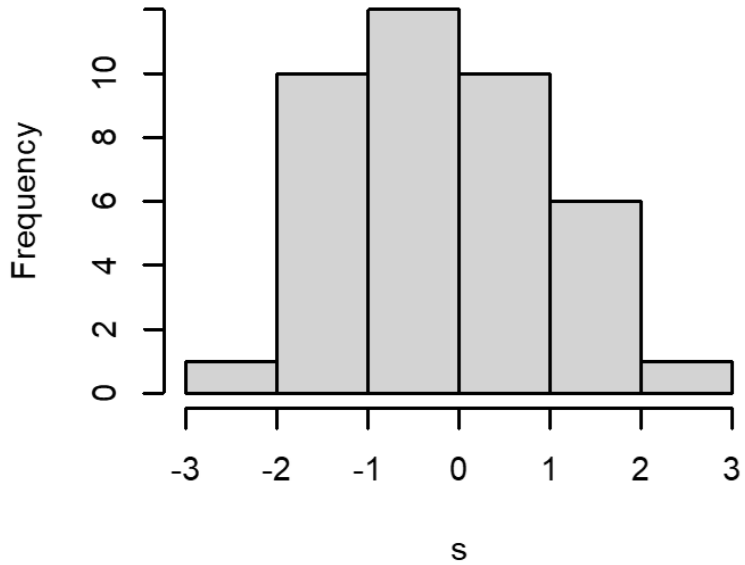
We will analyze real-world data on Covid-19 infections at the University of Miami (UM).

- Use data visualization techniques (histograms, boxplots) to understand the Covid-19 infections at UM
- Suppose that the positive cases for the next year will arrive at the same rate as the last year included in the data. What can you say about the number of positive cases in the next new year?
- Suppose that the infection rate for faculty will remain the same as that in the last year included in the data. What can you say about the number of positive cases in the next year among the 20 faculty/staff members in the Statistics department?

Refer to the separated PDF for details

Parameter Estimation

- Suppose we have a sample with size n from a population
- Histogram tells us that the sample is likely drawn from a normal population $N(\mu, \sigma^2)$
- How to estimate the population parameters (μ, σ^2)?



Assumptions:

- the sample is representative of the population
- the data are normally distributed

Parameter Space

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$ i.i.d. from a population distribution with unknown parameter θ and PMF or PDF $f(x;\theta)$.
- The parameter θ takes values in the **parameter space** Ω.

Example: Recall the normal distribution with PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty.$$

The corresponding parameter space is

$$\Omega = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$$

Estimator

- The observed values of $X_1, X_2, \ldots, X_n$ are denoted as $x_1, x_2, \ldots, x_n$.
- The statistics $u(X_1, X_2, \ldots, X_n)$ to estimate θ is also called an **estimator** of θ.
- The value of $u(x_1, x_2, \ldots, x_n)$ is said to be an **estimate** of θ.
- Since $u(X_1, X_2, \ldots, X_n)$ is a single value **estimator** of θ, it is called a **point estimator**.

Note that X_i is a random variable, while x_i represents the actual value of X_i in a sample

Likelihood Function

Definition (Likelihood function)

The likelihood function $L : \Omega \to [0, \infty)$ is equal to the joint density or joint mass function of the observations. L is treated as a function of the parameter θ, that is,

$$L(\theta) := f(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

The log-likelihood function is defined by

$$\ell(\theta) := \ln L(\theta).$$

Maximum Likelihood Estimator

Definition (Maximum likelihood estimator)

The maximum likelihood estimator (MLE), denoted by $\hat{\theta}$, is the value of θ that maximizes $L(\theta)$.

- To derive the expression of MLE, it is often easier to maximize $\ell(\theta)$ rather than to maximize $L(\theta)$.
- Since the natural logarithm function is a strictly increasing function, the resulting maximizer $\hat{\theta}$ from maximizing $\ell(\theta)$ is the same as that of $L(\theta)$.

Example: Bernoulli(p)

- Suppose $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ with pmf

$$f(x; p) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}, \quad 0 \leq p \leq 1.$$

- Parameter space: $\Omega = \{p : 0 \leq p \leq 1\}$.
- Likelihood:

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}.$$

- Task: Derive the MLE of p.
- Solution: $\hat{p} = \overline{X}$ with $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Example: Exponential(θ)

- $X_1, \ldots, X_n$ i.i.d. exponential with parameter θ:

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad 0 < x < \infty, \quad \theta \in (0, \infty).$$

- Likelihood and log-likelihood:

$$L(\theta) = \theta^{-n} \exp\left(-\frac{\sum_{i=1}^{n} x_i}{\theta} \right), \qquad \ell(\theta) = -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

- Task: Derive the MLE of θ.
- Solution: $\hat{\theta} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Example: Geometric(p)

- $X_1, \ldots, X_n$ i.i.d. geometric with pmf

$$f(x; p) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots, \quad p \in (0, 1).$$

- Likelihood and log-likelihood:

$$L(p) = p^n(1-p)^{\sum_{i=1}^n x_i - n}, \qquad \ell(p) = n \ln p + \Big(\sum_{i=1}^n x_i - n \Big) \ln(1-p).$$

- Task: Derive the MLE of p.
- Solution: $\hat{p} = \dfrac{1}{\overline{X}}, \quad \overline{X} = \dfrac{1}{n} \sum_{i=1}^n X_i.$

Example: Uniform$[0, \theta]$

- $X_1, \ldots, X_n$ i.i.d. uniform on $[0, \theta]$ with pdf

$$f(x; \theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta, \\ 0, & \text{otherwise,} \end{cases} \qquad \theta \in (0, \infty).$$

- Task: Derive the MLE of θ.
- Solution: $\hat{\theta} = X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Why Maximizing the Likelihood Function

Assumption (Regularity Conditions)

Regularity conditions (R1)-(R3) are given by
(R1) The pdfs are distinct: $\theta \neq \theta' \Rightarrow f(x_i;\theta) \neq f(x_i;\theta')$.
(R2) The pdfs have common support for all θ.
(R3) θ_0 is an interior point of Ω.

Theorem

Let θ_0 be the true parameter, under assumptions (R1) - (R3), we have

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}[L(\theta_0;x) > L(\theta;x)] = 1, \quad \forall\, \theta \neq \theta_0.$$

Note: This theorem is not required in this course. For the students who may have interest. To better understand the concept of MLE.

MLE

- The theorem implies that asymptotically the likelihood function L is maximized at the true value θ_0.
- So in considering estimates of θ_0, it is natural to consider the value of θ which maximizes the likelihood.
- Here comes the definition of Maximum Likelihood Estimator (MLE).

Definition (Maximum Likelihood Estimator)

We say that $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is a maximum likelihood estimator (MLE) of θ if

$$\hat{\theta} = \operatorname{argmax} L(\theta; X_1, \ldots, X_n),$$

where the notation argmax means that $L(\theta; X_1, \ldots, X_n)$ achieves its maximum value at $\hat{\theta}$.