

STAT1012 Statistics for Life Sciences

Quick Revision Notes

Fall, 2019

LEUNG Man Fung, Heman

(Reference: lecture and tutorial notes)

Contents

| | |
|---------------------------------|---|
| I) Descriptive Statistics | 3 |
| Central tendency | 3 |
| Dispersion | 3 |
| Graphical methods | 4 |
| II) Probability | 5 |
| Notations | 5 |
| Probability theory | 5 |
| Conditional probability | 5 |

I) Descriptive Statistics

Data type: Qualitative (Special: Categorical), Quantitative (Discrete, Continuous)

Population: the whole set of entities of interest

Sample: a subset of the population

Central tendency

Sample mean: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

Sequential update property: $\bar{X}_n = \frac{1}{n} [(n-1)\bar{X}_{n-1} + X_n]$

Mode: the value which has the greatest number of occurrence (may not be unique)

Median: the “middle” value, or the average of the two values closest to “middle” after sorting

Percentile: the p-th percentile ($V_{\frac{p}{100}}$) is a value such that p% of the data are less than or equal to $V_{\frac{p}{100}}$. In particular, upper quantile = $V_{0.75}$, median = $V_{0.5}$, lower quantile = $V_{0.25}$.

Denote the sorted data by $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ where $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. This is equivalent to saying that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest etc.

Median: $V_{0.5} = X_{(\frac{n+1}{2})}$ if n is odd or $\frac{1}{2} \left[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)} \right]$ if n is even

Percentile: $V_{\frac{p}{100}} = X_{(k)}$ where $k = \text{roundUp} \left(\frac{np}{100} \right)$ if $\frac{np}{100}$ is not an integer.

Otherwise, $V_{\frac{p}{100}} = \frac{1}{2} \left[X_{(\frac{np}{100})} + X_{(\frac{np}{100}+1)} \right]$

Dispersion

Symmetric: the left hand side of the distribution mirrors the right hand side

Unimodal: the mode is unique

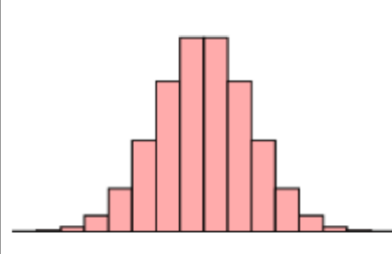
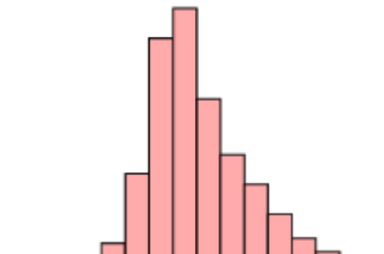
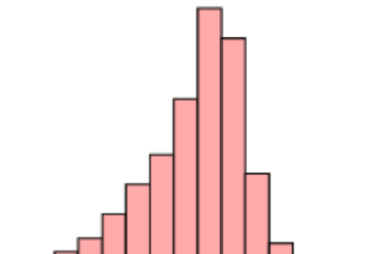
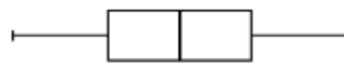

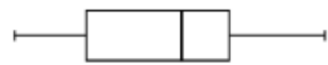
Skewness: measure of asymmetry

Left-skewed (negatively skewed): mean < median, have a few extreme small values

Right-skewed (positively skewed): mean > median, have a few extreme large values

Symmetric → mean = median (converse not true)

Symmetric + unimodal → mean = median = mode (converse not true)

| Symmetric | Skewed right (positive) | Skewed left (negative) |
|---|---|---|
|  |  |  |
|  |  |  |
| Q_1 and Q_3 should be approximately equally spaced from the median (Q_2). | Q_3 is farther from the median(Q_2) than Q_1 | Q_1 is farther from the median(Q_2) than Q_3 |

Range: maximum – minimum ($X_{(n)} - X_{(1)}$)

Interquartile range: $V_{0.75} - V_{0.25}$

Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ or $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i^2 - n\bar{X}^2)$

Sample standard deviation: $SD = \sqrt{S^2}$

[Graphical methods](#)

Bar graph: use for categorical data, show the number of observations in each category

Histogram: use for quantitative data, showing the number of observations in each range

Stem-and-leaf plot: ordered the data into a tree-like structure

Boxplot: show 5 numbers (min, Q_1 , median, Q_3 , max), help locate outliers (As a rule of thumb, some people define outliers as values $> Q_3 + 1.5 \cdot IQR$ or $< Q_1 - 1.5 \cdot IQR$)

II) Probability

Notations

Sample space: the set of all possible outcomes, often denoted as Ω

Outcome: a possible type of occurrence

Event: any set of outcomes of interest, can be denoted as $E \subset \Omega$

Probability (of an event): denoted by $P(E)$, always lies between 0 and 1 (both inclusive)

$$P(E) = \frac{\# \text{ of outcomes in } E}{\# \text{ of outcomes in } \Omega}$$

Union: either A or B occurs, or they both occurs, denoted by $A \cup B$ (logically equivalent to OR)

Intersection: both A and B occur, denoted by $A \cap B$ (logically equivalent to AND)

Complement: A does not occur, denoted by A^C (logically equivalent to NOT)

DeMorgan's laws: $(A \cup B)^C = A^C \cap B^C$, $(A \cap B)^C = A^C \cup B^C$

Probability theory

Mutually exclusive: A and B are mutually exclusive if $P(A \cap B) = 0$ (cannot co-occur)

Independence: $P(A \cap B) = P(A)P(B)$ iff A and B are independent. Their complements (A and B^C ; A^C and B; A^C and B^C) will be pairwise independent as well

Addition law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multiplication law: if A_1, \dots, A_k are mutually independent, then $P(A_1 \cap \dots \cap A_k) = P(A_1) \times \dots \times P(A_k)$

Conditional probability

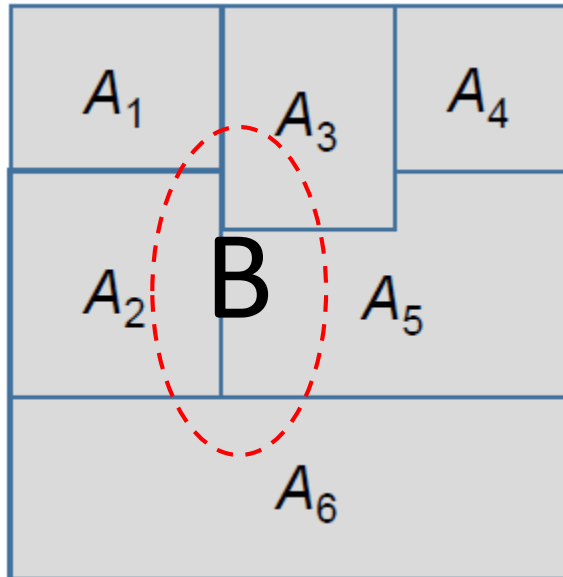
Conditional probability: $P(B|A) = \frac{P(A \cap B)}{P(A)}$, if $P(B|A) = P(B)$, then A and B are independent

Relative risk: $RR(B|A) = \frac{P(B|A)}{P(B|A^C)}$

Total probability rule: $P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$

Exhaustive: if A_1, \dots, A_k are exhaustive, then $A_1 \cup \dots \cup A_k = \Omega$ (at least one of them must occur)

Generalized total probability rule: let A_1, \dots, A_k be mutually exclusive and exhaustive events. For any event B , we have $P(B) = \sum_{i=1}^k P(B|A_i)P(A_i)$



Bayes' theorem: conditional probability + generalized total probability rule. let A_1, \dots, A_k be mutually exclusive and exhaustive events. For any event B ,

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^k P(B|A_j)P(A_j)}$$