STAT1012 Statistics for Life Sciences

Quick Revision Notes

Fall, 2019

LEUNG Man Fung, Heman

(Reference: lecture and tutorial notes)

Contents

I) Descriptive Statistics

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

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 $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. 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 $\frac{1}{2}\left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right]$ if n is even

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

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

Dispersion

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

Unimodal: the mode is unique

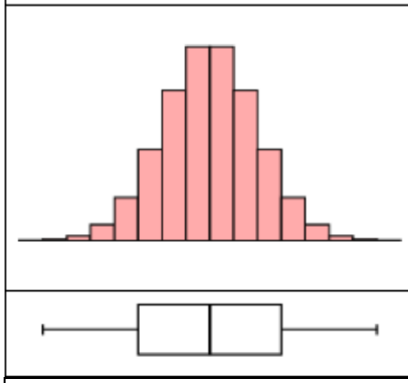
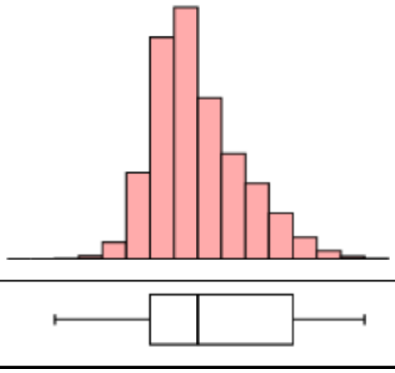
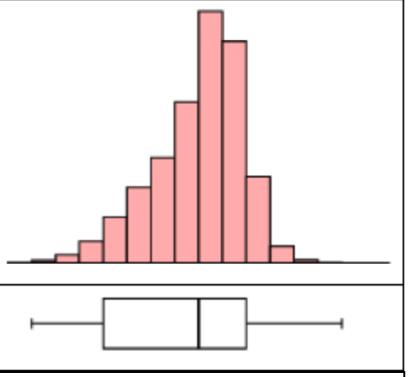
Skewness: measure of asymmetry

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, Q1, median, Q3, max), help locate outliers (As a rule of thumb, some people define outliers as values > Q3 + 1.5*IQR or < Q1 – 1.5*IQR)