STAT1012 Statistics for Life Sciences

Quick Revision Notes

Fall, 2019

LEUNG Man Fung, Heman

(Reference: lecture and tutorial notes)

Contents

[I) Descriptive Statistics 3](#_Toc20912045)

[Central tendency 3](#_Toc20912046)

[Dispersion 3](#_Toc20912047)

[Graphical methods 4](#_Toc20912048)

[II) Probability 5](#_Toc20912049)

[Notations 5](#_Toc20912050)

[Probability theory 5](#_Toc20912051)

[Conditional probability 5](#_Toc20912052)

[III) Discrete Probability Distributions 7](#_Toc20912053)

[Discrete random variables 7](#_Toc20912054)

[Binomial distribution 7](#_Toc20912055)

[Poisson distribution 8](#_Toc20912056)

[IV) Continuous Probability Distribution 9](#_Toc20912057)

[Continuous random variable 9](#_Toc20912058)

[Normal distribution 9](#_Toc20912059)

[V) Point Estimation 11](#_Toc20912060)

# 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:

Sequential update property:

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 () is a value such that p% of the data are less than or equal to . In particular, upper quantile = , median = , lower quantile = .

Denote the sorted data by where . This is equivalent to saying that is the smallest, is the second smallest etc.

Median: if n is odd or if n is even

Percentile: where if is not an integer.

Otherwise,

## 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)



Range: maximum – minimum ()

Interquartile range:

Sample variance: or

Sample standard deviation:

## 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)

# II) Probability

## Notations

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

Outcome: a possible type of occurrence

Event: any set of outcomes of interest, can be denoted as

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

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

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

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

DeMorgan’s laws:

## Probability theory

Mutually exclusive: A and B are mutually exclusive if (cannot co-occur)

Independence: iff A and B are independent. Their complements (A and BC; AC and B; AC and BC) will be pairwise independent as well

Addition law:

Multiplication law: if are mutually independent, then

## Conditional probability

Conditional probability: , if , then A and B are independent

Relative risk:

Total probability rule:

Exhaustive: if are exhaustive, then (at least one of them must occur)

Generalized total probability rule: let be mutually exclusive and exhaustive events. For any event B, we have



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

# III) Discrete Probability Distributions

Random variables: numeric quantities that take different values with specified probabilities

Discrete random variable: a R.V. that takes value from a discrete set of numbers

Continuous random variable: a R.V. that takes value over an interval of numbers

## Discrete random variables

Probability mass function: a pmf assigns a probability to each possible value x of the discrete random variable X, denoted by

(total probability rule)

Cumulative distribution function: a cdf gives the probability that X is less than or equal to the value x, denoted by

Expected value: (the idea is “probability weighted average”)

Variance: , alternatively

Translation/rescale: ,

Linearity of expectation:

## Binomial distribution

Factorial: , note that

Permutation (order is important):

Combination (order is not important): , also denoted as

Binomial distribution: probability distribution on the number of successes in independent experiments, each experiment has a probability of success , then

Pmf: for

Mean:

Variance:

Skewness: right-skewed if p<0.5, symmetric if p=0.5, left-skewed if p>0.5

## Poisson distribution

Poisson distribution: probability distribution on the number of occurrence (usually of a rare event) over a period of time or space with rate , then

Pmf: for

Mean:

Variance:

Skewness: right-skewed

Poisson limit theorem (poisson approximation to binomial): if where and , then where

# IV) Continuous Probability Distribution

## Continuous random variable

Probability density function: a pdf specifies the probability of the random variable falling within a particular range of values, denoted by

, which is the area under the curve from a to b

for all

(total probability rule)

Cumulative distribution function: a cdf gives the probability that X is less than or equal to the value x, denoted by

Expected value:

Variance:

## Normal distribution

Normal distribution: if follows normal distribution with mean and variance , then , often used to represent continuous random variable with unknown distributions

Pdf: for

Shape: bell-shape, symmetric about the mean, unimodal

Standard normal distribution:

Cdf of standard normal: denoted as

by symmetric property

Percentile of standard normal:

Standardization: if , then

De Moivre–Laplace theorem (normal approximation to binomial): if , where . The 0.5s are continuity correction

## Some remarks

Statistical parameter: a numerical characteristic of a statistical population or a statistical model. We are given these numbers (e.g. ) in previous chapters but in reality we do not know these numbers. These lead to the next part of our course: Statistical Inference

Why approximation: you may wonder why we need to approximate binomial probability using poisson/normal. One major reason is that calculating binomial probability involves combination and large factorials are hard/costly to compute in previous centuries

Variance of sum:

Tower rule of expectation:

Law of total variance (EVE):

Sum of poisson: if independently, then

Sum of normal: if independently, then

# V) Point Estimation