STAT1012 Statistics for Life Sciences

Quick Revision Notes

Fall, 2019

LEUNG Man Fung, Heman

(Reference: lecture and tutorial notes)

Contents

[I) Descriptive Statistics 3](#_Toc23345355)

[Central tendency 3](#_Toc23345356)

[Dispersion 3](#_Toc23345357)

[Graphical methods 4](#_Toc23345358)

[II) Probability 5](#_Toc23345359)

[Notations 5](#_Toc23345360)

[Probability theory 5](#_Toc23345361)

[Conditional probability 5](#_Toc23345362)

[III) Discrete Probability Distributions 7](#_Toc23345363)

[Discrete random variables 7](#_Toc23345364)

[Binomial distribution 7](#_Toc23345365)

[Poisson distribution 8](#_Toc23345366)

[Hypergeometric distribution (not required) 8](#_Toc23345367)

[Geometric distribution (not required) 8](#_Toc23345368)

[Negative binomial distribution (not required) 9](#_Toc23345369)

[IV) Continuous Probability Distributions 10](#_Toc23345370)

[Continuous random variables 10](#_Toc23345371)

[Uniform distribution 10](#_Toc23345372)

[Normal distribution 10](#_Toc23345373)

[Some remarks (not required) 11](#_Toc23345374)

[V) Point Estimation 12](#_Toc23345375)

[Sampling 12](#_Toc23345376)

[Point estimation 12](#_Toc23345377)

[Mean 13](#_Toc23345378)

[Variance 13](#_Toc23345379)

[VI) Interval Estimation 14](#_Toc23345380)

# 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)

Commutativity:

Associativity:

Distributive laws:

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

## Hypergeometric distribution (not required)

Hypergeometric distribution: probability distribution on the number of success in trials without replacement, from a finite population of size that contains trials classified as success, then

Pmf: for

Mean:

Variance:

## Geometric distribution (not required)

Geometric distribution: probability distribution on the number of trials when the first success occurs, each trial has a probability of success , then

Pmf: for

Mean:

Variance:

Memoryless: . Geometric distribution is the only discrete distribution with this property

## Negative binomial distribution (not required)

Negative binomial distribution: probability distribution on the number of times when the success occurs, each trial has a probability of success , then

Pmf: for

Mean:

Variance:

# IV) Continuous Probability Distributions

## Continuous random variables

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

(by the fundamental theorem of calculus)

Expected value:

Variance:

(Note: Calculus is NOT required in our course)

## Uniform distribution

Uniform distribution: if follows uniform distribution on the interval , then it has the same probability density at any point in the interval and we denote it by

Pdf: for , otherwise 0

Cdf: for

Mean:

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

Normal approximation to poisson: if where

## Some remarks (not required)

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: 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

Statistical inference: process of drawing conclusions from data that are subject to random variations

Estimation: estimate the values of specific population parameters based on the observed data

Hypothesis testing: test on whether the value of a population parameter is equal to some specific value based on the observed data

## Sampling

Sample: the data obtained after the experiments are performed, usually denoted by

Random sample: the data before the experiments are performed, usually denoted by

Non-probability sample: some elements of the population have no chance of being selected

Probability sample: all elements in the population has known nonzero chance to be selected

Simple random sample: all elements in the population has the same probability to be selected

Systematic sample: elements are selected at regular intervals through certain order

Stratified sample: all elements are classified into different stratums and each stratum is sampled as an independent sub-population

Cluster sample: all elements are divided into different clusters and a simple random sample of clusters is selected

Coverage error: exists if some groups are excluded from the frame and have no chance of being selected

Non-response error: people who do not respond may be different from those who do respond

Measurement error: due to weaknesses in question design, respondent error, and interviewer’s impact on the respondent

Sampling error: Chance (luck of the draw) variation from sample to sample

## Point estimation

Point estimator: a rule for calculating a single value to “best guess” an unknown population parameter of interest based on the observed data

(Note: estimator is random, estimate is fixed, estimand is the unknown parameter)

Unbiasedness:

Minimum variance:

Independent and identically distributed (i.i.d.): an assumption where the random variables are sampled such that they are independent and follows the same distribution

Central limit theorem (CLT, Lindeberg–Lévy): Let be i.i.d. random variables with mean and finite variance , then as n tends to infinity,

## Mean

Estimand:

Sample mean (estimator):

Expectation: (unbiased)

Variance: (by i.i.d.)

Distribution: . If , then this follows from the fact that sum of independent normal is normal (remarks in section IV). If follows some other distribution, then this follows from the CLT when n is large (usually ≥30)

## Variance

Estimand:

Sample variance (estimator):

Expectation: (unbiased, proof is not required)

Variance: (not required)

Distribution: (not required)

# VI) Interval Estimation