STAT1012 Statistics for Life Sciences

Quick Revision Notes

Spring, 2020

LEUNG Man Fung, Heman

(Reference: lecture and tutorial notes)

Contents

[I) Descriptive Statistics 4](#_Toc44424101)

[Central tendency 4](#_Toc44424102)

[Dispersion 4](#_Toc44424103)

[Graphical methods 5](#_Toc44424104)

[II) Probability 6](#_Toc44424105)

[Notations 6](#_Toc44424106)

[Probability theory 6](#_Toc44424107)

[Conditional probability 7](#_Toc44424108)

[III) Discrete Probability Distributions 8](#_Toc44424109)

[Discrete random variables 8](#_Toc44424110)

[Binomial distribution 8](#_Toc44424111)

[Poisson distribution 9](#_Toc44424112)

[Hypergeometric distribution (not required) 9](#_Toc44424113)

[Geometric distribution (not required) 9](#_Toc44424114)

[Negative binomial distribution (not required) 10](#_Toc44424115)

[IV) Continuous Probability Distributions 11](#_Toc44424116)

[Continuous random variables 11](#_Toc44424117)

[Uniform distribution 11](#_Toc44424118)

[Normal distribution 12](#_Toc44424119)

[Some remarks (not required) 13](#_Toc44424120)

[V) Point Estimation 14](#_Toc44424121)

[Sampling 14](#_Toc44424122)

[Point estimator 14](#_Toc44424123)

[Mean 15](#_Toc44424124)

[Variance 16](#_Toc44424125)

[Binomial proportion 16](#_Toc44424126)

[Poisson rate 16](#_Toc44424127)

[VI) Interval Estimation 17](#_Toc44424128)

[Confidence interval 17](#_Toc44424129)

[Mean 17](#_Toc44424130)

[Variance 18](#_Toc44424131)

[Binomial proportion 18](#_Toc44424132)

[Poisson rate 18](#_Toc44424133)

[VII) Hypothesis Testing 20](#_Toc44424134)

[Terminologies 20](#_Toc44424135)

[One-sample z-test 21](#_Toc44424136)

[One-sample t-test 21](#_Toc44424137)

[One-sample chi-squared test 22](#_Toc44424138)

[One-sample binomial proportion test 22](#_Toc44424139)

[Some remarks (not required) 22](#_Toc44424140)

[VIII) Extension (not required) 23](#_Toc44424141)

[Terminologies 23](#_Toc44424142)

[Difference of mean, two dependent samples 23](#_Toc44424143)

[Difference of mean, two independent samples 24](#_Toc44424144)

[Difference of proportion, two independent samples 24](#_Toc44424145)

[Ratio of variance, two independent samples 25](#_Toc44424146)

# 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

Mutual independence: iff A, B and C are mutually independent

Mutual independence does not imply pairwise, vice versa

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: (the idea is “probability weighted distance from mean”)

Alternatively

Translation/rescale: ,

Linearity of expectation:

Variance of sum under independence: if are independent

## 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 a 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: (the idea is “probability weighted average”)

Variance: (the idea is “probability weighted distance from mean”)

(Note: Calculus is NOT required in our course)

Translation/rescale: ,

Linearity of expectation:

Variance of sum under independence: if are independent

## 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

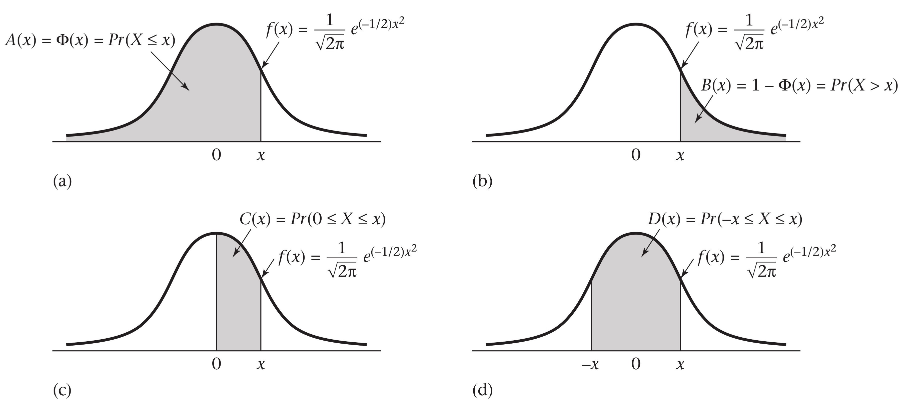
Standard normal distribution:

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

Condition for good approximation:

Normal approximation to poisson: if where

Condition for good approximation:

## 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

Square of standard normal: if , the

Sum of chi square: if , 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 estimator

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

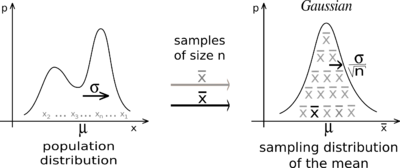
(Note: estimator is random, estimate is fixed, estimand is the unknown parameter)

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 (>30 in practice),

[[1]](#footnote-1)

## 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). Otherwise () we have  , where is a Student’s t-distribution with degree of freedom n-1.

## Variance

Estimand:

Sample variance (estimator): , if is known

Expectation: (unbiased)

Variance: (not required)

Distribution: (right-skewed)

## Binomial proportion

Estimand: where (similar to mean case)

Estimator:

Expectation: (unbiased)

Variance: (by i.i.d.)

Distribution: because the sampling distribution is binomial. For or , normal approximation gives

## Poisson rate

Estimand: where with as the total number of units

Estimator:

Expectation: (unbiased)

Variance: (by i.i.d.)

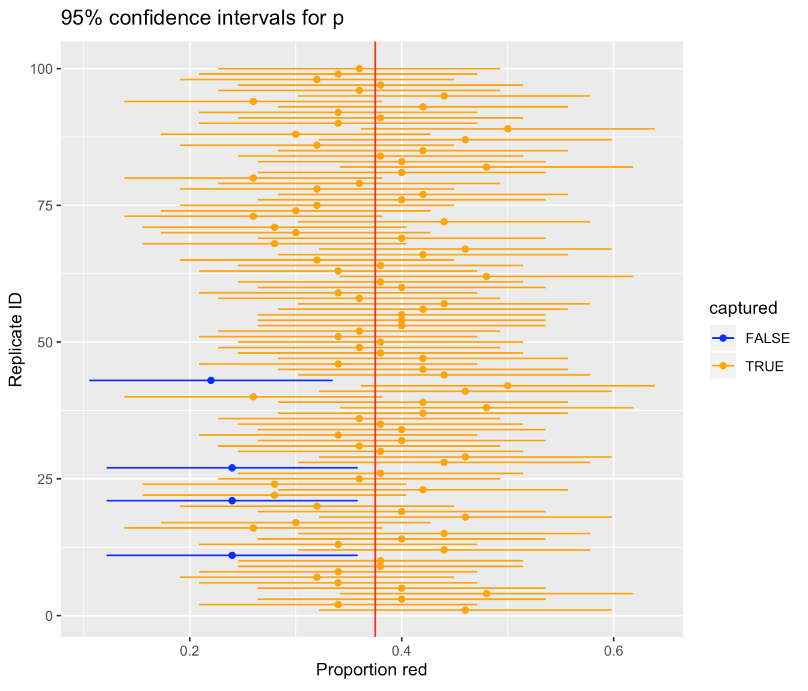
Distribution: because the sampling distribution is Poisson. For or , normal approximation gives

# VI) Interval Estimation

## Confidence interval

Confidence interval: an interval associated with a confidence level that may contain the true value of an unknown population parameter

Meaning of confidence level: in the long run, of all the confidence intervals that can be constructed will contain the unknown true parameter (NOT the probability that an interval will contain the parameter)

[[2]](#footnote-2)

Elements of confidence interval: , where is the point estimate, is the critical value from an asymptotic distribution under the confidence level , is the standard error of the point estimate

## Mean

Confidence interval ( is known):

Confidence interval ( is unknown, ):

Confidence interval ( is unknown, ): (differs in degree of freedom)

Margin of error: (width is which helps determine sample size)

Critical values: standard normal and t-distribution are symmetric around 0

Common normal critical value:

One-sided confidence interval: or

(Note: this is essentially adjusting the critical value, which arises naturally when we are not interested in the other bound, e.g. weight > 0 so negative lower bound is not interested)

## Variance

Confidence interval ( is unknown):

Confidence interval ( is known): (differs in d.f.)

Critical values: chi-squared distribution is not symmetric, so cannot simplify

## Binomial proportion

Confidence interval ( or ):

(Note: the standard error here is an approximated version from the lecture notes)

Confidence interval (exact method): solve for from where

## Poisson rate

Confidence interval (exact method): solve for from where

Confidence interval (bootstrap method): generate sample of size with replacement from . Calculate the point estimate from each bootstrap sample. Sort the means and the bootstrap confidence interval is given by the corresponding percentiles.

(Note: bootstrap is a very powerful method which can be applied to many statistical problems that do not require close form)

# VII) Hypothesis Testing

## Terminologies

Statistical hypothesis: a claim (assumption) about a population parameter

Null hypothesis: , the hypothesis to be tested (default position)

Alternative hypothesis: , a hypothesis challenge (against) (what we want to conclude)

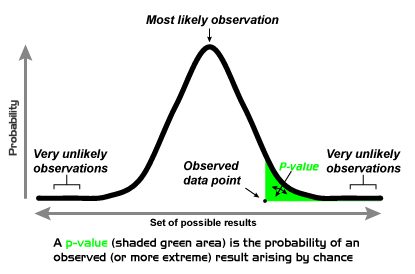
Hypothesis testing: a procedure to make decision on hypothesis based on some data samples. The idea is to assume is true first. If the population under is unlikely to generate the data sample, then we can make a decision to reject (and thus accept ).

Test statistics: a quantity (statistics) derived from the sample to help perform hypothesis test

Level of significance: , defines the unlikely value of the sample if is true

Critical value: cutoff values from the distribution of test statistic under given

p-value: probability of obtaining a test statistics at least as extreme as the observed sample value given is true

[[3]](#footnote-3)

“Accept the null hypothesis”: if we fail to reject , we cannot accept it because doing so violates the idea of prove by contradiction. It is possible that is not true but we have not collected enough data to reject it

Type I error: , reject when is true (false positive).

(Note: traditional statistical procedure controls type I error by the level of significance, so that’s why both of them are )

Type II error: , do not reject when is false (false negative)

|  |  |  |
| --- | --- | --- |
|  | is true | is false |
| Do not reject | Correct inference  (true negative, probability = 1-α) | Type II error  (false negative, probability = β) |
| Reject | Type I error  (false positive, probability = α) | Correct inference  (true positive, probability = 1-β) |

Duality of confidence interval with hypothesis test: is rejected at significance level if and only if the corresponding confidence interval does not contain the value claimed by with confidence level (true for common tests)

## One-sample z-test

Assumption: known , from normal distribution or of large size ()

Hypothesis: (1) or (2) or (3)

Test statistics: , under null

Note: the capital is not typo but indicates that it is random

Decision rule: reject if (1) ; (2) ; (3)

p-value: reject if where (1) ; (2); (3)

## One-sample t-test

Assumption: unknown

Hypothesis: (1) or (2) or (3)

Test statistics: , under null

Decision rule: reject if (1) ; (2) ; (3)

## One-sample chi-squared test

Assumption: unknown , from normal distribution

Hypothesis: (1) or (2) or (3)

Test statistics: , under null

Decision rule: reject if (1) or ; (2) ; (3)

## One-sample binomial proportion test

Assumption: binomial sample with or

Hypothesis: (1) or (2) or (3)

Test statistics: , under null

Decision rule: reject if (1) ; (2) ; (3)

## Some remarks (not required)

Power: . As higher power implies a lower type II error, traditional procedures usually fix the type I error and search for tests with high power

Bayesian inference: most procedures in this course are frequentist procedures. Taking interval estimation as an example, if we want our interval to have probability covering the unknown parameter, we should seek credible interval from Bayesian inference instead (confidence interval does not guarantee that). Consider taking more courses from our department if you are interested :)

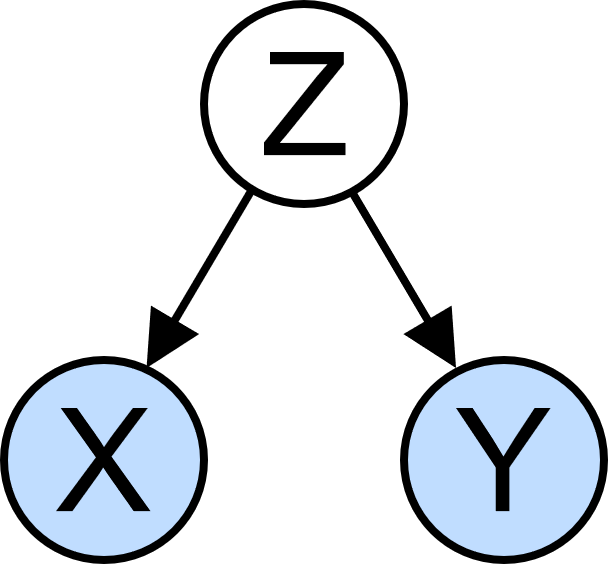
# VIII) Extension (not required)

## Terminologies

Longitudinal study: repeated observations of the same variables over a period of time

Cross-sectional study: observations from a population at a specific point in time

Confounder: a variable that influences both the dependent variable and independent variable

[[4]](#footnote-4)

## Difference of mean, two dependent samples

Assumption: both from normal, of large size () or difference is approximately normal

Hypothesis: (1) or (2) or (3) where

Z-test (known ): ,   under null

Decision rule: reject if (1) ; (2) ; (3)

T-test (unknown ): ,   under null,

Decision rule: reject if (1) ; (2) ; (3)

Confidence interval ( is known):

Confidence interval ( is unknown, ):

Confidence interval ( is unknown, ): (differs in degree of freedom)

Very similar to one-sample case due to duality of CI and testing

## Difference of mean, two independent samples

Assumption: both from normal or of large size ( though can be different)

Hypothesis: (1) or (2) or (3) where

Z-test (known ): ,   under null

Decision rule: reject if (1) ; (2) ; (3)

Approximate z-test (unknown but ): ,  ,   under null

Decision rule: reject if (1) ; (2) ; (3)

T-test (unknown ; ; both from normal): ,   under null

Pooled variance estimate: , by assumption

Decision rule: reject if (1) ; (2) ; (3)

T-test (unknown ; ; both from normal) ): ,   under null

Satterthwaite’s method: , which should fall between

Decision rule: reject if (1) ; (2) ; (3)

Behrens-Fisher problem: when but are unknown, how to test ?

Obviously Satterthwaite’s method is one possible solution but it may not be best

## Difference of proportion, two independent samples

Assumption: binomial sample with or

Hypothesis: (1) or (2) or (3) where

Test statistics: , under null,

Decision rule: reject if (1) ; (2) ; (3)

## Ratio of variance, two independent samples

Assumption: unknown and both from normal independently

Hypothesis: (1) or (2) or (3)

Test statistics: , under null

Decision rule: reject if (1) or ; (2) ; (3)

1. By Mathieu ROUAUD - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=60066898 [↑](#footnote-ref-1)
2. By Chester Ismay and Albert Y. Kim from Ch9 Confidence Intervals of Statistical Inference via Data Science [↑](#footnote-ref-2)
3. By Repapetilto - Adobe IllustratorPreviously published: Unpublished, CC BY-SA 3.0, https://en.wikipedia.org/w/index.php?curid=35569621 [↑](#footnote-ref-3)
4. By طاها - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31197358 [↑](#footnote-ref-4)