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)

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 $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. This is equivalent to saying that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest etc.

Median: $V_{0.5} = X_{\left(\frac{n+1}{2}\right)}$ 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

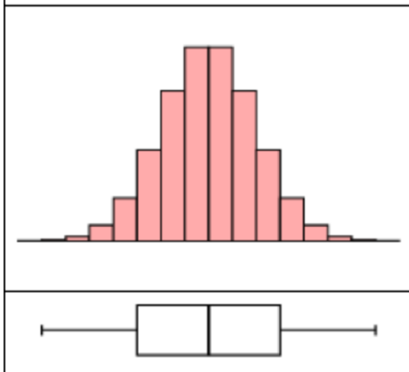
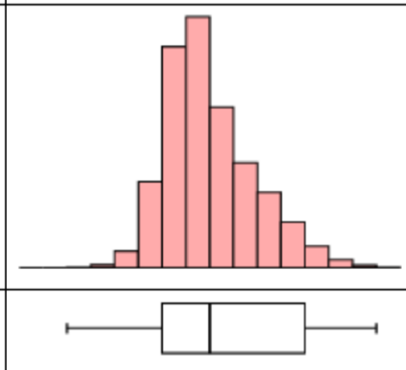
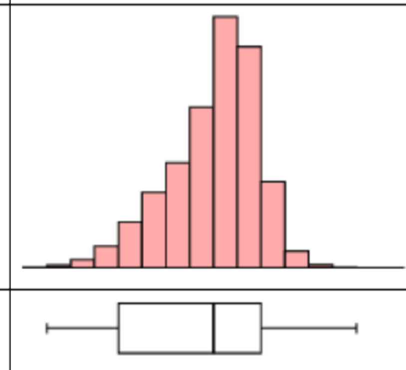
Skewness: measure of asymmetry

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)

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)

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

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)

Commutativity: $A \cup B = B \cup A,\ A \cap B = B \cap A$

Associativity: $(A \cup B) \cup C = A \cup (B \cup C),\ (A \cap B) \cap C = A \cap (B \cap C)$

Distributive laws: $(A \cap B) \cup C = (A \cup C) \cap (B \cup C),\ (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

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 $P(A_! \cap \dots \cap A_k) = P(A_1) \times \dots \times P(A_k)$

Conditional probability

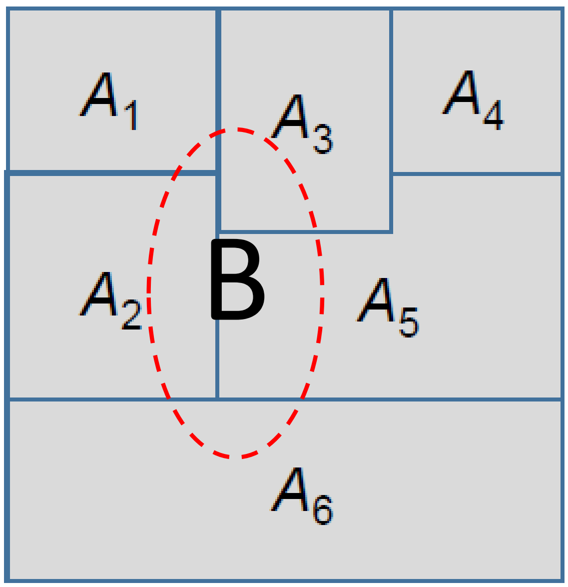
Conditional probability: $P(B|A) = \frac{P(A \cap B)}{P(A)}$, if $P(B|A) = P(B)$, then A and B are independent

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

Generalized total probability rule: let $A_1, \ldots, A_k$ 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 $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. For any event B,

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

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 $f(x) = P(X = x)$

$\sum_{i=1}^{n} f(x_i) = 1$ (total probability rule)

Cumulative distribution function: a cdf gives the probability that X is less than or equal to the value x, denoted by $F(x) = P(X \leq x)$

Expected value: $\mu = E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$ (the idea is "probability weighted average")

Variance: $\sigma^2 = Var(X) = \sum_{i=1}^{n}(x_i - \mu)^2 P(X = x_i)$, alternatively $Var(X) = E(X^2) - [E(X)]^2$

Translation/rescale: $E(aX + b) = aE(X) + b$, $Var(aX + b) = a^2 Var(X)$

Linearity of expectation: $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$

Binomial distribution

Factorial: $n! = n \times (n-1) \times \ldots \times 1$, note that $0! = 1$

Permutation (order is important): $P_k^n = \frac{n!}{(n-k)!}$

Combination (order is not important): $C_k^n = \frac{n!}{k!(n-k)!}$, also denoted as $\binom{n}{k}$

Binomial distribution: probability distribution on the number of successes X in n independent experiments, each experiment has a probability of success p, then $X \sim B(n, p)$

Pmf: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$

Mean: $E(X) = np$

Variance: $Var(X) = np(1-p)$

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 X (usually of a rare event) over a period of time or space with rate μ, then $X \sim Po(\mu)$

Pmf: $P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$ for $x = 0, 1, 2, \ldots$

Mean: $E(X) = \mu$

Variance: $Var(X) = \mu$

Skewness: right-skewed

Poisson limit theorem (poisson approximation to binomial): if $X \sim B(n, p)$ where $n \geq 20,\ p < 0.1$ and $np < 5$, then $X \approx Y \sim Po(\mu)$ where $\mu = np$

Hypergeometric distribution (not required)

Hypergeometric distribution: probability distribution on the number of success X in n trials without replacement, from a finite population of size $N_1 + N_2 = N \geq n$ that contains N_1 trials classified as success, then $X \sim Hypergeometric(N_1, N_2, n)$

Pmf: $P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$ for $x = \max(0, n - N_2), \ldots, \min(n, N_1)$

Mean: $E(X) = n\left(\frac{N_1}{N}\right)$

Variance: $Var(X) = n\left(\frac{N_1}{N}\right)\left(\frac{N_2}{N}\right)\left(\frac{N-n}{N-1}\right)$

Geometric distribution (not required)

Geometric distribution: probability distribution on the number of trials X when the first success occurs, each trial has a probability of success p, then $X \sim Geo(p)$

Pmf: $P(X = x) = (1-p)^{x-1}p$ for $x = 1, 2, \ldots$

Mean: $E(X) = \frac{1}{p}$

Variance: $Var(X) = \frac{1-p}{p^2}$

Memoryless: $P(X > k + j | X > k) = P(X > j)$. 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 X when the r success occurs, each trial has a probability of success p, then $X \sim NB(r, p)$

Pmf: $P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$ for $x = r, r+1, \ldots$

Mean: $E(X) = \frac{r}{p}$

Variance: $Var(X) = \frac{r(1-p)}{p^2}$

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 $f(x)$

$P(a \leq X \leq b) = \int_a^b f(x)dx$, which is the area under the curve from a to b

$P(X = a) = \int_a^a f(x)dx = 0$ for all a

$\int_{-\infty}^{\infty} f(x)dx = 1$ (total probability rule)

Cumulative distribution function: a cdf gives the probability that X is less than or equal to the value x, denoted by $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$

$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$ (by the fundamental theorem of calculus)

Expected value: $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

Variance: $\sigma^2 = Var(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

(Note: Calculus is NOT required in our course)

Uniform distribution

Uniform distribution: if X follows uniform distribution on the interval $[a, b]$, then it has the same probability density at any point in the interval and we denote it by $X \sim U(a, b)$

Pdf: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, otherwise 0

Cdf: $F(x) = \int_a^x \frac{1}{b-a} dt = \left[\frac{t}{b-a}\right]_a^x = \frac{x-a}{b-a}$ for $a \leq x \leq b$

Mean: $E(X) = \frac{a+b}{2}$

Variance: $Var(X) = \frac{(b-a)^2}{12}$

Normal distribution

Normal distribution: if X follows normal distribution with mean μ and variance σ^2, then $X \sim N(\mu, \sigma^2)$, often used to represent continuous random variable with unknown distributions

Pdf: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ for $-\infty < x < \infty$

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

Standard normal distribution: $Z \sim N(0,1)$

Cdf of standard normal: denoted as $\Phi(z) = P(Z \le z)$

$P(a \le Z \le b) = P(Z \le b) - P(Z \le a) = \Phi(b) - \Phi(a)$

$\Phi(-z) = 1 - \Phi(z)$ by symmetric property

Percentile of standard normal: $\Phi(1.645) = 0.95, \Phi(1.96) = 0.975$

Standardization: if $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$

$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

De Moivre–Laplace theorem (normal approximation to binomial): if $X \sim B(n,p)$, $P(a < X < b) \approx P(a + 0.5 \le Y \le b - 0.5)$ where $Y \sim N(np, np(1-p))$. The 0.5s are continuity correction

Normal approximation to poisson: if $X \sim Po(\lambda), P(X \le a) \approx P(Y \le a + 0.5)$ where $Y \sim N(\lambda, \lambda)$

Some remarks (not required)

Statistical parameter: a numerical characteristic of a statistical population or a statistical model. We are given these numbers (e.g. p, λ, μ) 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: $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Tower rule of expectation: $E(X) = E[E(X|Y)]$

Law of total variance (EVE): $Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$

Sum of poisson: if $X \sim Po(\lambda_1), Y \sim Po(\lambda_2)$ independently, then $X + Y \sim Po(\lambda_1 + \lambda_2)$

Sum of normal: if $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ independently, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Square of standard normal: if $X \sim N(\mu, \sigma^2)$, the $Z^2 = \left[\frac{X-\mu}{\sigma}\right]^2 \sim \chi_1^2$

Sum of chi square: if $X \sim \chi_n^2, Y \sim \chi_m^2$, then $X + Y \sim \chi_{n+m}^2$

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 $x_1, \dots, x_n$

Random sample: the data before the experiments are performed, usually denoted by $X_1, \dots, X_n$

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 $\hat{\theta}(X)$ is random, estimate $\hat{\theta}(x)$ is fixed, estimand θ is the unknown parameter)

Unbiasedness: $E(\hat{\theta}) = \theta$

Minimum variance: $Var(\hat{\theta}) \leq Var(\tilde{\theta})\ \forall\, \tilde{\theta} \in \Theta$

Independent and identically distributed (i.i.d.): an assumption where the random variables $X_1, \dots, X_n$ are sampled such that they are independent and follows the same distribution

Central limit theorem (CLT, Lindeberg–Lévy): Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and finite variance σ^2, then as n tends to infinity (>30 in practice), $\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$

Mean

Estimand: $\theta = \mu = E(X)$

Sample mean (estimator): $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Expectation: $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu$ (unbiased)

Variance: $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$ (by i.i.d.)

Distribution: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, then this follows from the fact that sum of independent normal is normal (remarks in section IV). If $X_1, \dots, X_n$ follows some other distribution, then this follows from the CLT when n is large (usually >30)

Variance

Estimand: $\theta = \sigma^2 = Var(X)$

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

Expectation: $E(S^2) = \sigma^2$ (unbiased)

Variance: $Var(S^2) = \frac{2\sigma^4}{n-1}$ (not required)

Distribution: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$ (right-skewed)

VI) Interval Estimation

Confidence interval

Confidence interval: an interval associated with a confidence level $1-\alpha$ that may contain the true value of an unknown population parameter

Meaning of confidence level: in the long run, $100(1-\alpha)\%$ 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)

Elements of confidence interval: $\left\{\hat{\theta}, c_\alpha, se\left(\hat{\theta}\right)\right\}$, where $\hat{\theta}$ is the point estimate, c_a is the critical value from an asymptotic distribution under the confidence level $1-\alpha$, $se(\hat{\theta})$ is the standard error of the point estimate

Mean

Confidence interval (σ is known): $\bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

Confidence interval (σ is unknown, $n > 30$): $\bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

Confidence interval (σ is unknown, $n \leq 30$): $\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

Margin of error: $E = c_\alpha \times se\left(\hat{\theta}\right)$ (width is $2E$ which helps determine sample size)

Critical values: standard normal and t-distribution are symmetric around 0 $\Rightarrow c_{1-\frac{\alpha}{2}} = c_{\frac{\alpha}{2}}$

Variance

Confidence interval (μ is known): $\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}\right)$

Confidence interval (μ is unknown): $\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$

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