STAT1012 POST-MIDTERM REVIEW

LEUNG MAN FUNG, HEMAN

SPRING, 2020

TOPIC SO FAR…

Central tendency
- Mean, mode, median
- Quartile, percentile

Dispersion
- Variance, SD
- Range, IQR, skewness

Graphical methods
- Bar graph, histogram
- Stem-and-leaf, boxplot

Ch1 Descriptive Statistics

Notation
- Union, intersect, complement
- DeMorgan's laws

Probability theory
- Mutually exclusive
- Independence
- Conditional probability, relative risk
- Total probability rule, exhaustive
- Bayes' theorem

Ch2 Probability

Discrete random variables
- Pmf, cdf
- Total probability rule
- Expectation, variance

Binomial distribution

Poisson distribution
- poisson approximation to binomial

Ch3 Discrete Probability Distributions

Continuous random variables
- Pdf, cdf
- Total probability rule
- Expectation, variance

Uniform distribution

Normal distribution
- Standardization
- Normal probability table
- Normal approximation to binomial
- Normal approximation to poisson

Ch4 Continuous Probability Distributions

WHAT IS IMPORTANT…

Ch1 Descriptive Statistics

Central tendency
- **Mean**, mode, median
- Quartile, percentile

Dispersion
- **Variance, SD**
- Range, IQR, skewness

Graphical methods
- Bar graph, **histogram**
- Stem-and-leaf, **boxplot**

Ch2 Probability

Notation
- Union, intersect, complement
- DeMorgan's laws

Probability theory
- Mutually exclusive
- **Independence**
- **Conditional probability, relative risk**
- **Total probability rule, exhaustive**
- **Bayes' theorem**

Ch3 Discrete Probability Distributions

Discrete random variables
- **Pmf, cdf**
- Total probability rule
- **Expectation, variance**

Binomial distribution
- **when to apply**

Poisson distribution
- **when to apply**
- poisson approximation to binomial

Ch4 Continuous Probability Distributions

Continuous random variables
- Pdf, cdf
- Total probability rule
- Expectation, variance

Uniform distribution

Normal distribution
- **Standardization**
- **Normal probability table**
- Normal approximation to binomial
- Normal approximation to poisson

AGENDA

- Sample and population
- Independence
- Bayes' theorem
- Discrete random variables
- Continuous random variables
- Normal distribution
- When to apply X distribution?
- Q&A

SAMPLE AND POPULATION

- Population: the whole set of entities of interest
  - Example 1: all Hong Kong citizen (Census)
  - Example 2: STAT1012 student in 2020 Spring
  - Example 3: current student taken STAT1012
- Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ (N is the population size)
- Variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

- Sample: a subset of the population
  - Example 1: 1000 randomly selected HK citizen
  - Example 2: STAT1012 year 1 student in 2020 Spring
  - Example 3: current student taken STAT1012 in 2019 Fall
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (n much smaller than N usually)
- Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - $\frac{1}{n-1}$ is the bias correction (to be taught in Ch5)

INDEPENDENCE

- We use event (A, B and C) for illustration
  - The idea of independence can be extended to other concepts like random variable etc.
- Pairwise independence
  - $P(A \cap B) = P(A) \times P(B) \Leftrightarrow A, B$ are pairwise independent
  - Same for the pair A, C and the pair B, C
- Mutual independence
  - $P(A \cap B \cap C) = P(A) \times P(B) \times P(C) \Leftrightarrow A, B, C$ are mutually independent
  - Extend independence to more than 2 events
  - Mutual independence does not imply pairwise independence, vice versa
- Independence is used as an assumption in sampling only later

BAYES' THEOREM

- $P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}$
- Why important?
  - Allow you to model the data with your belief (professional judgement) B
  - Basis of Bayesian inference (will not be covered in this course)
  - See quick revision notes section VII remarks
- Will not appear in later chapters but you should know this term
  - Example: people used Bayesian method to search MH730

DISCRETE RANDOM VARIABLES

- 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
  - Hence the values and probabilities of a discrete r.v. can be tabulated
- Probability mass function: a pmf assigns a probability to each possible value x of the discrete random variable X
  - Denoted by $f(x) = P(X = x)$
  - Attempt to relate values and probabilities of the table via a function
- Expected value: $\mu = E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$
  - The idea is "probability weighted average"
  - Population notation μ is used as you know everything about the r.v. X from the table/pmf
- Variance: $\sigma^2 = Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i)$
  - The idea is "probability weighted distance from mean"
  - Alternatively $Var(X) = E(X^2) - [E(X)]^2$

CONTINUOUS RANDOM VARIABLES

- Probability density function: a pdf specifies the probability of a r.v. falling within a particular range of values
  - denoted by $f(x)$
  - Attempt to relate probability via the area under pdf as pointwise probability is 0
    - $P(a \leq X \leq b) =$ the area under the pdf curve from a to b
    - $P(X = a) = 0$ for all a (in contrast this is what discrete r.v. use)
- Why do we teach the above even when calculus is not required?
  - For applying normal distribution
    - Visual “area under the pdf curve” when you check the normal probability table
    - Distinguish continuous r.v. from discrete as the later use pointwise definition
  - To give you basic concept if you want to further study statistics

NORMAL DISTRIBUTION

- Standard normal distribution: $Z \sim N(0,1)$
- Standardization: if $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$
- Normal probability table
  - Check probability when the lower and upper bounds are known
    - $P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$
  - Check the critical value when the probability is known (used in later chapters)
- Normal approximation: example of the central limit theorem (to be taught in Ch5)

WHEN TO APPLY X DISTRIBUTION?

- Binomial distribution
  - Number of success in a fixed number of independence trials (binary proportion)
  - Example: number of student answering yes in question 1 in a test (fixed total number of attendees)
- Poisson distribution
  - Number of occurrence over a fixed time/space (rate)
  - Example: number of customer getting into a supermarket in an hour (fixed time as an hour and space as supermarket)
- Uniform distribution
  - Fair outcome
  - Example: fair coin toss, fair dice thrown
- Normal distribution
  - Statistical inference due to central limit theorem (to be taught in Ch5)

Q&A

ALL MODELS ARE WRONG, BUT SOME ARE USEFUL