Week 1

- Probability = How likely something is to happen
  $$= \frac{\text{The number of events (A) that might happen}}{\text{Total number of event that might happen}}$$

- Statistics = Branch of mathematics dealing with the way of collecting, processing, analyzing, and presenting data conclusions.

- Statistics
  - Descriptive = Using data gathered on a group to describe or reach conclusions about that same group.
    - → Data collection, data presentation, characterize data
  - Inferential = Gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken.

- Population = All items of interest
  Characteristic-nya = Parameters

- Sample = Subset of population
  Representative of the population
  Randomly selected

- Sampling method
  - Random sampling : S By random
  - Stratified sampling : S By grouping data into several level
  - Systematic sampling : P By particular order, S By random
  - Cluster sampling : P By several parts/clusters, S By random

- Nominal ⊆ Ordinal ⊆ Interval ⊆ Ratio

- Data types
  - Categorical (Quali)
  - Numerical (Quanti)

- Measurement level
  - Nominal : Categorical, no order
    - ↳ address, name
  - Ordinal : Categorical, some order
    - ↳ dislike, neutral, like
  - Interval : Numerical, no true zero
    - ↳ temperature
  - Ratio : Numerical, has true zero
    - ↳ exam score

- Data collection techniques
  - Interview
  - Questionnaire
  - Observation
  - Test and objective scale
  - Projective method

Week 2

- Quantitative data graphs
  - Histogram
  - Frequency polygon
  - Ogive
  - Dot plot
  - Stem-and-leaf plot
  - Time series plot
  - Scatter plot

A. Histogram

- Series of contigous bars / rectangles that represent the frequency of data in given class intervals.

B. Frequency Polygon

- Each class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments.

C. Ogive / Cumulative Frequency Polygon

- Most useful when the decision maker wants to see running tools.

D. Dot Plot

- Useful for observing the overall shape of the distribution of data points along with identifying data values or intervals for which there are groupings and gaps in the data.

E. Stem-and-leaf Plot

- Constructed by seperating the digits for each number of the data into 2 groups.

- Leftmost digits = Stem = Higher valued digits
- Rightmost digits = Leaf = Lower values

F. Time Series Plot

- Used to visualize data sets against time

G. Scatter Plot

- Type of plot / mathematical diagram using Cartesian coordinates to display values for typically 2 variables for a set of data.

- Qualitative Data Graphs:
  - Pie charts
  - Bar charts
  - Pareto charts

A. Pie Charts

- A circle = 100% data, and slices of pie = Percentage breakdown of the sublevels
- Shows the relative magnitudes of the parts to the whole.

B. Bar Charts

- Contains 2 / more categories along 1 axis and a series of bars, one for each category, along the other axis.
- Length of bar = Magnitude of the measure for each category

C. Pareto Charts

- Produce vertical bar chart that display the most common types of defects, ranked in order of occurrence from left to right.

Week 3

- Random variables = Formed by assigning a numerical value to each outcome in the sample space of a particular experiment

- Continuous random variable = Take any value within a continuous intervals, take uncountably many values

- Discrete random variable = Used to express variables in discrete values finite values
  - e.g.:
    - Byk siswa dpt A
    - Byk mobil msk parking lot from 8-9am
    - Byk sisua bawa laptop to class

- p.m.f = Probability mass function (cari p_i utk setiap nilai x_i dgn random variable)
  = A set of probability values p_i assigned to each of the values x_i taken by the discrete random variable

- Syarat p.m.f:
  - $P(X = x_i) = p_i$
  - $0 \leq p_i \leq 1$
  - $\sum_i p_i = 1$

- c.d.f = Cumulative distribution function
  = Alternative way of specifying the probabilistic properties of a random variable X

- Syarat c.d.f:
  - $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy$
  - $F(x) = \sum_{y:y \leq x} p(X = y)$
  - $P(X = x) = F(x) - F(x^-)$
  - $P(a \leq x \leq b) = P(x \leq b) - P(x \leq a) = F(b) - F(a)$
  - $\lim_{x \to -\infty} F(x) = 0$ & $\lim_{x \to \infty} F(x) = 1$
  - $f(x) = \frac{dF(x)}{dx}$

- Bottom step = $F(x^-)$
  Top step = $F(x)$
  Height of step = $P(X=x)$

- Continuous random variables = Express variables whose values are continuous.
  arbitrary number
  ↳ Intinya ada interval

- The propabilistic properties of a continuous random variable defined through probability density function

- p.d.f = Probability density function

- Syarat p.d.f :
  - $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - $\int_{\text{state space}} f(x)\,dx = 1$
  - $f(x) \geq 0$

- Expectation = Mean = For summary = Average value of random variable

- $E(x) = \sum_i p_i x_i$ = Expected value of a discrete random variable with p.m.f $P(X=x_i)=p_i$

- $E(x) = \int_{\text{state space}} x \times f(x)dx$ = Expected value of a continuous random variable with p.d.f

- If a continuous random variable X has a p.d.f $f(x)$ that is symmetric about a point μ, so that $f(\mu+x)=f(\mu-x), \forall x \in R$, then $E(x)=\mu$, so that $E(x)$ = Point of symmetry.

- If a continuous random variable X has a p.d.f $f(x)$ that is symmetric about a point μ, then both the $Q_2 \wedge E(x)=\mu$

- Median / 2nd quartile : $F(x)=0.5$
  1st quartile : $F(x)=0.25$
  3rd quartile : $F(x)=0.75$

- $Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$
- Variance always positive
- Larger values of variance = Greater spread in distribution of random variable about the μ value
- Variance = Spread / Variability in the values taken by random variable.
- The pth quantile of a random variable X with a c.d.f $F(x)$ is defined : $F(x) = p$
  - First quartile $\Rightarrow p = 0.25 \rightarrow F(x) = 0.25$
  - Third quartile $\Rightarrow p = 0.75 \rightarrow F(x) = 0.75$

  } Find x
- A linear function of random variables X is another random variable $Y = aX + b$ for some numbers $a, b \in$ Real.
  - $E(Y) = a\,E(X) + b$
  - $Var(Y) = a^2\,Var(X)$

Linear Combinations of Random Variables

- Consider now a sequence of random variables *X*1, . . . , *Xn* together with some constants *a*1, . . . , *an* and *b*, and define a new random variable *Y* to be the *linear* combination

$$Y = a_1X_1 + \cdots + a_nX_n + b$$

- The **expectation and variance of the linear combination** is given as

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$$

$$\text{Var}(a_1X_1 + \cdots + a_nX_n + b) = a_1^2\,\text{Var}(X_1) + \cdots + a_n^2\,\text{Var}(X_n)$$

- Note : assumed that they are *independent* random variables.

Week 4

9) Binomial / Bernoulli Distributions

- A Bernoulli random variable with parameter p, $0 \leq p \leq 1$ with $P(X=0) = 1-p$ and $P(X=1) = p$, then $E(X) = p$, $Var(X) = p(1-p)$

- Consider an experiment consisting of:
  - n Bernoulli trials
  - independent
  - each have constant probability of success

  , then total number of success X is a random variable that has a binomial distribution with parameter n & p, written:

$$X \sim B(n,p)$$

- p.m.f of $B(n,p)$ random variable:

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

$$E(x) = np \qquad Var(X) = np(1-p)$$

9) Multinomial Distribution

- Mirip binomial distribution tapi bisa banyak outcome, tidak seperti binomial yang hanya 2 outcomes.

- p.m.f of multinomial distribution:

$$P(X = x_1, \ldots, X_k = k) = \frac{n!}{x_1! \ldots x_k!} \cdot p_1^{x_1} \cdot \ldots \cdot p_k^{x_k}, \text{ where } x_1 + x_2 + \ldots + x_k = n, \; p_1 + p_2 + \ldots + p_k = 1$$

$$E(x_i) = np_i \qquad Var(X_i) = np_i(1-p_i)$$

Poisson Distribution

- A random variable X distributed as Poisson random variable with parameter λ, written:

$$X \sim P(\lambda)$$

- p.m.f of Poisson distribution:

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}, \quad e = 2.71828\ldots, \text{ for } x = 0,1,2,3,\ldots$$

- Useful to model the number of times that a certain event occurs per unit of time, distance, or volume, and has mean, var equal to parameter λ.

Week 5

Uniform Distributions

- Random variable X with a flat p.d.f between 2 points a and b, so that:

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b, \quad \text{written} \quad X \sim U(a,b)$$

- The c.d.f is: $F(x) = \frac{x-a}{b-a}$ $\quad E(x) = \frac{a+b}{2}$ $\quad Var(X) = \frac{(b-a)^2}{12}$

- pth quantile: $(1-p)a + pb$
- interquartile range: $(b-a)/2$

Exponentials Distributions

- An exponential distribution with parameter $\lambda > 0$ has p.d.f:

$$f(x) = \lambda \cdot e^{-\lambda x}, \text{ for } x \geq 0, \text{ and } f(x) = 0 \text{ for } x < 0$$

- The c.d.f is: $F(x) = 1 - e^{-\lambda x}$
- Useful for modelling failure times & waiting times
- $E(x) = \frac{1}{\lambda}$ $\quad Var(x) = \frac{1}{\lambda^2}$

Gaussian / Normal Distribution

- The p.d.f is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, for $-\infty \leq x \leq \infty$
- $E(x) = \mu$ $\quad Var(X) = \sigma^2$
- Bell-shaped curve, written $X \sim N(\mu, \sigma^2)$

Standard Normal Distributions

- A normal distribution with mean $\mu = 0$ and $\sigma^2 = 1$. Its p.d.f has notation $\phi(x)$, is :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$$

- If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

- $P(a \leq x \leq b) = \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$

- Mean = Median

Correlation :

$$r = r_{xy} = \frac{n \cdot \sum_{i=1}^{n} X_i y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Covariance :

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{x})}{N}$$

Variance :

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}$$

Kurtosis :

$$\alpha_4 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n \cdot s^4}$$

Standard deviation :

$$\sigma = \sqrt{Varians}$$

Skewness :

$$\alpha = \frac{\bar{x} - modus}{\sigma} = \frac{3(\bar{x} - median)}{\sigma}$$

Ganjil

$$\begin{cases} Q_1 = \frac{1}{4}(n+1) \\ Q_2 = \frac{1}{2}(n+1) \\ Q_3 = \frac{3}{4}(n+1) \end{cases}$$

Genap

$$\begin{cases} Q_1 = \frac{1}{4}(n+2) \\ Q_2 = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \\ Q_3 = \frac{1}{4}(3n+2) \end{cases}$$

Interquartile Range = Hamparan

$$= H = IQR = x_{Q_3} - x_{Q_1}$$