Notes on Probability and Statistics

30.003 Probability and Statistics, Term 4 2019

Wei Min Cher

05 Jan 2020

Contents

1 W1: Probability and Statistics

1.1 Definitions

- Population: well defined collection of objects
- Sample: subset of population selected in certain manner
- Variable: any characteristic whose value may change from one object to another in population

- Probability: properties of populations known, question regarding sample taken from population are investigated **(deductive reasoning)**
- Statistics: characteristics of sample known from experiments, conclusions regarding population are made **(inductive reasoning)**

$$\begin{tikzcd} \text{Population} \arrow[r, bend left, "\text{Probability}"] & \text{Sample} \arrow[l, bend left, "\text{Statistics}"] \end{tikzcd}$$

- Descriptive statistics: techniques to describe a sample/population
- Inferential statistics: making predictions or inferences about population from observations and analyses of sample

1.2 Frequency

- Frequency: number of times value occurs in data set
- Relative frequency: fraction or proportion of times the value occurs

1.3 Range and mean

- Range: difference between largest and smallest sample values
- Mean: average of all values

- Population mean is denoted by μ
- Sample mean is denoted by $\bar{x}$, where

$$\bar{x} = \frac{\sum x_i}{n}, \text{ and } n \text{ denoting the number of data points}$$

1.4 Variance and standard deviation

- Variance: measures variability of data set
- Population variance is denoted by σ^2, where

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\,, \text{ and } N \text{ denoting the size of the population}$$

- Sample variance is denoted by s^2, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\,, \text{ and } n \text{ denoting the size of the sample}$$

- Standard deviation is denoted as σ for population variance and s for sample variance, and is calculated either by:

$$\sigma = \sqrt{\sigma^2}, \text{ or } s = \sqrt{s^2}$$

  where σ^2 is the population variance and s^2 is the sample variance

  - Shortcut to calculate population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2$$

1.5 Linear transformation of sample

Let $x_1, x_2, \ldots, x_n$ be a sample, with a and b being constants. If $y_i = ax_i + b$ is a linear transformation of x_i for $i = 1, 2, \ldots, n$, then

$$\bar{y} = a\bar{x} + b$$
$$s_y^2 = a^2 s_x^2$$

1.6 Median

- Median: the middle value in a data set
- Population median $\tilde{\mu}$

$$\tilde{\mu} = \begin{cases} x_m & N \text{ odd, } m = \frac{N+1}{2}; \\ \dfrac{x_m + x_{m+1}}{2} & N \text{ even, } m = \frac{N}{2}; \end{cases}$$

- Sample median $\tilde{x}$

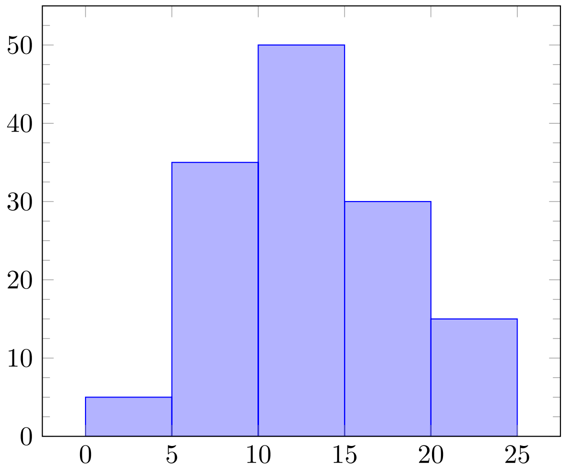
$$\tilde{x} = \begin{cases} x_m & n \text{ odd, } m = \frac{n+1}{2}; \\ \dfrac{x_m + x_{m+1}}{2} & n \text{ even, } m = \frac{n}{2}; \end{cases}$$

1.7 Percentage and percentile

- Percentage: number specifying proportion
- Percentile
  - value below which a given percentage of observations falls
  - data set is ordered as $x'_1 \leq x'_2 \leq \cdots \leq x'_n$,
    where x'_1 and x'_n are the smallest and largest data values respectively
  - x'_i corresponds to the $\frac{100(i-0.5)}{n}$th percentile

1.8 Histogram

- A graphical representation of the distribution of data



1.9 Sample space and events

- Sample space: the set of all possible outcomes of an experiment
  1. Collectively exhaustive
     - Contain all possible outcomes
  2. Mutually exclusive
     - Each outcome in sample space should be unique
- Event: collection of outcomes contained in sample space Ω
  1. Simple event: exactly one outcome e.g. *value of die rolled*
  2. Compound event: > 1 outcome e.g. *event that outcome is even*

1.10 Sample Space vs Population

- Sample space: contains mutually exclusive events
- Population: events can repeat many times

1.11 Set Theory

- Complement of event A, A^c: set if outcomes in Ω that are not in A
- Intersection of 2 events A and B, $A \cap B$: all outcomes that are in A and B
- Union of 2 events A and B, $A \cup B$: all outcomes that are either in A or B
- Null event, $\varnothing$: event with no outcome

- Events A and B are mutually exclusive/disjoint if $A \cap B = \varnothing$
- Events $A_1, A_2, A_3, \ldots$ are mutually exclusive (or pairwise disjoint) if no 2 events have any outcome in common

1.12 De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$
$$A \cup B = A + B - A \cap B$$

- P(A): probability that event A will occur

1.13 Axiom of Probability

1. For any event A, P(A) $\leq$ 0.
2. P(Ω) = 1
3. Any infinite collection of mutually exclusive/disjoint events $A_1, A_2, A_3, \ldots, A_n$ satisfies

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = \sum_{i=1}^{\infty} P(A_i)$$

1.14 Properties of Probability

- For any event A, $P(A) + P(A^c) = 1$ **OR** $P(A) = 1 - P(A^c)$.
- $P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$
  $\because$ A and A^c are disjoint
- For any event A, $P(A) \leq 1$.
- For a null event $\varnothing$, $P(\varnothing) = 0$
  - Does **NOT** suggest A = $\varnothing$
- Similarly, P(A) = 1 does **NOT** suggest A = Ω

1.15 Equally likely outcomes

$$P(\text{equally likely event}) = \frac{1}{n}, \text{ where n is the number of equally likely events}$$

1.16 Simple and compound events

- Simple event: Find out how many outcomes in sample space
- Compound event: Find out how many outcomes in event

2 W1: Counting Technique

2.1 Finding probability

- Computing probability → counting

$$P(A) = \frac{N(A)}{N}$$

  - where $N(A)$ is the number of outcomes for event A,
    and N is the number of outcomes in the sample space

2.2 Tuple

- Group of k elements: k-tuple
- The 1^{st} element is selected in n_1 ways; the 2^{nd} element is selected in n_2 ways; the k^{th} element is selected in n_k ways; such that *the elements are selected independently.*

2.3 Permutation

- Ordered subset
- Number of permutations of size k formed from n objects:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

2.4 Combination

- Unordered subset of a group
- Number of combinations of size k formed from n objects:

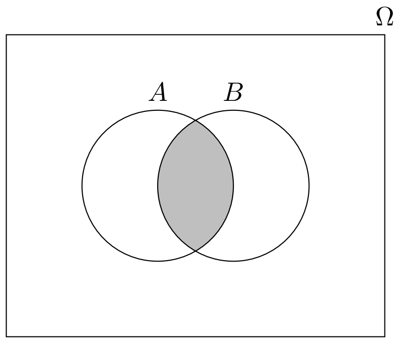
$$\binom{n}{k} \text{ or } C_{k,n} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

- Disregards the different outcomes due to order

3 W2: Conditional Probability

- Probability of event A given that event B has occurred: P(A|B)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



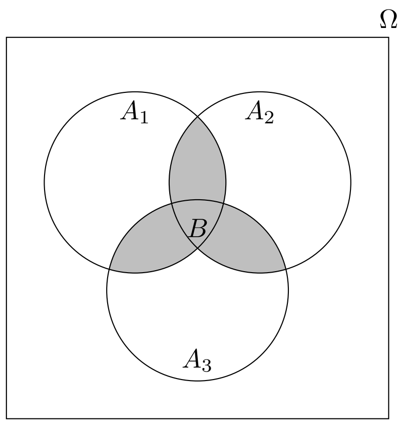
3.1 Law of Total Probability

- Events $A_1, A_2, \ldots, A_k$ are exhaustive if one A_i must occur, i.e. $A_1 \cup A_2 \cup \ldots \cup A_k = \Omega$.

- Let $A_1, A_2, \ldots, A_k$ be mutually exclusive and exhaustive events.

  For any other event B,

$$P(B) = \sum_{i=1}^{k} P(B \mid A_i) P(A_i)$$



3.2 Bayes' Theorem

- Let $A_1, A_2, \ldots, A_k$ be mutually exclusive and exhaustive events with prior unconditional probabilities $P(A_i), i = 1, 2, \ldots, k$

- For any other event B with P(B) > 0, the conditional posterior probability of A_j given that B has occurred is

$$\begin{aligned} P(A_j \mid B) &= \frac{P(A_j \cap B)}{P(B)} \\ &= \frac{P(B \cap A_j)}{P(B)} \\ &= \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)} \end{aligned}$$

3.3 Independence of Random Variables

- Independence: occurrence/non-occurrence of one event has no bearing on the chance that the other will occur
  - $P(A \mid B) = P(A)$: A and B are independent
  - $P(A \mid B) \neq P(A)$: A and B are not independent
- Independence of A and B also implies $P(B \mid A) = P(B)$ if $P(A) > 0$

3.3.1 Multiplication Rule

- A and B are independent iff. $P(A \cap B) = P(A)P(B)$

3.3.2 Independence of several events

- Events $A_1, A_2, \ldots, A_n$ are mutually independent if for every $k \in \{2, 3, \ldots n\}$ and every subset of indices $i_1, i_2, \ldots, i_k$:

$$P(A_{i1} \cap A_{i2} \cap \ldots \cap A_{ik}) = P(A_{i1})P(A_{i2}) \ldots P(A_{ik})$$

- Events are mutually independent if probability of the intersection of any subset of the n events is equal to the product of the individual probabilities.

3.3.3 Disjoint and independent events

- Disjointness: set-theory concept
  - Sets of each group of outcomes share nothing in common
- Independence: probability concept
  - Event is not influenced by the outcome of another event

4 W2: Discrete Random Variable

4.1 Random Variable (RV)

- Random variable (RV): a variable depending on outcomes of a random phenomenon
- Discrete RV: possible values make up a finite set or "countable" in finite set
- Continuous RV: possible values make up an infinite set
- Bernoulli RV: any RV whose only possible values are 0 and 1

4.2 Probability Mass Function (PMF) for Discrete RV

- Known as probability mass function (pmf)
  - e.g. $p(0) = \frac{1}{8},\ p(1) = \frac{3}{8},\ p(2) = \frac{3}{8},\ p(3) = \frac{1}{8}$
- Completely describes probabilistic properties of RV X
- For any pmf, $p(x) \leq 0$ and $\sum_{\text{all possible x}} p(x) = 1$

4.3 Parameter of probability distribution

- Possible value(s) which p(x) depends on
- Different value(s) determine a different probability distribution
- Collection of all probability distributions for different parameters: *family of probability distributions*

4.4 Bernoulli RV

- pmf of any Bernoulli RV:

$$p(x; \alpha) = \begin{cases} 1 - \alpha, & \text{if } x = 0 \\ \alpha, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

- α is a parameter, where $0 < \alpha < 1$
- Each different value of α between 0 and 1 determines a different member of the Bernoulli family of distributions

4.5 Bernoulli process

- A process with repeated independent trials
- 2 outcomes: 1 (success), 0 (failure)
- Success rate of trials is the same

4.6 Binomial distribution

- pmf of binomial RV:

$$p(x; n, p) = \begin{cases} C_{x,n} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

  - where n is the number of trials, and p is the success rate of each trial

- Since $\sum_{\text{all possible x}} p(x) = 1$,

$$\sum_{x=0}^{n} p(x; n, p) = \sum_{x=0}^{n} C_{x,n} p^x (1-p)^{n-x} = 1$$

4.7 Geometric distribution

- Probability distribution of number of Bernoulli trials X needed to get 1 success
- If $X = x$, $x - 1$ failures followed by success
- pmf of geometric RV:

$$p(x) = \begin{cases} p(1-p)^{x-1}, & x = 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

  - where p is the success rate of each trial

- Since $\sum_{\text{all possible x}} p(x) = 1$,

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = p \sum_{i=0}^{\infty} (1-p)^i = \frac{p}{1-(1-p)} = 1$$

4.8 Poisson distribution

- Used to model the number of occurrences of events in a time interval,
  where the average occurrence is λ
- pmf of Poisson RV:

$$p(x; \lambda) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0,\ 1,\ \ldots \\ 0, & \text{otherwise} \end{cases}$$

  - where λ is the parameter of Poisson distribution

- Since $\sum_{\text{all possible x}} p(x) = 1$,

$$\sum_{n=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

4.9 Cumulative Distribution Function (CDF)

- CDF $F(x)$ of discrete RV X with pmf $p(x)$:

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

- $F(x)$ is the probability that the observed value is at most x

- Graph of $F(x)$ for discrete RV X is the linear combination of step functions, such that

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1$$

5 W3: Expectation

5.1 Expected Value

- Expected value $E(X)$

$$E(X) = \mu_x = \sum_{x \in D} x \cdot p(x), \text{ provided that } \sum_{x \in D} |x| \cdot p(x) < \infty$$

- Expected value of a function $E[h(X)]$

$$E[h(X)] = \mu_{h(x)} = \sum_{x \in D} h(x) \cdot p(x)$$

- Expected value of a linear function $aX + b$

$$E(aX + b) = aE(X) + b$$

5.2 Variance

- Variance $V(X)$

$$V(X) = \textstyle\sum_{x \in D} (x - \mu)^2 p(x) = E[(X - \mu)^2], \text{ provided that the expectation exists}$$

OR

$$\text{Population variance, } \sigma^2 = V(X) = E(X^2) - [E(X)]^2$$

- Variance of a function $V[h(X)]$

$$V[h(X)] = \sum_{x \in D} \{h(x) - [E(X)]\}^2 \cdot p(x)$$

- Variance of a linear function $aX + b$

$$V(aX + b) = a^2 V(X)$$

$$\sigma_{aX+b} = |a| \sigma_x$$

5.3 Expected Value and Variance of Discrete PMFs

5.3.1 Bernoulli RV

Expected value $E(X)$

$$\begin{aligned} E(X) &= \sum_{x \in D} x \cdot p(x) \\ &= 0(1 - p) + 1(p) \\ &= p \end{aligned}$$

Variance $V(X)$

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= 0^2(1 - p) + 1^2(p) - p^2 \\ &= p - p^2 \\ &= p(1 - p) \end{aligned}$$

5.3.2 Binomial RV

The complete proof for expected value and variance can be found here:
https://www.math.ubc.ca/∼feldman/m302/binomial.pdf

Expected value $E(X)$

$$E(X) = np$$

Variance $V(X)$

$$V(X) = np(1-p)$$

5.3.3 Geometric RV

The complete proof for expected value and variance can be found here:
https://semath.info/src/st-geometric-distribution.html

Expected value $E(X)$

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

Variance $V(X)$

$$V(X) = \frac{1-p}{p^2}$$

5.3.4 Poisson RV

The complete proof for expected value and variance can be found here:
https://www.statlect.com/probability-distributions/Poisson-distribution

Expected value $E(X)$

$$E(X) = \lambda$$

Variance $V(X)$

$$V(X) = \lambda$$

6 W3: Continuous Random Variable

6.1 Definition

- Continuous RVs can take on any value in a continuous range (e.g. real numbers)
  - In contrast, discrete RVs can take on a discrete list of values

6.2 Probability Density Function (PDF) for Continuous RV

- Probability described by the probability density function (pdf), measured between an interval

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

6.3 Uniform Distribution

$$\text{pdf } f(x; a, b) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

6.4 Exponential Distribution

$$\text{pdf } f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

6.5 Normal/Gaussian Distribution

$$\text{pdf } f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

6.6 Cumulative Distribution Function (CDF)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

- Capital F means CDF, while small F means PDF
- For any a: $P(x > a) = 1 - F(a)$
- Between a and b: $P(a \leq X \leq b) = F(b) - F(a)$

6.6.1 Obtaining PDF from CDF

$$f(x) = F'(x)$$

- The PDF is the derivative of the CDF.

6.7 Expected Value

- Expected value $E(X)$

$$E(X) = \mu_x = \sum_{x \in D} x \cdot p(x), \text{ provided that } \int_{\infty}^{\infty} |x| \cdot p(x) < \infty$$

- Expected value of a function $E[h(X)]$

$$E[h(X)] = \mu_{h(x)} = \int_{-\infty}^{\infty} h(x) f(x) dx$$

- Expected value of a linear function $aX + b$

$$E(aX + b) = aE(X) + b$$

6.8 Variance

- Variance $V(X)$

$$\begin{aligned} V(X) = \mu_X^2 &= E[(X - \mu)^2] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

- Variance of a linear function $aX + b$

$$V(aX + b) = a^2 V(X)$$

$$\sigma_{aX+b} = |a| \sigma_x$$

6.9 Expected Value and Variance of Continuous PDFs

6.9.1 Uniform RV

The complete proof for expected value and variance can be found here:
https://www.statlect.com/probability-distributions/uniform-distribution

Expected value $E(X)$

$$E(X) = \frac{1}{2}(a + b)$$

Variance $V(X)$

$$V(X) = \frac{1}{12}(b - a)^2$$

6.9.2 Exponential RV

The complete proof for expected value and variance can be found here:
https://www.statlect.com/probability-distributions/exponential-distribution

Expected value $E(X)$

$$E(X) = \frac{1}{\lambda_E}$$

Variance $V(X)$

$$V(X) = \frac{1}{\lambda^2}$$

7 W4: Useful Distributions

7.1 Poisson Approximation of Binomial Distributions

For any binomial distribution where n is large and p is small, such that $np > 0$,

$$b(x; n, p) \approx p(x; \lambda), \text{ where } \lambda = np$$

- Approximation can be safely applied if $n > 50$ and $np < 5$

7.2 Poisson and Exponential Distributions

7.2.1 Poisson Distribution

- Often used to model the number of occurrence of events in a time interval
- e.g. number of buses at a bus stop between 3 and 4 pm

$$\text{pmf } p(x; \lambda) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0,\, 1,\, \ldots \\ 0, & \text{otherwise} \end{cases}$$

7.2.2 Exponential Distribution

- Often used to model the elapsed time between two successive events
- e.g. waiting time for a bus

$$\text{pdf } f(x; \alpha) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

7.2.3 Relationship between Poisson and Exponential Distributions

Let $X_1, X_2, \ldots$ be the time when the 1st, 2nd, ... event occur.
The probability of waiting not more than t for the first event is $P(X_1 \leq t)$.

Deriving via Poisson Distribution

$$\begin{aligned} P(X_1 \leq t) &= 1 - P(X_1 > t) \\ &= 1 - P(\text{no event in [0, t]}) \\ &= 1 - \frac{\lambda^0 e^{-\lambda}}{0!} \\ &= 1 - e^{-\lambda} \\ &= 1 - e^{-\alpha t}, \text{where } \lambda = \alpha t \end{aligned}$$

Deriving via Exponential Distribution

$$\begin{aligned} P(X_1 \leq t) &= 1 - P(X_1 > t) \\ &= 1 - \int_t^{\infty} \alpha e^{-\alpha x} dx \\ &= 1 - \left[\frac{\alpha}{-\alpha} e^{-\alpha x} \right]_t^{\infty} \\ &= 1 - e^{-\alpha t} \end{aligned}$$

The rate of occurrence α in the Poisson distribution is the parameter of the exponential distribution.

7.3 Memoryless Property of Exponential Distribution

- Distribution of waiting time until a certain event does not depend on how much time has elapsed
- e.g. P(bulb can last for 600 h) = P(bulb can last for 900 h | bulb can last for 300 h)

7.4 Normal Distribution

- Parameters: mean μ, variance σ^2
- Abbreviated $X \sim N(\mu, \sigma^2)$
- pdf of X:

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

7.5 Standard Normal Distribution

- Parameters: mean $\mu = 0$, variance $\sigma^2 = 1$
- Abbreviated $Z \sim N(0, 1)$
- pdf of Z:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

- cdf of Z:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(u)\ du$$

  - Result can be found using standard normal table

7.5.1 z_α Notation

- Denotes value on the z axis for which α of the area under the z curve lies to the **RIGHT** of z_α
- $100(1 - \alpha)$th percentile of the standard normal distribution

7.6 Standardizing A Normal Distribution

- Normal RV: $X \sim N(\mu, \sigma^2)$
- Standard Normal RV: $Z = \frac{X-\mu}{\sigma}$
- Similarly,

$$
\begin{aligned}
P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)
\end{aligned}
$$

8 W4: Joint Probability Distribution

8.1 Joint Probability Mass Function

The joint probability mass function p(x, y) is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

It must satisfy the following conditions:

1. $p(x, y) \geq 0$
2. $\sum_x \sum_y p(x, y) = 1$

The probability P[(X, Y) ∈ A] is obtained by summing the joint pmf over pairs in A:

$$P[(X, Y) \in A] = \sum_{(x,y)} \sum_{\in A} p(x, y)$$

8.2 Marginal Probability Mass Function

The marginal probability mass function of x, $p_X(x)$ is given by

$$p_X(x) = \sum_{y:p(x,y)>0} p(x, y) \text{ for each possible value of } x.$$

Similarly, the marginal probability mass function of y, $p_X(x)$ is given by

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x, y) \text{ for each possible value of } y.$$

- The word "marginal" indicates that the pmf is obtained from the joint probability distribution.
- We can obtain the marginal pmf from the joint pmf, however the reverse is not always true.

8.3 Joint Probability Density Function

The joint probability density function f(x, y) for two different RV is satisfies two conditions:

1. $f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\ dx\ dy = 1$

For any two dimensional set A, where $a \leq x \leq b,\ c \leq y \leq d$,

$$\begin{aligned} P[(X, Y) \in A] &= \iint_A f(x, y)\ dx\ dy \\ &= \int_a^b \int_c^d f(x, y)\ dx\ dy \end{aligned}$$

- $P[(X, Y] \in A]$ is the volume beneath the surface above the region A

8.4 Marginal Probability Density Function

The marginal probability density function of X and Y, denoted by $f_X(x)$ and $f_Y(y)$ respectively, are

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\ dy, \quad -\infty < x < \infty$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\ dx, \quad -\infty < y < \infty$$

- Marginal pdf of X is the pdf of X
- The word "marginal" indicates that the pdf is obtained from the joint probability distribution.
- We can obtain the marginal pdf from the joint pdf, however the reverse is not always true.

8.5 Multiple Random Variables

If $X_1, X_2, \ldots, X_n$ are all discrete RVs, the joint pmf of the variables is

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

If $X_1, X_2, \ldots, X_n$ are all continuous RVs, the joint pdf of the variables with intervals $[a_1, b_1], \ldots, [a_n, b_n]$ is

$$\begin{aligned} &P(a_1 \le X_1 \le b_1, a_2 \le X_2 \le b_2, \ldots, a_n \le X_n \le b_n) \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(x_1, x_2, \ldots, x_n)\ dx_n \cdots dx_2\ dx_1 \end{aligned}$$

8.6 Independence of Random Variables

Two RVs X and Y are said to be independent if for **every pair** of x and y values:

$$p(x,y) = p_X(x) \cdot p_Y(y) \quad \text{for discrete RV}$$
$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \text{for continuous RV}$$

If the above is not satisfied for all (x, y), then X and Y are dependent.

9 W5: Conditional Distribution

9.1 Conditional Probability Mass Function

Let X and Y be two discrete RVs with pmf $p(x, y)$.

For any value x for which $p(x) > 0$, the conditional probability mass function of Y given that $X = x$ is

$$p_{Y|X}(y \mid x) = \frac{p(x, y)}{p_X(x)}$$

where $p_X(x)$ is the marginal pmf of X.

9.2 Conditional Probability Density Function

Let X and Y be two continuous RVs with pdf $f(x, y)$. For any value x for which $f(x) > 0$, the conditional probability density function of Y given that $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

where $f_X(x)$ is the marginal pdf of X.

9.3 Conditional Distribution

- The summation of the conditional pmf or pdf over the entire sample space is 1.

$$\sum_y p_{Y|X}(y \mid x) = 1 \quad \text{for discrete RVs X and Y}$$
$$\int_{-\infty}^{\infty} f_{Y|X}(y \mid x) dy = 1 \quad \text{for continuous RVs X and Y}$$

9.4 Conditional Expectation

Let X and Y be jointly distributed RVs with pmf $p(x, y)$ or pdf $f(x, y)$. The expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X,Y)}$ is given by

$$E[(h(X, Y)] = \begin{cases} \sum_x \sum_y h(x, y)p(x, y) & \text{for discrete RVs X and Y} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)f(x, y) & \text{for continuous RVs X and Y} \end{cases}$$

9.5 Conditional Mean

Let X and Y be jointly distributed RVs with pmf $p(x, y)$ or pdf $f(x, y)$. The conditional mean of Y, given that $X = x$, denoted by $\mu_{Y|x}$ is given by

$$\mu_{Y|x} = E(Y \mid x) = \begin{cases} \sum_y yp(y \mid x) & \text{for discrete RVs X and Y} \\ \sum_y h(y)f(y \mid x)dy & \text{for continuous RVs X and Y} \end{cases}$$

9.6 Conditional Variance

Let X and Y be jointly distributed RVs with pmf $p(x, y)$ or pdf $f(x, y)$. The conditional mean of Y, given that $X = x$. denoted by $\sigma^2_{Y|x}$ is given by

$$\begin{aligned}\sigma^2_{Y|x} &= E\{[Y - E(Y \mid x)]^2\} \\ &= E(Y^2 \mid x) - [E(Y \mid x)]^2\end{aligned}$$

9.7 Law of Total Expectation

If X is a RV, and Y is a RV in the same probability space, then

$$E\left[E(X \mid Y)\right] = E(X)$$

i.e. expected value of the conditional expected value of X given Y is the = expected value of X

9.8 Covariance

The covariance between two variables X and Y, denoted by $\sigma_{X,Y}$ is given by

$$\begin{aligned}\sigma_{X,Y} &= K(X, Y) = E[(X - \mu_x)(Y - \mu_y)] \\ &= \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_x)(y - \mu_y)\ p(x, y) & \text{for discrete RVs X and Y} \\ \displaystyle\int_x \int_y (x - \mu_x)(y - \mu_y)\ f(x, y)\ dx\ dy & \text{for continuous RVs X and Y} \end{cases}\end{aligned}$$

- Shortcut formula: $K(X, Y) = E(XY) - E(X)E(Y)$
- Value of covariance:
  - Positive $\sigma_{X,Y}$: positive linear relationship between X and Y
  - Near-zero $\sigma_{X,Y}$: no linear relationship between X and Y
  - Negative $\sigma_{X,Y}$: negative linear relationship between X and Y

9.9 Correlation

- Correlation coefficient $\rho_{X,Y}$: measure of degree of linear relationship between two RVs X and Y

$$\rho_{X,Y} = \widetilde{K}(X,Y) = \frac{K(X,Y)}{\sigma_X \sigma_Y}$$

  - It is always true that $-1 \leq \rho_{X,Y} \leq 1$

- If X and Y, then $\rho_{X,Y} = 0$

  - **BUT** $\rho_{X,Y}$ does not imply independence between X and Y

- Measure of linear relationship:

  - $|\rho| = 1$: Strong linear relationship between X and Y
  - $|\rho| \neq 1$: Not completely linear relationship between X and Y; could be strong non-linear relationship
  - $\rho = 0$: X and Y are uncorrelated

10 W5: Central Limit Theorem

10.1 Linear Combination of One RV

For a linear combination of one RV X, denoted by $aX + b$, the mean and variance are as follows:

- Mean, $E(aX + b) = aE(X) + b$
- Variance, $V(aX + b) = a^2E(X)$

10.2 Linear Combination of Two RVs

For a linear combination of two RVs X and Y, where $W = aX + bY$,
the mean and variance are as follows:

	X, Y independent	**X, Y dependent**
Mean, $E(W)$	$aE(X) + bE(Y)$	
Variance, $V(W)$	$a^2V(X) + b^2V(Y)$	$a^2V(X) + b^2V(Y) + 2abK(X, Y)$

10.3 Linear Combination of Multiple RVs

For a linear combination of multiple RVs $X_1, X_2, \ldots, X_n$, where $W = \sum_{i=1}^{n} a_i x_i$,
the mean and variance are as follows:

	RVs independent	**RVs dependent**
Mean, $E(W)$	$\sum_{i=1}^{n} a_i E(X_i)$	
Variance, $V(W)$	$\sum_{i=1}^{n} a_i^2 V(X_i)$	$\sum_{i=1}^{n} a_i^2 V(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_i a_j K(X_i, X_j)$

10.4 Linear Combination of Independent and Identically Distributed RVs

For a linear combination of independent and identically distributed (iid) RVs $X_1, X_2, \ldots, X_n$ where $W = \sum_{i=1}^{n} X_i$ with mean μ and variance σ^2, the mean and variance are as follows:

- Mean, $E(W) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\mu$
- Variance, $V(W) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$

10.5 Linear Combination of Normal RVs

For two normal RVs X and Y, where $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, the linear combination $W = X + Y$ is also a normal RV with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$, i.e.

$$W \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

10.6 Sample Mean

Let $X_1, X2, \ldots, X_n$ be iid RVs with mean μ and variance σ^2.
The sample mean $\overline{X}$ can be calculated using the formula $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.
The mean and variance of $\overline{X}$ is as follows:

- Mean, $E(\overline{X}) = \mu$
- Variance, $V(\overline{X}) = \dfrac{\sigma^2}{n}$

10.7 Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. The sample mean $\overline{X}$ can be calculated using the formula $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

For a sufficiently large n, i.e. $\mathbf{n \leq 30}$, $\overline{X}$ has approximately a normal distribution with mean $E(\overline{X})$ and variance $V(\overline{X})$ as follows:

- Mean, $E(\overline{X}) = \mu$
- Variance, $V(\overline{X}) = \dfrac{\sigma^2}{n}$

If the distribution is close to a normal pdf, a small n yields a good approximation to a normal distribution.

11 W8: Statistics and Their Distributions

11.1 Definitions

- Population: all observations
- Sample: subset of population
- Statistic: quantity whose value can be calculated from sample data
  - A random variable

11.2 Order statistic

For iid RVs $X_1, X_2, \ldots, X_n$ of unknown distribution, they can be rearranged in an increasing order:

$$X_{(1)} \leq X_{(2)} \leq \ldots X_{(k)} \ldots \leq X_{(n)}$$

where

- $X_{(1)} = \min\{X_1, \ldots, X_n\}$ is the smallest order statistic;
- $X_{(k)}$ is the k-th order statistic; and
- $X_{(n)} = \max\{X_1, \ldots, X_n\}$ is the largest order statistic.

11.3 Sample range

The sample range R is the distance between the largest and smallest order statistic.
It is also a random variable, and can be calculated by:

$$R = X_{(n)} - X_{(1)}$$

11.4 Distribution of a statistic

The distribution of a statistic can be obtained by either 1 of the 2 methods:

1. Derive the probability distribution analytically via order statistics
2. Simulate the probability distribution using Monte Carlo simulation

11.5 Distribution of $\overline{X}$

For a sufficiently large n, i.e. $\mathbf{n \leq 30}$, $\overline{X}$ has approximately a normal distribution with mean $E(\overline{X})$ and variance $V(\overline{X})$ as follows:

- Mean, $E(\overline{X}) = \mu$
- Variance, $V(\overline{X}) = \dfrac{\sigma^2}{n}$

11.6 Distribution of smallest order statistic $X_{(1)}$

- pdf:

$$f_{(1)}(x) = n\,[1 - F_X(x)]^{n-1}\, f_X(x)$$

- cdf:

$$\begin{aligned}
F_{(1)}(x) &= P(X_{(1)} \leq x) \\
&= 1 - P(X_i > x, \forall i) \\
&= 1 - \prod_{i=1}^{n} P(X_i > x) \quad \text{(independent)} \\
&= 1 - [P(X_i > x)]^n \quad \text{(identically distributed)} \\
&= 1 - [1 - P(X_i \leq x)]^n \\
&= 1 - [1 - F_X(x)]^n
\end{aligned}$$

11.7 Distribution of largest order statistic $X_{(n)}$

- pdf:

$$f_{(n)}(x) = n\,[F_X(x)]^{n-1}\, f_X(x)$$

- cdf:

$$\begin{aligned}
F_{(n)}(x) &= P(X_{(n)} \leq x) \\
&= P(X_i \leq x, \forall i) \\
&= \prod_{i=1}^{n} P(X_i \leq x) \quad \text{(independent)} \\
&= [P(X_i \leq x)]^n \quad \text{(identically distributed)} \\
&= [F_X(x)]^n
\end{aligned}$$

11.8 Distribution of k-th order statistic $X_{(k)}$

- pdf:

$$f_{(k)}(x) = \frac{n!\,[F_X(x)]^{k-1}\,[1 - F_X(x)]^{n-k}\, f_X(x)}{(k-1)!(n-k)!}$$

12 W9: Point Estimation

12.1 Point estimate

- Statistic, function of data to infer value of unknown parameter
- A random variable
  - e.g. point estimate of θ is $\hat{\theta}$

12.2 Principle of Unbiased Estimator

- Choose an unbiased estimator among several candidates
- Point estimate $\hat{\theta}$ is an unbiased estimator if $E(\hat{\theta}) = \theta$ for every possible value of θ
- Can be obtained from biased estimator by using making $E(\hat{\theta}) = \theta$

12.3 Principle of Minimum Variance Unbiased Estimation

- Among all the unbiased estimators of θ, choose the estimator with the minimum variance.
- Estimator with the minimum variance is the minimum variance unbiased estimator (MVUE) of θ.

13 W9: Method of Moments Estimator (MME)

13.1 Notations

- Random sample of size n: $X_1, X_2, \ldots, X_n$
- Distribution: $f(x, \theta)$ or $p(x, \theta)$, where θ is the parameter

13.2 Moments

- Measure something relative to center of values
- k-th population moment, $\mu_k = E(X^k)$
  - Depends on unknown parameters
- k-th sample moment, $M_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$
  - Function of random sample

13.3 Method of Moments

- Assumes that sample moments provide good estimates of the corresponding population moments
- Does **NOT** guarantee to produce an unbiased estimator

13.4 Method of Moments Estimator (MME)

To calculate the MME(s) of $\hat{\theta}$:

1. Find m population moments, where m is the number of unknown parameters.
2. Find m sample moments.
3. Equate each population moment to its corresponding sample moments
4. Solve for $\theta = (\theta_1, \ldots, \theta_m)$ to obtain the MMEs for θ.

14 W10: Maximum Likelihood Estimator (MLE)

14.1 Intuition

- Find parameters of the distribution that would most likely produce observed data
- If a sample is observed, the probability of having such a sample should be maximized because it has actually occurred

14.2 Likelihood function

Let $X_1, X_2, \ldots, X_n$ have a joint pdf or pmf:

$$L(\theta_1, \ldots, \theta_m) = f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$$

The likelihood function is given by

$$L(\theta) = P(X_1 = x, \ldots, X_n = x_n) = \begin{cases} \prod_{i=1}^{n} p(x_i, \theta) & \text{for discrete RVs} \\ \prod_{i=1}^{n} f(x_i, \theta) & \text{for continuous RVs} \end{cases}$$

14.3 Maximizing the likelihood

- The maximum likelihood estimator (MLE) $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are values that maximize the likelihood function such that
$$L(\hat{\theta}_1, \ldots, \hat{\theta}_m) \leq L(\theta_1, \ldots, \theta_m)$$

14.4 Maximimum Likelihood Estimator (MLE)

To calculate the MME of θ:

1. Find the likelihood function $L(\theta)$ based on the distribution.
2. Differentiate $L(\theta)$ with respect to θ, and equate the derivative to 0.
    - The natural logarithm of $L(\theta)$ could simplify calculations.
3. Solve for the MME of θ.
4. Check if the value is maximum by taking the second derivative of $L(\theta)$.

15 W10: Confidence Interval

- Quantifies the confidence interval of a point estimate $\hat{\theta}$

$$l(X_1, \ldots, X_n) < \hat{\theta}(X_1, \ldots, X_n) < u(X_1, \ldots, X_n)$$

  - where $l(\ldots)$ is the lower bound and $u(\ldots)$ is the upper bound respectively.

- The interval contains θ with a confidence interval p:

$$P\{\theta \in [l(X_1, \ldots, X_n), u(X_1, \ldots, X_n)]\} = p$$

- The confidence interval p is often set to a high value e.g. 0.95, 0.99 in practice

15.1 Equivalent expressions for Confidence Interval

The following expressions are equivalent in describing a 90% confidence interval (CI) for μ.

$$\begin{aligned} P\left(|\overline{X} - \mu| < \frac{1.65\sigma}{\sqrt{n}}\right) &= 0.90 \\ P\left(\overline{X} - \frac{1.65\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{1.65\sigma}{\sqrt{n}}\right) &= 0.90 \\ P\left[\mu \in \left(\overline{X} - \frac{1.65\sigma}{\sqrt{n}}, \overline{X} + \frac{1.65\sigma}{\sqrt{n}}\right)\right] &= 0.90 \end{aligned}$$

- Replace 1.64 with:
  - 1.96 if CI is 95%
    - Closest Z-score of area 0.97500 in standard normal table
  - 2.58 if CI is 99%
    - Closest Z-score of area 0.99500 in standard normal table
  - **Rule of thumb:**
    - Search for Z score of area $p + \frac{1-p}{2}$ in the standard normal table, where p is the CI.

15.2 Interpretation of Confidence Interval

- e.g. 95% CI for μ
  - As the number of samples collected tend to infinity, 95% of the samples will contain μ.

15.3 Properties of Confidence Interval

- As population variance σ increases, the width of CI increases.
- As sample size n increases, the width of CI decreases.
- As the confidence interval p increases, the width of CI increases.
- At a fixed confidence interval,
  - Large width of CI $\rightarrow$ low precision
  - Small width of CI $\rightarrow$ high precision

16 W11: Hypothesis Testing 1

16.1 Statistical hypothesis

- A claim about values of parameters/form of probability distribution

16.2 Null and Alternative Hypotheses

- Null hypothesis, H_0
  - Claim that is initially assumed to be true
  - H_0 is **always** $H_0 : \theta = \theta_0$
- Alternative hypothesis, H_a
  - Claim that contradicts the null hypothesis H_0
  - H_a has 3 forms with implicit hypothesis
    - $H_a : \theta > \theta_0$ (implicit hypothesis: $\theta \leq \theta_0$)
    - $H_a : \theta < \theta_0$ (implicit hypothesis: $\theta \leq \theta_0$)
    - $H_a : \theta \neq \theta_0$ (implicit hypothesis: $\theta = \theta_0$)

16.3 Hypothesis Testing

- Method to decide whether to accept or reject the null hypothesis, H_0
- Comprises 2 components:
  - Test statistic
    - Function of sample data to make a decision
  - Rejection region
    - Set of values for which the null hypothesis H_0 will be rejected
    - If test statistic falls in rejection region, H_0 will be rejected

16.4 Errors in Hypothesis Testing

- Type I error (α): Rejecting the null hypothesis H_0 when H_0 is true

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

- Type II error (β): Accepting the null hypothesis H_0 when H_a is true

$$\beta = P(\text{accept } H_0 \mid H_a \text{ is true})$$

- Good rejection region yields small α and β
  - Typical approach: specify largest value of α that can be tolerated, then back-calculate for the rejection region

16.5 Hypothesis Testing using Rejection Region

1. Figure out appropriate H_0 and H_a.

2. Figure out appropriate test statistic.

$$\overline{X} = \frac{1}{n}\sum X_i \quad \Longrightarrow \quad Z = \begin{cases} \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

3. Calculate the rejection region based on type I error/significance level α:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

4. Calculate the normalized sample mean z using sample mean $\overline{x}$.

$$z = \begin{cases} \dfrac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

5. Compare the normalized sample mean z with the rejection region.

   Reject H_0 if z falls in the rejection region.

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - Rejection region: $Z < -z_\alpha$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - Rejection region: $Z > -z_\alpha$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - Rejection region: $Z < -z_{\alpha/2} \ \cup \ Z > z_{\alpha/2}$

17 W11: Hypothesis Testing 2

17.1 Hypothesis Testing of Difference between 2 Populations

1. Figure out appropriate H_0 and H_a.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

2. Figure out appropriate test statistic.

$$\overline{X_1} - \overline{X_2} = \frac{1}{n} \sum (X_{1i} - X_{2i})$$

$$\implies \quad Z = \begin{cases} \dfrac{\overline{X_1} - \overline{X_2}}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{X_1} - \overline{X_2}}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

3. Calculate the rejection region based on type I error/significance level α:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

4. Calculate the normalized sample mean z using sample mean $\overline{x_1} - \overline{x_2}$.

$$z = \begin{cases} \dfrac{\overline{x_1} - \overline{x_2}}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{x_1} - \overline{x_2}}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

5. Compare the normalized sample mean z with the rejection region.
   Reject H_0 if z falls in the rejection region.

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - Rejection region: $Z < -z_\alpha$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - Rejection region: $Z > -z_\alpha$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - Rejection region: $Z < -z_{\alpha/2} \ \cup \ Z > z_{\alpha/2}$

17.2 P-value

- A probability, calculated assuming that H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.
- Also known as *observed significance level* (OSL) for the data
  - Data is significant if H_0 is rejected
  - Data is not significant if H_0 is accepted

17.3 Hypothesis Testing using P-value

1. Figure out appropriate H_0 and H_a.

2. Calculate the test statistic value of sample z.

$$z = \begin{cases} \dfrac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

3. Determine range of test statistic values as contradictory to H_0 as the above value of z.

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - Range: $Z < z$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - Range: $Z > z$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - Range: $Z > z \ \cup \ Z < -z$

4. Calculate probability of getting that range, assuming H_0 is true:

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - P-value $= P(Z < z \mid H_0$ is true)
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - P-value $= P(Z > z \mid H_0$ is true)
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - P-value $= P(Z > z \ \cup \ Z < -z \mid H_0$ is true)

5. Compare the P-value against the significance level α.

   - Reject H_0: P-value $\leq \alpha$
   - Accept H_0: P-value $> \alpha$

17.4 Comparison between Hypothesis Testing Methods

- The two procedures – the rejection region method and P-value method – are equivalent.
  - The same conclusion will be reached via either of the two procedures.

18 W12: Linear Regression

18.1 Least-squares method

- Estimates unknown parameters of a function based on known data

18.2 Estimating β_0 and β_1

1. Define an error function to minimize.

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)^2$$

2. Take the partial derivative of the error function with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and solve for the unknowns.

$$\begin{aligned} \frac{\partial f}{\partial \hat{\beta}_1} = 0 : -2\sum (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)(-x_i) = 0 \\ \sum x_i (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) = 0 \\ \Rightarrow \sum (\hat{\beta}_1 x_i^2 + \hat{\beta}_0 x_i) = \sum (x_i y_i) \end{aligned}$$

$$\begin{aligned} \frac{\partial f}{\partial \hat{\beta}_0} = 0 : -2\sum (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)(-1) = 0 \\ \sum (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) = 0 \\ \Rightarrow \sum (\hat{\beta}_1 x_i + \hat{\beta}_0) = \sum y_i \end{aligned}$$

$$\text{Design matrix of error function: } \sum_{i=1}^{n} \begin{bmatrix} x_i^2 & x_i \\ x_i & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} x_i y_i \\ y_i \end{bmatrix}$$

3. Examine the Hessian matrix to determine if the solutions are at a minimum, i.e.

$$\begin{bmatrix} \dfrac{\partial f}{\partial \hat{\beta}_0^2} & \dfrac{\partial^2 f}{\partial \hat{\beta}_0 \hat{\beta}_1} \\ \dfrac{\partial^2 f}{\partial \hat{\beta}_0 \hat{\beta}_1} & \dfrac{\partial f}{\partial \hat{\beta}_1^2} \end{bmatrix} \text{ is positive definite.}$$

18.3 Least-squares estimates for β_0 and β_1

$$\begin{aligned} \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\overline{x}^2} \end{aligned}$$

- $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the estimated regression line or least-squares line

18.4 Residuals and fitted values

- Residual, $y_i - \hat{y}_i$
  - The difference between the observed value y_i and the fitted value $\hat{y}_i$
  - Positive residual $\rightarrow$ observed point lies above the least-squares line
  - Negative residual $\rightarrow$ observed point lies below the least-squares line
- Sum of residuals, $y_i - \hat{y}_i$
  - For an estimated regression line obtained by the least-squares method, the sum of residuals is zero:

$$\sum_{i=1}^{n} y_i - \hat{y}_i = 0$$

- Fitted values $\hat{y}_i$
  - Obtained by substituting x_i into the regression line equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

18.5 The simple linear regression model

- The simple linear regression model can be described by the model equation

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where ε represents uncertainty of the model and is a normal N(0, σ^2) RV.

- The line $y = \beta_0 + \beta_1 x$ is called the true/population regression line.
- Mean of Y, $E(Y)$

$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\ &= \beta_0 + \beta_1 x + E(\varepsilon) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

- Variance of Y, $V(Y)$

$$\begin{aligned} V(Y) &= V(\beta_0 + \beta_1 x + \varepsilon) \\ &= 0 + V(\varepsilon) \\ &= \sigma^2 \end{aligned}$$

18.6 Sum of squared error (SSE)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

- Measures discrepancy between the data and the estimation model
- Small SSE $\rightarrow$ tight fit of estimation model to data

18.7 Estimating σ^2 of regression model

- An unbiased estimate for σ^2 in the regression model is s^2:

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

- Estimating β_0 and β_1 results in the loss of 2 degrees of freedom
  - Thus the denominator for s^2 is $n - 2$