50.034 Notes

Koh Jing Yu

April 11, 2018

Abstract

My personal notes for 50.034 Probability and Statistics.

Contents

1 Week 1

- **Probability:** properties of the populations are known, questions regarding a sample taken from the population are investigated (deductive reasoning).
- **Statistics:** characteristics of a sample are known from the experiment, and conclusions regarding the population are made (inductive reasoning).
- **Inferential statistics:** making predictions or inferences about a population from observations and analyses of a sample.
- **Descriptive statistics:** techniques with describing a sample or population.

1.1 Population

A well defined collection of objects. **E.g.** All registered students in this course, all aircrafts sold by Lockheed Martin last year.

1.2 Sample

A subset of the population selected in some prescribed manner. **E.g.** "Students in this class" is a sample of "registered students in this course".

1.3 Variable

Any characteristic whose value may change from one object to another in the population. **E.g.** $x =$ gender of a student

1.4 Descriptive Statistics

1.4.1 Range:

Difference between largest and smallest sample values. **E.g.** For $1, 3, 5$, the range is $5 - 1 = 4$.

1.4.2 Mean:

Average of all values.

Population mean usually represented by μ. Sample mean often denoted by $\bar{x} = \frac{\sum x_i}{n}$

Properties of Mean

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

$$\bar{y} = a\bar{x} + b$$

$$s_y^2 = a^2 s_x^2$$

1.4.3 Variance:

Variability of a data set.
Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Shortcut to obtain σ^2 is $(\frac{1}{N}\sum_{i=1}^{N} x_i^2) - \mu^2$. Proof: Just expand!

Why is it $n-1$? $\bar{x}$ is an estimation of population mean μ but is biased towards the chosen sample. Dividing by n produces an estimation of population variance smaller than the actual value. Dividing by $n-1$ corrects the deviation.

$$\begin{aligned}
E[S^2(n)] &= E[\frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2] \\
&= \frac{1}{n-1}E[\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + n\bar{x}^2] \\
&= \frac{1}{n-1}E[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2] \\
&= \frac{n}{n-1}(E[x^2] - E[\bar{x}^2(n)]) \\
&= \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 \\
&= \sigma^2
\end{aligned}$$

1.4.4 Median

Population median is defined as:

$$\widetilde{\mu} = \begin{cases} x_m, & \text{N is odd, } m = (N+1)/2. \\ (x_m + x_{m+1})/2, & \text{N is even, } m = N/2. \end{cases}$$

Sample median is defined as:

$$\widetilde{x} = \begin{cases} x_m, & \text{n is odd, } m = (n+1)/2. \\ (x_m + x_{m+1})/2, & \text{n is even, } m = n/2. \end{cases}$$

1.4.5 Percentile

Percentile is a value corresponding to a percentage of all data below the value. A dataset is ordered $(x_1, x_2, \ldots, x_n)$ as

$$x'_1 \leq x'_2 \leq x'_3 \leq \ldots \leq x'_n$$

where x'_1 and x'_n are the smallest and largest data. Then, x'_i corresponds to $\frac{100(i-0.5)}{n}$th percentile.

For a 10-data sample: $1, 3, 5, 7, 9, 11, 13, 15, 17, 19$, 20th percentile is 4. (3 is 15th percentile and 5 is 25th percentile).

Note: For a given percentile, the corresponding value may not exist in the data set. This is similar to median when the data set has an even sample size.

1.4.6 Frequency and Relative Frequency

Frequency of any particular x value is its occurrence time in the data set.

Relative Frequency of a value is the fraction or proportion of times the value occurs.

This can be represented in a *histogram*.

Steps to construct a histogram of frequency distribution:

1. Number of classes for variable x. No hard and fast rule, but between 5 and 20 classes is good for most data sets. If larger than 20, a reasonable rule of thumb is: number of classes $\approx \sqrt{\text{number of observations}}$
2. Determine the frequency and relative frequency of each x value or class.
3. Mark x values or class intervals on a horizontal scale.
4. Above each class, draw a rectangle whose height is the relative frequency/percent or frequency/count of each class

1.5 Sample Space, Event

1.5.1 Sample Space

Sample space, denoted by Ω, is the set of ALL possible outcomes of that experiment. **E.g. of experiments:** Rolling a die, checking the age of a student, recording if it rains in SUTD at 10am.

1. **Collectively exhaustive:** Sample space should contain all possible outcomes.
2. **Mutually exclusive:** Each outcome in the sample space should be unique.

Sample space vs population: Sample space is unique. For example, in the sample space of the age of all students, the age of 18 only appears once. In the population, there may be several students whose age is 18.

1.5.2 Event

An event is a collection (subset) of outcomes contained in a sample space Ω. An event is **simple** if it consists of exactly one outcome, and **compound** if it consists of more than one outcome.

1.6 Set Theory

The **complement** of an event A is denoted by A^c, the set of all outcomes in Ω not contained in A.

The **intersection** of two events A and B denoted by $A \cap B$, the event consisting of all outcomes in both A and B.

The **union** of two events A and B denoted by $A \cup B$, the event consisting of all outcomes that are either in A or B.

1.6.1 Null event

The null event, denoted by $\varnothing$ is the event that consists of no outcome.

Events A and B are said to be mutually exclusive or disjoint events if $A \cap B = \varnothing$. Events $A_1, A_2, A_3, \ldots$ are mutually exclusive (or pairwise disjoint) if no two events have any outcome in common.

1.6.2 De Morgan's Laws

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

1.7 Probability

The probability of an event A is given by a number $P(A)$ which is a measure of the chance that A will occur.

1.8 Axiom of Probability

- For any event A, $P(A) \geq 0$, and $P(A) \leq 1$
- $P(\Omega) = 1$
- Any infinite collection of disjoint events satisfies: $P(A_1 \cup A_2 \cup A_3 ...) = \sum_{i=1}^{\infty} P(A_i)$

1.9 Properties of Probability

- $P(A) + P(A^c) = 1$
- $P(A) = 1 - P(A^c)$
- $P(A) \leq 1$
- $P(\varnothing) = 0$ **BUT** $P(A) = 0 \nRightarrow A = \varnothing$
- $P(A) = 1 \nRightarrow A = \Omega$

2 Week 2

2.1 Motivation

When the various outcomes of an experiment are equally likely, the problem is reduced to counting. Let N denote the number of outcomes in Ω and $N(A)$ represent the number of outcomes contained in an event A:

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

2.2 Independence

Two events being independent means that the occurrence or non-occurrence of one event has no bearing on the chance of the other.

Two events are **indepedent** iff. $P(A \cap B) = P(A)P(B)$. This is equivalent to: $P(A|B) = P(A)$.

Two events are **dependent** if $P(A|B) \neq P(A)$.

Disjoint events are not independent, as $P(A \cap B) = 0$, but $P(A)P(B) \geq 0$

2.2.1 Independence of Several Events

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

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

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

Question: Events A_1, A_2 and A_3 are pairwise independent. Are they mutually independent? Answer: Not necessarily.

2.3 Product Rule

Consider a group of two elements. If the first element of a group can be selected in n_1 ways, and for each of these n_1 ways the second element can be selected in n_2 ways, then the number of pairs is $n_1 n_2$.

2.3.1 Tuple

A group of k elements is called a k-tuple. A pair is a 2-tuple, and a triple is a 3-tuple.

If the 1st element can be selected in n_1 ways, the second in n_2 ways, the k-th element in n_k ways, and the elements are selected independently, then there are $n_1 n_2 \dots n_k$ possible k-tuples.

2.4 Permutation

An ordered subset. Number of permutations of size k that can be formed from n objects is denoted by $P_{k,n}$. $P_{k,n}$ can be determined based on the product rule of k-tuple.

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

2.5 Combination

An unordered subset of a group is called a combination. The number of combinations of size k that can be formed from n objects is $C_{k,n}$ or $\binom{n}{k}$.

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

This is equivalent to the number of corresponding permutations disregarding the different outcomes due to order.

2.6 Conditional Probability

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

Given that B has already occurred, the relevant sample space is no longer Ω but B.

2.7 Law of Total Probability

The events $A_1, A_2, \dots, A_k$ are exhaustive if one A_i **must** occurr, i.e. $A_1 \cup A_2 \cup \dots \cup A_k = \Omega$

Let $A_1, A_2, \dots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

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

2.8 Bayes' Theorem

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

For any other event B with $P(B) \geq 0$, the posterior probability of A_j given that B has occurred is:

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

2.9 Random Variable

For any sample space Ω of some experiment, a random variable (R.V.) is any rule that associates a number with each outcome in Ω. it is a function whose domain is the sample space and range is a set of real numbers.

Discrete: Possible values constitute either a finite set of "countably" infinite set.

Continuous: Possible values constitute either a single interval on the real line or a union of disjoint intervals on the real line. $P(X = x) = 0$ for any possible value x.

2.10 Probability Distribution

The probabily distribution of a discrete R.V. X is called probability mass function (pmf). It is defined for every number x as $p(x) = P(X = x) = P[w \in \Omega : X(w) = x]$

The pmf complete describes the probabilistic properties of X. For any pmf, $p(x) \geq 0$, and $\sum_{allpossible x} p(x) = 1$.

If $p(x)$ depends on a quantity that can be assigned a number of possible values, which each value determining a different probability distribution, it is called a parameter of the distribution.

The collection of all probability distributions for different values of the parameter is called a family of probability distributions.

2.10.1 Bernoulli Distribution

The pmf of any Bernoulli R.V. can be expressed as $p(1) = p$ and $p(0) = 1 - p$, where $0 \leq p \leq 1$. pmf depends on p, so we often write $p(x; p)$ rather than just $p(x)$:

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

Every different value of p between 0 and 1 determines a different member of the Bernoulli family of distributions.

The **Bernoulli process** is an experiment of n repeated trials. The trials are:

1. Independent.
2. Only has two outcomes: 1 (success) and 0 (failure).
3. The success rate/probability of the trials is the same (denoted as p).

For a Bernoulli distribution, $\mu = p$ and $\sigma^2 = p(1-p)$.

2.10.2 Binomial Distribution

The pmf of a binomial R.V. is given by:

$$p(x;n,p) = \begin{cases} C_{x,n}p^x(1-p)^{n-x}, & \text{x=0,1,..., n.} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Where n is the number of trials and p is success rate of each trial.

It can be verified that:

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

For a Binomial distribution, $\mu = np$ and $\sigma^2 = np(1-p)$.

2.10.3 Geometric Distribution

Consider an experiment consisting of Bernoulli trials, each with success probability p. The total number of trials X up to and including the first success follows a Geometric distribution.

If $X = x$, there must be $x-1$ failures followed by a success, so the pmf of X is:

$$p(x) = \begin{cases} (1-p)^{x-1}p, & \text{x=1,2,...} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

It can be verified that:

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

A geometric distribution has the memoryless (forgetfulness) property:

$$P(X > s+t|X > t) = P(X > s)$$

or equivalently:

$$P(X > s+t) = P(X > s)P(X > t)$$

.

Proof. Using the conditional probability property:

$$\begin{aligned} P(X > s+t|X > t) &= \frac{P(X > s+t, X > t)}{P(X > t)} \\ &= \frac{P(X > s+t)}{P(X > t)} \\ &= \frac{(1-p)^{s+t}}{(1-p)^{t}} \\ &= (1-p)^{s} \\ &= P(X > s) \end{aligned}$$

□

For a Geometric distribution, $\mu = \frac{1}{p}$ and $\sigma^2 = \frac{1-p}{p^2}$.

2.10.4 Poisson Distribution

The Poisson distribution is often used to model the number of occurrences of events in a time interval (e.g. the number of buses through a bus stop from 3 to 4pm.). The average occurrence is λ.

$$p(x;\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & \text{x=0,1,}\ldots. \\ 0, & \text{otherwise.} \end{cases} \quad , \quad \lambda > 0 \tag{4}$$

It can be shown that $\sum_{x=0}^{\infty} p(x;\lambda) = 1$ using the $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$.

For a Poisson distribution, $\mu = \lambda$ and $\sigma^2 = \lambda$.

2.11 Cumulative Distribution Function

The cumulative distribution function (cdf) $F(x)$ of a discrete R.V. X with pmf $p(x)$ is defined for every number x:

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

$F(x)$ is the probability that the observed value of X is at most x. For a discrete R.V. X, the graph of $F(x)$ is a step function with:

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

2.12 Expected Value

Let X be a discrete R.V. with possible values in D and pmf $p(x)$. The expected value or mean of X, denoted by $E[X]$ or μ_X or just μ is:

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

Provided that $\sum_{x \in D} |x| \cdot p(x) < \infty$. If the sum diverges, the expected value is undefined.

2.12.1 Expected Value of a Function

Given a R.V. X, one may be interested in another R.V. $Y = h(X)$, which is a function of X. If the R.V. X has possible values in D and pmf $p(x)$, the expected value of any function $h(X)$ is:

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

2.12.2 Expected Value of a Linear Function

$$E[aX + b] = aE[X] + b$$

For 2 R.V. X and Y and $Z = X + Y$,

$$E[Z] = E[X] + E[Y]$$

2.12.3 Multiplicated Random Variables

Let X and Y be **independent** R.V. with finite expectations $E[X]$ and $E[Y]$. Then $E[XY] = E[X]E[Y]$

2.13 Variance of X

Let X have pmf $p(x)$ and mean μ; then the variance of X, denoted by $\text{var}(X)$ or σ_X^2 or just σ^2, is:

$$\text{var}(X) = \sum_{x \in D} (x - \mu_x)^2 p(x) = E[(x - \mu_x)^2]$$

Provided the expectation exists. A shorter way to calculate the variance is:

$$\text{var}(X) = E[X^2] - (E[X])^2 = \sum x^2 p(x) - \mu_X^2$$

If X and Y are two **independent** R.V., and $Z = X + Y$,

$$\text{var}(Z) = \text{var}(X) + \text{var}(Y)$$

The standard deviation of X is:

$$\sigma_X = \sqrt{\sigma_X^2}$$

2.13.1 Variance of a Function

If the R.V. X has possible values in D and pmf $p(x)$, the variance of any function $h(X)$, denoted by $\text{var}(h(X))$ or $\sigma_{h(X)}$ is:

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

For $h(X) = aX + b$ (linear function of X), the variance and standard deviation of $h(X)$ are:

$$\text{var}(aX + b) = \sigma_{aX+b}^2 = a^2 \sigma_X^2 = a^2 \text{var}(X),$$

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

3 Week 3

3.1 Continuous R.V.

A R.V. is continuous if:

1. Possible values constitute either a single interval on the real line or a union of disjoint intervals on the real line.
2. No possible value has positive probability: $P(X = x) = 0 \qquad \forall x$

The probability distribution or probability density function (pdf) of a continuous R.V. is a function $f(x)$, s.t. for any a and b with $a \leq b$,

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

For $f(x)$ to be a legitimate pdf, it must satisfy:

1. $f(x) \geq 0 \qquad \forall x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Boundary issue: Since the probability of any value is 0, the following are equal:

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

3.2 Uniform Distribution

A continuous R.V. X is said to have a uniform distribution on the interval $[a, b]$ if the pdf of X is:

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

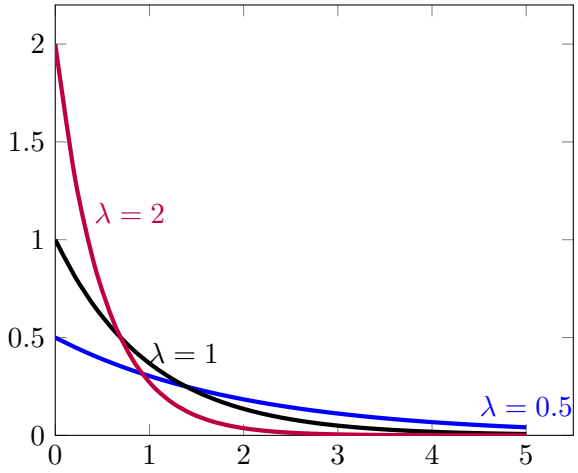
For a uniform distribution, $\mu = \frac{b+a}{2}$, $\sigma^2 = \frac{(b-a)^2}{12}$

3.3 Exponential Distribution

A continuous R.V. X is said to follow an exponential distribution with parameter $\lambda(\lambda > 0)$ if the pdf of X is:

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

An exponential distribution has $\mu = \frac{1}{\lambda}$, and $\sigma^2 = \frac{1}{\lambda^2}$.



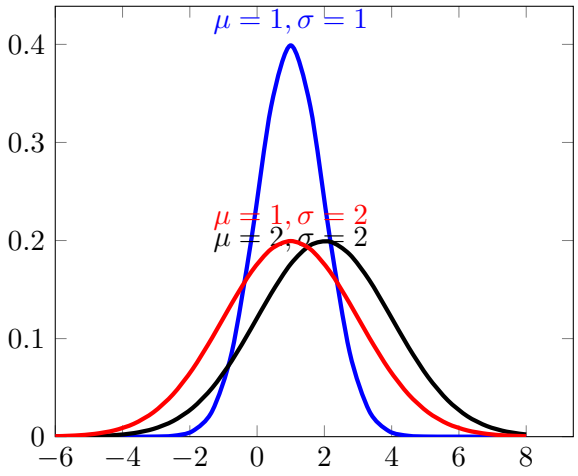
The exponential distribution has the memoryless property.

3.4 Gaussian / Normal Distribution

A continuous R.V. X is said to follow a Gaussian or normal distribution with parameter μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$ if the pdf of X is:

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

A normally distributed R.V. X with parameters μ and σ^2 can be abbreviated as $X \sim N(\mu, \sigma^2)$



3.4.1 Standard Normal Distribution

The normal distribution is more often used, which has $\mu = 0$ and $\sigma = 1$. A R.V. having a standard normal distribution is called a standard normal random variable and is denoted by Z. The pdf of Z is:

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

To transform $X \sim N(\mu, \sigma^2)$ into the standard normal distribution, we can do:

$$Z = \frac{X - \mu}{\sigma}$$

3.4.2 cdf of Standard Normal Distribution

The cdf of a standard normal R.V. Z is denoted by $\phi(z)$

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

The probability, $\phi(z)$, associated with different z under a standard normal distribution curve are available in a table. Calculating this can be done without a calculator!

3.4.3 z_α Notation

Notation: z_α denotes the value on the z axis for which α of the area under the z curve lies to the right of z_α.

z_α is the $100(1-\alpha)$th percentile of the standard normal distirbution.

3.4.4 Nonstandard to Standard

For $X \sim N(\mu, \sigma^2)$, we can write $P(a \leq zx \leq b)$ in terms of $\phi(.)$.

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

3.5 Cumulative Distribution Function

The cumulative distribution function (cdf) $F(x)$ of a continuous random variable R.V. X is defined for every number x:

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

For each x, $F(x)$ is the area under the density curve to the left of x.

3.5.1 Using F(x) to compute probability

Let $F(x)$ be the cdf of R.V. X. For any number a,

$$P(X > a) = 1 - F(a)$$

For any two numbers a and b,

$$P(a \leq X \leq b) = F(b) - F(a)$$

3.5.2 Obtaining f(x) from F(x)

Let X be a continuous R.V. with pdf $f(x)$ and cdf $F(x)$. At every x where the derivative $F'(x)$ exists,

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

3.6 Expected Value

The expected value or mean of a continuous R.V. X with pdf $f(x)$ is:

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx$$

provided that

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty$$

If the integral diverges, the expected value is undefined.

3.6.1 Expected value of a function

If X is a continuous R.V. with pdf $f(x)$ and $h(X)$ is any function of X, then

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

The expectation satisfies the linearity property.

3.7 Variance

The variance of a continuous R.V. X with pdf $f(x)$ and mean μ is:

$$V(X) = \sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx$$

Provided that

$$\int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx < \infty$$

3.8 Poisson Distribution as a Limit

Proposition: in a binomial distribution with pmf $b(x; n, p)$, let $n \to \infty$ and $p \to 0$ in such a way that np approaches $\lambda > 0$; then $b(x; n, p) \to p(x; \lambda)$, where $p(x; \lambda)$ is the pmf of a Poisson distribution.

For any binomial distribution where n is large ($n > 50$), and p is small $np < 5$, $b(x; n, p) \approx p(x : \lambda)$, where $\lambda = np$.

3.9 Poisson and Exponential Distributions

Poisson distribution is often used to model the number of occurrence of events in a time interval. Exponential distribution is often used to model the elapsed time between two successive events (e.g. the waiting time of a bus).

The two distributions are profoundly related!

Suppose the number of events occurring in a time interval t has a Poisson distribution with parameter $\lambda_p = \alpha t$ where α is the rate of occurrence (e.g. number of occurrence per hour). We want to study the probability that the waiting time for the first event is not more than t.

Let $T_1, T_2, \ldots$ be the time when 1st, 2nd, ... events occur. The probability of the waiting time for the first event not more than t is $P(T_1 \leq t)$.

$$P(T_1 \leq t) = 1 - P(T_1 > t)$$

$P(T_1 > t)$ suggests that no event occurs in $[0, t]$. Now we can apply the Poisson pmf:

$$P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - P(\text{no event in } [0, t]) = 1 - e^{-\alpha t}$$

This is exactly the cdf of the exponential distribution with $\lambda = \alpha$!

Taking the derivative w.r.t t:

$$f(t) = \alpha e^{-\alpha t}$$

Comparing this to the expression of the exponential pdf, $t = x$, $\alpha_{Exponential} = \alpha$.

Conclusion: the rate of occurrence α in the Poisson distribution is the parameter of the exponential distribution. α can be obtained from the average number of occurrence, $\lambda_{Poisson}$, and time interval t.

4 Week 4

4.1 Joint Probability Mass Function

Let X and Y be two discrete R.V.s defined on the sample space Ω of an experiment. The joint probability mass function $p(x, y)$ is defined for each pair of numbers (x, y) by

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

A legitimate joint pmf must satisfy

$$p(x, y) \geq 0 \text{ and } \sum_x \sum_y p(x, y) = 1$$

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

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

4.2 Marginal Probability Mass Function

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

$$p_X(x) = \sum_{y:p(x,y)>0} p(x, y)$$

for each possible value of x.
Similarly, the marginal probability mass function of Y is

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x, y)$$

for each possible value of y.

Marginal pmf of X is exactly the pmf of X. The word "marginal" indicates that the pmf is obtained from a joint probability distribution.

A legitimate marginal pmf must satisfy the same conditions as a pmf.

From a joint pmf, we can obtain the marginal pmf of each R.V.; however, the reverse is not always true. You can easily recover the joint pmf if they're independent.

4.3 Joint Probability Density Function

Let X and Y be two continuous R.Vs. A joint joint probability density function $f(x, y)$ of X and Y is a function satisfying

$$f(x, y) \geq 0 \text{ and } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) = 1$$

For any two dimensional set A

$$P[(X, Y) \in A] = \int \int_{A} f(x, y) dx dy$$

We can think of $f(x, y)$ as specifying a surface. $P[(X, y) \in A]$ is the volume beneath the surface above the region A.

If A is a two-dimensional region

$$(x, y) : a \leq x \leq b, c \leq y \leq d, \text{ then}$$
$$P[(X, Y) \in A] = \int_{a}^{b} \int_{c}^{d} f(x, y) dx dy$$

4.4 Marginal Probability Density Function

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

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

4.5 More than 2 R.Vs

If $X_1, X_2, \ldots, X_n$ are all discrete R.Vs, then the joint pmf of the variables is the function

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

If $X_1, X_2, \ldots, X_n$ are all continuous R.Vs, then the joint pdf of the variables is the function $f(x_1, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], [a_2, b_2], \ldots, [a_n, b_n]$

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots a_n \leq X_n \leq b_n) \tag{5}$$

$$= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(x_1, x_2, \ldots, x_n)\, dx_n \ldots dx_2 dx_1 \tag{6}$$

4.6 Independence of R.V.

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

$$p(x, y) = p_X(x) \cdot p_Y(y) \text{ for discrete R.V.}$$

$$f(x, y) = f_X(x) \cdot f_Y(y) \text{ for continuous R.V.}$$

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

Two R.Vs are independent if their joint pmf or pdf is the product of the two R.Vs' marginal pmf or pdf.

5 Week 5

5.1 Sampling Distribution of the Mean

The probability distribution of a statistic is called a **sampling distribution**. If $\bar{X}$ is a random variable representing sample mean, the proabbility distribution of $\bar{X}$ is called the **sampling distribution of the mean**.

The sampling distribution of $\bar{X}$ with sample size n is the distribution of the many $\bar{X}$ values that arise from conducting an experiment repeatedly, always with sample size n. This describes the *variability* of sample averages around the population mean μ.

Proposition Let $X_1, X_2, \ldots, X_n$ be a random sample from a *normal distribution* with mean μ and variance σ^2. Then, for any n,

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

2. $W = \sum_{i=1}^{n} X_i$, is normally distributed with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$.

Proof

$$\begin{aligned}\text{Var}[W] &= \text{Var}[\sum_{i=1}^{n} X_i] \\ &= \text{Var}[X_1] + \text{Var}[X_2] + \dots \\ &= n\text{Var}[X_1] \\ &= n\sigma^2\end{aligned}$$

Hence, standard deviation $\sigma = \sqrt{n}\sigma$.

$$\begin{aligned}E[W] &= E[\sum_{i=1}^{n} X_i] \\ &= E[X_1] + E[X_2] + \dots \\ &= nE[X_1] \\ &= n\mu\end{aligned}$$

The distribution of W becomes more spread out as the sample size n increases, while the distribution of $\bar{X}$ becomes more concentrated about the μ as n increases.

The standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ is often called the **standard error of the mean**, it indicates the magnitude of a typical deviation of the sample mean from the population mean.

5.2 Central Limit Theorem

Let $X_1, X_2, \dots X_n$ be a random sample from *any* distribution with mean μ and variance σ^2. Then, if n is sufficiently large:

1. $\bar{X}$ has approximately a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$
2. W also has approximately a normal distribution with $\mu_W = n\mu$ and $\sigma^2_W = n\sigma^2$

The larger the value of n, the better the approximation. CLT is regarded as one of the most important theorems of probability!

5.2.1 The Median Theorem

For fixed quantiles of a R.V. X (such as the median), they have asymptotic normality.

Meaning for large enough samples, the medians of the samples follow approximately a normal distribution. The mean of the sample medians will provide an estimate of the population median.

6 Week 6

6.1 Point Estimation

Statistical inference is concerned primarily with understanding the quality of parameter estimates.

The sample mean is a **point estimate** of the population mean.

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to:

$$SE = \frac{\sigma}{\sqrt{n}}$$

In general, a point estimate of a parameter θ is a single number that can be regarded as a *sensible* value for θ. A point estimate is obtained by selected a suitable statistic and computing its values from the given sample data. This is called the **point estimate** of θ.

6.1.1 Unbiased Estimator

A point esimator $\hat{\theta}$ is said to be an **unbiased esimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$.

6.1.2 Principle of Minimum Variance Unbiased Estimation

Among all the estimators of θ that are unbiased, choose the one that has the *minimum* variance. The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

6.1.3 The Method of Moments

Basic idea: equate certain sample characteristics, such as the mean, to the corresponding population expected values. Solving these equations for unknown parameters yields the

estimators.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a pdf (or pmf) $f(x)$. For $k = 1, 2, \ldots$, the k-th population (theoretical) moment μ_k is

$$\mu_k = E[X^k]$$

The k-th sample moment M_k is

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a pdf (or pmf) $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots \theta_m$ are parameters (e.g. mean, variance, etc) whose values are unknown.

Then, the **method of moment estimators** (MME) $\hat{\theta_1}, \ldots, \hat{\theta_m}$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \ldots, \theta_m$.

6.2 Test of Hypotheses

A test of hypothesis is a method to decide whether to **reject** H_0 or **fail to reject** H_0. This contains two components:

1. **Test Statistic** = A function of the sample data used to make a decision.
2. **Rejection region** = A set of all test statistic(s) values for which H_0 will be rejected.

The null hypothesis always contains the equality sign. The alternative hypothesis is true if the null hypothesis is false.

6.2.1 Type I and II Errors

A **type I** error involves rejecting the null hypothesis H_0, when it is actually true. The maximum probability of type I error that can be tolerated is usually denoted by α, the significance level.
A **type II** error involves failing to reject the null hypothesis H_0 when the alternative hypothesis H_a is actually true. The probability of type II error is denoted by β.

7 Week 8

7.1 Hypothesis Testing

General Steps

1. Identify parameter of interest, describe in context of problem.
2. Determine and state null hypothesis.
3. State appropriate alternative hypothesis.
4. Give the formula for the computed value of test statistic (substituting known value of parameters, but not sample quantities)
5. State rejection region for selected significance level α.
6. Compute necessary sample quantities and substitute. Compute statistic value.
7. Decide whether H_0 should be rejected, and state conclusion in problem context.

7.2 Calculating Probability of β and Sample Size Determination

Consider the upper-tailed test, where H_0 is rejected iff $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$, i.e. when $\bar{X} > \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$. Let μ' denote a particular value of μ, s.t. $\mu' > \mu_0$. Then,

$$\begin{aligned}
\beta(\mu') &= P(H_0 \text{ is not rejected when } \mu = \mu') \\
&= P(\bar{X} \leq \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ when } \mu = \mu') \\
&= P(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} + z_\alpha \text{ when } \mu = \mu') \\
&= \phi(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}})
\end{aligned}$$

7.3 Determining Sample Size

To determine the sample size, observe that two requirements need to be met:

$$P(\text{type I error}) = \alpha \text{ and } \beta(\mu') = \beta$$

To meet first requirement, set $c = z_\alpha$. To meet the second, draw a sample size n s.t.

$$\beta = \phi(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}})$$

This implies that $-z_\beta = z$ critical value that captures lower-tail area. Solving for n:

$$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2$$

This result is for an upper-tailed or lower-tailed test only. For a two-tailed test, an approximate solution is:

$$n = \left\lceil \frac{\sigma(z_{\alpha/2} + z_{\beta})}{\mu_0 - \mu'} \right\rceil^2$$

7.4 Calculating the p-value

The p-value is the probability of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.

- The smaller the p-value, the more evidence there is in the sample against the null hypothesis. (H_0 should be rejected in favor of H_a when the p-value is sufficiently small.
- Decision based on p-value involves selecting a significance level α. H_0 is rejected if p-value $\leq \alpha$

7.5 Linear Regression

The simplest relationship between two variables x and y is a linear relationship: $y = \beta_0 + \beta_1 x$.

If we account for uncertainty, a more general model is:

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

An observed value of Y is deterministic, and will be denoted by y. It is referred to as a **dependent** or **response variable**.

Definition: There are parameters β_0, β_1 and σ^2 such that x (independent variable) and y (dependent variable) are related through the **model equation**

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ is called **random deviation** or **random error term** in the model.

The line $y = \beta_0 + \beta_1 x$ is called the **true** (or **population**) **regression line**.

7.5.1 When to use Linear Regression

- Theory suggests that x and y have a linear relationship with ϵ representing measurement error
- The scatter plot exhibits a linear pattern

7.6 Distribution

Random observation y corresponds to the particular value of $x = x^*$ equals to $y = \beta_0\beta_1 x^* + \epsilon$. Hence, Y has a normal distribution, with mean

$$E[\beta_0 + \beta_1 x^* + \epsilon] = \beta_0 + \beta_1 x^* + E[\epsilon] = \beta_0 + \beta_1 x^*$$

and variance

$$\text{Var}[\beta_0 + \beta_1 x^* + \epsilon = \text{Var}[\beta_0 + \beta_1 x^*] + \text{Var}[\epsilon] = \sigma^2$$

Note: This mean and variance is obtained for a specific value $x = x^*$.

7.7 Least Square Principle

The discreprency between the line $y = b_0 + b_1 x$ and the points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ is given by

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, are called the **least square estimates**. In other words, $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1) \quad \forall b_0, b_1$.

The **estimated regression line** or **least squares line** is then

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

7.8 Normal Equations

To find $\hat{\beta}_0$ and $\hat{\beta}_1$, take the partial derivatives of $f(b_0, b_1)$ w.r.t. b_0 and b_1, set them to 0, and solve for b_0 and b_1. This gives:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2/n$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

7.8.1 Estimate σ^2

σ^2 leads us to determine the amount of variability in a regression model. Small σ^2 leads to the observed (x_i, y_i) being close to the regression line.

The **error sum of squares** is

$$\text{SSE} = \sum (y_i - \hat{y_i})^2 = \sum_i [y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)]^2$$

And an estimate for σ^2 is

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

7.8.2 Coefficient of Determination

The **total sum of squares**, SST, gives a quantitative measure of the total amount of variation in observed y values:

$$\text{SST} = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

The **coefficient of determination**, denoted by r^2 or R^2, is given by

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

This is interpreted as the proportion of observed y variation that can be explained by the simple linear model.

7.9 Managing Data

Right-skewed data: use log operation to make it more normal.
Left-skewed data: square or cube data to make it more normal.

8 Week 10

The long term goal of the next lectures: Given a set of data points $x_1, \ldots, x_n$, and a set of probability distributions, P_θ which can be used to assign a probability $P_\theta(x_1, \ldots, x_n)$ to the dta points, and which depend on a parameter θ, find a value of θ such that the probability function P_θ is a good model for the data points.

8.1 Math Concepts

- Compute a probability as a sum of a discrete set A:

$$P(A) = \sum_{(x_1,x_2,x_3)\in A} P(x_1, x_2, x_3)$$

- Find normalizing constant when probability density is given proportional to function, i.e. what is c?

$$c\int_0^1 xe^{-x}dx = 1$$

- Compute probability over set A when we have a density. For example, given $A = (x_1, x_2) : x_1 \leq x_2, x_i \in [0, 1]$

$$P(A) = \int_A 4x_1x_2dx_1dx_2 = ?$$

- Compute expectations, higher moments for joint distributions

$$E[X_2] = \int_A x_2 f(x_1, x_2, x_3)dx_1dx_2dx_3$$

$$E[X_3^2] = \int_A x_3^2 f(x_1, x_2, x_3)dx_1dx_2dx_3$$

8.2 Joint and Marginal, 2 Variable Discrete Case

Let X_1 and X_2 be two **discrete R.V.**. The **joint probability mass function** $p(x_1, x_2)$ is defined for each pair of experiment outcomes (x_1, x_2) by

$$p(x_1, x_2) = P(X_1 = x_1 \text{ and } X_2 = x_2) = P(X_1 = x_1, X_2 = x_2)$$

x_1 and x_2 satisfy the following properties:

$$p(x_1, x_2) \geq 0$$

$$\sum_{x_1}\sum_{x_2} p(x_1, x_2) = 1$$

For two variables, we can represent this in a table.

8.2.1 Joint Probability Over a Set

Let A be any set consisting of pairs of (x_1, x_2) values. Then the probability $P((X_1, X_2) \in A)$ is obtained by summing the joint pmf over all pairs (x_1, x_2) in A:

$$P((X_1, X_2) \in A) = \sum_{(x_1,x_2)\in A} p(x_1, x_2)$$

8.2.2 Marginal Probability Distribution

The **marginal probability mass function** of X_1, denoted by $p_{x_1}(x_1)$, is given by the following for each possible value of x_1:

$$p_{X_1}(x_1) = \sum_{x_2} p(x_1, x_2)$$

Similarly, for X_2:

$$p_{X_2}(x_2) = \sum_{x_1} p(x_1, x_2)$$

The importance for marginal probabilities is three-fold:

1. They provide probabilities defined over a subset of variables.
2. They are used to help define conditional probabilities.
3. They can be used to validate independence.

8.3 Joint and Margin, 2 Variable Continuous Case

Note: even if a set is continuous, you can have a subset that is discrete!

8.3.1 Joint Probability Over a Set

Let X_1 and X_2 be **continuous R.V.**. Then their joint probability density function is a function $f(x_1, x_2)$ satisfying

$$f(x_1, x_2) \geq 0$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)dx_1dx_2 = 1$$

For any two-dimensional set A the probability to see samples (x_1, x_2) in a set A is defined as:

$$P((X_1, X_2) \in A) = \int_A f(x_1, x_2)dxdy = \int_{(x_1,x_2)\in A} f(x_1, x_2)dx_1dx_2$$

In particular, if A is a two-dimensional rectangle $[a_1, b_1] \times [a_2, b_2]$, then

$$\begin{aligned} P((X_1, X_2) \in A) &= P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2) \\ &= \int_{x_1=a_1}^{x_1=b_1} \int_{x_2=a_2}^{x_2=b_2} f(x_1, x_2)dx_1dx_2 \end{aligned}$$

Comparing with the discrete case, we replace summation by integration.

8.3.2 Margin Probability Distribution

The marginal probability density functions of X_1 and X_2, denoted by $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, respectively, are given by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2$$
$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1$$

The get the marginal in the variable of interest, we must sum/integrate over the variable that we are not interested in.

8.4 Joint Distribution: n Variables

Suppose we have n variables: $X_1, X_2, \ldots X_n$. Then, the joint and marginal distributions are defined analogously.

For discrete R.V., the joint distribution is given by:

$$P((X_1, X_2, \ldots, X_n) \in A) = \sum_{(x_1,x_2,\ldots,x_n)\in A} p(x_1, x_2, \ldots, x_n)$$

For continuous R.V., the joint distribution is given by:

$$P((X_1, X_2, \ldots, X_n) \in A) = \int_A f(x_1, x_2, \ldots, x_n)dx_1dx_2 \ldots dx_n$$

When A is a product of intervals, $A = [c_1, d_1] \times [c_2, d_2] \times \ldots [c_n, d_n]$, the last integral can be written

$$\int_{x_1=a_1}^{b_1} \int_{x_2=a_2}^{b_2} \dots \int_{x_n=a_n}^{b_n} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

For sets A that do not have such shapes, it can be computed using the **integral substitution theorem** for n dimensions.

8.5 Marginal Distribution: n Variables

For a pmf $P(X_1, X_2, \dots, X_n)$, and pdf $f(X_1, X_2, \dots, X_n)$, we can compute marginals with $n-1, n-2, n-3, \dots, 3, 2, 1$ variables.

For example, in the discrete case:

$$P(X_2, \dots, X_n) = \sum_{X_1} P(X_1, X_2, \dots, X_n)$$

and in the continuous case:

$$f(x_2, \dots, x_n) = \int_{x_1 \in \text{Def}(X_1)} f(x_1, x_2, \dots, x_n) dx_1$$

where $\text{Def}(X_1)$ is the space on which X_1 is defined. This density can be used to compute a joint probability over a set of variables $(X_2, X_3, \dots, X_n)$ as done above.

Whether a probability density is marginal or joint density depends on the set of variables you consider. For example, $f(x_2, \dots, x_n)$ is a marginal density for variables $(X_1, X_2, \dots, X_n)$, but a joint density of variable set $(X_2, \dots, X_n)$.

For a k-dimensional marginal, there exist $\binom{n}{k}$ possibilities. For a k-dimensional joint distribution, there exist $\binom{\sum_{i=0}^{k} n}{i=2^n}$.

8.6 Joint Cumulative Distribution: 2 Variables

If X_1 and X_2 are continuous R.V., then the joint cumulative distribution F of two variables is given as:

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = \int_{y_1=-\infty}^{x_1} \int_{y_1=-\infty}^{x_1} f(y_1, y_2) dy_1 dy_2$$

The relationship between joint cdf and joint pdf is:

$$\frac{\partial^2 F}{\partial x_1 \partial x_2}(x_1, x_2) = f(x_1, x_2)$$

8.7 Expectation of a function of N variables $h(X_1, \ldots, X_n)$

The expected value of a unfction $h(X_1, X_2)$, is as follows:

$$E[h(X_1, X_2)] = \sum_{x_1} \sum_{x_2} h(x_1, x_2) \cdot p(x_1, x_2)$$

or for continuous R.V.:

$$E[h(X_1, X_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x_1, x_2) \cdot p(x_1, x_2) dx_1 dx_2$$