STA247 Notes

Jenci Wei

Fall 2021

Lecture 1A - New Terminology & Event Relations

Terminology

Random experiment: Process that allows us to gather data or observations

- Can be repeated multiple times provided conditions do not change where outcomes are **random**
- Set of possible outcomes are known
- Outcome of a specific experiment is not known

Sample space: the set of all possible outcomes from a random experiment

- Denoted Ω or S
- Elements are determined by the *outcome of interest*

Event: a set of outcomes, subset of the sample space

- Denoted by an uppercase letter, e.g. A
- Simple event: 1 outcome; compound event: > 1 outcomes

Complement event: set of outcomes in Ω and are *not* in A

- Denoted $A^c, \bar{A}, A'$

Operations on Sets

Union of two events A and B is the set of outcomes that are elements of A, B, or both

- Denoted $A \cup B$
- E.g. $A \cup A^c = \Omega$

Intersection of two events A and B is the set of outcomes that are common to both A and B

- Denoted $A \cap B$, AB
- E.g. $A \cap A^c = \emptyset$

Events A and B are **disjoint** if their intersection is empty

Event Relations

Two events A and B are **mutually exclusive** if the events cannot both occur as an outcome of the experiment

- A and B are also called **disjoint** events
- i.e. $A \cap B = \emptyset$

Two events A and B are **independent** if the occurrence of one event does not alter the probability of occurrence of the other

Commutative, Associative, and Distributive Laws

Commutative law: $A \cup B = B \cup A$

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

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

DeMorgan's Laws

For two events A and B:

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

For the set events $\{A_1, A_2, \ldots, A_n\}$

- $\left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} A_i^c$
- $\left(\bigcap_{i=1}^{n} A_1 \right)^c = \bigcup_{i=1}^{n} A_i^c$

Lecture 1B - What is Probability?

Intro

In a random experiment with sample space Ω, the **probability** of an event A, denoted $P(A)$, is a function that measures the chance that event A will occur

Certain *axioms* that must hold for probability functions:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. For a set of disjoint events $A_1, \ldots, A_n$ in Ω, $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

Probability Functions

Suppose the sample space Ω can be represeented with a finite sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$ or a countably infinite sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$, then the probability function P is a function on Ω with the following properties:

1. $P(\omega) \geq 0$ for all $\omega \in \Omega$
2. $\sum_{\omega \in \Omega} P(\omega) = 1$
3. For all events $A \subseteq \Omega$, $P(A) = \sum_{\omega \in A} P(\omega)$

Probability and Event Relations

Complement:

- $P(\Omega) = P(A \cup A^c)$
- $1 = P(A) + P(A^c)$
- $P(A) = 1 - P(A^c)$

$P(A \cup B) = P(A) + P(B)$ if A and B are disjoint since $P(A \cap B) = 0$

Inclusion-Exclusion Principle:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) = &\sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots \\ &+ (-1)^{r+1} \sum_{i_1 < I_2 < \ldots < i_r} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) + \cdots \\ &+ (-1)^{n+1} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) \end{aligned}$$

Lecture 2A - Counting Permutations

Permutations

The number of ways to order n *distinct* items is

$$n! = n \cdot (n-1) \cdots 2 \cdot 1$$

Then number of ways to select an ordered subset of k items from a group of n *distinct* items is

$${}_nP_k = \frac{n!}{(n-k)!} = n \cdot (n-1) \cdots (n-k+2) \cdot (n-k+1)$$

Lecture 2B - Counting Combinations

Combinations

The number of ways to select an *unordered* subset of k items from a group of n *distinct* items without replacement is

$$\binom{n}{k} = {}_nC_k = \frac{n!}{(n-k)! \cdot k!}$$

- Divide by $(n-k)!$ to remove the ways of ordering the remaining items
- For every ${}_nP_k$ ordering of distinct objects, there exists $k!$ orderings of the same collection of k objects; thus, divide $k!$

Lecture 3A - Conditional Probability and Independence

Conditional Probability

$P(A|B)$ - the probability of A given the condition that event B has occurred

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{Provided that } P(B) > 0$$
$$P(A \cap B) = P(A|B) \cdot P(B)$$
$$P(A \cap B) = P(B|A) \cdot P(A)$$

Conditional probabilities are probability distributions on a *restricted sample space*

Consider a random experiment with sample space Ω. Let B be an event with $P(B) > 0$. Let b denote the elements of event B. Then:

1. $P(b|B) \geq 0 \quad \text{for all } b \in B$
2. $\sum_{b \in B} P(b|B) = 1$
3. For $A \subseteq B, P(A|B) = \sum_{b \in A} P(b|B)$

Independent Events

Two events A and B are **independent** if the occurrence of A *does not alter the chances* of B, i.e.:

$$P(A|B) = P(A), \quad \text{provided that } P(B) > 0$$
$$P(B|A) = P(B), \quad \text{provided that } P(A) > 0$$

When event A is independent of event B, then $P(A \cap B) = P(A) \cdot P(B)$

Mutually exclusive: the occurrence of A excludes the occurrence of B

- $P(A \cap B) = 0$
- The events are dependent

If events A and B are independent, then so are their complements, A^c and B^c

For a collection of n events, $A_1, \ldots, A_n$:

- If all n events are independent, then:

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \times \cdots \times P(A_n)$$

- $A_1, \ldots, A_n$ are **mutually independent** if for any subset of k events, $k = 2, \ldots, n$:

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \times \cdots \times P(A_{i_k})$$

Lecture 3B - Law of Total Probability and Bayes Rule

Law of Total Probability

Suppose we have a sample space consisting *only* of events $A, B_1, B_2, \ldots, B_k$ where B_is *partitions* the sample space, i.e. the B_is are disjoint and $\bigcup_{i=1}^{k} B_i = \Omega$. Then:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_k) \\ &= P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) + \cdots + P(A|B_k) \cdot P(B_k) \end{aligned}$$

Law of Total Probability: if $B_1, \ldots, B_k$ is a collection of mutually exclusive (i.e. disjoint) and exhaustive events, then for any event A:

$$P(A) = \sum_{i=1}^{k} P(A|B_i) \cdot P(B_i)$$

Bayes' Rule

Let $B_1, \ldots, B_k$ form a partition of the sample space and let A be an event in Ω. Then:

$$\begin{aligned} P(B_i|A) &= \frac{P(A \cap B_i)}{P(A)} \\ &= \frac{P(A|B_i) \cdot P(B_i)}{\sum_{i=1}^{k} P(A|B_i) \cdot P(B_i)} \end{aligned}$$

Lecture 4A - Intro to Discrete RVs - PMF

Random Variable

A **random variable** is a real-valued function that assigns a numerical value to each event in Ω arising from a random experiment

A random variable X is a function $X : \Omega \to \mathbb{R}$ such that $\forall \omega \in \Omega, X(\omega) = x \in \mathbb{R}$.

Converts each outcome into a number

Support of X: the 'domain' of X

Discrete Random Variable

A **discrete** of a random variable X is one that can take on only a finite number or a countably infinite number of possible values x.

A random variable X is **continuous** if its domain is an interval of real numbers

Proabability Mass Function

A **probability mass function** (PMF) of a discrete random variable is one that assigns a probability to each value $x \in X$ such that:

1. $0 \leq P(X = x) \leq 1$
2. $\sum\limits_{x \in X} P(X = x) = 1$

Lecture 4B - Features of a Distribution

Characteristics of Random Variables

Expected Value: the theoretical average

- If a random experiment were to be conducted n times, then as $n \to \infty$, the *average* of outcomes converges to the expected value
- Denoted μ
- $\mu = E(X) = \sum_{x \in X} x \cdot P(X = x)$
- For a given transformation of X, $g(X)$, then

$$E[g(X)] = \sum_{x \in X} g(x) \cdot P(X = x)$$

- $E[g(X)] \neq g(E(X))$ except when $g(X)$ is a linear transformation

Variance or Standard Deviation: measures of the spread and variability of a random variable

- Standard deviation is the *square root* of variance
- Variance is denoted as σ^2
- Standard deviation is denoted as σ
- $\sigma^2 = V(X) = E[(X - \mu)^2] = \sum_{x \in X} (x - \mu)^2 \cdot P(X = x)$
- Variance captures the spread in units^2
- Standard deviation has same units as the random variable X

Properties of Expectation

For any constant a, b, and discrete random variable X, then:

- $E[a] = a$
- $E[X + a] - E[X] + a = \mu + a$, i.e. increasing in $x \in X$ will shift the average by the same amount
- $E[aX] = a \cdot E[X] = a \cdot \mu$
- $E[aX + b] = a \cdot E[X] + b = a \cdot \mu + b$
- $E[X + Y] = E[X] + E[Y]$
- $E[XY] \neq E[X] \cdot E[Y]$ *unless* X and Y are independent
- $V(a) = 0$ since constants do not vary
- $V(a+X) = V(X) = \sigma^2$, i.e. increasing each $x \in X$ will not change how spread out the random variable is
- $V(aX) = a^2 \cdot V(X) = a^2 \cdot \sigma^2$
- $V(aX + b) = a^2 \cdot V(X) = a^2 \cdot \sigma^2$
- $V(X + Y) \neq V(X) + V(Y)$ *unless* X and Y are independent

Variance

WE can calculate the variance of a discrete random variable X with PMF $f(x)$:

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

where $E[X^2] = \sum\limits_{x \in X} x^2 \cdot f(x)$

Lecture 4C - The Cumulative Distribution Function

Cumulative Distribution Function

The **cumulative distribution function** (CDF) $F(x)$ of a discrete random variable with probability mass function $P(x)$ or $f(x)$ is the cumulative probability up to and including $X = x$ ("left tail probability")

$$F(b) = P(X \leq b) = \sum_{x \in \{x \leq b\}} P(x)$$

where $F(b)$ is the CDF at $X = b$

For a discrete random variable X with CDF $F(X)$:

- The graph of the CDF will be a *non-decreasing step-function*, i.e. for $a < b, F(a) \leq F(b)$
- The graph of the CDF is *right continuous*, i.e. $\lim\limits_{x \to c^+} F(x) = F(c)$
- $\lim\limits_{x \to \infty} F(x) = 1$
- $\lim\limits_{x \to -\infty} = 0$

Lecture 5A - Using Features to Describe Probable Outcomes

Markov's Inequality

Let X be a non-negative RV with mean $E(X)$. Then for some constant $a > 0$:

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Chebyshev's Inequality

Let X be an RV with mean μ and finite variance σ^2. Then for any positive k

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

- Chance of observing an RV outcome X that is a value in the $(\mu - k\sigma, mu + k\sigma)$ interval
- As k increase, the interval expands
- Estimation is usually conservative

Lecture 5B - Bernoulli and Binomial Random Variables

Common Discrete Distributions

A **Bernoulli trial** is a random experiment consisting of exactly 1 trial involving 2 possible outcomes (i.e. *success* or *failure*), Let X be the outcome of a Bernoulli trial where

- $X = 1$ if the outcome is a success
- $X = 0$ if the outcome is a failure

We define p to be the probability of success, and $q = 1 - p$ to be the probability of failure. The *probability mass function* is

$$f(x) = p^x \cdot (1-p)^{1-x}$$

A **Binomial experiment** consists of n independent and identical Bernoulli trials

- The probability of success, p, is fixed for each trial

Let X be the RV representing the *number of successes* among the n trials. Then X can be modelled by the **binomial distribution** withg parameters n and p, denoted as $X \sim Bin(n, p)$. The binomial distribution has PMF:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Binomial Distribution

Let $X \sim Bin(n, p)$. Then $E(X) = np$ and $V(X) = np(1-p)$

Lecture 5C - Negative Binomial and Hypergeometric Distributions

Negative Binomial Distribution

The discrete RV X that models the *number of failed independent Bernoulli trials* to achieve a fixed, predefined number of successes r has PMF

$$P(X = x) = \binom{x + r - 1}{x} p^r (1 - p)^x$$

with expected value $E(X) = \frac{r(1-p)}{p}$ and $V(X) = \frac{r(1-p)}{p^2}$

- Total: $x + r$
- r successes, x failures
- The final trial i the rth success

A related variable Y can model the *total number of independent Bernoulli trials* instead with PMF:

$$P(Y = y) = \binom{y - 1}{r - 1} p^r (1 - p)^{y-r}$$

with expected value $E(Y) = \frac{r(1-p)}{p} + r$ and variance $V(Y) = \frac{r(1-p)}{p^2}$

- $Y = X + r$
- Final trial is the rth success

Notation: $X \sim NB(r, p)$.

The case of *number of failures* to achieve the first success is the special case of the negative binomial distribution with $r = 1$ an dis modelled by the *Geometric* distribution

Geometric distribution: X is a discrete RV representing the number of failed independent and identical Bernoulli trials *before* the first success is achieved, with probability of success being noted by p and probability of failure by $q = 1 - p$. The PMF is given by

$$P(X = x) = q^x p$$

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

- Has **memoryless property** where $P(X \geq a + b | X \geq a) = P(X \geq b)$

Hypergeometric Distribution

A discrete RV X represents the number of desirable objects in a random sample of n from a finite pool of N objects, of which M are desirable. The PMF of X is

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

with expected value $E(X) = \frac{nM}{N}$ and variance $V(X) = n\left(\frac{N-n}{N-1}\right)\frac{M}{N}\left(1 - \frac{M}{N}\right)$

- When $N, M, N - M$ are large relative to n, the hypergeometric distribution can be approximated by the binomial distribution

Lecture 6A - Intro to Continuous RVs

Probability Density Function

The **probability density function** (PDF) of a cts RV X is a function $f(x)$ that has the following properties:

1. $f(x) \geq 0$ for all x in the support of X
2. $\displaystyle\int_{\infty}^{-\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \displaystyle\int_{a}^{b} f(x)dx$

- $f(x) \neq P(X = x)$ for a cts RV X
- The area under the PDF correspons to the probability over the interval
- $f(x)$ not upper bounded, can be > 1

Memoryless Property

$P(X \geq a + b | X \geq a) = P(X \geq b)$ for constants a and b in the support of X

Cumulative Distribution Function

The **cumulative distribution function** (CDF) of a cts RV X is the function $F(x)$ s.t.

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

The derivative of the CDF $F'(x)$ is the PDF of X, i.e. $F'(x) = f(x)$

Properties of CDF:

- $P(X = c) = \displaystyle\int_{c}^{c} f(x)dx = 0$
- $P(X \leq c) = P(X < c)$
- $P(a \leq X \leq b) = \displaystyle\int_{a}^{b} f(x)dx = F(b) - F(a)$
- $\displaystyle\lim_{x \to \infty} F(x) = 1$
- $\displaystyle\lim_{x \to -\infty} F(x) = 0$

Note:

- $f(x)$ usually denotes PMF or PDF
- $F(x)$ usually denotes CDF, which is $P(X \leq x)$

CDF and Percentiles

The kth percentile is for which $k\%$ of values are less than or equal to it.

For a random RV X with PDF $f(x)$, the kth percentile x_k is:

$$\frac{k}{100} = P(X \leq x_{k/100}) = \int_{-\infty}^{x_{k/100}} f(x)dx = F(x_{k/100})$$

$$x_{k/100} = F^{-1}\left(\frac{k}{100}\right)$$

Special percentiles:

- *Median*: 50th percentile
- *first quartile* and *third quartile*: 25th and 75th percentiles, respectively

Interquartile range (IQR): difference between the 75th percentile and 25th percentile

Lecture 6B - Features of Continuous Distributions

Expected Value

The **expected value** (aka **mean**) of a cts RV X with PDF $f(x)$ is given by

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

For any real-valued function $g(X)$ of X:

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

Variance

The **variance** of a cts RV X with PDF $f(x)$ is given by

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

And

$$V(X) = E(X^2) - E(X)^2$$

- The variance is the average *squared* deviation of X from its mean
- The **standard deviation** of a RV is the square root of the variance (SD = $\sqrt{V(X)}$)
- Notation: σ^2 for variance and σ for standard deviation

Properties of E(X) and V(X)

For any two RVs X and Y and constants a and b:

- $E(aX + b) = aE(X) + b$
- $E(aX + bY) = aE(X) + bE(Y)$
- $E(XY) = E(X)E(Y)$ only if X and Y are *independent*
- $V(aX + b) = a^2 V(X)$
- $V(aX + bY) = a^2 V(X) + b^2 V(Y)$ only if X and Y are *independent*

Chebyshev's Inequality

For a RV X with expected value $\mu = E(X)$ and finite variance $\sigma^2 = V(X)$:

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Lecture 7A - Uniform and Exponential Distributions

Common Continuous RV - Uniform

A cts RV X follows a **uniform distribution** on the interval $a \leq X \leq b$ if it has PDF:

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

- Distribution with constant density
- Ay interval of the same width in the support have equal probability of occurrence

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

Poisson Distribution

A discrete RV X denoting the number of events of interests in an interval, with λ the average rate of occurrences *per unit interval.* PMF:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Expectation and variance are both λ

Models the *quantity* of arrivals in an interval

Common Continuous RV - Exponential

A cts R X is **exponentially distributed** with *mean parameter* $\theta > 0$ (or rate of $\lambda > 0$) if it has the PDF

$$f(x) = \begin{cases} \frac{1}{\theta}e^{-\frac{x}{\theta}} \quad \text{or} \quad \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{elsewise} \end{cases}$$

Mean and variance:

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

We say $X \sim Exp(\theta)$ or $X \sim Exp(\lambda)$ to define the distribution

- θ: average X until occurrence, e.g. 5 min per customer
- λ: average number of occurrence per interval, e.g. 12 customers per hour

CDF:

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-\frac{x}{\theta}}, & \text{if } x > 0 \\ 0, & \text{elsewise} \end{cases}$$

Memoryless property:

$$\begin{aligned}
P(X \geq a+b|x \geq a) &= \frac{P(X \geq a+b \cap X \geq a)}{P(X \geq a)} \\
&= \frac{P(X \geq a+b)}{P(X \geq a)} \\
&= \frac{1-P(X < a+b)}{1-P(X < a)} \\
&= \frac{1-F(a+b)}{1-F(a)} \\
&= \frac{1-\left(1-e^{-\frac{a+b}{\theta}}\right)}{1-\left(1-e^{-\frac{a}{\theta}}\right)} \\
&= e^{-\frac{b}{\theta}} \\
&= P(X \geq b)
\end{aligned}$$

Lecture 8A - Normal Distributions

Normal Distribution

A normal distribution with **location parameter** μ and **scale parameter** $\sigma^2 > 0$ for a cts RV X has PDF:

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

We say $X \sim N(\mu, \sigma^2)$ if X is normally distributed with parameters μ and σ^2

Properties:

- $E(X) = \mu$ is the centre of the distribution
- $V(X) = \sigma^2$ is the spread of the distribution
  - Larger variance means flatter and wider bell
  - Smaller variance means taller and narrower bell
- The normal distribution is *symmetric* about the mean μ, i.e. $P(X \geq \mu) = P(X \leq \mu) = 0.5$
- The **standard normal** is the normal distribution with $\mu = 0$ and $\sigma^2 = 1$, and is represented as $Z \sim N(0, 1)$
- The CDF of a $N(\mu, \sigma^2)$ distribution $P(X \leq c) = \int_{-\infty}^{c} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$ has no closed form solution and is denoted $\Phi(c)$
  - Compute in R using `pnorm()`
  - Use a normal probability table to compute
- The linear combination of normal RVs (independent or not) results in a normal RV
- Every normal RV $X \sim N(\mu, \sigma^2)$ is a linear transformation of the *standard normal* $Z \sim (0, 1)$:

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

  This transformation is called **standardizing a normal RV**
- Since $Z \sim N(0, 1)$, all z-scores are equivalently expressing outcomes in terms of its distance from mean, in units of standard deviation
- Can be used for **continuity correction**, which treats a discrete RV as a normal RV

Sum of two normal RV If we have $W = aX + bY$ where

- $X \sim N(\mu_1, \sigma_1^2)$
- $Y \sim N(\mu_2, \sigma_2^2)$
- $a, b \in \mathbb{R}$

Then

$$\begin{aligned} E(W) &= a\mu_1 + b\mu_2 \\ V(W) &= a^2V(X) + b^2V(Y) \\ W &\sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2) \end{aligned}$$

Lecture 8B - Moment Generating Functions

Moment

Moments of a RV are the expected values of powers of the RV, e.g. $E(X^k)$ is called the **kth moment** of X

- The expected value is $E(X^1)$ and is also known as the *1st moment*
- The variance can be computed using $E(X^2) - E(X)^2$, thus the variance is a *function of the 1st and 2nd moments* but is not a moment itself

The **moment fenerating function** (MGF) is defined to be:

$$M(t) = E(e^{tX}) = \begin{cases} \int_{x \in X} e^{tx} f(x) dx, & \text{if } X \text{ is cts} \\ \sum\limits_{x \in X} e^{tx} f(x), & \text{if } X \text{ is discrete} \end{cases}$$

If $M(t)$ exists and is differentiable in the neighbourhood of $t = 0$, then we can generate the kth moment of X, $E(X^k)$ by:

$$E(X^k) = M^{(k)}(0)$$

We can use the MGF to find any kth moment of X by using Leibniz's integral rule:

- If we have a function $f(x, t)$, then

$$\frac{d}{dt} \int_a^b f(x, t) dx = \int_a^b \frac{d}{dt} f(x, t) dx$$

- First moment

$$\begin{aligned} M'(t) &= \frac{d}{dt} M(t) \\ &= \frac{d}{dt} \int_{x \in X} e^{tx} f(x) dx \\ &= \int_{x \in X} \frac{d}{dt} e^{tx} f(x) dx \\ &= \int_{x \in X} x e^{tx} f(x) dx \\ M'(0) &= \int_{x \in X} x e^{0x} f(x) dx \\ &= \int_{x \in X} x f(x) dx \\ &= E(X) \end{aligned}$$

MGF - Uses and Properties

- The MGF can be used to find moments of a RV
- MGFs are unique, i.e. if X and Y have the same MFG, then X and Y have the same distributions
- We can classify the distribution of a RV by matching it to a known MGF
- Sps a RV X has MGF $M_X(t)$. Then $Y = aX + b$ has MGF $M_Y(t) = e^{tb} M_X(at)$, where Y is a linear transformation of X with $a, b \in \mathbb{R}$
- If X and Y are two independent RVs, then $M_{X+Y}(t) = M_X(t) M_Y(t)$

Lecture 9A - Transformations

Continuous Transformations - Distribution Method

Let $Y = g(X)$ be a function of a RV X.

1. Find the corresponding support of Y
2. Begin by deriving the CDF of Y by relating it back to X

$$\begin{aligned} F_Y(y) &= P(Y \leq y) && \text{Definition of CDF} \\ &= P(g(X) \leq y) && Y \text{ is the RV, } y \text{ is a specific value of } Y \\ &= P(X \leq g^{-1}(y)) && \text{Express in terms of known RV } X \\ &= \int_{-\infty}^{g^{-1}(y)} f_X(x)dx && g^{-1}(y) \text{ is some value of RV } X \end{aligned}$$

3. The PDF of Y is given by $f_Y(y) = \frac{dF_Y(y)}{dy}$

Lecture 10A - Marginal PDFs (Continuous)

Marginal Distributions

The **marginal distribution functions** can be extracted from the joint distribution, which returns the probability mass/density of one variable only:

$$f_X(x) = P(X = x) = \sum_{y \in Y} f(x, y) \quad \text{or} \quad \int_Y f(x, y)dy$$

$$f_Y(y) = P(Y = y) = \sum_{x \in X} f(x, y) \quad \text{or} \quad \int_X f(x, y)dx$$

If $\forall(x, y), f(x, y) = f_X(x) \cdot f_Y(y)$, then X and Y are **independent**. Otherwise, X and Y are dependent.

Joint Distribution Function/CDF

The **joint distribution function** of two RVs X, Y defined as

$$F(a, b) = P(X \leq a, Y \leq b)$$

When X, Y are discrete:

$$F(a, b) = \sum_{x=-\infty}^{a} \sum_{y=-\infty}^{b} f(x, y)$$

When X, Y are cts:

$$F(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)dydx$$

Properties:

1. $\lim_{x \to -\infty} \lim_{y \to -\infty} F(x, y) = 0$
2. $\lim_{x \to \infty} \lim_{y \to \infty} F(x, y) = 1$
3. $F(x, y)$ is non-decreasing
4. The distribution function, for a fixed x or y, is right-cts for the remaining variable, e.g. for fixed x, $\lim_{h \to 0^+} F(x, y + h) = F(x, y)$

Lecture 10B - Covariance and Correlation

Expected Value

If X, Y are two joint RVs with PMF/PDF $f(x, y)$, then the **expected value of** XY is given by:

$$E(XY) = \sum_{x \in X} \sum_{y \in Y} xyf(x, y) \qquad \text{if } X, Y \text{ are discrete}$$

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dydx \qquad \text{if } X, Y \text{ are cts}$$

- $E(XY) \neq E(X)E(Y)$ except when X, Y are independent

For any real-valued function $g(X, Y)$, we can find its expected value using the joint PMF/PDF:

$$E(g(X, Y)) = \sum_{x \in X} \sum_{y \in Y} g(x, y)f(x, y)$$

$$E(g(X, Y)) = \int_{x \in X} \int_{y \in Y} g(x, y)f(x, y)dydx$$

Covariance

The **covariance** is a measure that all0ws us to assess the *association* between X and Y. If X, Y are two RVs with a joint probability mass/density function $f(x, y)$, then the covariance is given by:

$$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$

The covariance measure depends on the units and scale of X and Y. In order to get a unitless measure of *linear association* between X and Y, we have the **correlation**

$$\rho = \frac{cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

- ρ is the **correlation coefficient**
- $\rho \in [-1, 1]$
- $\rho = \pm 1$ indicates a perfect positive/negative linear association

Properties of Covariance

Let $a, b, c, d \in \mathbb{R}$ and X, Y, Z be cts RVs:

- $Cov(X, X) = var(X) = \sigma_X^2$
- $Cov(aX + b, cY + d) = acCov(X, Y) = ac\sigma_{XY}$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(aX, bY + cZ) = abCov(X, Y) + acCov(X, Z)$
- If X, Y are independent, then $Cov(X, Y) = 0$
  - The converse is only true when X, Y are normally distributed

Properties of Expected Values and Variance

Let X, Y be RVs:

- $E(aX + bY + c) = aE(X) + bE(Y) + c$
- $E(XY) = E(X)E(Y)$ only if X, Y are independent
- $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$
  - And so $V(X + Y) = V(X) + V(Y)$ only if X, Y are independent

Lecture 11A - Central Limit Theorem

The Sample Mean

Defined as

$$\overline{X}_n = \frac{\sum\limits_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The *sampling distribution* of the sample mean when the sample size is large enough is approximately norma, as long as the X_is are *independent and identically distributed* (i.i.d)

- $E(\overline{X}_n) = E(X)$
- $V(\overline{X}_n) = \frac{V(X)}{n}$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. RVs with $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$. Then

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to a standard normal RV (i.e. as the sample size grow)

- The denominator is $\frac{\sigma}{\sqrt{n}} = \sqrt{V(X)/n} = \sqrt{V(\overline{X}_n)}$, which is the sd of $\overline{X}_n$

Equivalently:
Let $X_1, X_2, \ldots, X_n$ be i.i.d RVs with $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$. Then for large enough n,

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Guidelines for 'Large Enough Sample'

- RVs with symmetric distributions require smaller sample sizes before convergence in distribution occurs for the sample mean (sometimes as low as $n = 5, n = 10$)
- The greater the skew/asymmetry iof the distribution of the RV, the larger the sample size we will need before the sample mean 'stabilizes' and has a distribution that can be approximated by a normal distribution (i.e. $n = 25, n = 30$)
- The above rules only apply when considering *quantitative* RVs