

STAB52
Summary Of

Probability and Statistics

The Science of Uncertainty

Second Edition

By: Michael J. Evans and Jerrey S. Rosenthal
University of Toronto



Computer & Mathematical Sciences
UNIVERSITY OF TORONTO
S C A R B O R O U G H

Instructor: Dr. Daniel Roy
Email: droy@utsc.utoronto.ca
Office: IC462
Office Hours: TU 11:00 - 12:00

1 Probability Basics

1.1 Probability Models

Sample space, often written S . This is any set that lists all possible outcomes of some unknown experiment. Collections of events are subsets of S , to which probabilities can be assigned.

Finally, a probability model requires a probability measure, usually written P . This must assign to each event A , a probability $P(A)$ with the following properties:

1. $P(A)$ is always a non-negative real number, between 0 and 1 inclusive.
2. $P(\emptyset) = 0$
3. $P(S) = 1$
4. P is countably additive, where for disjoint events A_1, A_2, A_3, \dots
we have $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$

1.2 Venn Diagrams and Subsets

The complement of a set A , denoted set $A^c = \{s | s \notin A\}$

The intersection of two sets A, B , denoted $A \cap B = \{s | s \in A \wedge s \in B\}$

The union of two sets A, B , denoted $A \cup B = \{s | s \in A \vee s \in B\}$

We also have properties $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

1.3 Properties of Probability Models

For any event A , A and A^c are always disjoint.

Furthermore, their union is always the entire sample space: $A \cup A^c = S$

And since we have $P(S) = 1$. $P(A^c) = 1 - P(A)$

Suppose that A_1, A_2, \dots are disjoint events that form a partition of the sample space i.e., $A_1 \cup A_2 \cup \dots = S$.

For any event B , $P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \dots$

Principle of inclusion-exclusion, Let A, B be two events. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

1.4 Uniform Probability on Finite Spaces

If a sample space S is finite, then one possible probability measure on S is the uniform probability measure, which assigns probability $\frac{1}{|S|}$ to each outcome. By additivity, we see that for any event A , $P(A) = \frac{|A|}{|S|}$

1. Multiplication Principle

With m in A and n elements in B , there are $m \times n$ total possible ordered pairs of elements from both sets, $C = \{(a_i, b_j) | a_i \in A, b_j \in B\}$, $|C| = m \times n$

2. Permutation Principle

Ordered arrangement of k objects, chosen without replacement from n possible objects.

The number of these ordered arrangements is $P_k^n = \frac{n!}{(n-k)!}$

3. Combination Principle

Unordered arrangement of k objects, chosen without replacement from n possible object.

The number of these unordered arrangement is $C_k^n = \binom{n}{k} = \frac{P_k^n}{k!} = \frac{n!}{k!(n-k)!}$

1.5 Conditional Probability and Independence

Given two events A, B with $P(B) > 0$, the conditional probability of A given B written $P(A|B)$ denotes the fraction of time that A occurs once we know that B has occurred.

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

Then the law of total probability can be rewritten: Let A_1, A_2, \dots be events that form a partition of the sample space S , each of positive probability.

Then for any event B , $P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \dots$

Let A, B be two events, each of positive probability. Then $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

Two events A, B are independent if $P(A \cap B) = P(A)P(B)$

Three events A, B, C are independent if **all** of the following equations hold:

1. $P(A \cap B) = P(A)P(B)$
2. $P(A \cap C) = P(A)P(C)$
3. $P(B \cap C) = P(B)P(C)$
4. $P(A \cap B \cap C) = P(A)P(B)P(C)$

If only 1 - 3 hold, then the set is called **pairwise independent**.

2 Random Variables and Distributions

2.1 Random Variables

A random variable is a function from the sample space S to \mathbb{R} .

Constant Random Variables

let c be any constant and also also a function, by saying $c(s) = c, \forall s \in S$. Thus, 5 is a random variable, as is 3 or -21.6.

Indicator Functions

If A is any event, then we can define the indicator function I_A to be the random variable such that:

$$I_A(s) = \begin{cases} 1 & s \in A \\ 0 & s \notin A \end{cases}$$

which is equal to 1 on A and is equal to 0 on A^C .

2.2 Distribution of Random Variables

Since random variables are defined to be functions of the outcome s , and because the outcome s is assumed to be random, it follows that the value of a random variable will itself be random.

However, if X is a random variable, then the probability that X will equal to some particular value x is precisely when the outcome of s is chosen such that $X(s) = x$.

If X is a random variable, then the distribution of X is the collection of probabilities $P(X \in B)$ for all subsets B of the real numbers.

2.3 Discrete Distributions

For many random variables X , if we have $P(X = x) > 0$ for certain x values. This means there is positive probability that the variable will be equal to certain particular values.

If $\sum_{x \in \mathbb{R}} P(X = x) = 1$, which says all of the probability assigned with the random variable X sums to 1, this random variable X is discrete.

We can formalize this as: A random variable X is discrete if there is a finite or countable sequence x_1, x_2, \dots of distinct real numbers, and a corresponding sequence p_1, p_2, \dots of non-negative real numbers, such that $P(X = x_i) = p_i$ for all i , and $\sum_i p_i = 1$.

For a discrete random variable X , its probability function is the function $P_X : \mathbb{R} \rightarrow [0, 1]$ defined by $P_X(y) = P(X = y)$

Distributions

Bernoulli

Consider flipping a coin that has probability θ of coming up heads, and probability of $1 - \theta$ of coming up tails, where $\theta \in [0, 1]$.

Let $X = 1$ if the coin is heads, and $X = 0$ otherwise. then $P_X(1) = P(X = 1) = \theta$ and $P_X(0) = P(X = 0) = 1 - \theta$. The random variable X is said to have the Bernoulli(θ) distribution; we write this as $X \sim \text{Bernoulli}(\theta)$.

Binomial

Consider flipping n coins, each of which has independent probability of θ of coming up heads, and probability $1 - \theta$ of coming up tails, where $\theta \in [0, 1]$.

Let X be the total number of heads showing, then for each $y = 1, 2, 3, \dots, n$,

$$P_X(y) = P(X = y) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} = \frac{n!}{(n-y)!y!} \theta^y (1 - \theta)^{n-y}$$

The random variable X is said to have the Binomial(n, θ) distribution; we write this as $X \sim \text{Binomial}(n, \theta)$. The Bernoulli(θ) distribution corresponds to the special case of the Binomial(n, θ) distribution where $n = 1$.

Geometric

Consider repeatedly flipping a coin that has probability θ of coming up heads and probability $1 - \theta$ of coming up tails, where again $0 < \theta < 1$. Let X be the number of tails that appear before the first head.

Then for $k \geq 0$, $X = k$ if and only if the coin shows exactly k tails followed by a head. The probability of this is equal to $(1 - \theta)^k \theta$

Negative-Binomial Distribution

Consider repeatedly flipping a coin that has probability θ of coming up heads and probability $1 - \theta$ of coming up tails. Let r be a positive integer, and let Y be the number of tails that appear before the r -th head.

Then for $k \geq 0$, $Y = k$ if and only if the coin shows exactly $r - 1$ heads and k tails on the first $r + k - 1$ flips, then shows a head on the $r + k$ -th flip. The probability of this is equal to:

$$P_Y(k) = \binom{r+k-1}{r-1} \theta^{r-1} (1-\theta)^k \theta = \binom{r+k-1}{r-1} \theta^r (1-\theta)^k$$

The random variable Y is said to have the Negative-Binomial(r, θ) distribution; we write this as $Y \sim \text{Negative-Binomial}(r, \theta)$. Of course, the special case $r = 1$ is the Geometric(θ) distribution.

The Poisson Distribution

We say that if a random variable Y has the poisson distribution, and write $Y \sim \text{Poisson}(\lambda)$, if

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

For $y = 0, 1, 2, 3, \dots$. We note that since $\sum_{y=0}^{\infty} \lambda^y / y! = e^y$ (From Calculus), then $\sum_{y=0}^{\infty} P(Y = y) = 1$

We motivate the Poisson distribution as follows. Suppose $X \sim \text{Binomial}(n, \theta)$, then for $0 \leq x \leq n$
 $P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$,

The Hyper-geometric Distribution

Suppose an urn contains N total balls, M white balls and $N - M$ black balls. If we were to select a total n balls from the urn. What is the probability that x of those n balls are white?

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

2.4 Continuous Distributions

A random variable X is continuous if $P(X = x) = 0, \forall x \in \mathbb{R}$, this is to say that X takes on uncountably infinite many values.

PDF of Continuous Random Variables

For a continuous RV X its Probability Density Function (PDF) is a function $f_X(\cdot)$ such that

$$P(X \in [a, b]) = \int_a^b f_X(x) dx$$

Properties of PDFs

1. $0 \leq f_X(x), \forall x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x) dx = 1 \equiv P(X \in (-\infty, \infty)) = P(S) = 1$
3. $F_X(x) = \int_{-\infty}^x f_X(u) du \Rightarrow f_X(x) = F'_X(x)$ Where F is the CDF.

Cumulative Distribution Function

1. $\lim_{x \rightarrow -\infty} F_X(x) = 0$
2. $\lim_{x \rightarrow \infty} F_X(x) = 1$
3. $\forall x_1 < x_2 \in \mathbb{R}, F_X(x_1) \leq F_X(x_2)$

There are two main kinds of continuous distributions:

Uniform

Uniform RV X takes values in an interval $[l, u]$ where $l < u \in \mathbb{R}$. So the probability of any sub-interval (a, b) is proportional to its length.

$$P(X \in (a, b)) = \frac{b-a}{u-l}, \forall l \leq a \leq b \leq u$$

We denote this $X \sim \text{Uniform}(l, u)$

$$\text{Then } f_X(x) = \begin{cases} \frac{1}{u-l}, & l \leq x \leq u \\ 0, & \text{otherwise} \end{cases} \text{ and } F(x) = \begin{cases} 0, & x < l \\ \frac{x-l}{u-l}, & l \leq x \leq u \\ 1, & x > u \end{cases}$$

Exponential Distribution

An Exponential RV X takes positive values according to PDF $f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$ for some $\lambda > 0$

It then follows that $F_X(x) = 1 - e^{-\lambda x}$

We denote this as $X \sim \text{Exponential}(\lambda)$

Poisson Distribution

A Poisson RV X counts the number of successes in some continuous interval. where the PMF is $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \dots$

where $\lambda > 0$ denotes the average number of successes over the interval.

A Poisson(λ) is approximated by Binomial($n, \lambda/n$) as $n \rightarrow \infty$

Gamma Distribution

A job consisting of a tasks, each completed in a sequence according to independent Exponential(λ) times.

$$\text{The PDF of a Gamma distribution is } f(x) = \begin{cases} \frac{1}{\Gamma(a)} \lambda^a x^{a-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$\text{Where } \Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

To make computation easier, we have the following properties

- For $a > 1, \Gamma(a) = (a-1)\Gamma(a-1)$
- For $a \in \mathbb{Z}, \Gamma(a) = (a-1)!$
- For $a = 1, X \sim \text{Gamma}(a, \lambda) \sim \text{Exponential}(\lambda)$

Normal Distribution

Normal (Gaussian) Distribution is the typical way to describe how a continuous RV X is distributed around its center with some spread.

Typically, the average of RVs will converge to Normal.

$$\text{The PDF of a Normal distribution is } f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$$

Where $\mu \in \mathbb{R}$ is the **mean**, representing center. $\sigma > 0$ is called the standard deviation, representing spread. We say this $X \sim \text{Normal}(\mu, \sigma^2)$

2.5 Expectation

Expected Value

The Expected Value of a random variable X is what we expect to observe on average over a large number of repetitions of the experiment.

Consider a random variable X with PMF $P(X = x) = P_X(x)$ and assume you draw value n times.

It's expected value denoted $E(X)$ is defined as: $E(X) = \sum_{x_i} x P(X = x) = \sum_{\text{all } x} x P_X(x)$

Consider an indicator RV $I_A(s) = \begin{cases} 1, & s \in A \\ 0, & s \notin A \end{cases}$

Then $E(I_A) = P(A)$

Expected Value - Continuous RVs

For a continuous RV X with PDF $f_X(x)$, expected value is defined as: $E(X) = \int_{-\infty}^{\infty} xf_X(x)dx$

For functions of RVs, $E(Y) = E(h(X)) = \begin{cases} \sum h(x)P_X(x) & \text{discrete } X \\ \int_{-\infty}^{\infty} h(x)f_X(x)dx & \text{continuous } X \end{cases}$

Properties of Expected Values

Linearity of expectations: for any linear function $h(X) = a + bX$

Where $E(a + bX) = a + bE(X)$

More generally: $E(g(X) + h(X)) = E(g(X)) + E(h(X))$

Expectation measures the center of a Random Variable. But sometimes we are interested in the spread.

Variance is a measure of spread defined as: $V(X) = E(X^2) - E(X)^2$

Then, the standard deviation is the square root of the variance.

3 Change of RV - 1D

Assume Random Variable X follows a certain distribution and Random Variable $Y = h(X)$

We can find the distribution of Y based on that of X with 3 methods:

3.1 General Method

Say we know the distribution of X and $P(X \in B), \forall B \subseteq \mathbb{R}$, and $Y = h(X)$

Then Probability $P(Y \in A) = P(X \in h^{-1}[A])$

where $h^{-1}[A] = \{x \in \mathbb{R} | h(x) \in A\}$ is the **inverse image** of A . This method is incredibly useful for discrete RVs.

3.2 CDF Method

Suppose we are given a continuous random variable X with pdf $f_X(x)$, and we know $Y = h(X)$ where h is an invertible function.

Recall that the cdf of a random variable R is given by the probability that $P(R \leq r)$. We will also need to use the fact that the **PDF** is the derivative of the **CDF**.

We start with the cdf of Y , where $F_Y(y) = P(Y \leq y) = P(h(X) \leq y) = P(X \leq h^{-1}(y)) = F_X(h^{-1}(y))$

Finally we derive $F_Y(y) = F_X(h^{-1}(y))$

Furthermore, the **PDF** of Y : $F_Y(y) = \frac{d}{dy}F_X(h^{-1}(y)) = f_X(h^{-1}(y)) \left\| \frac{d}{dy}h^{-1}(y) \right\|$

3.3 PDF Method

Assume again that $Y = h(X)$ for a continuous **one-to-one** function h , where **PDF** of X exists, but there is no closed form **CDF**.

Then the **PDF** of Y is given by $f_Y(y) = \frac{f_X(h^{-1}(y))}{|h'(h^{-1}(y))|}$

4 Discrete 2D Distributions

Suppose we have multiple Random Variables defined in a random experiment.

E.g. roll two 6-sided dice and let: X be the value of the 1st die, Y be the value of the 2nd die.

Then each event in S maps to coordinates in a 2D space.

For any two Random Variables X, Y , their joint (bivariate) distribution is the collection of all the probabilities of the form

$$P((X, Y) \in B) = P(\{s \in S | (X(s), Y(s)) \in B\}), \forall B \subseteq \mathbb{R}^2$$

Furthermore, the joint **PMF** is defined as: $P_{X,Y}(x, y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\})$

Also, $P_{X,Y}(x, y) \geq 0$ and $\sum_{x,y} P_{X,Y}(x, y) = 1$

5 Continuous 2D Distributions

For arbitrary Random Variables X, Y , joint CDF: $F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P(\{X \leq x\} \cap \{Y \leq y\})$

We can find marginal CDFs: $F_X(x) = F_{X,Y}(x, \infty) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y)$

For continuous Random Variables X, Y , the **joint PDF** is a function $f_{x,y}(x, y)$ such that:

$P((X, Y) \in R) = \int \int_R f_{X,Y}(x, y) dx dy$, Where the probability is the volume contained under function $f_{X,Y}$ over region R .

Properties of joint PDFs

- $f_{X,Y}(x, y) \geq 0, \forall x, y \in \mathbb{R}$
- $\int \int_{\mathbb{R}} f_{X,Y}(x, y) dx dy = 1$

Relationship between joint PDF and joint CDF

- $F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(s, t) dt ds, \forall x, y \in \mathbb{R}$
- $f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$

6 Independent Distributions

Two Random Variables X, Y are independent denoted $(X \perp Y)$

If $\forall A, B \subseteq \mathbb{R}, P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

Independent CDFs,

Independence has implications for joint CDF & PMF/PDF

For independent Random Variables X, Y , joint CDF factorizes as:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

Joint PMF factorizes as:

$$P_{X,Y}(x, y) = P(X = x, Y = y) = P(X = x)P(Y = y) = P_X(x)P_Y(y)$$

For absolutely continuous independent Random Variables X, Y ,

joint PDF factorizes as:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$$

If joint PMF/PDF factorizes over all values in \mathbb{R}^2 , then discrete/continuous Random Variables X, Y are independent.

6.1 Conditional Distributions

Consider Random Variables X, Y with some joint distribution, **conditional distributions** describe probabilities of RVs given some condition on their values.

$$P((X, Y) \in A | (X, Y) \in B) = \frac{P(\{(X, Y) \in A\} \cap \{(X, Y) \in B\})}{P((X, Y) \in B)}$$

For discrete Random Variables X, Y , conditional probabilities can be found as:

$$P((X, Y) \in A | (X, Y) \in B) = \sum_{(X, Y) \in A} P(X = x, Y = y | (X, Y) \in B) = \sum_{(X, Y) \in A} \frac{P((X = x, Y = y), (X, Y) \in B)}{P((X, Y) \in B)}$$

Conditional PMF:

Most often, we condition on a specific value of one RV and look at the probability of the other.

Conditional PMF of X given $Y = y$

$$P_{X|Y}(x|y) = P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P_{X,Y}(x, y)}{P_Y(y)}$$

Note that a conditional PMF is still a proper PMF. and all properties hold.

if $X \perp Y$, then the conditional PMF = the marginal PMF

$$\text{i.e., } P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)} = \frac{P_X(x)P_Y(y)}{P_Y(y)} = P_X(x)$$

Conditional CDF of $X|Y = y$

$$F_{X|Y}(x, y) = P(X \leq x | Y = y) = \sum_{i \leq x} P_{X|Y}(i|y)$$

Conditional PDF

let continuous Random Variables X, Y have joint PDF $f_{X,Y}(x, y)$

$$f_{X|Y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

this effectively cuts a slice of $f_{X,Y}(x, y)$ at $Y = y$ and scales it by a $1/f_Y(y)$ so that it integrates to 1.

$$f_{X|Y}(x|y) \geq 0 \quad \& \quad \int_{\mathbb{R}} f_{X|Y}(x|y) dx = 1, \forall y$$

$$\text{If } X \perp Y, \text{ then conditional = marginal PDF, } f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

Conditional PDF can also be used to define joint PDFs: $f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x)$

Conditional CDF

Conditional PDF of $X|Y = y$ can be integrated to find conditional probabilities: $P(X \in A | Y = y) =$

$$\int_A f_{X|Y}(x|y) dx$$

7 2D Change of Variables

Consider a function of Random Variables X, Y , defining transformed Random Variables Z, W

For example: $h\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = Z$ or $h\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} h_1(X, Y) \\ h_2(X, Y) \end{bmatrix} = \begin{bmatrix} Z \\ W \end{bmatrix}$

If we want to find the joint/marginal distribution of transformed Random Variables Z, W based on joint distribution of original Random Variables X, Y we can use the following three methods:

7.1 PMF Method

Supposed we have discrete Random Variables X, Y with joint PMF $P_{X,Y}(x, y)$, and we wish to calculate joint PMF of $Z = h_1(X, Y), W = h_2(X, Y)$

Then we can write $P_{Z,W}(z, w) = P(Z = z, W = w) = P(h_1(X, Y) = z, h_2(X, Y) = w)$

Let $h^{-1}(z, w) = \{(x, y) | h_1(x, y) = z, h_2(x, y) = w\}$

This is to say that $h^{-1}(z, w)$ is the set of all (x, y) such that $h_1(x, y) = z$ and $h_2(x, y) = w$

Then $P_{Z,W}(z, w) = \sum_{(x,y) \in h^{-1}(z,w)} P_{X,Y}(x, y)$

7.2 CDF Method

In certain cases, we can calculate the CDF of transformed Random Variables Z and W in terms of original Random Variables X, Y joint CDF.

$F_{Z,W}(z, w) = P(Z \leq z, W \leq w) = P(h_1(X, Y) \leq z, h_2(X, Y) \leq w)$

Then if we let $h^{-1}((-\infty, z] \times (-\infty, w]) = \{(x, y) | h_1(X, Y) \leq z, h_2(X, Y) \leq w\}$

$P_{Z,W}(z, w) = P((X, Y) \in h^{-1}((-\infty, z] \times (-\infty, w]))$

7.3 PDF Method

For continuous Random Variables X, Y CDF method involves integrals, but often times we cannot solve integral in closed form.

For X, Y with joint PDF $f_{X,Y}(x, y)$ and differentiable 1 to 1 transformation random variables $(Z, W) = h(X, Y)$

Then if the joint PDF of X, Y is known, we can find the joint pdf of some $Z, W = h(X, Y)$ for a **differentiable** and 1 to 1 function $h: \mathbb{R}^2 \rightarrow \mathbb{R}^2$

The joint pdf is given by $f_{Z,W}(z, w) = \frac{f_{X,Y}(h^{-1}(z, w))}{|J(h^{-1}(z, w))|}$

8 Sum & Order Statistics

Two specific transformation of Random Variables X_1, X_2, X_3, \dots are of practical importance for Statistics & CS

- Sum of Random Variables: $X_1 + X_2 + \dots$
- Order Statistics: min/max, or 2nd, 3rd, \dots , etc largest value of some X_1, X_2, X_3, \dots

8.1 Sum of Random Variables

Consider RVs X, Y with joint PDF $f_{X,Y}(x, y)$, and let $Z = X + Y$. Then we can find the distribution of Z via the CDF method:

$$P(Z \leq z) = P(X + Y \leq z) = P(Y \leq z - X) = \int \int_R f_{X,Y}(x, y) dx dy$$

Where R describes the area where $y \leq z - x$ on the cartesian plane.

But we can also consider the **Convolution Method**

For random Variables X, Y with joint PDF $f_{X,Y}(x, y)$ and $Z = X + Y$

$$\text{The PDF of } Z \text{ is given by: } f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y) dy$$

Similarly for discrete random variables: the **PMF** of such Z can be found by

$$P_Z(z) = \sum_x P_{X,Y}(x, z-x) = \sum_x P_{X,Y}(z-y, y)$$

8.2 Order Statistics

For random variables X_1, \dots, X_n

The k -th order statistic ($X_{(k)}$) is the k -th smallest variable such that:

- $X_{(1)} = \min(X_1, \dots, X_n)$
- $X_{(n)} = \max(X_1, \dots, X_n)$
- $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n-1)} \leq X_{(n)}$

Marginal distribution of any order statistics can easily be derived when random variables are identically and independently distributed and their CDF $F_X(x)$ is known

The easiest order statistic distributions to derive are for $X_{(1)}$ i.e. **minimum**, and $X_{(n)}$ i.e. **maximum**.

For identical and independently distributed random variables X_1, \dots, X_n , with CDF $F_X(x)$ and PDF $f(x)$, the distribution of the maximum $X_{(n)}$ is:

$$F_{(n)} = [F(x)]^n \text{ \& } f_{(n)}(x) = n[F(x)]^{n-1}f(x)$$

Then the distribution of the minimum $X_{(1)}$ is:

$$F_{(1)} = 1 - [1 - F(x)]^n \text{ \& } f_{(1)}(x) = n[1 - F(x)]^{n-1}f(x)$$

9 Moments

Let $Z = g(X, Y)$ be a function of random variables X, Y

The expected value of $g(X, Y)$ can be found either by using the distribution of Z or using the joint distribution of X, Y

- Discrete Case: $E[g(X, Y)] = \sum_x \sum_y g(x, y) P_{X,Y}(x, y)$
- Continuous Case: $E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$

Note that linearity of Expectations still hold, further more, for independent random variables X, Y :

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

The moments of a Random Variable X are the expected values of different **powers** of X , or functions thereof

- The **r**-th moment of X is defined as $E[X^r]$, Where the 1-st moment is the mean, $E[X] = \mu$
- The **r**-th central moment of X is defined as $E[(X - \mu)^r]$, In particular the second central moment is the variance.

Consider Random Variables X, Y with means & variances $\begin{cases} \mu_X = E[X], \mu_Y = E[Y] \\ \sigma_X^2 = V(X), \sigma_Y^2 = V(Y) \end{cases}$

- The **covariance** of X and Y : $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
- The **correlation** of X and Y : $\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}, \rho_{X,Y} \in [-1, 1]$

Covariance Properties

- $Cov(X, X) = V(X)$
- $Cov(X, Y) = E(XY) - \mu_X \mu_Y$
- $X \perp Y \Rightarrow Cov(X, Y) = 0$

For Random Variables X_1, \dots, X_n & Y_1, \dots, Y_m and constants a_1, \dots, a_n & b_1, \dots, b_m

We define linear functions $Z = \sum_{i=1}^n a_i X_i$ and $W = \sum_{j=1}^m b_j Y_j$

- $E(Z) = \sum_{i=1}^n a_i E[X_i]$
- $V(Z) = \sum_{i=1}^n (a_i)^2 V(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j Cov(X_i, X_j)$

If X_1, \dots, X_n are independent, then $V(Z) = \sum_{i=1}^n (a_i)^2 V(X_i)$

- $Cov(Z, W) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \cdot Cov(X_i, Y_j)$

9.1 Moment Generating Function

$$m(t) = E(e^{tX})$$

In particular, MGF allows calculation of all moments of X .

$$E(X^k) = m^{(k)}(0) = \frac{d^k}{dt^k} m(0)$$

$$\text{Recall that } e^X = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Proof that $m^{(k)}(0) = E(X^k)$

$$E(e^{tX}) = E\left(\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E(X^k) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E(X^k) = 1 + tE(X) + \frac{t^2}{2} E(X^2) + \dots$$

9.2 MGF Method

MGF uniquely characterizes the distribution of a Random Variable.

For Random Variables X, Y with MGFs $m_X(t), m_Y(t)$, $m_X(t) = m_Y(t) \Leftrightarrow X \sim Y$

MGF can also be used to find distribution of functions of Random Variables:

Let $Y = g(X_1, \dots, X_n)$ with $m_Y(t) = E[e^{tY}] = E[e^{t \cdot g(X_1, \dots, X_n)}]$

Then if $m_Y(t)$ is a moment generating function of some known distribution then Y follows that distribution.

Let $Y = a_1 X_1 + \dots + a_n X_n$, Where X_1, \dots, X_n are independent with MGFs X_{X_i}

$$\text{Then } m_Y(t) = E[e^{tY}] = E[e^{t(\sum_{i=1}^n a_i X_i)}] = E[\prod_{i=1}^n e^{ta_i X_i}]$$

$$\text{Then by independence, we have: } \prod_{i=1}^n E[e^{ta_i X_i}] = \prod_{i=1}^n m_{X_i}(a_i t)$$

In particular, for i.i.d. X_1, \dots, X_n and $Y = X_1 + \dots + X_n$, $m_Y(t) = (m_X(t))^n$

10 Conditional Expectation

Consider a random variable X and a related event A . Such that conditioning on A changes the distribution of X and it's expected value.

The **Conditional Expectation** of any function $g(X)$ given A is:

- Discrete Case: $E[g(X)|A] = \sum_x g(x)p(x|A)$
- Continuous Case: $E[g(X)|A] = \int_{-\infty}^{\infty} g(x)f(x|A)dx$

More often, we are interested in the expected value of Y conditional on some value of another random variable X

The **Conditional Expectation** of any function $g(Y)$ given $X = x$ is:

- Discrete Case: $E[g(Y)|X = x] = \sum_y g(y)p_{Y|X}(y|x)$
- Continuous Case: $E[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y)f_{Y|X}(y|x)dy$

If $X \perp Y \Rightarrow E[Y|X = x] = E[Y]$

For any value $X = x$, conditional expectation $E[g(Y)|X = x]$ returns another value, we can think of this as a function of x , $h(x) = E[g(Y)|X = x]$

Furthermore, if the value of X is not specified, we can define the conditional expectation of $g(Y)|X$ as a function of Random Variable X

$$E[g(Y)|X] = h(X)$$

10.1 Laws of Total Expectation

This is known from the Law of Total Probability.

$$\text{i.e. } P(Y \in A) = \begin{cases} \sum P(Y \in A|X = x)p_X(x) \\ \int_{\mathbb{R}} P(Y \in A|X = x)f_X(x)dx \end{cases}$$

Similarly for expectations, the Law of Total Expectation holds:

$$E[g(Y)] = E[E[g(Y)|X]] = \begin{cases} \sum E[g(Y)|X = x]p_X(x) \\ \int_{\mathbb{R}} E[g(Y)|X = x]f_X(x)dx \end{cases}$$

Proof of Discrete Case

$$E[E[X|Y]] = E\left[\sum_x x \cdot P(X = x|Y)\right] = \sum_y \left[\sum_x x \cdot P(X = x|Y)\right]P(Y = y)$$

$$\text{Then we have } \sum_y \sum_x x \cdot P(X = x, Y = y)$$

switching around the summation, we get

$$\sum_x x \sum_y P(X = x, Y = y) = \sum_x xP(X = x) = E(X)$$

Proof of Continuous Case

$$E[E[X|Y]] = E\left[\int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|Y)dx\right] = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)dx\right) f_Y(y)dy$$

$$\text{But since we know that } f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \text{ we have } \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \cdot f_{X,Y}(x,y)dx\right) dy$$

Switching orders of integration,

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \cdot f_{X,Y}(x,y)dy\right) dx = \int_{-\infty}^{\infty} x \cdot \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right) dx = \int_{-\infty}^{\infty} x f_X(x)dx = E[X]$$

10.2 Laws of Total Variance

Define conditional variance in terms of conditional expectation, $V(Y|X) = E[Y^2|X] - [E(Y|X)]^2$

The law of total variance $v(Y) = E[V(Y|X)] + V[E(Y|X)]$

11 Inequalities

Recall these topics:

- **Moments:** $E[X], E[X^2], \dots, E[X^k]$
- **Moment Generating Function:** $m_X(t) = E[e^{tX}], E[X^k] = m_X^{(k)}(t)|_{t=0}$

Expectations are related to underlying distributions, we look at 4 inequalities related to expectations:

- Markov Inequality
- Chebyshev Inequality
- Cauchy-Schwarz Inequality
- Jensen Inequality

Markov and **Chebyshev** is used for probabilities.

Cauchy-Schwarz and **Jensen** is used for expected value of functions of random variables.

11.1 Markov Inequality

For a positive random variable X , probability of right tail is bound by mean:

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

Proof of Continuous Case

$$\begin{aligned} E[X] &= \int_0^\infty x f_X(x) dx = \int_0^a x f_X(x) dx + \int_a^\infty x f_X(x) dx \\ &\geq \int_a^\infty x f_X(x) dx \geq \int_a^\infty a f_X(x) dx = a \int_a^\infty f_X(x) dx = a P(X \geq a) \end{aligned}$$

$$\text{Thus, } E[X] \geq a P(X \geq a) \Leftrightarrow P(X \geq a) \leq \frac{E[X]}{a}$$

Chernoff Bound

consider any Random Variable X with MGF $m_X(t)$

Show that $P(X \geq a) \leq \frac{m_X(t)}{e^{ta}}$. Let $g(x) = e^{tx}$

Then applying Markov's Inequality, $P(e^{tX} \geq e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}} \Rightarrow P(X \geq a) \leq \frac{m_X(t)}{e^{ta}}$

11.2 Chebyshev Inequality

For any Random Variable X , the probability of both tails is bounded by the variance.

$$P(|X - E[X]| \geq a) \leq \frac{V(X)}{a^2}$$

Proof: Apply Markov's inequality to $g(X) = (X - \mu)^2 \geq 0$

$$\text{Then } P(g(x) \geq a^2) \leq \frac{E[g(x)]}{a^2} \Rightarrow P((X - \mu)^2 \geq a^2) \leq \frac{E[(X - \mu)^2]}{a^2} \Rightarrow P(|X - \mu| \geq a) \leq \frac{V(X)}{a^2}$$

11.3 Cauchy-Schwarz Inequality

For any two random variables, The expectation of their product is bound by the geometric average of their 2-nd moments

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]} \text{ This provides range of covariance. } |Cov(XY)| \leq \sqrt{V(X)V(Y)}$$

11.4 Jensen Inequality

In general, we know that $E[g(X)] \neq g(E[X])$

However, for any random variable X and a **convex** function g , we have $E[g(X)] \geq g(E[X])$

12 Law of Large Number & Central Limit Theorem

Many probability problems involve sequences of Random Variables X_1, X_2, \dots where we are interested in limiting behaviour of X_i as $i \rightarrow \infty$

We often look at two important results for **sum/averages** of increasing numbers of Random Variables:

- Weak Law of Large Numbers
- Central Limit theorem

13 Averages of Random Variables

Consider sequence of **independent** Random Variables X_1, \dots, X_n with common mean μ and variance σ^2 , we define their average $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$

Mean $E[\bar{X}_n] = E\left[\frac{1}{n}(X_1 + \dots + X_n)\right] = \frac{1}{n}E[X_1 + \dots + X_n] = \mu$

Variance $V(\bar{X}_n) = V\left(\frac{1}{n}(X_1 + \dots + X_n)\right) = \frac{1}{n^2}(V(X_1) + V(X_2) + \dots + V(X_n)) = \frac{\sigma^2}{n}$

13.1 Weak Law of Large Numbers

Weak Law of Large Numbers (WLLN): averages of independent Random Variables with finite variance "converges" to their common mean μ .

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0, \forall \epsilon > 0$$

This can be proven with Chebychev's inequality, $P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{V(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$

Then we see that when $n \rightarrow \infty$, $\frac{\sigma^2}{n\epsilon^2} \rightarrow 0$

Weak Law of Large Numbers has important applications in the following:

- Statistics: Estimate mean $\mu = E(X)$ of an **unknown** distribution by averaging random values X_1, X_2, \dots (AKA samples)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu \text{ as we take } n \rightarrow \infty$$

- Simulation: Approximate probability of an event A by repeating the experiment and counting the average number of times event occurs.

Let I_A be the indicator that A occur,

$$\bar{P}_n = \frac{1}{n} \sum_{i=1}^n I_i(A) \rightarrow P(A) \text{ as we take } n \rightarrow \infty$$

13.2 Types of Convergence

Consider the following sequence of continuous random variables X_1, X_2, \dots and random variable Y .

- \bar{X}_n **converges in probability** to Y , as $n \rightarrow \infty$, if
$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - Y| \geq \epsilon) = 0, \forall \epsilon > 0, \text{ denoted } \bar{X}_n \rightarrow^P Y$$

- \bar{X}_n **converges in distribution** to y , as $n \rightarrow \infty$, if
$$\lim_{n \rightarrow \infty} P(\bar{X}_n \leq x) = P(Y \leq x), \forall x \in \mathbb{R} \text{ denoted } \bar{X}_n \rightarrow^D Y$$

13.3 Central Limit Theorem

Consider a sequence of **independent** Random Variables X_1, \dots, X_n with common mean μ and variance σ^2 , and define their standardized average $Z_n = \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$

Standardizing: subtracting mean and divide by standard deviation.

Standardized Random Variables always have a mean of 0 and variance of 1