Overview of Math 477

Even before reading all this, the most important thing to keep in mind is that nobody can learn probability by only studying the theory. This is intended as a guide that puts the entire material together, so you can see the bigger picture and how things fit together. But you will only gain familiarity with the topics by practicing problems from the book and other sources. Also notice that the order presented here does not always agree with the order in which everything was covered.

The good news: Students will be allowed to bring a formula sheet to the final exam, as long as it:

- is prepared by the student;
- is handwritten;
- contains only one page (just one side of the paper);
- contains only formulas found in this guide;
- gets approved by the instructor prior to the exam;
- is turned in together with the exam at the end.

The point is that, if you want a formula sheet, you are supposed to go through the trouble of writing it yourself. You cannot make copies of another student's sheet.

Contents

---

1 Combinatorics

1.1 Counting

- **Basic Principle of Counting:** If an experiment has n_1 possible outcomes, and a second independent experiment has n_2 possible outcomes, then the number of different combined outcomes is $n_1 n_2$.
- **Permutations:** The number of ways to put n different objects in order is $n!$.
- **Combinations:** The number of ways to select a group of r different objects from a set of n different objects is $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.
- **Arrangements:** The number of ways to select r different objects from a set of n different objects if the order in which the selection is made matters is $(n)_r = \frac{n!}{(n-r)!}$ (we just have to think of number of possibilities for each member of this group of r that we are choosing).
- **Multinomials:** The number of ways to split n different objects into k different groups having $n_1, n_2, \ldots, n_k$ members (where $n_1 + n_2 + \ldots + n_k = n$) is $\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$.
- **Distributing *equal* objects:** The number of ways to give n *equal* objects to r different people is $\frac{n+r-1}{r-1}$.
- **Distributing *different* objects:** The number of ways to give n *different* objects to r different people is r^n (we just need to think of "where each object can go").
- There are no easy ways to handle distribution of objects to *equal* recipients. This is the *Theory of Partitions*.

1.2 Properties of Binomial Coefficients

- **Symmetry:** $\binom{n}{r} = \binom{n}{n-r}$

  Example: $\binom{20}{7} = \binom{20}{13}$
- **Sum of two consecutive:** $\binom{n}{r} + \binom{n}{r+1} = \binom{n+1}{r+1}$

  Example: $\binom{16}{7} + \binom{16}{8} = \binom{17}{8}$
- **Sum of entire row:** $\sum_{i=0}^{n} \binom{n}{i} = 2^n$

  Example: $\binom{10}{0} + \binom{10}{1} + \binom{10}{2} + \cdots + \binom{10}{10} = 2^{10}$.
- **Sum of column up to a certain point:** $\sum_{i=r}^{n} \binom{i}{r} = \binom{n+1}{r+1}$

  Example: $\binom{4}{4} + \binom{5}{4} + \binom{6}{4} + \cdots + \binom{23}{4} = \binom{24}{5}$.
- **Binomial Theorem:** $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$ for any $x, y \in \mathbf{R}$ and $n \in \mathbf{N}$

  Example: $(2x+5)^4 = \binom{4}{0}(2x)^0 5^4 + \binom{4}{1}(2x)^1 5^3 + \binom{4}{2}(2x)^2 5^2 + \binom{4}{3}(2x)^3 5^1 + \binom{4}{4}(2x)^4 5^0$

2 Events and Probability

2.1 Operations with Events

- **Associative Laws:**
$$(E \cup F) \cup G = E \cup (F \cup G) = E \cup F \cup G$$
$$(E \cap F) \cap G = E \cap (F \cap G) = E \cap F \cap G$$
- **Distributive Laws:**
$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$$
$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$$
- **De Morgan's Laws:**
$$(E \cup F)^c = E^c \cap F^c$$
$$(E \cap F)^c = E^c \cup F^c$$

2.2 Equally Likely Outcomes

If all outcomes of the sample space are equally likely, then $P(E) = \dfrac{|E|}{|S|}$ for any event $E \subseteq S$.

2.3 Properties of Probability Functions

- **Boundedness:** $0 \leq P(E) \leq 1$ for any event E.
- **Probability of sample space:** $P(S) = 1$
- **Probability of the null event:** $P(\emptyset) = 0$
- **Probability of the complement:** $P(E^c) = 1 - P(E)$
- **Monotonicity:** $P(E) \leq P(F)$ if $E \subseteq F$.
- **Probability of disjoint union:** $P(E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + P(E_3) + \cdots + P(E_n)$ if the events E_i are **MUTUALLY DISJOINT**.
- **Independence:** Events $E_1, E_2, E_3, \ldots, E_n$ are **INDEPENDENT** if
$$P\left(\bigcap_{j=1}^{k} E_{i_j}\right) = \prod_{j=1}^{k} P(E_{i_j}) \quad , \quad \text{for all } i_1 < i_2 < \ldots i_k \text{ and } k = 1, 2, \ldots, n$$

  This means probability of intersection of any number of them is product of corresponding probabilities.

  Example with two events: E and F are independent if $P(E \cap F) = P(E)P(F)$

  Example with three events: E, F and G are independent if
$$P(E \cap F) = P(E)P(F) \quad , \quad P(E \cap G) = P(E)P(G) \quad , \quad P(F \cap G) = P(F)P(G) \quad ,$$
$$P(E \cap F \cap G) = P(E)P(F)P(G)$$

  Independence is sometimes assumed or given (then you can **use** that probabilities of intersections are products of probabilities), and sometimes asked (then you need to **show** that probabilities of intersections are products of probabilities).

- **Principle of Inclusion-Exclusion:** For any finite collection of events $E_1, E_2, \ldots, E_n$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} \sum_{i_1<i_2<\ldots<i_k} (-1)^{k+1} P\left(\bigcap_{j=1}^{k} E_{i_j}\right)$$

$$\begin{aligned}
&= \text{sum of probabilities of each } E_i && n \text{ terms}\\
&- \text{sum of probabilities of intersections of pairs} && \binom{n}{2} \text{ terms}\\
&+ \text{sum of probabilities of intersections of triples} && \binom{n}{3} \text{ terms}\\
&- \text{sum of probabilities of intersections of quadruples} && \binom{n}{4} \text{ terms}\\
&+ (\cdots)\\
&\pm \text{sum of probabilities of intersections of groups of } n-1 \text{ events} && \binom{n}{n-1} \text{ terms}\\
&\mp P(E_1 \cap E_2 \cap E_3 \cap \cdots \cap E_n) && 1 \text{ term}
\end{aligned}$$

Example with 2 events:
$P(E \cup F) = P(E) + P(F) - P(E \cap F)$
Example with 3 events:
$P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(E \cap F) - P(E \cap G) - P(F \cap G) + P(E \cap F \cap G)$

2.4 Conditional Probability

- **Definition of the conditional:** The probability of E given F is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

whenever $P(F) \neq 0$. When a problem gives you the information that event F has happened and asks you to find the probability of event E, it is asking for $P(E|F)$. The number $P(E)$ does **not** change if we know F happened or not, but the thing you are looking for is not $P(E)$, it is $P(E|F)$.

- **Properties of conditional:** All usual properties apply **to the first argument**, not the second. For example, these are true:

$$P(E^c|F) = 1 - P(E|F) \quad , \quad P(A \cup B|F) = P(A|F) + P(B|F) - P(A \cap B|F)$$

And these are **NOT TRUE**:

$$P(E|F^c) = 1 - P(E|F) \quad , \quad P(A|E \cup F) = P(A|E) + P(A|F) - P(A|E \cap F) \quad \textbf{(NO!)}$$

- **Law of Total Probability:** If events $A_1, A_2, \ldots, A_n$ are *alternatives* (which is to say that they are **MUTUALLY DISJOINT** and their union is the **ENTIRE SAMPLE SPACE S**), then

$$P(E) = \sum_{i=1}^{n} P(E|A_i)P(A_i)$$

Example with three alternatives A, B, C: $P(E) = P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C)$
Example with two alternatives A, A^c: $P(E) = P(E|A)P(A) + P(E|A^c)P(A^c)$

- **Bayes' Law:**

$$P(E|F) = \frac{P(E)}{P(F)} P(F|E)$$

Generally used when we want $P(E|F)$ but $P(F|E)$ is easier to find. Also used in problems that want to update probabilities given that something happened: We have different alternatives and we know

something has happened, and we want the probability of one of the alternatives to have happened. *Example:* 10% *of people have a disease. A test returns positive* 90% *of the time when tested on diseased people and* 5% *of the time when tested on healthy people. If someone's test returned positive, what's the probability that they have the disease?*

- **Double Conditioning** or **Conditional Law of Total Probability:** If events $A_1, A_2, \ldots, A_n$ are *alternatives* (which means they are **MUTUALLY DISJOINT** and their union is the **ENTIRE SAMPLE SPACE S**), then

$$P(E|F) = \sum_{i=1}^{n} P(E|A_i \cap F)P(A_i|F)$$

Example with three alternatives A, B, C:

$P(E|F) = P(E|A \cap F)P(A|F) + P(E|B \cap F)P(B|F) + P(E|C \cap F)P(C|F)$

Example with two alternatives A, A^c:

$P(E|F) = P(E|A \cap F)P(A|F) + P(E|A^c \cap F)P(A^c|F)$

Generally used in Bayes type problems where instead of probability of a certain alternative to have happened we want probability that the given thing will happen again. *Example:* 10% *of people have a disease. A test returns positive* 90% *of the time when tested on diseased people and* 5% *of the time when tested on healthy people. If someone's test returned positive and they re-take it, what's the probability that it will return positive again?*

3 Random Variables

3.1 Discrete

Suppose $x_1, x_2, x_3, \ldots$ (potentially infinitely many values) are the values that a discrete random variable X can take. Some of the things associated to X are:

- **Probability mass function (pmf):** $f_X(x_i) = P(X = x_i)$

  This is the function that gives the probabilities. The sum of all its values is 1, that is, $\sum_i f_X(x_i) = 1$.

- **Cumulative distribution function (cdf):** $F_X(x) = P(X \leq x)$

  This is the function that gives the cumulative sum of all probabilities up to and including the probability of x. It is increasing from 0 to 1, that is, $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$.

- **Expected Value:** $E[X] = \sum_i x_i f_X(x_i) = x_1 P(X = x_1) + x_2 P(X = x_2) + x_3 P(X = x_3) + \cdots$

- **Variance:** $\mathrm{Var}(X) = E[X^2] - E[X]^2 \geq 0$

  We need to be able to calculate $E[X^2]$ before, by doing

$$E[X^2] = \sum_i x_i^2 f_X(x_i) = x_1^2 P(X = x_1) + x_2^2 P(X = x_2) + x_3^2 P(X = x_3) + \cdots$$

- **Standard Deviation:** $\sigma(X) = \sqrt{\mathrm{Var}(X)}$

- **Moments:** $1 = E[X^0] \quad , \quad E[X] \quad , \quad E[X^2] \quad , \quad E[X^3] \quad , \quad \ldots$

- **Moment generating function (mgf):** $M_X(t) = E[e^{tX}] \quad , \quad t \in \mathbf{R}$

  What that means is

$$\begin{aligned} M_X(t) &= \sum_i e^{x_i t} f_X(x_i) \\ &= e^{x_1 t} P(X = x_1) + e^{x_2 t} P(X = x_2) + e^{x_3 t} P(X = x_3) + \cdots \end{aligned}$$

Here are examples of discrete random variables.

Binomial $\mathrm{Bin}(n,p)$ (for $n \in \mathbf{N}$ and $0 \le p \le 1$)

$$f_X(i) = \binom{n}{i} p^i (1-p)^{n-i} \quad , \quad i = 0, 1, \ldots, n$$

$$E[X] = np \quad , \quad \mathrm{Var}(X) = np(1-p) \quad , \quad M_X(t) = (1-p+pe^t)^n$$

This RV counts number of successes in n trials, if each success has probability p of happening independently from the other trials. A Binomial of parameters n and p is a sum of n independent Bernoullis of parameter p, and that can be used to prove the formulas for $E[X]$ and $\mathrm{Var}(X)$.

Bernoulli $\mathrm{Ber}(p)$ (for $0 \le p \le 1$)

$$f_X(0) = 1-p \quad , \quad f_X(1) = p$$

$$E[X] = p \quad , \quad \mathrm{Var}(X) = p(1-p) \quad , \quad M_X(t) = 1-p+pe^t$$

This is an RV that can only take values 0 and 1.

Geometric $\mathrm{Geo}(p)$ (for $0 \le p \le 1$)

$$f_X(i) = (1-p)^{i-1} p \quad , \quad i = 1, 2, 3, \ldots$$

$$E[X] = \frac{1}{p} \quad , \quad \mathrm{Var}(X) = \frac{1-p}{p^2} \quad , \quad M_X(t) = \frac{pe^t}{1-(1-p)e^t} \quad (t < -\ln(1-p))$$

This RV counts number of trials necessary until a success happens, if each success has probability p of happening independently from the other trials.

Negative Binomial $\mathrm{NegBin}(r,p)$ (for $r \in \mathbf{N}$ and $0 \le p \le 1$)

$$f_X(i) = \binom{i-1}{r-1} (1-p)^{i-r} p^r \quad , \quad i = 1, 2, 3, \ldots$$

$$E[X] = \frac{r}{p} \quad , \quad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

This RV counts number of trials necessary until r successes have happened, if each success has probability p of happening independently from the other trials.

Hypergeometric $\mathrm{HypGeo}(n,N,m)$ (for $n, N, m \in \mathbf{N}$ and $m, n \le N$)

$$f_X(i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad , \quad i \text{ from } \max(0, m-N+n) \text{ to } \min(n,m)$$

$$E[X] = \frac{mn}{N}$$

This RV counts the number of special items in a random selection of n items from a box that contains N items which are m special and the rest ordinary.

Poisson $\mathrm{Poi}(\lambda)$ (for $\lambda > 0$)

$$f_X(i) = e^{-\lambda} \frac{\lambda^i}{i!} \quad , \quad i = 0, 1, 2, \ldots$$

$$E[X] = \lambda \quad , \quad \mathrm{Var}(X) = \lambda \quad , \quad M_X(t) = e^{-\lambda} e^{\lambda e^t}$$

This RV is used to approximate probabilities involving *Poisson processes*: Counting number of occurrences of a certain type when it's reasonable to assume that each individual occurrence has low probability and their average over time remains constant. $\mathrm{Poi}(\lambda)$ approximates $\mathrm{Bin}(n,p)$ with $\lambda = np$ when p is small and np is moderately large.

3.2 Continuous

Let X be a continuous random variable. Some of the things associated to X are:

- **Probability density function (pdf):** This is the function that we integrate to find probabilities. This time there is no simple definition of it like

$$f_X(x_i) = P(X = x_i) \qquad \textbf{(NO!)}$$

because this would be zero. We don't have a direct definition of this function, but it should be thought as measuring the likelihood of all possible values of X. Usually it will be defined only for x in a certain range, which explains why we need to worry about endpoints of integration in the formulas. The total probability should still be one: $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$.

- **Cumulative distribution function (cdf):** $F_X(x) = P(X \leq x)$

This is the function that gives the cumulative sum of all probabilities up to x. It is increasing from 0 to 1, that is, $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$. The relationship between pdf and cdf is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t \quad , \quad f_X(x) = F_X'(x)$$

- **Expected Value:** $E[X] = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$
- **Variance:** $\mathrm{Var}(X) = E[X^2] - E[X]^2 \geq 0$

We need to be able to calculate $E[X^2]$ before, by doing $E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\mathrm{d}x$.

- **Standard Deviation:** $\sigma(X) = \sqrt{\mathrm{Var}(X)}$
- **Moments:** $1 = E[X^0] \quad , \quad E[X] \quad , \quad E[X^2] \quad , \quad E[X^3] \quad , \quad \ldots$
- **Moment generating function (mgf):** $M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\mathrm{d}x \quad , \quad t \in \mathbf{R}$

Here are examples of discrete random variables.

Uniform $\mathrm{Uni}(\alpha, \beta)$ (for $\alpha < \beta$)

$$f_X(x) = \frac{1}{\beta - \alpha} \quad , \quad \alpha \leq x \leq \beta$$

$$E[X] = \frac{\alpha + \beta}{2} \quad , \quad \mathrm{Var}(X) = \frac{(\beta - \alpha)^2}{12} \quad , \quad M_X(t) = \frac{e^{\beta t} - e^{\alpha t}}{(\beta - \alpha)t}$$

This represents a random variable that is *uniformly distributed* over $[\alpha, \beta]$, which means it has equal chances of being anywhere in this interval.

Exponential $\mathrm{Exp}(\lambda)$ (for $\lambda > 0$)

$$f_X(x) = \lambda e^{-\lambda x} \quad , \quad 0 \leq x < \infty$$

$$F_X(x) = 1 - e^{-\lambda x} \quad , \quad 0 \leq x < \infty$$

$$E[X] = \frac{1}{\lambda} \quad , \quad \mathrm{Var}(X) = \frac{1}{\lambda^2} \quad , \quad M_X(t) = \frac{\lambda}{\lambda - t} \quad (t < \lambda)$$

This usually models time until something happens. The "*property of no memory*" says that probabilities involving X conditioned on $X \geq a$ for some $a > 0$ are equal to the same probabilities with no conditioning, as if time started counting again at instant a. For example, $P(X \leq 8 | X \geq 2) = P(X \leq 6)$.

Normal $\text{Nor}(\mu, \sigma^2)$ (for $\sigma > 0$)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \quad x \in \mathbf{R}$$

$$E[X] = \mu \quad , \quad \text{Var}(X) = \sigma^2 \quad , \quad M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

This models many different types of processes, discrete or continuous, that have a clear average and may deviate from it in a symmetrical way to the right or the left. We cannot integrate the pdf to find exact probabilities, since the function does not have an elementary integral. We must **renormalize** and use a **standard normal table**. Renormalization means that $Z = \dfrac{X-\mu}{\sigma}$ is a *standard normal random variable*, that is, a normal of mean 0 and variance 1. The cdf of a standard normal, F_Z, is usually denoted by Φ.

Also important to remember is that the sum of independent normals is a normal with sum of the means and sum of the variances:

$$X_i \sim \text{Nor}(\mu_i, \sigma_i^2) \quad , \quad i = 1, \ldots, n \quad \Longrightarrow \quad \sum_{i=1}^{n} X_i \sim \text{Nor}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

This we showed using the mgf.

3.3 Functions of a Random Variable

Discrete If we know $f_X(x)$ and want to find $f_Y(y)$ where $Y = g(X)$, we usually can just plug-in $y = g(x)$ into the pmf (as long as g is *injective*):

$$f_Y(g(x)) = f_X(x)$$

For example, if $f_X(3) = 0.4$ and $f_X(4) = 0.6$, then the pmf of $Y = X^2$ is $f_Y(9) = 0.4$ and $f_Y(16) = 0.6$.

The expected value of $Y = g(X)$ can be found by

$$E[g(X)] = \sum_i g(x_i) f_X(x_i) = g(x_1)P(X = x_1) + g(x_2)P(X = x_2) + g(x_3)P(X = x_3) + \cdots$$

For example we do this to find $E[X^2]$ before finding $\text{Var}(X)$.

Continuous If we know $f_X(x)$ and want to find $f_Y(y)$ where $Y = g(X)$, what we **CANNOT DO** is simply plug-in $y = g(x)$ into the pdf. Instead:

- Find $F_X(x) = \displaystyle\int_{-\infty}^{x} f_X(t)\mathrm{d}t$.
- Find $F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le h(y))$, where h is the inverse function of g.
- Find $f_Y(y) = F_Y'(y)$.

The expected value of $Y = g(X)$ can be found by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\mathrm{d}x$$

For example we do this to find $E[X^2]$ before finding $\text{Var}(X)$.

4 Joint Distribution

When the problem involves two random variables, we need to know how they are distributed jointly.

4.1 Discrete Case

Joint pmf: The main function from which everything can be deduced is the joint probability mass function

$$f_{XY}(x, y) = P(X = x \cap Y = y)$$

These values are usually written in a table.

	$X = x_1$	$X = x_2$	$X = x_3$	$\dots$
$Y = y_1$	$f_{XY}(x_1, y_1)$	$f_{XY}(x_2, y_1)$	$f_{XY}(x_3, y_1)$	$\dots$
$Y = y_2$	$f_{XY}(x_1, y_2)$	$f_{XY}(x_2, y_2)$	$f_{XY}(x_3, y_2)$	$\dots$
$Y = y_3$	$f_{XY}(x_1, y_3)$	$f_{XY}(x_2, y_3)$	$f_{XY}(x_3, y_3)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

They are the probability that (X, Y) has the value (x, y). The sum of all values is one, which is to say that, if $x_1, x_2, \dots$ are the values of X and $y_1, y_2, \dots,$ are the values of Y, then

$$\sum_i \sum_j f_{XY}(x_i, y_j) = 1$$

Marginals: The marginals are the pmf's of X and Y, obtained by summing rows and columns of the table:

$$f_X(x) = \sum_j f_{XY}(x, y_j) \quad , \quad f_Y(y) = \sum_i f_{XY}(x_i, y)$$

	$X = x_1$	$X = x_2$	$X = x_3$	$\dots$	
$Y = y_1$	$f_{XY}(x_1, y_1)$	$f_{XY}(x_2, y_1)$	$f_{XY}(x_3, y_1)$	$\dots$	$\to f_Y(y_1)$
$Y = y_2$	$f_{XY}(x_1, y_2)$	$f_{XY}(x_2, y_2)$	$f_{XY}(x_3, y_2)$	$\dots$	$\to f_Y(y_2)$
$Y = y_3$	$f_{XY}(x_1, y_3)$	$f_{XY}(x_2, y_3)$	$f_{XY}(x_3, y_3)$	$\dots$	$\to f_Y(y_3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
	$\downarrow$	$\downarrow$	$\downarrow$		
	$f_X(x_1)$	$f_X(x_2)$	$f_X(x_3)$		

Independence: The random variables X and Y are **INDEPENDENT** if

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

This means the product of the sums in the margins gives the entries of the table. Independence is sometimes assumed or given (then you can **use** that joint pmf is product of marginals), and sometimes asked (then you need to **show** that joint pmf is product of marginals).

Conditional pmf: The conditional pmf of X given $Y = y$, and the conditional pmf of Y given $X = x$ are

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

They give the probability distribution of one variable given that the other has the given value. They are the entries of the table divided by the corresponding marginal of the given variable.

Conditional expectation: The conditional expected value of X given $Y = y$, and the conditional expected value of Y given $X = x$ are

$$E[X|Y = y] = \sum_i x_i f_{X|Y}(x_i|y) = x_1 f_{X|Y}(x_1|y) + x_2 f_{X|Y}(x_2|y) + x_3 f_{X|Y}(x_3|y) + \dots$$

$$E[Y|X=x] = \sum_j y_j f_{Y|X}(y_j|x) = y_1 f_{Y|X}(y_1|x) + y_2 f_{Y|X}(y_2|x) + y_3 f_{Y|X}(y_3|x) + \dots$$

These are the usual formulas for expected values, but we use *conditional pmf's* instead of the regular pmf's.

Law of total expectation: This tells us how to find $E[X]$ by conditioning on the possible values of Y, and vice-versa:

$$E[X] = \sum_j E[X|Y=y_j]P(Y=y_j) = E[X|Y=y_1]P(Y=y_1) + E[X|Y=y_2]P(Y=y_2) + \dots$$

$$E[Y] = \sum_i E[Y|X=x_i]P(X=x_i) = E[Y|X=x_1]P(X=x_1) + E[Y|X=x_2]P(X=x_2) + \dots$$

The numbers $P(X=x_i)$ and $P(Y=y_j)$ are of course the marginals $f_X(x_i)$ and $f_Y(y_j)$.

Covariance: The covariance between X and Y is

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

This is **positive** when knowing that X is large makes Y more likely to be large, or knowing that X is small makes Y more likely to be small. This is **negative** when X large implies Y small and vice-versa. This is **zero** when X and Y are independent (although it can also be zero even if they are not independent). The RV's must have finite expected values for this to be defined.

Correlation: The correlation between X and Y is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

We always have $-1 \le \rho(X,Y) \le 1$. Correlation is only ± 1 when there is a linear relationship between X and Y, that is, $Y = aX + b$. The RV's must have finite expected values and nonzero variances for this to be defined.

Functions of two RV's: The way to obtain the pmf of $Z = g(X,Y)$ is usually to list out its possible values and find its probabilities from the table. Once we have this we can for example calculate

$$E[Z] = \sum_i z_i f_Z(z_i) = z_1 P(Z=z_1) + z_2 P(Z=z_2) + z_3 P(Z=z_3) + \cdots$$

This could also have been calculated using the joint pdf:

$$E[g(X,Y)] = \sum_i \sum_j g(x_i, y_j) f_{XY}(x_i, y_j)$$

This means calculate the value of $g(X,Y)$ on each entry of the table and multiply that by the probability which is that entry, then sum over all possible entries. Such an expected value is needed for example in calculating $\mathrm{Cov}(X,Y)$, because for that we need to know $E[XY]$.

4.2 Continuous Case

Joint pdf: The main function from which everything can be deduced is the joint probability density function f_{XY}. There is no formula for it, but it should still be thought of as the function that gives a measure of likelihood for the possible values of the random vector (X,Y). We integrate it to find probabilities. The total probability should still be one:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x\mathrm{d}y = 1$$

Marginals: The marginals are the pdf's of X and Y, obtained by integrating out the *other* variable:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y \quad , \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x$$

Independence: The random variables X and Y are **INDEPENDENT** if

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

Independence is sometimes assumed or given (then you can **use** that joint pdf is product of marginals), and sometimes asked (then you need to **show** that joint pdf is product of marginals).

Conditional pdf: The conditional pdf of X given $Y = y$, and the conditional pdf of Y given $X = x$ are

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

They give the probability distribution of one variable given that the other has the given value. They are the functions that you integrate in order to find conditional probabilities when the condition given is of the form $X = a$ or $Y = b$ for some number a or b. For example:

$$P(X \geq 5 \mid Y = 3) = \int_5^\infty f_{X|Y}(x|3)\mathrm{d}x \quad , \quad P(-1 \leq Y \leq 2 \mid X = 0) = \int_{-1}^2 f_{Y|X}(y|0)\mathrm{d}y$$

This is because, if we tried to use the formula for conditional probability, we would get 0 over 0, since $P(Y = a) = 0$ in the continuous case:

$$P(X \geq 5 \mid Y = 3) = \frac{P(X \geq 5 \cap Y = 3)}{P(Y = 3)} \qquad \textbf{(NO!)}$$

Conditionals involving *inequalities* as conditions should still be treated by the definition of conditional probability. For example:

$$P(X \leq 2 \mid Y \geq 4) = \frac{P(X \leq 2 \cap Y \geq 4)}{P(Y \geq 4)} = \frac{\int_{-\infty}^2 \int_4^\infty f_{XY}(x,y)\mathrm{d}y\mathrm{d}x}{\int_4^\infty f_Y(y)\mathrm{d}y}$$

Conditional expectation: The conditional expected value of X given $Y = y$, and the conditional expected value of Y given $X = x$ are

$$E[X|Y = y] = \int_{-\infty}^\infty x f_{X|Y}(x|y)\mathrm{d}x$$

$$E[Y|X = x] = \int_{-\infty}^\infty y f_{Y|X}(y|x)\mathrm{d}y$$

These are functions of y and x, respectively. These are the usual formulas for expected values, but we use *conditional pdf's* instead of the regular pdf's.

Law of total expectation: This tells us how to find $E[X]$ by conditioning on the possible values of Y, and vice-versa:

$$E[X] = \int_{-\infty}^\infty E[X|Y = y] f_Y(y)\mathrm{d}y$$

$$E[Y] = \int_{-\infty}^\infty E[Y|X = x] f_X(x)\mathrm{d}x$$

Covariance and Correlation: Exactly the same as in the discrete case. Only now we need integrals to find the necessary expected values.

Functions of two RV's: The way to obtain the pdf of $Z = g(X,Y)$ is:

- Find $F_Z(z) = P(Z \leq z) = P(g(X,Y) \leq z)$ by integrating the joint pdf in the appropriate region defined by the inequality $g(X,Y) \leq z$.
- Find $f_Z(z) = F_Z'(z)$.

A particular example is when $Z = X + Y$ for **INDEPENDENT** X and Y. This procedure will yield the formula

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) \mathrm{d}t = \int_{-\infty}^{\infty} f_X(z-t) f_Y(t) \mathrm{d}t$$

Once we have the pdf of $g(X, Y)$ we can for example calculate

$$E[Z] = \int_{-\infty}^{\infty} z f_Z(z) \mathrm{d}z$$

This could also have been calculated using the joint pdf:

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) \mathrm{d}x \mathrm{d}y$$

Such an expected value is needed for example in calculating $\mathrm{Cov}(X, Y)$, because for that we need to know $E[XY]$.

5 Properties of Expectation, Variance and Covariance

- **Linearity:**

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

$$\text{Cov}(aX, bY) = ab\text{Cov}(X,Y)$$

Variance is **NOT** linear: $\text{Var}(aX+b) = a^2\text{Var}(X)$

- **Sums of random variables:**

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$ only when the random variables $X_1, \ldots, X_n$ are **INDEPENDENT**

A good way to find variance of a sum when the RV's are not independent is to use

$$E[X^2] - E[X] = 2\sum_{i<j} E[X_i X_j] \qquad \left(X = \sum_{i=1}^{n} X_i\right)$$

to find $E[X^2]$ and then find the variance. This is particularly useful when the RV's X_i are Bernoullis. The sum on the right hand side is over all pairs of the random variables, so there are $\binom{n}{2}$ terms. Here is an example of this formula with 3 RV's:

$$E[(X+Y+Z)^2] - E[X+Y+Z] = 2\Big(E[XY] + E[XZ] + E[YZ]\Big)$$

- **Relationship between variance and covariance**:

$$\text{Var}(X) = \text{Cov}(X,X)$$

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$$

This last formula is another way to find variance of a sum if it's easy to find the covariances. The last sum contains $\binom{n}{2}$ terms, as before. Here are examples of this formula:

With two RV's: $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$

With three RV's: $\text{Var}(X+Y+Z) = \text{Var}(X)+\text{Var}(Y)+\text{Var}(Z)+2\Big(\text{Cov}(X,Y)+\text{Cov}(X,Z)+\text{Cov}(Y,Z)\Big)$

- **Moment generating function**

The mgf is important because it can be used to calculate:

  - Moments of one random variable:

$$M_X'(0) = E[X] \quad , \quad M_X''(0) = E[X^2] \quad , \quad M_X'''(0) = E[X^3] \quad , \ldots$$

  - The mgf of a sum of two **INDEPENDENT** random variables:

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

6 Estimation Inequalities and Limit Theorems

6.1 Inequalities

For these inequalities we usually don't worry about the difference between $\leq$ and $<$, or between $\geq$ and $>$, because they are usually applied to continuous RV's or to approximate discrete RV's by continuous ones.

- **Markov's Inequality:** If X is a random variable that can only be ≥ 0, then

$$P(X \geq a) \;\leq\; \frac{E[X]}{a} \quad , \quad \text{for all } a > 0$$

This also implies

$$P(X \leq a) \;\geq\; 1 - \frac{E[X]}{a} \quad , \quad \text{for all } a > 0$$

These are used when we want to bound probabilities but we only have information about the mean.

- **Chebyshev's Inequality:** If X is a random variable with $E[X] < \infty$ and $\mathrm{Var}(X) < \infty$, then

$$P(|X - \mu| \geq k) \;\leq\; \frac{\sigma^2}{k^2} \quad , \quad \text{for all } k > 0$$

This also implies

$$P(|X - \mu| \leq k) \;\geq\; 1 - \frac{\sigma^2}{k^2} \quad , \quad \text{for all } k > 0$$

In here we denote $E[X]$ by μ and $\mathrm{Var}(X)$ by σ^2, even though X may not be a normal random variable. These are used when we want to bound probabilities but we only have information about the mean and the variance.

6.2 Limit Theorems

These deep results say what happens to the average of a large number of random variables. They give meaning to our interpretation of probability as a measure of how often things happen in the long run.

- **Weak Law of Large Numbers:** Let $X_1, X_2, \ldots$ be independent random variables having a common probability distribution (pdf or pmf). Let us denote their common mean by μ. Then

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| < \varepsilon\right) = 1 \quad , \quad \text{for all } \varepsilon > 0$$

This is saying that, when the same random experiment is performed several times, the average result will have a high probability of being close to the expected value of each experiment.

- **Strong Law of Large Numbers:** Let $X_1, X_2, \ldots$ be independent random variables having a common probability distribution (pdf or pmf). Let us denote their common mean by μ. Then

$$P\left(\lim_{n\to\infty}\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| = 0\right) = 1$$

This is a slightly stronger way to say something which has the same interpretation as above.

- **Central Limit Theorem:** Let $X_1, X_2, \ldots$ be independent random variables having a common probability distribution (pdf or pmf). Let us denote their common mean by μ and their common variance by σ^2. Then

$$\lim_{n\to\infty} P\left(X_1 + \cdots + X_n \leq a\right) = \frac{1}{\sigma\sqrt{2\pi n}} \int_{-\infty}^{a} e^{-(x-n\mu)^2/2n\sigma^2} \mathrm{d}x \quad , \quad \text{for all } a \in \mathbf{R}$$

This is saying that the sum of a large number, n, of identical random variables is always approximately a normal with mean $n\mu$ and variance $n\sigma^2$, independently of what type of random variables they were. Because of this we can approximate probabilities related to such a sum by normalizing it first:

$$Z = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \text{ is approximately the standard normal.}$$