Northwestern University

Probability

MATH 310-1

summary or som

Author:
Elliott Yoon

6 December 2022

1 Combinatorial Analysis

If an experiment consists of two events A and B, there are n outcomes in event A and m outcomes in event B, then there are nm possible outcomes of the experiment. This is called the **multiplication principle**.

There are $n! = n(n-1)\cdots 3 \cdot 2 \cdot 1$ possible linear orderings of n items, where $0! = 1$. The number of ways to choose a subgroup of size i from a set of size n (called the **binomial coefficient** is

$$\binom{n}{i} = \frac{n!}{(n-i)!i!}$$

when $0 \leq i \leq n$, and is 0 otherwise.

For $n_1, \dots, n_r$ summing to n, the number of divisions of n items into r distinct disjoint subgroups of sizes $n_1, n_2, \dots, n_r$ is

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

2 Axioms of Probability

For each event A of the sample space S, we suppose that the probability of A, $P(A)$ is defined such that

1. $0 \leq P(A) \leq 1$,
2. $P(S) = 1$,
3. For mutually exclusive events $A_i, i \geq 1$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Theorem 2.1 (Inclusion-exclusion)

For events A, B,

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

which can be generalized to give

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{i<j} P(A_i A_j) + \sum\sum\sum_{i<j<k} P(A_i A_j A_k + \cdots + (-1)^{n+1} P(A_1 \cdots A_n).$$

3 Conditional Probability and Independence

Definition 3.1. For events E and F, the conditional probability of E given F has occurred is

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

Theorem 3.2 (Multiplication Rule)
For events $E_1, \ldots, E_n$:

$$P(E_1, E_2 \cdots E_n) = P(E_1)P(E_2|E_1) \cdots P(E_n|E_1 \cdots E_{n-1}).$$

Remark 3.3. An important identity is

$$P(E) = P(E|F)P(F) + P(E|F^c)P(F^c),$$

which can be used to compute $P(E)$ by conditioning on whether F occurs.

Theorem 3.4 (Bayes's Formula)
If $F_i, i = 1, \ldots, n$ are mutually exclusive events whose union is the entire sample space, then

$$P(F_j|E) = \frac{P(E|f_j)P(F_j)}{\sum_{i=1}^n P(E|F_i)P(F_i)}.$$

Definition 3.5. We say E and F are **independent** if $P(EF) = P(E)P(F)$.

Remark 3.6. This is equivalent to $P(E|F) = P(E)$ and $P(F|E) = P(F)$.

The events $E_1, \ldots, E_n$ are independent if, for any subset $E_{i_1}, \ldots, E_{i_r}$ of them,

$$P(E_{i_1}, \ldots, E_{i_r}) = P(E_{i_1} \cdots P(E_{i_r}).$$

4 Random Variables

Definition 4.1. A real-valued function defined on the outcome of a probability experiment is called a **random variable**.

If X is a random variable, the **distribution function** $F(x)$ of X is defined

$$F(x) = P\{X \leq x\}.$$

A random variable whose set of possible values is either finite or countably infinite is called **discrete**, with **probability mass function**

$$p(x) = P\{X = x\}.$$

The **expected value** (or *mean*) of X is

$$E[X] = \sum_{x:p(x)>0} xp(x).$$

Theorem 4.2

$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$

Definition 4.3. The **variance** of a random variable X is defined by

$$\text{Var}(X) = E[(X - E[X]^2] = E[X^2] - (E[X])^2.$$

4.1 Important Probability Distributions

Definition 4.4. The **binomial random variable** can be interpreted as being the number of successes that occur when n independent trials, each of which has probability of success p, are performed. It has probability mass function

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

and mean and variance

$$E[X] = np \quad \mathrm{Var}(X) = np(1-p).$$

Remark 4.5. If X is a binomial random variable,

$$E[X^2] = np[(n-1)p + 1].$$

Definition 4.6. The **Poissson random variable** with parameter λ can be used to approximate binomial random variables, where $\lambda = np$. It has probability mass function (giving probability $p(X)$ of X successes):

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x \geq 0$$

and mean and variance

$$E[X] = \mathrm{Var}(X) = \lambda.$$

Remark 4.7. If X is a Poisson random variable,

$$E[X^2] = \lambda(\lambda + 1).$$

Definition 4.8. The **geometric random variable** represents the number of independent trials of probability p it takes for the first success. Its probability mass function is

$$p(i) = p(1-p)^{i-1} \quad i = 1, 2, \dots$$

and has mean and variance

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

Remark 4.9. If X is a geometric random variable,

$$E[X^2] = \frac{q+1}{p^2}.$$

4.2 Alone. He's just like me!

Theorem 4.10 (Mean of the sum is the sum of the means)

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

5 Continuous Random Variables

Definition 5.1. A random variable X is *continuous* if there is a nonnegative function f, called the **probability density function** of X, such that for any set B:

$$P\{X \in B)\} = \int_B f(x)\,dx.$$

If X is continuous, its distribution function F is differentiable and

$$\frac{d}{dx}F(x) = f(x).$$

The expected value of a continuous random variable X is defined by

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx.$$

Theorem 5.2

For any function g,

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

Remark 5.3. Just like in the discrete case, the variance of X is defined to be

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

5.1 Important Probability Distributions

Definition 5.4. A random variable X is said to be **uniform** over the interval (a, b) if its probability density function is given by

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

It has mean and variance

$$E[X] = \frac{a+b}{2} \qquad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Definition 5.5. A random variable X is said to be **normal** with parameters μ, σ^2 if its probability density function is given by

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

It has mean and variance

$$E[X] = \mu \quad \text{Var}(X) = \sigma^2.$$

If X is normal with mean μ and variance σ^2, then Z, defined by

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

is normal with mean 0 and variance 1.

> **Remark 5.6.** When n is large, the probability distribution function of a binomial random variable with parameters n and p can be approximated by that of a normal variable with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

Definition 5.7. An **exponential random variable** with parameter λ has probability density function of the form

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

with mean and variance

$$E[X] = \frac{1}{\lambda} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

An exponential random variable X intuitively represents the time it takes for the first success in a collection of independent events. See this stack exchange post for a more solid intuition.

> **Remark 5.8.** The exponential random variable is the *only* random variable that is **memoryless**, meaning that for $s, t > 0$
>
> $$P\{X > s + t | X > t\} = P\{X > s\}.$$
>
> If X represents the life of an item, then the memoryless property states that, for any t, the remaining life of a t-year-old item has the same probability distribution as the life a new item.

Definition 5.9. Let X be a nonnegative continuous random variable with distribution function F and density function f. The function

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \quad t \geq 0$$

is called the **hazard rate** (or *failure rate*) of F. Notice that if X is exponential with parameter λ, then $\lambda(t) = \lambda$. In fact, the exponential distribution uniquely has constant hazard rate.

Definition 5.10. A random variable is said to have **gamma** distribution with parameters α and λ if its probability density function is equal to

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \quad x \geq 0$$

and 0 otherwise. Note that the gamma function $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1}\, dx.$$

The mean and variance of a gamma random variable are

$$E[X] = \frac{\alpha}{\lambda} \quad \text{Var}(X) = \frac{\alpha}{\lambda^2}.$$

Whereas exponential random variables represent the time it takes for the first occurrence of a given event, the gamma random variable represents the time it takes for the α-th occurrence.

6 Jointly Distributed Random Variables

Definition 6.1. The **joint cumulative probability distribution function** of the pair of random variables X and Y is defined by

$$F(x, y) = P\{X \leq x, Y \leq y\} \quad -\infty < x, y < \infty.$$

To find the individual probability distribution functions of X and Y, use

$$F_X(x) = \lim_{y \to \infty} F(x, y) \quad F_Y(y) = \lim_{x \to \infty} F(x, y).$$

- If X, Y are both discrete random variables, then their **joint probability mass function** is defined by

  $$p(i, j) = P\{X = i, Y = j\}.$$

  The individual mass functions are

  $$P\{X = i\} = \sum_j p(i, j) \quad P\{Y = j\} = \sum_i p(i, j).$$

- The random variables X, Y are *jointly continuous* if there is a function $f(x, y)$, called the **joint probability density function** such that for any two dimensional set C,

  $$P\{(X, Y) \in C\} = \iint_C f(x, y)\, dx\, dy.$$

Theorem 6.2 (Marginal Density Functions)
If X, Y are jointly continuous, they are individually continuous with (marginal) density functions

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

Theorem 6.3 (Independence of Jointly Continuous Random Variables)
Random variables X and Y are *independent* if for all sets A, B,

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$$

This holds generally for $X_1, \ldots, X_n$.

> **Remark 6.4.** IF the joint distribution function (or joint probability mass function in the discrete case) factors into a part depending only on x and a part depending only on y, then X and Y are independent

Theorem 6.5 (Convolutions)
If X, Y are independent continuous random variables, then the distribution function of their sum can be obtained as follows:

$$F_{X+Y}(a) = \int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\, dy.$$

> **Remark 6.6.** This follows from
> $$F_{X+Y}(a) = \iint_{X+Y\le a} f(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)f_Y(y)\,dx\,dy = \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\,dy.$$

6.1 Sums of Specific Distribution

Theorem 6.7

If $X_i, i = 1, \ldots, n$ are independent..

1. normal random variables with respective parameters μ_i and σ_i^2, then $\sum_{i=1}^n X_i$ is normal with parameters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$.
2. Poisson random variables with respective parameter λ_i, then $\sum_{i=1}^n S_i$ is Poisson with parameter $\sum_{i=1}^n \lambda_i$.

6.2 Conditional Probability

Definition 6.8. If X, Y are discrete random variables, then the **conditional probability mass function** of X given that $Y = y$ is defined by
$$P\{X = x | Y = y\} = \frac{p(x,y)}{p_Y(y)}$$
where p is their joint probability mass function.

Definition 6.9. If X, Y are independent continuous random variables, then the **conditional probability density function** of X given that $Y = y$ is defined by
$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}.$$

7 Properties of Expectation

If X and Y have a joint probability mass function $p(x,y)$, then
$$E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$$
whereas if they have joint density function $f(x,y)$, then
$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy.$$

Corollary 7.1

It follows immediately then that
$$E[X+Y] = E[X] + E[Y]$$
and, more generally,
$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i].$$

Definition 7.2. The **covariance** between random variables X and Y is

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

Remark 7.3. A useful identity is

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

Definition 7.4. The **correlation** between X and Y, denoted $\rho(X,Y)$ is defined by

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

7.1 Conditional Expectation

Definition 7.5.

- If X, Y are jointly discrete random variables, then the **conditional expected value** of X, given that $Y = y$, is
$$E[X|Y=y] = \sum_{x} xP\{X = x|Y = y\}.$$
- If X, Y are jointly continuous random variables, then
$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx$$
where $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$ is the conditional probability of X given that $Y = y$.

Remark 7.6. Conditional expectations satisfy all the properties of ordinary expectations.

Theorem 7.7

Let $E[X|Y]$ denote the function of Y whose value at $Y = y$ is $E[X|Y = y]$. Then

$$E[X] = E[E[X|Y]].$$

1. For discrete random variables,
$$E[X] = \sum_{y} E[X|Y=y]P\{Y=y\}$$
2. For continuous random variables,
$$E[X] = \int_{-\infty}^{\infty} E[X|Y=y]f_Y(y)\,dy$$

We can use these equations to obtain $E[X]$ by first "conditioning" on the value of some other random variable Y. Also, for any event A, $P(A) = E[I_A]$, where I_A is 1 if A occurs and 0 otherwise, so we can use these equations to compute probabilities.

Definition 7.8. The conditional variance of X, given $Y = y$, is defined

$$\text{Var}(X|Y = y) = E[(X - E[X|Y = y])^2|Y = y].$$

Letting $\text{Var}(X|Y)$ be the function of Y whose value at $Y = y$ is $\text{Var}(X|Y = y)$,

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]).$$

7.2 Moment Generating Functions

Definition 7.9. The **moment generating function** of X is defined as

$$M(t) = E[e^{tX}].$$

The moments of X, i.e. $E[X], E[X^2], \ldots, E[X^n]$, are obtained by successively differentiating $M(t)$ and then evaluating the resulting quantity at $t = 0$. Specifically, we have

$$E[X^n] = \frac{d^n}{dt^n} M(t)\bigg|_{t=0} \qquad n = 1, 2, \ldots$$

Remark 7.10. Two useful results arise from moment generating functions:

1. The MGF *uniquely* determines the distribution function of the random variable, and
2. The MGF of the sum of independent random variables is equal to the product of *their* moment generating functions.

8 Limit Theorems

8.1 Probability Bounds

Using the following two theorems, we can derive bounds on probabilities when only the mean (or both the mean and the variance) are known.

Theorem 8.1 (Markov's Inequality)

If X is a random variable that takes only non-negative values, then for any $a > 0$,

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

Theorem 8.2 (Chebyshev's Inequality)

If X is a random variable with finite mean μ and variance σ^2, then, for any value $k > 0$,

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}.$$

8.2 The Big Kahunas

Theorem 8.3 (The Weak Law of Large Numbers)
Let $X_1, X_2, \ldots,$ be a sequence of independent and identically distributed random variables, each having finite mean $E[X_i] = \mu$. Then, for any $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \quad \text{as } n \to \infty.$$

Remark 8.4. This requires only that the random variables in the sequence have a finite mean μ. It states that, with probability 1, the average of the first n of them will converge to μ as n goes to infinity.

This implies that if A is any specified event of an experiement for which independent replications are performed, then the limiting proportion of experiments whose outcomes are in A will, with probability 1, equal $P(A)$.

After all this dum dum probability hoo-hah, we get to the real deal:

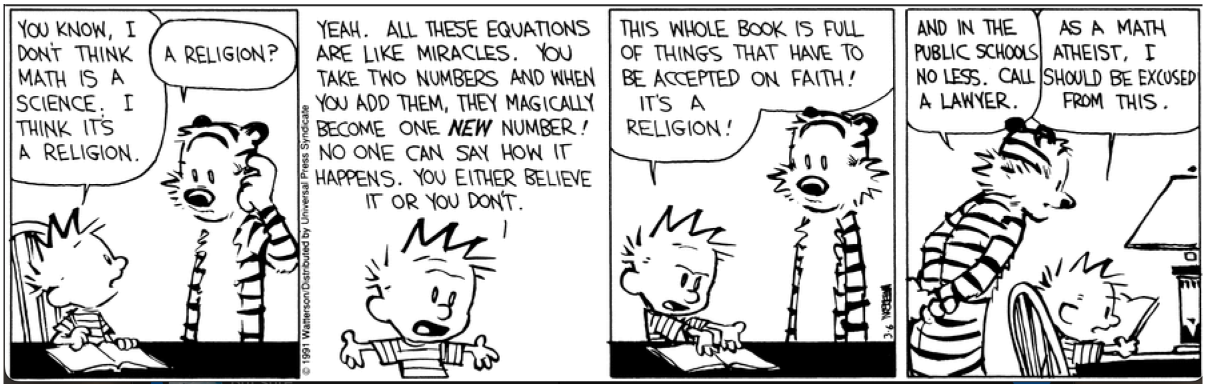
Theorem 8.5 (The Central Limit Theorem)
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\, dx \quad \text{as } n \to \infty.$$

Remark 8.6. This says that if the random variables have a finite mean μ and a finite variance σ^2, then the distribution of the sum of the first n of them is, for large n, approximately that of a normal random variable with mean $n\mu$ and variance $n\sigma^2$.


YOU KNOW, I DON'T THINK MATH IS A SCIENCE. I THINK IT'S A RELIGION.
A RELIGION?
YEAH. ALL THESE EQUATIONS ARE LIKE MIRACLES. YOU TAKE TWO NUMBERS AND WHEN YOU ADD THEM, THEY MAGICALLY BECOME ONE NEW NUMBER! NO ONE CAN SAY HOW IT HAPPENS. YOU EITHER BELIEVE IT OR YOU DON'T.
© 1991 Watterson/Distributed by Universal Press Syndicate
THIS WHOLE BOOK IS FULL OF THINGS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND IN THE PUBLIC SCHOOLS NO LESS. CALL A LAWYER.
AS A MATH ATHEIST, I SHOULD BE EXCUSED FROM THIS.