Stat 110 Notes

Elvin Lo

Fall 2023

2024.9.25

Preface

These notes follow the second edition of Blitzstein and Hwang, *Introduction to Probability*, the text accompanying STAT 110 at Harvard College.

Contents

1 Probability and counting **1**
- 1.1 Why study probability? . . . 1
- 1.2 Sample spaces and Pebble World . . . 1
- 1.3 Naive definition of probability . . . 1
- 1.4 How to count . . . 1
- 1.5 Story proofs . . . 1
- 1.6 Non-naive definition of probability . . . 1

2 Conditional probability **3**
- 2.1 The importance of thinking conditionally . . . 3
- 2.2 Definition and intuition . . . 3
- 2.3 Bayes' rule and the law of total probability . . . 3
- 2.4 Conditional probabilities are probabilities . . . 4
- 2.5 Independence of events . . . 4
- 2.6 Coherency of Bayes' rule . . . 5
- 2.7 Conditioning as a problem solving strategy . . . 6
- 2.8 Pitfalls and paradoxes . . . 7

3 Random variables **9**
- 3.1 Random variables . . . 9
- 3.2 Distributions and probability mass functions . . . 9
- 3.3 Bernoulli and Binomial . . . 9
- 3.4 Hypergeometric . . . 10
- 3.5 Discrete uniform . . . 10
- 3.6 Cumulative distribution functions . . . 10
- 3.7 Functions of random variables . . . 11
- 3.8 Independence of random variables . . . 11
- 3.9 Connections between binomial and hypergeometric . . . 12

4 Expectation **13**
- 4.1 Definition of expectation . . . 13
- 4.2 Linearity of expectation . . . 13
- 4.3 Geometric and negative binomial . . . 13
- 4.4 Indicator r.v.s and the fundamental bridge . . . 15
- 4.5 Law of the unconscious statistician (LOTUS) . . . 16
- 4.6 Variance . . . 16
- 4.7 Poisson . . . 17
- 4.8 Connections between Poisson and Binomial . . . 18
- 4.9 Using probability and expectation to prove existence . . . 19

5 Continuous random variables **20**
- 5.1 Probability density functions . . . 20
- 5.2 Uniform . . . 21
- 5.3 Universality of the uniform . . . 21
- 5.4 Normal . . . 22
- 5.5 Exponential . . . 23

5.6 Poisson processes . . . 24
5.7 Symmetry of i.i.d. continuous r.v.s . . . 25

6 Moments **26**
6.1 Summaries of a distribution . . . 26
6.2 Interpreting moments . . . 27
6.3 Sample moments . . . 28
6.4 Moment generating functions . . . 29
6.5 Generating moments with MGFs . . . 30
6.6 Sums of independent r.v.s via MGFs . . . 31
6.7 Probability generating functions . . . 31

7 Joint distributions **32**
7.1 Joint, marginal, and conditional . . . 32
7.2 2D LOTUS . . . 36
7.3 Covariance and correlation . . . 37
7.4 Multinomial . . . 38
7.5 Multivariate normal . . . 39

8 Transformations **41**
8.1 Change of variables . . . 41
8.2 Convolutions . . . 41
8.3 Beta . . . 42
8.4 Gamma . . . 45
8.5 Beta-Gamma connections . . . 47
8.6 Order statistics . . . 47

9 Conditional expectation **49**
9.1 Conditional expectation given an event . . . 49
9.2 Conditional expectation given an r.v. . . . 49
9.3 Properties of conditional expectation . . . 50
9.4 Geometric interpretation of conditional expectation . . . 50
9.5 Conditional variance . . . 50
9.6 Adam and Eve examples . . . 51

10 Inequalities and limit theorems **52**
10.1 Inequalities . . . 52
10.2 Law of large numbers . . . 53
10.3 Central limit theorem . . . 54
10.4 Chi-Square and Student-t . . . 55

11 Markov chains **57**
11.1 Markov property and transition matrix . . . 57
11.2 Classification of states . . . 58
11.3 Stationary distribution . . . 59
11.4 Reversibility . . . 60

12 Markov chains Monte Carlo **62**
12.1 Metropolis-Hastings . . . 62

A Table of distributions **63**

B Problem-solving strategies **64**

C Math **64**

1 Probability and counting

1.1 Why study probability?

1.2 Sample spaces and Pebble World

Definition 1.1. Sample space and event
The sample space S of an experiment is the set of all possible outcomes of the experiment. An event A is a subset of the sample space S, and we say that A occurred if the actual outcome is in A.

Theorem 1.2. De Morgan's laws

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

1.3 Naive definition of probability

Definition 1.3. Naive definition of probability
Let A be an event for an experiment with a finite sample space S. The naive probability of A is

$$P_{\text{naive}}(A) = \frac{|A|}{|S|}.$$

1.4 How to count

1.5 Story proofs

1.6 Non-naive definition of probability

Definition 1.4. General definition of probability
A probability space consists of a sample space S and a probability function P which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output. The function P must satisfy the following axioms:

1. $P(\emptyset) = 0, P(S) = 1$.
2. If $A_1, A_2, \ldots$ are disjoint events, then

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

Remark 1.5. Frequentist and Bayesian views
There exist two main views of interpreting probability probability functions:

- The frequentist view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment: if we say a coin has probability $1/2$ of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.
- The Bayesian view of probability is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses like "candidate A will win the election" or "the defendant is guilty" even if it isn't possible to repeat the same election or the same crime over and over again.

Theorem 1.6. Properties of probability
Probability has the following properties, for any events A and B.

1. $P(A^c) = 1 - P(A)$.
2. If $A \subseteq B$, then $P(A) \leq P(B)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

2 Conditional probability

2.1 The importance of thinking conditionally

2.2 Definition and intuition

Definition 2.1. Conditional probability
If A and B are events with $P(B) > 0$, then define

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

Theorem 2.2. Probability of the intersection of n events
For any events $A_1, \ldots, A_n$ with $P(A_1, A_2, \ldots, A_{n-1}) > 0$,

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \ldots P(A_n|A_1, \ldots, A_{n-1}).$$

2.3 Bayes' rule and the law of total probability

Theorem 2.3. Bayes' rule

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

Proof. Follows from the equality $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$. □

Definition 2.4. Odds
The odds of an event A are

$$\text{odds}(A) = P(A)/P(A^c).$$

Of course we can also convert from odds back to probability:

$$P(A) = \text{odds}(A)/(1 + \text{odds}(A)).$$

Theorem 2.5. Odds form of Bayes' rule
For any events A and B with positive probabilities, the odds of A after conditioning on B are

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \frac{P(A)}{P(A^c)}.$$

In words, this says that the posterior odds $P(A|B)/P(A^c|B)$ are equal to the prior odds $P(A)/P(A^c)$ times the factor $P(B|A)/P(B|A^c)$, which is known in statistics as the likelihood ratio.

Proof. Take the Bayes' rule expression for $P(A|B)$ and divide it by the Bayes' rule expression for $P(A^c|B)$. □

Theorem 2.6. Law of total probability, LOTP
Let $A_1, \ldots, A_n$ be a partition of the sample space S with $P(A_i) > 0$ for all i. Then

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

Remark 2.7.
Note that applying Bayes' rule and the LOTP together is a common strategy.

2.4 Conditional probabilities are probabilities

Remark 2.8.
When we condition on an event E, the laws of probability operate just as before. As basic examples:

- $P(A^c|E) = 1 - P(A|E)$;
- PIE: $P(A \cup B|E) = P(A|E) + P(B|E) - P(A \cap B|E)$.

There are also forms of Bayes' and the law of total probability with "extra conditioning" (see page 60 of the text).

2.5 Independence of events

Definition 2.9. Independence of two events
Events A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

If $P(A) > 0$ and $P(B) > 0$, then this is equivalent to $P(A|B) = P(A)$ or $P(B|A) = P(B)$.

Proposition 2.10.
If A and B are independent, then all the following are independent:

- A and B^c;
- A^c and B;
- A^c and B^c.

As a sidenote, independence is not transitive.

Proof. Let A and B be independent. We will first show that A and B^c are independent. If $P(A) = 0$, then A is independent of every event, including B^c. So assume $P(A) \neq 0$. Then

$$P(B^c|A) = 1 - P(B|A) = 1 - P(B) = P(B^c),$$

so A and B^c are independent. Swapping the roles of A and B, we have that A^c and B are independent. Using the fact that A, B independent implies A, B^c independent, with A^c playing the role of A, we also have that A^c and B^c are independent. □

Definition 2.11. Independence of many events
For n events A_i to be independent, we require pairwise independence, triple-wise independence, and similarly for all quadruples, quintuples, and so on.
This can quickly become unwieldy, but later we will discuss other ways to think about independence. For infinitely many events, we say that they are independent if every finite subset of the events is independent.

Remark 2.12. Pairwise independence alone does not imply independence
Consider an example with three events A, B, C. It is possible that just learning about A or just learning about B is of no use in predicting whether C occurred, but learning that both A and B occurred could still be highly relevant for C.

Definition 2.13. Conditional independence
Defined analogously: Events A and B are conditionally independent given E if

$$P(A \cap B|E) = P(A|E)P(B|E).$$

However, note that two events can be conditionally independent given E, but not independent, and vice versa. Two events independent given E also may not be independent given E^c. (For examples to build intuition, see page 65 of the text.)

2.6 Coherency of Bayes' rule

Theorem 2.14. Coherency of Bayes'
Bayes' rule is coherent. If we wish to update our probabilities with multiple pieces of information, we may either

- update sequentially, i.e., update with one piece of information at a time, and use that probability as a prior for the following update;
- update simultaneously, using all the evidence at once.

Example 2.15. Bayes' with two pieces of information
For example, suppose we want to calculate with Bayes' the probability

$$P(A \mid B \cap C).$$

- Updating sequentially, we may consider that we have first observed B, and then everything following will be conditioned on B.

$$P(A \mid B \cap C) = \frac{P(C \mid A, B)}{P(C \mid B)} P(A \mid B).$$

- Updating simultaneously, we simply have

$$P(A \mid B \cap C) = \frac{P(B \cap C \mid A)P(A)}{P(B \cap C)}.$$

2.7 Conditioning as a problem solving strategy

Example 2.16. Monty Hall
Consider the Monty Hall problem and assume WLOG that we begin by choosing door 1. Suppose we decide preemptively that we want to switch doors. Now let us calculate the unconditional and conditional probabilities of getting the car.
Define events

- W that we get the car;
- C_i that the car is behind door i;
- M_j that Monty opens door j to reveal a goat (for $j = 2, 3$).

Then we may calculate the unconditional probability by LOTP:

$$P(W) = P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) = 0 + \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

Supposing M_j happens for $j = 2, 3$, we may calculate the conditional probability using Bayes',

$$P(C_1|M_j) = \frac{P(M_j|C_1)P(C_1)}{P(M_j)} = \frac{(1/2)(1/3)}{1/2} = \frac{1}{3},$$

so then the conditional probability of winning is again $2/3$ if we switch.

Example 2.17. Gambler's ruin
Two gamblers, A and B, make a sequence of \$1 bets. In each bet, gambler A has probability p of winning, and gambler B has probability $q = 1 - p$ of winning. Gambler A starts with i dollars and gambler B starts with $N - i$ dollars.

We may visualize this game as a random walk which terminates with A's victory at position N or terminates with B's victory at position 0. To solve this problem, we may use states to derive what we call a difference equation, and then construct and solve the characteristic polynomial of the difference equation. Solving, we have that the probability of A winning with a starting wealth of i is

$$p_i = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N} & \text{if } p \neq 1/2 \\ \frac{i}{N} & \text{if } p = 1/2 \end{cases}.$$

See textbook Example 2.7.3 on page 72 for full details.

2.8 Pitfalls and paradoxes

Example 2.18. Simpson's paradox
This is a statistical phenomenon where an association between two variables in a population emerges, disappears, or reverses when the population is divided into subpopulations.

General form. Given events A, B, C, we say that we have a Simpson's paradox if

$$P(A|B,C) < P(A|B^c,C),$$
$$P(A|B,C^c) < P(A|B^c,C^c).$$

but

$$P(A|B) > P(A|B^c).$$

Explanation. The LOTP shows why this can happen:

$$P(A|B) = P(A|C,B)P(C|B) + P(A|C^c,B)P(C^c|B),$$
$$P(A|B^c) = P(A|C,B^c)P(C|B^c) + P(A|C^c,B^c)P(C^c|B^c).$$

The above equations express $P(A|B)$ as a weighted average of $P(A|C,B)$ and $P(A|C^c,B)$, and $P(A|B^c)$ as a weighted average of $P(A|C,B^c)$ and $P(A|C^c,B^c)$. If the corresponding weights were the same in both of these weighted averages, then Simpson's paradox could not occur. But the weights may be very different, e.g.,

$$P(C|B) < P(C|B^c) \text{ and } P(C^c|B) > P(C^c|B^c),$$

and if the differences in weights are sufficiently drastic, then Simpson's may occur.

Example 2.19. Example case of Simpson's
Case study. To build intuition, consider the classical example: we have two doctors, Dr. Hibbert and Dr. Nick, each performing two types of surgeries: heart surgery and Band-Aid removal. Let A be the event of a successful surgery, B be the event that Dr. Nick is the surgeon, and C be the event that the surgery is a heart surgery.

Simpson's paradox may arise here if the probability of a successful surgery is lower under Dr. Nick than under Dr. Hibbert whether we condition on heart surgery or on Band-Aid removal, but the overall probability of success is higher for Dr. Nick.

3 Random variables

3.1 Random variables

Definition 3.1. Random variable
Given an experiment with sample space S, a random variable (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$.

3.2 Distributions and probability mass functions

Definition 3.2. Discrete random variable, support
A random variable X is said to be discrete if there is a finite list of values $a_1, \ldots, a_n$ or an infinite list of values $a_1, a_2, \ldots$ such that $P(X = a_j \text{ for some } j) = 1$.

If X is a discrete r.v., then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the support of X.

Definition 3.3. Probability mass function
The probability mass function (PMF) of a discrete r.v. X is the function p_X given by $p_X(x) = P(X = x)$. Note that this is positive if x is in the support of X, and 0 otherwise.

3.3 Bernoulli and Binomial

Definition 3.4. Bernoulli distribution
An r.v. X is said to have the Bernoulli distribution with parameter p if

$$P(X = 1) = p,$$
$$P(X = 0) = 1 - p,$$

where $0 < p < 1$. We write this as

$$X \sim \text{Bern}(p).$$

Definition 3.5. Indicator random variable
The indicator random variable I_A or $I(A)$ of an event A is the r.v. which equals 1 if A occurs and 0 otherwise. Note that

$$I_A \sim \text{Bern}(p)$$

with $p = P(A)$.

Definition 3.6. Binomial distribution
Suppose that n independent Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes. The distribution of X is called the Binomial distribution with parameters n and p, denoted

$$X \sim \text{Bin}(n, p),$$

and its PMF is given by

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

for $k \in \{0, \ldots, n\}$ and $P(X = k) = 0$ otherwise. We may verify these probabilities sum to 1 with the Binomial Theorem.

3.4 Hypergeometric

Definition 3.7. Hypergeometric distribution
Consider an urn with w white balls and b black balls. We draw n balls out of the urn at random without replacement. Let X be the number of white balls in the sample. Then X is said to have the Hypergeometric distribution with parameters w, b, n, denoted

$$X \sim \text{HGeom}(w, b, n),$$

and its PMF is

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$, and $P(X = k) = 0$ otherwise. We may verify these probabilities sum to 1 with the Vandermonde's identity.

3.5 Discrete uniform

Definition 3.8. Discrete uniform distribution
Let C be a finite, nonempty set of numbers. Choose one of these numbers uniformly at random. Call the chosen number X. Then X is said to have the Discrete Uniform distribution with parameter C, denoted

$$X \sim \text{DUnif}(C),$$

and its PMF is given by

$$P(X = x) = \frac{1}{|C|}$$

for $x \in C$ and 0 otherwise.

3.6 Cumulative distribution functions

Definition 3.9. Cumulative distribution function, CDF
The CDF of an r.v. X is the function F_X given by $F_X(x) = P(X \leq x)$.

Theorem 3.10. Properties of valid CDFs
Any CDF F has the following properties:

- Increasing
- Right-continuous: CDFs are continuous except possibly for some jumps, which are still right continuous. Formally, for any a we have

$$F(a) = \lim_{x \to a^+} F(x).$$

- Convergence to 0 and 1 in the limits:

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$$

3.7 Functions of random variables

Definition 3.11. Function of an r.v. is an r.v.
For an experiment with sample space S, an r.v. X, and a function $g : \mathbb{R} \to \mathbb{R}$, $g(X)$ is the r.v. that maps s to $g(X(s))$ for all $s \in S$.

Theorem 3.12. PMF of $g(X)$
Let X be a discrete r.v. and $g : \mathbb{R} \to \mathbb{R}$. Then the support of $g(X)$ is the set of all y such that $g(x) = y$ for at least one x in the support of X, and the PMF of $g(X)$ is

$$P(g(X) = y) = \sum_{x : g(x) = y} P(X = x).$$

In the case that g is one-to-one, then we simply have

$$P\big(g(X) = g(x)\big) = P(X = x).$$

3.8 Independence of random variables

Definition 3.13. Independence of two r.v.s
Random variables X and Y are said to be independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),$$

for all $x, y \in \mathbb{R}$.

In the discrete case, this is equivalent to the condition

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

for all x, y with x in the support of X and y in the support of Y.

Definition 3.14. Independence of many r.v.s
Random variables $X_1, \ldots, X_n$ are independent if

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \ldots P(X_n \leq x_n),$$

for all $x_1, \ldots, x_n \in \mathbb{R}$. For infinitely many r.v.s, we say that they are independent if every finite subset of the r.v.s is independent.

To show some r.v.s $X_1, \ldots, X_n$ are not independent merely requires exhibiting values $x_1, \ldots, x_n$ in their respective supports such that

$$P(X_1 = x1, \ldots, X_n = x_n) = P(X_1 = x_1) \ldots P(X_n = x_n).$$

Theorem 3.15. Functions of independent r.v.s
If X and Y are independent r.v.s, then any function of X is independent of any function of Y.

Definition 3.16. Conditional independence of r.v.s
Random variables X and Y are conditionally independent given an r.v. Z if for all $x, y \in \mathbb{R}$ and all z in the support of Z,

$$P(X \leq x, Y \leq y \mid Z = z) = P(X \leq x \mid Z = z)P(Y \leq y \mid Z = z).$$

For discrete r.v.s, an equivalent definition is to require

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z).$$

As with independence of events, independence of r.v.s does not imply conditional independence, nor vice versa.

Definition 3.17. Conditional PMF
For any discrete r.v.s X and Z, the function $P(X = x \mid Z = z)$, when considered as a function of x for fixed z, is called the conditional PMF of X given $Z = z$.

3.9 Connections between binomial and hypergeometric

Theorem 3.18. Binomial to hypergeometric
If $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p)$, and X is independent of Y, then the conditional distribution of X given $X + Y = r$ is $\text{HGeom}(n, m, r)$.

Theorem 3.19. Hypergeometric to binomial
The binomial distribution is a limiting case of the hypergeometric. If $X \sim \text{HGeom}(w, b, n)$ and $N = w + b \to \infty$ such that $p = {}^{w}/_{(w+b)}$ remains fixed, then the PMF of X converges to the $\text{Bin}(n, p)$ PMF.

4 Expectation

4.1 Definition of expectation

Notation 4.1.
We often abbreviate $E(X)$ to EX, $E(X^2)$ to EX^2, and $E(X^n)$ to EX^n. Unless parentheses indicate otherwise, the exponent modifies the random variable and not the expected value, i.e., $E(X-1)^4$ is $E\big((X-1)^4\big)$, not $\big(E(X-1)\big)^4$.

4.2 Linearity of expectation

Remark 4.2. Linearity of expectation holds even for dependent r.v.s
As an example, consider the expected value of a r.v. following the hypergeometric distribution. Then the expected value of each of the n trials is $w/(w+b)$. Even though these trials are not independent, linearity still holds and thus the expected value of the random variable is $nw/(w+b)$.

Proposition 4.3. Monotonicity of expectation
Let X and Y be r.v.s such that $X \geq Y$ with probability 1. Then $E(X) \geq E(Y)$, with equality holding iff $X = Y$ with probability 1.

Proof. While this result holds for all r.v.s, here we will prove it only for discrete r.v.s. Define r.v. $Z = X - Y$ which is nonnegative (with probability 1), so $E(Z) \geq 0$ because $E(Z)$ is a sum of nonnegative terms. By linearity,

$$E(X) - E(Y) = E(X - Y) \geq 0,$$

as desired. If $E(X) = E(Y)$, then by linearity we also have $E(Z) = 0$, which implies that $P(X = Y) = P(Z = 0) = 1$ since if even one term in the sum defining $E(Z)$ is positive, then the whole sum is positive. □

4.3 Geometric and negative binomial

Definition 4.4. Geometric distribution
Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, with trials performed until a success occurs. Let X be the number of failures before the first successful trial. Then X has the Geometric distribution with parameter p, denoted

$$X \sim \text{Geom}(p),$$

and its PMF is

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

for $k \in \{0, \dots\}$, where $q = 1 - p$.

Theorem 4.5. Geometric CDF
If $X \sim \text{Geom}(p)$, then the CDF of X is

$$F(x) = \begin{cases} 1 - q^{\lfloor x \rfloor} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

where as usual $q = 1 - p$.
From the expectation of the First Success distribution (see below), we may easily derive $E(X)$ as follows:

$$E(X) = \frac{1}{p} - 1 = \frac{q}{p}.$$

Remark 4.6. Conventions for the Geometric
In this book, the Geometric distribution excludes the success, and the First Success distribution includes the success.

Definition 4.7. First Success distribution
The First Success distribution is the Geometric distribution except we now consider the first success in the number of trials, i.e.,

$$Y \sim \text{FS}(p) \implies Y - 1 \sim \text{Geom}(p)$$

and

$$P(Y = k) = q^{k-1}p.$$

We may derive $E(Y)$ as follows:

$$E(Y) = p \cdot 1 + E(X)(1 - p) \implies E(X) = \frac{1}{p}.$$

Definition 4.8. Negative Binomial distribution
In a sequence of independent Bernoulli trials with success probability p, if X is the number of failures before the r^{th} success, then X is said to have the Negative Binomial distribution with parameters r and p, denoted

$$X \sim \text{NBin}(r, p),$$

and its PMF is given by

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

for $n \in \{0, \dots, \}$.

Theorem 4.9. Negative Binomial r.v. is the sum of Geometric r.v.s
Let $X \sim \text{NBin}(r, p)$. Then

$$X = X_1 + \cdots + X_r,$$

where the X_i are i.i.d. Geom(p). Then

$$E(X) = \frac{rq}{p}$$

follows from linearity of expectation.

Proof. Let X_1 be the number of failures until the first success, X_2 be the number of failures between the first success and the second success, and in general X_i be the number of failures between the $(i-1)^{\text{st}}$ success and the i^{th} success.

Then clearly each X_i follows the geometric distribution, and we can sum them to get the total number of failures. □

Example 4.10. St. Petersburg paradox
Suppose a wealthy stranger offers to play the following game with you. You will flip a fair coin until it lands Heads for the first time, and you will receive \$2 if the game lasts for 1 round, \$4 if the game lasts for 2 rounds, \$8 if the game lasts for 3 rounds, and in general, \$2n if the game lasts for n rounds. What is the fair value of this game (the expected payoff)? How much would you be willing to pay to play this game once?

Lesson: This paradox illustrates the danger of confusing $E(g(X))$ with $g(E(X))$ when g is not linear.

Proof. Let X be your winnings from playing the game. By definition, $X = 2^N$ where N is the number of rounds that the game lasts. We have

$$E(X) = E(2^N) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \cdots = \infty,$$

even though $E(N) = 2$ and $2^{E(N)} = 4$.

The ∞ in this paradox is driven by an infinite "tail" of extremely rare events where you get extremely large payoffs. Cutting off this tail at some point, which makes sense in the real world, dramatically reduces the expected value of the game. □

4.4 Indicator r.v.s and the fundamental bridge

Theorem 4.11. Fundamental bridge between probability and expectation
There is a one-to-one correspondence between events and indicator r.v.s, and the probability of an event A is the expected value of its indicator r.v. I_A.

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

This allows us to express any probability as an expectation.

Remark 4.12. Strategy of decomposing r.v.s into indicators
We can often express a complicated discrete r.v. whose distribution we don't know as a sum of indicator r.v.s, which are extremely simple. The fundamental bridge lets us find the expectation of the indicators; then, using linearity, we obtain the expectation of our original r.v. This is the strategy we previously used to derive the expectations of the Binomial and Hypergeometric distributions.

4.5 Law of the unconscious statistician (LOTUS)

Theorem 4.13. LOTUS
If X is a discrete r.v. and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then

$$E(g(X)) = \sum_x g(x)P(X = x),$$

where the sum is taken over all possible values of X.

4.6 Variance

Definition 4.14. Variance and standard deviation
Given an r.v. X,

$$\text{Var}(X) = E(X - EX)^2 = E(X^2) - (EX)^2,$$
$$\text{SD}(X) = \sqrt{\text{Var}(X)}.$$

Proposition 4.15. Properties of variance
Given r.v. X and constant c, the following hold:

- $\text{Var}(X + c) = \text{Var}(X)$;
- $\text{Var}(cX) = c^2 \text{Var}(X)$;
- If X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ (but this does not hold if they are dependent!);
- $\text{Var}(X) \geq 0$, with equality iff $P(X = a) = 1$ for some a.

Remark 4.16. Strategy of decomposing r.v.s into indicators (again)
Once again, the strategy of decomposing discrete r.v.s into indicators is very useful for calculating variances, for it is easy to verify

$$I \sim \text{Bern}(p) \implies \text{Var}(I) = p(1-p).$$

For example, we might decompose a Binomial r.v. into a sum of n indicators to see that the Binomial distribution has variance $np(1-p)$.

4.7 Poisson

Definition 4.17. Poisson distribution
Given parameter $\lambda > 0$, we denote

$$X \sim \text{Pois}(\lambda)$$

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

for $k \in \{0, 1, \dots\}$. This is a valid PMF because

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}.$$

Deriving $E(X)$ is easy; deriving $\text{Var}(X)$ is requires some manipulation:

$$E(X) = \lambda,$$
$$\text{Var}(X) = \lambda.$$

The Poisson distribution is often used in situations where we are counting the number of successes in a particular region or interval of time, and there are a large number of trials, each with a small probability of success. For example, the following random variables could follow a distribution that is approximately Poisson. Then the parameter λ is interpreted as the rate of occurrence of these rare events, i.e., we would expect λ occurrences in this interval.

Definition 4.18. Poisson paradigm
Let $A_1, \ldots, A_n$ be events with $p_j = P(A_j)$, where n is large, the p_j are small, and the A_j are independent or weakly dependent. Let

$$X = \sum_{j=1}^{n} I(A_j),$$

then X is approximately distributed as Pois(λ) for $\lambda = \sum_j p_j$.

Of course, the number of events that might occur isn't exactly Poisson because a Poisson r.v. has no upper bound while at most n of $A_1, \ldots, A_n$ might occur. But the Poisson distribution often gives good approximations, and the conditions for the Poisson paradigm are flexible:

- the n trials can have different success probabilities;
- the trials don't have to be independent, though they should not be very dependent.

4.8 Connections between Poisson and Binomial

Theorem 4.19. Sum of independent Poissons
If $X \sim \text{Pois}(\lambda_1), Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then

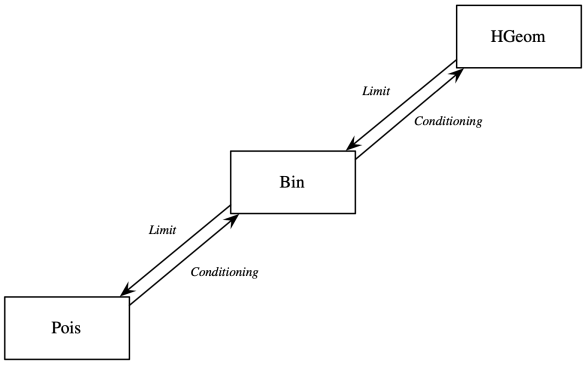
$$X + Y \sim \text{Pois}(\lambda_1 + \lambda_2).$$

See 4.8.1 of the text for the proof.

Theorem 4.20. Poisson approximation to Binomial
If $X \sim \text{Bin}(n, p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains fixed, then the PMF of X converges to the Pois(λ) PMF. More generally, the same conclusion holds if $n \to \infty$ and $p \to 0$ in such a way that np converges to a constant λ.

See 4.8.3 of the text for proof.

Remark 4.21. Relationship between Binomial, Hypergeometric, and Poisson See the figure below. In the rest of the book, we'll continue to introduce new named distributions and add them to this family tree, until everything is connected!



4.9 Using probability and expectation to prove existence

An interesting section, but less practical and so skipped for now.

5 Continuous random variables

5.1 Probability density functions

Definition 5.1. Continuous r.v.
An r.v. has a continuous distribution if its CDF is differentiable. We also allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else. A continuous random variable is a random variable with a continuous distribution.

Definition 5.2. Probability density function
For a continuous r.v. X with CDF F, the probability density function (PDF) of X is the derivative f of the CDF, given by $f(x) = F'(x)$. The support of X, and of its distribution, is the set of all x where $f(x) > 0$.

Proposition 5.3. PDF to CDF
Let X be a continuous r.v. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

More generally, to get a desired probability, integrate the PDF over the appropriate range.

Definition 5.4. Expectation of a continuous r.v.
The expected value of a continuous r.v. X with PDF f is

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

Theorem 5.5. LOTUS, continuous
If X is a continuous r.v. with PDF f and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then

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

5.2 Uniform

Definition 5.6. Uniform distribution
A continuous r.v. U is said to have the Uniform distribution on the interval (a, b),

$$U \sim \text{Unif}(a, b),$$

if its PDF is

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

We may derive

$$E(U) = \frac{a+b}{2}$$
$$\text{Var}(U) = \frac{(b-a)^2}{12}.$$

Proposition 5.7.
Let $U \sim \text{Unif}(a, b)$, and let (c, d) be a subinterval of (a, b). Then the conditional distribution of U given $U \in (c, d)$ is $\text{Unif}(c, d)$.

Definition 5.8. Location-scale transformation
Let X be an r.v. and $Y = \sigma X + \mu$ for constants $\sigma > 0$ and μ. Then we say that Y has been obtained as a location-scale transformation of X.

In a location-scale transformation, Y is a linear function of X and so Uniformity is preserved. But a nonlinear transformation of X will not be Uniform in general.

Remark 5.9. Deriving expectation and variance with location-scale transformations
Suppose $U \sim (0, 1)$, then we may derive $E(U) = \frac{1}{2}$ and $\text{Var}(U) = \frac{1}{12}$.

Then to generalize to any Uniform distribution, we may scale the expectation of U by σ and add μ, while we scale the variance of U by σ^2.

5.3 Universality of the uniform

Theorem 5.10. Universality of the Uniform
Let F be a CDF which is a continuous function and strictly increasing on the support of the distribution. This ensures that the inverse function F^{-1} exists, as a function from $(0, 1)$ to $\mathbb{R}$. We then have:

- Let $U \sim \text{Unif}(0, 1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.
- Let X be an r.v. with CDF F. Then $F(X) \sim \text{Unif}(0, 1)$.

Definition 5.11. Survival function
The survival function of an r.v. X with CDF F is the function G given by $G(x) = 1 - F(x) = P(X > x)$.

Theorem 5.12. Expectation by integrating the survival function
Given nonnegative r.v. X we have

$$E(X) = \sum_0^\infty P(X > x)dx.$$

Note that this result holds for any nonnegative r.v., not just for continuous nonnegative r.v.s.

5.4 Normal

Definition 5.13. Normal distribution
A continuous r.v. Z is said to have the standard Normal distribution,

$$Z \sim \mathcal{N}(0, 1),$$

if its PDF φ is given by

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

Its CDF is given by

$$\Phi(z) = \int_{-\infty}^{z} \varphi(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}dt.$$

Unfortunately, it is impossible to find a closed-form expression for the antiderivative of φ.

Note that in general, $X = \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$ and has PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Proposition 5.14. Properties of the normal
Given $Z \sim \mathcal{N}(0, 1)$, we have:

- Symmetry of PDF: φ is even,
$$\varphi(z) = \varphi(-z)$$
- Symmetry of tails: $\varphi(z) = 1 - \varphi(-z)$
- Symmetry of Z and $-Z$:
$$-Z \sim \mathcal{N}(0, 1).$$

Remark 5.15. Sum of independent normally distributed r.v.s is normally distributed
Though we will not prove it here, note that the sum of independent normally distributed r.v.s is normally distributed.

5.5 Exponential

Definition 5.16. Exponential distribution
Given continuous r.v. X, we have

$$X \sim \text{Expo}(\lambda)$$

with parameter $\lambda > 0$ if its PDF is

$$f(x) = \lambda e^{-\lambda x}, \quad x \in (0, \infty)$$

and corresponding CDF is

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

Of course, given $X \sim \text{Expo}(1)$, we may obtain $Y = \text{Expo}(\lambda) = X/\lambda$ by scaling. And integration by parts gives

$$E(X) = 1, \quad \text{Var}(X) = 1.$$

As is easily verified, Exponential distribution has the memoryless property described in the following theorem, implying

$$E(X \mid X \geq s) = s + E(X) = s + \frac{1}{\lambda}.$$

Also note that the Geometric converges to the Exponential where the Bernoulli trials are performed faster and faster but with smaller and smaller success probabilities.

Theorem 5.17. Memoryless property
A continuous distribution is said to have the memoryless property if a random variable X from that distribution satisfies for all $s, t \geq 0$,

$$P(X \geq s + t \mid X \geq s) = P(X \geq t).$$

Intuitively, supposing your wait time follows a memoryless distribution, then even if you have waited for hours or days without success, the success is not any more likely to arrive soon.

If X is a positive continuous random variable with the memoryless property, then X has an Exponential distribution. See Theorem 5.5.3 (page 241) for the proof.

Proposition 5.18. Comparing Exponentials of different rates
If $T_1 \sim \text{Expo}(\lambda_1)$ and $T_2 \sim \text{Expo}(\lambda_2)$, then

$$P(T_1 < T_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

5.6 Poisson processes

Definition 5.19. Poisson process
A process of arrivals in continuous time, here considering the interval $(0, \infty)$, is called a Poisson process with rate λ if:

1. The number of arrivals that occur in an interval of length t is a $\text{Pois}(\lambda t)$ random variable.
2. The numbers of arrivals that occur in disjoint intervals are independent of each other.

Observe count-time duality: Given discrete r.v. N_t counting the number of arrivals and continuous r.v. T_n marking the time of the n^{th} arrival, then $T_n > t$ and $N_t < n$ are the same event.

It is easy to see that $P(T_1 > t) = e^{-\lambda t}$ and so $T_1 \sim \text{Expo}(\lambda)$. Having that the intervals $(0, T_1), (T_1, T_2), \ldots$ are disjoint and thus independent by construction, we may apply the memoryless property to have

$$\begin{aligned} T_1 &\sim \text{Expo}(\lambda) \\ T_2 - T_1 &\sim \text{Expo}(\lambda) \\ &\vdots \end{aligned}$$

Example 5.20. Minimum of independent Expos
Let $X_1, \ldots, X_n$ be independent, with $X_j \sim \text{Expo}(\lambda_j)$. Let $L = \min(X_1, \ldots, X_n)$. Then

$$L \sim \text{Expo}(\lambda_1 + \cdots + \lambda_n).$$

Proof. Simply derive the CDF. Intuitively, we can interpret the λ_j as the rates of n independent Poisson processes. Then L is the waiting time for the event to occur in any of the processes, so it makes sense that L has a combined rate of $\lambda_1 + \cdots + \lambda_n$. □

Example 5.21. Maximum of 3 independent Exponentials
Three students are working independently on their probability homework. All 3 start at 1 pm on a certain day, and each takes an Exponential time with mean 6 hours to complete the homework. What is the earliest time at which all 3 students will have completed the homework, on average?

As a more general example, see homework 6 problem 7.

Proof. Solution: Label the students as $1, 2, 3$, and let X_j be how long it takes student j to finish the homework. Let $\lambda = 1/6$, and let T be the time when all 3 students will have completed the homework, so $T = \max(X_1, X_2, X_3)$ with $X_i \sim \text{Expo}(\lambda)$.

Finding $E(T)$ by integrating $tf_T(t)$ is possible but not especially pleasant. Instead, we may use the memoryless property and the fact that the minimum of independent Exponentials is Exponential. We can decompose

$$T = T_1 + T_2 + T_3$$

where $T_1 = \min(X_1, X_2, X_3)$ is how long it takes for one student to complete the homework, T_2 is the additional time it takes for a second student to complete the homework, and T_3 is the additional time until all 3 have completed the homework. Then $T_1 \sim \text{Expo}(3\lambda)$, by the previous result.

By the memoryless property, at the first time when a student completes the homework the other two students are starting from fresh, so $T_2 \sim \text{Expo}(2\lambda)$. Again by the memoryless property, $T_3 \sim \text{Expo}(\lambda)$. The memoryless property also implies that T_1, T_2, T_3 are independent (which would be very useful if we were finding $\text{Var}(T)$). By linearity,

$$E(T) = \frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda} = 2 + 3 + 6 = 11,$$

which shows that on average, the 3 students will have all completed the homework at midnight, 11 hours after they started. □

5.7 Symmetry of i.i.d. continuous r.v.s

Proposition 5.22.
Let $X_1, \ldots, X_n$ be i.i.d. from a continuous distribution. Then

$$P(X_{a_1} < X_{a_2} < \cdots < X_{a_n}) = \frac{1}{n!}$$

for any permutation $a_1, a_2, \ldots, a_n$ of $1, 2, \ldots, n$.

Proof. Follows easily from

$$P(X_i = X_j) = 0$$

for $i \neq j$, which holds from the definition of independent continuous r.v.s. □

6 Moments

6.1 Summaries of a distribution

Definition 6.1. Median
Define c as a median of an r.v. X if

$$P(X \leq c) \geq \frac{1}{2} \text{ and } P(X \geq c) \geq \frac{1}{2}.$$

Of course the simplest way this can happen is if the CDF of X is $1/2$ exactly at c, but some CDFs have jumps.

Definition 6.2. Mode
The modes of a discrete r.v. are the values maximizing the PMF,

$$P(X = c) \geq P(X = x)$$

for all x. Similarly, the modes of a continuous r.v. X are the values maximizing the PDF

$$f(c) \geq f(x)$$

for all x.

Remark 6.3. Understanding measures of center intuitively
Considering the idea of mass,

- the mean is the center of mass;
- the median is the value such that half the mass of the distribution falls on either side of c (or as close to half as possible for discrete r.v.s);
- the mode is a value that has the greatest mass or density out of all values in the support.

Proposition 6.4.
Let X be an r.v. with mean μ and m a median of X. Then

- the mean squared error $E(X-c)^2$ is minimized by $c = \mu$,
- the mean absolute error $E|X-c|$ is minimized by $c = m$.

Proof. We will first prove a useful identity:

$$E(X-c)^2 = \text{Var}(X) + (\mu - c)^2.$$

To see this, observe that

$$\text{Var}(X) = \text{Var}(X-c) = E(X-c)^2 - (E(X-c))^2 = E(X-c)^2 - (\mu - c)^2.$$

Then it follows easily that $c = \mu$ is the unique choice that minimizes $E(X-c)^2$.

Next, let's consider the mean absolute error. Let $a \neq m$. We need to show that $E|X-m| \leq E|X-a|$, which is equivalent to $E(|X-a|-|X-m|) \geq 0$. Assume that $m < a$ (the case $m > a$ can be handled similarly). If $X \leq m$ then

$$|X-a|-|X-m| = a - X - (m-X) = a-m,$$

and if $X > m$ then

$$|X-a|-|X-m| \geq X - a - (X-m) = m-a.$$

Let

$$Y = |X-a|-|X-m|.$$

We can split the definition of $E(Y)$ into 2 parts based on whether $X \leq m$ occurs, using indicator r.v.s. Let I be the indicator r.v. for $X \leq m$, so $1-I$ is the indicator r.v. for $X > m$. Then

$$\begin{aligned} E(Y) &= E(YI) + E(Y(1-I)) \\ &\geq (a-m)E(I) + (m-a)E(1-I) \\ &= (a-m)P(X \leq m) + (m-a)P(X > m) \\ &= (a-m)P(X \leq m) - (a-m)(1-P(X \leq m)) \\ &= (a-m)(2P(X \leq m) - 1). \end{aligned}$$

By definition of median, we have $2P(X \leq m) - 1 \geq 0$. Thus, $E(Y) \geq 0$, which implies $E(|X-m|) \leq E(|X-a|)$. □

6.2 Interpreting moments

Definition 6.5. Moments
Suppose r.v. X has mean μ and variance σ^2. For any positive integer n, define (if it exists) the

- n^{th} moment $E\left(X^n\right)$,
- n^{th} central moment $E\left((X-\mu)^n\right)$,
- n^{th} standardized moment $E\left(\left(\frac{X-\mu}{\sigma}\right)^n\right)$.

Definition 6.6. Skewness
The skewness of X is the third standardized moment of X

$$\text{Skew}(X) = E\left(\frac{X-\mu}{\sigma}\right)^3.$$

We standardize first so that $\text{Skew}(X)$ does not depend on the location or the scale of X, which are already described by μ and σ.

Definition 6.7. Symmetry of an r.v.
We say X is symmetric about μ if $X - \mu$ has the same distribution as $\mu - X$. It is easy to verify that μ must equal the mean and median should it exist.

Equivalently, X is symmetric about μ iff for all x

$$f(x) = f(2\mu - x).$$

Proposition 6.8. Odd central moments of a symmetric distribution
If X is symmetric, then for any odd m, the m^{th} central moment $E((X-\mu)^m)$ is 0 if it exists.

Thus we may use odd central moments to measure skewness, and we choose to use the third central moment.

Proof. Since $X - \mu$ has the same distribution as $\mu - X$, they must have the same m^{th} moment (if it exists). Then we have

$$E((X-\mu)^m) = E((\mu - X)^m),$$

but of course also

$$(X-\mu)^m = -(\mu - X)^m,$$

and so

$$E((X-\mu)^m) = E((\mu - X)^m) = 0,$$

as desired. □

Definition 6.9. Kurtosis
The kurtosis X is a shifted version of the fourth standardized moment:

$$\text{Kurt}(X) = E\left(\frac{X-\mu}{\sigma}\right)^4 - 3$$

Kurtosis measures tailedness of a distribution relative to the normal distribution (we subtract 3 such that any normal distribution has kurtosis 0).

6.3 Sample moments

Definition 6.10. Sample moments
Let $X_1, \ldots, X_n$ be i.i.d. random variables. The k^{th} sample moment is the r.v.

$$M_k = \frac{1}{n}\sum_{j=1}^{n} X_j^k.$$

The law of large numbers (chapter 10) shows that the k^{th} sample moment converges to the k^{th} moment $E\left(X_1^k\right)$ as $n \to \infty$.
Also, the expected value of the k^{th} sample moment is the k^{th} moment, ie., it is an unbiased estimator.

Proposition 6.11. Mean and variance of sample mean
Let $X_1, \ldots, X_n$ be i.i.d. r.v.s. Then the sample mean $\bar{X}_n$ is unbiased for estimating μ,

$$E\left(\bar{X}_n\right) = \mu,$$

and we have by additivity of variances of independent r.v.s that

$$\text{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}.$$

Definition 6.12. Sample variance and sample standard deviation
Let $X_1, \ldots, X_n$ be i.i.d. random variables. The sample variance is the r.v.

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}_n\right)^2,$$

and of course the sample standard deviation is the square root of the sample variance.

This mimics the usual variance definition by averaging the squared distances of the X_j from the sample mean, except dividing by $n-1$ rather than n such that S_n^2 is unbiased for estimating σ^2.

6.4 Moment generating functions

Definition 6.13. Moment generating function
The MGF $M(t)$ of X is

$$M(t) = E\left(e^{tX}\right)$$

if this is finite on some open interval $(-a, a)$ containing 0. Otherwise we say the MGF of X does not exist.

Note that t has no interpretation in particular, but is simply a bookkeeping device introduced to induce differentiability instead of working with a discrete sequence of moments. Also note that as a quick validity check, we must have $M(0) = 1$ for any valid MGF M.

Definition 6.14. Moments via derivatives of the MGF
Given the MGF of X, we can get the n^{th} moment of X by evaluating the n^{th} derivative of the MGF at 0:

$$E\left(X^n\right) = M^{(n)}(0).$$

Proof. Proof. This can be seen by noting that the Taylor expansion of $M(t)$ about 0 is

$$M(t) = \sum_{n=0}^{\infty} M^{(n)}(0)\frac{t^n}{n!}$$

while on the other hand, we also have

$$M(t) = E\left(e^{tX}\right) = E\left(\sum_{n=0}^{\infty} X^n \frac{t^n}{n!}\right).$$

We are allowed to interchange the expectation and the infinite sum because certain technical conditions are satisfied (this is where we invoke the condition that $E\left(e^{tX}\right)$ is finite in an interval around 0), so

$$M(t) = \sum_{n=0}^{\infty} E\left(X^n\right) \frac{t^n}{n!}$$

Matching the coefficients of the two expansions, we get $E\left(X^n\right) = M^{(n)}(0)$. □

Proposition 6.15. MGF determines the distribution
The MGF of an r.v. determines its distribution: if two r.v.s have the same MGF, they must have the same distribution. In fact, if there is even a tiny interval $(-a, a)$ containing 0 on which the MGFs are equal, then the r.v.s must have the same distribution.

Proposition 6.16. MGF of a sum of independent r.v.s
If X and Y are independent, then the MGF of $X + Y$ is the product of the individual MGFs:

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

This is true because if X and Y are independent, then $E\left(e^{t(X+Y)}\right) = E\left(e^{tX}\right) E\left(e^{tY}\right)$.

Remark 6.17. Not all r.v.s have an MGF
Some r.v.s X don't even have $E(X)$ exist, or don't have $E\left(X^n\right)$ exist for some $n > 1$, in which case the MGF clearly will not exist. But even if all the moments of X exist, the MGF may not exist if the moments grow too quickly. Luckily, there is a way to fix this: inserting an imaginary number! The function $\psi(t) = E\left(e^{itX}\right)$ with $i = \sqrt{-1}$ is called the characteristic function by statisticians and the Fourier transform by everyone else. It turns out that the characteristic function always exists. In this book we will focus on the MGF rather than the characteristic function, to avoid having to handle imaginary numbers.

Proposition 6.18. MGF of location-scale transformation
If X has MGF $M(t)$, then the MGF of $a + bX$ is

$$E(e^{t(a+bX)}) = e^{at} E\left(e^{btX}\right) = e^{at} M(bt).$$

6.5 Generating moments with MGFs

We now give some examples of using an MGF to find moments. We saw above that we can get moments by differentiating the MGF and evaluating at 0, rather than doing a complicated sum or integral by LOTUS. Better yet, in some cases we can simultaneously find all the moments of a distribution via a Taylor expansion, rather than repeated differentiation.

Example 6.19. Deriving moments of exponential distribution
Let $X \sim \text{Expo}(1)$. The MGF of X is $M(t) = 1/(1-t)$ for $t < 1$. Let us derive the moments of X via a Taylor expansion.

Recognize $1/(1-t)$ as a geometric series, valid in an interval around 0. For $|t| < 1$,

$$M(t) = \frac{1}{1-t} = \sum_{n=0}^{\infty} t^n = \sum_{n=0}^{\infty} n! \frac{t^n}{n!}.$$

On the other hand, we know that $E(X^n)$ is the coefficient of the term involving t^n in the Taylor expansion of $M(t)$:

$$M(t) = \sum_{n=0}^{\infty} E(X^n) \frac{t^n}{n!}.$$

Thus we can match coefficients to conclude that $E(X^n) = n!$ for all n. Of course, to find the moments of $Y \sim \text{Expo}(\lambda)$, use a scale transformation: we can express

$$E(Y^n) = \frac{n!}{\lambda^n}.$$

6.6 Sums of independent r.v.s via MGFs

Example 6.20. Sum of independent Poissons
Using MGFs, we can easily show that the sum of independent Poissons is Poisson. First let's find the MGF of $X \sim \text{Pois}(\lambda)$:

$$E\left(e^{tX}\right) = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(\lambda e^t\right)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda\left(e^t - 1\right)}.$$

Now let $Y \sim \text{Pois}(\mu)$ be independent of X. The MGF of $X + Y$ is

$$E\left(e^{tX}\right) E\left(e^{tY}\right) = e^{\lambda\left(e^t-1\right)} e^{\mu\left(e^t-1\right)} = e^{(\lambda+\mu)\left(e^t-1\right)},$$

which is the Pois $(\lambda + \mu)$ MGF. Since the MGF determines the distribution, we have proven that $X + Y \sim \text{Pois}(\lambda + \mu)$. Contrast this with the proof from Chapter 4 (Theorem 4.8.1), which required using the law of total probability and summing over all possible values of X. The proof using MGFs is far less tedious.

6.7 Probability generating functions

7 Joint distributions

The table below gives an overview of chapter 7:

	Two discrete r.v.s	**Two continuous r.v.s**
Joint CDF	$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$	$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$
Joint PMF/PDF	$P(X = x, Y = y)$ • Joint PMF is nonnegative. • Joint PMF sums to 1. • $P((X,Y) \in A) = \sum\sum_{(x,y)\in A} P(X = x, Y = y)$.	$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$ • Joint PDF is nonnegative. • Joint PDF integrates to 1. • $P((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy$.
Marginal PMF/PDF	$P(X = x) = \sum_y P(X = x, Y = y)$ $= \sum_y P(X = x \mid Y = y)P(Y = y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ $= \int_{-\infty}^{\infty} f_{X\mid Y}(x\mid y) f_Y(y)dy$
Conditional PMF/PDF	$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$ $= \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$	$f_{Y\mid X}(y\mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ $= \frac{f_{X\mid Y}(x\mid y) f_Y(y)}{f_X(x)}$
Independence	$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x and y.	$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all x and y.
LOTUS	$E(g(X,Y)) = \sum_y \sum_x g(x,y)P(X = x, Y = y)$	$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy$

7.1 Joint, marginal, and conditional

7.1.1 Discrete

Definition 7.1. Joint PMF
The joint PMF of discrete r.v.s X and Y is the function $p_{X,Y}$ given by

$$p_{X,Y}(x,y) = P(X = x, Y = y).$$

The joint PMF of n discrete r.v.s is defined analogously, as is the joint CDF (which is determined by the joint PMF).

Definition 7.2. Marginal PMF
For discrete r.v.s X and Y, the marginal PMF of X is

$$P(X = x) = \sum_y P(X = x, Y = y).$$

The marginal PMF of X is the PMF of X, viewing X individually rather than jointly with Y. The above equation follows from the axioms of probability (we are summing over disjoint cases). The operation of summing over the possible values of Y in order to convert the joint PMF into the marginal PMF of X is known as marginalizing out Y.

Definition 7.3. Independence of discrete r.v.s
Random variables X and Y are independent if for all x and y,

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

If X and Y are discrete, this is equivalent to the condition

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

and it is also equivalent to the condition

$$P(Y = y \mid X = x) = P(Y = y)$$

for all x, y such that $P(X = x) > 0$.

Example 7.4. Chicken-egg
Suppose a chicken lays a random number of eggs, N, where $N \sim \text{Pois}(\lambda)$. Each egg independently hatches with probability p and fails to hatch with probability $q = 1 - p$. Let X be the number of eggs that hatch and Y the number that do not hatch, so $X + Y = N$. What is the joint PMF of X and Y?

Solution. We seek the joint PMF $P(X = i, Y = j)$ for nonnegative integers i, j. Conditional on the total number of eggs N, the eggs are independent Bernoulli trials with probability of success p, so we have $(X|N = n) \sim \text{Bin}(n, p)$ and $(Y|N = n) \sim \text{Bin}(n, q)$. Thus we may derive

$$\begin{aligned} P(X = i, Y = j) &= \sum_{n=0}^{\infty} P(X = i, Y = j|N = n)P(N = n) \\ &= P(X = i, Y = j|N = i + j)P(N = i + j) \\ &= P(X = i|N = i + j)P(N = i + j) \\ &= \binom{i+j}{i} \frac{e^{-\lambda}\lambda^{i+j}}{(i+j)!} \\ &= \frac{e^{-\lambda p}(\lambda p)^i}{i!} \cdot \frac{e^{-\lambda q}(\lambda q)^j}{j!}. \end{aligned}$$

where the third equality follows from the fact that conditional of $N = i + j$, $X = i$ and $Y = j$ are the same event.

The final equality reveals that the joint PMF is a product of the Pois(λp) PMF (as a function of i) and the Pois(λq) PMF (as a function of j). This tells us two elegant facts: (1) X and Y are independent, since their joint PMF is the product of their marginal PMFs, and (2) $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$.

At first it may seem deeply counterintuitive that X is independent of Y. Doesn't knowing that a lot of eggs hatched mean that there are probably not so many that didn't hatch? For a fixed number of eggs, this independence would be impossible: knowing the number of hatched eggs would perfectly determine the number of unhatched eggs. But in this example, the number of eggs is random, following a Poisson distribution, and this happens to be the right kind of randomness to make X and Y unconditionally independent. This is a very special property of the Poisson.

Theorem 7.5.
If $X \sim \text{Pois}(\lambda p), Y \sim \text{Pois}(\lambda q)$, and X and Y are independent, then $N = X + Y \sim \text{Pois}(\lambda)$ and $(X|N = n) \sim \text{Bin}(n, p)$.

This is a result from chapter 4. The chicken-egg story above now gives us the converse in the following theorem.

Theorem 7.6.
If $N \sim \text{Pois}(\lambda)$ and $(X|N = n) \sim \text{Bin}(n, p)$, then $X \sim \text{Pois}(\lambda p), Y = N - X \sim \text{Pois}(\lambda q)$, and X and Y are independent.

7.1.2 Continuous

Definition 7.7. Joint PDF
Given X and Y continuous, define the joint PDF

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

We integrate the joint PDF to get the probability of a two-dimensional region, i.e., given a region $A \subseteq \mathbb{R}^2$,

$$P((X, Y) \in A) = \iint_A f_{X,Y}(x, y) dx dy.$$

Valid joint PDFs are nonnegative and integrate to 1.

Definition 7.8. Marginal PDF
For continuous r.v.s X and Y with joint PDF $f_{X,Y}$, the marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

This is the PDF of X, viewing X individually rather than jointly with Y.

Definition 7.9. Conditional PDF
For continuous r.v.s X and Y with joint PDF $f_{X,Y}$, the conditional PDF of Y given $X = x$ is

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

for all x with $f_X(x) > 0$. This is considered as a function of y for fixed x. As a convention, in order to make $f_{Y|X}(y \mid x)$ well-defined for all real x, let $f_{Y|X}(y \mid x) = 0$ for all x with $f_X(x) = 0$.

Theorem 7.10. Continuous Bayes' and LOTP
For continuous r.v.s X and Y, we have Bayes'

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)}, \text{ for } f_X(x) > 0,$$

and LOTP

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy.$$

Theorem 7.11. Variations of Bayes' and LOTP
Here are the four versions of Bayes' rule, summarized in a table.

	Y discrete	Y continuous
X discrete	$P(Y = y \mid X = x) = \frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y \mid X = x) = \frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X continuous	$P(Y = y \mid X = x) = \frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y \mid x) = \frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

And here are the four versions of LOTP, summarized in a table. The top row gives expressions for $P(X = x)$, while the bottom row gives expressions for $f_X(x)$.

	Y discrete	Y continuous
X discrete	$\sum_y P(X = x \mid Y = y)P(Y = y)$	$\int_{-\infty}^{\infty} P(X = x \mid Y = y) f_Y(y) dy$
X continuous	$\sum_y f_X(x \mid Y = y)P(Y = y)$	$\int_{-\infty}^{\infty} f_{X\|Y}(x \mid y) f_Y(y) dy$

Definition 7.12. Independence of continuous r.v.s
Random variables X and Y are independent if for all x and y,

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

This is equivalent to the condition

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

for all x, y, and it is also equivalent to the condition

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

for all x, y such that $f_X(x) > 0$.

Proposition 7.13. Factoring the joint PDF
Suppose that the joint PDF $f_{X,Y}$ of X and Y factors as

$$f_{X,Y}(x,y) = g(x)h(y)$$

for all x and y, where g and h are nonnegative functions. Then X and Y are independent. Also, if either g or h is a valid PDF, then the other one is a valid PDF too and g and h are the marginal PDFs of X and Y, respectively. (The analogous result in the discrete case also holds.)

7.2 2D LOTUS

Theorem 7.14. 2D LOTUS
Let g be a function from $\mathbb{R}^2$ to $\mathbb{R}$. If X and Y are discrete, then

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

If X and Y are continuous with joint PDF $f_{X,Y}$, then

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy.$$

7.3 Covariance and correlation

Definition 7.15. Covariance
The covariance between r.v.s X and Y is

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\big((X-EX)(Y-EY)\big)\\ &= E(XY)-E(X)E(Y)\end{aligned}$$

Intuitively, if X and Y tend to move in the same direction, then $X-EX$ and $Y-EY$ will tend to be either both positive or both negative, giving positive covariance.

Covariance is a measure of linear association; note that r.v.s can be dependent in nonlinear ways and still have zero covariance.

Theorem 7.16. Independent r.v.s are uncorrelated
If X and Y are independent, then they are uncorrelated, i.e., $\mathrm{Cov}(X,Y)=0$. (But note the converse is false, for as described above, covariance is a measure only of linear association.)

Proposition 7.17. Properties of covariance
Covariance has the following key properties.

- $\mathrm{Cov}(X,X)=\mathrm{Var}(X)$.
- $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$.
- $\mathrm{Cov}(X,c)=0$ for any constant c.
- $\mathrm{Cov}(aX,Y)=a\,\mathrm{Cov}(X,Y)$ for any constant a.
- $\mathrm{Cov}(X+Y,Z)=\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)$.
- $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$. For n r.v.s $X_1,\dots,X_n$,

$$\mathrm{Var}\left(X_1+\cdots+X_n\right)=\mathrm{Var}\left(X_1\right)+\cdots+\mathrm{Var}\left(X_n\right)+2\sum_{i<j}\mathrm{Cov}\left(X_i,X_j\right).$$

Definition 7.18. Correlation
Define the correlation

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

undefined in the degenerate cases $\mathrm{Var}(X)=0$ or $\mathrm{Var}(Y)=0$.

The correlation is invariant under shifting and scaling of X and Y, does not depend on the units of measurement, and is always between -1 and 1.

7.4 Multinomial

Definition 7.19. Multinomial distribution
Consider if each of n objects is independently placed into one of k categories, with probability p_j that a given object is placed into category j, such that $\sum_{j=1}^{k} p_j = 1$. Let X_i be the number of objects in category i, so that $X_1 + \cdots + X_k = n$. Then we say

$$\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$$

for $\mathbf{X} = (X_1, \ldots, X_k)$ and $\mathbf{p} = (p_1, \ldots, p_k)$.
It is easy to see that the joint PMF of $\mathbf{X}$ is

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \cdot p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

for $n_1, \ldots, n_k$ satisfying $n_1 + \cdots + n_k = n$.

Theorem 7.20. Multinomial marginals, multinomial lumping
The marginals of a Multinomial are Binomial. Specifically, if $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then $X_j \sim \text{Bin}(n, p_j)$.

More generally, whenever we merge multiple categories together in a Multinomial random vector, we get another Multinomial random vector. Specifically, for any distinct i and j, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$, and the random vector of counts obtained from merging categories is still Multinomial, e.g.,

$$(X_1 + X_2, X_3, \ldots, X_k) \sim \text{Mult}_{k-1}(n, (p_1 + p_2, p_3, \ldots, p_k)).$$

Theorem 7.21. Multinomial conditioning
If $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then

$$(X_2, \ldots, X_k) \mid X_1 = n_1 \sim \text{Mult}_{k-1}\left(n - n_1, \left(p_2', \ldots, p_k'\right)\right)$$

where $p_j' = p_j / (p_2 + \cdots + p_k)$.

Theorem 7.22. Covariance in a Multinomial
Let $(X_1, \ldots, X_k) \sim \text{Mult}_k(n, \mathbf{p})$, where $\mathbf{p} = (p_1, \ldots, p_k)$. For $i \neq j$, $\text{Cov}(X_i, X_j) = -np_ip_j$.

Proof. For concreteness, let $i = 1$ and $j = 2$. Using the lumping property and the marginal distributions of a Multinomial, we know $X_1 + X_2 \sim \text{Bin}(n, p_1 + p_2)$, $X_1 \sim \text{Bin}(n, p_1)$, and $X_2 \sim \text{Bin}(n, p_2)$. Therefore

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\,\text{Cov}(X_1, X_2)$$

becomes

$$n(p_1 + p_2)(1 - (p_1 + p_2)) = np_1(1 - p_1) + np_2(1 - p_2) + 2\,\text{Cov}(X_1, X_2).$$

Solving for $\text{Cov}(X_1, X_2)$ gives $\text{Cov}(X_1, X_2) = -np_1p_2$. By the same logic, for $i \neq j$, we have $\text{Cov}(X_i, X_j) = -np_ip_j$, and of course the components are negatively correlated as we would expect. □

7.5 Multivariate normal

Definition 7.23. Multivariate Normal distribution
The random vector $\mathbf{X} = (X_1, \ldots, X_k)$ has a MVN distribution if every linear combination of the X_j has a Normal distribution, i.e.,

$$t_1X_1 + \cdots + t_kX_k$$

has a Normal distribution for any constants $t_1, \ldots, t_k$.

- The marginal distribution of each X_i is Normal, since we may consider $t_i = 1$ and all other $t_j = 0$. However, it is possible to have Normally distributed r.v.s $X_1, \ldots, X_k$ such that $\mathbf{X}$ is not MVN.
- Each subvector, say, (X_1, X_2), is of course MVN.

A MVN distribution has two parameters, the mean vector and the covariance matrix. (Note that the diagonal entries of the covariances gives the variances.)

Theorem 7.24.
If independent $\mathbf{X}$ and $\mathbf{Y}$ are MVN, the concatenated random vector $\mathbf{W} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is MVN, for the sum of independent Normal r.v.s is Normal.

Definition 7.25. Joint MGF
The joint MGF of a random vector $\mathbf{X}$ is

$$M(\mathbf{t}) = E\left(e^{\mathbf{t}\cdot\mathbf{X}}\right) = E\left(e^{t_1X_1+\cdots+t_kX_k}\right),$$

for $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{R}^k$. We require this expectation to be finite in a box containing the origin in $\mathbb{R}^k$; otherwise we say the joint MGF does not exist.

For a MVN random vector, we may recall that any Normal r.v. W has

$$E\left(e^W\right) = e^{E(W)+\frac{1}{2}\text{Var}(W)},$$

and then we have the joint MGF of an MVN $\mathbf{X}$

$$M(\mathbf{t}) = E\left(e^{t_1X_1+\cdots+t_kX_k}\right) = \exp\left(t_1E(X_1) + \cdots + t_kE(X_k) + \frac{1}{2}\text{Var}(t_1X_1 + \cdots + t_kX_k)\right).$$

Theorem 7.26. Correlation and independence in an MVN random vector
Within an MVN random vector, uncorrelated implies independent. That is, if $\mathbf{X} \sim$ MVN can be written as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are subvectors, and every component of $\mathbf{X}_1$ is uncorrelated with every component of $\mathbf{X}_2$, then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent.

In particular, if (X, Y) is Bivariate Normal and $\text{Corr}(X, Y) = 0$, then X and Y are independent. This may be proved by joint MGFs.

Remark 7.27.
Random variables with continuous distributions cannot have a probability mass, i.e., we cannot have $P(X = x) > 0$ for any x. (Note this result is not stated explicitly in the textbook.)

8 Transformations

8.1 Change of variables

Theorem 8.1. Change of variables in one dimension
Let X be a continuous r.v. with PDF f_X, and let $Y = g(X)$, where g is differentiable and strictly increasing (or strictly decreasing). Then the PDF of Y is given by

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

where $x = g^{-1}(y)$. The support of Y is all $g(x)$ with x in the support of X.

The change of variables formula (in the strictly increasing g case) is easy to remember when written in the form

$$f_Y(y)dy = f_X(x)dx.$$

Proof. Let g be strictly increasing. The CDF of Y is

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right) = F_X(x),$$

so by the chain rule, the PDF of Y is

$$f_Y(y) = f_X(x)\frac{dx}{dy}$$

The proof for g strictly decreasing is analogous. In that case the PDF ends up as $-f_X(x)\frac{dx}{dy}$, which is nonnegative since $\frac{dx}{dy} < 0$ if g is strictly decreasing. Using $\left|\frac{dx}{dy}\right|$, as in the statement of the theorem, covers both cases. □

8.2 Convolutions

Definition 8.2.
A convolution is a sum of independent random variables.

Theorem 8.3. Convolution sums and integrals
Let X and Y be independent r.v.s and $T = X + Y$ be their sum. If X and Y are discrete, then the PMF of T is

$$\begin{aligned} P(T = t) &= \sum_x P(Y = t - x)P(X = x) \\ &= \sum_y P(X = t - y)P(Y = y). \end{aligned}$$

If X and Y are continuous, then the PDF of T is

$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} f_Y(t - x)f_X(x)dx \\ &= \int_{-\infty}^{\infty} f_X(t - y)f_Y(y)dy. \end{aligned}$$

Proof. For the discrete case, we use LOTP, conditioning on X:

$$\begin{aligned} P(T=t) &= \sum_x P(X+Y=t \mid X=x)P(X=x) \\ &= \sum_x P(Y=t-x \mid X=x)P(X=x) \\ &= \sum_x P(Y=t-x)P(X=x), \end{aligned}$$

where the last inequality follows from independence of X, Y.

In the continuous case, since the value of a PDF at a point is not a probability, we first find the CDF, and then differentiate to get the PDF. By LOTP,

$$\begin{aligned} F_T(t) = P(X+Y \leq t) &= \int_{-\infty}^{\infty} P(X+Y \leq t \mid X=x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} P(Y \leq t-x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} F_Y(t-x) f_X(x) dx \end{aligned}$$

Again, we need independence to drop the condition $X = x$. To get the PDF, we then differentiate with respect to t, interchanging the order of integration and differentiation. This gives

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t-x) f_X(x) dx.$$

□

8.3 Beta

Definition 8.4. Beta distribution
For $a > 0$ and $b > 0$, we have

$$X \sim \text{Beta}(a, b)$$

if X has PDF

$$f(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1,$$

where the normalizing constant $\beta(a, b)$ is chosen to make the PDF integrate to 1. As derived in Example 8.5.2, then for $W \sim \text{Beta}(a, b)$ we have

$$E(W) = \frac{a}{a+b}.$$

Note that for $a = b = 1$, Beta$(1, 1)$ is constant on $(0, 1)$, so Beta$(1, 1)$ and Unif$(0, 1)$ are the same. For varying the values of a and b, the shape of Beta has a few general patterns:

- If $a < 1$ and $b < 1$, the PDF is U-shaped and opens upward. If $a > 1$ and $b > 1$, the PDF opens down.
- If $a = b$, the PDF is symmetric about $1/2$. If $a > b$, the PDF favors values larger than $1/2$; if $a < b$, the PDF favors values smaller than $1/2$.

Example 8.5. Bayes' billiards
For any integers k and n with $0 \leq k \leq n$, we have

$$\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}.$$

Proof. We will construct an r.v. X in two different ways, calculating $P(X = k)$ for both constructions, and then equating.

- Suppose we have $n+1$ balls, of which 1 is black and n are white. Throw all balls onto the interval [0,1] such that the positions of the balls are i.i.d. Unif$(0,1)$. Let X be the number of white balls to the left of the black ball, and B the position of the black ball. Given $B = p$, observe that the number of white balls on the left of the black ball is binomial, so we may calculate with LOTP that

$$P(X = k) = \int_0^1 P(X = k \mid B = p) f(p) dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp.$$

- Now suppose we have $n+1$ balls, all white. Randomly throw each ball onto the unit interval; then choose one ball at random and paint it gray. Again, let X be the number of white balls to the left of the gray ball. By symmetry, any one of the $n+1$ balls is equally likely to be painted gray, so for $k = 0, 1, \ldots, n$, we have

$$P(X = k) = \frac{1}{n+1}.$$

Equating, we are done. □

Substituting $k = a - 1$ for $n - k = b - 1$, we have that for positive integer values of a and b,

$$\beta(a,b) = \frac{1}{(a+b-1)\binom{a+b-2}{a-1}} = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

Proposition 8.6. Beta-Binomial conjugacy
Suppose a coin lands Heads with probability p, and we use Bayesian inference to infer the value of p given the number X of heads in the first n tosses of the coin. That is, we effectively assign a random variable to p, assuming it follows some prior distribution, and update by conditioning on X.

Suppose the prior distribution on p is Beta, say, $p \sim \text{Beta}(a, b)$ for constants a, b. Of course X is conditionally Binomial given p, i.e.,

$$X \mid p \sim \text{Bin}(n, p),$$

but to get the posterior distribution of p, we use Bayes' on the prior distribution $f(p)$ to get

$$\begin{aligned} f(p \mid X = k) &= \frac{P(X = k \mid p) f(p)}{P(X = k)} \\ &= \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{1}{\beta(a,b)} p^{a-1}(1-p)^{b-1}}{P(X = k)}. \end{aligned}$$

The denominator, which is the marginal PMF of X, is given by

$$P(X = k) = \int_0^1 P(X = k \mid p) f(p) dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} f(p) dp,$$

and for $a = b = 1$ (i.e., a Unif$(0, 1)$ prior on p), we showed in the Bayes' billiards story that $P(X = k) = 1/(n + 1)$, i.e., X is Discrete Uniform on $\{0, 1, \ldots, n\}$.

In general, we have

$$f(p \mid X = k) \propto p^{a+k-1}(1-p)^{b+n-k-1},$$

implying that

$$p \mid X = k \sim \text{Beta}(a + k, b + n - k).$$

In other words, if we have a Beta prior distribution on p and data that are conditionally Binomial given p, then when going from prior to posterior, we simply update the parameters of the Beta distribution. We say that the Beta is the conjugate prior of the Binomial.

8.4 Gamma

Definition 8.7. Gamma function
For real numbers $a > 0$, the gamma function Γ is defined by

$$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x}.$$

It is convenient to write the $\frac{dx}{x}$ factor because we often make transformations of form $u = cx$ and use $\frac{du}{u} = \frac{dx}{x}$.

Here are two important properties of the gamma function.

- $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$. This follows from integration by parts:

$$\Gamma(a+1) = \int_0^\infty x^a e^{-x} dx = -\, x^a e^{-x}\big|_0^\infty + a \int_0^\infty x^{a-1} e^{-x} dx = 0 + a\Gamma(a)$$

- $\Gamma(n) = (n-1)!$ if n is a positive integer, as is easily verified by induction.

Definition 8.8. Gamma distribution
From the definition of the gamma function, observe that

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

and we say that this integrand, being a valid PDF on $(0, \infty)$, is the PDF of Gamma$(a, 1)$.

We obtain the general Gamma distribution by a scale transformation: given $X \sim$ Gamma$(a, 1)$, we have $Y = X/\lambda \sim$ Gamma(a, λ) with PDF

$$f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right| = \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{\lambda y} \left|\frac{d}{dy} \lambda y\right| = \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{y}, \quad y > 0.$$

Summarizing, for parameters $a > 0$ and $\lambda > 0$, we have the Gamma distribution

$$Y \sim \text{Gamma}(a, \lambda),$$

$$f(y) = \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{y}, \quad y > 0.$$

Note that Gamma$(1, \lambda)$ is distributed as Expo(λ).

Proposition 8.9. Properties of Gamma
To calculate mean, variance, and moments of $X \sim \text{Gamma}(a, 1)$, we may use properties of the gamma function:

$$\begin{aligned} E(X) &= \int_0^\infty \frac{1}{\Gamma(a)} x^{a+1} e^{-x} \frac{dx}{x} = \frac{\Gamma(a+1)}{\Gamma(a)} = a, \\ \text{Var}(X) &= E(X^2) - a^2 = \frac{\Gamma(a+2)}{\Gamma(a)} - a^2 = a(a+1) - a^2 = a, \\ E(X^c) &= \frac{\Gamma(a+c)}{\Gamma(a)}. \end{aligned}$$

Then for the general gamma distribution $Y \sim \text{Gamma}(a, \lambda)$, we may simply transform X:

$$\begin{aligned} E(Y) &= \frac{1}{\lambda} E(X) = \frac{a}{\lambda}, \\ \text{Var}(Y) &= \frac{1}{\lambda^2} \text{Var}(X) = \frac{a}{\lambda^2}, \\ E\left(Y^c\right) &= \frac{1}{\lambda^c} E\left(X^c\right) = \frac{1}{\lambda^c} \cdot \frac{\Gamma(a+c)}{\Gamma(a)}, \quad c > -a. \end{aligned}$$

Theorem 8.10.
Let $X_1, \ldots, X_n$ be i.i.d. Expo(λ). Then

$$X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda).$$

This can be proven with MGFs or induction.

As a corollary, observe that the n^{th} arrival time in a Poisson process of rate λ is distributed as $\Gamma(n, \lambda)$. (Note that even though the times between each arrival are independent, the times the T_i themselves are of course not independent.)

Proposition 8.11. Gamma-Poisson conjugacy
Suppose we have a Poisson process with unknown rate λ that we assume to have prior $\text{Gamma}(r_0, b_0)$. Intuitively, r_0 is interpreted as the number of prior arrivals and b_0 as the total time required for those prior arrivals.

We update λ based on observations of Y_t, the number of arrivals in some given time interval t. We have that

$$\begin{aligned} \lambda &\sim \text{Gamma}(r_0, b_0), \\ Y_t \mid \lambda &\sim \text{Pois}(\lambda t). \end{aligned}$$

Then we have the posterior

$$\lambda \mid Y_t = k \sim \text{Gamma}(r_0 + y, b_0 + t).$$

See Story 8.4.5 (page 393) for the full derivation.

8.5 Beta-Gamma connections

Example 8.12. Bank-post office
While running errands, you need to go to the bank, then to the post office. Let

$$\begin{aligned} X &\sim \text{Gamma}(a, \lambda), \\ Y &\sim \text{Gamma}(b, \lambda) \end{aligned}$$

be your waiting times at the bank and at the post office, respectively. Assuming X and Y are independent, we have that

$$\begin{aligned} T = X + Y &\sim \text{Gamma}(a + b, \lambda) \\ W = \frac{X}{X + Y} &\sim \text{Beta}(a, b), \end{aligned}$$

and that T and W are independent. See story 8.5.1 (page 396) for the full derivation.

8.6 Order statistics

Definition 8.13. Order statistics
Given r.v.s $X_1, \ldots, X_n$, the i^{th} order statistic is the r.v. $X_{(i)}$ that is the i^{th}-smallest of $X_1, \ldots, X_n$. We are mainly interested in the case that $X_1, \ldots, X_n$ are i.i.d.

If n is odd, then $X_{(n+1/2)}$ is called the sample median.

Theorem 8.14. CDF of order statistic
Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with CDF F. Then the CDF of the j th order statistic $X_{(j)}$ is

$$P\left(X_{(j)} \leq x\right) = \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.$$

Theorem 8.15. PDF of order statistic
Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with CDF F and PDF f. Then the marginal PDF of the j th order statistic $X_{(j)}$ is

$$f_{X_{(j)}}(x) = n\binom{n-1}{j-1} f(x) F(x)^{j-1} (1 - F(x))^{n-j}.$$

Proof. What is the probability $f_{X_{(j)}}(x)dx$ of this extremely specific event?

- First, we choose which of the n r.v.s X_i will fall into the infinitesimal interval around x, each of which occurs with probability $f(x)dx$, where f is the PDF of the X_i.
- Next, we choose $j-1$ of the remaining $n-1$ to fall to the left of x, and each possible combination occurs with probability $F(x)^{j-1}(1 - F(x))^{n-j}$.

Together, we have

$$f_{X_{(j)}}(x)dx = n\binom{n-1}{j-1} f(x)dx F(x)^{j-1} (1 - F(x))^{n-j},$$

and we may drop the dx's to finish. □

9 Conditional expectation

9.1 Conditional expectation given an event

Definition 9.1. Conditional expectation given an event
Let A be an event with positive probability. If Y is a discrete r.v., then the conditional expectation of Y given A is

$$E(Y \mid A) = \sum_y yP(Y = y \mid A),$$

where the sum is over the support of Y. If Y is a continuous r.v. with PDF f, then

$$E(Y \mid A) = \int_{-\infty}^{\infty} yf(y \mid A)dy$$

where the conditional PDF $f(y \mid A)$ is defined as the derivative of the conditional CDF $F(y \mid A) = P(Y \leq y \mid A)$, and can also be computed by a hybrid version of Bayes' rule:

$$f(y \mid A) = \frac{P(A \mid Y = y)f(y)}{P(A)}$$

Theorem 9.2. Law of total expectation
Let $A_1, \ldots, A_n$ be a partition of a sample space, with $P(A_i) > 0$ for all i, and let Y be a random variable on this sample space. Then

$$E(Y) = \sum_{i=1}^{n} E(Y \mid A_i) P(A_i).$$

In fact, since all probabilities are expectations by the fundamental bridge, LOTP is simply LOTE on some indicator variable.

9.2 Conditional expectation given an r.v.

Definition 9.3. Conditional expectation given an r.v.
Let $g(x) = E(Y \mid X = x)$. Then the conditional expectation of Y given X, denoted $E(Y \mid X)$, is defined to be the random variable $g(X)$. In other words, if after doing the experiment X crystallizes into x, then $E(Y \mid X)$ crystallizes into $g(x)$.

9.3 Properties of conditional expectation

Proposition 9.4. Properties of conditional expectation
Conditional expectation has these properties:

- Dropping what's independent: For independent X and Y,

$$E(Y \mid X) = E(Y).$$

- Taking out what's known: For any function h,

$$E(h(X)Y \mid X) = h(X)E(Y \mid X).$$

- Linearity

- Adam's law:

$$E(E(Y \mid X)) = E(Y).$$

  Note that we may also apply extra conditioning:

$$E(E(Y \mid X, Z) \mid Z) = E(Y \mid Z).$$

- Projection interpretation: The r.v. $Y - E(Y \mid X)$, called the residual from using X to predict Y, is uncorrelated with $h(X)$ for any function h.

9.4 Geometric interpretation of conditional expectation

Skipped for now.

9.5 Conditional variance

Definition 9.5. Conditional variance
The conditional variance of Y given X is

$$\text{Var}(Y \mid X) = E\left((Y - E(Y \mid X))^2 \mid X\right)$$

This is equivalent to

$$\text{Var}(Y \mid X) = E\left(Y^2 \mid X\right) - (E(Y \mid X))^2$$

Theorem 9.6. Law of total variance, Eve's law, EVVE
For any r.v.s X and Y,

$$\text{Var}(Y) = E(\text{Var}(Y \mid X)) + \text{Var}(E(Y \mid X)).$$

Intuitively, we may consider a population where each person has a value of X and a value of Y. Then we may divide this population into subpopulations for each possible value of X, and then the total variance $\text{Var}(Y)$ is the sum of the mean variation $E(\text{Var}(Y \mid X))$ within each group and the variation $\text{Var}(E(Y \mid X))$ between groups. Ideally there would be no variation within groups, so within-group variation may be considered unexplained while between-group variation is explained.

Proof. Let $g(X) = E(Y \mid X)$. By Adam's law, $E(g(X)) = E(Y)$. Then

$$E(\text{Var}(Y \mid X)) = E\left(E\left(Y^2 \mid X\right) - g(X)^2\right) = E\left(Y^2\right) - E\left(g(X)^2\right),$$
$$\text{Var}(E(Y \mid X)) = E\left(g(X)^2\right) - (Eg(X))^2 = E\left(g(X)^2\right) - (EY)^2.$$

Adding these equations, we have Eve's law. □

9.6 Adam and Eve examples

10 Inequalities and limit theorems

10.1 Inequalities

Theorem 10.1. Cauchy-Schwarz: a marginal bound on a joint expectation
For any r.v.s X and Y with finite variances,

$$|E(XY)| \leq \sqrt{E\left(X^2\right) E\left(Y^2\right)}.$$

Theorem 10.2. Jensen: an inequality for convexity
Let X be a random variable. If g is a convex function, then $E(g(X)) \geq g(E(X))$. If g is a concave function, then $E(g(X)) \leq g(E(X))$. In both cases, the only way that equality can hold is if there are constants a and b such that $g(X) = a + bX$ with probability 1.

The next three inequalities are Markov, Chebyshev, Chernoff. They are bounds on tail probabilities.

Theorem 10.3. Markov
For any r.v. X and constant $a > 0$,

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

Markov's is usually a weak bound because it requires no assumptions about X; the RHS could be greater than 1 or even infinite. But Chebyshev and Chernoff can be easily derived from Markov's inequality and give stronger bounds.

Proof. Let $Y = \frac{|X|}{a}$. We need to show that $P(Y \geq 1) \leq E(Y)$. Note that

$$I(Y \geq 1) \leq Y$$

since if $I(Y \geq 1) = 0$ then the inequality reduces to $Y \geq 0$, and if $I(Y \geq 1) = 1$ then $Y \geq 1$ (because the indicator says so). Taking the expectation of both sides, we have Markov's inequality. □

Theorem 10.4. Chebyshev
Let X have mean μ and variance σ^2. Then for any $a > 0$

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

Substituting $c\sigma$ for a, for $c > 0$, we have the following equivalent form of Chebyshev's inequality:

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2},$$

giving us an upper bound on the probability of an r.v. being more than c standard deviations away from its mean, e.g., there can't be more than a 25% chance of being 2 or more standard deviations from the mean.

Proof. By Markov's inequality,

$$P(|X-\mu| \geq a) = P\left((X-\mu)^2 \geq a^2\right) \leq \frac{E(X-\mu)^2}{a^2} = \frac{\sigma^2}{a^2}.$$

□

Theorem 10.5. Chernoff
For any r.v. X and constants $a > 0$ and $t > 0$,

$$P(X \geq a) \leq \frac{E\left(e^{tX}\right)}{e^{ta}}.$$

Chernoff's bound offers two very nice features:

1. The right-hand side can be optimized over t to give the tightest upper bound, as in the proof of Cauchy-Schwarz.
2. If the MGF of X exists, then the numerator in the bound is the MGF, and some of the useful properties of MGFs can come into play.

Proof. The transformation g with $g(x) = e^{tx}$ is invertible and strictly increasing. So by Markov's inequality, we have

$$P(X \geq a) = P\left(e^{tX} \geq e^{ta}\right) \leq \frac{E\left(e^{tX}\right)}{e^{ta}}.$$

□

10.2 Law of large numbers

Definition 10.6. Sample mean
Assume $X_1, X_2, X_3, \ldots$ are i.i.d. with finite mean μ and finite variance σ^2. Define the sample mean of X_1 through X_n as

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n},$$

which it itself an r.v., with mean μ and variance σ^2/n. The LLN and CLT describe the behavior of the sample mean of i.i.d. r.v.s as the sample size grows.

Theorem 10.7. Law of large numbers (LLN)
As n grows, the sample mean $\bar{X}_n$ converges to the true mean μ. The LLN comes in two versions with slightly different definitions of what it means for a sequence of r.v.s to converge to a number:

- **Strong LLN:** $\bar{X}_n$ converges to the true mean μ pointwise with probability 1,

$$P\left(\bar{X}_n \to \mu\right) = 1.$$

  Recalling that r.v.s are functions from the sample space S to $\mathbb{R}$, this form of convergence says that $\bar{X}_n(s) \to \mu$ for each point $s \in S$, except that the convergence is allowed to fail on some set B_0 of exceptions, as long as $P(B_0) = 0$.

- **Weak LLN (convergence in probability):** For all $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) = 0.$$

10.3 Central limit theorem

Theorem 10.8. Central limit theorem
For large n, the distribution of $\bar{X}_n$ after standardization approaches a standard Normal distribution. As $n \to \infty$,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to \mathcal{N}(0,1).$$

In other words, the CDF of the LHS converges to Φ. In approximation form, this means that for large n, the distribution of $\bar{X}_n$ is approximately $\mathcal{N}\left(\mu, \sigma^2/n\right)$.

Proof. We prove the CLT assuming that the MGF of the X_j exists, though the CLT holds more generally. Let $M(t) = E\left(e^{tX_j}\right)$, and WLOG let us standard the X_j such that $\mu = 0, \sigma^2 = 1$. Then $M(0) = 1, M'(0) = \mu = 0$, and $M''(0) = \sigma^2 = 1$.

As may be proven with analysis, if $Z_1, Z_2, \ldots$ are r.v.s whose MGFs converge to the MGF of a continuous r.v. Z, then the CDF of Z_n converges to the CDF of Z. (This is intuitive given that the MGF of an r.v. determines its distribution.) Thus we wish to show that the MGF of

$$\sqrt{n}\bar{X}_n = \left(X_1 + \cdots + X_n\right)/\sqrt{n}$$

converges to the MGF of $\mathcal{N}(0,1)$, which is $e^{t^2/2}$.

By properties of MGFs,

$$\begin{aligned} E\left(e^{t(X_1+\cdots+X_n)/\sqrt{n}}\right) &= E\left(e^{tX_1/\sqrt{n}}\right) E\left(e^{tX_2/\sqrt{n}}\right) \ldots E\left(e^{tX_n/\sqrt{n}}\right) \\ &= \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n. \end{aligned}$$

Letting $n \to \infty$, we get the indeterminate form 1^∞, so instead we should take the limit of the

logarithm, $n \log M\left(\frac{t}{\sqrt{n}}\right)$, and then exponentiate at the end. This gives

$$\begin{aligned}
&\lim_{n\to\infty} n \log M\left(\frac{t}{\sqrt{n}}\right) = \lim_{y\to 0} \frac{\log M(yt)}{y^2} \quad \text{where } y = 1/\sqrt{n} \\
&= \lim_{y\to 0} \frac{tM'(yt)}{2yM(yt)} \quad \text{by L'Hôpital's rule} \\
&= \frac{t}{2} \lim_{y\to 0} \frac{M'(yt)}{y} \quad \text{since } M(yt) \to 1 \\
&= \frac{t^2}{2} \lim_{y\to 0} M''(yt) \quad \text{by L'Hôpital's rule} \\
&= \frac{t^2}{2},
\end{aligned}$$

as desired. □

Example 10.9. Binomial convergence to Normal, continuity correction
Let $Y \sim \text{Bin}(n, p)$. This is a sum of n i.i.d. Bern(p) r.v.s, and so for large n, we have

$$Y \sim \mathcal{N}(np, np(1-p)).$$

To account for the discreteness of Y, we may write

$$P(Y = k) = P(k - \frac{1}{2} < Y < k + \frac{1}{2}) \approx \Phi\left(\frac{k + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}}\right).$$

The Poisson approximation (Chapter 4) works best when p is small, while this Normal approximation works best when n is large and p is around $1/2$, so that the distribution of Y is symmetric.

10.4 Chi-Square and Student-t

Definition 10.10. Chi-Square distribution
An r.v. V has the Chi-Square distribution with n degrees of freedom,

$$V \sim \chi_n^2,$$

if we have

$$V = Z_1^2 + \cdots + Z_n^2$$

for $Z_1, \ldots, Z_n$ i.i.d. $\mathcal{N}(0, 1)$. Equivalently, the χ_n^2 distribution is the Gamma $\left(\frac{n}{2}, \frac{1}{2}\right)$ distribution.

When our random variables are i.i.d. Normals, the distribution of the sample variance after appropriate scaling is Chi-Square, and so the Chi-Square distribution is important for estimating the true variance of a distribution.

Definition 10.11. Student-t distribution (t distribution)
An r.v. T has the t distribution with n degrees of freedom, written

$$T \sim t_n,$$

if we have

$$T = \frac{Z}{\sqrt{V/n}},$$

where $Z \sim \mathcal{N}(0,1), V \sim \chi^2_n$, and Z is independent of V. The t_n distribution has the following properties:

- Symmetry: If $T \sim t_n$, then $-T \sim t_n$ as well.
- Cauchy as special case: The t_1 distribution is the same as the Cauchy distribution, introduced in Example 7.1.25.
- Convergence to Normal: As $n \to \infty$, the t_n distribution approaches the standard Normal distribution.

Though it is not generally useful, the PDF of t_n is

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + t^2/n\right)^{-(n+1)/2}.$$

This looks similar to a standard Normal, except with heavier tails (much heavier for small n, and not much heavier for large n). The t distribution forms the basis for hypothesis testing procedures known as t-tests.

11 Markov chains

In this chapter, we focus on discrete-state, discrete-time, and time-homogeneous Markov chains, with a finite state space.

11.1 Markov property and transition matrix

Definition 11.1. Markov chain
A sequence of r.v.s $X_0, X_1, X_2, \ldots$ taking values in the state space $\{1, 2, \ldots, M\}$ is called a Markov chain if any particular transition probability depends only on the most recent state, i.e., for all $n \geq 0$,

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i).$$

This relation is called the Markov property, and the RHS probability is called a transition probability.

Definition 11.2. Transition matrix
Let $X_0, X_1, X_2, \ldots$ be a Markov chain with state space $\{1, 2, \ldots, M\}$, and let $q_{ij} = P(X_{n+1} = j \mid X_n = i)$ be the transition probability from state i to state j. The $M \times M$ matrix $Q = (q_{ij})$ is called the transition matrix of the chain, in which each row sums to 1.

Remark 11.3. Markov chains for n^{th} order relations
Suppose we had the state space $\{R, S\}$ whose transition probabilities had second order dependency. Then we could modify our transition matrix to the states

$$(R, S),\ (R, R),\ (S, S),\ (S, R),$$

to capture these second order dependencies, and similarly for n^{th} order dependencies.

Definition 11.4. n-step transition probability
The n-step transition probability from i to j is the probability of being at j exactly n steps after being at i. We denote this by $q_{ij}^{(n)}$:

$$q_{ij}^{(n)} = P(X_n = j \mid X_0 = i)$$

Note that

$$q_{ij}^{(2)} = \sum_k q_{ik} q_{kj},$$

and since the RHS is the (i, j) entry of Q^2 by definition of matrix multiplication, we conclude that the matrix Q^2 gives the two-step transition probabilities. By induction, the n^{th} power of the transition matrix gives the n-step transition probabilities:

$$q_{ij}^{(n)} \text{ is the } (i, j) \text{ entry of } Q^n.$$

Proposition 11.5. Marginal distribution of X_n
Define $\mathbf{t} = (t_1, t_2, \ldots, t_M)$ by $t_i = P(X_0 = i)$, and view $\mathbf{t}$ as a row vector. Then the marginal distribution of X_n is given by the vector $\mathbf{t}Q^n$. That is, the j^{th} component of $\mathbf{t}Q^n$ is $P(X_n = j)$.

Proof. By the law of total probability, conditioning on X_0, the probability that the chain is in state j after n steps is

$$P(X_n = j) = \sum_{i=1}^{M} P(X_0 = i) P(X_n = j \mid X_0 = i) = \sum_{i=1}^{M} t_i q_{ij}^{(n)},$$

which is the j^{th} component of $\mathbf{t}Q^n$ by definition of matrix multiplication. □

11.2 Classification of states

Definition 11.6. Recurrent and transient states
State i of a Markov chain is recurrent if starting from i, the probability is 1 that the chain will eventually return to i. Otherwise, the state is transient, which means that if the chain starts from i, there is a positive probability of never returning to i.

In fact, although the definition of a transient state only requires that there be a positive probability of never returning to the state, we can say something stronger: as long as there is a positive probability of leaving i forever, the chain eventually will leave i forever. Moreover, we can find the distribution of the number of returns to the state.

Proposition 11.7. Number of returns to transient state is Geometric
Let i be a transient state of a Markov chain. Suppose the probability of never returning to i, starting from i, is a positive number $p > 0$. Then, starting from i, the number of times that the chain returns to i before leaving forever is distributed Geom(p).

Definition 11.8. Irreducible and reducible chain
A Markov chain with transition matrix Q is irreducible if for any two states i and j, it is possible to go from i to j in a finite number of steps (with positive probability). That is, for any states i, j there is some positive integer n such that the (i, j) entry of Q^n is positive. A Markov chain that is not irreducible is called reducible.

Proposition 11.9. Irreducible implies all states recurrent
In an irreducible Markov chain with a finite state space, all states are recurrent.

Proof. It is clear that at least one state must be recurrent; if all states were transient, the chain would eventually leave all states forever and have nowhere to go! So assume without loss of generality that state 1 is recurrent, and consider any other state i. We know that $q_{1i}^{(n)}$ is positive for some n, by definition of irreducibility. Thus, every time the chain is at state 1, there is a positive probability that after n more steps it will be at state i.

Since the chain visits state 1 infinitely often, we know the chain will eventually reach state i from state 1; think of each visit to state 1 as starting a trial, where "success" is defined as reaching state i in at most n steps. From state i, the chain will return to state 1 because state 1 is recurrent, and by the same logic, it will eventually reach state i again. By induction, the chain will visit state i infinitely often. Since i was arbitrary, we conclude that all states are recurrent. □

Definition 11.10. Period of a state, periodic and aperiodic chain
The period of a state i in a Markov chain is the greatest common divisor (gcd) of the possible numbers of steps it can take to return to i when starting at i. That is, the period of i is the greatest common divisor of numbers n such that the (i, i) entry of Q^n is positive. (The period of i is ∞ if it's impossible ever to return to i after starting at i.) A state is called aperiodic if its period equals 1, and periodic otherwise. The chain itself is called aperiodic if all its states are aperiodic, and periodic otherwise.

Proposition 11.11. Periods in an irreducible chain
In an irreducible Markov chain, all states have the same period.

11.3 Stationary distribution

Definition 11.12. Stationary distribution; existence and uniqueness of stationary distribution
Intuitively, this is the long-run behavior of the chain, regardless of its initial conditions. A row vector $\mathbf{s} = (s_1, \ldots, s_M)$ such that $s_i \geq 0$ and $\sum_i s_i = 1$ is a stationary distribution for a Markov chain with transition matrix Q if

$$\sum_i s_i q_{ij} = s_j$$

for all j. This system of linear equations can be written as one matrix equation:

$$\mathbf{s}Q = \mathbf{s}.$$

For any irreducible Markov chain, there exists a unique stationary distribution. In this distribution, every state has positive probability.

Theorem 11.13. Convergence to stationary distribution
Let $X_0, X_1, \ldots$ be an irreducible, aperiodic Markov chain with stationary distribution $\mathbf{s}$ and transition matrix Q. Then $P(X_n = i)$ converges to s_i as $n \to \infty$. In terms of the transition matrix, Q^n converges to a matrix in which each row is $\mathbf{s}$.

Theorem 11.14. Expected time to return
Let $X_0, X_1, \ldots$ be an irreducible Markov chain with stationary distribution $\mathbf{s}$. Let r_i be the expected time it takes the chain to return to i, given that it starts at i. Then

$$s_i = \frac{1}{r_i}.$$

11.4 Reversibility

Definition 11.15. Reversibility
Let $Q = (q_{ij})$ be the transition matrix of a Markov chain. Suppose there is $\mathbf{s} = (s_1, \ldots, s_M)$ with $s_i \geq 0$ and $\sum_i s_i = 1$, such that

$$s_i q_{ij} = s_j q_{ji}$$

for all states i and j. This equation is called the reversibility or detailed balance condition, and we say that the chain is reversible with respect to $\mathbf{s}$ if it holds.

The term "reversible" comes from the fact that a reversible chain, started according to its stationary distribution, behaves in the same way regardless of whether time is run forwards or backwards.

Proposition 11.16. Reversible implies stationary
Suppose that $Q = (q_{ij})$ is the transition matrix of a Markov chain that is reversible with respect to a nonnegative vector $\mathbf{s} = (s_1, \ldots, s_M)$ whose components sum to 1. Then $\mathbf{s}$ is a stationary distribution of the chain.

Proof. We have

$$\sum_i s_i q_{ij} = \sum_i s_j q_{ji} = s_j \sum_i q_{ji} = s_j$$

where the last equality is because each row sum of Q is 1. So s is stationary. □

Proposition 11.17.
If each column of the transition matrix Q sums to 1, then the uniform distribution over all states, $(1/M, 1/M, \ldots, 1/M)$, is a stationary distribution. (A nonnegative matrix such that the row sums and the column sums are all equal to 1 is called a doubly stochastic matrix.)

Proof. Assuming each column sums to 1, the row vector $\mathbf{v} = (1, 1, \ldots, 1)$ satisfies $\mathbf{v}Q = \mathbf{v}$. It follows that $(1/M, 1/M, \ldots, 1/M)$ is stationary. □

Example 11.18. Random walk on an undirected network
This is a collection of nodes with bidirectionally-traversable edges such that from node i, the probabilities of traversing any edge at i are equal. Self-loops are allowed.

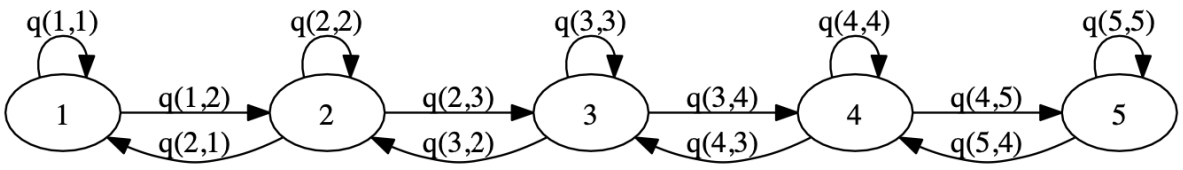
The degree of a node is the number of edges attached to it, and the degree sequence $(d_1, \ldots, d_n)$ lists all the degrees. Observe that

$$d_i q_{ij} = d_j q_{ji}$$

for all i, j, because both sides are 1 if $\{i, j\}$ is an edge and 0 otherwise. Thus the reversible (and thus stationary) distribution is proportional to the degree sequence.

Example 11.19. Birth-death chain
A birth-death chain on states $\{1, 2, \ldots, M\}$ is a Markov chain with transition matrix $Q = (q_{ij})$ such that $q_{ij} > 0$ if $|i - j| = 1$ and $q_{ij} = 0$ if $|i - j| \geq 2$. This says it's possible to go one step to the left and possible to go one step to the right (except at boundaries) but impossible to jump further in one step.

For example, the chain shown below is a birth-death chain if the labeled transitions have positive probabilities, except for the loops from a state to itself, which are allowed to have 0 probability.



Solution. We will now show that any birth-death chain is reversible, and construct the stationary distribution. Let s_1 be a positive number, to be specified later. Since we want $s_1 q_{12} = s_2 q_{21}$, let

$$s_2 = s_1 q_{12}/q_{21}.$$

Then since we want $s_2 q_{23} = s_3 q_{32}$, let

$$s_3 = s_2 q_{23}/q_{32} = s_1 q_{12} q_{23}/\left(q_{32} q_{21}\right)$$

Continuing in this way, let

$$s_j = \frac{s_1 q_{12} q_{23} \cdots q_{j-1,j}}{q_{j,j-1} q_{j-1,j-2} \cdots q_{21}}$$

for all states j with $2 \leq j \leq M$. Choose s_1 so that the s_j sum to 1. Then the chain is reversible with respect to $\mathbf{s}$, since $q_{ij} = q_{ji} = 0$ if $|i - j| \geq 2$ and by construction $s_i q_{ij} = s_j q_{ji}$ if $|i - j| = 1$. Thus, $\mathbf{s}$ is the stationary distribution.

12 Markov chains Monte Carlo

12.1 Metropolis-Hastings

Theorem 12.1. Metropolis-Hastings
Let $\mathbf{s} = (s_1, \ldots, s_M)$ be a desired stationary distribution on state space $\{1, \ldots, M\}$. Assume that $s_i > 0$ for all i (if not, just delete any states i with $s_i = 0$ from the state space). Suppose that $P = (p_{ij})$ is the transition matrix for a Markov chain on state space $\{1, \ldots, M\}$. Intuitively, P is a Markov chain that we know how to run but that doesn't have the desired stationary distribution.

Our goal is to modify P to construct a Markov chain $X_0, X_1, \ldots$ with stationary distribution s. We will give a Metropolis-Hastings algorithm for this. Start at any state X_0 (chosen randomly or deterministically), and suppose that the new chain is currently at X_n. To make one move of the new chain, do the following.

1. If $X_n = i$, propose a new state j using the transition probabilities in the i^{th} row of the original transition matrix P.

2. Compute the acceptance probability

$$a_{ij} = \min\left(\frac{s_j p_{ji}}{s_i p_{ij}}, 1\right).$$

3. Flip a coin that lands Heads with probability a_{ij}.

4. If the coin lands Heads, accept the proposal (i.e., go to j), setting $X_{n+1} = j$. Otherwise, reject the proposal (i.e., stay at i), setting $X_{n+1} = i$.

That is, the Metropolis-Hastings chain uses the original transition probabilities p_{ij} to propose where to go next, then accepts the proposal with probability a_{ij}, staying in its current state in the event of a rejection.

A Table of distributions

Table of Important Distributions for Stat 110

Name	Param.	PMF or PDF	Mean	Variance
Bernoulli	p	$P(X=1)=p, P(X=0)=q$	p	pq
Binomial	n,p	$\binom{n}{k}p^k q^{n-k}$, for $k \in \{0,1,\dots,n\}$	np	npq
FS	p	pq^{k-1}, for $k \in \{1,2,\dots\}$	$1/p$	q/p^2
Geom	p	pq^k, for $k \in \{0,1,2,\dots\}$	q/p	q/p^2
NBin	r,p	$\binom{r+n-1}{r-1}p^r q^n, n \in \{0,1,2,\dots\}$	rq/p	rq/p^2
HGeom	w,b,n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$, for $k \in \{0,1,\dots,n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1})\mu(1-\frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!}$, for $k \in \{0,1,2,\dots\}$	λ	λ
Uniform	$a<b$	$\frac{1}{b-a}$, for $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ,σ^2	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
Log-Normal	μ,σ^2	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}, x>0$	$\theta = e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$
Expo	λ	$\lambda e^{-\lambda x}$, for $x>0$	$1/\lambda$	$1/\lambda^2$
Gamma	a,λ	$\Gamma(a)^{-1}(\lambda x)^a e^{-\lambda x}x^{-1}$, for $x>0$	a/λ	a/λ^2
Beta	a,b	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$, for $0<x<1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
Chi-Square	n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$, for $x>0$	n	$2n$

The function Γ is given by

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

for all $a>0$. For any $a>0$, $\Gamma(a+1) = a\Gamma(a)$. We have $\Gamma(n) = (n-1)!$ for n a positive integer, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

B Problem-solving strategies

Remark B.1.
Problem-solving strategies:

- Try simple and extreme cases
- Name/index objects
- Draw diagram
- Make up a story to decrease abstraction

C Math

Theorem C.1. Taylor

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$$

Theorem C.2. Vandermonde's
As is easily verified with combinatorial argument,

$$\sum_{k=0}^{r} \binom{m}{k}\binom{r}{r-k} = \binom{m+n}{r}.$$

Proposition C.3.

$$\lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n = e^x$$

Proof. Using L'Hopital on the log we have

$$\lim_{n\to\infty} \frac{\log(1+\frac{x}{n})}{1/n} = \lim_{n\to\infty} \frac{n}{n+x}\frac{-x}{n^2}(-n^2) = x.$$

□

Proposition C.4. Harmonic series

$$\sum_{k=1}^{n} \frac{1}{k} \approx \log(n) + \gamma$$

for n large, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant.