Introduction to Probability

Zeming WANG

2025-01-26

Table of contents

Overview

This book is adapted from *MAT921: Probability* at Southwest University of Finance and Economics (RIEM). It is an introductory probability course that aims to be not boring. The course emphasizes:

- Conventional teaching
- Interesting puzzles
- Data-oriented practical skills

Course Instructor's email: gamma12@126.com Please indicate your class and student ID when you email me.

PS: $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ is one of my favorite equation!

Syllabus

Topic 1: Classical probabilities
How likely were some of your classmates born on the same day as you?

Topic 2: Data and random variables
Why is your exam score in this class a random variable?

Topic 3: Discrete distributions
How many earthquakes are likely to happen in a random year?

Topic 4: Expectation and variance
How old are you expected to live?

Topic 5: Continuous distributions
How long are you expected to wait in the queue at a restaurant?

Topic 6: Limiting theorems
Why a lottery company never loses?

Topic 7: Sampling distribution
How do I know I am taller than an average person?

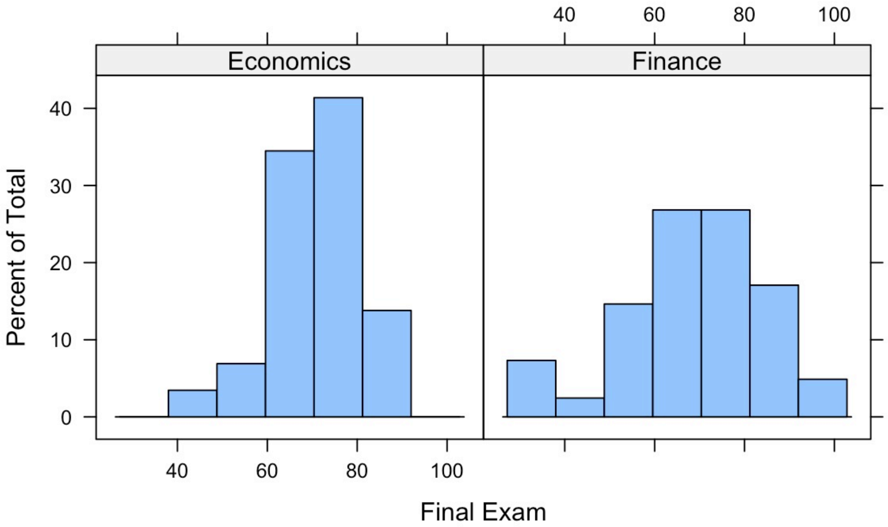
Assessment

Quiz (25%). There will be an arbitrary number of in-class quizzes. The date for each quiz will be announced in advance. Each quiz will consist of 1-2 questions based on material covered in previous weeks. Every quiz is mandatory; there will be no make-up quizzes under any circumstances.

Project (25%). The goal of the projects is to encourage students to apply the knowledge learned in this course to solve practical problems. Projects are usually open-ended and may involve data analysis, simulations, or exploring real-world applications of probability. Essays that present innovative perspectives and use the data persuasively to support their conclusions will receive higher marks. Selective students may be invited to present their findings to the class.

Final exam (50%). The final exam will be a closed-book, paper-and-pencil exam scheduled on Week 17. It will not simply repeat lecture material but will assess your ability to apply the knowledge you have gained to solve novel problems. To perform well, you must have a deep understanding of the concepts and acquire some degree of problem-solving skills. The average score of the past exam is 69 with a standard deviation 15. The pass rate (>=60) is about 80%.

Class participation (5%). Additional 5 marks for class participation on top of the above. Regular attendance and active participation in class discussions are encouraged (though not mandatory) and will be recognized.



Lecture notes

All lecture materials will be published through this online website. You are not required to read any textbook. For students who insist on a textbook, it would be DeGroot and Schervish's *Probability and Statistics (4th edition).*

It is recommended to use the textbook as a supplement not a replacement of the lecture note. For students who prefer to read the textbook. There are two key differences between this lecture note and the textbook. First, the sections are arranged differently. Second, the examples and exercises are entirely different despite the key definitions and theorems are the same.

Homework

There is no homework assignment in this course. We will do in-class exercises instead. However, problem solving is essential for learning math. You are encouraged to practice the exercises in DeGroot and Schervish's textbook after class. But it is not mandatory.

Statistical software

Statistical software is indispensable for modern statistics. For practical consideration, it is beneficial to start learn it as early as possible. We will demonstrate how to do statistics in R, which is a widely-used open-source statistical programming language. It is highly recommended that you try it yourself while learning this course.

Reference

1. Schervish, M. J., & DeGroot, M. H. (2014). *Probability and statistics.* Pearson Education.
2. Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability.* Chapman and Hall/CRC.
3. Grimmett, G., & Stirzaker, D. (2020). *Probability and random processes.* Oxford University Press.

Online playground

Probability Playground: Interactive Probability Distributions

Statistics: Unlocking the Power of Data

Exam score lookup

Student number

Find scores

Loading exam scores...

Copyright ©

Part I

Probability Basics

1 What is probability?

> Probability is the most important concept in modern science, especially as nobody has the slightest notion of what it means. —— B. Russell

What is probability? We all talk about probabilities in everyday life, but mostly in vague languages. This course is to introduce probability as a logical framework for quantifying uncertainty and randomness.

> Mathematics is the logic of certainty; probability is the logic of uncertainty. —— J. Blitzstein

The earliest development of probability is rooted in gambling. The famous Monte Carlo method in statistics, invented by Stanislaw Ulam in the late 1940s, takes its name from the *Monte Carlo Casino* in Monaco, where Ulam's uncle went to gamble.

Today, probability theory has been applied to almost every field of human knowledge. It is the foundation of statistics, machine learning, and artificial intelligence. It also plays a crucial role in everyday decision-making, from stock investments to effective strategies to combat an infectious disease.

The first formal definition of probability is often attributed to Pierre-Simon Laplace in the 18th century. In his work *Theorie analytique des probabilites*, published in 1812,

> The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

We will soon discover that this definition is obsolete. We start the journey of modern probability theory by introducing the basic concepts of events and sample spaces.

2 Event and sample space

Definition 2.1. We use *sets* to build the foundational concepts in probability:

- **Experiment**: a procedure with an uncertain outcome
- **Event** A: a set of possible outcomes
- **Sample space** S: the set of all possible outcomes

Anything (a gamble, an exam, a financial year, …) with uncertain outcomes can be an experiment. The sample space can be finite, countably infinite, or uncountably infinite. It is convenient to visualize events with *Venn diagrams*.

> ℹ Don's confuse events with outcomes
>
> Outcomes are individual results, while events are groups of outcomes that we are interested in. Outcomes are elements of the sample space, and events are subsets of this space.

Example 2.1. A coin is flipped twice. We write "H" if a coin lands Head, and "T" if a coin lands Tail.

- The sample space (all possible outcomes): $S = \{HH, HT, TH, TT\}$
- Let A_1 be the event that the first flip is Heads, $A_1 = \{HH, HT\}$
- Let A_2 be the event that the second flip is Heads, $A_2 = \{HH, TH\}$
- Let B be the event that at least one flip is Heads, $B = A_1 \cup A_2$
- Let C be the event that all the flips are Heads, $C = A_1 \cap A_2$
- Let D be the event that no flip is Heads, $D = B^c$

Use set language to describe events

English	Sets
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$

English	Sets
A or B occurs	$A \cup B$
Both A and B occur	$A \cap B$
A does not occur	A^c
at least one of $A_1, \dots, A_n$ occur	$A_1 \cup \dots \cup A_n$
all of $A_1, \dots, A_n$ occur	$A_1 \cap \dots \cap A_n$
A implies B	$A \subseteq B$

Definition 2.2. A and B are **disjoint** (mutually exclusive) if $A \cap B = \phi$.

Definition 2.3. $A_1, \dots, A_n$ are a **partition** of S if

- $A_1 \cup \dots \cup A_n = S$, and
- $A_i \cap A_j = \phi$ for $i \neq j$.

Simulating coin flipping

Randomly sampling from the set $\{H, T\}$:

```
# 10 draws with equal prob with replacement
sample(c('H', 'T'), 10, replace = T)
```

```
 [1] "H" "T" "T" "H" "T" "T" "T" "H" "T" "H"
```

Compute the probability of HH when tossing two coins:

```
# simulate coin tossing 10000 times
toss <- sample(c('H','T'), 10000, replace =T)

# group them into pairs
toss.pair <- paste0(toss[-length(toss)], toss[-1])

# number of HH
n_HH <- sum(toss.pair == 'HH')

# total number of tosses
n_total <- length(toss.pair)
```

```r
# compute the probability
prob <- n_HH / n_total

cat("Prob of HH: ", prob)
```

```
Prob of HH:  0.2465247
```

3 Classical probability

Definition 3.1. Classical (naive) definition of probability:

$$P(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes in A}}$$

assuming the outcomes are finite and equally likely.

> ℹ Don't misuse the classical definition
>
> The classical probability is very restrictive. It only applies to scenarios such as flipping coins or rolling dice where the outcomes are equally likely. It has often been misapplied by people who assume equally likely outcomes without justification. For example, if one wants calculate the probability of a rainy day, it would be misleading to assume every day is equally likely to rain and compute $\frac{\text{rainy days}}{365}$.

Calculating the naive probability of an event A often involves counting the number of outcomes in A and the number of outcomes in the sample space S, which usually involve some counting methods. We now review some of the counting methods (multiplications, factorials, permutations, combinations) that was introduced in high schools.

Counting methods

Multiplications. Consider a compound experiment consisting of two sub-experiments, Experiment A and Experiment B. Suppose that Experiment A has a possible outcomes, and for each of those outcomes Experiment B has b possible outcomes. Then the compound experiment has $a \times b$ possible outcomes.

Exponentiation. Consider n objects and making k choices from them, one at a time with replacement. Then there are n^k possible outcomes.

Factorials. Consider n objects $1, 2, \ldots, n$. A permutation of $1, 2, \ldots, n$ is an arrangement of them in some order, e.g., $3, 5, 1, 2, 4$ is a permutation of $1, 2, 3, 4, 5$. The are $n!$ permutations of $1, 2, \ldots, n$.

Permutations. Consider n objects and making k choices from them, one at a time without replacement. Then there are $P_n^k = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$ possible outcomes, for $k \le n$. (Ordering matters in this case, e.g. $1,2,3$ is considered different from $2,3,1$)

Combinations. Consider n objects and making k choices from them, one at a time without replacement, without distinguishing between the different orders in which they could be chosen (e.g. $1,2,3$ is considered no different from $2,3,1$). Then there are $C_n^k = \frac{n(n-1)\cdots(n-k+1)}{k!}$ possible outcomes. In modern math, we prefer the notation

$$\binom{n}{k} \equiv C_n^k,$$

which reads as "n choose k".

The following table summarizes the counting methods.

	order matters	order doesn't matter
with replacement	n^k	$\binom{n+k-1}{k}$
non-replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

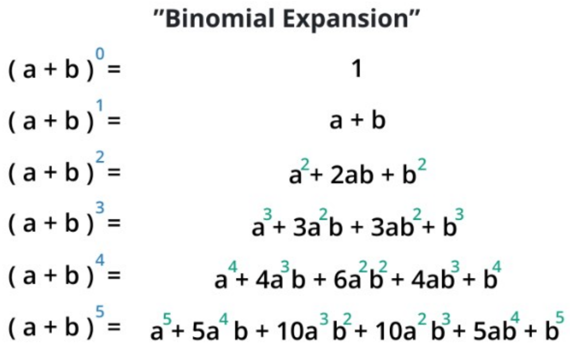
The upper-right corner case is equivalent to putting k indistinguishable balls into n distinguishable baskets (e.g. two balls in Basket 3 means the 3rd object is chosen twice). Therefore, the number of possible arrangements is $\binom{k+n-1}{n-1}$.

Binomial coefficient

The Binomial coefficient $\binom{n}{k}$ counts the number of subsets of size k for a set of size n. It is also the coefficient of x^k when expanding the polynomial $(x+y)^n$.

Theorem 3.1 (Binomial theorem)**.**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$



"Binomial Expansion"

$(a+b)^0 = 1$

$(a+b)^1 = a + b$

$(a+b)^2 = a^2 + 2ab + b^2$

$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$

$(a+b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$

The coefficients form an infinite triangle called the **Pascal triangle**. By observing the patterns in the triangle, it is not hard to conclude the following recursive formula:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

The value of the binomial coefficient is only defined for non-negative integers n and k with $0 \le k \le n$. But mathematics is about **generalization**. We can generalize the notion of "n choose k" for negative n:

$$\begin{aligned}\binom{-n}{k} &= \prod_{i=0}^{k-1} \frac{-n-i}{k-i} = (-1)^k \prod_{i=0}^{k-1} \frac{n+i}{k-i} \\ &= (-1)^k \frac{n(n+1)\dots(n+k-1)}{k!} \\ &= (-1)^k \frac{(n+k-1)!}{k!(n-1)!} \\ &= (-1)^k \binom{n+k-1}{k}\end{aligned}$$

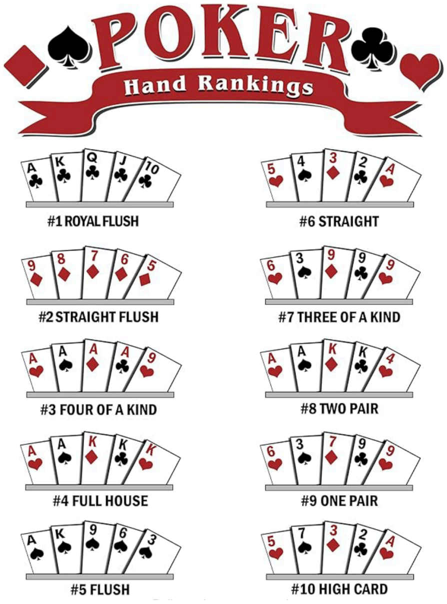
We can also extend the formula to real numbers and even complex numbers:

$$\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}$$

where $\Gamma(x+1) = x!$ is a generalization of factorials. We will come back to the Gamma function when we discuss Gamma distributions.

4 Gambling problems

Example 4.1. Texas hold'em is one of the most popular variant of the card game of poker. Essentially, the players in the game bet on the rankings of their hand of five cards (illustrated in the figure below). For the game to be fair, a hand of higher values must have lower probability than a hand of lower values. Compute the probability for each type of hand.



Solution. To apply Definition 3.1, we first determine the total number of possible five-card hands from a standard 52-card deck. The total number of combinations is:

$$\binom{52}{5} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2,598,960$$

We then find the number of combinations for each hand type. As an illustration, we only compute the case of *Full House*.

A full house consists of three cards of one rank and two cards of another rank. The number of ways to choose the rank for the triplet is $\binom{13}{1} = 13$, and the number of ways to choose 3 cards of that rank is $\binom{4}{3} = 4$. The number of ways to choose the rank for the pair is $\binom{12}{1} = 12$, and the number of ways to choose 2 cards of that rank is $\binom{4}{2} = 6$. Therefore, the total number of full houses is: $\binom{13}{1}\binom{4}{3} \cdot \binom{12}{1}\binom{4}{2} = 3,744$.

Thus, the probability is:

$$P(\text{Full House}) = \frac{3,744}{2,598,960} \approx 0.144\%$$

Example 4.2. (Newton-Pepys problem) The Newton-Pepys problem is a classical probability problem involving dice rolls, posed in correspondence between Samuel Pepys, a famous diarist and government official, and Isaac Newton in 1693. The problem concerns which of the following three events is most likely when rolling fair dice:

1. At least one 6 appears when 6 fair dice are rolled.
2. At least two 6's appear when 12 fair dice are rolled.
3. At least three 6's appear when 18 fair dice are rolled.

Solution. We compute the probabilities for each scenario:

1. Probability of at least one 6 in six rolls of a single die:

   The probability of not rolling a 6 in six rolls is:

   $$P(\text{no 6 in six rolls}) = \frac{5^6}{6^6}$$

   Thus, the probability of at least one 6 in six rolls is:

   $$P(\text{at least one 6}) = 1 - \frac{5^6}{6^6} \approx 0.67$$

2. Probability of at least two 6s in twelve rolls of a single die:

   We use the complement, finding the probabilities of getting 0 or 1 six. The probability of getting exactly 0 sixes in twelve rolls is similar as above. The probability of getting exactly 1 six is:

   $$P(1 \text{ six}) = \binom{12}{1}\frac{5^{11}}{6^{12}}$$

   Thus, the probability of at least two 6s is:

   $$P(\text{at least two 6s}) = 1 - P(0 \text{ six}) - P(1 \text{ six}) \approx 0.62$$

3. Probability of at least three 6s in eighteen rolls of a single die:

   Similarly, we calculate the complement, finding the probabilities of getting fewer than 3 sixes.

$$P(\text{at least three 6s}) = 1 - P(0\text{ six}) - P(1\text{ six}) - P(2\text{ sixes})$$

$$= 1 - \frac{5^{18} + \binom{18}{1}5^{17} + \binom{18}{2}5^{16}}{6^{18}} \approx 0.60$$

Thus, $P(\text{one 6 in 6 rolls}) > P(\text{two 6s in 12 rolls}) > P(\text{three 6s in 18 rolls})$.

Intuitively, this is because as the number of dice increases, the likelihood of matching higher thresholds does not keep pace with the probability of rolling individual sixes. This is perhaps contrary to most people's common sense: the more dice rolled, the more likely to see the sixes.

Find probability by simulation

Let's redo Example 4.1 by simulation.

```
# generate a deck of cards
deck.grid <- expand.grid(c(1:10,'J','Q','K','A'), c(' ',' ',' ',' '))

# convert the grid to a vector
deck <- do.call(paste0, deck.grid)

# total number of simulations
N <- 100000

# number of target hand
K <- 0

# for random generator
set.seed(1000)

for (j in 1:N) {

  # a random five-cards hand
  hand <- sample(deck, 5)

  # drop the color
  num <- substr(hand, 1, nchar(hand)-1)
  # Full House have only two distinguished numbers
  if ( length(unique(num)) == 2 ) {
    K <- K + 1
  }
}

# compute the probability
P <- K / N

sprintf("Prob of Full House: %.3f %%", P*100)
```

```
[1] "Prob of Full House: 0.146 %"
```

5 Birthday paradox

The birthday problem is a classic probability puzzle that demonstrates how likely it is for at least two people in a group to share the same birthday. While it might seem intuitive that the probability is low in small groups, the results are surprising.

Example 5.1. In a group of k people, what is the probability that at least two people share the same birthday? Assume (1) there are 365 possible birthdays; (2) birthdays are evenly distributed across the year; (3) people are equally likely to be born on any given day.

Solution. If $k > 365$, the probability is 1. Assume $k \leq 365$ for the rest. Instead of directly calculating the probability of at least two people sharing a birthday, we first compute the complementary probability, $P(\text{no match})$, where no two people in the group have the same birthday.

For the first person, there are 365 choices for their birthday. For the second person, to avoid a shared birthday, there are 364 remaining choices. For the third person, there are 363 choices, and so on. For k people, the total number of arrangements (no shared birthday) is:

$$365 \times 364 \times 363 \times \cdots \times (365 - k + 1)$$

The total number of possible arrangements (with or without shared birthdays) is 365^k. Thus, the probability of no shared birthday is:

$$P(\text{no match}) = \frac{365 \cdot 364 \cdots (365 - k + 1)}{365^k}$$

Thus, the probability of at least one matched birthday is:

$$P(\text{match}) = 1 - P(\text{no match}) = \begin{cases} 50.7\% & k = 23 \\ 70.6\% & k = 30 \\ 97.0\% & k = 50 \\ 99.9\% & k = 70 \end{cases}$$

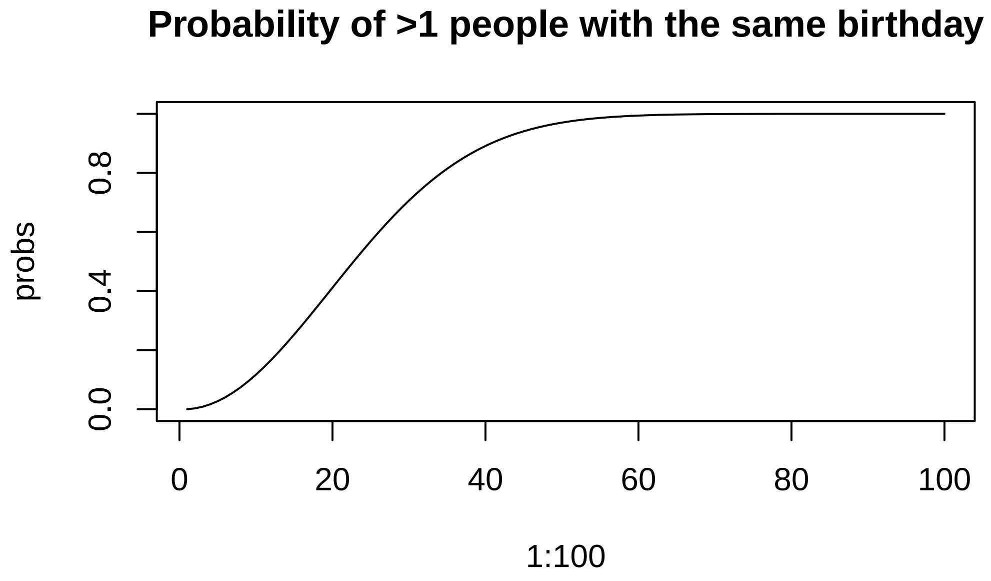
R simulation

```r
# a class of k people
k <- 30

# number of experiments
N <- 1000

# number of matches
m <- 0

for (i in 1:N) {

  # draw k random numbers from 1 to 365 with replacement
  birthdays <- sample(1:365, k, replace = T)

  # number of duplicated birthdays
  dups <- duplicated(birthdays)

  # if duplicated birthdays found
  if (any(dups)) m <- m + 1

}

cat("Prob of at least one match: ", m / N)
```

```
Prob of at least one match:  0.708
```

There is even a built-in function `pbirthday` that computes the probability of birthday coincidence. We may utilize this function to plot the probability as the number of people increases.

```r
# compute the probability of birthday match for 30 people
prob <- pbirthday(30)

# compute a vector of probabilities for 1,2...100 people
probs <- sapply(1:100, pbirthday)

# make a plot
plot(1:100, probs, type="l",main = "Probability of >1 people with the same birthday")
```


Probability of >1 people with the same birthday
probs
0.0
0.4
0.8
0
20
40
60
80
100
1:100

6 Axiomatic probability

Definition 6.1. A *probability space* consists of S and P, where S is a sample space, and P is a function which takes an event $A \subseteq S$ as input and returns $P(A) \in [0, 1]$ such that

- $P(\phi) = 0$,
- $P(S) = 1$,
- $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if $A_1, A_2, \dots, A_n$ are disjoint.

Note that this Definition does not imply any particular interpretation of probability. In fact, any function P that satisfies the axioms are valid "probabilities". Thus, the theories of probability do not depend on any particular interpretation. It is purely axiomatic. From the three axioms, we can derive any property of probabilities. The interpretation also matters, but it is more of a philosophical debate.

ℹ Two interpretations of probability

- 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.

Proposition 6.1. *For any events A and B, we have*

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

Proof. We prove the three properties by just using the Axioms.

1) Since A and A^c are disjoint and their union is S, apply the third axiom:

$$P(S) = P(A \cup A^c) = P(A) + P(A^c);$$

By the second axiom, $P(S) = 1$. So $P(A) + P(A^c) = 1$.

2) The key is to break up the set into disjoint sets. If $A \subseteq B$, then $B = A \cup (B \cap A^c)$ where A and $B \cap A^c$ are disjoint (draw a Venn diagram for intuition). By the third axiom, we have

$$P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c) \geq P(A).$$

3) We can write $A \cup B$ as the union of the disjoint set A and $B \cap A^c$. Then by the third axiom,

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

It suffices to show that $P(B \cap A^c) = P(B) - P(A \cap B)$. Since $B \cap A$ and $B \cap A^c$ are disjoint, we have

$$P(B) = P(B \cap A) + P(B \cap A^c).$$

So $P(B \cap A^c) = P(B) - P(A \cap B)$ as desired.

□

The last property is a very useful formula for finding the probability of a union of events when the events are not necessarily disjoint. We can generalize it to n events.

Theorem 6.1. *For any events $A_1, A_2, \dots, A_n$, it holds that*

$$P(A_1 \cup A_2 \cdots \cup A_n) = \sum_{j=1}^{n} P(A_j) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots$$
$$(-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

This formula can be proved by induction using the axioms. Below is a famous application (known as de Montmort's problem, named after French mathematician Pierre Remond de Montmort) of the inclusion-exclusion theorem.

Example 6.1 (Matching problem)**.** Suppose there are n people who each check in a hat at a party. The hats are randomly returned to them without any concern for whose hat is whose. What is the probability that at least one person gets their own hat back?

Solution. Let A_j be the event: the j-th person gets his own hat. The problem is equivalent to find $P(A_1 \cup A_2 \cup \cdots \cup A_n)$.

Since all position are equally likely, $P(A_j) = \frac{1}{n}$. The probability of there being two matches is: $P(A_1 \cap A_2) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$. Similarly, the probability of there being k matches is:

$P(A_1 \cap \cdots \cap A_k) = \frac{(n-k)!}{n!} = \frac{1}{n(n-1)\cdots(n-k+1)}$. Using the property of the union of events,

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) =& n \cdot \frac{1}{n} - \binom{n}{2} \frac{1}{n(n-1)} + \binom{n}{3} \frac{1}{n(n-1)(n-2)} - \cdots \\ =& 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{n+1} \frac{1}{n!} \\ \approx& 1 - \frac{1}{e}. \end{aligned}$$

Pattern matching with Taylor series

Pattern matching is a very useful technique. In the last step, we recognize that the Taylor series of e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Therefore, $e^{-1} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots$

Example 6.2 (Infinite monkey theorem)**.** A monkey hitting keys independently and at random on a typewriter keyboard for an infinite amount of time will almost surely type any given text (e.g. the complete works of William Shakespeare). In other words, *an infinite random sequence of letters contain every finite string infinitely often with probability 1.*

Proof. Let's compute the probability of the monkey typing "banana" correctly with random strokes. Suppose there are 50 keys on the keyboard. The monkey typed correctly "banana" is $\frac{1}{50^6}$. Suppose the monkey tried n times, the probability that he did not typed the correct text is

$$X_n = \left(1 - \frac{1}{50^6}\right)^n$$

For finite n (even n is very large), X_n would be very close to 1. For example, when $n = 10^6$, $X_n \approx 0.9999$. But as $n \to \infty$, $X_n \to 0$. That means, if n is infinitely large, the probability that the monkey produced the correct text goes to 1.

The theorem reminds us that infinite limits behave very differently from large finite numbers. It is also used as a metaphor: given enough time, randomness can generate order, structure, or meaning. □

7 Conditional probability

The probability of A **conditioned on** B is the updated probability of event A after we learn that event B has occurred. Since events contain information, the occurring of a certain event may change our believes on probabilities of other relevant events. The updated probability of event A after we learn that event B has occurred is the conditional probability of A given B.

Definition 7.1. If A and B are events with $P(B) > 0$, then the conditional probability of A given B is defined as

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

> Don't confuse P(A|B) with P(A,B)
>
> $P(A|B)$ is the probability of A occurring given that B has already occurred. While $P(A, B) = P(A \cap B)$ is the probability that A and B occur simultaneously.

Proposition 7.1. *Properties of conditional probability:*

- $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$
- $P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1 \dots A_{n-1})$

Theorem 7.1 (Bayes' rule)**.** *Assume* $P(B) > 0$*, we have*

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

> Thomas Bayes and "causes and effects"
>
> The Bayes' rule is named after Thomas Bayes (18th century) who wanted to know how to infer causes from effects. Human intelligence wants to know the cause given its effects. However, we are only able to observe the effects given the cause. Here is Bayes' reasoning. Suppose we have a *prior* belief about the cause of something we want to learn. We may not be able to learn the true cause directly, but after we observe its effects (the *Data*), we would update our belief based on the new information we have learned from the data.

The updated belief (the *posterior*) is therefore somewhat closer to the “truth”.

$$\underbrace{P(\text{Belief} \mid \text{Data})}_{\text{Posterier}} = \frac{\overbrace{P(\text{Data} \mid \text{Belief})}^{\text{Likelihood}} \overbrace{P(\text{Belief})}^{\text{Prior}}}{P(\text{Data})}$$

Theorem 7.2 (Law of total probability (LOTP))**.** *Let $B_1, ..., B_n$ be a partition of the sample space S (i.e., the B_i are disjoint events and their union is S), with $P(B_i) > 0$ for all i. Then*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Theorem 7.3 (Conditional version of LOTP)**.** *The law of total probability has an analog conditional on another event C, namely,*

$$P(A|C) = \sum_{i=1}^{n} P(A|B_i \cap C)P(B_i|C).$$

Example 7.1. Get a random 2-card hand from a standard deck. Find the probability of (a) Both cards are aces given that at least one of them (not necessarily the first one) is an ace; (b) Getting another ace given the first draw is an ace of spade.

Solution. The example shows the subtleness of conditional probabilities. The seemingly indifferent probabilities are in fact different:

$$\begin{aligned} P(\text{two aces} \mid \text{one ace}) &= \frac{P(\text{both aces})}{P(\text{one ace})} \\ &= \frac{\binom{4}{2}/\binom{52}{2}}{1 - \binom{48}{2}/\binom{52}{2}} \\ &= \frac{1}{33}; \end{aligned}$$

$$\begin{aligned} P(\text{another ace} \mid \text{ace of spade}) &= \frac{P(\text{ace of spade \& another ace})}{P(\text{ace of spade})} \\ &= \frac{\binom{3}{1}/\binom{52}{2}}{\binom{51}{1}/\binom{52}{2}} \\ &= \frac{1}{17}. \end{aligned}$$

Note that, in the first case, the denominator is interpreted as “at least one ace”; whereas in the second case, it is “ace of space + another card”.

Example 7.2. A disease has a prevalence rate of 10% (i.e., the probability of being infected is 10%). A diagnostic test for the disease has an accuracy of 98%, meaning it correctly identifies infected individuals as positive 98% of the time. Calculate the probability that an individual is infected given that the test result is positive.

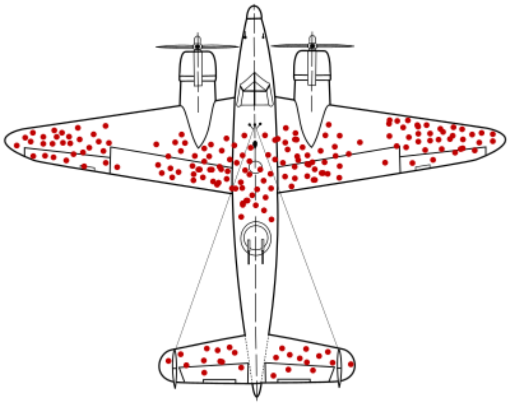
Solution. Let D denote being actually infected by the disease; and T denote a positive test. The test accuracy means: $P(T|D) = 98\%$. It also means $P(T|D^C) = 2\%$. We also know that $P(D) = 0.1$. We want to find $P(D|T)$. Note they are two different conditional probabilities, though we mostly confuse the two in everyday life. The two conditional probabilities are associated with Bayes' rule:

$$\begin{aligned}P(D|T) =& \frac{P(T|D)P(D)}{P(T)} \\ =& \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)} \\ =& \frac{0.98 \times 0.1}{0.98 \times 0.1 + 0.02 \times 0.9} \approx 84\%.\end{aligned}$$

Note that how $P(T|D)$ is far away from $P(D|T)$!

Thinking conditionally

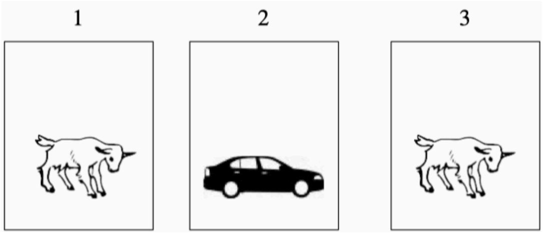
Abraham Wald, the renowned statistician, was hired by the Statistical Research Group (SRG) at Columbia University to figure out how to minimize the damage to bomber aircraft. The data they had comprised aircraft returning from missions with bullet holes on their bodies. If asked which parts of the aircraft should be armored to enhance survivability, the obvious answer seemed to be to armor the damaged parts. However, Wald suggested the exact opposite—to armor the parts that were not damaged. Why? Because the observed damage was conditioned on the aircraft returning. If an aircraft had been damaged on other parts, it likely would not have returned. Thinking conditionally completely changes the answer!



> See The Soul of Statistics by Professor Joseph Blitzstein.

8 Monty Hall problem

Example 8.1 (Monty Hall problem)**.** Suppose you are on Monty Hall's TV show. There are three doors. One of them has a car behind it. The other two doors have goats. Monty knows which one has the car. Monty now asks you to pick one door. You will win whatever is behind the door. After you pick one door. Monty opens another door that shows a goat. Monty then asks you if you want to switch. Is it optimal to switch?



We present two solutions to the problem. The first one is using the law of total probability. Let S: succeed assuming switch; D_j: door j has the car, $j \in 1, 2, 3$. Without loss of generality, assume the initial pick is Door 1. Monty will always open the door with a goat. By the law of total probability,

$$\begin{aligned} P(S) = & \underbrace{P(S|D_1)}_{\text{switch from initial pick}} P(D_1) + \underbrace{P(S|D_2)}_{\text{Monty opens door 3}} P(D_2) + \underbrace{P(S|D_3)}_{\text{Monty opens door 2}} P(D_3) \\ = & 0 + 1 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

The problem can also be solved using the Bayes' rule. Let D_j: door j has the car; M_j: Monty opens door j, $j \in 1, 2, 3$. Assume the initial pick is Door 1. If Monty opens door 3, the probability of winning the car assuming switching is

$$\begin{aligned} P(D_2|M_3) = & \frac{P(M_3|D_2)P(D_2)}{P(M_3)} \\ = & \frac{P(M_3|D_2)P(D_2)}{P(M_3|D_1)P(D_1) + P(M_3|D_2)P(D_2) + P(M_3|D_3)P(D_3)} \\ = & \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0} = \frac{2}{3}. \end{aligned}$$

Note that, if door 1 has the car, Monty will open door 2 and 3 with equal probability, thus $P(M_3|D_1) = \frac{1}{2}$. And Monty will never open the door with the car, therefore $P(M_3|D_3) = 0$. Similarly, if Monty opens door 2, we have $P(D_3|M_2) = \frac{2}{3}$. Therefore, the optimal choice is always to switch. Intuitively, because Monty knows which door has the car, the fact that he always opens the door without the car gives additional information regarding the choice of the door.

R simulation

```r
# Number of simulations
N <- 1000

# Number of total wins
W <- 0

# Prizes offered
prizes <- c('Car', 'Goat 1', 'Goat 2')

for (i in 1:N) {

  # Prizes are in random order behind the doors
  doors <- sample(prizes, 3)

  # Guest picks a random door
  pick <- sample(doors, 1)

  # Monty opens the door that is not picked by the guest
  # nor does it has the Car behind it
  open <- sample(setdiff(doors, c(pick, 'Car')), 1)

  # If the Guest swithes, he chooses the door that is not
  # his initial pick nor the one opened by Monty
  switch <- setdiff(doors, c(pick, open))

  # The guest wins if his final choice has the Car
  win <- switch == 'Car'

  # Increase the winning counter
  W <- W + win
}
```

```
# Frequency of winning the game if always switching
W/N
```

```
[1] 0.654
```

9 Simpson's paradox

Example 9.1. (Simpson's paradox). There are two doctors, Dr. Lee and Dr. Wong, performing two types of surgeries — heart surgery (hard) and band-aid removal (easy). Dr. Lee has higher overall surgery success rate. Is Dr. Lee necessarily a better doctor than Dr. Wong?

No. Consider the following example:

	Dr. Lee			Dr. Wong		
	Heart	Band-Aid	Total	Heart	Band-Aid	Total
Success	2	81	83	70	10	80
Failure	8	9	17	20	0	20
Success rate	20%	90%	83%	78%	100%	80%

The truth is Dr. Lee has overall higher success rate because he only does easy surgeries (band-aid removal). Dr. Wong does mostly hard surgeries and thus has a lower overall success rate. Yet, he is better at each single type of surgery. To formalize the argument, let S: successful surgery; D: treated by Dr. Lee, D^c: treated by Dr. Wong; E: heart surgery, E^c: band-aid removal. Dr. Wong is better at each type of surgery,

$$P(S|D, E) < P(S|D^c, E)$$
$$P(S|D, E^c) < P(S|D^c, E^c);$$

But, Dr. Lee has a higher overall successful rate,

$$P(S|D) > P(S|D^c).$$

This is because there is a "confounder" E:

$$P(S|D) = \underbrace{P(S|D, E)}_{<P(S|D^c,E)} \underbrace{P(E|D)}_{\text{weight}} + \underbrace{P(S|D, E^c)}_{<P(S|D^c,E^c)} \underbrace{P(E^c|D)}_{\text{weight}}.$$

A **confounder** is a variable that influences with both explanatory variable and the outcome variable, which therefore "confounds" the correlation between the two. In our example, the type of surgery (E) is associated with both the doctor and the outcome. Without the confounder being controlled, it is impossible to draw valid conclusions from the statistics.

In general terms, Simpson's paradox refers to the paradox in which a trend that appears across different groups of aggregate data is the reverse of the trend that appears when the aggregate data is broken up into its components. It is one of the most common sources of statistical misuse. Here is another example.

Example 9.2. (UC Berkeley gender bias). One of the best-known examples of Simpson's paradox comes from a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Male		Female	
	Applicants	Admitted	Applicants	Admitted
Total	8,442	44%	4,321	35%

However, when taking into account the information about departments being applied to, the conclusion turns to the opposite: in most departments, the admission rate for women is higher than men. The lower overall admission rate is caused by the fact that women tended to apply to more competitive departments with lower rates of admission, whereas men tended to apply to less competitive departments with higher rates of admission.

Department	Male		Female	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
Total	2691	45%	1835	30%

See https://setosa.io/simpsons for a really good illustration of the Simpson's paradox.

R demostration

```
# R has a built-in dataset `UCBAdmissions`
# we convert it to data frame for analysis
data <- as.data.frame(UCBAdmissions)
```

```
# browse the first a few rows
head(data)
```

```
     Admit Gender Dept Freq
1 Admitted   Male    A  512
2 Rejected   Male    A  313
3 Admitted Female    A   89
4 Rejected Female    A   19
5 Admitted   Male    B  353
6 Rejected   Male    B  207
```

```
# subset of the data with only admissions
data <- subset(data, Admit == 'Admitted')

# number of admissions by Gender
aggregate(Freq ~ Gender, data = data, FUN = sum)
```

```
  Gender Freq
1   Male 1198
2 Female  557
```

```
# number of admissions by Gender and Department
aggregate(Freq ~ Gender + Dept, data = data, FUN = sum)
```

```
   Gender Dept Freq
1    Male    A  512
2  Female    A   89
3    Male    B  353
4  Female    B   17
5    Male    C  120
6  Female    C  202
7    Male    D  138
8  Female    D  131
9    Male    E   53
10 Female    E   94
11   Male    F   22
12 Female    F   24
```

ℹ The importance of conditional thinking

Whenever we talk about probability or statistics, always remind ourselves what we are the conditioning on. Any statistical reasoning without specifying the conditions can be misleading. We are prone to such fallacies everyday everywhere.

- “10 millions new jobs were added during the term of President X.” But it doesn’t tell you this was achieved conditioned on that the economy had just had the worst recession.
- “Private schools’ graduates earned 50% more than those graduated from public schools.” But it doesn’t tell you the background of those students who enrolled in private schools.

Be vigilant to these claims when you see them next time.

10 Independence

Definition 10.1 (Independence for two events)**.** If event B's occurrence does not change the probability of A, then we say A and B are independent. That is to say A and B are independent if

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

Assuming $P(B) > 0$, this is equivalent to

$$P(A|B) = P(A)$$

Theorem 10.1. *If events A and B are independent, then*

- *A and B^c are independent;*
- *A^c and B^c are independent.*

A and B are independent means they do not provide information to each other in the sense that conditional probability is not different from the unconditional probability. It is not an intuitive idea as it seems. It will become clearer when we discuss random variables in later chapters.

Definition 10.2 (Independence for three events)**.** Events A, B, and C are said to be (mutually) independent if all of the following equations hold:

$$\begin{aligned}
P(A \cap B) =& P(A)P(B), \\
P(A \cap C) =& P(A)P(C), \\
P(B \cap C) =& P(B)P(C), \\
P(A \cap B \cap C) =& P(A)P(B)P(C).
\end{aligned}$$

Definition 10.3 (Independence for n events)**.** For n events $A_1, A_2, \dots, A_n$ to be (mutually) independent, we require any pair to satisfy $P(A_i \cap A_j) = P(A_i)P(A_j)$ (for $i \neq j$), any triplet to satisfy $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$ (for i, j, k distinct), and similarly for all quadruplets, quintuplets, and so on.

Don’t assume independence without justification

Independence provides lots of nice properties, which are not necessarily true without independence. A common mistake is to assume independence without justification. Be careful when you apply properties that assume independence.

Common confusions

Independence is not the same as disjointness.

A and B are disjoint means if A occurs, B cannot occur. But independence means A occurs has nothing to do with B.

Pairwise independence does not imply independence.

In Definition 10.2, If the first three conditions hold, we say that A, B, and C are *pairwise independent*. Pairwise independence does not imply independence. Convince yourself with the following example.

Example 10.1. Consider two fair, independent coin tosses, and let A be the event that the first is Heads, B the event that the second is Heads, and C the event that both tosses have the same result. Show that A, B, and C are pairwise independent but not independent.

Solution. For each event, $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$. Consider the two events together, there are four possible outcomes: HH, HT, TH, TT. $P(C) = P(HH) + P(TT) = \frac{1}{2}$. Thus,

$$\begin{aligned}
P(A \cap B) = P(HH) = \frac{1}{4} = P(A)P(B) \\
P(A \cap C) = P(HH) = \frac{1}{4} = P(A)P(C) \\
P(B \cap C) = P(HH) = \frac{1}{4} = P(B)P(C)
\end{aligned}$$

But A, B, C are not independent, because

$$P(A \cap B \cap C) = P(HH) = \frac{1}{4} \neq P(A)P(B)P(C).$$

11 Optimal mating problem*

The question

How many people should you date before you settle down with someone for marriage? The answer is you should date 37% of your potential options and choose the next one who is better. [1]

> The 37% rule
>
> The 37% Rule, also known as the **Optimal Stopping Theory**, provides a strategy to maximize the chances of making the best choice when faced with a sequence of options where decisions are irreversible. It suggests that you should review and reject the first 37% of the total options without selecting any, then choose the next option that is better than all those previously considered.

Mathematical framework

Let's assume there's a pool of N people out there from which you are choosing. We'll also assume that you have a clear-cut way of rating people. You know who is the best to be your partner. We will call that person Mr/Ms X. The people that you will meet show up one by one in random order. X may show up anywhere in the sequence. Sadly, a person you have dated and then rejected isn't available to you any longer later on. So you cannot date all of them and pick the best one.

Your dating strategy is to date M of the N people and then settle with the next person who is better. Our task is to find the optimal M. If M is too small, you will likely land with someone before X shows up. If M is too large, X will likely pass X and pick someone less optimal. Of course, there is no perfect solution. We want to find the M that maximizes the probability of landing X.

Let $P(M,N)$ be the probability of successfully picking X if you date M people out of N and then go for the next person who is better than the previous ones. Let S be the event of successfully picking X, and X_j means X is in the jth position in the sequence. The overall probability is:

$$P(M,N) = P(S|X_1)P(X_1) + P(S|X_2)P(X_2) + \cdots + P(S|X_n)P(X_n)$$

[1] See Kissing the frog: A mathematician's guide to mating and Strategic dating: The 37% rule for reference.

For a given value of M, if X is among the first M people you date, then you have missed your chance. The probability of settling with X is zero. Therefore, the first M terms are all zero.

If X is in $M+1$, you're in luck: since X is better than all others so far, you will pick X for sure. Therefore,

$$P(S|X_{M+1})P(X_{M+1}) = 1 \cdot P(X_{M+1}) = \frac{1}{N}$$

Since X is equally likely to be in any position, the probability of X being in $M+1$ out of N people is $1/N$.

If X is in $M+2$, you'll pick him/her up as long as the $(M+1)$st person didn't have a higher rating than all the previous M people. In other words, you would pass the $(M+1)$st person and pick X if the best one out of the $(M+1)$ people has shown up among the first M people. The change is $M/(M+1)$. Thus,

$$P(S|X_{M+2})P(X_{M+2}) = \frac{M}{M+1}\frac{1}{N}$$

Similarly, if X shows up in $M+3$, you'll pick him/her up to as long as neither the $(M+1)$st nor the $(M+2)$nd person have a higher rating than all the previous M people. In other words, the best one out of the first $(M+2)$ people has to show up among the first M people. The chance is $M/(M+2)$. Thus,

$$P(S|X_{M+3})P(X_{M+3}) = \frac{M}{M+2}\frac{1}{N}$$

Putting them all together, we have

$$\begin{aligned} P(M,N) &= \frac{1}{N} + \frac{M}{N(M+1)} + \frac{M}{N(M+2)} + \cdots + \frac{M}{N(N-1)} \\ &= \frac{M}{N}\left(\frac{1}{M} + \frac{1}{M+1} + \frac{1}{M+2} + \cdots + \frac{1}{N-1}\right) \end{aligned}$$

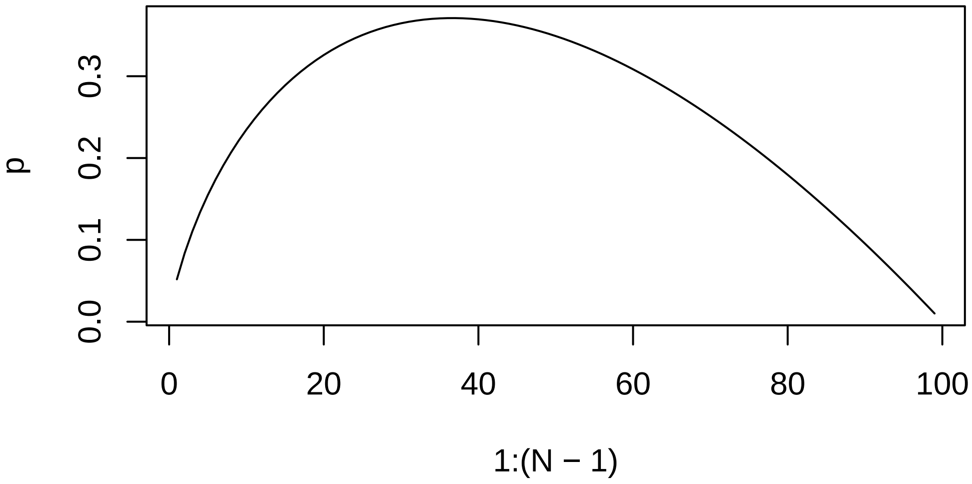
Maximizing your chance of success

Assuming $P(M,N)$ is strictly concave (fortunately this is the case), the M the maximizes the chance satisfies

$$P(M-1,N) < P(M,N) \text{ and } P(M+1,N) < P(M,N)$$

We can ask the computer to find the solution:

```r
N <- 100
p <- sapply(1:(N-1), function(m) m/N*sum(1/seq(m, N-1)))
plot(1:(N-1), p, type="l")
```



For $N = 100$, the highest probability if achieved at $M = 37$.

The limiting solution

We can find the solution analytically if M, N are large. For large n, the harmonic sequence can be approximated by the logarithm function:

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \ln(n) + \gamma$$

where γ is a constant.

We rewrite the function $P(M, N)$ as

$$P(M,N) = \sum_{k=M}^{N-1} \frac{M}{N} \cdot \frac{1}{k} = \frac{M}{N} \sum_{k=M}^{N-1} \frac{1}{k} = \frac{M}{N}\Big(H_{N-1} - H_{M-1}\Big)$$

For large M and N, it is approximated by

$$P(M,N) = \frac{M}{N}(H_{N-1} - H_{M-1}) \approx \frac{M}{N}\left[\ln(N-1) - \ln(M-1)\right] \approx \frac{M}{N} \ln\left(\frac{N}{M}\right)$$

Let $x = \frac{M}{N} \in (0, 1)$. We want to maximize

$$f(x) = x \ln\left(\frac{1}{x}\right) = -x \ln x$$

Differentiate:

$$f'(x) = \ln\left(\frac{1}{x}\right) - 1 = -\ln x - 1$$

Set $f'(x) = 0 \Rightarrow -\ln x - 1 = 0 \Rightarrow \ln x = -1 \Rightarrow x = e^{-1}$. Second derivative $f''(x) = -1/x < 0$, so it's a maximum.

Therefore, the maximizing fraction satisfies

$$\frac{M^\star}{N} \longrightarrow \frac{1}{e} \quad \text{as } N \to \infty,$$

and the maximal success probability tends to

$$f(1/e) = \frac{1}{e}.$$

12 Review of calculus*

Calculus is a prerequisite to work with continuous distributions. The following chapters assume readers are proficient in calculus. We nonetheless review some basic concepts here as a warm-up. This review is not exhaustive, so please refer to a specific textbook if needed for a more comprehensive understanding.

Differentiation

We define the derivative of a function $f(x)$ to be

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$

Loosely speaking, a function is continuous if there is no jump in the graph, differentiable if the curve is smooth. Some commonly used derivatives:

$$\begin{aligned}
\frac{d}{dx}(x^n) &= nx^{n-1} \\
\frac{d}{dx}(e^x) &= e^x \\
\frac{d}{dx}(\ln(x)) &= \frac{1}{x} \\
\frac{d}{dx}(\sin(x)) &= \cos(x) \\
\frac{d}{dx}(\cos(x)) &= -\sin(x) \\
(fg)' &= f'g + fg' \\
\left(\frac{f}{g}\right)' &= \frac{f'g - fg'}{g^2} \\
[f(g(x))]' &= f'(g(x))g'(x)
\end{aligned}$$

When dealing with limits of the form "$\frac{0}{0}$" or "$\frac{\infty}{\infty}$", the L'Hospital rule is very handy.

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

One important application of derivatives is the Taylor's theorem, which gives the approximation of a function around a given point by polynomials. Assume function f is at least k times differentiable, then

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \cdots$$

which means we can approximate a function arbitrarily well by higher order polynomials. Some commonly used Taylor series (expanding around $a = 0$):

$$\begin{aligned}
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots \quad \text{for } |x| < 1 \\
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\
\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad \text{for } |x| < 1 \\
\arctan(x) &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \quad \text{for } |x| \le 1
\end{aligned}$$

Approximating π with Taylor series

Taylor series are one of the most amazing results in calculus. For example, in the last formula, if we let $x = 1$:

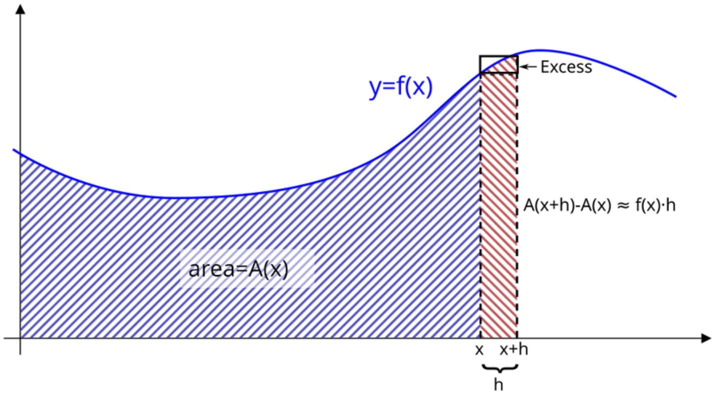
$$\frac{\pi}{4} = \arctan(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

Therefore, we can approximate π by summing up a sequence of fractions:

$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\right).$$

Integration

Integration is the inverse operation of differentiation. Integral has the geometric interpretation as the area under the curve. Let $A(x)$ be the area under the curve of $y = f(x)$. Thus $A(x) = \int_0^x f(t)dt$. The change of the area resulted from a tiny little change of x is approximated by $A(x+h) - A(x) \approx f(x)h$. That is $\frac{A(x+h)-A(x)}{h} = f(x)$. If the change is infinitesimal, $h \to 0$, we have $A'(x) = f(x)$.



The Fundamental Theorem of Calculus: if F is the anti-derivative of f, then

$$F(x) = \int_a^x f(t)dt$$

$$\int_a^b f(x)dx = F(b) - F(a)$$

One interpretation of the integral is — the integral of a rate of change of a quantity gives the net change in that quantity. Think about speed and distance: $\int_a^b v(t)dt = s(b) - s(a)$.

Because the integral is just a sum over infinitely many approximating rectangles, $\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^n f(x_i)\Delta x$. Integrals behave just like sums. For example, $\frac{1}{b-a}\int_a^b f(x)dx$ has the interpretation of the average of $f(x)$ from a to b.

Indefinite integrals are the general antiderivatives without specifying the interval of the integration. It always comes with a constant C. Some commonly used integrals:

$$
\begin{aligned}
\int dx \quad &= x + C \\
\int x^n dx \quad &= \frac{x^{n+1}}{n+1} + C \\
\int e^x dx \quad &= e^x + C \\
\int \frac{1}{x} dx \quad &= \ln|x| + C \\
\int \cos(x) dx \quad &= \sin(x) + C \\
\int \sin(x) dx \quad &= -\cos(x) + C \\
\int \frac{1}{1+x^2} dx \quad &= \arctan(x) + C
\end{aligned}
$$

Two common integration techniques are *substitution* and *integration by parts.*

Example 12.1 (Integration by substitution)**.** Find $\int \sqrt{3x+2}dx$.

Solution. Let $u = 3x + 2$, then $du = 3dx$. Then

$$\int \sqrt{3x+2}dx = \frac{1}{3}\int \sqrt{u}du = \frac{2}{9}u^{3/2} + C = \frac{2}{9}(3x+2)^{3/2} + C.$$

Example 12.2 (Integration by parts)**.** Find $\int x \sin x dx$.

Solution. Integration by parts follows the formula:

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx$$

Let $f(x) = x$, $g'(x) = \sin x$. Then $g(x) = -\cos x$. Then,

$$\int x \sin x dx = -x\cos x - \int (-\cos x)dx = -x\cos x + \sin x + C.$$

13 R tutorial*

Variables

```r
a = 5
b <- 5
c <- "Hello"
```

Assignment

```r
a <- a + 1 # assignment
a == a + 1 # math equal
```

Vectors

```r
u <- c(1,2,3,4,5)
v <- 6:10
b <- c('good', 'night', ' ')
```

Matrices

```r
A <- matrix(c(1,2,3,4), nrow = 2, ncol = 2)
B <- matrix(c(5,6,7,8), nrow = 2, ncol = 2)
```

Linear algebra

```r
u * v   # element-wise
u %*% v # dot product
A * B   # element-wise
A %*% B # matrix multiplication
t(A)    # transpose
det(A)  # determinant
```

Random numbers

```
# a random number from 0 to 100
runif(1, 0, 100)

# generate 10 random numbers
runif(10, 0, 100)

# random sampling
sample(1:100, 10)
```

Conditional statement

```
# draw a random integer
x <- sample(1:100, 1)

# if the remainder divided by 2 is 0
if (x %% 2 == 0) {

  # display it is an even number
  print("even number")

  } else {

    # otherwise, it is an odd number
    cat("odd number")

}
```

Loop

```
# for-loop
# loop for a given number of times
for (k in 1:10) {

  # the code to be repeated
  print("Hello!")

}

# while-loop
# loop on condition
k <- 0
while (k < 10) {
  # the code to be repeated
  print("Hello!")

  # keep track of the
  k <- k + 1
}
```

```
# 1+2+...+100=?
s = 0;
k = 1;
while (k <= 100) {
  s <- s + k;
  k <- k + 1;
}
```

Functions

```
# built-in functions
sin(pi/2)
log(100)
exp(2.3)
```

```
# custom functions
square <- function(a) {
  a * a
}

# call the function
square(10)
```

```
# function as a reusable code block
seqsum <- function(begin, end) {
  s = 0;
  k = begin;
  while (k <= end) {
    s <- s + k;
    k <- k + 1;
  }
  return(s)
}
```

```
# function call
seqsum(1, 100)
```

Part II

Random Variables

14 What is a random variable?

In the previous chapter, we have been working with *events*, which is a conceptualization of real world outcomes occurred with probabilities. In this chapter, we introduce a much more powerful conceptualization that deals with uncertain outcomes — random variables, which is the foundation of all probability and statistical studies.

Informally, a random variable differs from a normal variable as it is "random". A random variable, say X, is never associated with a certain value. It could different values *probabilistically.* For example, X may take the value 1 with probability 0.4, and take the value 2 with probability 0.3. The formal definition of a random variable is as follows.

Numeric encoding of events

Definition 14.1 (Random variable)**.** Given an experiment with sample space S, a random variable is a function from the sample space S to the real numbers $\mathbb{R}$.

As an example, flipping a coin twice, let X be the number of heads. Then $X(\cdot)$ is a functions that maps events in $\{HH, HT, TH, TT\}$ into real numbers. In our case, the mapping goes like

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.$$

X is therefore an encoding of events in the sample space into real numbers. We could, of course, have different encoding. Consider the random variable Y as the number of tails. Then we have $Y = 2 - X$.

$$Y(HH) = 0, Y(HT) = 1, Y(TH) = 2, Y(TT) = 2.$$

We could also define Z as the number heads in the 1st toss only. The encoding goes like

$$Z(HH) = 1, Z(HT) = 1, Z(TH) = 0, Z(TT) = 0.$$

We have listed three ways of "encoding" the same experiment as random variables. All of them are valid random variables, but they map the outcomes into different numbers. We can say that, a random variable is a numeric "summary" of an aspect of an experiment.

Notation for random variables

We usually use capital letters, such as X, Y, Z, to denote random variables. We use small letters, such as x, y, z, to denote specific values. $P(X = x)$ means the probability of X taking the value x. Don't confuse the random variable X with the number x.

Don't confuse random variables, numbers, and events

Random variables are never fixed numbers. Functions of random variables, such as X^2, $|X|$, e^X, are also random variables. Random variables are not events. It does not make sense to write $P(X)$, because X is not an event. But $X = a$ is an event, it makes sense to write $P(X = a)$.

Definition 14.2 (Distribution)**.** Let X be a random variable. The distribution of X is the collection of all probabilities of the form $P(X \in C)$ for all sets C of real numbers such that $\{X \in C\}$ is an event.

A distribution specifies the probabilities associated with all values of a random variable. In the above example, the distribution of X is given by

$$P(X = 0) = \frac{1}{4}, P(X = 1) = \frac{1}{2}, P(X = 2) = \frac{1}{4}.$$

The distribution of Y is given by

$$P(Y = 0) = \frac{1}{4}, P(Y = 1) = \frac{1}{2}, P(Y = 2) = \frac{1}{4}.$$

The distribution of Z is given by

$$P(Z = 0) = \frac{1}{2}, P(Z = 1) = \frac{1}{2}.$$

You may have noted that the probabilities in a distribution always sums up to 1, as all possible events constitute the entire sample space.

Specifying the distribution

Listing all the values is not a smart way to specify a distribution. We like to use a function (if possible), such as $f(x) \overset{?}{=} e^{-x}$, to specify the probability of a random variable X taking the value x. This is convenient, because once we know the function, we know all the probabilities. But how to specify this function depends on whether a random variable is discrete or continuous.

Conceptualization of uncertain outcomes

Many real-world processes have uncertain outcomes. For example, the outcome of tossing a coin or the temperature of tomorrow. In many applications like this, we simply do not have perfect information to predict the future with certainty. In such cases, we model the uncertain outcome as an RV, which takes uncertain values with probabilities. The exact distribution of many applications may be unknown, but we can approximate it with frequencies observed in samples.

Experiment:	**Tossing a coin**	
	Conceptualization	**Observations**
Random variable	X with support $\{0, 1\}$	$\{0, 1, 1, 0, 0, 1, ...\}$
Distribution	$P(X = i) = 0.5, i \in \{0, 1\}$	Proportion of 1s $= 0.45$
Experiment:	**Taking an exam**	
	Conceptualization	**Observations**
Random variable	Z with support $\{0, 1, 2, ..., 100\}$	$\{80, 69, 75, 60, 92, ...\}$
Distribution	$Z \sim N(80, 10)$ (assumed)	Proportion of 80+ $= 0.14$

Deterministic vs probabilistic models

In high school, mathematical models are typically presented as if they operate with certainty. For example, the time it takes an object to fall from a height h to the ground is given by $t = \sqrt{\frac{2h}{g}}$, where g denotes the gravitational constant. The outcome here is *deterministic*: once the values of the variables are specified, the result follows with certainty. While the variables may or may not be known in practice, they are not *random* in the sense that the outcome is fully determined once inputs are given. Errors can only arise from frictions or measurement inaccuracies.

By contrast, many real-world processes are inherently uncertain. Consider tomorrow's temperature or stock market returns: such outcomes can only be predicted probabilistically. This uncertainty does not reflect randomness in the nature of the universe itself, but rather the limits of human knowledge. In principle, with perfect information about the climate system, tomorrow's temperature could be predicted exactly. However, given informational constraints, the only feasible approach is to incorporate uncertainty into mathematical models. *Probabilistic models* thus arise from the deliberate or unavoidable abstraction from complete information. The concept of the *random variable* provides the mathematical foundation for formalizing such uncertainty.

Exercise 14.1. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, with $P(\omega_1) = P(\omega_2) = P(\omega_3) = \frac{1}{3}$. Define random variables $X, Y, Z : \Omega \to \mathbb{R}$ by

$$\begin{aligned} X(\omega_1) &= 1, \quad X(\omega_2) = 2, \quad X(\omega_3) = 3 \\ Y(\omega_1) &= 2, \quad Y(\omega_2) = 3, \quad Y(\omega_3) = 1 \\ Z(\omega_1) &= 2, \quad Z(\omega_2) = 2, \quad Z(\omega_3) = 1 \end{aligned}$$

1. Show that X and Y have the same distribution.
2. Find the distribution of $X + Y$, XY, and X/Y.

15 Data descriptives*

A random variable is a mathematical abstraction that provides a bridge between theoretical probability and real-world data. Every dataset can be viewed as observations from random variables.

Despite the outcome of any one event being uncertain, we can use patterns from past observations to predict the general behavior of these variables. By collecting data, we can figure out how often certain outcomes occur and connect them to theoretical distributions.

$$\begin{array}{ccccccc} \begin{array}{c}\text{Question with}\\ \text{uncertainty}\end{array} & \rightarrow & \text{Data} & \rightarrow & \text{Patterns} & & \\ & & \downarrow & & \downarrow & & \\ & & \text{RVs} & \rightarrow & \text{Distributions} & \rightarrow & \text{Predictions} \end{array}$$

Columns as random variables

In a dataset, we view every column as a random variable.

```r
# Load a dataset from a CSV file
exam <- read.csv("../dataset/exam.csv")

# View the data: each column is a random variable
head(exam)
```

```
  id gender     major hw mid final overall
1  1 Female Economics 85  89    74      81
2  2 Female   Finance 90  84    79      83
3  3 Female Economics 90  71    51      65
4  4 Female   Finance 86  84    68      76
5  5   Male   Finance 80  84    67      75
6  6 Female   Finance 96 100    99      99
```

Summary statistics

We can describe the distribution of a variable with summary statistics: such as quartiles, deciles and percentiles.

```
summary(exam$overall)
```

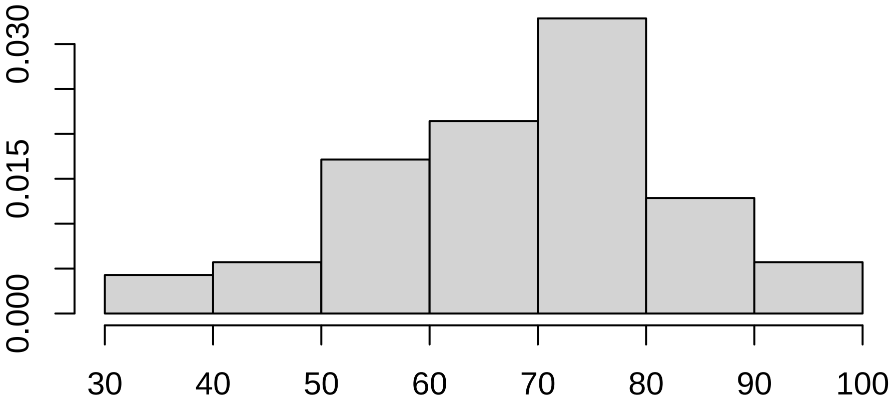
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  45.00   72.00   77.50   76.56   84.00   99.00 
```

Histograms

One way to visualize the distribution of a variable is to plot a histogram. A histogram groups data points into intervals, showing how often data values fall within each range. The horizontal axis represents the intervals (or bins), and the vertical axis shows the frequency or count of data points in each bin.

A histogram gives an approximation of the true (unknown) distribution. It is not itself the distribution. A distribution refers to the *theoretical* assignment of probabilities across all possible outcomes, whereas a histogram represents the *empirical* frequencies observed in a finite sample.

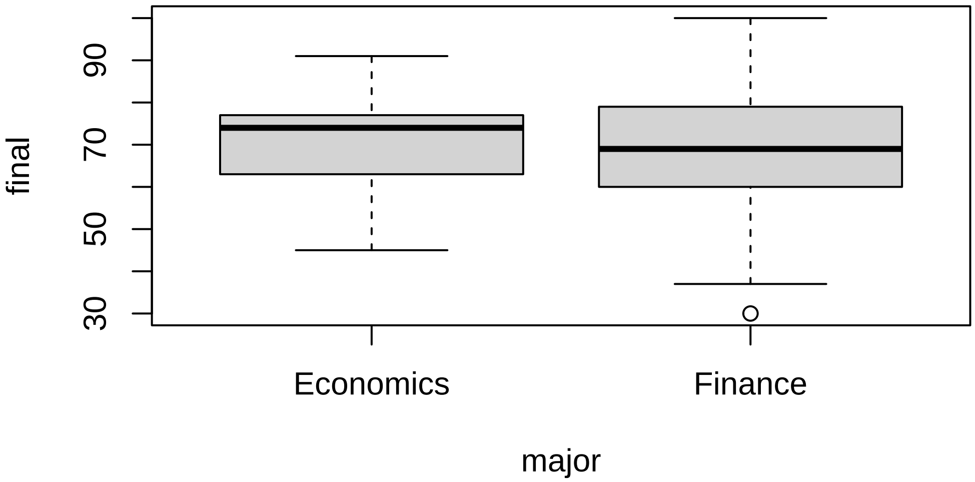
```
hist(exam$final, prob = T, ann=F)
```



Boxplots

A boxplot, also known as a box-and-whisker plot, displays the median, quartiles, and range of the data. The box represents the interquartile range (IQR), which contains the middle 50% of the data, with the lower and upper edges corresponding to the first (Q1) and third quartiles (Q3). Whiskers extend from the box to indicate the range of values within 1.5 times the IQR from Q1 and Q3, while points beyond this range are considered outliers.

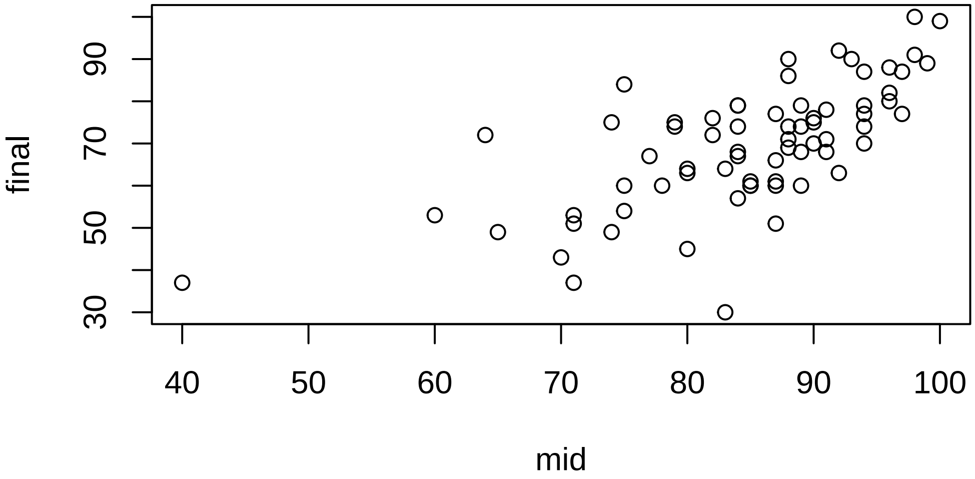
```
boxplot(final ~ major, exam)
```



Scatter plots

To observe the relationship between variables, it is straightforward to make a scatter plot of Y against X. An upward-sloping pattern in the scatter plot indicates that the variables tend to move together, whereas a downward-sloping pattern suggests that they move in opposite directions. A flat slope implies the absence of correlation between the two variables.

```
plot(final ~ mid, exam)
```



16 Discrete RVs

Definition 16.1 (Discrete random variable)**.** We say X is a discrete random variable if X can take a finite or countable number of values $x_1, x_2, \dots, x_n$.

Definition 16.2 (Support)**.** The finite or countably infinite set of values x such that $P(X = x) > 0$ is called the support of X.

Definition 16.3 (Probability mass function)**.** If a random variable X has a discrete distribution, the probability mass function (PMF) of X is defined as the function $f : \mathbb{R} \to [0, 1]$ such that

$$f(x) \equiv P(X = x).$$

Note that the PMF $f(x)$ is a discrete function which can only take values in the support $\{x_1, x_2, \dots, x_n\}$.

> **i** Notation for PMF
>
> Throughout this course, we use PMF to refer to the probability function for a discrete random variable. Some textbooks may call it the *probability function* (p.f.), while others may use the term *mass function.* All these terms describe the same concept.
> Note that how $f(x)$ differs from the probability function $P(\cdot)$. $f(x)$ is a real-valued function, whereas $P(\cdot)$ is the probability operator. The two should not be confused even when the notation $p(x)$ is used to represent a PMF.
> We may want to use a subscript to distinguish PMFs for different RVs. For example, f_X is the PMF for random variable X, f_Y is the PMF for random variable Y.

Proposition 16.1. *A probability mass function $f : \mathbb{R} \to [0, 1]$ satisfies*

1. *$f(x) \geq 0$ for all x and $f(x) \neq 0$ if and only if x is in the support.*
2. *$\sum_i f(x_i) = 1$ where i indexes every value in the support.*

There are different ways to represent a PMF. We can (1) list all the possible values and their associated probabilities; (2) write a formula for the PMF; or (3) visualize it in a graph.

Example 16.1 (Bernoulli distribution)**.** A random variable X is said to have the Bernoulli distribution if X has only two possible values, 0 and 1, and $P(X = 1) = p$, $P(X = 0) = 1 - p$.

The PMF of a Bernoulli random variable X is given by

$$f(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

This can also be expressed as

$$f(k) = p^k (1 - p)^{1-k}, \quad k \in \{0, 1\}\,.$$

Example 16.2. A student is trying to connect to the campus Wi-Fi network. Each attempt is independent, and:

- With probability p the attempt is successful.
- With probability $1 - p$ the attempt fails, and the student tries again.

The student will keep trying until the first success.

1. Define A_k = "the first successful connection occurs on the k-th attempt." Find $P(A_k)$.
2. Define a random variable X = "the number of attempts needed until the first success." What is the support of X?
3. Derive the probability mass function (PMF) of X.
4. Show that this is a valid PMF (Proposition 16.1).

17 Continuous RVs

Definition 17.1 (Continuous random variable)**.** We say a random variable X has a continuous random variable if the possible values of X takes the form of a continuum.

Definition 17.2 (Probability density function)**.** For a continuous random variable X, the probability density function (PDF) of X is a real-valued function $f : \mathbb{R} \to [0, \infty)$ such that

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

Continuous random variables are usually measurements. Examples include height, weight, temperature, the amount of money and so on.

> Density is not probability
>
> PDF differs from the discrete PMF in important ways:
>
> - For a continuous random variable, $P(X = x) = 0$ for all x;
> - The quantity $f(x)$ is not a probability. To get the probability, we integrate the PDF (probability is the area under the PDF):
>
> $$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx.$$
>
> - Since any single value has probability 0, including or excluding endpoints does not matter.
>
> $$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

Proposition 17.1. *If X has density function f then*

1. $P(X = x) = 0$ *for all* $x \in \mathbb{R}$
2. $P(a \leq X \leq b) = \int_a^b f(x)dx$
3. $\int_{-\infty}^{\infty} f(x)dx = 1$

Example 17.1 (Uniform distribution)**.** A uniform distribution is a probability distribution where all values within a specified interval $[a, b]$ are equally likely to occur, and its probability density function (PDF) is given by:

$$f(x) = \frac{1}{b-a} \quad \text{for} \quad a \leq x \leq b$$

and $f(x) = 0$ otherwise.

Don't confuse a random variable with its distribution

If random variable X has distribution $f(x)$, the distribution of X^2 is not $f^2(x)$. To get the distribution of $X + Y$, you can't just add up $f_X(x) + f_Y(y)$. The right way to do it will be discussed in later chapters (transformation and convolution).

Example 17.2. Every morning, a student waits for the elevator in their dormitory. The waiting time (in minutes) is equally likely to be anywhere between 0 and 3 minutes, depending on when the elevator arrives.

1. Define X = "the student's elevator waiting time." What is the support of X?
2. Derive the probability density function (PDF) of X.
3. Compute $P(X \leq 1)$, i.e. the probability that the waiting time is at most 1 minute.
4. If the student must wait more than 2 minutes, they decide to take the stairs instead. Define a new indicator random variable Y, which equals 1 if $X > 2$ and 0 otherwise. Compute $P(Y = 1)$.

18 Cumulative distribution

Definition 18.1 (Cumulative distribution function)**.** The cumulative distribution function (CDF) of a random variable X is the function F given by $F(x) = P(X \leq x)$.

For discrete random variables, $F(x) = \sum_{k \leq x} p(k)$.

For continuous random variables, $F(x) = \int_{-\infty}^{x} f(t)dt$. We thus have $\frac{dF(x)}{dx} = f(x)$.

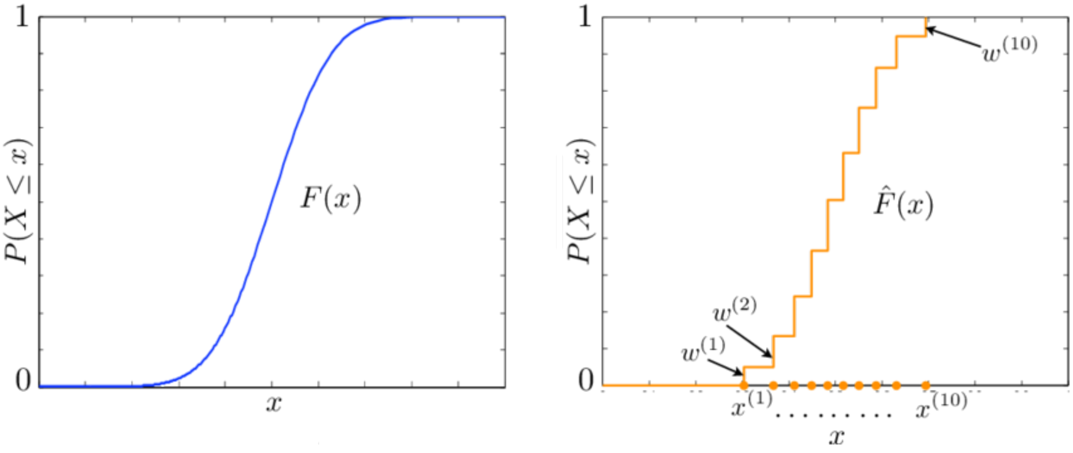
Unlike PMF or PDF, a cumulative distribution function can be defined for both discrete and continuous random variables. CDF gives the full distribution of a random variable. Given the CDF, we can figure out any probability distribution of the random variable:

$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1).$$

Proposition 18.1. *Any CDF has the following properties:*

- $P(X > x) = 1 - F(x)$
- $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$
- *Increasing: if* $x_1 \leq x_2$*, then* $F(x_1) \leq F(x_2)$*.*
- *Right-continuous: for any* a*,* $F(a) = \lim_{x \to a+} F(x)$*.*
- $F(x) \to 0$ *as* $x \to -\infty$*;* $F(x) \to 1$ *as* $x \to +\infty$*.*

The CDF for a continuous random variable is differentiable, while the CDF for a discrete random variable consists of jumps and flat regions.


1
0
$P(X \leq x)$
$F(x)$
x
1
0
$P(X \leq x)$
$\hat{F}(x)$
$w^{(1)}$
$w^{(2)}$
$w^{(10)}$
$x^{(1)}$
$x^{(10)}$
x

Part III

Discrete Distributions

19 Binomial distribution

Definition 19.1 (Binomial distribution)**.** Suppose $X_1, X_2, \dots, X_n$ are independent and identical $\text{Bern}(p)$ distributions. Let X be the total number of successes of the n independent trials. That is, $X = X_1 + X_2 + \dots + X_n$. Then X has the Binomial distribution, $X \sim \text{Bin}(n, p)$.

The PMF of X directly follows from the combination theory:

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

This is a valid PMF because, by the Binomial theorem, we have

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

Binomial distribution and Binomial theorem

You may have noticed the connection between Binomial distribution and Binomial theorem. Consider using the polynomial $px + q$ to represent the outcome of a single Bernoulli trial, where x is the indicator for a success. Then $(px + q)^n$ is the outcome for n independent trials. The coefficient of x^k gives the probability of there being exactly k successes.

Theorem 19.1. *Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ be two independent Binomial random variables. Then $X + Y \sim Bin(n + m, p)$.*

Proof. By the definition of the Binomial distribution, $X = \sum_{i=1}^{n} X_i$ where $X_i \sim \text{Bern}(p)$; $Y = \sum_{j=1}^{m} Y_j$ where $Y_j \sim \text{Bern}(p)$. Therefore,

$$X + Y = \sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j = \sum_{k=1}^{n+m} Z_k$$

where $Z_k \sim \text{Bern}(p)$. Since X_i and Y_j are identical Bernoulli random variables,

$$Z_k = \begin{cases} X_k, & \text{for } k = 1, \dots, n \\ Y_{k-n}, & \text{for } k = n+1, \dots, n+m \end{cases}$$

By definition, $X + Y = \sum_{k=1}^{n+m} Z_k \sim \text{Bin}(n + m, p)$. □

Coin tossing problem

Example 19.1. In the previous example of tossing two coins, we compute the distribution of X by counting the equally likely outcomes in an event. We can get the same result by realizing it is a Binomial distribution. $X \sim \text{Bin}(2, 1/2)$. Since each coin tossing is an independent Bernoulli trial. The probabilities come directly from the PMF.

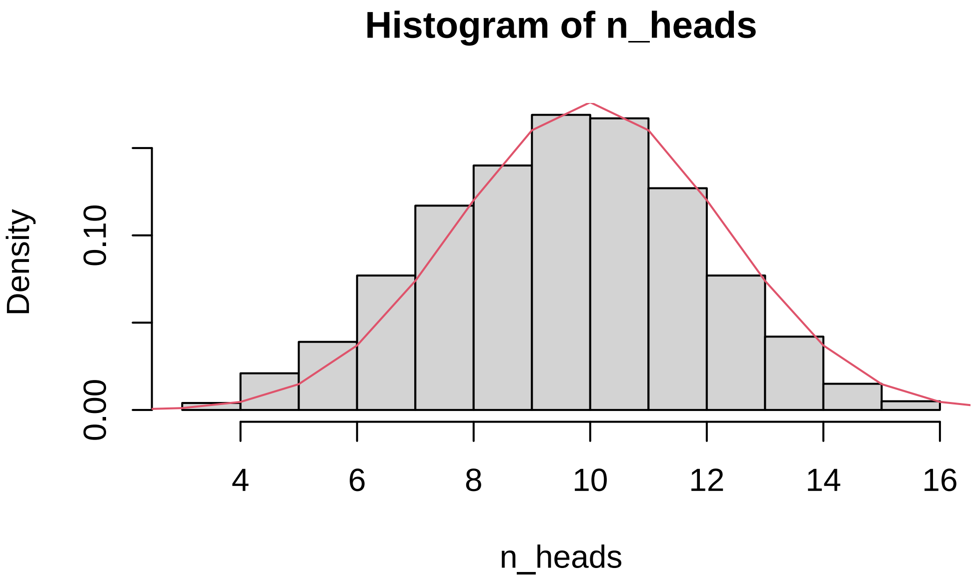
$$
\begin{aligned}
P(X=0) &= \binom{2}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^2 = \frac{1}{4};\\
P(X=1) &= \binom{2}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1 = \frac{1}{2};\\
P(X=2) &= \binom{2}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0 = \frac{1}{4}.
\end{aligned}
$$

Utilizing the Binomial distribution also allows us to generalize the problem. Suppose we are tossing n coins, we want to find the probability of getting k heads. It is almost impossible to count all the possible outcomes, but the answer immediately follows from the Binomial PMF.

The following code simulates the number of heads if tossing N coins:

```r
# Number of simulations
k <- 1000

# Number of coins
n <- 20

# Store the results
n_heads <- numeric(k)

# Initialize random generator
set.seed(100)

# Run simulations
for (i in 1:k) {
  toss <- sample(c('H','T'), n, replace = T)
  n_heads[i] <- sum(toss == 'H')
}

# Plot distribution
hist(n_heads, probability=TRUE)
```

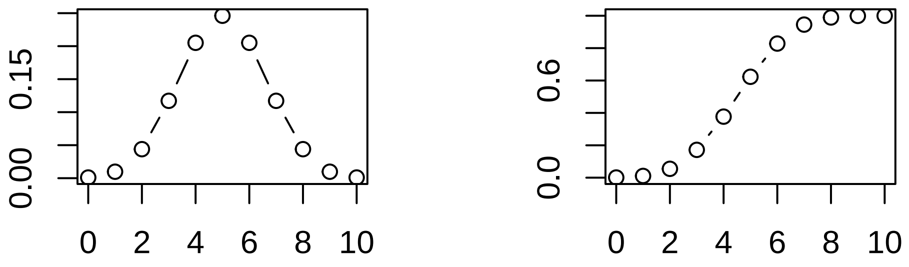
```
# Overlay the Binomial PMF
curve(choose(n,x)* 0.5^x * 0.5^(n-x), from=1, to=n, n=n, col=2, add=T)
```



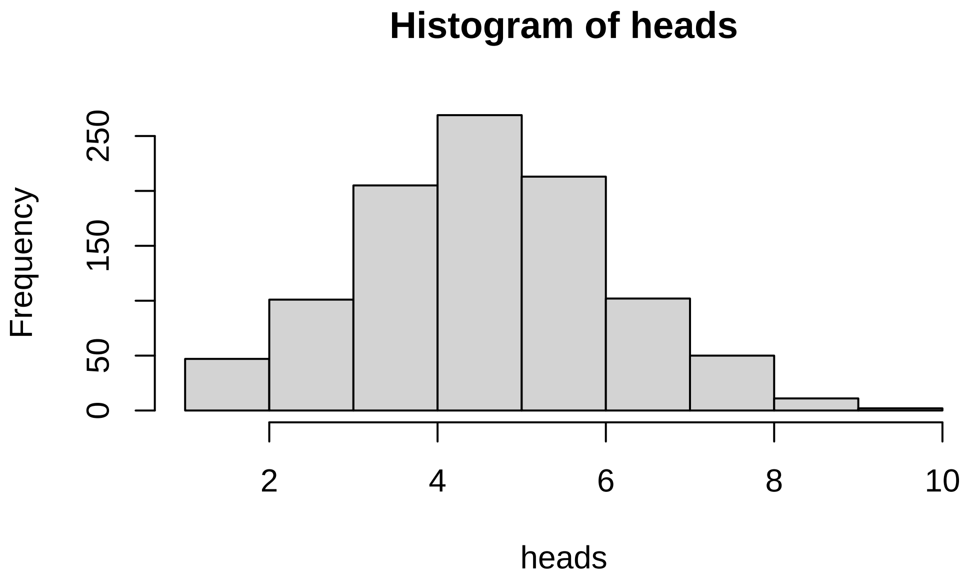
Binomial functions in R

There are built-in functions in R to work with Binomial distributions.

```
# computes P(X=5) for Bin(10,0.5)
p <- dbinom(5, 10, 0.5)

par(mfrow=c(1,2))

# plot the PMF for Bin(10,0.5)
curve(dbinom(x, 10, 0.5), from=0, to=10, n=11, type="b", ann=F)

# `pbinom` computes the CDF
curve(pbinom(x, 10, 0.5), from=0, to=10, n=11, type="b",ann=F)
```



```r
# draw a random value from a given Binomial distribution
# this allows us to simulate a random experiment
# e.g. the number of heads when flipping 10 fair coins
outcome <- rbinom(1, 10, 0.5)

# Repeat the experiment 1000 times
heads <- rbinom(1000, 10, 0.5)

# the histogram will converge to the ideal Binomial distribution
# if the experiment is repeated a large number of times
hist(heads)
```



Exam survival problem

Example 19.2. An exam consists of 20 multiple-choice questions, each with four choices and exactly one correct answer. Suppose a student answers every question by guessing at random.

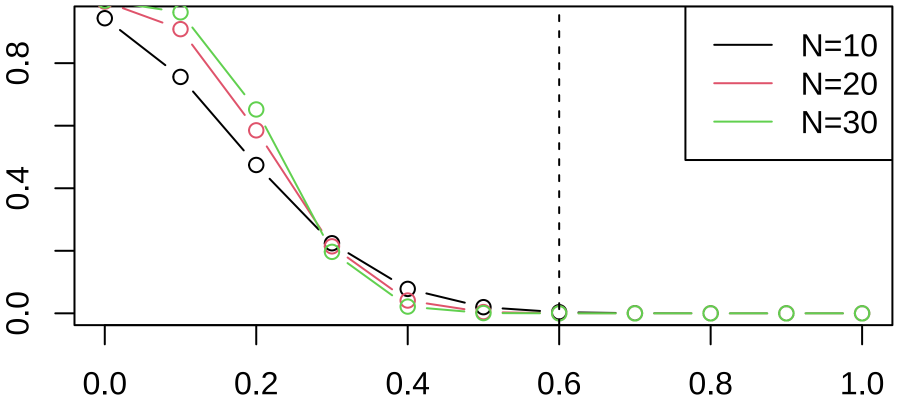
What is the probability that the student passes the exam, defined as answering more than 60% of the questions correctly?

Solution. The probability of correctly answering one question is $p = 1/4$. Let N be the total number of questions, $N = 20$. Let X be the number of correctly answered questions, $X \leq N$. Then X follows the Binomial distribution $X \sim B(N, 1/4)$. The probability of passing the exam is therefore

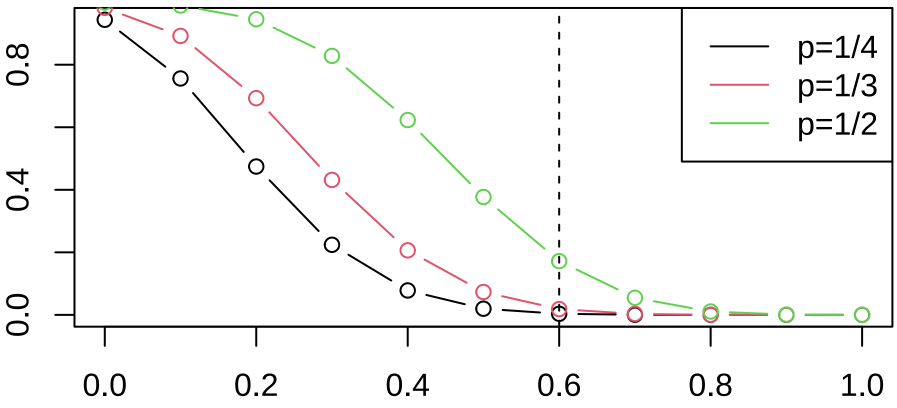
$$P(X \geq 12) = 1 - P(X \leq 11) = 1 - \text{CDF}^{\text{Bin}}(11) \approx 0.001.$$

Now we compare the survival probability for different choice of N and p :

```r
# Percentage of correct answers
x <- seq(0, 1, .1)

# Survival probabilities for different N
y1 <- 1 - pbinom(10*x, 10, .25)
y2 <- 1 - pbinom(20*x, 20, .25)
y3 <- 1 - pbinom(30*x, 30, .25)

# Compare the curves for different N
plot(x, y1, type="b", col=1, ann=F)
lines(x, y2, type="b", col=2)
lines(x, y3, type="b", col=3)

# Indicating passing the exam
abline(v=0.6, lty=2)

# Add a legend at the top right corner of the plot
legend("topright", c("N=10", "N=20", "N=30"), lty=1, col=1:3)
```



```
# Percentage of correct answers
x <- seq(0, 1, .1)

# Survival probabilities for different p (number of choices)
y1 <- 1 - pbinom(10*x, 10, .25)
y2 <- 1 - pbinom(10*x, 10, .33)
y3 <- 1 - pbinom(10*x, 10, .5)

# Compare the curves for different p
plot(x, y1, type="b", col=1, ann=F)
lines(x, y2, type="b", col=2)
lines(x, y3, type="b", col=3)

# Indicating passing the exam
abline(v=0.6, lty=2)

# Add a legend at the top right corner of the plot
legend("topright", c("p=1/4", "p=1/3", "p=1/2"), lty=1,col=1:3)
```



20 Discrete expectation

Definition 20.1 (Expectation of a discrete random variable)**.** Let X be a discrete random variable. The expectation of X (or the mean of X) is defined as:

$$E(X) = \sum_{\text{all } x} xP(X = x).$$

In other words, the expected value of X is a *weighted average* of the possible values that X can take on, weighted by their probabilities.

> **i** Note
>
> The expected value of X is a fixed number, $E(X) \in \mathbb{R}$. It is not a random variable such as $g(X)$.

Sometimes, we would like to omit the parentheses for simplicity and write $EX := E(X)$. We also like to denote expectation by the Greek letter $\mu := E(X)$.

Example 20.1. The expectation of a Bernoulli random variable $X \sim \text{Bern}(p)$:

$$E(X) = 1 \times P(X = 1) + 0 \times P(X = 0) = p.$$

Example 20.2. The expectation of a Binomial random variable $X \sim \text{Bin}(n, p)$:

$$\begin{aligned} E(X) =& \sum_{k=0}^{n} kp(k) \\ =& \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k q^{n-k} \\ =& \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} p^k q^{n-k} \\ =& np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} \\ =& np \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j}}_{\text{another Binomial PMF}} \\ =& np. \end{aligned}$$

Proposition 20.1. *Expectation has the following properties:*

- $E(X+Y) = E(X) + E(Y)$
- $E(aX+b) = aE(X) + b$

Example 20.3. Redo the expectation of $X \sim \text{Bin}(n,p)$ with properties of expectation:

$$E(X) = E(X_1 + \dots + X_n) = nE(X_i) = np$$

where $X_i \sim \text{Bern}(p)$.

Law of averages*

You may wonder what is the difference between $E(X)$ defined in Definition 20.1 and the average of values defined as $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$.

The short answer is this: $E(X)$ is a theoretical value, while $\bar{X}$ is an approximation to $E(X)$ with finite observations. They are associated by the following theorem.

Law of averages

The law of averages (or the law of large numbers) states that if you repeat a random experiment, such as tossing a coin or rolling a die, a very large number of times, your individual outcomes, when averaged, should be very close to the theoretical mean (a constant parameter). In mathematical language,

$$\bar{X}_n \to^p \mu \text{ when } n \to \infty.$$

where $\to^p$ reads as "converge in probability".

There is another fundamental difference. In probability theory, we treat $E(X)$ as a fixed number; while $\bar{X}$ is another random variable! Because the sample $\{X_1, X_2, \dots, X_n\}$ is generated randomly. Consider the coin flipping example, while $E(X) = 0.5$ is a constant, each time you compute the average of, say, 10 flips, you get a different number. We will come back to this point later.

21 Hypergeometric dist

Example 21.1. Let's explore an example that appears to be Binomial but is, in fact, not a Binomial distribution. Given a 5-card hand. Find the distribution of the number of aces.

Let X be the number of aces. It is tempting to say $X \sim \text{Bin}(5, p)$. But this not correct. Because having one ace is NOT independent from having another ace. We need to use the classical approach:

$$P(X = k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}.$$

This is a Hypergeometric distribution.

Suppose we have a box filled with w white and b black balls. We draw n balls out of the box with replacement. Let X be the number of white balls. Then $X \sim \text{Bin}(n, w/(w+b))$. Since the draws are independent Bernoulli trials, each with probability $w/(w+b)$ of success. If we instead sample without replacement, then the number of white balls follows a **Hypergeometric distribution**. We denote this by $X \sim \text{HGeom}(w, b, n)$.

Theorem 21.1. *If $X \sim HGeom(w, b, n)$, then the PMF of X is*

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

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$, and $p_X(k) = 0$ otherwise.

In Example 21.1, the number of aces in the hand has the HGeom$(4, 48, 5)$ distribution, which can be seen by thinking of the aces as white balls and the non-aces as black balls. The probability of having exactly three aces is 0.0017%.

Example 21.2. Let $X \sim \text{HGeom}(w, b, n)$. Find $E(X)$ the expected number of white balls. Similarly, we can decompose X:

$$X = I_1 + \cdots + I_n$$

where I_j equals 1 if the jth ball is white and 0 otherwise. We have said that $\{I_j\}$ are not independent, but the linearity of expectation still holds:

$$E(X) = E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n).$$

Meanwhile we have

$$E(I_j) = P(j\text{-th ball is white}) = \frac{w}{w+b}$$

since unconditionally the jth ball is equally likely to be any of the balls. Thus, $E(X) = \frac{nw}{w+b}$.

The Binomial and Hypergeometric distributions are often confused. Both are discrete distributions taking on integer values between 0 and n for some n, and both can be interpreted as the number of successes in n Bernoulli trials. However, a crucial part of the Binomial story is that the Bernoulli trials involved are independent. The Bernoulli trials in the Hypergeometric story are dependent, since the sampling is done without replacement.

22 Geometric distribution

Definition 22.1 (Geometric distribution)**.** Consider a sequence of independent Bernoulli trials, each with the same success probability p. Let X be the number of failures before the first successful trial. Then X has a Geometric distribution: $X \sim \text{Geom}(p)$.

Let's derive the PMF for the Geometric distribution. By definition,

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

where $q = 1 - p$. This is a valid PMF because

$$\sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = \frac{p}{1-q} = 1.$$

The expectation of X is given by

$$E(X) = \sum_{k=0}^{\infty} k \cdot q^k p = p \sum_{k=0}^{\infty} k q^k = p \frac{q}{p^2} = \frac{q}{p}.$$

To see why this holds, taking derivative with respect to q on both sides of $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ yields

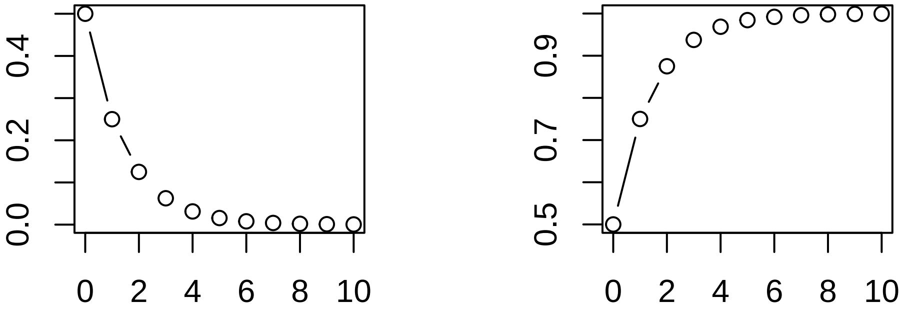
$$\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2};$$

Then multiply both sides by q:

$$\sum_{k=1}^{\infty} k q^k = \frac{q}{(1-q)^2} = \frac{q}{p^2}.$$

Plot the PMF and CDF

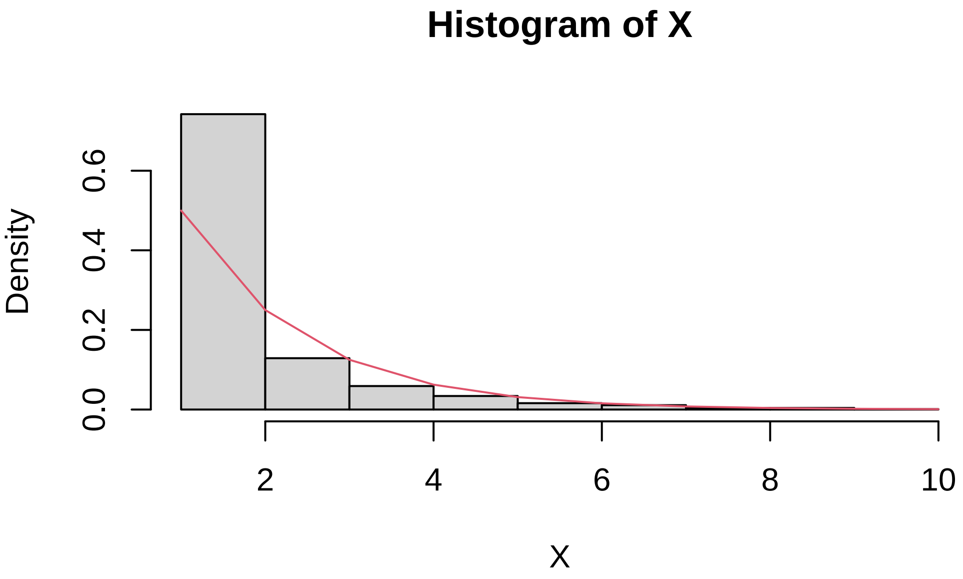
```
par(mfrow=c(1,2))

# PMF for Geom(0.5)
curve(dgeom(x, 0.5), from=0, to=10, n=11, type="b", ann=F)

# CDF for Geom(0.5)
curve(pgeom(x, 0.5), from=0, to=10, n=11, type="b", ann=F)
```



Example 22.1 (Coin flip until Head)**.** Flipping a fair coin, what is the expected number of flips before the first Head?

Let X be the number of flips until the first head. We know $X - 1 \sim \text{Geom}(0.5)$ as geometric distribution models the number of failures excluding the success. Thus, $E(X-1) = 0.5/0.5 = 1$, $E(X) = 2$. Let's compare the theoretical value with results from simulations.

```
# number of simulations
N <- 1000

# X: number of flips until first head
# stores value of X in each simulation
X <- numeric(N)

set.seed(100)

# run simulations
for (i in 1:N) {
  x <- 0
  # repeat until first head
  while(TRUE) {
    x <- x + 1
    t <- sample(c('H','T'), 1, F)
    if (t == 'H') break
  }
  # record the number
  X[i] <- x
}

# plot distribution of X
hist(X, probability=T)

#overlay with geometric distribution
curve(dgeom(x-1,.5), from=1, to=10,n=10,add=T,col=2)
```



```
cat("Average number of flips until Head:", mean(X))
```

```
Average number of flips until Head: 2.021
```

23 Coin flip: HH vs HT*

Flip a coin indefinite times. Let X denote the number of flips until you see **HH**. Let Y denote the number of flips until you see **HT**. Find $E(X)$ and $E(Y)$.

It is tempting to think they are the same, since either H or T happens with probability 1/2. But the answer is extremely counter-intuitive: $E(X) > E(Y)$!

HH case. Let E_0 = E(X|No H observed), and E_1 = E(X|One H observed). Then

$$E_0 = 1 + \frac{1}{2}E_1 + \frac{1}{2}E_0$$

The first term is we need to flip once. If the first flip is H, the additional expected number of flips is E_1. If the first flip is T, we have to start over again (E_0).

$$E_1 = 1 + \frac{1}{2}(0) + \frac{1}{2}E_0$$

Once we have observed an H, we do another flip. If it is another H, we are done. If it is a T, we have to start over again (E_0).

Solve the two equations, we have $E_0 = 6$, $E_1 = 4$. Thus, $E(X) = 6$.

HT case. Let E_0 = E(Y|No H observed), and E_1 = E(Y|One H observed). Then

$$E_0 = 1 + \frac{1}{2}E_1 + \frac{1}{2}E_0$$

If the first flip is H, we need E_1. If the first flip is T, we have wasted the flip, so it is E_0 again.

$$E_1 = 1 + \frac{1}{2}(0) + \frac{1}{2}E_1$$

If we have a T by 1/2 chance, we are done (the first term). If it is an H, we get another E_1.

In this case, we have $E_0 = 4$, $E_1 = 2$. Thus, $E(Y) = 4$.

```r
# number of simulations
N <- 1000

# X: number of flips until HH
X <- numeric(N)

set.seed(100)

# run simulations
for (i in 1:N) {
  x <- 0
  # repeat until first head
  while(TRUE) {
    x <- x + 1
    t <- sample(c('H','T'), 1, F)
    if (x >=2 && t == 'H' && tt == 'H') break
    else tt <- t  # store last toss
  }
  # record the number
  X[i] <- x
}

cat("Average number of flips until HH:", mean(X))
```

```
Average number of flips until HH: 5.789
```

```r
# number of simulations
N <- 1000

# X: number of flips until HT
Y <- numeric(N)

set.seed(100)

# run simulations
for (i in 1:N) {
  y <- 0
  # repeat until first head
  while(TRUE) {
    y <- y + 1
    t <- sample(c('H','T'), 1, F)
    if (y >= 2 && t == 'T' && tt == 'H') break
    else tt <- t  # store last toss
  }
  # record the number
  Y[i] <- y
}

cat("Average number of flips until HT:", mean(Y))
```

```
Average number of flips until HT: 4.047
```

24 Negative Binomial

Definition 24.1 (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 a Negative Binomial distribution, denoted $X \sim \mathrm{NBin}(r, p)$.

The PMF for Negative Binomial distribution is given by

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

To compute the expectation, let $X = X_1 + \cdots + X_r$ where X_i is the number of failures between the $(i - 1)$-th success and the i-th success, $1 \leq i \leq r$. Then $X_i \sim \mathrm{Geom}(p)$. By linearity of expectations,

$$E(X) = E(X_1) + \cdots + E(X_r) = r\frac{1 - p}{p}.$$

Theorem 24.1. *Let $X_1, X_2, ..., X_n$ be independent geometric distributions with the same parameter p. Then*

$$X = X_1 + X_2 + \cdots + X_n$$

is a negative binomial distribution, namely $X \sim NBin(n, p)$.

The theorem is straightforward if we interpret X_1 as the number of failures before the 1st success, X_2 as the number of failures between the 1st and 2nd successes, and X_n be the number of failures between the $(n - 1)$-th and n-th successes.

Example 24.1 (Toy collection)**.** There are n types of toys. Assume each time you buy a toy, it is equally likely to be any of the n types. What is the expected number of toys you need to buy until you have a complete set?

Solution. Define the following random variables:

$$\begin{aligned}
T =& T_1 + T_2 + \cdots + T_n \\
T_1 =& \text{number of toys until 1st new type} \\
T_2 =& \text{additional number of toys until 2nd new type} \\
T_3 =& \text{additional number of toys until 3rd new type} \\
&\vdots
\end{aligned}$$

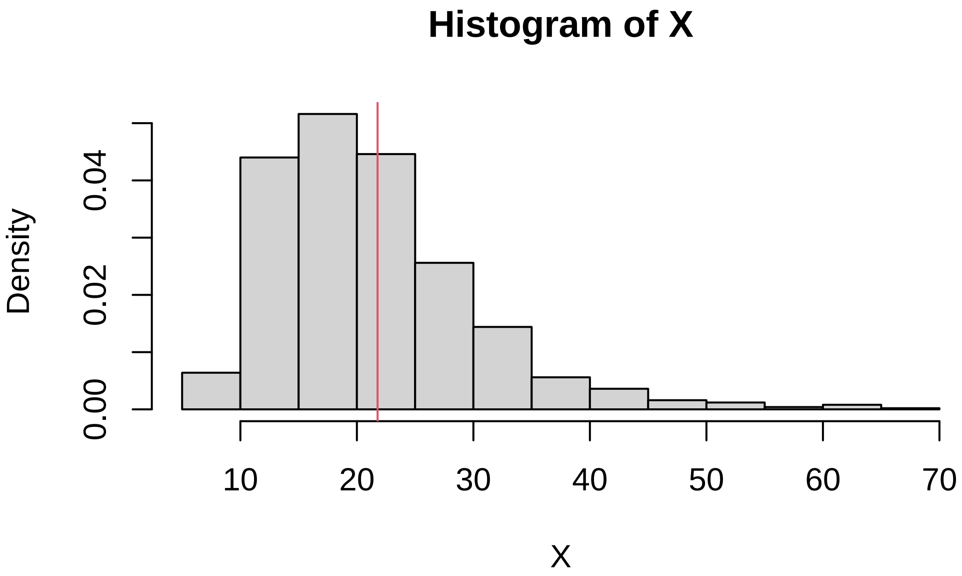
We know, $T_1 = 1$, $T_2 - 1 \sim \text{Geom}\left(\frac{n-1}{n}\right)$,..., $T_j - 1 \sim \text{Geom}\left(\frac{n-(j-1)}{n}\right)$. Thus,

$$\begin{aligned}
E(T) =& E(T_1) + E(T_2) + \cdots + E(T_n) \\
=& 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{1}{n} \\
=& n(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}) \\
\to& n(\log n + 0.577).
\end{aligned}$$

If $n = 5$, $E(T) \approx 11$; if $n = 10$, $E(T) \approx 29$.

```r
# number of simulations
N <- 1000

# number of toys bought in each simulation
X <- numeric(N)

# the set of toys
Toys <- c(' ',' ',' ',' ',' ',' ',' ',' ')

set.seed(100)

# run simulation
for (i in 1:N) {
  C <- c() # the collection of toys bought
  # repeat until a full set is collected
  while(TRUE) {
    t <- sample(Toys, 1, replace=T)
    C <- c(C, t)
    if(setequal(C, Toys)) break
  }
  X[i] <- length(C)
}
```

```
hist(X, probability = T)
abline(v = mean(X), col = 2)
```



25 Bivariate distribution

We need a tool to study collections of variables. Knowledge of each individual PMF is of little help. Because variables can be dependent on each each other (they are not necessarily independent). We need to know their inter-relationship. Joint distribution gives the probability that two or more random variables simultaneously takes particular values.

Definition 25.1 (Joint distribution)**.** The joint PMF of random variables (X, Y) is given by

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

The joint CDF of random variables (X, Y) is given by

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

Theorem 25.1. *The discrete random variables X and Y are* ***independent*** *if and only if*

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

for all possible values of x, y.

Equivalently, the condition can be stated with CDF: the random variables X and Y are ***independent*** *if and only if*

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

for all possible values of x, y.

Proof. X and Y are independent implies the event $\{X = x\}$ and $\{Y = y\}$ are independent for any x, y. By Definition 10.1, we have

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

□

Note

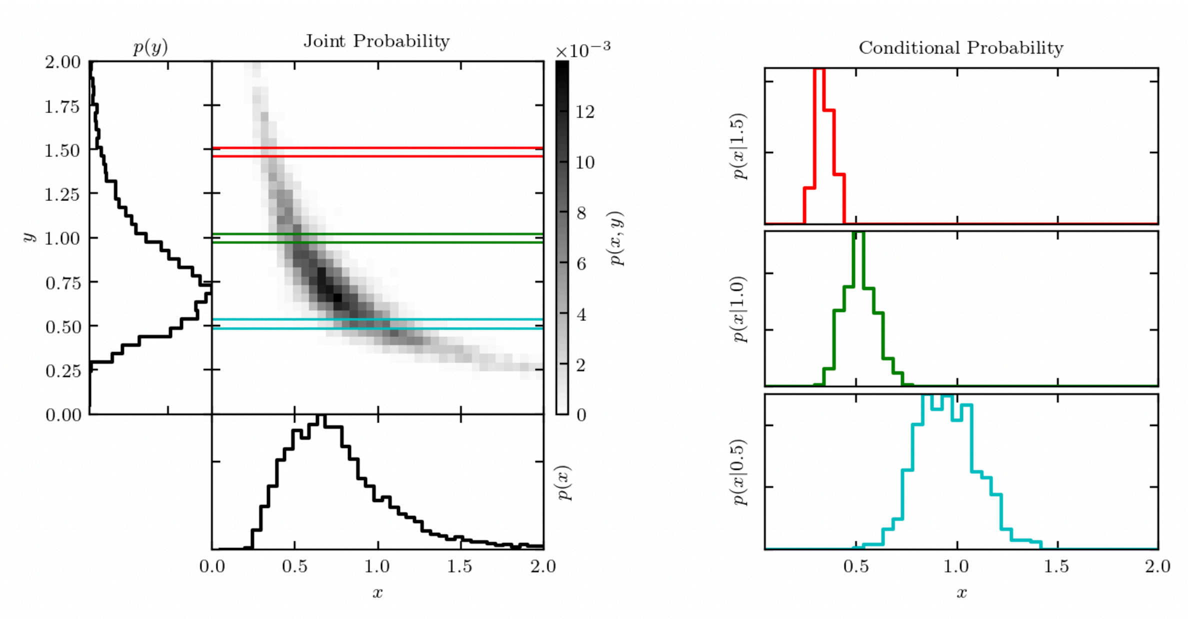
If X and Y are independent, then any function of X is independent of any function of Y.

Definition 25.2 (Marginal distribution)**.** The marginal distribution gives the distribution of a subset of variables in a joint distribution without reference to the values of the other variables.

The marginal PMF of X given the joint PMF of (X, Y) is given by

$$f_X(x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x, y).$$

> Note
>
> It is easy to compute the marginal distribution given the joint distribution. However, in general, we cannot deduce the joint distribution from the marginal distribution. Unless the random variables are independent, the joint distribution is **not** the product of marginal distributions.



Example 25.1. Let X be an indicator of an individual being a current smoker. Let Y be the indicator of his developing lung cancer at some point in his life. The joint PMF of X and Y is as specified in the table below.

	$Y = 1$	$Y = 0$	**Total**
$X = 1$	0.05	0.20	**0.25**
$X = 0$	0.03	0.72	**0.75**

Total	**0.08**	**0.92**	**1**

The marginal PMF for having lung cancer is

$$
\begin{aligned}
P(Y=1) =& P(Y=1, X=0) + P(Y=1, X=1) = 0.08, \\
P(Y=0) =& P(Y=0, X=0) + P(Y=0, X=1) = 0.92.
\end{aligned}
$$

In this example, X, Y are not independent, because

$$P(X=1, Y=1) \neq P(X=1)P(Y=1).$$

Definition 25.3. If a given number of random variables are independent and have the same distribution, we call them **independent and identically distributed**, or **i.i.d** for short.

- Independent and identically distributed (X, Y independent die rolls)
- Independent and not identically distributed (X: die roll; Y: coin flip)
- Dependent and identically distributed (X: number of Heads; Y: number of Tails)
- Dependent and not identically distributed (X: economic growth; Y: presidential election)

Note

We view random sample as a collection of i.i.d random variables from the same population distribution. For example, let X_i be the test score of student i. We say $X_1, X_2, \dots, X_n \overset{iid}{\sim} G$ where G is the (unknown) population distribution for test scores.
The **independent** assumption means that one observation does not influence another, while the **identically distributed** assumption ensures all observations follow the same probability law. This perspective simplifies statistical analysis and is foundational for many statistical inference.

Exercise 25.1 (Benford's law)**.** The distribution of first two digits in many real-life data (e.g. annual accounts of a corporation) can be approximated by the joint mass function:

$$f(x, y) = \log_{10}\left(1 + \frac{1}{10x + y}\right), \quad 1 \leq x \leq 9, 0 \leq y \leq 9.$$

1. Verify this is valid joint PMF.
2. Find the marginal PMF of X.
3. Give an approximation to $E(X)$.

26 Conditional distribution

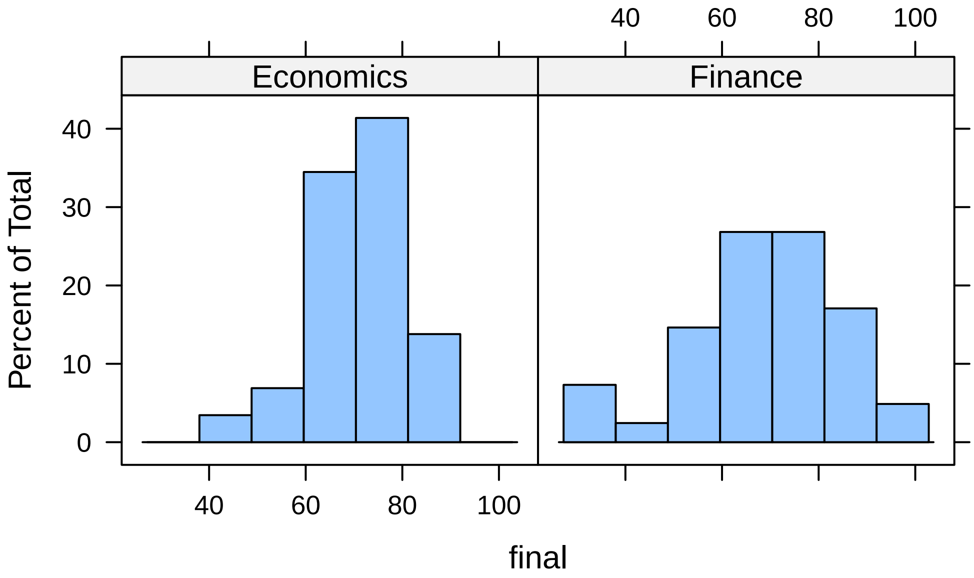
Definition 26.1 (Conditional distribution)**.** The conditional PMF of Y given $X = x$ is defined as

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

for any x such that $P(X = x) > 0$.

Plot conditional distribution

```r
library(lattice) # easy to make conditional plots

# conditional distribution of exam scores
exam <- read.csv("../dataset/exam.csv")

# distribution of exam scores conditioned on major
histogram(~ final | major, data = exam)
```



Definition 26.2 (Conditional expectation)**.** The conditional expectation of Y given $X = x$ is defined as

$$\psi(x) = E(Y|X = x) = \sum_{\text{all } y} y \ f_{Y|X}(y|x).$$

$\psi(x)$ depends on the value of x taken by X, so it can be thought of as a function $\psi(X)$ of X itself.

$$\psi(X) = E(Y|X)$$

is called the conditional expectation of Y given X.

> **Note**
>
> Although $E(X)$ is a number, $E(Y|X)$ is a random variable. It is a function of random variable X, and therefore it is a random variable itself.

Conditional distribution is a key concept in probability, describing how the distribution of one random variable depends on the values of other variables—an idea central to many practical applications. For instance, we might be interested in how income distributions vary by education level or how the probability of a disease changes with age.

Conditional expectation gives the expected value of one variable given the value of another. It is frequently used for making predictions, such as predicting your earnings given that you graduate from a this college.

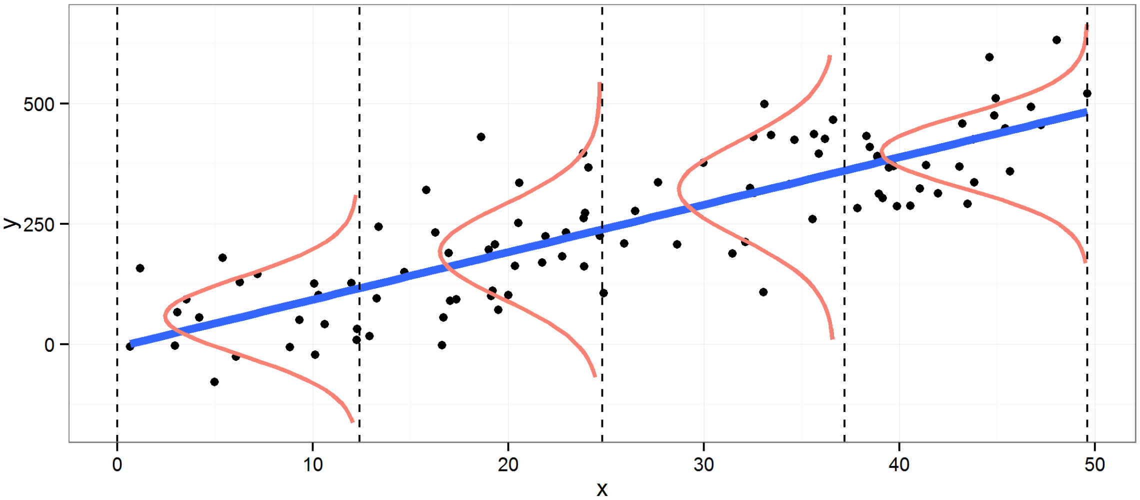


Figure 26.1: Given each value of X, there is a distribution of Y|X. E(Y|X) is a function of X.

27 Poisson distribution

Now we introduce arguably the most popular discrete distribution—Poisson distribution. Poisson distribution is used to model independent events occurring at a constant mean rate. It is like the Binomial distribution in the sense that they both model the number of occurrence of events, but it is parametrized on the "rate" of the event (how many times an event occurs in a unit of time on average) rather than the total number of events and the probability of each event. It is therefore more practical in real-world modeling since we mostly observe the rate rather than the totality. We introduce the Poisson distribution by showing that it is a limiting case of the Binomial distribution.

Question: Suppose we are studying the distribution of the number of visitors to a certain website. Every day, a million people independently decide whether to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. What is the probability of getting k visitors on a particular day?

We can model the problem with a Binomial distribution. Let $X \sim \mathrm{Bin}(n, p)$ be the number of visitors, where $n = 10^6$ and $p = 2 \times 10^{-6}$. But it is easy to run into computational difficulties with such a large n and small p. This is not uncommon, if we want to model the number of emails one receives per day, or the number of phone calls in a service center. In such cases, we could reasonably assume $n \to \infty$ and $p \to 0$ while $np = \lambda$ is a constant. We may call λ — the "rate", as it can be interpreted as the average visitors per day.

Take limit of the Binomial distribution:

$$\begin{aligned}
P(X = k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \lim_{n\to\infty} \underbrace{\frac{n!}{n^k(n-k)!}}_{\to 1} \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}.
\end{aligned}$$

This is the PMF of the Poisson distribution.

ℹ The limiting definition of exponential function

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

Definition 27.1 (Poisson distribution)**.** A random variable X has the Poisson distribution with parameter λ if the PMF of X is

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

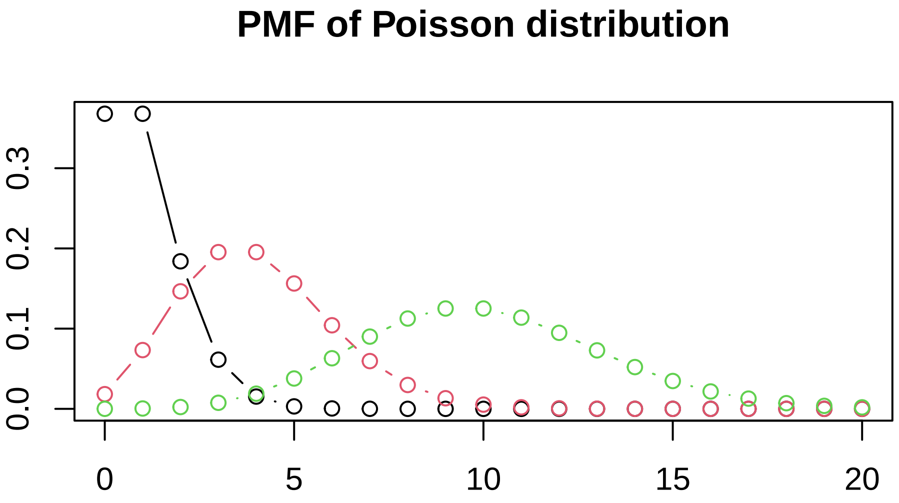
We denote this as $X \sim \text{Pois}(\lambda)$.

We can easily verify this is a valid PMF because $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda}$.

Theorem 27.1. *If $X \sim Bin(n, p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains fixed, then the PMF of X converges to the PMF of $Pois(\lambda)$.*

The expectation of the Poisson distribution is

$$\begin{aligned} E(X) =& \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda}\lambda^k}{k!} \\ =& e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ =& \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ =& \lambda e^{-\lambda}e^{\lambda} = \lambda. \end{aligned}$$



Example 27.1. Continued with the website visiting example, there are one million people visiting the site every day, each with probability $p = 2 \times 10^{-6}$. Give an approximation for the probability of getting at least three visitors on a particular day.

Let X be the number of visitors. Since n is large, p is small, $np = 2$ is fixed, X is well approximated by $\mathrm{Pois}(2)$. Therefore,

$$\begin{aligned} P(X \geq 3) = 1 - P(X < 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-2} - 2e^{-2} - \frac{2^2}{2!}e^{-2} \\ &= 1 - 5e^{-2} \approx 0.32. \end{aligned}$$

The Poisson distribution is often used in situations where we are counting the number of successes in a particular region or interval of time, where there are a large number of trials, each with a small probability of success. The Poisson paradigm says in situations like this, we can approximate the number of successes by a Poisson distribution. It is more general than Theorem 27.1, as we relax the assumption of independence and identical events.

Proposition 27.1 (Poisson paradigm)**.** *Let $A_1, \dots, A_n$ 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. Then $X = \sum_{j=1}^n I(A_j)$, that is how many of the A_j occur, is approximately distributed as Pois(λ) with $\lambda = \sum_{j=1}^n p_j$.*

The Poisson paradigm is also called the *law of rare events.* The interpretation of “rare” is that the p_j are small, but λ is relatively stable. The number of events that occur may not be exactly Poisson, but the Poisson distribution often gives good approximations. Note that the conditions for the Poisson paradigm to hold are fairly flexible: the n trials can have different success probabilities, and the trials don’t have to be independent, though they should not be very dependent. So there are a wide variety of situations that can be cast in terms of the Poisson paradigm. This makes the Poisson a very popular model.

Poisson distribution is also used to model the number of **events occurring randomly over time** with **constant rate**, such as the number of customers visiting a store, the number of phone calls to a call center, and so on.

Why the random occurrence of events has anything to do with the Poisson distribution? Consider in this way: one can divide the time line into infinitely small intervals (e.g. milliseconds). In each interval, an event either happens or not. The chance that an event occurs in a millisecond is very small. While there are infinitely many trials. So counting events occurring randomly at a fixed average rate over time is mathematically equivalent to counting rare events in many trials.

Definition 27.2 (Poisson process)**.** A sequence of arrivals in continuous time is a Poisson process with rate λ per unit of time if

- The number of arrivals in an interval of length t is distributed $\text{Pois}(\lambda t)$;
- The numbers of arrivals in disjoint time intervals are independent.

28 Birthday problem revisited

The beauty if approximating discrete problems by continuous function is that it makes calculation easier. Now we revisit the birthday problem with Poisson distribution.

Example 28.1. If we have m people and $\binom{m}{2}$ pairs. Each pair of people has probability $p = 1/365$ of having the same birthday. Find the probability of at least one match.

Solution. The probability of match is small, and the number of pairs is large. We consider using the Poisson paradigm to approximate the number X of birthday matches. $X \approx Pois(\lambda)$ where $\lambda = \binom{m}{2}\frac{1}{365}$. Then the probability of at least one match is

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda}.$$

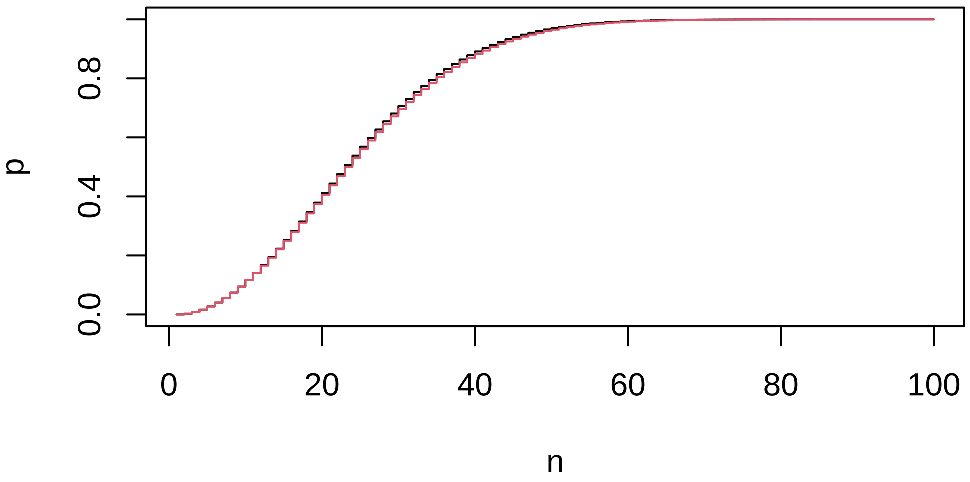
For $m = 23$, $\lambda = 253/365$ and $1 - e^{-\lambda} \approx 0.5$, which agrees with our previous finding that we need 23 people to have 50% chance of a birthday match.

Example 28.2. Continued with the assumption above. What's the probability of two people who were born not only on the same day, but also at the same hour and the same minute?

Solution. This is the birthday problem with $c = 365 \cdot 24 \cdot 60 = 525600$ categories rather than 365 categories. By Poisson approximation, the probability of at least one match is approximately $1 - e^{-\lambda_1}$ where $\lambda_1 = \binom{m}{2}\frac{1}{525600}$. This would require $m = 854$ to reach the break even point, 50% chance of getting a match.

You may wonder how good the Poisson approximation is. We can compare it with the true values.

```r
# compute the probability of coincidences for 1,2...100 people
n <- 1:100
p <- sapply(n, pbirthday)

# approximate the probability by Poisson paradigm
lambda <- choose(n, 2)/365
q <- 1 - exp(-lambda)

# black line is the true probability
# red line is the Poisson approximation
plot(n, p, type = "s")
lines(n, q, col = 2, type="s")
```



29 Convolution

A convolution is a sum of independent random variables. The main task in this section is to determine the distribution of $T = X + Y$, where X and Y are independent random variables whose distributions are known.

Theorem 29.1 (Convolution)**.** *If X and Y are independent discrete random variables, then the PMF of their sum $T = X + Y$ 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 independent continuous random variables, then the PDF of their sum $T = X+Y$ is

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

Theorem 29.2 (Sum of Binomial random variables)**.** *Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ be two independent Binomial random variables. Then $X + Y \sim Bin(n + m, p)$.*

Proof. We have proved the theorem in Theorem 19.1. Here is another way to prove it using convolution.

$$
\begin{aligned}
P(X+Y=k) &= \sum_{i=0}^{k} P(X=i)P(Y=k-i) \\
&= \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \binom{m}{k-i} p^{k-i} (1-p)^{m-k+i} \\
&= \sum_{i=0}^{k} \binom{n}{i} \binom{m}{k-i} p^k (1-p)^{m+n-k} \\
&= p^k (1-p)^{m+n-k} \sum_{i=0}^{k} \binom{n}{i} \binom{m}{k-i} \\
&= p^k (1-p)^{m+n-k} \binom{n+m}{k}.
\end{aligned}
$$

The last step: $\binom{n+m}{k} = \sum_{i=0}^{k} \binom{n}{i}\binom{m}{k-i}$
is known as the Vandermonde's identity. □

Example 29.1 (Sum of Poisson random variables)**.** If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, and X, Y are independent, then $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Proof. Intuitively, X is the number of events occurring at rate λ_1; Y is the number of events occurring at rate λ_2. Therefore, $X + Y$ should be events occurring at rate $\lambda_1 + \lambda_2$.

To get the PMF of $X + Y$, condition on X and use the law of total probability:

$$
\begin{aligned}
P(X+Y=k) &= \sum_{j=0}^{k} P(Y=k-j)P(X=j) \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!} \cdot \frac{e^{-\lambda_1}\lambda_1^{j}}{j!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1+\lambda_2)^k.
\end{aligned}
$$

We thus arrive at the PMF for $\text{Pois}(\lambda_1 + \lambda_2)$. Intuitively, if there are two different types of events occurring at rates λ_1 and λ_2, independently, then the overall event rate is $\lambda_1 + \lambda_2$. □

30 Dice rolling formula*

The Binomial distribution gives the formula for the probability of observing k heads when flipping n coins. Can we find a formula for the probability of getting a total of p points when rolling n dice?

The probability of obtaining p points on n s-sided dice can be computed as the coefficient of x^p in

$$f(x) = (x + x^2 + ... + x^s)^n$$

since each possible arrangement contributes one term.

$$f(x) = x^n(1 + x + \cdots + x^{s-1})^n = x^n \left(\frac{1 - x^s}{1 - x}\right)^n$$

To obtain the coefficient of x^p, expand the binomial power:

$$x^n(1 - x^s)^n(1 - x)^{-n} = x^n \sum_{k=0}^{n} (-1)^k \binom{n}{k} x^{sk} \sum_{l=0}^{\infty} \binom{n + l - 1}{l} x^l$$

The coefficient of x^p include all terms with $p = n + sk + l$. Therefore,

$$c_p = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{p - sk - 1}{p - sk - n}$$

But $p - sk - n > 0$ only when $k < (p - n)/s$, so the other terms do not contribute. Furthermore, applying the symmetric property of the binomial formula, we have

$$\binom{p - sk - 1}{p - sk - n} = \binom{p - sk - 1}{n - 1}$$

Therefore, the probability of getting p points when rolling n s-sided dice is given by

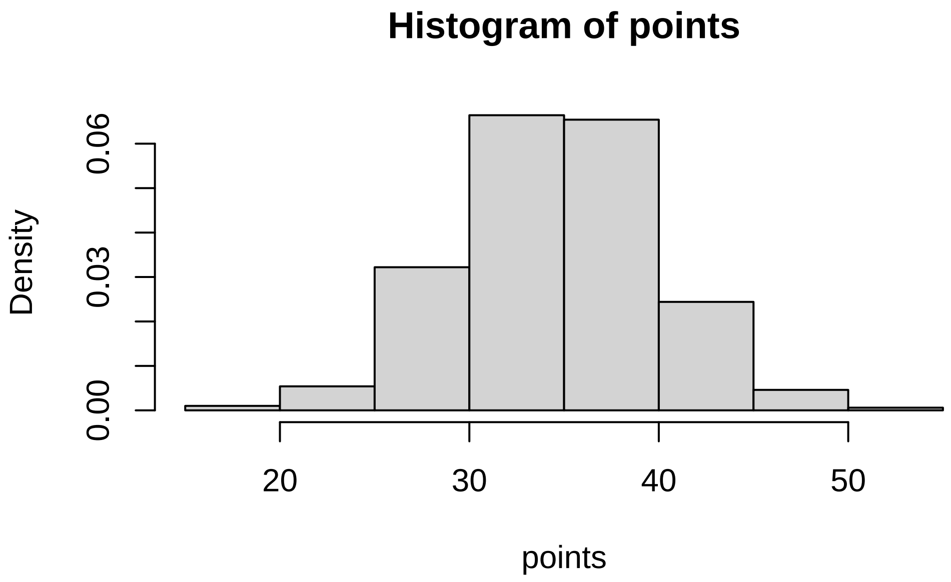
$$f(p,n,s) = \sum_{k=0}^{\lfloor (p-n)/s \rfloor} (-1)^k \binom{n}{k} \binom{p-sk-1}{n-1}.$$

Binomial formula for negative n

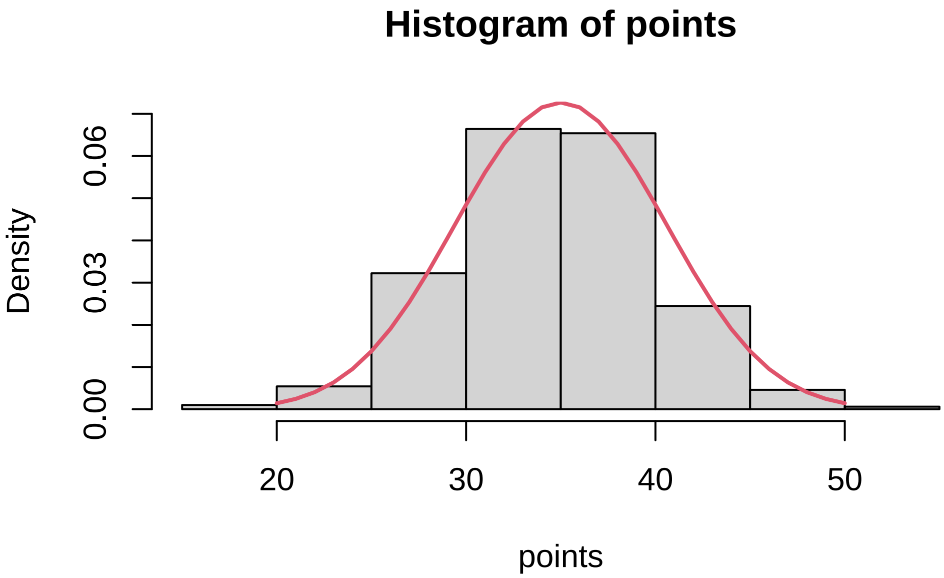
$$\begin{aligned}
\binom{-n}{k} &= \prod_{i=0}^{k-1} \frac{-n-i}{k-i} = (-1)^k \prod_{i=0}^{k-1} \frac{n+i}{k-i} \\
&= (-1)^k \frac{n(n+1)\ldots(n+k-1)}{k!} \\
&= (-1)^k \frac{(n+k-1)!}{k!(n-1)!} \\
&= (-1)^k \binom{n+k-1}{k}
\end{aligned}$$

We can verify our formula by simulating the dice rolling game.

```r
set.seed(0)

# simulates rolling n dice and returns the sum
roll_dice <- function(n, s=6) {
  sum(sample(seq(1,s), n, replace = T))
}

# rolling 10 dice 1000 times and collect the results
points <- replicate(1000, roll_dice(10))

# distribution of the sum of points
hist(points, freq = F)
```



```
# formula for computing probability of dice points
dice_formula <- function(p, n, s=6) {
  prob <- 1/s^n*sum(
    sapply(seq(0, floor((p-n)/s)),
           function(k) (-1)^k*choose(n,k)*choose(p-s*k-1,n-1)))
}

# computing the probability of getting 20-50 when rolling 10 dice
x <- 20:50;
y <- sapply(x, function(p) dice_formula(p,n=10))

# overlay the formula on the histogram
# it turns out the formula does a nice job!
hist(points, ylim = c(0, 0.07), freq = F)
lines(x, y, col = 2, lwd=2)
```


Histogram of points
Density
0.00
0.03
0.06
20
30
40
50
points

31 Application: seller ratings*

This example involves multiple types of discrete distributions. The technique used to solve this problem aligns with Bayesian inference, which is beyond the scope of this course. However, it remains an interesting case. The procedure illustrates the process of statistical modeling: we begin with an assumption and a proposed statistical model, then update it with new data. Finally, we draw inferences based on the model, typically addressing the question we aim to answer. You are not required to understand everything in this example. Nonetheless, it helps to develop a mindset of statistical inference early in the study.

Suppose you are shopping a product online. There are three sellers with the following ratings:

- Seller 1: 100% positive out of 10 reviews
- Seller 2: 96% positive out of 50 reviews
- Seller 3: 93% positive out of 200 reviews

Which seller is likely to give the best service?

The problem is intriguing because it is obvious that higher ratings do not necessarily means higher satisfaction. We have to weight in the number of reviews. The more reviews, the more trustworthy the ratings are. Let $X_j^{(i)}$ be a random variable that means consumer j is satisfied with seller i, where $i \in \{1, 2, 3\}$. Assume $X_j^{(i)}$ follows a Bernoulli distribution:

$$X_j^{(i)} = \begin{cases} 1 & \text{satisfied with probability } \theta_i \\ 0 & \text{otherwise} \end{cases}$$

where θ_i is an unknown parameter of seller i that captures their “genuine” satisfaction rate. We assume the consumers independently write their ratings. The overall positive rate of seller i is therefore $R_i = \frac{1}{n_i} \sum_j X_j^{(i)}$ where n_i is the total number of reviews. We want to infer the value of θ_i from their observed positive rate R_i. From now on we drop the seller index i to simply the notation since it is symmetric for all sellers.

Because we have no prior knowledge about θ. We assume that θ takes any value from $[0, 1]$ equally likely, i.e. $\theta \sim \text{Unif}(0, 1)$. Assuming each X_j is independent and identical, then

$$S = X_1 + X_2 + \cdots + X_n$$

follows the Binomial distribution with PMF:

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Our goal is to find: $p(\theta|k)$. Recall that the Bayes' rule allows us to invert the conditional probability:

$$p(\theta|k) = \frac{p(k|\theta)p(\theta)}{p(k)} = \frac{p(k|\theta)p(\theta)}{\int_{-\infty}^{\infty} p(k|\theta)p(\theta)d\theta}$$

Since $\theta \sim \text{Unif}(0,1)$, we have

$$p(\theta) = \begin{cases} 1 & \text{if } \theta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

We now focus on $\theta \in [0,1]$, since the probability is 0 otherwise. Substitute in the PMF of the Binomial distribution,

$$p(\theta|k) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}$$

The hard part is to evaluate the integral. We state without proof (this is known as the Beta function, which we will prove in later chapters):

$$\int_0^1 \theta^k(1-\theta)^{n-k} = \frac{k!(n-k)!}{(n+1)!}$$

Therefore,

$$p(\theta|k) = \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

Now suppose you are the next customer. The probability that you would be satisfied is

$$\begin{aligned}
P(X_{n+1}=1|S=k) = & \int_0^1 P(x_{n+1}=1|\theta)p(\theta|k)d\theta \\
= & \int_0^1 \theta \times \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}d\theta \\
= & \frac{(n+1)!}{k!(n-k)!}\int_0^1 \theta^{k+1}(1-\theta)^{(n+1)-(k+1)}d\theta \\
= & \frac{(n+1)!}{k!(n-k)!} \times \frac{(k+1)!(n-k)!}{(n+2)!} \\
= & \frac{k+1}{n+2}.
\end{aligned}$$

Now we substitute the ratings for the three sellers:

- Seller 1: $n = 10, k = 10$
- Seller 2: $n = 50, k = 48$
- Seller 3: $n = 200, k = 186$

The probabilities that you would be satisfied with each seller are: 92%, 94%, 93%. The result is known as the **Laplace's rule of succession**. The rule of thumb is, pretending we have too more reviews: one is positive, the other is negative. Compute the satisfaction rate as $\frac{k+1}{n+2}$.

Part IV

Expectation and Variance

32 Expectation revisited

Definition 32.1. For discrete random variable X, the expectation of X is defined as

$$E(X) = \sum_{\text{all } x} xP(X = x);$$

For continuous random variable X with density function $f(x)$, the expectation is defined as

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

Proposition 32.1 (Linearity). *For random variables $X_1, X_2, \dots, X_n$, regardless of their dependencies, it holds that*

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

Proof. We prove the simplest case $E(X + Y) = E(X) + E(Y)$.

$$\begin{aligned}
E(X+Y) &= \sum_{z=x+y} zP(X+Y=z) \\
&= \sum_x \sum_y (x+y)P(X=x, Y=y) \\
&= \sum_x \sum_y xP(X=x, Y=y) + \sum_x \sum_y yP(X=x, Y=y) \\
&= \sum_x x \sum_y P(X=x, Y=y) + \sum_y y \sum_x P(X=x, Y=y) \\
&= \sum_x xP((X=x) \cap \bigcup_{\text{all } y} (Y=y)) + \sum_y yP(\bigcup_{\text{all } x} (X=x) \cap (Y=y)) \\
&= \sum_x xP(X=x) + \sum_y yP(Y=y) \\
&= E(X) + E(Y).
\end{aligned}$$

□

Proposition 32.2. *Further properties on the linearity of expectations:*

- *If* $Y = aX + b$*, then* $E(Y) = aE(X) + b$*.*
- $E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$

Proposition 32.3 (Multiplication)**.** *If* X *and* Y *are independent, we have*

$$E(XY) = E(X)E(Y).$$

In general, if $X_1, \dots, X_n$ *are independent, we have*

$$E(X_1X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n).$$

Proof. For discrete and independent X, Y,

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \quad \text{if independent} \\ &= \sum_x xP(X = x) \sum_y yP(Y = y) \\ &= E(X)E(Y). \end{aligned}$$

□

🔥 Multiplication does not hold without independence

It is misleadingly natural to extend the generality of the addition rule to multiplication. But the multiplication rule of expectation is very restrictive. Always remember to check independence before applying the multiplication rule.

ℹ Sufficient but not necessary condition

If X, Y are independent, it follows that $E(XY) = E(X)E(Y)$. However, the latter does not imply independence. Consider a counter-example,

$$X = \begin{cases} 1 & \text{with prob. } 1/2 \\ 0 & \text{with prob. } 1/2 \end{cases}, \quad Z = \begin{cases} 1 & \text{with prob. } 1/2 \\ -1 & \text{with prob. } 1/2 \end{cases};$$

Then

$$Y = XZ = \begin{cases} -1 & \text{with prob. } 1/4 \\ 0 & \text{with prob. } 1/2 \\ 1 & \text{with prob. } 1/4 \end{cases}.$$

We have $E(X) = 1/2$, $E(Y) = 0$, $E(XY) = 0$. So $E(XY) = E(X)E(Y)$. But clearly X, Y are not independent.

Proposition 32.4 (Law of total expectation)**.** *Let $\{A_i\}$ be a finite (or countable) partition of the sample space, then*

$$E(X) = \sum_i E(X|A_i)P(A_i).$$

Theorem 32.1 (Law of the unconscious statistician (LOTUS))**.** *Let X be a random variable, and g be a real-valued function of a real variable. If X has a discrete distribution, then*

$$E[g(X)] = \sum_{all\ x} g(x)P(X = x).$$

LOTUS says we can compute the expectation of $g(X)$ without knowing the PMF of $g(X)$.

Example 32.1. Compute $E(X)$ and $E(X^2)$ given the following distribution:

$$f(x) = \begin{cases} \frac{1}{4}, & x = 0 \\ \frac{1}{2}, & x = 1 \\ \frac{1}{4}, & x = 2 \end{cases}$$

Solution. Compute the expectations of X by definition:

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

Compute the expectations of X^2 by LOTUS:

$$E(X^2) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = \frac{3}{2}.$$

Note that $E(X^2) \neq [E(X)]^2$.

Don’t pull non-linear functions out of expectation

In general, $E[g(X)] \neq g(E(X))$. Linearity implies $E[g(X)] = g(E(X))$ if g is a linear function. For a nonlinear function g, you can’t pull function g out of expectation E. The right way to find $E[g(X)]$ is with LOTUS.

Example 32.2 (St. Petersburg Paradox)**.** Flip a fair coin over and over again until the head lands the first time. You will win 2^k dollars if the head lands in the k-th trial (including the successful trial). What is the expected payoff of this game?

Solution. Let $X = 2^k$. We want to find $E(X)$. The probability of the first head showing up in the k-th trial is $\frac{1}{2^k}$. Therefore,

$$E(X) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty$$

The expected payoff is infinitely high! This is against most people's intuition, which is likely to be a small number. This is because we mistakenly go through the calculation $E(X) = E(2^k) = 2^{E(k)}$ in our mind. $E(k)$ the expected number of flips before a head is 2. Thus, $2^{E(k)} = 4$.

Another way to resolve the paradox is that we don't typically reason about infinity. No one would play this game infinitely many times. For finite number of plays, the probability of getting very large payoff, say 2^{100}, is none. We can demonstrate this with a simulation.

```r
# number of simulations
N <- 1000

# store simulated results
X <- numeric(N)

set.seed(0)

# run simulation
for (i in 1:N) {
  # start with the initial reward
  x <- 2
  # flip a coin until it lands tails
  while (runif(1) < 0.5) {
    x <- x * 2
  }
  # store the reward for this simulation
  X[i] <- x
}

cat("Expected Reward:", mean(X))
```

```
Expected Reward: 12.938
```

33 Life expectancy

Life expectancy is the average number of years a person is expected to live. It is a crucial indicator of the quality of living and one of the three components of the Human Development Index (HDI) (the other two components are education and per capita GDP). Here is a toy example to compute life expectancy with hypothetical data.[1]

(1)	(2)	(3)	(4)		(5)	(6)
Age	Population	Mortality rates	# Survive		# Died at age	P(Age)
0	200	1%	1000		10	1%
20	300	2%	990	=1000(1-1%)	20	2%
40	250	10%	970	=990(1-2%)	97	10%
60	150	20%	873	=970(1-10%)	175	17%
80	100	100%	699	=873(1-20%)	699	70%
Total	1000					

To simplify our analysis, we will assume there are only five possible ages: 0, 20, 40, 60, and 80. A baby is born at age 0, and can either die at that age or survive to age 20. We intentionally exclude intermediate ages such as 5 and 10 for the sake of computational simplicity.

It's important to note that life expectancy is not the same as the average age of the population. For instance, based on the hypothetical data presented, the average age can be calculated as:

$$\overline{\text{Age}} = (0 \times 200 + 20 \times 300 + 40 \times 250 + 60 \times 150 + 80 \times 100)/1000 = 33.$$

However, the expected age, denoted as $E(\text{Age})$, is defined as:

$$E(\text{Age}) = \sum \text{Age} \times P(\text{Age}).$$

To compute this expected value, we need to determine $P(\text{Age})$, the probability of living to a specific age or dying at that age. This requires consideration of the mortality rate at each age, which is given in Column 3.

[1] This is an overly simplified example that only serves to clarify the definition of expectation. See this tutorial from MEASURE Evaluation for the actual computation of life expectancy.

Assuming 1000 babies are born at age 0, with a mortality rate of 1% at that age, we find that 99% of the babies survive to age 20. Thus, the number of babies that survive to age 20 is: $1000 \times (1-1\%) = 990$. We can apply similar calculations to determine the number of survivors at each subsequent age.

The number of individuals who die at a specific age (Column 5) is the difference between the number of survivors at that age and the next (Column 4). To find the probability of living to a specific age, we compute: $P(\text{Age}) = \text{Column } 4/1000$.

Finally, we compute the expected value of age (or life expectancy) as follows:

$$E(Age) = 0 \times 1\% + 20 \times 2\% + 40 \times 10\% + 60 \times 17\% + 80 \times 70\% = 70.6.$$

This figure differs from the average age. Since the mortality rate is low at younger ages, the probabilities $P(\text{Age})$ for these ages are also low, while they are higher for older ages. This example illustrates the distinction between average and expected values. In everyday conversation, we may use these terms interchangeably, but in certain contexts, expected values can significantly differ from averages.

34 Two envelope paradox*

Example 34.1 (Two-envelope paradox)**.** Imagine you are given two identical envelopes, each containing money. One contains twice as much as the other. You may pick one envelope and keep the money it contains. Having chosen an envelope at will, but before inspecting it, you are given the chance to switch envelopes. Should you switch?

The paradox arises when you try to solve the expectation. Let A denote the amount of money in the envelope you have chosen, and B denote the amount of money in the other envelope.

We know B is either twice as much as A, or half as much as A. Each with probability 1/2. So

$$E(B) = \frac{1}{2}(2A) + \frac{1}{2}(A/2) = \frac{5}{4}A$$

Since $E(B) > A$, you should always switch! However, after you switch to B, by the same argument, you should switch back to A. You you switch back and forth indefinitely!

Where do things go wrong? The error in this calculation lies in a subtle misunderstanding: the two As in the calculation actually represent different values, that are incorrectly equated. In particular, the $2A$ represents the expected value in the other envelope given that it is the larger one, and the $A/2$ represents the expected value in the other envelope given that it is the smaller one.

$$E(B) = E(B|B < A)P(B < A) + E(B|B > A)P(B > A)$$

Suppose the amount of money in the two envelopes are a and $2a$ respectively. $E(B|B < A) = a$ and $E(B|B > A) = 2a$. Therefore,

$$E(B) = \frac{1}{2}a + \frac{1}{2}2a = \frac{3}{2}a.$$

The same calculation applies to $E(A)$. Thus, $E(A) = E(B)$.

35 Linearity and indicators

Definition 35.1 (Indicator variable)**.** An indicator variable $\mathbb{1}_A$ for an event A is a random variable defined as:

$$\mathbb{1}_A = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{if event } A \text{ does not occur.} \end{cases}$$

The indicator variable $\mathbb{1}_A$ "indicates" whether the event A happens (1) or not (0).

The expected value of an indicator variable is equal to the probability of the event A:

$$E[\mathbb{1}_A] = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

This is known as the **fundamental bridge**, as it allows us to convert between probability and expectation.

Indicator variables are often used in linearity of expectation calculations. This allows us to break down a problem into easy-to-solve small problems. For example, if $X = \sum_{i=1}^{n} \mathbb{1}_{A_i}$, then:

$$E[X] = \sum_{i=1}^{n} E[\mathbb{1}_{A_i}] = \sum_{i=1}^{n} P(A_i)$$

Example 35.1. In a group of n people, what is the expected number of distinct birthdays among the n people (the expected number of days on which at least one of the people was born)? What is the expected number of people sharing a birthday (any day)?

Solution. Let X be the number of distinct birthdays, and write $X = I_1 + \cdots + I_{365}$, where

$$I_j = \begin{cases} 1 & \text{if someone was born on day } j \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\begin{aligned} E(I_j) &= P(\text{someone was born on day } j) \\ &= 1 - P(\text{no one was born on day } j) \\ &= 1 - \left(\frac{364}{365}\right)^n . \end{aligned}$$

Then by linearity,

$$E(X) = 365 \left(1 - \left(\frac{364}{365}\right)^n\right).$$

Let Y be the number of people sharing a birthday, and $Y = J_1 + \cdots + J_n$ where J_k is an indicator that the j-th person shares his birthday with somebody else.

$$\begin{aligned} E(J_k) &= P(\text{someone shares birthday with } k) \\ &= 1 - P(\text{no one shares birthday with } k) \\ &= 1 - \left(\frac{364}{365}\right)^{n-1}. \end{aligned}$$

Therefore,

$$E(Y) = \sum_{k=1}^{n} E(J_k) = n\left(1 - \left(\frac{364}{365}\right)^{n-1}\right).$$

For some numeric values, $E(Y) = 2.3$ if $n = 30$; $E(Y) = 6.3$ if $n = 50$.

Example 35.2. Let Π be a permutation over $\{1, 2, \ldots, n\}$. That is a reordering of the numbers. A fixed point of a permutation are the points not moved by the permutation. For example, in the permutation below

$$\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ \Pi & 2 & 4 & 3 & 1 \end{array}$$

The fixed point is 3. Find the expected number of fixed points of a random permutation.

Solution. Let X be the number of fixed points of a random permutation. Then $X = \sum_{k=1}^{n} 1_{\Pi(k)=k}$ where $1_{\Pi(k)=k}$ indicates the k-th number stays the same after the permutation. By linearity,

$$E(X) = E\left(\sum_{k=1}^{n} 1_{\Pi(k)=k}\right) = \sum_{k=1}^{n} E\left(1_{\Pi(k)=k}\right) = \sum_{k=1}^{n} \frac{1}{n} = 1.$$

Example 35.3 (Buffon's needle)**.** A plan is ruled by the lines $y = 0, \pm 1, \pm 2, \ldots$ and a needle of unit length is cast randomly on to the plane. What is the probability that it intersects some line?

Solution. Here we sketch an intuitive approach. A more rigorous one will be given in Example 44.2.

Let X be the number of times the needle crosses a line. The needle can cross a line either 1 or 0 times. Thus, $P(\text{intersection}) = E(X)$.

Consider dropping any (continuous) curve of unit length onto the surface. Divide the curve into N straight line segments, each of length $1/N$. Let X_i be the indicator for the i-th segment crossing a line. Then,

$$E(X) = E\left(\sum X_i\right) = \sum E(X_i) = N \cdot E(X_i).$$

We don't necessarily have to compute this expectation, but by this line of reasoning: $E(X)$ is proportional to the length of the curve, *regardless* the shape of the curve. If we can compute $E(X)$ for some curve, then $E(X)$ is the same for all curves with the same length.

Consider a circle of diameter $d = 1$. The circle always crosses the lines twice for sure. That is, $E(X_{\text{circle}}) = 2$. The length of the circle is π. Therefore, for any curve (including a needle) of length π we have $E(X_\pi) = 2$.

For a needle of unit length, scale it down by π, we have $E(X) = 2/\pi$. Therefore,

$$P(\text{intersection}) = \frac{2}{\pi}.$$

```
# number of simulations
N <- 10000

# number of crossings
X <- numeric(N)

set.seed(0)

# run simulation
for (i in 1:N) {

  # randomly generate the position of the needle's midpoint
  # distance from the nearest line
  y <- runif(1, min = 0, max = 1/2)

  # randomly generate the angle of the needle (in radians)
   <- runif(1, min = 0, max = pi/2)

  # check if the needle crosses a line
  if (y <= 1/2 * sin( )) X[i] <- 1
  else X[i] <- 0
}

 <- 2 / mean(X)

cat("Estimated value of  :", )
```

```
Estimated value of  : 3.140704
```

36 Median and mode

The mean is called a measure of *central tendency* because it tells us something about the center of a distribution, specifically its center of mass. Other measures of central tendency that are commonly used in statistics are the median and the mode, which we now define.

Definition 36.1 (Median)**.** We say that c is a median of a random variable X if

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

Intuitively, the median is a value c such that half the mass of the distribution falls on either side of c (or as close to half as possible, for discrete random variables). Note that the condition given above is more general than

$$P(X \leq c) = P(X \geq c) = \frac{1}{2}$$

Consider a discrete distribution as follows:

$$P(X = k) = \begin{cases} \frac{1}{3}, & k = 1 \\ \frac{1}{2}, & k = 2 \\ \frac{1}{6}, & k = 3 \end{cases}$$

In this case, 2 is a median since $P(X \leq 2) = 5/6 \geq 1/2$ and $P(X \geq 2) = 2/3 \geq 1/2$. However, $P(X \leq 2) \neq P(X \geq 2)$. For strictly continuous random variable X, Definition 36.1 does imply

$$P(X \leq c) = P(X \geq c) = \frac{1}{2}$$

Since the CDF of X satisfies $F(c) \geq 1/2$ and $1 - F(c) \geq 1/2$, which implies $F(c) = 1/2$. Moreover, if the CDF of X is strictly increasing, $F^{-1}(1/2)$ is the unique median.

Definition 36.2 (Mode)**.** For a discrete random variable X, we say that c is a mode of X if it maximizes the PMF:

$$P(X = c) \geq P(X = x) \quad \text{for all } x.$$

For a continuous random variable X with PDF f, we say that c is a mode if it maximizes the PDF:

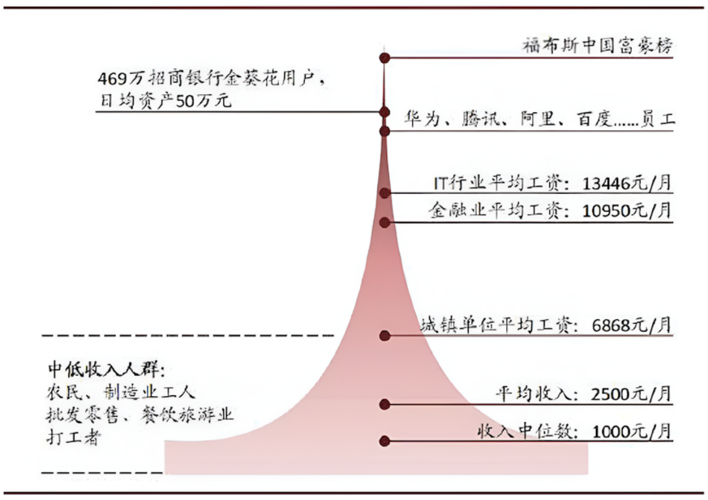
$$f(c) \geq f(x) \quad \text{for all } x.$$

Intuitively, the mode is a value that has the greatest mass or density out of all values in the support of X.

Note

A distribution can have multiple medians and multiple modes. Medians have to occur side by side; modes can occur all over the distribution.

Example 36.1 (Income distribution)**.** The main reason why the median is sometimes preferred over the mean is that the median is more robust to extreme values. A typical example is the income distribution. Higher incomes are rare, but their absolute values are high. Thus, the mean income tends be higher than what the mass of the population would earn. But the median is more robust to extreme values and is closer to the earnings of an "average" person. For example, the average monthly income in China is ¥$2,500$ in 2019, but the median is only ¥$1,000$.



注：城镇单位平均工资为2018年数据，其他指标为2019年数据。

Theorem 36.1. *Let X be an random variable with mean μ , and let m be a median of X.*

- *A value of c that minimizes the mean squared error $E\left(X-c\right)^2$ is $c=\mu$.*
- *A value of c that minimizes the mean absolute error $E\left|X-c\right|$ is $c=m$.*

Proof.

1) Minimizing the mean squared error $E[(X-c)^2]$. Expand the mean squared error:

$$E[(X-c)^2] = E[X^2 - 2cX + c^2] = E[X^2] - 2cE[X] + c^2.$$

To find the value of c that minimizes this expression, take the derivative with respect to c and set it to zero:

$$\frac{d}{dc}E[(X-c)^2] = -2E[X] + 2c = 0$$

This implies $c = \mu$. We can confirm with second-order condition that $c = \mu$ is indeed a minimizer.

2) Minimizing the mean absolute error $E\,|X - c|$. We prove this assuming X is continuous.

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

Take derivative with respect to c, applying the Leibniz's rule:

$$(c-x)f(x)\frac{d}{dc}c + \int_{-\infty}^{c} f(x)dx - (x-c)f(x)\frac{d}{dc}c + \int_{c}^{\infty} (-f(x))dx = 0$$

The first-order condition resolves to

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

which is exactly the definition of a median.

□

37 Variance

Expectation is the most commonly used summary of a distribution, as it indicates where values are likely centered. However, it provides limited insight into the distribution's overall shape. For example, two random variables might have the same mean, yet one could have values spread far from the mean while the other has values tightly clustered around it. Variance, on the other hand, describes how far values in a distribution typically deviate from the mean, offering a measure of the distribution's dispersion.

Definition 37.1 (Variance)**.** The variance of a random variable X is defined as

$$Var(X) = E\left[X - E(X)\right]^2.$$

By convention, variance is also denoted by Greek letter σ^2, where $\sigma = \sqrt{Var(X)}$ is called the **standard deviation**.

Variance measures how far X typically deviates from its mean, but instead of averaging the differences, we average the squared differences to ensure both positive and negative deviations contribute. The expected deviation, $E(X - E(X))$, is always zero, so squaring avoids this cancellation. Since variance is in squared units, we take the square root to get the standard deviation, restoring the original units.

Why squared deviation?

We can measure the dispersion of a distribution in different ways. For example, $E(|X - E(X)|)$ is also a possible choice. But it is less common because the absolute value function isn't differentiable. Besides, squaring connects to geometric concepts like the distance formula and Pythagorean theorem, which have useful statistical meanings.

Sample variance

Definition 37.1 gives the theoretical variance of a distribution. With finite sample from the distribution, we estimate the variance with sample observations:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Note that we divide by $n-1$ not n. Why? Because we want an unbiased estimator. We

will discuss this later in detail. But here is a sketch of the reasoning.
First note:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Take expectations of both sides:

$$E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = (n-1)\sigma^2$$

Dividing by $n-1$ makes $E(s^2) = \sigma^2$.

Theorem 37.1. *For any random variable X,*

$$Var(X) = E(X^2) - (EX)^2.$$

Proof. Let $\mu = E(X)$. By definition,

$$\begin{aligned} Var(X) &= E(X-\mu)^2 = E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2. \end{aligned}$$

□

Example 37.1. Find the variance for $X \sim \text{Bern}(p)$.

$$Var(X) = E(X^2) - E^2(X) = p - p^2 = p(1-p).$$

Proposition 37.1. *Variance has the following properties:*

- $Var(X) \geq 0$
- $Var(X + c) = Var(X)$
- $Var(cX) = c^2 Var(X)$
- *If X, Y are independent, $Var(X+Y) = Var(X) + Var(Y)$.*
- *If $X_1, X_2, \dots, X_n$ are independent, $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$.*

Example 37.2 (Variance of Binomial distribution)**.** Find the variance for $X \sim \text{Bin}(n,p)$. $X = X_1 + \cdots + X_n$ where X_i are *i.i.d* Bernoulli distributions

$$Var(X) \overset{iid}{=} \sum_{i=1}^{n} Var(X_i) = np(1-p).$$

Example 37.3 (Variance of Poisson distribution)**.** Let $X \sim \text{Pois}(\lambda)$. To find the variance, we first compute $E(X^2)$. By LOTUS,

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!}$$

Differentiate $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ on both sides with respect to λ and multiply (replenish) again by λ:

$$\sum_{k-1}^{\infty} k \frac{\lambda^k}{k!} = \lambda e^{\lambda}$$

Repeat:

$$\sum_{k-1}^{\infty} k^2 \frac{\lambda^k}{k!} = \lambda(e^{\lambda} + \lambda e^{\lambda})$$

Therefore, we have

$$E(X^2) = e^{-\lambda}(\lambda + \lambda^2)e^{\lambda} = \lambda + \lambda^2$$

Finally,

$$Var(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

38 Covariance

For more than one random variable, it is also of interest to know the relationship between them. Are they dependent? How strong is the dependence? Covariance and correlation are intended to measure that dependence. But they only capture a particular type of dependence, namely linear dependence.

Definition 38.1 (Covariance)**.** The covariance between random variables X and Y is defined as

$$Cov(X, Y) = E[(X - EX)(Y - EY)].$$

The covariance between X and Y reflects how much X and Y *simultaneously* deviate from their respective means.

Theorem 38.1. *For any random variables X and Y,*

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

Proof. Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$. By definition,

$$\begin{aligned} Cov(X, Y) &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

□

Theorem 38.2. *If X, Y are independent, they are uncorrelated. But the converse is false.*

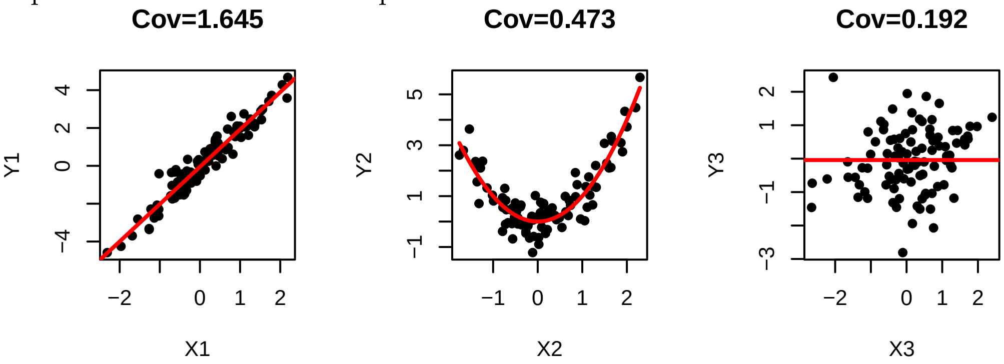
Proof.

1. $Cov(X, Y) = E(XY) - E(X)E(Y)$. Independence implies $E(XY) = E(X)E(Y)$. Thus, $Cov(X, Y) = 0$.
2. $Cov(X, Y) = 0$ does not necessarily imply independence. Consider the following counter example. Let X be a random variable that takes three values -1, 0, 1 with equal probability. And $Y = X^2$. X and Y are clearly dependent. But they their covariance is 0. Since $E(X) = 0$, $E(Y) = 2/3$, $E(XY) = E(X^3) = 0$, $Cov(X, Y) = 0$.

□

Linear dependency

Covariance and correlation provide measures of the extend to which two random variables are linearly related. It is possible that the covariance is 0 even when X and Y are dependent but the relationship is nonlinear.



Proposition 38.1. *Covariance has the following properties:*

- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(cX, Y) = Cov(X, cY) = c\,[Cov(X, Y)]$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Proof. We only prove the variance-covariance property:

$$\begin{aligned} Var(X+Y) &= E[(X+Y-\mu_X-\mu_Y)^2] \\ &= E[(X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y)] \\ &= Var(X)+Var(Y)+2Cov(X,Y). \end{aligned}$$

□

Theorem 38.3. *For random variables $X_1, X_2, \ldots, X_n$, it holds that*

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

If $X_1, X_2, \ldots, X_n$ are identically distributed and have the same covariance relationships (symmetric), then

$$Var\left(\sum_{i=1}^{n} X_i\right) = nVar(X_1) + 2\binom{n}{2} Cov(X_1, X_2).$$

While Cov(X, Y) quantifies how X and Y vary together, its magnitude also depends on the absolute scales of X and Y (multiply X by a constant c, the covariance will be different). To establish a measure of association between X and Y that is unaffected by arbitrary changes in the scales of either variable, we introduce a "standardized covariance" called correlation.

Definition 38.2 (Correlation)**.** The correlation between random variables X and Y is defined as

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

By convention, we denote correlation by Greek letter $\rho \equiv Corr(X, Y)$.

Unlike covariance, scaling X or Y has no effect on the correlation. We can verify this:

$$Corr(cX, Y) = \frac{Cov(cX, Y)}{\sqrt{Var(cX)Var(Y)}} = \frac{cCov(X, Y)}{c\sqrt{Var(X)Var(Y)}} = Corr(X, Y).$$

Theorem 38.4. *For any random variable X and Y,*

$$-1 \leq Corr(X, Y) \leq 1.$$

Proof. Without loss of generality, assume X, Y both have variance 1, since scaling does not change the correlation. Let $\rho = Corr(X, Y) = Cov(X, Y)$. Then

$$\begin{aligned} Var(X + Y) &= Var(X) + Var(Y) + 2Cov(X, Y) = 2 + 2\rho \geq 0, \\ Var(X - Y) &= Var(X) + Var(Y) - 2Cov(X, Y) = 2 - 2\rho \geq 0. \end{aligned}$$

Thus $-1 \leq \rho \leq 1$. □

- X and Y are **positively correlated** if $\rho_{XY} > 0$;
- X and Y are **negatively correlated** if $\rho_{XY} < 0$;
- X and Y are **uncorrelated** if $\rho_{XY} = 0$.

Theorem 38.5. *Suppose that X is a random variable and $Y = aX + b$ for some constants a, b, where $a \neq 0$. If $a > 0$, then $\rho_{XY} = 1$. If $a < 0$, then $\rho_{XY} = -1$.*

Proof. If $Y = aX + b$, then $E(Y) = aE(X) + b$. Thus, $Y - E(Y) = a(X - E(X))$. Therefore,

$$Cov(X, Y) = aE[(X - EX)^2] = aVar(X).$$

Since $Var(Y) = a^2Var(X)$, $\rho_{XY} = \frac{a}{|a|}$. The theorem thus follows. □

Correlation analysis

A **correlation matrix** shows the pairwise correlation coefficients between variables. It's one of the most common tools for exploring relationships in multivariate data.

```
# variables for analysis
vars <- mtcars[, 1:4]

# compute the correlation matrix
print(cor(vars))
```

```
        mpg    cyl   disp     hp
mpg   1.000 -0.852 -0.848 -0.776
cyl  -0.852  1.000  0.902  0.832
disp -0.848  0.902  1.000  0.791
hp   -0.776  0.832  0.791  1.000
```

Example 38.1. Let $X \sim \text{HGeom}(w, b, n)$. Find $Var(X)$.

Solution. Interpret X as the number of white balls in a sample of size n from an box with w white and b black balls. We can represent X as the sum of indicator variables, $X = I_1 + \cdots + I_n$, where I_j is the indicator of the j-th ball in the sample being white. Each I_j has mean $p = w/(w+b)$ and variance $p(1-p)$, but because the I_j are dependent, we cannot simply add their variances. Instead,

$$\begin{aligned}
Var(X) &= Var\left(\sum_{j=1}^{n} I_j\right) \\
&= Var(I_1) + \cdots + Var(I_n) + 2\sum_{i<j} Cov(I_i, I_j) \\
&= np(1-p) + 2\binom{n}{2} Cov(I_i, I_j)
\end{aligned}$$

In the last step, because of symmetry, for every pair i and j, $Cov(I_i, I_j)$ are the same.

$$\begin{aligned}
Cov(I_i, I_j) &= E(I_i I_j) - E(I_i)E(I_j) \\
&= P(i \text{ and } j \text{ both white}) - P(i \text{ is white})P(j \text{ is white}) \\
&= \frac{w}{w+b} \cdot \frac{w-1}{w+b-1} - p^2 \\
&= p\frac{Np-1}{N-1} - p^2 \\
&= \frac{p(p-1)}{N-1}
\end{aligned}$$

where $N = w + b$. Plugging this into the above formula and simplifying, we eventually obtain

$$Var(X) = np(1-p) + n(n-1)\frac{p(p-1)}{N-1} = \frac{N-n}{N-1}np(1-p).$$

This differs from the Binomial variance of $np(1-p)$ by a factor of $\frac{N-n}{N-1}$. This discrepancy arises because the Hypergeometric story involves sampling without replacement. As $N \to \infty$, it becomes extremely unlikely that we would draw the same ball more than once, so sampling with or without replacement essentially become the same.

39 Portfolio allocation*

In the world of finance, one of the most well-established principles is the idea of **diversification**. By combining assets with varying levels of risk and return, investors can reduce the overall risk of their portfolio.

Consider a portfolio of two assets, Asset A and Asset B. Both assets have the same expected return and individual risks (standard deviations), and they are weighted equally in the portfolio.

Asset	Return μ	Risk σ	Weight w
A	10%	15%	50%
B	10%	15%	50%

The expected return of the portfolio is:

$$\mu_P = w_A \mu_A + w_B \mu_B = 10\%$$

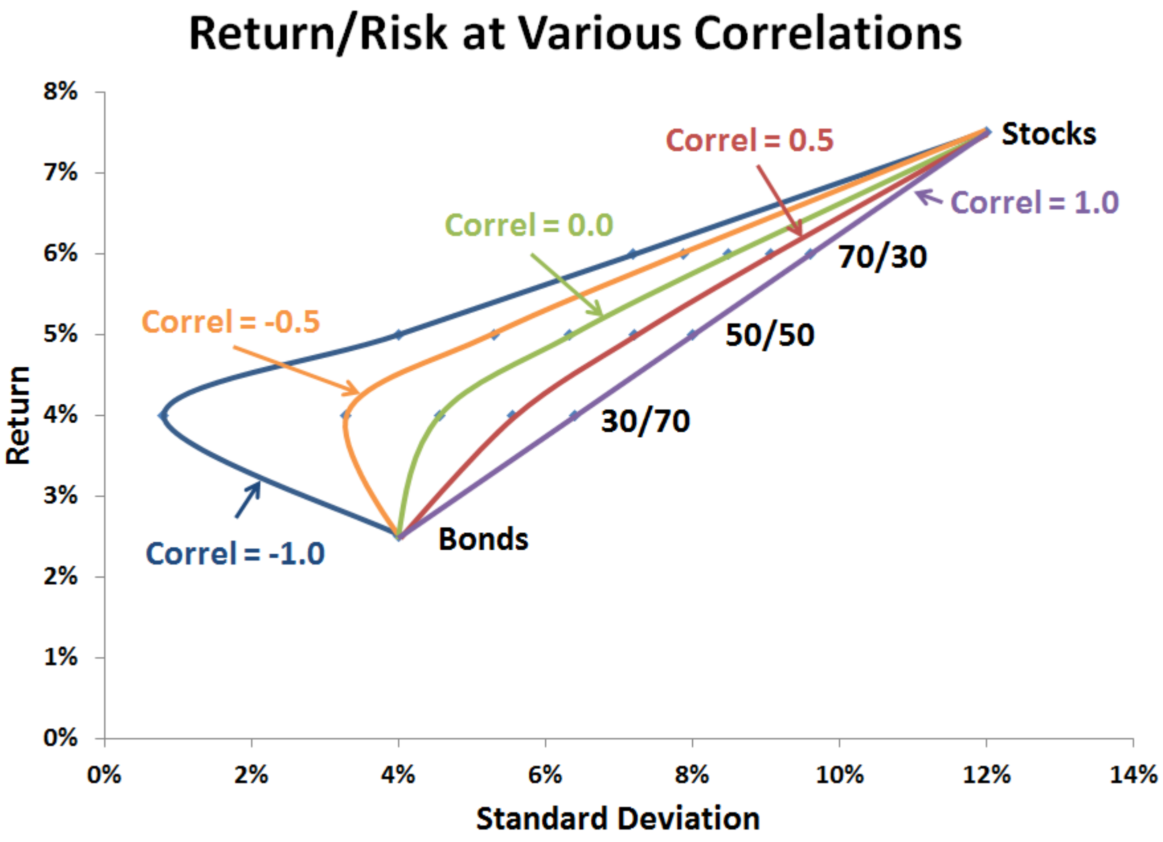
Let's consider the portfolio risk. First, assuming high correlation, $\rho_{AB}^H = 0.8$. The portfolio risk is:

$$\sigma_P^H = \sqrt{w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2 w_A w_B \sigma_A \sigma_B \rho_{AB}^H} \approx 14.2\%$$

If assuming low correlation, $\rho_{AB}^L = 0.2$. The portfolio risk is:

$$\sigma_P^L = \sqrt{w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2 w_A w_B \sigma_A \sigma_B \rho_{AB}^L} \approx 11.6\%$$

As we see, by reducing the correlation between the two assets, we reduced the portfolio risk, though the expected return remains the same. Therefore, diversification is often referred to as a "free lunch" in finance because it allows investors to reduce portfolio risk without sacrificing expected returns.


Return/Risk at Various Correlations
8%
7%
6%
5%
4%
3%
2%
1%
0%
Return
0%
2%
4%
6%
8%
10%
12%
14%
Standard Deviation
Stocks
Correl = 0.5
Correl = 1.0
Correl = 0.0
70/30
Correl = -0.5
50/50
30/70
Bonds
Correl = -1.0

40 Conditional expectation

We have introduced conditional expectation in Definition 26.2. Here we reiterate the definition with continuous random variables.

Definition 40.1 (Conditional expectation)**.** Let X and Y be continuous random variables with joint density $f_{X,Y}(x,y)$, X's density $f_X(x)$, and conditional density $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$. The conditional expectation of Y given $X = x$ is

$$\begin{aligned} E(Y|X = x) &= \int_{-\infty}^{\infty} y \ f_{Y|X}(y|x)dy \\ &= \int_{-\infty}^{\infty} y \ \frac{f_{X,Y}(x,y)}{f_X(x)} dy \end{aligned}$$

When the denominator is zero, the expression is undefined.

Note that conditioning on a continuous random variable is not the same as conditioning on the event $\{X = x\}$ as it was in the discrete case. The probability of the event is zero, but we define the conditional expectation in terms of the density function.

Theorem 40.1 (Law of iterated expectation)**.** *For any random variable X and Y, it holds that*

$$E(E(Y|X)) = E(Y).$$

Proof. Note that $E(Y|X) = g(X)$ is a function of X. Apply LOTUS:

$$\begin{aligned} E(E(Y|X)) &= \int g(x)f(x)dx \\ &= \int \left(\int yf(y|x)dy \right) f(x)dx \\ &= \int \int yf(y|x)f(x)dydx \\ &= \int y \int f(y,x)dx \, dy \\ &= \int_{-\infty}^{\infty} yf(y)dy \\ &= E(Y). \end{aligned}$$

□

Theorem 40.2. *For any random variable X and Y, and any function g, we have*

$$E(g(X)Y|X) = g(X)E(Y|X).$$

Proof. For any specific value of $X = x$, $g(x)$ is a constant. Thus, $E(g(x)Y|X = x) = g(x)E(Y|X = x)$. This is true for all values of x. □

Theorem 40.3 (Best predictor)**.** *Conditional expectation $E(Y|X)$ is the best predictor for Y using X (minimized the square loss function).*

Proof. Let $g(X)$ be a predictor for Y using X. We want to find the g such that minimizes $E(Y - g(X))^2$.

$$\begin{aligned} E(Y - g(X))^2 &= E(Y - E(Y|X) + E(Y|X) - g(X))^2 \\ &= E(Y - E(Y|X))^2 + 2\underbrace{E(Y - E(Y|X)}_{E(Y)=E(E(Y|X))}((E(Y|X) - g(X)) \\ &\quad + E(E(Y|X) - g(X))^2 \\ &= E(Y - E(Y|X))^2 + E(E(Y|X) - g(X))^2 \\ &\geq E(Y - E(Y|X))^2. \end{aligned}$$

Therefore, $E(Y - g(X))^2$ is minimized when $g(X) = E(Y|X)$. □

Definition 40.2 (Linear conditional expectation model)**.** An extremely widely used method for data analysis in statistics is linear regression. In its most basic form, we want to predict the mean of Y using a single explanatory variable X. A linear conditional expectation model assumes that $E(Y|X)$ is linear in X:

$$E(Y|X) = a + bX,$$

or equivalently,

$$Y = a + bX + \epsilon,$$

with $E(\epsilon|X) = 0$. The intercept and the slope is given by

$$b = \frac{Cov(X,Y)}{Var(X)}, a = E(Y) - bE(X).$$

We first show the equivalence of the two expressions of the model. Let $Y = a + bX + \epsilon$, with $E(\epsilon|X) = 0$. Then by linearity,

$$E(Y|X) = E(a|X) + E(bX|X) + E(\epsilon|X) = a + bX.$$

Conversely, suppose that $E(Y|X) = a + bX$, and define

$$\epsilon = Y - (a + bX).$$

Then $Y = a + bX + \epsilon$, with

$$E(\epsilon|X) = E(Y|X) - E(a + bX|X) = E(Y|X) - (a + bX) = 0.$$

To derive the expression for a and b, take covariance between X and Y,

$$\begin{aligned} Cov(X, Y) &= Cov(X, a + bX + \epsilon) \\ &= Cov(X, a) + bCov(X, X) + Cov(X, \epsilon) \\ &= bVar(X) + Cov(X, \epsilon) \end{aligned}$$

Note that $Cov(X, \epsilon) = 0$ because

$$\begin{aligned} Cov(X, \epsilon) &= E(X\epsilon) - E(X)E(\epsilon) \\ &= E(E(X\epsilon|X)) - E(X)E(E(\epsilon|X)) \\ &= E(XE(\epsilon|X)) - E(X)E(E(\epsilon|X)) \\ &= 0 \end{aligned}$$

Therefore,

$$Cov(X, Y) = bVar(X)$$

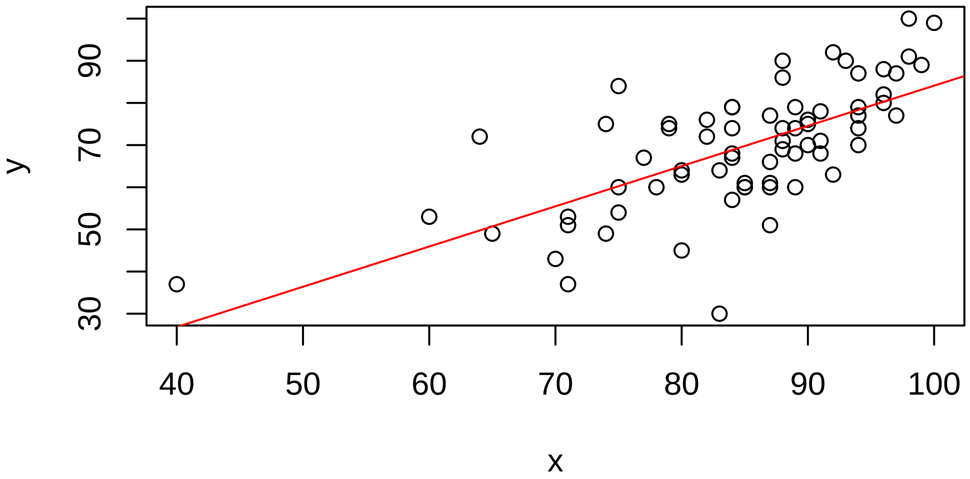
Thus,

$$\begin{aligned} b &= \frac{Cov(X, Y)}{Var(X)}, \\ a &= E(Y) - bE(X) = E(Y) - \frac{Cov(X, Y)}{Var(X)}E(X). \end{aligned}$$

In practice, we don't know the true value of $Cov(X, Y)$ or $Var(X)$. We have to estimate it with sample observations. Thus, we compute $\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$. By definition, b gives the marginal change of $E(Y|X)$ with respect to X.

```r
# load data
exam <- read.csv("../dataset/exam.csv")

# midterm score
x <- exam$mid
```

```r
# final score
y <- exam$final

# regress y on x, compute coefficients
b <- cov(x,y)/var(x)
a <- mean(y) - b*mean(x)

# plot the data and the regression line
plot(x,y)
abline(a,b,col="red")
```



Linear regression is the simple yet powerful modeling tool in statistics. It is useful whenever we want to predict one variable with another. When the assumptions are met (though this is rare), the model gives the best predictor (conditional expectation). If the assumptions are not met, regression gives a linear approximation.

41 Moments and MGF

Definition 41.1 (Moment)**.** Let X be a random variable with mean μ and variance σ^2 . For any positive integer n, the n-th **moment** of X is $E(X^n)$, the n-th **central moment** is $E(X-\mu)^n$, and the n-th **standardized moment** is $E\left(\frac{X-\mu}{\sigma}\right)^n$.

In accordance with this terminology, $E(X)$ is the first moment of X, $Var(X)$ is the second central moment of X. It is natural to ask if there are higher order moments. The answer is yes.

Definition 41.2 (Skewness)**.** Let X be a random variable with mean μ, standard deviation σ, and finite third moment. The skewness of X is defined as

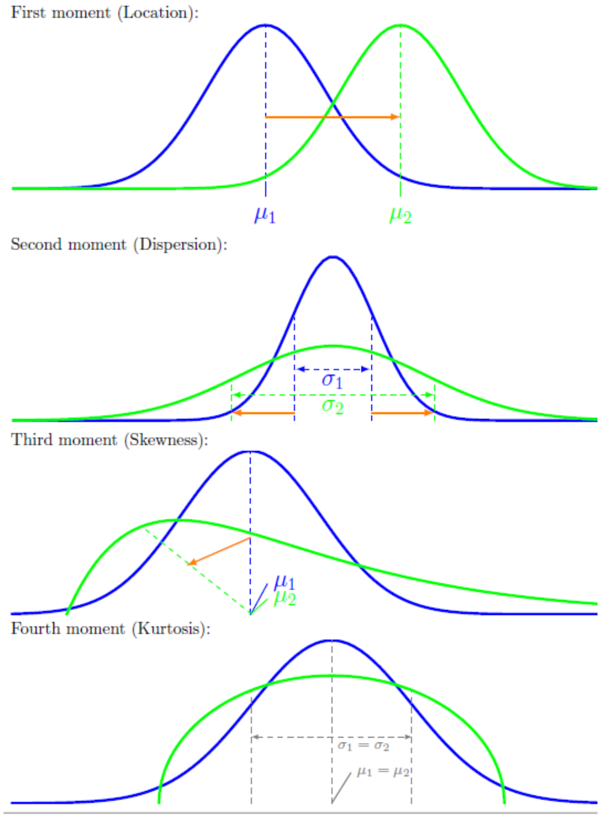
$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right].$$

Definition 41.3 (Kurtosis)**.** The Kurtosis of X is defined as

$$\text{Kurt}(X) = \left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

Skewness is the measure of the lopsidedness of the distribution; any symmetric distribution will have a third central moment, if defined, of zero. A distribution that is skewed to the left (the tail of the distribution is longer on the left) will have a negative skewness. A distribution that is skewed to the right (the tail of the distribution is longer on the right), will have a positive skewness.

Kurtosis is a measure of the heaviness of the tail of the distribution. If a distribution has heavy tails, the kurtosis will be high; conversely, light-tailed distributions have low kurtosis.



We see that moments give information about the shape of a distribution. Different orders of moments captures different aspects of the distribution. As higher and higher moments are calculated, they reveal more and more aspects of the distribution. Loosely speaking, it is somewhat like the Taylor theorem in the probability theory. We can approximate a distribution by "expectation of polynomials": $E(X), E(X^2), E(X^3), \ldots$

Definition 41.4 (Moment generating function)**.** Let X be a random variable. For each real number t, define the moment generating function (MGF) as

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

To see why it is "generating" moments, take the Taylor expansion of the exponential function:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

Hence,

$$M_X(t) = E\left(e^{tX}\right) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + \cdots$$

A natural question at this point is: What is the interpretation of t? The answer is that t has no interpretation in particular; it's just a bookkeeping device that we introduce in order to *encode* the sequence of moments in a differentiable function.

Theorem 41.1. *Let $M_X(t)$ be the MGF of X. Then the n-th moment of X is given by $E(X^n) = M_X^{(n)}(0)$, where $M_X^{(n)}$ denotes the n-th derivative of the MGF.*

Theorem 41.2. *If the MGFs of two random variables X_1 and X_2 are finite and identical for all values of t in an open interval around the point $t = 0$, then the probability distributions of X_1 and X_2 must be identical.*

Theorem 41.3. *If X and Y are independent, then the MGF of $X + Y$ is the product of the individual MGFs:*

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

Example 41.1. For $X \sim \text{Bern}(p)$, e^{tX} takes on the value e^t with probability p and the value 1 with probability q, so $M(t) = E\left(e^{tX}\right) = pe^t + q$. Since this is finite for all values of t, the MGF is defined on the entire real line.

Example 41.2. The MGF of a $\text{Bin}(n, p)$ random variable is $M(t) = (pe^t + q)^n$, since it is the product of n independent Bernoulli MGFs.

42 Inequalities*

This section introduces some of the most popular inequality in statistics and general mathematics. Interestingly, our probability theories can shed light on these inequalities that are otherwise hard to explain. We don't show formal proofs here, but just point out how these inequalities can be useful in statistics.

Theorem 42.1 (Cauchy-Schwarz inequality)**.**

$$\left|\sum x_i y_i\right| \leq \sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$$

Proof. If X, Y have zero means, their correlation can be written as

$$\rho_{XY} = \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}}$$

Since $|\rho_{XY}| \leq 1$, we always have

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

Consider $\{x_i\}$ and $\{y_i\}$ as realizations of X and Y with equal probabilities, such that $E(X) = \frac{1}{n}\sum x_i$. The original inequality is thus proved. □

Theorem 42.2 (Jensen's inequality)**.** *For a convex function f, we have*

$$\frac{1}{n}\sum f(x_i) \geq f\left(\frac{1}{n}\sum x_i\right);$$

If f is concave, then

$$\frac{1}{n}\sum f(x_i) \leq f\left(\frac{1}{n}\sum x_i\right).$$

We do not intend to prove it, but offer a special case in statistics that helps to understand Jensen's inequality. Consider

$$Var(X) = E(X^2) - (E(X))^2 \geq 0$$

We have

$$E(X^2) \geq (E(X))^2.$$

Note that $f(X) = X^2$ is a convex function, and $E(*) = \frac{1}{n}\sum *$, we have shown the first inequality. The concave case is the opposite.

In general, 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.

Theorem 42.3 (Markov inequality). *Let X be a random variable, then*

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

That is, the probability of $|X|$ deviating from its mean by a multiple of a must be less than $1/a$.

Proof. Define a random variable

$$I_{|X|\geq a} = \begin{cases} 1 & \text{if } |X| \geq a \\ 0 & \text{if } |X| < a \end{cases}$$

Note that $P(|X| \geq a) = E(I_{|X|\geq a})$. It always holds that

$$a \cdot I_{|X|\geq a} \leq |X|$$

Therefore,

$$E\left[a \cdot I_{|X|\geq a}\right] \leq E|X|$$

Hence,

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

□

For an intuitive interpretation, let X be the income of a randomly selected individual from a population. Taking $a = 2E(X)$, Markov's inequality says that $P(X \geq 2E(X)) \leq 1/2$, i.e., it is impossible for more than half the population to make at least twice the average income. This is clearly true, since if over half the population were earning at least twice the average income, the average income would be higher. Similarly, $P(X \geq 3E(X)) \leq 1/3$: you can't have more than 1/3 of the population making at least three times the average income, since those people would already drive the average above what it is.

Theorem 42.4 (Chebyshev inequality)**.** *Let X be a random variable with mean μ and standard deviation σ, then*

$$P\left(|X - \mu| > c\sigma\right) \leq \frac{1}{c^2}$$

That is, the probability of X deviating from its mean by a times the standard deviation must be less than $1/a^2$.

Proof. We first show

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

This is true by taking squares and applying the Markov inequality,

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

Substitute $c\sigma$ for a, we have the original inequality. □

This gives us an upper bound on the probability of a random variable 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. Given the mean and standard deviation of a random variable X, we know that $\mu \pm 2\sigma$ captures 75% of its possible values; $\mu \pm 3\sigma$ captures 90% of the possible values.

Part V

Continuous Distributions

43 Definition revisited

Continuous random variables, in many ways, are more versatile and useful than discrete distributions. One key reason is that many quantities in the physical world, such as temperature, height, weight, and time, are inherently continuous in nature. Additionally, the probability density functions (PDFs) of continuous distributions are often defined by smooth, differentiable functions. This mathematical structure allows us to apply calculus for analysis.

Definition 43.1. A random variable has a continuous distribution if its CDF is *differentiable.*

Remark.

- For a continuous random variable, $P(X = x) = 0$ for all x;
- The density function $f(x)$ is not a probability. To get the probability, we integrate the PDF (probability is the area under the PDF):
$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx.$$
- Since any single value has probability 0, including or excluding endpoints does not matter.
$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$
- The PDF of a continuous random variable satisfies the property:
$$\int_{-\infty}^{\infty} f(x)dx = 1.$$
- The CDF is the integral of PDF:
$$F(X) = \int_{-\infty}^{x} f(x)dx, \quad f(x) = F'(x).$$

Definition 43.2. The expectation of a continuous random variable X with PDF f is

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

Theorem 43.1. *If X is a continuous random variable with PDF f and $g : \mathbb{R} \to \mathbb{R}$. The LOTUS applies*

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

	Discrete	Continuous
PMF/PDF	$P(X = x) = p(x)$	$P(a \leq X \leq b) = \int_a^b f(x)dx$
CDF	$F(x) = P(X \leq x) = \sum_{k \leq x} p(k)$	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
Expectation	$E(x) = \sum_x xP(X = x)$	$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$
LOTUS	$E[g(x)] = \sum_x g(x)P(X = x)$	$E[g(x)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$

44 Uniform distribution

Definition 44.1 (Uniform distribution)**.** Let a and b be two given real numbers such that $a < b$. Let X be a random variable such that it is known that $a \leq X \leq b$ and, for every subinterval of $[a, b]$, the probability that X will belong to that subinterval is proportional to the length of that subinterval. We then say that the random variable X has the Uniform distribution on the interval $[a, b]$. The PDF of X is

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

This is a valid PDF since

$$\int_{-\infty}^{+\infty} f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b dx = 1.$$

The CDF of X is

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_a^x f(t)dt = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}.$$

The expectation of X:

$$E(X) = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}.$$

To figure out the variance, first compute

$$E(X^2) = \int_a^b x^2\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{a^2+ab+b^2}{3}$$

Thus,

$$Var(X) = E(X^2) - E^2(X) = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

Example 44.1 (Longer piece of a broken stick)**.** A stick of unit length is broken at a random point X. What is the expected length of the longer piece?

Solution. The lengths of the two pieces are X and $1-X$, with $X \sim U(0,1)$. The longer piece is $\max(X, 1-X)$. For $X < 0.5$, the longer piece is $1-X$, and for $X \geq 0.5$, it is X. The expected value is:

$$E[\max(X, 1-X)] = \int_0^{0.5} (1-X)\,dx + \int_{0.5}^{1} X\,dx = \frac{3}{4}.$$

Intuition might suggest that since the stick is broken at a random point, the longer piece should be "somewhat larger" than the shorter piece, but not as large as 3/4. However, the uniform distribution of the break point means that the longer piece can sometimes be much larger than the shorter piece, especially when the break point is close to one end.

```r
# number of simulations
N <- 1000

# simulate random break point
X <- runif(N, min = 0, max = 1)

# length of the longer piece
L <- pmax(X, 1 - X)

cat("Expected Length of Longer Piece:", mean(L))
```

```
Expected Length of Longer Piece: 0.7544274
```

Example 44.2 (Buffon's needle revisited)**.** A plan is ruled by the lines $y = 0, \pm 1, \pm 2, \dots$ and a needle of unit length is cast randomly on to the plane. What is the probability that it intersects some line?

Solution. Let Z be the distance from the needle's center to the nearest line beneath it. Let Θ be the angle made by the needle and the x-axis. The fact that the needle is cast randomly means $Z \sim U(0,1)$, $\Theta \sim U(0,\pi)$ and Z and Θ are independent. Thus the joint density function of (Z, Θ) is

$$f(z, \theta) = \frac{1}{\pi}, \quad 0 \leq z \leq 1, 0 \leq \theta \leq \pi.$$

An intersection occurs if and only if (draw a diagram to see this):

$$z \leq \frac{1}{2}\sin\theta \text{ or } 1 - z \leq \frac{1}{2}\sin\theta$$

Hence

$$\begin{aligned} P(\text{intersection}) &= \frac{1}{\pi}\int_0^{\pi}\left(\int_0^{\frac{1}{2}\sin\theta} dz + \int_{1-\frac{1}{2}\sin\theta}^{1} dz\right) d\theta \\ &= \frac{2}{\pi}. \end{aligned}$$

45 Special integrals*

There are many reasons to learn integrals. But the most compelling reason is that math is no longer the same with integrals. We can have many amazing results with integrals that were otherwise not imaginable. This section introduces two integrals that are of special importance to continuous distributions.

Proposition 45.1 (Guassian integral)**.**

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

This is known as Gaussian integral, which is the kernel of the PDF of the normal distribution. It also amazingly relates two of the most famous constants in mathematics. It is not integrable by normal integration techniques. But it can be proved by switching to the polar coordinate.

Proof. Let $I = \int_{-\infty}^{+\infty} e^{-x^2} dx$.

$$\begin{aligned}
I^2 =& \int_{-\infty}^{+\infty} e^{-x^2} dx \int_{-\infty}^{+\infty} e^{-y^2} dy \\
=& \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dxdy \\
=& \int_0^{2\pi} \int_0^{\infty} e^{-r^2} rdrd\theta \qquad\qquad dA = dxdy = rdrd\theta \\
=& \int_0^{2\pi} \int_0^{\infty} \frac{1}{2} e^{-u} dud\theta \qquad\qquad \text{let } u = r^2 \\
=& \frac{1}{2} \int_0^{2\pi} d\theta = \pi.
\end{aligned}$$

Thus, $I = \sqrt{\pi}$. □

Proposition 45.2 (Gamma function)**.**

$$\int_0^{\infty} t^n e^{-t} dt = n!$$

$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is known as the Gamma function, which is definitely one of the most interesting functions in mathematics. It is the extension of factorials to real numbers or even complex numbers. It also has many interesting properties, such as $\Gamma(n) = (n-1)!$, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(3/2) = \sqrt{\pi}/2$, $\Gamma'(1) = -\gamma$ and so on. The $(n-1)$ in the Gamma function is due to historical reasons and does not matter in our case. We will prove the integral with n instead of $(n-1)$.

Proof. There are many ways to prove this. One is to discover the recursive relationship $\Gamma(n+1) = n\Gamma(n)$. But it does not give a clue why we need this integral to approximate the factorial. We start with an elementary integral

$$\int_0^\infty e^{at}dt = -\frac{1}{a}$$

where $a < 0$. Differentiate both sides n times with respect to a:

$$\begin{aligned}
\int_0^\infty e^{at}tdt = & \quad -(-1)a^{-2} \\
\int_0^\infty e^{at}t^2dt = & \quad -(-1)(-2)a^{-3} \\
\int_0^\infty e^{at}t^3dt = & \quad -(-1)(-2)(-3)a^{-4} \\
\vdots & \\
\int_0^\infty e^{at}t^ndt = & \quad (-1)^{n+1}n!a^{-(n+1)}
\end{aligned}$$

Let $a = -1$, we have

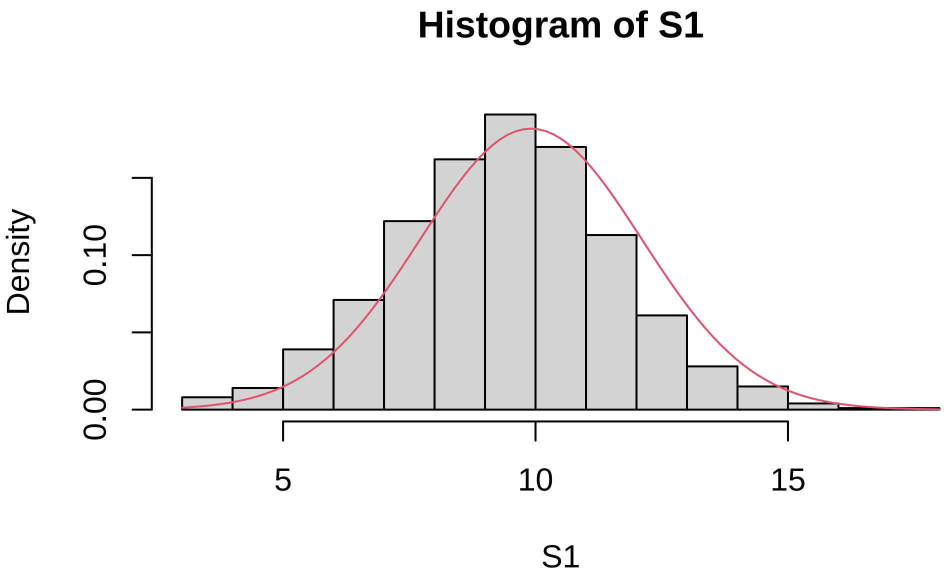
$$\int_0^\infty e^t t^n = n!$$

□

46 Sum of RVs

We know the sum of coin heads and the sum of dice points follow a Binomial distribution.

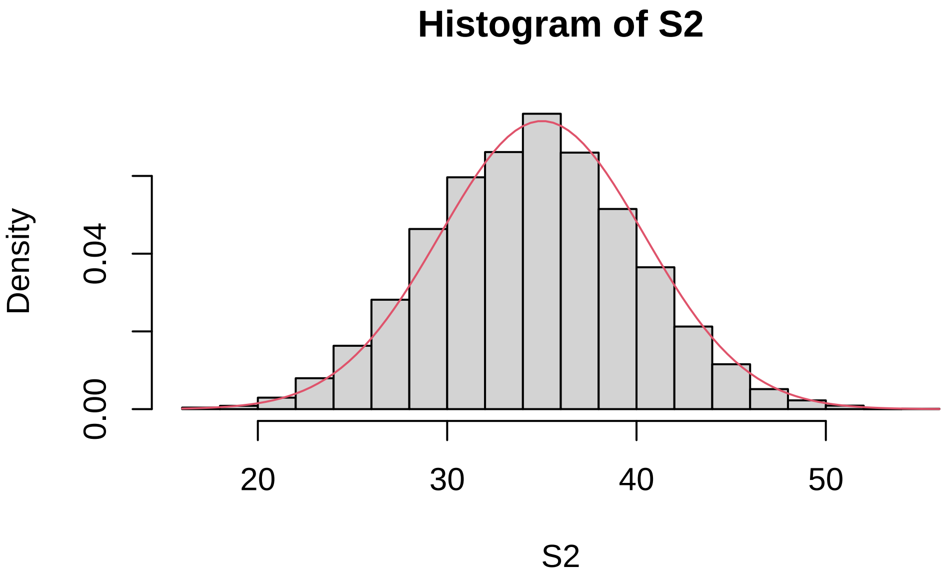
```
set.seed(0)

# Repeat experiment: sum of heads tossing ten coins
S1 <- replicate(1000, { sum(sample(0:1, 20, TRUE)) })

hist(S1, prob=TRUE)

# Bell-shaped curve
curve(dnorm(x, mean(S1), sd(S1)), col=2, add=TRUE)
```



The distribution of the sum of dice has a similar shape:

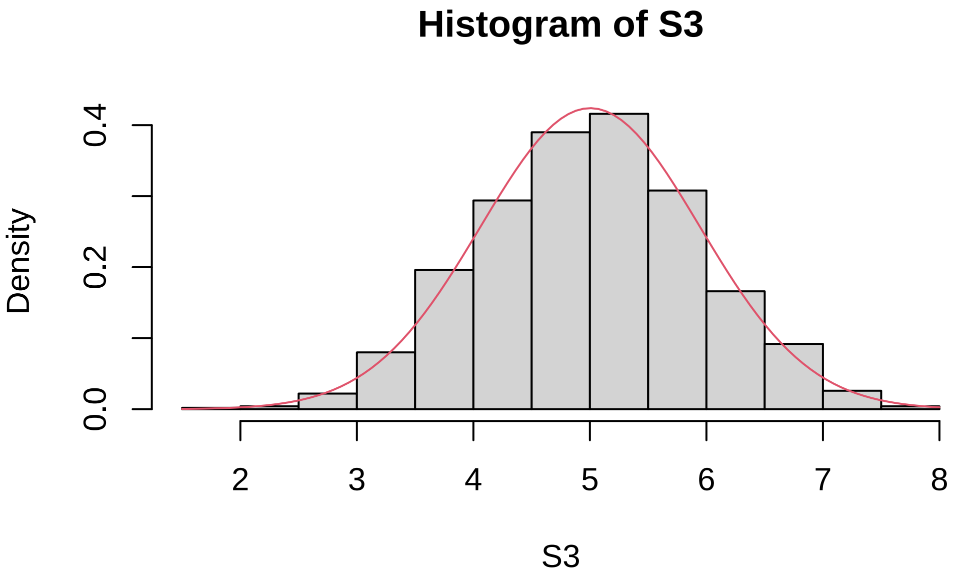
```
set.seed(0)

# Repeat experiment: sum of rolling ten dice
S2 <- replicate(10000, { sum(sample(1:6, 10, TRUE)) })
```

```
hist(S2, prob=TRUE)

# Bell-shaped curve
curve(dnorm(x, mean(S2), sd(S2)), col=2, add=TRUE)
```



This is not a coincidence. In fact, the sum of random variables from any distribution would reveal a similar shape.

```
set.seed(0)

# Repeat experiment: sum of ten uniform variables
S3 <- replicate(1000, { sum(runif(10)) })

hist(S3, prob=TRUE)

# Bell-shaped curve
curve(dnorm(x, mean(S3), sd(S3)), col=2, add=TRUE)
```



The bell-shaped curve comes from Normal distributions. The distribution of the *sum* of random variables always converges to Normal distribution.

ℹ Sum of random variables approaches Normal

Let $X_1, X_2, \dots, X_n$ be a sequence of independent and identical random variables. Let

$$S_n = X_1 + X_2 + \cdots + X_n$$

Then, as $n \to \infty$, S_n converges to Normal in distribution. That is

$$S_n \to^d \text{Normal.}$$

47 Normal distribution

The most widely used model for random variables with continuous distributions is the family of normal distributions. One reason is that many real world samples appears to be normally distributed (the mass centered around the mean). The other reason is because of the Central Limit Theorem (will be discussed in later chapters), which essentially says the sum (or mean) or any random samples are approximately normal.

Definition 47.1 (Standard Normal)**.** A random variable Z has the standard Normal distribution with mean 0 and variance 1, denoted as $Z \sim N(0, 1)$, if Z has a PDF that follows

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

This is a valid PDF because $\int_{-\infty}^{\infty} f(z)dz = 1$, which directly follows from the Gaussian integral. We further verify its mean and variance:

$$E(Z) = \int_{-\infty}^{+\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0 \quad \text{by symmetry.}$$

$$\begin{aligned}
Var(Z) &= E(Z^2) - (EZ)^2 = E(Z^2) \\
&= \int_{-\infty}^{+\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \underbrace{z}_{u} \cdot \underbrace{z e^{-z^2/2} dz}_{dv} \\
&= \frac{2}{\sqrt{2\pi}} \left\{ \left[z(-e^{-z^2/2}) \right]_0^{\infty} + \underbrace{\int_0^{\infty} e^{-z^2/2} dz}_{\sqrt{2\pi}/2} \right\} \\
&= 1.
\end{aligned}$$

Definition 47.2 (The Φ function)**.** The CDF of standard normal distribution is usually denoted by Φ. Therefore,

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

By symmetry, we have $\Phi(-z) = 1 - \Phi(z)$.

To find the value of $\Phi(z)$, we need to use the normal probability table or statistical software.

Definition 47.3 (General Normal)**.** Let $X = \mu + \sigma Z$ where $Z \sim N(0,1)$. Then we say X has the Normal distribution with mean μ and variance σ^2, denoted as $X \sim N(\mu, \sigma^2)$. The PDF of X is given by

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

The mean and variance of X can be easily verified by the properties of expectation and variance.

$$E(X) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu,$$
$$Var(X) = Var(\mu + \sigma Z) = \sigma^2 Var(Z) = \sigma^2.$$

To verify the PDF, we utilize the standard normal CDF:

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

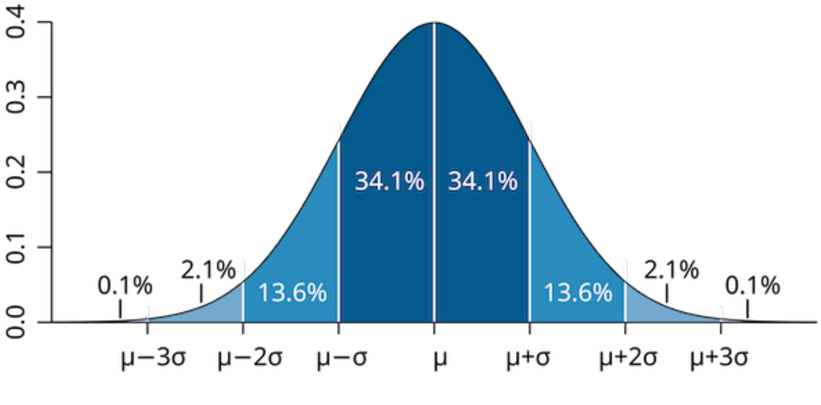
The PDF is the derivative of the CDF,

$$f(x) = \frac{1}{\sigma}\Phi'\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Proposition 47.1 (Three-sigma rule)**.** *The normal distribution has the following properties:*

$$P(|X-\mu| \le \sigma) \approx 0.68$$
$$P(|X-\mu| \le 2\sigma) \approx 0.95$$
$$P(|X-\mu| \le 3\sigma) \approx 0.997$$

Critical values in standard normal: $\Phi(-1) \approx 0.16, \Phi(-2) \approx 0.025, \Phi(-3) \approx 0.0015$.



Theorem 47.1 (Standardization)**.** *Let X have the Normal distribution with mean μ and variance σ^2. Let F be the CDF of X. Then the standardization of X*

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

has the standard normal distribution, and, for all x:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Example 47.1. Suppose the test score of a class of 50 students is normally distributed with mean 80 and standard deviation 20 (the total mark is 100). A student has scored 90. What is his percentile in the class?

Solution. $X \sim N(80, 20)$. We want to find $P(X < 90)$. Standardize the distribution:

$$P(X < 90) = P\left(\frac{X - 80}{20} < \frac{90 - 80}{20}\right) = \Phi(0.5) \approx 0.69.$$

Theorem 47.2 (The MGF of Normal)**.** *The MGF of $X \sim N(\mu, \sigma^2)$ is:*

$$M_X(t) = E[e^{tX}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Proof. Let $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0, 1)$. The MGF of Z is:

$$\begin{aligned} E(e^{tZ}) &= \int_{-\infty}^{\infty} e^{tz} \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2 + tz} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2} dz \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

Therefore,

$$\begin{aligned} E(e^{tX}) &= E[e^{t(\mu + \sigma Z)}] \\ &= e^{t\mu} \cdot E(e^{\sigma t Z}) \\ &= e^{t\mu} M_Z(\sigma t) \\ &= e^{t\mu} \cdot e^{\frac{1}{2}(\sigma t)^2} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \end{aligned}$$

□

Theorem 47.3 (Linear transformation of Normal variables)**.** *Suppose $X \sim N(\mu, \sigma^2)$. If $Y = aX + b$, then Y has the Normal distribution $Y \sim N(a\mu + b, a^2\sigma^2)$.*

Proof. The MGF of X is:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

The MGF of $Y = aX + b$ is:

$$\begin{aligned} M_Y(t) = E(e^{t(aX+b)}) &= e^{tb} \cdot E(e^{atX}) = e^{tb} M_X(at) \\ &= \exp(tb) \cdot \exp\left(\mu at + \frac{1}{2}\sigma^2 (at)^2\right) \\ &= \exp\left((a\mu + b)t + \frac{1}{2}(a^2\sigma^2)t^2\right) \end{aligned}$$

which indicates $Y \sim N(a\mu + b, a^2\sigma^2)$. □

Theorem 47.4 (Sum of Normal variables)**.** *If the random variables $X_1, \dots, X_k$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$. Then*

$$X_1 + \dots + X_k \sim N(\mu_1 + \dots + \mu_k, \sigma_1^2 + \dots + \sigma_k^2).$$

Proof. We prove the case $k = 2$. By independence, the MGFs satisfy:

$$\begin{aligned} M_{X_1+X_2}(t) &= M_{X_1}(t) M_{X_2}(t) \\ &= \exp\left\{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right\} \cdot \exp\left\{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right\} \\ &= \exp\left\{(\mu_1 + \mu_2)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2\right\}. \end{aligned}$$

□

Example 47.2. The average height of men and women in America is given in the table below (Source: CDC's National Health and Nutrition Examination Survey):

Ethnic Group	Men	Women
White	177.4 cm	163.3 cm
Black	175.5 cm	162.5 cm
Hispanic	169.5 cm	157.5 cm
Asian	169.7 cm	156.3 cm

The standard deviation for both men and women is about 7 cm. Find the probability (approximately) that a randomly selected Asian woman be taller than a man.

Solution. Suppose the heights of women and men independently follow the normal distributions, $X \sim N(156.3, 7^2)$, $Y \sim N(169.7, 7^2)$.Then $W = X - Y \sim N(-13.4, 98)$. Therefore we have:

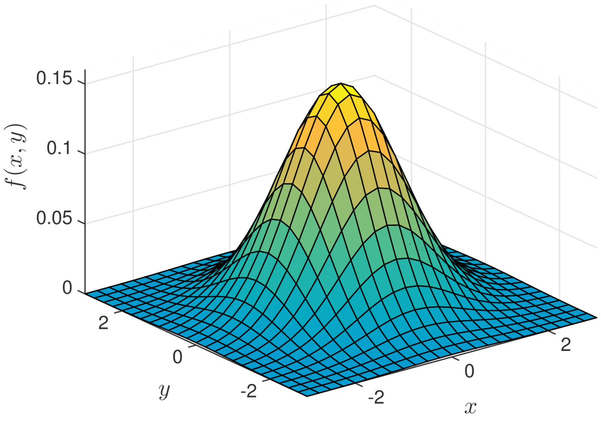
$$\begin{aligned} P(W > 0) &= P\left(\frac{W + 13.4}{\sqrt{98}} > \frac{-13.4}{\sqrt{98}}\right) \\ &= P(Z > -1.35) \\ &= 1 - \Phi(-1.35) \\ &\approx 0.09. \end{aligned}$$

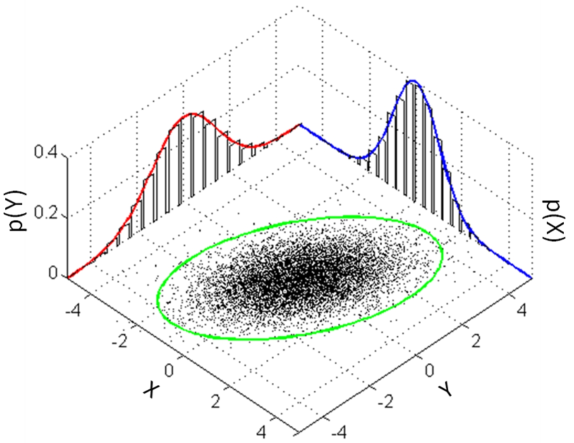
48 Multivariate normal*

Definition 48.1 (Bivariate normal distribution)**.** (X, Y) is said to have a Bivariate Normal distribution if the joint PDF satisfies

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right)$$

where $\rho \in (-1, 1)$ is the correlation between X and Y.



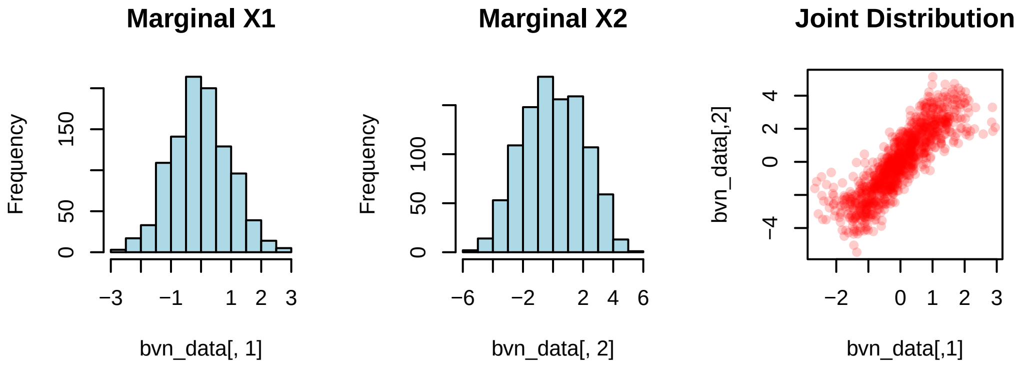


Multivariate Normal (MVN) is an extension of the bivariate normal distribution to n-dimensional variables. We skip the joint PDF here since it is too complicated. But like the bivariate case, an MVN is fully specified by knowing the mean of each component, the variance of each component, and the covariance between any two components.

> Marginal normality does not imply joint normality
>
> If $(X_1, ..., X_k)$ is MVN, then the marginal distribution of every X_j is Normal. However, the converse is false: it is possible to have Normally distributed $X_1, ..., X_k$ such that $(X_1, ..., X_k)$ is not Multivariate Normal.

```r
# Load necessary library
library(MASS)
```

```
# Set seed for reproducibility
set.seed(123)

# Generate bivariate normal data
bvn_data <- mvrnorm(n = 1000,
                    mu = c(0, 0),
                    Sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))

# Modify the joint distribution: apply a nonlinear transformation
bvn_data[, 2] <- bvn_data[, 2] + 2 * sin(bvn_data[, 1])

# The marginal distribution remains normal
par(mfrow = c(1, 3))
hist(bvn_data[, 1], main = "Marginal X1", col = "lightblue")
hist(bvn_data[, 2], main = "Marginal X2", col = "lightblue")

# But the joint distribution is not normal
plot(bvn_data, main = "Joint Distribution", pch = 16, col = rgb(1,0,0,.2))
```



Theorem 48.1. *A random vector $(X_1, ..., X_k)$ is Multivariate Normal if every linear combination of the X_j has a Normal distribution (X_j do not have to be independent). That is, we require $t_1X_1 + \cdots + t_kX_k$ to have a Normal distribution for any choice of constants $t_1, ..., t_k$.*

Theorem 48.2. *In general, uncorrelated does not imply independent. But with an MVN random vector, uncorrelated implies independent. In particular, if (X, Y) is Bivariate Normal and $\rho_{XY} = 0$, then X and Y are independent.*

Theorem 48.3. *If (X, Y) is Bivariate Normal, then the conditional expectation satisfies*

$$E(Y|X) = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X)).$$

In other words,

$$E(Y|X) = a + bX$$

where $b = \frac{Cov(X,Y)}{Var(X)}$ *and* $a = E(Y) - bE(X)$*.*

This is exactly the case in Definition 40.2, where we assume the conditional expectation $E(Y|X)$ is a linear function of X. This assumption is true when (X, Y) are jointly normal. Otherwise, the assumption might not be reasonable. In practice, we don't know precisely the joint distribution of variables. The linear model is just a simplified assumption.

49 χ^2 and t-distribution

This section introduces distributions that are closely related to normal distributions. More specifically, they are distributions of functions of normal random variables.

Definition 49.1 (χ^2 distribution)**.** Let $X_1, \dots, X_m$ be independent random variables from $N(0, 1)$. Then

$$X_1^2 + \dots + X_m^2 \sim \chi^2(n),$$

which is known as the χ^2 distribution with n degrees of freedom.

Theorem 49.1 (Sample variance distribution)**.** *Suppose that $X_1, \dots, X_n$ form a random sample from $N(\mu, \sigma^2)$. Then*

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n).$$

If μ is replaced by the sample mean $\bar{X}$, one degree of freedom is lost:

$$\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Definition 49.2 (Student-t distribution)**.** Let Y and Z be two independent random variables, for which $Y \sim \chi^2(n)$ and $Z \sim N(0, 1)$. Then

$$\frac{Z}{\sqrt{Y/n}} \sim t(n),$$

which is known as the t-distribution with n degrees of freedom.

Theorem 49.2 (Sample mean distribution)**.** *Suppose that $X_1, \dots, X_n$ form a random sample from $N(\mu, \sigma^2)$. Let $\bar{X}$ denote the sample mean, and define the sample variance*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Then we have

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1).$$

ℹ Sampling distribution of normal variables

Statistics such as $\bar{X}$ and s^2 are functions of the random variables and are thus random variables themselves. The distribution of a statistic is called **sampling distribution**, as the distribution arises due to the sampling process.
Theorem 49.2 and Theorem 49.1 give the sampling distributions associated with $\bar{X}$ and s^2 respectively. They allow us to gauge how accurate the statistics are as measures of the true parameters.
For example, the t-distribution gives the probability:

$$P\left(t_1 \le \frac{\bar{X}_n - \mu}{s/\sqrt{n}} \le t_2\right) = P\left(\bar{X} - t_2\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_1\frac{s}{\sqrt{n}}\right)$$

So we know how close $\bar{X}$ is to the true mean μ.
Similarly, since $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$, we know:

$$P\left(\chi_1^2 \le \frac{(n-1)s^2}{\sigma^2} \le \chi_2^2\right) = P\left(\frac{(n-1)s^2}{\chi_2^2} \le \sigma^2 \le \frac{(n-1)s^2}{\chi_1^2}\right)$$

which gives the probable range of the true σ^2.
It is crucial to stress that the above inferences are only valid when the sample is drawn from a normal distribution.

50 Exponential distribution

Imagine you are a shop owner that waits for your next customer. The customers arrive randomly. What interests us is the waiting time until the next customer arrives.

Let X represent the waiting time. Since the customers arrives randomly, the likelihood of it coming in the next moment is the same whether you've been waiting for one minute or ten minutes. In other words, the probability of waiting another t minutes is the same no matter how long you've already waited. Therefore,

$$P(X \geq s+t \mid X \geq s) = P(X \geq t), \quad \forall s, t \geq 0.$$

The conditional probability can be rewritten using the definition of conditional probabilities:

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

Thus, the memoryless property implies:

$$\frac{P(X \geq s+t)}{P(X \geq s)} = P(X \geq t).$$

Let the survival function $S(x)$ represent $P(X \geq x)$. Substituting $S(x)$ into the equation gives:

$$\frac{S(s+t)}{S(s)} = S(t).$$

This reminds us of the exponential function. In fact, the only continuous and non-negative solution to this equation is:

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

where λ is a positive constant. This solution represents the probability that the waiting time exceeds x , and λ determines how quickly the probability decreases over time.

The CDF of X is exactly the opposite of $S(x)$:

$$F(x) = 1 - S(x) = 1 - e^{-\lambda x}.$$

Take derivative to get the PDF:

$$f(x) = F'(x) = \lambda e^{-\lambda x}.$$

Definition 50.1 (Exponential distribution)**.** A random variable X is said to have the Exponential distribution with parameter λ if its PDF is

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

We denote this as $X \sim \text{Exp}(\lambda).\lambda$ is interpreted as the "rate", i.e. number of events per unit of time.

To compute the expectation and variance, we first standardize the exponential distribution. Let $Y = \lambda X$, then $Y \sim \text{Exp}(1)$, because

$$P(Y \leq y) = P(X \leq y/\lambda) = 1 - e^{-y}.$$

It follows that,

$$E(Y) = \int_0^\infty ye^{-y}dy = [-ye^{-y}]_0^\infty + \int_0^\infty e^{-y}dy = 1;$$

$$Var(Y) = E(Y^2) - (EY)^2 = \int_0^\infty y^2e^{-y}dy - 1 = 1.$$

For $X = Y/\lambda$, we have $E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$.

Theorem 50.1 (Memoryless property)**.** *If X has the exponential distribution with parameter λ, and let $t > 0$, $h > 0$, then*

$$P(X \geq t + h | X \geq t) = P(X \geq h).$$

Proof. For $t > 0$ we have

$$P(X \geq t) = \int_t^\infty \lambda e^{-\lambda x}dx = e^{-\lambda t}.$$

Hence for each $t > 0$ and each $h > 0$,

$$P(X \geq t + h | X \geq t) = \frac{P(X \geq t + h)}{P(X \geq t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = P(X \geq h).$$

□

The memoryless property is a very special property of the Exponential distribution. In fact, the Exponential is the only memoryless continuous distribution (with support $(0, \infty)$); and Geometric distribution is the only memoryless discrete distribution (with support $0, 1, ...$).

Theorem 50.2 (Poisson-Exponential connection)**.** *Let T be the time between two consecutive events in Poisson process Pois(λt). Then T follows Exponential distribution $T \sim$ Exp(λ).*

Proof. The waiting time $T > t$ is equivalent to no event occurred during period t. Therefore,

$$P(T > t) = P(N_t = 0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}$$

where N_t is the number of events occurred in $[0, t]$, which follows a Poisson distribution. The CDF of T is

$$F(t) = 1 - P(T > t) = 1 - e^{-\lambda t}$$

The PDF of T is

$$f(t) = F'(t) = \lambda e^{-\lambda t}$$

This indicates $T \sim \text{Exp}(\lambda)$. □

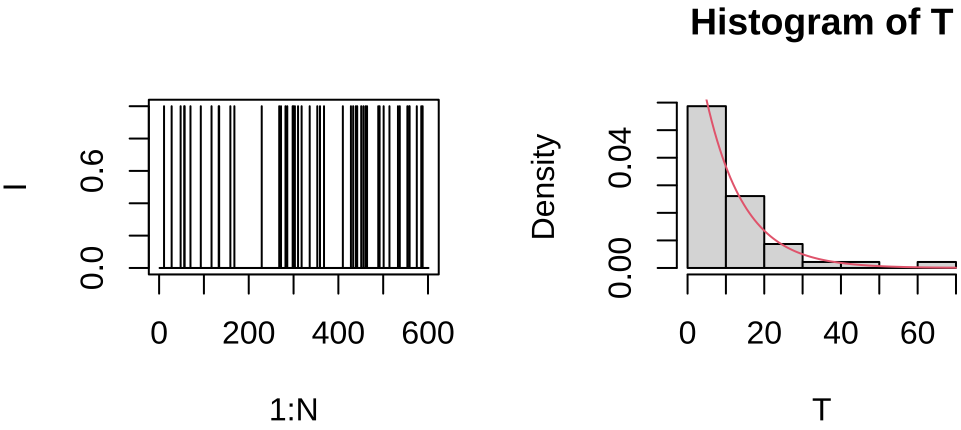
Example 50.1 (Bus arrivals)**.** We try to model the waiting time at a bus station. Suppose the bus arrives at random time but on average there will be one bus per 10 minutes. You arrive at the bus stop at a random time, not knowing how long ago the previous bus came. What is the distribution of your waiting time for the next bus? What is the mean waiting time? What is the median waiting time?

Solution. The bus arrivals in a period of time is best modeled by a Poisson distribution. Let X be the waiting time and we know it is an Exponential distribution. Since $E(X) = 1/\lambda = 10$, $X \sim \text{Exp}(1/10)$. Thus, The average waiting time is always 10 minutes.

The CDF of X is $F(x) = 1 - e^{-\lambda x}$. The median m satisfies $F(m) = 1/2$. Thus, $m = \log(2)/\lambda \approx$ 6.9 minutes. So the typical waiting experienced by most passengers is less than 10 minutes.

```r
# Simulate random arrivals and inter-arrival time

N <- 600  # total simulation time
p <- .1   # prob of occurrence per unit time

set.seed(0)

# I[t] = 1 if an event occurs in time t
I <- 1* (runif(N) < p)

par(mfrow=c(1,2))

# plot the random occurrence
plot(1:N, I, type = "h")

# inter-arrival time
```

```r
T <- diff( (1:N)[I==1] )

# distribution if waiting time
hist(T, prob=TRUE)

# Overlay the exponential function
curve(exp(-x/10)/10, col = 2, add = TRUE)
```



51 Gamma distribution

The Gamma distribution is a continuous distribution on the positive real line; it is a generalization of the Exponential distribution. While an Exponential RV represents the waiting time for the first event to occur, we shall see that a Gamma RV represents the total waiting time for n events to occur.

Let's start with a simple case. Suppose we want to find out the total waiting until the 2nd event occurred. Let $Y = X_1 + X_2$ where $X_1, X_2 \sim \text{Exp}(\lambda)$ independently. If Y is discrete, we have $P(Y = y) = \sum_{k=0}^{y} P(X_1 = k, X_2 = y - k)$. For continuous y, we have

$$\begin{aligned} f_Y(y) &= \int_0^y f_X(x) f_X(y-x) dx = \int_0^y \lambda e^{-\lambda x} \lambda e^{-\lambda(y-x)} dx \\ &= \int_0^y \lambda^2 e^{-\lambda y} dx = \lambda^2 e^{-\lambda y} y. \end{aligned}$$

If there is a third variable,

$$\begin{aligned} f_Z(z) &= \int_0^z f_X(x) f_Y(z-x) dx = \int_0^z \lambda e^{-\lambda x} \lambda^2 e^{-\lambda(z-x)} (z-x) dx \\ &= \lambda^3 e^{-\lambda z} \int_0^z (z-x) dx = \lambda^3 e^{-\lambda z} z^2/2. \end{aligned}$$

The general pattern is the Gamma distribution.

Definition 51.1 (Exponential distribution)**.** An random variable X is said to have the Gamma distribution with parameters a and λ, $a > 0$ and $\lambda > 0$, if it has the PDF

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

We write $X \sim \text{Gamma}(a, \lambda)$.

Verify this is a valid PDF:

$$\int_0^\infty \frac{1}{\Gamma(a)} (\lambda x)^a e^{-\lambda x} \frac{dx}{x} \overset{u=\lambda x}{=} \frac{1}{\Gamma(a)} \int_0^\infty u^a e^{-u} \frac{du}{u} = \frac{\Gamma(a)}{\Gamma(a)} = 1.$$

Taking $a = 1$, the Gamma$(1, \lambda)$ PDF is $f(x) = \lambda e^{-\lambda x}$, which is the same as Exp(λ). So Exponential distribution is a special case of Gamma distribution.

Let's find the expectation and variance of the Gamma distribution. Let $Y \sim$ Gamma$(a, 1)$. Recall Γ function has the property $\Gamma(a+1) = a\Gamma(a)$.

$$E(Y) = \int_0^\infty y \cdot \frac{1}{\Gamma(a)} y^{a-1} e^{-y} dy = \frac{1}{\Gamma(a)} \int_0^\infty y^a e^{-y} dy = \frac{\Gamma(a+1)}{\Gamma(a)} = a.$$

Apply LOTUS to evaluate the second moment:

$$E(Y^2) = \int_0^\infty y^2 \cdot \frac{1}{\Gamma(a)} y^{a-1} e^{-y} dy = \frac{1}{\Gamma(a)} \int_0^\infty y^{a+1} e^{-y} dy = \frac{\Gamma(a+2)}{\Gamma(a)} = (a+1)a.$$

Therefore,

$$Var(Y) = (a+1)a - a^2 = a.$$

So for $Y \sim$ Gamma$(a, 1)$, $E(Y) = Var(Y) = a$. For the general case $X \sim$ Gamma(a, λ), we now show that $X = \frac{Y}{\lambda}$. Note that

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(Y \leq x/\lambda) = F_Y(x/\lambda) \\ f_X(x) &= \frac{dF_X}{dx} = \frac{\partial F_Y}{\partial y} \frac{dy}{dx} = f_Y(y)\lambda \end{aligned}$$

Therefore,

$$f_X(x) = \frac{1}{\Gamma(a)} y^{a-1} e^{-y} \lambda = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}.$$

Hence, we have $E(X) = \frac{a}{\lambda}$, $Var(X) = \frac{a}{\lambda^2}$.

Theorem 51.1 (Exponential-Gamma connection)**.** *Let $X_1, \ldots, X_n$ be independent and identical Exp(λ). Then*

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

Proof. Let's prove by showing the MGFs are equivalent.

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda$$

Thus, the MGF of $Y = X_1 + \cdots + X_n$ is $M_Y(t) = (M_X(t))^n = \left(\frac{\lambda}{\lambda - t}\right)^n$. We verify this is the MGF of a Gamma distribution. Suppose $Y \sim$ Gamma(n, λ), it has MGF:

$$
\begin{aligned}
M_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty}\frac{\lambda^n}{\Gamma(a)}y^{n-1}e^{-\lambda y}dy \\
&= \frac{\lambda^n}{(\lambda-t)^n}\int_0^\infty \frac{1}{\Gamma(a)}((\lambda-t)y)^{n-1}e^{-(\lambda-t)y}(\lambda-t)dy \\
&= \frac{\lambda^n}{(\lambda-t)^n}\int_0^\infty \frac{1}{\Gamma(a)}u^{n-1}e^{-u}du \qquad u=(\lambda-t)y \\
&= \left(\frac{\lambda}{\lambda-t}\right)^n .
\end{aligned}
$$

□

Thus, if X_i represents the i.i.d inter-arrival time. Y has the interpretation of the arrival time until the n-th event.

$$
Y = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (\text{time of the i-th arrival}) \sim \text{Gamma}(n, \lambda).
$$

Example 51.1 (Service time in a queue)**.** Customer i must wait time X_i for service once reaching the head of the queue. The average service rate is 1 customer per 10 minutes. Assume the service for each customer is independent. If you are the 5th in the queue. What is the expected waiting to be served?

Solution. $X_i \sim \text{Exp}(0.1)$. Then $E(X_i) = 10$. Let Y be the time until you are served. Then $Y \sim \text{Gamma}(5, 0.1)$. Thus, $E(Y) = \frac{5}{0.1} = 50$ minutes. The probabilities of some selected values:

$$
P(Y \le t) = \begin{cases} 5\% & t = 20 \\ 18\% & t = 30 \\ 71\% & t = 60 \end{cases} .
$$

52 Beta distribution*

The Beta distribution is a continuous distribution on the interval $(0, 1)$. It is a generalization of the Unif$(0, 1)$ distribution, allowing the PDF to be non-constant on $(0, 1)$.

Definition 52.1 (Beta distribution)**.** A random variable X is said to have the Beta distribution with parameters a and b, $a > 0$ and $b > 0$, if its PDF is

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

where the constant $\beta(a, b)$ is chosen to make the PDF integrate to 1. We write this as $X \sim$ Beta(a, b).

The Beta distribution takes different shapes for different a and b values. Here are some general patterns:

- If $a = b = 1$, the Beta$(1, 1)$ PDF is constant on $(0, 1)$, equivalent to Unif$(0, 1)$.
- If $a < 1$ and $b < 1$, the PDF is U-shaped and opens upward. If $a > 1$ and $b > 1$, the PDF opens downward.
- 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 $1/2$.

To make the PDF integrates to 1, the constant $\beta(a, b)$ has to satisfy

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx.$$

We now try to find this integral:

$$
\begin{aligned}
\beta(a,b) &= \int_0^1 \underbrace{x^{a-1}}_{f}\,\underbrace{(1-x)^{b-1}}_{g'}\,dx \\
&= \left[-x^{a-1}\frac{(1-x)^b}{b}\right]_0^1 + \int_0^1 (a-1)x^{a-2}\frac{(1-x)^b}{b}dx \\
&= \frac{a-1}{b}\beta(a-1,b+1) \\
&= \frac{a-1}{b}\cdot\frac{a-2}{b+1}\beta(a-2,b+2) \\
&= \frac{a-1}{b}\cdot\frac{a-2}{b+1}\cdot\frac{a-3}{b+2}\beta(a-3,b+3) \\
&\vdots \\
&= \frac{(a-1)!}{b(b+1)(b+2)\cdots(b+a-2)}\underbrace{\beta(1,a+b-1)}_{\frac{1}{a+b-1}} \\
&= \frac{(a-1)!}{\frac{(b+a-2)!}{(b-1)!}}\cdot\frac{1}{a+b-1} \\
&= \frac{(a-1)!(b-1)!}{(a+b-1)!} \\
&= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.
\end{aligned}
$$

Example 52.1. Let $X_1, \dots, X_n$ be independent random variables with the uniform distribution on the interval $[0,1]$. Find the distribution of $Y = \max(X_1, \dots, X_n)$.

Solution. Let's find the CDF of Y:

$$
\begin{aligned}
P(Y \le y) &= P(X_1 \le y \cap X_2 \le y \cap \cdots \cap X_n \le y) \\
&\overset{iid}{=} P(X_1 \le y)P(X_2 \le y)\cdots P(X_n \le y) \\
&= y^n
\end{aligned}
$$

for $y \in [0,1]$. Hence,

$$
F_Y(y) = P(Y \le y) = \begin{cases} 0 & y < 0 \\ y^n & 0 \le y \le 1 \\ 1 & y > 1 \end{cases}
$$

The PDF of Y is

$$
f_Y(y) = F_Y'(y) = \begin{cases} ny^{n-1} & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}
$$

Thus, $Y \sim Beta(n,1)$.

Beta distributions are often used as *priors* for parameters in Bayesian inference. We do not cover Bayesian inference in this book. Nonetheless we illustrate this with an example.

Example 52.2 (Beta-Binomial conjugacy)**.** We have a coin that lands Heads with probability p, but we don't know what p is. Our goal is to infer the value of p after observing the outcomes of n tosses of the coin. The larger that n is, the more accurately we should be able to estimate p.

Solution. We model the unknown parameter p as a Beta distribution, $p \sim \text{Beta}(a, b)$. Since we are completely ignorant about this p, we can also model it as the uniform distribution. But we will see that using the Beta distribution is even simpler than the uniform distribution. Let X be the number of heads in n tosses of the coin. Then

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

Apply the Bayes' rule to inverse the conditioning:

$$\begin{aligned} f(p|X = k) &= \frac{P(X = k|p)f(p)}{P(X = k)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k} \cdot \frac{1}{\beta(a,b)}p^{a-1}(1-p)^{b-1}}{\int_0^1 \binom{n}{k}p^k(1-p)^{n-k}f(p)dp} \\ &\propto p^{a+k-1}(1-p)^{b+n-k-1} \end{aligned}$$

This the kernel of $\text{Beta}(a + k, b + n - k)$. The rest is just a normalizing constant. Therefore,

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

The *posterior* distribution of p after observing $X = k$ is still a Beta distribution! This is a special relationship between the Beta and Binomial distributions called *conjugacy*: 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 don't leave the family of Beta distributions. We say that the Beta is the conjugate prior of the Binomial.

53 Transformation

Example 53.1 (Min/max of random variables)**.** Let $X_1, X_2, \dots, X_n$ be i.i.d random variables, each following a uniform distribution on the interval $[0, 1]$. Find the distribution of $\max(X_1, X_2, \dots, X_n)$.

Solution. Let $M = \max(X_1, X_2, \dots, X_n)$. The CDF of M, denoted $F_M(m)$, is the probability that $M \leq m$. For $M \leq m$ to hold, all X_i must satisfy $X_i \leq m$. Since the X_i are independent and identically distributed:

$$\begin{aligned} F_M(m) = P(M \leq m) &= P(X_1 \leq m, X_2 \leq m, \dots, X_n \leq m) \\ &= P(X_1 \leq m) \cdot P(X_2 \leq m) \cdots P(X_n \leq m) \end{aligned}$$

For a uniform distribution on $[0, 1]$, $P(X_i \leq m) = m$ for $0 \leq m \leq 1$. Thus:

$$F_M(m) = m^n.$$

The PDF of M, denoted $f_M(m)$, is the derivative of the CDF:

$$f_M(m) = \frac{d}{dm} F_M(m) = \frac{d}{dm}(m^n) = nm^{n-1}.$$

Example 53.2 (Chi-square PDF)**.** Let $X \sim N(0, 1)$, $Y = X^2$. The distribution of Y is an example of a Chi-Square distribution. Find the PDF of Y.

Solution. Again, we try to find the CDF first, and differentiate to the PDF.

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

Therefore,

$$f_Y(y) = 2\varphi(\sqrt{y}) \cdot \frac{1}{2} y^{-1/2} = \varphi(\sqrt{y}) y^{-1/2}, \quad y > 0.$$

Theorem 53.1 (Transformation)**.** *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 $x = g^{-1}(y)$.

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

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)) = F_X(x)$$

By the chain rule, the PDF of Y is

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

If g is strictly decreasing,

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)) = 1 - F_X(x)$$

Then the PDF of Y is

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

But in this case, $dx/dy < 0$. So taking absolute value covers both cases. □

Example 53.3 (Log-Normal PDF)**.** Let $X \sim N(0, 1)$, $Y = e^X$. Then the distribution of Y is called the Log-Normal distribution. Find the PDF of Y.

Solution. Since $g(x) = e^x$ is strictly increasing. Let $y = e^x$, so $x = \log y$ and $dy/dx = e^x$. Then

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \varphi(x)\frac{1}{e^x} = \varphi(\log y)\frac{1}{y}, \quad y > 0.$$

Note that after applying the change of variables formula, we write everything on the right-hand side in terms of y, and we specify the support of the distribution. To determine the support, we just observe that as x ranges from $-\infty$ to∞, e^x ranges from 0 to ∞.

Theorem 53.2 (Transformation of multi-variables*)**.** *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a continuous random vector with joint PDF $f_{\mathbf{X}}$, and let $\mathbf{Y} = g(\mathbf{X})$ where g is an invertible function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Let $\mathbf{y} = g(\mathbf{x})$. Define the Jacobian matrix:*

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}.$$

Also assume that the determinant of the Jacobian matrix is never 0. Then the joint PDF of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|,$$

where $\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|$ is the absolute value of the determinant of the Jacobian matrix.

Example 53.4. Suppose $X, Y \overset{iid}{\sim} \text{Exp}(1)$. Find the distribution of $X/(X + Y)$.

Solution. Let $U = \frac{X}{X+Y}$, $V = X + Y$. Then $X = UV$, $Y = V - UV$. The determinant of the Jacobian matrix is

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

Thus, the joint distribution of (U, V) is

$$f_{UV}(u,v) = f_{XY}(x,y)|v| = f_X(x)f_Y(y)v = e^{-(x+y)}v = e^{-v}v.$$

The distribution of $X/(X+Y)$ is equivalent to the marginal distribution of U:

$$f_U(u) = \int_0^{\infty} e^{-v}vdv = 1$$

for $0 \leq u \leq 1$. Hence U is a Uniform distribution over $[0, 1]$.

54 Additional problems

We extend the concepts of joint, marginal and conditional distribution to continuous random variables.

	Discrete	Continuous
Joint CDF	$F_{XY}(x,y) = P(X \leq x, Y \leq y)$	$F_{XY}(x,y) = P(X \leq x, Y \leq y)$
Joint PMF/PDF	$p_{XY}(x,y) = P(X = x, Y = y)$	$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$
	$\sum_x \sum_y P(X = x, Y = y) = 1$	$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x,y)dxdy = 1$
Marginal PMF/PDF	$P(X = x) = \sum_y P(X = x, Y = y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy$
Conditional PMF/PDF	$P(X = x \vert Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)}$	$f_{X\vert Y}(x\vert y) = \frac{f_{XY}(x,y)}{f_Y(y)}$
Independence	$P(X = x, Y = y) = P(X = x)P(Y = y)$	$f_{XY}(x,y) = f_X(x)f_Y(y)$
	$P(X = x \vert Y = y) = P(X = x)$	$f_{X\vert Y}(x\vert y) = f_X(x)$
	$F_{XY}(x,y) = F_X(x)F_Y(y)$	$F_{XY}(x,y) = F_X(x)F_Y(y)$
Bayes' rule	$P(Y = y \vert X = x) = \frac{P(X=x\vert Y=y)P(Y=y)}{P(X=x)}$	$f_{Y\vert X}(y\vert x) = \frac{f_{X\vert Y}(x\vert y)f_Y(y)}{f_X(x)}$
LOTP	$P(X = x) = \sum_y P(X = x \vert Y = y)P(Y = y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{X\vert Y}(x\vert y)f_Y(y)dy$
LOTUS	$E(g(X,Y)) = \sum_x \sum_y g(x,y)P(X = x)P(y = y)$	$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f_{XY}(x,y)dxdy$

Example 54.1. Suppose X and Y are uniformly distributed on a disk $\{(x,y) : x^2 + y^2 \leq 1\}$. Find the joint PDF, marginal distributions and conditional distributions. Are X and Y independent?

Solution. The area of the disk is π, therefore

$$f(x,y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal distributions are

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1+x^2}} \frac{1}{\pi} dy = \frac{2}{\pi}\sqrt{1-x^2}, \qquad -1 \leq x \leq 1$$

$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1+y^2}} \frac{1}{\pi} dx = \frac{2}{\pi}\sqrt{1-y^2}, \qquad -1 \leq y \leq 1$$

The conditional distributions are

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-x^2}} = \frac{1}{2\sqrt{1-x^2}}$$

Therefore, $Y|X \sim \text{Unif}(-\sqrt{1-x^2}, \sqrt{1-x^2})$.

Since $f(x,y) \neq f_X(x)f_Y(y)$, X and Y are not independent. This is because knowing the value of X constrains the value of Y.

Example 54.2. Suppose $X, Y \overset{iid}{\sim} U(0,1)$. Find the density function of $X + Y$.

Solution. Let $S = X + Y$, $0 \leq S \leq 1$. Apply convolution:

$$f_S(s) = \int f_X(x) f_Y(s-x) dx$$

For X, Y only take values between 0 and 1. So we require $0 \leq x \leq 1$ and $0 \leq s - x \leq 1$ $(s - 1 \leq x \leq s)$. If $0 \leq s \leq 1$, the valid range for x is $[0, s]$:

$$f_S(s) = \int_0^s 1\ dx = s.$$

If $1 \leq s \leq 2$, the valid range for x is $[s-1, 1]$:

$$f_S(s) = \int_{s-1}^1 1\ dx = 2 - s.$$

Thus,

$$f_S(s) = \begin{cases} s, & \text{if } 0 \leq s \leq 1 \\ 2 - s, & \text{if } 1 < s \leq 2. \end{cases}$$

The problem can also be approached geometrically. $P(X + Y \leq s)$ can be computed by the area surrounded by $y = s - x$ and $0 \leq x, y \leq 1$. Draw a diagram to see it.

Example 54.3. Suppose $X, Y \overset{iid}{\sim} U(0,1)$. Find the density function of XY.

Solution. We cannot use convolution in this case, since it is not the sum of random variables. Let $T = XY$, $0 \leq T \leq 1$. We first compute the CDF:

$$P(XY \leq t) = \int_0^1 P(XY \leq t|X = x)f_X(x)dx = \int_0^1 P(Y \leq t/x)dx$$

The probability depends on the values of t and x. If $x > t$, it is proportional to the value t/x. If $x \leq t$, it is always 1. Thus,

$$P(XY \leq t) = \int_0^t 1 \ dx + \int_t^1 \frac{t}{x}dx = t(1 - \ln t).$$

Differentiate to get the density function:

$$f_T(t) = \begin{cases} -\ln t, & \text{if } 0 \leq t \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Draw a diagram to get the same conclusion. The probability can be found by the area surrounded by $y = \frac{t}{x}$ and $0 \leq x, y \leq 1$.

Example 54.4. For $X, Y \overset{iid}{\sim} U(0, 1)$, find $E(|X - Y|)$.

Solution. Apply 2D LOTUS:

$$\begin{aligned} E(|X - Y|) =& \int_0^1 \int_0^1 |x - y|dxdy \\ =& \int_0^1 \int_y^1 (x - y)dxdy + \int_0^1 \int_0^y (y - x)dxdy \\ =& 2\int_0^1 \int_y^1 (x - y)dxdy \\ =& \frac{1}{3}. \end{aligned}$$

Part VI

Estimation and Sampling Distributions

55 Law of large numbers

Definition 55.1 (Independent and identical RVs)**.** Random variables $X_1, X_2, \ldots, X_n$ are independently and identically distributed (or **i.i.d** for short) if they are independently drawn from the same distribution (with the same parameters).

Definition 55.2 (Random sample)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables from distribution F. We call the collection $\{X_1, X_2, \ldots, X_n\}$ a **random sample** of F. F is known as the **population distribution**, or just the **population**.

ℹ Population as an abstraction

The definition above is more general than the concept of population in everyday life (e.g. members of a group). It is an mathematical abstraction of the underlying "truth" we want to learn about.

Definition 55.3 (Converge in probability)**.** A sequence $Z_1, Z_2, \ldots$ of random variables converges to b in probability if for every number $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Z_n - b| < \epsilon) = 1.$$

The property is denoted by $Z_n \to_p b$.

Theorem 55.1 (Law of large numbers)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution for which the mean is μ and the variance is finite. Let $\bar{X}_n$ denote the* ***sample mean****, i.e. $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\bar{X}_n \to_p \mu.$$

Proof. Since $X_1, X_2, \ldots, X_n$ are i.i.d random variables from the same distribution. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1, 2, \ldots, n$. For every number $\epsilon > 0$, by the Chebyshev inequality,

$$P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

Hence,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

□

ℹ Sample mean as an random variable

Note that sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, is the weighted sum of random variables. It is therefore itself a random variable. It is easy to see that

$$\begin{aligned} E(\bar{X}_n) &= \mu \\ Var(\bar{X}_n) &= \frac{\sigma^2}{n} \to 0, \quad \text{as } n \to \infty \end{aligned}$$

The random variable $\bar{X}_n$ becomes fixed at μ as n becomes large. Thus, it converges to μ in a probabilistic sense.

Theorem 55.2 (Continuous function of RVs)**.** *If $Z_n \to_p b$, and $g(\cdot)$ is a continuous function, then $g(Z_n) \to_p g(b)$.*

The theorem implies that LLN does not only apply to the sample mean, but also applies to any moments of a random variable,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \to_p E(X^k).$$

This is the foundation of the method of moments estimation that we will discuss later.

Applications of LLN

It might seem that the LLN just states the obvious. But it has wide applications in daily life that you might not even realize. The LLM implies that: the uncertainty at the individual level becomes certain at aggregate level; the risks that are unmanageable at the individual level becomes manageable collectively.

Example 55.1 (Lottery)**.** A lottery company is designing a game with a 6-digit format. Each time someone buys a ticket, they receive a randomly generated 6-digit number. Only one number will win the grand prize of 1 million dollars. What should the company charge per ticket to break even?

Solution. Let $X_i \sim \text{Bern}(p)$ be the random variable that indicates whether the i-th ticket is a winner, where $p = 1/10^6$. For each individual who buys a ticket, the outcome is highly uncertain. But collectively, by the Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_p p.$$

Meaning that when n is large, the proportion of winners should be very close to p. Therefore, the total number of winners should be very close to np.

If it is estimated that there will be 10 million tickets sold. There would be almost exactly $10^7 p = 10$ winners. The total cost of the company is therefore 10 million. If the company charge \$1 per ticket, it is just enough to cover the cost. Any price above \$1 would make the business profitable.

Example 55.2 (Insurance)**.** Insurance is anther great application of the LLN. It is essentially the same as the the lottery game but most people do not realize it. Suppose there is a disease with mortality rate 1 out of a million. The medical expenditure to cure the disease is 1 million dollars. How much the insurance company should charge per customer to cover this disease?

Solution. The solution is essentially the same as above. As long as the number of customers is large enough, by the LLN, the number of claims should be very close to np. Thus the risks that are unmanageable at the individual level becomes manageable when pooled together.

Insurance is a great invention because it not only provides cover for individuals but also improve the capital efficiency of the society as a whole. This about this: Without the insurance, each individual has to set aside 1 million dollars pre-cautiously for the disease (if he is rich enough) or be exposed to the risk completely uncovered. The insurance product enables everyone to get covered at the fraction of the cost and thus frees up capital for more productive uses.

56 Central limit theorem

Definition 56.1 (Convergence in distribution)**.** Let $X_1, X_2, \dots, X_n$ be a sequence of random variables, and let F_n denote the CDF of X_n. It is said the sequence $X_1, X_2, \dots$ converges in distribution F if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for all x at which F(x) is continuous. The property is denoted as

$$F_n \to_d F.$$

F is called the **asymptotic distribution** of X_n.

Theorem 56.1 (Central limit theorem)**.** *Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a distribution with mean μ and finite variance σ^2. Let $\bar{X}_n$ be the sample mean. Then,*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to_d N(0,1).$$

Note that the CLT can be equivalently expressed as:

$$\begin{aligned}\sum_{i=1}^{n} X_i &\to_d N(n\mu, n\sigma^2) \\ \bar{X}_n &\to_d N\left(\mu, \frac{\sigma^2}{n}\right) \\ \frac{\bar{X}_n - \mu}{\sigma} &\to_d N\left(0, \frac{1}{n}\right)\end{aligned}$$

Proof. We will prove the CLT assuming the MGF of the X_i exists, though the theorem holds under much weaker conditions. Without loss of generality let $\mu = 1, \sigma^2 = 1$ (since we standardize it anyway). We show that the MGF of $\sqrt{n}\bar{X}_n = (X_1 + \dots + X_n)/\sqrt{n}$ converges to the MGF of the $N(0,1)$.

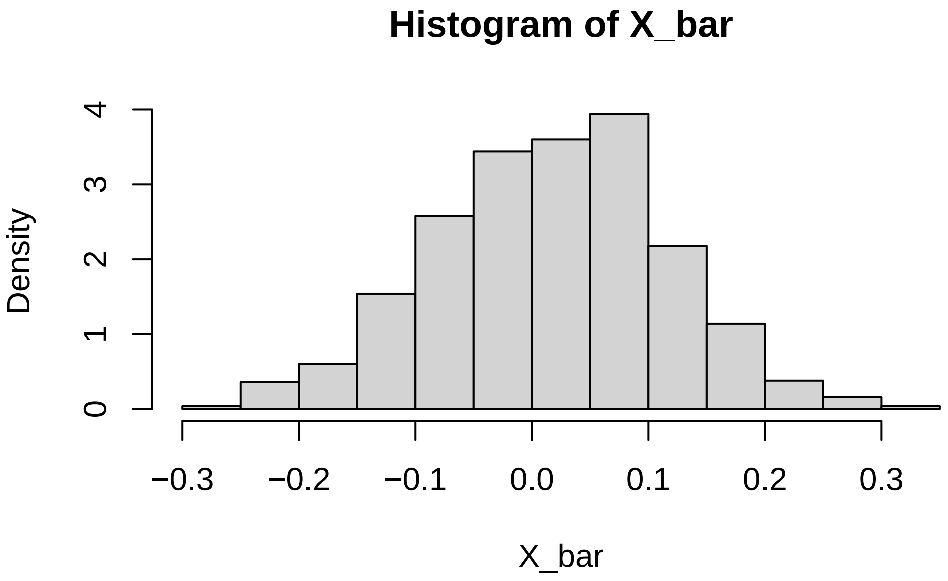
$$
\begin{aligned}
E(e^{\sqrt{n}\bar{X}_n}) &= E(e^{t(X_1+\cdots+X_n)/\sqrt{n}}) \\
&= E(e^{tX_1/\sqrt{n}})E(e^{tX_2/\sqrt{n}})\cdots E(e^{tX_n/\sqrt{n}}) \\
&= \left[E(e^{tX_i/\sqrt{n}})\right]^n \qquad \text{since } i.i.d \\
&= \left[E\left(1+\frac{tX_i}{\sqrt{n}}+\frac{t^2X_i^2}{2n}+o(n^{-1})\right)\right]^n \\
&= \left[1+\frac{t}{\sqrt{n}}E(X_i)+\frac{t^2}{2n}E(X_i^2)+o(n^{-1})\right]^n \\
&= \left[1+\frac{t^2}{2n}+o(n^{-1})\right]^n \\
&= \left[1+\frac{t^2/2}{n}+o(n^{-1})\right]^n \\
&\to e^{t^2/2} \qquad \text{as } n\to\infty
\end{aligned}
$$

Therefore, the MGF of $\sqrt{n}\bar{X}_n$ approaches the MGF of the standard normal. Since MGF determines the distribution, the distribution of $\sqrt{n}\bar{X}_n$ also approaches the standard normal distribution. □

ℹ Distributuion of $\bar{X}$ vs Distribution of X

The distribution of the sample mean $\bar{X}$ and the distribution of the random variables X_i are fundamentally different. The former is a theoretical distribution of the averages we would get if we drew repeated samples of size n. What the CLT says is that, regardless the distribution of X_i, the distribution of $\bar{X}$ always approaches Normal.

```r
set.seed(0)

# population of uniform distribution
G <- rnorm(10000)

# repeatedly sample from the population and compute sample mean
X_bar <- replicate(1000, mean(sample(G, 100)))

# distribution of the sample mean
hist(X_bar, prob = TRUE)
```



The CLT gives the distribution of the sample mean regardless of the underlying distribution. It allows us to gauge the accuracy of our sample mean as an estimate of the true mean:

$$P\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq 1\right) = .68$$

$$P\left(\bar{X}-\sigma^2/\sqrt{n} \leq \mu \leq \bar{X}+\sigma^2/\sqrt{n}\right) = .68$$

Thus, the interval $[\bar{X}-\sigma^2/\sqrt{n}, \bar{X}+\sigma^2/\sqrt{n}]$ contains the true mean 68% of the chance. This is known as **confidence intervals** that we will come back later. But we need one more theorem to make this actually works.

Theorem 56.2 (Slutsky's theorem). *Let X_n and Y_n be sequences of random variables. If $X_n \to_p c$ and $Y_n \to_d Y$, then*

- $X_n + Y_n \to_d c + Y$
- $X_n Y_n \to_d cY$
- $Y_n / X_n \to_d Y/c$, *provided c is invertible.*

In practice, the true variance σ^2 is unknown. But if we have an estimate from our sample such that $s^2 \to_p \sigma^2$, the Slutsky's theorem implies:

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \to_d N(0,1).$$

Thus we can feasibly construct the confidence interval with the sample variance: $[\bar{X}-s^2/\sqrt{n}, \bar{X}+s^2/\sqrt{n}]$.

57 Statistical inference

Definition 57.1 (Statistical model)**.** A statistical model is a formalized structure used for quantifying uncertainty, which consists of

1. a set of random variables of interest: $X_1, ..., X_n$
2. a specification of the joint distribution of the random variables $f(X_1, ..., X_n; \theta_1, ..., \theta_k)$
3. the parameters $\theta_1, ..., \theta_k$ that determine the distribution
4. (possibly) the distributions of the parameters $h(\theta_1, ..., \theta_k)$.

Definition 57.2 (Statistical inference)**.** A statistical inference is a procedure that produces a probabilistic statement about some or all parts of a statistical model (e.g. a parameter, a conditional distribution, etc).

Definition 57.3 (Parameter space)**.** The characteristics that determine the joint distribution of the random variables of interest is called the parameters of the distribution. The set of all possible values of the parameters is called the parameter space.

> Parameters as random variables
>
> There is a debtate over wheter unknown parameters should be treated as random variables or merely as numbers. Treating parameters as numbers is typically adopted by the **Frequentist framework**. We use the observables to provide the best estimate of the unknown parameter. For example,
>
> $$\bar{X}_n \to_p \mu$$
>
> Whereas treating parameters as random variables is typically associated with the **Bayesian framework**. We update the distribution of the parameters with the information provided by the observables:
>
> $$p(\theta|\bar{X}_n) \propto p(\bar{X}_n|\theta)p(\theta).$$

Definition 57.4 (Statistic)**.** Suppose we observe random variables $X_1, ..., X_n$. Let g be a real-valued function of n variables. Then the random variable $T = g(X_1, ..., X_n)$ is called a **statistic**.

Definition 57.5 (Sampling distribution)**.** Statistic T is a function of random variables, therefore it is itself a random variable. The distribution of T is called the **sampling distribution** of T.

The name comes from the fact that T depends on the sampling process (a different sample gives to a different value of the statistic).

Definition 57.6 (Estimator)**.** Let θ be a parameter of interest in a statistical model. An **estimator** of parameter θ is a statistic $\hat{\theta} = g(X_1, ..., X_n)$ constructed to learn about θ. If $X_1 = x_1, ..., X_n = x_n$ are observed, then $g(x_1, ..., x_n)$ is called the **estimate** of θ.

Constructing an estimator

How to construct an estimator is an art in itself, which needs to be guided by principles. We introduction the methods to construct estimators in the following sections.

Definition 57.7 (Method of moments)**.** Let $X_1, ..., X_n$ be a random sample from a distribution with at least k finite moments. Let $m_j = E(X^j; \theta)$ be the j-th order moment, $j = 1, ..., n$. Suppose the parameter of interest θ can be expressed as a function of the moments: $\theta = M(m_1, ..., m_k)$. The method-of-moments estimator of θ is given by

$$\hat{\theta} = M(\hat{m}_1, ..., \hat{m}_k)$$

where $\hat{m}_j$ is the sample moment: $\hat{m}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ for $j = 1, ..., k$.

The method of momment is guided by the LLN, as the sample moments converges to the true moments for large samples. As an example, the estimator for population mean μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The estimator for population variance σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

The LLN ensures $\hat{\sigma}^2 \to \sigma^2$ when the sample size is large. But we know $\hat{\sigma}^2$ is biased for small samples.

58 Maximum likelihood

Definition 58.1 (Likelihood function)**.** Let x be the observed values of a vector of random variables. Let $f(x;\theta)$ be the joint probability (density) function, where θ is the parameter(s) that governs the distribution. The likelihood function is f expressed as a function of θ:

$$\mathcal{L}(\theta; x) = f(x;\theta),$$

which represents the likelihood of observing x given θ.

Definition 58.2 (Maximum likelihood estimator)**.** The maximum likelihood estimator (MLE) of the parameter θ is found by maximizing the likelihood function:

$$\widehat{\theta}_{\text{MLE}} = \text{argmax}_\theta \ \mathcal{L}(\theta).$$

We express $\widehat{\theta}_{\text{MLE}}$ as a function of the random variables. Whereas the **estimate** given x is the value of θ that maximizes the chance of observing x.

Example 58.1 (MLE of the Bernoulli parameter)**.** Let $X_1, ..., X_n \overset{iid}{\sim} \text{Bern}(\theta)$ where θ is unknown. Find $\widehat{\theta}_{\text{MLE}}$.

First write down the joint probability function (also the likelihood funnction):

$$\mathcal{L}(\theta) = f(x;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

It is usually easier to maximize the log of $\mathcal{L}(\theta)$ since logarithm is a monotonic transformation:

$$\begin{aligned} l(\theta) = \ln \mathcal{L}(\theta) &= \sum_{i=1}^{n}[x_i \ln\theta + (1-x_i)\ln(1-\theta)] \\ &= \left(\sum_{i=1}^{n} x_i\right)\ln\theta + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-\theta). \end{aligned}$$

The maximum is achieved when $l'(\theta) = 0$, which occurs at $\theta = \bar{x}_n$. Therefore,

$$\widehat{\theta}_{\text{MLE}} = \bar{X}_n.$$

Example 58.2 (MLE of the mean from a Normal distribution)**.** Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. Find $\hat{\mu}_{\text{MLE}}$.

The likelihood function is:

$$\begin{aligned}\mathcal{L}(\mu; x, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] \\ &= \frac{1}{\sqrt{2\pi}^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]\end{aligned}$$

It is maximized by the value of μ by minimizing

$$Q(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2.$$

$Q(\mu)$ is minimized when $\mu = \bar{x}_n$. Therefore,

$$\hat{\mu}_{\text{MLE}} = \bar{X}_n.$$

Example 58.3 (MLE of the mean from a Uniform distribution)**.** Let $X_1, ..., X_n \overset{iid}{\sim} U(\theta)$ where θ is unknown. The population mean is $\mu = \frac{\theta}{2}$. Find $\hat{\mu}_{\text{MLE}}$.

The PDF $f(x; \theta)$ of each observation takes the form:

$$f(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, the joint PDF or the likelihood function is:

$$\mathcal{L}(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 \leq x_i \leq \theta \ (i = 1, ..., n), \\ 0, & \text{otherwise.} \end{cases}$$

The MLE of θ if found for which $\mathcal{L}$ is maximized. Since $\mathcal{L}$ is a decreasing function of θ, the estimate will be the smallest value of θ such that $\theta \geq x_i$ for $i = 1, ..., n$. Therefore, $\theta = \max\{x_1, ..., x_n\}$. The MLE of the mean μ is:

$$\hat{\mu}_{\text{MLE}} = \frac{\max\{X_1, ..., X_n\}}{2}.$$

Note that as $n \to \infty$, $\hat{\mu}_{\text{MLE}} \to \frac{\theta}{2}$ because the largest X_i would be very close to θ.

🔥 The importance of the underlying distribution

In statistics, you cannot simply assume that the arithmetic average $\bar{x}$ is always the best way to estimate the center of a population. The arithmetic mean is the MLE when the data comes from a normal distribution. However, it is not always the best estimate if the data is distributed otherwise. That's why understanding the distribution is the prerequisite of statistical inference.

59 Estimator properties

Definition 59.1 (Bias)**.** The bias of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$\text{Bias}[\hat{\theta}] = E(\hat{\theta}) - \theta.$$

An estimator is **biased** means its sampling is incorrectly centered. An estimator is **unbiased** if

$$E(\hat{\theta}) = \theta.$$

Theorem 59.1 (Sample mean)**.** *The sample mean, defined as:*

$$\bar{X}_n = \sum_{i=1}^{n} X_i,$$

is an ***unbiased*** *estimator of* $\mu = E(x)$*.*

Proof.

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

□

Theorem 59.2 (Sample variance)**.** *The sample variance, defined as:*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

is an unbiased estimator of σ^2 *(*$\sigma^2 < \infty$*).*

Proof. First note:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Take expectations of both sides:

$$E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = (n-1)\sigma^2$$

Dividing by $n-1$ makes $E(s^2) = \sigma^2$. □

Definition 59.2 (Mean absolute error)**.** The mean absolute error (MAE) of an estimator is defined as:

$$\mathrm{MSE}[\hat{\theta}] = E\left|\hat{\theta} - \theta\right|.$$

Definition 59.3 (Mean square error)**.** The mean square error (MSE) of an estimator is defined as:

$$\mathrm{MSE}[\hat{\theta}] = E\left[(\hat{\theta} - \theta)^2\right].$$

Theorem 59.3 (Bias-variance trade-off)**.** *For any estimator with a finite variance, we have*

$$MSE[\hat{\theta}] = Var[\hat{\theta}] + (Bias[\hat{\theta}])^2.$$

Proof. By expanding the MSE we find that

$$\begin{aligned}
\mathrm{MSE}[\hat{\theta}] &= E\left[(\hat{\theta} - \theta)^2\right] \\
&= E\left[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2\right] \\
&= E\left[(\hat{\theta} - E[\hat{\theta}])^2\right] + 2E(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta) + (E[\hat{\theta}] - \theta)^2 \\
&= Var[\hat{\theta}] + (\mathrm{Bias}[\hat{\theta}])^2.
\end{aligned}$$

Thus, the MSE is the variance plus the squared bias. A good estimator balances the variance (precision) and the bias (correctly centered). □

Definition 59.4 (Consistency)**.** An estimator is consistent if $\mathrm{MSE}[\hat{\theta}] \to 0$ as $n \to \infty$.

> ℹ Unbiasedness vs consistency
>
> Bias is the property of an estimator for finite samples. Consistency is the property of an estimator when the sample size gets large. A consistent estimator behaves well for large sample size. Whereas an unbiased estimator is correct centered even for small samples.

For unbiased estimator, MSE is solely determined by the variance of the estimator. Consider the sample mean $\bar{X}$ as an estimator for the population mean. Suppose the sample is large enough so that CLT holds. Then

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

But this is not a very useful in practice because σ^2 is usually unknown. So we replace it with its sample estimator.

Definition 59.5 (Standard error)**.** The standard error of an estimator $\hat{\theta}$ is defined as

$$SE(\hat{\theta}) = \hat{\sigma}(\hat{\theta}).$$

The standard error of $\bar{X}_n$ is

$$SE(\bar{X}_n) = \frac{s}{\sqrt{n}}$$

where s^2 is the sample variance.

The standard error indicates the "precision" of the estimator, thereby carrying a sense of "error". The smaller the standard error, the more precise the estimator.

60 Confidence intervals

Confidence intervals provide a method of adding more information to an estimator $\widehat{\theta}$ when we wish to estimate an unknown parameter θ. We can find an interval (A, B) that we think has high probability of containing θ. The length of such an interval gives us an idea of how closely we can estimate θ.

Definition 60.1. A $100(1-\alpha)\%$ confidence interval (CI) for θ is an interval $[L(\theta), U(\theta)]$ such that the probability that the interval contains the true θ is $(1-\alpha)$.

Due to randomness we rarely seek a confidence interval with 100% coverage as this would typically need to be the entire parameter space. Instead we seek an interval which includes the true value with reasonably high probability. Standard choices are $\alpha = 0.05$ and 0.10, corresponding to 95% and 90% confidence.

Confidence intervals are reported to indicate the degree of precision of our estimates. The narrower the confidence interval, the more precise the estimate. Because a small range of values contains the true parameter with high probability.

With the help of the CLT, it is not hard to find the CI for the sample mean $\bar{X}_n$. Let's set $\alpha = 5\%$, that is, we are trying to find the CI that contains the true mean 95% of the times. Assume our sample size n is large enough to invoke the CLT, we thus have

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Let's find the interval $[a, b]$ such that

$$P\left(a \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq b\right) = 1 - 2\Phi(L) = 0.95$$

since the normal distribution is symmetric, $b = -a$. By looking at the CDF of standard normal, we get $a = -1.96$, $b = 1.96$. Thus,

$$P\left(-1.96 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

With a little rearrangement, we have

$$P\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Therefore, the interval $\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$ contains the true mean 95% of the times.

Theorem 60.1. *The* $100(1-\alpha)\%$ *confidence interval for the sample mean* $\bar{X}_n$ *is* $\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$*, where* $z_{\alpha/2}$ *is the critical value such that* $\Phi(z_{\alpha/2}) = \frac{\alpha}{2}$*.*

In practice, because we do not know $\sigma/\sqrt{n}$, we replace it with the standard error $s/\sqrt{n}$. Thus, we compute the confidence interval as $\bar{X}_n \pm z_{\alpha/2}SE$. However, this replacement is not without risk. When the sample size is small, s is a very poor estimate of σ. For the approximation to be valid, we require either the sample size is large enough ($n \geq 30$ at least) or the population distribution is nearly normal. Some commonly used confidence levels:

- 90% CI: $\alpha = 0.1$, $z_{0.05} = 1.645$
- 95% CI: $\alpha = 0.05$, $z_{0.025} = 1.96$
- 99% CI: $\alpha = 0.01$, $z_{0.005} = 2.58$

We go through some common misunderstandings about confidence intervals through an example. Suppose we have a sample fo size 50 with mean 3.2 and standard deviation 1.74. We construct the 95% confidence interval as

$$\bar{X} \pm 1.96 \times \frac{1.74}{\sqrt{50}} \approx 3.2 \pm 0.5 = (2.7, 3.7).$$

Now check the following interpretations (true or false):

1. We are 95% confident that the sample mean is between 2.7 and 3.7.
   False. The CI definitely contains the sample mean $\bar{X}$.
2. 95% of the population observations are in 2.7 to 3.7.
   False. The CI is about covering the population mean, not for covering 95% of the entire population.
3. The true mean falls in the interval (2.7, 3.7) with probability 95%.
   False. The true mean μ is a fixed number, not a random one that happens with a probability.
4. If a new random sample is taken, we are 95% confident that the new sample mean will be between 2.7 and 3.7.
   False. The confidence interval is for covering the population mean, not for covering the mean of another sample.

5. This confidence interval is not valid if the population or sample is not normally distributed.
   False. The construction of the CI only uses the normality of the sampling distribution of the sample mean (by the CLT). Neither the population nor the sample is required to be normally distributed.

So what is exactly the thing that has a 95% change to happen? It is the procedure to construct the 95% interval. About 95% of the intervals constructed following the procedure will cover the true population mean μ. After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn't. But if we were able to take many such samples and reconstruct the interval many times, 95% of the intervals will contain the true mean.

61 Hypothesis testing

Confidence interval allows us to construct an interval estimate of a population parameter. Hypothesis testing allows us to test specific hypothesis about a population parameter with the evidence obtained from a sample. The earliest use of statistical hypothesis testing is generally credited to the question of whether male and female births are equally likely (null hypothesis), which was addressed in the 1700s by John Arbuthnot and later by Pierre-Simon Laplace.

Let p be the population ratio (defined as the ratio of boys to the total number of babies). We hypotheses that

$$H_0 : p = 0.5$$

This is called the **null hypothesis**, which is the hypothesis we want to test. If the null hypothesis is false, we have

$$H_1 : p \neq 0.5$$

This is called the **alternative hypothesis**. How am I able to test which hypothesis is true? I can answer this question by collecting a small sample. Suppose I have collected a sample of 50 babies computed a sample ratio of $\hat{p} = 0.55$. Does it prove or disprove the hypothesis?

Note that the ratio $\hat{p}$ can be regarded as a sample mean. Let X_i be a random variable that equals 1 if the i-th baby is a boy and 0 otherwise. Then, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The variance of $\hat{p}$ is given by

$$Var(\hat{p}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{p(1-p)}{n}$$

since X_i is a Bernoulli random variable. By the Central Limit Theorem, we have

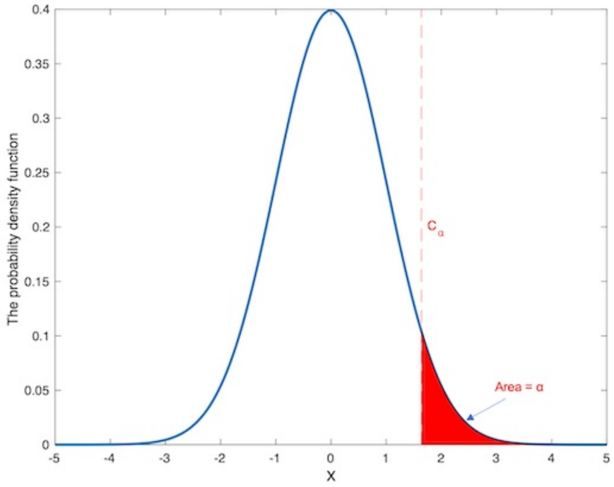
$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$$

Suppose H_0 is true, then we know the distribution of $\hat{p}$. In particular, there is 95% chance that $\hat{p}$ would be in the interval

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.5 \pm 0.14$$

Our observed sample mean $\hat{p} = 0.55$ is not outrageous. It is well within this interval. That means the evidence is not against the null hypothesis. It does not mean H_0 is true. But it is reasonable given we have observed a sample mean $\hat{p} = 0.55$.

Suppose we have observed $\hat{p} = 0.65$. This piece of evidence does not seem to be consistent with the null hypothesis. Because if H_0 is true, we only have less than 5% chance of observing this sample mean. It is extremely unlikely. Based on this sample, we are more inclined to reject the H_0. Rejecting the null hypothesis does not mean it is false, but it means our evidence does not support this hypothesis.



***p*-value**: the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis. Thus, The smaller the p-value, the stronger the evidence against the H_0.

In some studies, we can simply report the p-value and let people judge whether the evidence is strong enough. In other studies, we prefer to select a cut-off value α, call the **significance level**, and follow the rule:

- If the p-value $< \alpha$, reject H_0;
- If the p-value $> \alpha$, do not reject H_0.

Commonly used significance levels: 0.05 and 0.01. And we like to use the word "significant" to describe the test result:

- A test with p-value < 0.05 is said to be (statistically) **significant**;
- A test with p-value < 0.01 is said to be highly **significant**.

When we make a decision about accepting or rejecting a hypothesis, there are chances that we make a mistake. There are two types of mistakes: **Type 1 error** and **Type 2 error**.

	Decision	
	Reject H_0	Fail to reject H_0
H_0 is true	Type 1 error	✓
H_0 is false	✓	Type 2 error

Type 1 error is rejecting the H_0 when it is true. **Type 2 error** is failing to reject the H_0 when it is false. Usually, it is more important to control the Type 1 error than the the Type 2 error. That is, we want to minimize the chance of falsely rejecting the null hypothesis.

In the example above, we reject the null hypothesis on the ground that there is only 2.3% of the chance that we could observe this sample. Therefore, the probability of Type 1 error is only 2.3%.

If we make decisions based on a significance level, the significance level is the Type 1 error rate. In other words, when using a 5% significance level, there is 5% chance of making a Type 1 error.

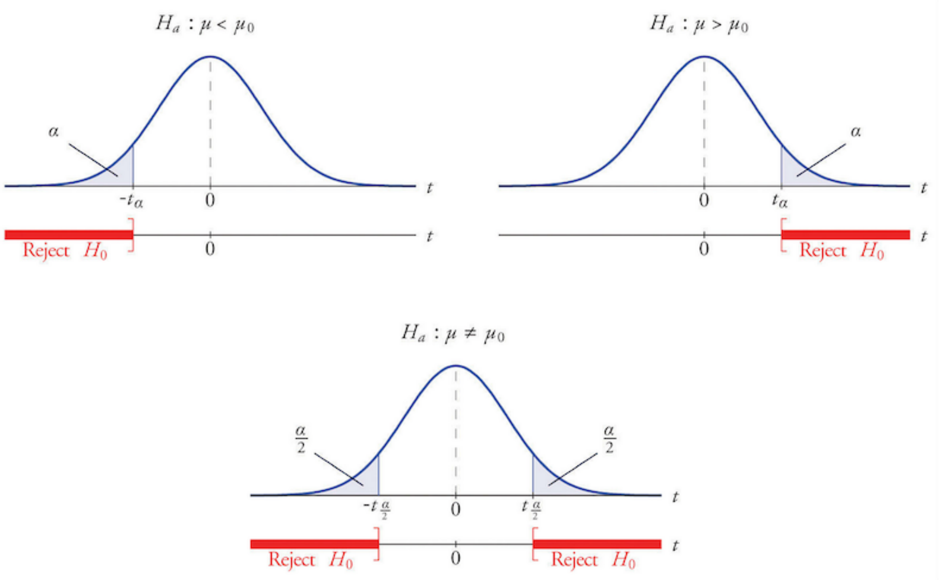
$$P(\text{Type 1 error}|H_0 \text{ is true}) = \alpha$$

This is why we prefer small values of α—smaller α reduces the Type 1 error rate. However, significance level doesn't control Type 2 error rate.

Hypothesis testing with z-statistics

We may have noticed that, in the above example, the assumption that the population σ is known is unrealistic. In practice, we approximate it with the standard error $s/\sqrt{n}$. The approximate is valid if the the sample size is large enough or the underlying distribution is nearly normal. If this is not the case, we would opt for a t-test. Here we summarize the steps of testing for a population mean with z-statistics.

We notice that the **two-sided** hypothesis tests are very closed related to the concept of confidence intervals. A two-sided test means we are interested in rejection regions on both sides of the tail distribution. Typically, the alternative hypothesis is $H_1 : \mu \neq \mu_0$.



Suppose we are doing a hypothesis test under the significance level α, the region of accepting the H_0 is

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{SE} \leq z_{\alpha/2}$$

such that the rejection region (p-value) has probability α. This is equivalent to

$$\bar{X} - z_{\alpha/2} SE \leq \mu \leq \bar{X} + z_{\alpha/2} SE$$

which is exactly the $100(1-\alpha)\%$ confidence interval of $\bar{X}$. Therefore, for a two-sided test, we have the rule:

- Reject H_0 if μ is not in the $100(1-\alpha)\%$ CI: $\bar{X} \pm z_{\alpha/2} SE$

We conclude this chapter by reiterating a couple of critical points that could be easily misunderstood.

Rejecting H_0 doesn't means we are 100% sure that H_0 is false. We might make Type 1 errors. Setting a significance level just guarantee we won't make Type 1 error too often.

Failing to reject H_0 does not necessarily mean H_0 is true. We could make a type 2 error when failing to reject H_0. Moreover, unlike type 1 error rate is controlled at a low level, type 2 error rate is usually quite high. When we fail to reject H_0, it just means the data are not able to distinguish between H_0 and H_1. That's why we say *fail to reject*. p-value is not the probability that the H_0 is true.

Saying that results are statistically significant just informs the reader that the findings are unlikely due to chance alone. However, it says nothing about the practical importance of the finding. For example, rejecting the H_0: $\mu = \mu_0$ does not tell us how big the difference $|\mu - \mu_0|$ is. Mostly in practice we care more about the magnitude of this difference, rather than the fact that they are indeed different. It is possible that the difference is too small to be relevant even if it is significant.

Hypothesis testing with t-statistics

When the sample size is small, we opt for t-test for more reliable hypothsis testing. Define test statistics

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

where s is the sample standard deviation. For small samples, this test statistics follows a Student t-distribution with n degrees of freedom, $T \sim t(n)$.

Why Student-t distribution? Recall the definition of Student-t distribution: when the underlying distribution of $X_1, X_2, \dots, X_n$ is Normal, sample variance s^2 follows a χ^2 distribution. T follows t distribution by definition regardless of the sample size. However, if the underlying distribution is not normal, this argument loses ground. We use t-test mainly as a convention. But t distribution has heavier tails than standard normal, meaning that we are more likely to reject a hypothesis based on t distribution. In other words, t-test is a more conservative choice than z-test for small samples.

one-tail α	0.05	0.025	0.005
two-tail α	0.10	0.05	0.01
d.f.			
10	1.812	2.228	3.169
20	1.725	2.086	2.845
30	1.697	2.042	2.750
z value	1.645	1.960	2.576

The table shows a few critical values for t-test with different degrees of freedom (d.f.). We can see as the sample size gets larger, t distribution converges to standard normal.

62 Bayesian inference

Bayesians take a different approach to statistical inference. Bayesians treat the unknown parameters as random variables associated with distributions. Instead of trying to estimate the "true" value of the parameters, the distribution of the parameters gets updated with the information contained in the data.

Definition 62.1 (Prior distribution)**.** Suppose we have a statistical model with parameter θ. If we treat θ as *random*, then the distribution that one assigns to θ before observing the other random variables of interest is called its prior distribution, denoted as $p(\theta)$.

Definition 62.2 (Posterior distribution)**.** Consider a statistical inference problem with parameter θ and the vector observables $x = (x_1, ..., x_n)$. The conditional distribution of θ given x is called the posterior distribution of θ, denoted as $p(\theta|x)$.

Theorem 62.1 (Bayesian inference)**.** *Suppose random variables $X_1, ..., X_n$ has joint probability (density) function $f(x_1, ..., x_n|\theta)$. The parameter θ has prior distribution $p(\theta)$. Then the posterior distribution of θ is*

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \propto f(x|\theta)p(\theta).$$

$f(x|\theta)p(\theta)$ is also known as the likelihood function.

The essence of Bayesian inference is to *update* the distribution of the parameter with the information in the data. The posterior distribution is a function of θ, the denominator bahaves like a normalizing constant. So we don't lost anything if we only focus on the likelihood function and the prior.

Definition 62.3 (Conjugate prior)**.** In Bayesian inference, if, given the likelihood function $f(x|\theta)$, the posterior distribution $p(\theta|x)$ is in the same **probability distribution family** as the prior distribution $p(\theta)$, the prior and posterior are then called conjugate distributions with respect to that likelihood function. The prior is called a conjugate prior.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise, numerical integration would be necessary.

Theorem 62.2 (Beta-Binomial conjugacy)**.** *Let $X \sim Bin(n, \theta)$. Assume θ has prior distribution: $p(\theta) \sim Beta(a, b)$. We observe $X = k$. Then the posterior distribution is:*

$$p(\theta|X = k) \sim Beta(a + k, b + n - k).$$

Proof. The likelihood function of Binomial distribution is:

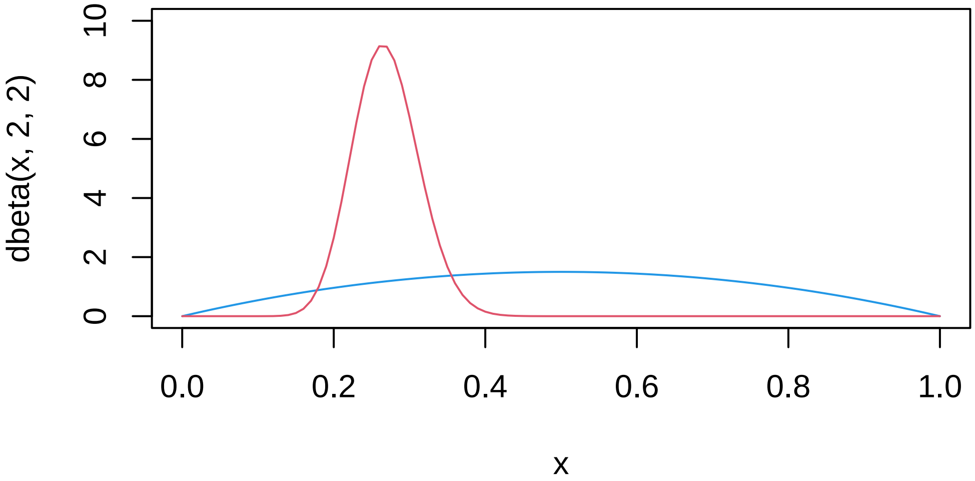
$$f(k|\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}.$$

Combining the likelihood and the prior:

$$\begin{aligned} p(\theta|k) &\propto f(k|\theta)p(\theta) \\ &\propto \binom{n}{k} \theta^k (1 - \theta)^{n-k} \cdot \frac{1}{\beta(a, b)} \theta^{a-1} (1 - \theta)^{b-1} \\ &\propto \theta^{a+k-1} (1 - \theta)^{b+n-k-1}. \end{aligned}$$

This is the kernel of Beta$(a + k, b + n - k)$. □

```
# unknown parameter
p <- 0.3

# number of observations
n <- 100

# generate Bernoulli observations
X <- 1* (runif(n) < p)

# sum of positive outcomes
k <- sum(X)

# the prior distribution (blue)
curve(dbeta(x, 2, 2), col=4, ylim=c(0,10))

# the posterior distribution (red)
curve(dbeta(x, 2+k, 2+n-k), col=2, add=TRUE)
```



Theorem 62.3 (Poisson-Gamma conjugacy)**.** *Let $X \sim Pois(\lambda)$. Assume the unknown parameter λ has a prior distribution: $p(\lambda) \sim Gamma(a, b)$. We observe $X = k$. Then the posterior distribution is:*

$$p(\lambda|k) \sim Gamma(a + k, b + 1).$$

Proof. The likelihood function (Poisson PMF):

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The prior (Gamma PDF):

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

We multiply the components, focusing only on the parts involving λ:

$$p(\lambda|k) \propto (\lambda^k e^{-\lambda}) \cdot (\lambda^{a-1} e^{-b\lambda})$$

Combine the λ terms:

$$p(\lambda|k) \propto \lambda^{(a+k)-1} e^{-(b+1)\lambda}$$

We recognize this is the kernel of Gamma$(a + k, b + 1)$. □

Theorem 62.4 (Normal-normal conjugacy)**.** *Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. Assume the prior distribution is also normal: $p(\mu) \sim N(\mu_0, v_0^2)$. Then the posterior distribution is also normal:*

$$p(\mu|x) \sim N\left(\frac{\sigma^2 \mu_0 + n v_0^2 \bar{x}_n}{\sigma^2 + n v_0^2}, \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2}\right).$$

Proof. The Prior distribution, ignoring the constant, is:

$$p(\mu) \propto \exp\left(-\frac{1}{2v_0^2}(\mu-\mu_0)^2\right)$$

The likelihood of the independent sample:

$$p(x|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Multiply the prior and likelihood exponentials:

$$p(\mu|x) \propto \exp\left(-\frac{1}{2}\left[\frac{(\mu-\mu_0)^2}{v_0^2}+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\right]\right)$$

Group terms by μ^2 and μ and ignore the constant:

$$p(\mu|x) \propto \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{1}{v_0^2}+\frac{n}{\sigma^2}\right)-2\mu\left(\frac{\mu_0}{v_0^2}+\frac{n\bar{x}}{\sigma^2}\right)\right]\right)$$

We want the posterior to look like a Normal distribution $N(\mu_n, v_n^2)$, which has the form:

$$\exp\left(-\frac{1}{2v_n^2}(\mu-\mu_n)^2\right) \propto \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{v_n^2}-\frac{2\mu\mu_n}{v_n^2}\right]\right)$$

We match the coefficients from our derived equation to this standard form:

$$\frac{1}{v_n^2}=\frac{1}{v_0^2}+\frac{n}{\sigma^2}=\frac{\sigma^2+nv_0^2}{v_0^2\sigma^2}$$

$$\frac{\mu_n}{v_n^2}=\frac{\mu_0}{v_0^2}+\frac{n\bar{x}}{\sigma^2}=\frac{\sigma^2\mu_0+nv_0^2\bar{x}}{v_0^2\sigma^2}$$

Solving for μ_n and v_n^2 gives the answer desired. □

Frequentist estimator vs Bayesian posterior

Feature	**Frequentist Estimator**	**Bayesian Posterior**
What is the Parameter? (θ)	**A Fixed, Unknown Constant.**The true mean is a specific number (e.g., 5.2) hidden in nature. It does not move or fluctuate.	**A Random Variable.**The parameter is uncertain. We describe it using a probability distribution that reflects our state of knowledge.
What is the Data? (X)	**Random & Repeatable.**We imagine the data is just one of infinite possible samples we could have drawn.	**Fixed.**The data is the only solid reality we have. We condition our beliefs on this specific dataset.
The Output	**A Point Estimate.**A single number (like $\bar{x}$) or a Confidence Interval.	**A Probability Distribution.**A full curve (the Posterior) showing which values are most likely.
Probability Meaning	**Long-Run Frequency.**The frequency of occurrence out of infinitely many trials.	**Degree of Belief.**How certain we are about something given our knowledge.
Use of Prior Knowledge	**Forbidden.**The data must speak for itself. Bringing in outside beliefs is considered “bias.”	**Required (The Prior).**You start with an initial belief ($p(\theta)$) and update it with data to get the Posterior.