Lecture 1

Probability basics

- Random phenomena are all around us: outcome of a coin flip, weather patterns, stock market, getting disease
- Random means non-deterministic, i.e. phenomena where there is an associated uncertainty about the outcome; we cannot predict perfectly (next day weather, roll of a die, price of Apple stock)
- Inability to predict perfectly could be due to lack of full information (e.g. flipping coin) or intrinsic (quantum mechanics)
- Error in measurements: anything we measure has uncertainty associated with it
- Randomness does not imply complete lack of pattern (e.g. a multiple coin toss will have about half heads and half tails), meteorologists can assess (and quantify!) whether the chance of rain is high or low
- Probability is the branch of mathematics that allows us to model randomness and studies properties of random phenomena
- Probability has application in computer science, machine learning, artificial intelligence, finance, statistics
- Probability allows as to model observations/data arising from phenomena/processes that are or can be modeled as random
- Statistics can be thought of as providing solutions to the inverse problem of probability: given observations/data infer properties about the underlying random phenomenon/process that generated the data
- Probability can be defined formally (at different levels of mathematical rigor) or can be treated more intuitively

Sample space

Definition: the set of all possible outcomes of an experiment of interest

Examples:

- Single coin tossing: $\Omega = \{H, T\}$
- Month of birth of a randomly chosen person: $Omega = \{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec\}$
- Whether a Youtube video will be clicked when presented to a potential viewer $Omega = \{YES, NO\}$

Event space

Event Space: all possible events (collection of outcomes) we will consider.

For discrete sample spaces event space is typically all possible subsets of Ω

- Example: single coin toss
  - Sample space: $\Omega = \{H, T\}$
  - Event space: $\mathcal{F} = \{\emptyset, \{T\}, \{H\}, \{H, T\}\}$
- Example: birth month of a randomly chosen person
  - $\Omega = \{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec\}$
  - $\mathcal{F} = \{\emptyset, \{Jan\}, \ldots, \{Dec\}, \{Jan, Feb\}, \ldots, \{Jan, \ldots, Dec\}\}$
- Q: how many elements in $\mathcal{F}$?
  - If your sample space Ω has n elements (outcomes), then the event space $\mathcal{F}$ contains 2^n elements.

We require that union of events and intersection of events are also events:

$$A, B \in \mathcal{F} \implies A \cup B \text{ and } A \cap B \in \mathcal{F} \tag{1}$$

$\mathcal{F}$ is closed under unions and intersections

E.g. $A = \text{First semester} = \{\text{Jan}, \text{Feb}, \text{Mar}, \text{Apr}, \text{May}, \text{Jun}\}$

$B = \{\text{May}, \text{Jun}, \text{July}\}$

$A \text{ or } B = A \cup B = \{\text{Jan}, \text{Feb}, \text{Mar}, \text{Apr}, \text{May}, \text{Jun}, \text{Jul}\}$

$A \text{ and } B = A \cap B = \{\text{May}, \text{Jun}\}$

Complements

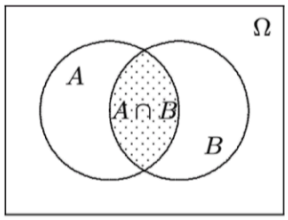
$$\text{not } A = A^c = \{Jul, Aug, Sep, Oct, Nov, Dec\} \tag{2}$$

$$A \setminus B = A - B = \{Jan, Feb, Mar\} \tag{3}$$

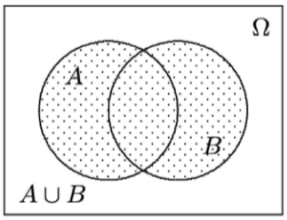
$$A^c = \Omega \setminus A = \Omega - A \tag{4}$$

$\mathcal{Z}$ is closed under union, intersection, and set difference

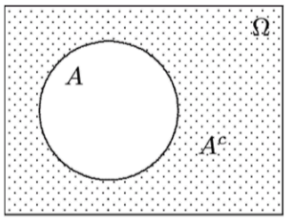
Venn Diagrams:



Intersection $A \cap B$



Union $A \cup B$



Complement A^c

Disjoint events

$$A \text{ and } B = \emptyset \quad \text{i.e., no elements in common} \tag{5}$$

E.g.

- $A = \text{First semester} = \{Jan, Feb, Mar, Apr, May, Jun\}$
- $B = \text{Second semester} = \{Jul, Aug, Sep, Oct, Nov, Dec\}$

DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c \tag{6}$$

$$(A \cap B)^c = A^c \cup B^c \tag{7}$$

Example: Let J be the event "John is guilty" and M the event "Mary is guilty."

- $(J \cap M)^c = \text{Not true that both John and Mary are guilty}$
- $J^c \cup M^c = \text{Either John or Mary are not guilty}$
- $(J \cup M)^c = \text{Not true that either John or Mary (or both) are guilty}$
- $J^c \cap M^c = \text{Neither John nor Mary are guilty}$

(prove DeMorgan's Laws to practice with set operations)

Probability functions

A probability function is a 'set function' that assigns a real number to each event in $\mathcal{Z}$:

$$P : \mathcal{Z} \to \mathbb{R} \text{ such that:} \tag{8}$$

1.

$$P(A) \geq 0 \tag{9}$$

2.

$$P(\Omega) = 1 \tag{10}$$

3.

$$P(A \cup B) = P(A) + P(B) \text{ if } A \cap B = \emptyset \quad (\text{i.e. additive on disjoint sets}) \tag{11}$$

The probability reflects the chances an event occurs, 0 being impossible and 1 being certain

Example: Fair coin

$$P(\{H\}) = P(\{T\}) = \frac{1}{2} \quad (\text{we will simplify notation as } P(H) = P(T)) \tag{12}$$

$$P(\emptyset) = 0 \tag{13}$$

$$P(\{H, T\}) = P(\Omega) = 1 \tag{14}$$

Example: birth month of a randomly chosen person

$$P(Jan) = P(Feb) = \cdots = P(Dec) = \frac{1}{12} \tag{15}$$

Or perhaps a more reasonable assignment of probabilities would be proportional to the number of days in each month:

$$P(Jan) = \frac{31}{365}, P(Feb) = \frac{28}{365}, \ldots \tag{16}$$

(This shows that it is us, the users who assign probabilities; probabilities are not 'laws of nature')

Any probability in a discrete sample space can be constructed like in the previous examples:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\} \tag{17}$$

$$p_1 + p_2 + \cdots + p_n = 1, \quad p_1 \geq 0, p_2 \geq 0, \ldots p_n \geq 0 \tag{18}$$

$$P(A) = \sum_{w_i \in A} p_i, \quad \forall A \subset \Omega \tag{19}$$

is a probability function.

Exercise: show this

Property 3 holds for any number of disjoint events:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C), \quad \text{if } A \cap B = \emptyset, A \cap C = \emptyset, B \cap C = \emptyset \tag{20}$$

E.g. the sets:

- $A = \{\text{First Trimester}\} = \{Jan, Feb, Mar\}$
- $B = \{\text{Second Trimester}\} = \{Apr, May, Jun\}$
- $C = \{Nov\}$

Are pairwise disjoint

$$P(A \cup B \cup C) = P(\{Jan, Feb, Mar, Apr, May, Jun, Nov\}) = \frac{7}{12} \tag{21}$$

$$P(A) = P(B) = \frac{3}{12}, \quad P(C) = \frac{1}{12} \tag{22}$$

In general:

$$P(A_1 \cup \ldots \cup A_n) = P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \tag{23}$$

Provided the events are pairwise disjoint:

$$A_k \cup A_l = \emptyset, \forall k, l \in \{1 \ldots n\}, k \neq l \tag{24}$$

Derived properties

If $A, B \in \mathcal{Z}$

$$P(\emptyset) = 0 \tag{25}$$

$$0 \leq P(A) \leq 1 \tag{26}$$

$$P(A^c) = 1 - P(A) \tag{27}$$

$$P(A - B) = P(A) - P(A \cap B) \tag{28}$$

$$B \subset A \Rightarrow P(A - B) = P(A) - P(B) \tag{29}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{30}$$

(Good practice exercise to show these)

Repeated experiments/Product of sample spaces

E.g. Flip a coin twice:

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\} = \{H,T\} \times \{H,T\} = \{H,T\}^2 \tag{31}$$

E.g. Flip a coin n times:

$$\Omega = \{H,T\}^n \quad \text{all } n\text{-tuples with elements in } H,T \tag{32}$$

E.g. Flip a coin and then pick a month at random:

$$\Omega_1 = \{H,T\}, \quad \Omega_2 = \{Jan, Feb, Mar, \ldots, Dec\} \tag{33}$$

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2), \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\} \tag{34}$$

Q: How many elements in Ω?

Repeated experiments/Product of sample spaces

If we have a probability function P_1 defined in Ω_1 and a probability P_2 defined in Ω_2 we can naturally defined a probability P in $\Omega^1 \times \Omega^2$ as:

$$P(\{\omega^1, \omega^2\}) = P_1(\omega^1) P_2(\omega^2) \tag{35}$$

E.g.

$$P(\{H, Jul\}) = P(H) \times P(Jul) = \frac{1}{2} \times \frac{1}{12} \tag{36}$$

(This is how we model independence, which will cover next week)

Uniform probability spaces

In many applications it makes sense to assign the same probability to all elements of a finite sample space

E.g. two coin flip:

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\} = \{H,T\} \times \{H,T\} \tag{37}$$

$$P(H,H) = P(H,T) = P(T,H) = P(T,T) = \frac{1}{4} \tag{38}$$

In general, a uniform probability space with $|\Omega| = n$, has:

$$P(\omega) = \frac{1}{n} \quad \forall \omega \in \Omega \tag{39}$$

And the probability of an event is the number of elements in the event divided by the total number of elements in the sample space:

$$P(A) = \frac{|A|}{|\Omega|} \tag{40}$$

E.g. Pick a single card from a well shuffled standard 52-card deck:

$$P(\text{Ace}) = \frac{4}{52} \tag{41}$$

$$P(\text{diamond suit}) = P(\lozenge) = \frac{13}{52} = \frac{1}{4} \tag{42}$$

Multiplicative counting principle

Many problems in probability theory require that we count the number of ways that a particular event can occur. This kind of counting falls under the area of mathematics called combinatorics.

The Multiplicative counting principle (MP).

Suppose that we perform r experiments such that the k^{th} experiment has n_k possible outcomes, for $k = 1, 2, \ldots, r$. Then there are a total of:

$$n_1 \times n_2 \times n_3 \times \cdots \times n_r \tag{43}$$

possible outcomes for the sequence of r experiments.

Example 1: Need to choose a password for an online account. Password must consist of two lowercase letters (a to z) followed by one capital letter (A to Z) followed by four digits (0,1, \dots, 9).

Example 2: How many subsets does a set with n elements have?

Permutations

How many five-card hands are possible from a standard fifty-two card deck? (if order matters)

$$52 \times 51 \times 50 \times 49 \times 48 = \frac{52!}{47!} = 311,875,200 \text{ by the MP} \tag{44}$$

In general, a $k\text{-permutation}$ of n distinct objects is a way to arrange k objects out of the n in a row (order matters).

The number of $k\text{-permutations}$, $p(n,k)$, is given by:

$$p(n,k) = \frac{n!}{(n-k)!} \tag{45}$$

$n\text{-permutations}$ are often refer to as just permutations. There are:

$$\frac{n!}{(n-n)!} = n! \text{ permutations} \tag{46}$$

Combinations

How many five-card hands are possible from a standard fifty-two card deck? (if order does not matter)

$$52 \times 51 \times 50 \times 49 \times 48 = \frac{52!}{47!} \quad \text{ordered arrangements} \tag{47}$$

Each arrangement was counted $5!$ times so the number of unordered arrangements is:

$$\frac{52!}{47!5!} = 2,598,960 \tag{48}$$

In general, a k-combination of n distinct objects is a way to arrange k objects out of the n when order does not matter.

The number of k-combinations, $c(n, k)$, is given by:

$$c(n, k) = \frac{n!}{(n-k)!k!} = \binom{n}{k} \tag{49}$$

$$p(n, k) = c(n, k) \times k! \tag{50}$$

Card problem

Suppose we deal a 5-card hand from a regular 52-card deck. Which is larger, P(One king) or P(Two hearts)?

$$P(\text{One king}) = \frac{\binom{4}{1} \times \binom{48}{4}}{\binom{52}{5}} \tag{51}$$

$$P(\text{Two hearts}) = \frac{\binom{13}{2} \times \binom{39}{3}}{\binom{52}{5}} \tag{52}$$

```
4 * choose (48, 4) / choose (52, 5)
## [1] 0.2994736
choose(13, 2) * choose (39, 3) / choose (52, 5)
## [1] 0.2742797
```

De Mere's problem

- Dice game that played an important role in the historical development of probability.
- Chevalier de Méré had been betting that, in four rolls of a die, at least one six would turn up.
- He was winning consistently and, to get more people to play, he changed the game to bet that, in 24 rolls of two dice, a pair of sixes would turn up.
- De Méré lost with 24 and felt that 25 rolls were necessary to make the game favorable.
- Was De Méré right?

Simulating De Mere's problem

Single die roll in R:

```
sample (1:6, size = 1, replace = TRUE)
[1] 2
```

Four rolls of one die

```
sample (1:6, size = 4, replace = TRUE)
[1] 1 2 4 3
```

Simulating De Mere's problem

Checking if a six came up

```
any(sample(1:6, size = 4, replace = TRUE) == 6)
## [1] FALSE
```

Full simulation:

```
nreps = 1000
set.seed (2021)
results = numeric(0)
for (i in 1:nreps) results[i] = any (sample(1:6, size = 4, replace = TRUE) == 6)
mean(results)

## [1] 0.507
```

De Mere's problem

Questions:

1. Based on this simulation result, do you think the bet's favorable?
2. Derive/compute the actual probability (hint: use that all outcomes of the four rolls of a die are equally likely)
3. Simulate the second scheme (24 rolls of two dice). What can you say about the favorability of the bet?
4. Derive/compute the actual probability. How about for 25 rolls of two dice?

Birthday 'paradox'

How many people n do we need to have in a room to make it a favorable bet (probability of success greater than $1/2$) that two people in the room will have the same birthday?

Assume all 365 b-days are equally likely.

1. Perform a simulation in R to answer this question (hint: use the base R function 'duplicate' to check that whether there are matching b-days)
2. Compute the probability by mathematical derivation and plot the probability as a function of n. (Hint: use the multiplication principle to count)

Infinite (countable) sample spaces

E.g. flip a coin until first heads appears:

What is the right probability space for this experiment?

$$\Omega = \{1, 2, 3, \ldots\} = \mathbb{N} \tag{53}$$

Sample space has to be infinite because no guarantee experiment will terminate in a finite number of steps!

If we assume that after k flips all k-tuples are equally likely what should the probability $P(k)$ be?

$$k = 1 = \{H\} \Rightarrow P(1) = \frac{1}{2} \tag{54}$$

$$k = 2 = \{T, H\} \Rightarrow P(2) = \frac{1}{4} \tag{55}$$

$$\vdots \tag{56}$$

$$k = \{\underbrace{T, \ldots, T}_{\text{k-1 times}}, H\} \Rightarrow P(k) = \frac{1}{2^k} \tag{57}$$

Does this result in a probability function?

Infinite (countable) sample spaces

For infinite sample spaces need to change additivity rule to countably additivity rule:

2'.

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \quad \text{provided they are disjoint } (A_k \cap A_l = \emptyset \quad \forall k, l \in \mathbb{N}, k \neq l) \tag{58}$$

$$\Omega = \{H, T\}^{\infty} \quad \text{all infinite sequences with elements in } \{H, T\} \tag{59}$$

Verification of $P(\Omega) = 1$:

$$P(\Omega) = P(\{1, 2, 3, \ldots\}) = P(1) + P(2) + P(3) + \cdots = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1 \tag{60}$$

(Used that for a geometric series:

$$1 + r + r^2 + \cdots = \sum_{k=0}^{\infty} r^k = \frac{1}{1-r} \quad \text{if } 0 \leq r < 1 \tag{61}$$

)

Q: What's the probability that it'll take an even number of tosses until the first heads?

Finite and countable sets: a mathematical aside

A set is finite if its elements can be put in one-to-one correspondence with with:

$$\{1, 2, \ldots, n\} \quad \text{for some } n \in \mathbb{N} \tag{62}$$

E.g., the set of students in the classroom, the set of inhabitants in the world, the set of stars in the Milky Way.

A set is countable if its elements can be put in one-to-one correspondence with with the natural numbers:

$$\mathbb{N} = \{1, 2, 3, \ldots\} \tag{63}$$

E.g. The set of natural numbers, the set of odd numbers $(n \to 2n + 1)$, the set of even numbers $(n \to 2n)$, the set of primes, the set of rational numbers ($\mathbb{Q}$)**!!**

Examples of infinite non-countable sets:

$$\mathbb{R}, \quad \text{the set of irrational numbers } \mathbb{R} - \mathbb{Q}, \quad \mathbb{R}^2 \tag{64}$$

Lecture 2 - Conditional Probability and Independence

Review from Last Class

- Sample spaces (finite)
- Probability functions on discrete spaces
- Uniform probability spaces
- Multiplicative counting principle
- Permutations/combinations

Repeated Experiments/Product of Sample Spaces

E.g. Flip a coin twice:

$$\Omega_1 = \{H, T\} \tag{65}$$

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\} = \Omega_1 \times \Omega_1 = \Omega_1^2 \tag{66}$$

E.g. Flip a coin n times:

$$\Omega = \Omega_1 \times \cdots \times \Omega_1 = \Omega_1^n \quad \text{all n-tuples with elements in } \{H, T\} \tag{67}$$

E.g. Flip a coin and then pick a month at random:

$$\Omega_1 = \{H, T\}, \quad \Omega_2 = \{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec\} \tag{68}$$

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2), \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\} \tag{69}$$

Q: How many elements in Ω? (Answer: 24)

Product of Probability Spaces

If we have a probability function P_1 defined in Ω_1 and a probability P_2 defined in Ω_2 we can naturally define a probability P in $\Omega_1 \times \Omega_2$ as:

$$P(\{\omega_1, \omega_2\}) = P_1(\omega_1) P_2(\omega_2) \tag{70}$$

E.g.

$$P(\{H, Jul\}) = P(H) \times P(Jul) = \frac{1}{2} \times \frac{1}{12} \tag{71}$$

This is how we model **independence**.

Infinite (Countable) Sample Spaces

E.g. number of coin flips until first heads appears:

What is the right probability space for this experiment?

$$\Omega = \{1, 2, 3, \ldots\} = \mathbb{N} \tag{72}$$

Sample space has to be infinite because no guarantee experiment will terminate in a finite number of steps!

If we assume that after k flips all k-tuples are equally likely what should the probability $P(k)$ be?

$$\{k = 1\} = \{H\} \Rightarrow P(1) = \frac{1}{2} \tag{73}$$

$$\{k = 2\} = \{T, H\} \Rightarrow P(2) = \frac{1}{4} \tag{74}$$

$$\vdots \tag{75}$$

$$\{k\} = \{\underbrace{T, \ldots, T}_{k-1 \text{ times}}, H\} \Rightarrow P(k) = \frac{1}{2^k} \tag{76}$$

Does this result in a proper probability function?

For infinite sample spaces need to change additivity rule to **countably additivity rule**:

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \quad \text{provided they are disjoint } (A_k \cap A_l = \emptyset \; \forall k, l \in \mathbb{N}, k \neq l) \tag{77}$$

Verification of $P(\Omega) = 1$:

$$P(\Omega) = P(\{1, 2, 3, \ldots\}) = P(1) + P(2) + P(3) + \cdots = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1 \tag{78}$$

(Used that for a geometric series: $1 + r + r^2 + \cdots = \sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ if $0 \leq r < 1$)

By extension of the rule for finite sample spaces, the probability defined above is a proper probability function.

Conditional Probability

Example: You roll a fair 6-faced die. Let A be the event that the outcome is an odd number, $A = \{1, 3, 5\}$. Let B be the event that the outcome is less than 4, $B = \{1, 2, 3\}$. What is the probability of A? What is the probability of A given B?

$$P(A) = \frac{|A|}{|S|} = \frac{|\{1, 3, 5\}|}{|S|} = \frac{3}{6} = \frac{1}{2} \tag{79}$$

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{2}{3} \tag{80}$$

We can write:

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|S|}}{\frac{|B|}{|S|}} = \frac{P(A \cap B)}{P(B)} \tag{81}$$

Definition

If A and B are events, and $P(B) > 0$ the conditional probability of A given B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{82}$$

- Fraction of the probability B also in the event A
- Tells us how to update probability in the presence of new information

Example:

- What is the probability that two cards drawn at random from a deck of playing cards will both be aces?

$$\frac{\binom{4}{2}}{\binom{52}{2}} = \frac{4 \times 3}{52 \times 51} \tag{83}$$

- What is the probability that two cards drawn at random from a deck of playing cards will both be aces if after dealing the first card it is an Ace?

$$\frac{3}{51} \tag{84}$$

Bayes Rule

From the definition we get the properties:

Multiplication rule:

$$P(A \cap B) = P(A|B)P(B) \tag{85}$$

Bayes rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (\text{if } P(A) > 0) \tag{86}$$

Example: There are approximately 2.6 physicians per 1,000 people in the US (from world public health data by country)

Probability of choosing a physician if randomly choose a US inhabitant = $\frac{2.6}{1000} = 0.0026$

- $A = \{\text{Being a Physician in the US}\} = \{\text{Physician}\}$
- $B = \{\text{Being a Woman in the US}\} = \{\text{Woman}\}$
- $P(\text{Physician}) = \frac{2.6}{1000}$
- $P(\text{Woman}) = 0.508$ (From US Census)
- $P(\text{Woman}|\text{Physician}) = 0.36$ (From Labor statistics)

$$P(\text{Physician}|\text{Woman}) = \frac{P(\text{Woman}|\text{Physician})P(\text{Physician})}{P(\text{Woman})} = \frac{2.6}{1000} \times \frac{0.36}{0.508} = \frac{1.6}{1000} \tag{87}$$

Some Special Cases

- A and B disjoint $\Rightarrow P(A|B) = P(B|A) = 0$
- $B \subset A \Rightarrow P(A|B) = 1$
- $A \subset B \Rightarrow P(A|B) = \frac{P(A)}{P(B)}$

Conditional Probability is a Probability Function

For fixed C with $P(C) > 0$ the conditional probability $P_C(\cdot) = P(\cdot|C)$ is a probability function:

1. $P_C(A) = P(A|C) \geq 0$
2. $P_C(\Omega) = P(\Omega|C) = 1$
3. $P_C(A \cup B) = P_C(A) + P_C(B)$ if $A \cap B = \emptyset$

Law of Total Probability

If the sample space can be partitioned as $\Omega = \bigcup_{i=1}^{n} A_i$, with $A_1, \ldots, A_n$ disjoint, then:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{88}$$

(holds even for a countable partition)

In particular, for any event A, the sample space can be partitioned as $\Omega = A \cup A^c$:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) \tag{89}$$

Law of total probability example

The probability of infection from a certain virus upon exposure is 10% for children age < 13, 5% for ages 13-60, and 15% for ages 60+. What is the probability that a random individual is infected upon exposure in a population where $P(\text{Age} < 13) = 0.2$, $P(13 \leq \text{Age} \leq 60) = 0.6$, $P(\text{Age} > 60) = 0.2$?

Let I denote the event of infection:

$$P(I) = P(I|\text{Age} < 13)P(\text{Age} < 13) + P(I|13 \leq \text{Age} \leq 60)P(13 \leq \text{Age} \leq 60) + P(I|\text{Age} > 60)P(\text{Age} > 60) \quad (90)$$

$$= 0.1 \times 0.2 + 0.05 \times 0.6 + 0.15 \times 0.2 = 0.08 \quad (91)$$

Medical Testing Example

A diagnostic test has 99% sensitivity and 98% specificity.

If the population prevalence of the disease is 3%, what is the probability that an individual who tests positive is affected with the disease?

- $\text{Sensitivity} = P(\text{Test+}|\text{Affected}) = 0.99$
- $\text{Specificity} = P(\text{Test-}|\text{Not Affected}) = 1 - P(\text{Test+}|\text{Not Affected})$
- $(P(\text{Test+}|\text{Not Affected}) = 1 - \text{Specificity} = 0.02)$
- $\text{Prevalence} = P(\text{Affected}) = 0.03$

- $P(\text{Affected}|\text{Test+}) = ?$
- $P(\text{Test+}) = P(\text{Test+}|\text{Affected})P(\text{Affected}) + P(\text{Test+}|\text{Not Affected})P(\text{Not Affected})$
- $= 0.99 \times 0.03 + 0.02 \times 0.97 = 0.0297 + 0.0194 = 0.0491$
- $P(\text{Affected}|\text{Test+}) = \frac{P(\text{Test+}|\text{Affected})P(\text{Affected})}{P(\text{Test+})} = \frac{0.0297}{0.0491} = 0.605$

Monty Hall Problem

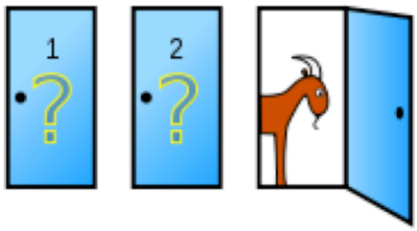
You're on a game show, and you're given the choice of three doors:

- Behind one door is a car
- Behind the others, goats

You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.

He then says to you, "Do you want to switch to door No. 2 or keep prize behind door No. 1?"

Should you switch? `Answer: Yes! Switching gives you 2/3 probability of winning`



Independence

Two events A and B are **independent** iff (if and only if):

$$P(A \cap B) = P(A)P(B) \quad (92)$$

E.g. fair coin tossed twice:

$$P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4} \tag{93}$$

$$P(\{H \text{ in first toss}\}) = P(HH) + P(HT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \tag{94}$$

$$P(\{H \text{ in second toss}\}) = P(HH) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \tag{95}$$

$$P(\{H \text{ in first toss}\} \cap \{H \text{ in second toss}\}) = P(\{H \text{ in first toss}\})P(\{H \text{ in second toss}\}) \tag{96}$$

The events $\{H \text{ in first toss}\}$ and $\{H \text{ in second toss}\}$ are independent.

(in fact any first toss outcome is independent of any second toss outcome)

Independence between A and B is equivalent to:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. A and B^c are independent (or A^c and B are independent, or A^c and B^c are independent)

Independence of Multiple Events

$A_1, \ldots, A_n$ are independent iff:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n) \tag{97}$$

And also the equation above holds replacing any number of the A_is by their complements (2^n equations!)

Pairwise independence does not imply independence!!

Example: Two tosses of a fair coin

- $A = \{H \text{ in first toss}\}$
- $B = \{H \text{ in second toss}\}$
- $C = \{\text{two tosses are equal}\}$

`These are pairwise independent but not mutually independent`

Lecture 3 - Discrete Random Variables

Review from Last Class

- Countable sample spaces (e.g. flipping a coin until first head, $\Omega = \{1, 2, \ldots\} = \mathbb{N}$)
- Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Law of total probability: $P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)$ for a (disjoint) partition $\Omega = A_1 \cup \cdots \cup A_n$
- Bayes theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Independence: $P(A \cap B) = P(A)P(B)$ (or more intuitively $P(A|B) = P(A)$)
- For $n > 2$ events independence requires satisfying 2^n equations

Q: Is disjoint the same as independent? No! Disjoint events cannot both occur, while independent events don't affect each other's probabilities.

Discrete Random Variables

A discrete random variable is a function $X : \Omega \to \mathbb{R}$ that takes a finite or countable number of values $x_1, x_2, \ldots$

E.g. The number in the upper face of the rolled die, the sum of two dice

Notation:

- $\{\omega : X(\omega) = x\} = \{X = x\}$
- $P(\{X = x\}) = P(X = x)$

Probability Mass Function (pmf)

The pmf of a discrete random variable taking values $x_1, x_2, \ldots$ is the function:

$$p : \mathbb{R} \to [0, 1] \tag{98}$$

$$p(x) = P(X = x) \quad (\text{Sometimes also denoted } p_X(x) \text{ or } f_X(x)) \tag{99}$$

If X takes on the values $x_1, x_2, \ldots$ then:

- $p(x_i) > 0$, and $p(x_1) + p(x_2) + \cdots = 1$
- $p(x) = 0$ for all other x

E.g. Fair coin flip:

$$X = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases} \tag{100}$$

$$p(1) = \frac{1}{2}, \quad p(0) = \frac{1}{2}, \quad p(0.5) = p(\pi) = 0 \tag{101}$$

Cumulative Distribution Function (cdf)

The cdf of a discrete random variable taking values $x_1, x_2, \ldots$ is the function:

$$F : \mathbb{R} \to [0, 1] \quad (\text{sometimes denoted } F_X) \tag{102}$$

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R} \tag{103}$$

If X takes on the values $x_1, x_2, \ldots$ then:

$$F(x) = \sum_{x_i \leq x} p(x_i) \tag{104}$$

E.g. Fair coin flip:

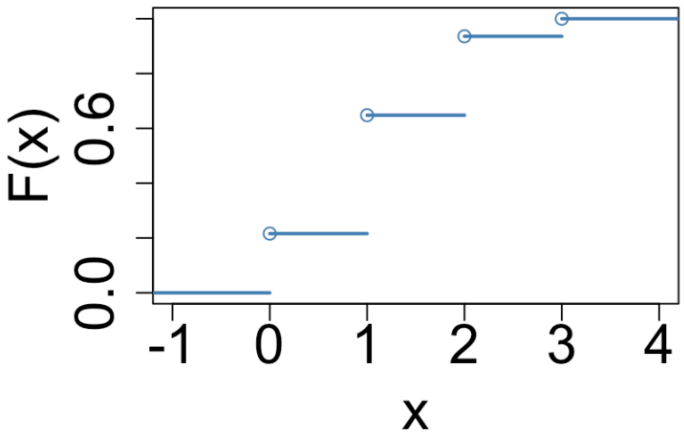
$$X = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases} \tag{105}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases} \tag{106}$$

Both the pmf and the cdf completely characterize all the **probabilistic** information about a random variable (two random variables can have the same pmf and cdf and be different).

Properties of the Cumulative Distribution Function

- $F(x)$ is increasing: $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$
- $0 \leq F(x) \leq 1$, $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$
- F is right continuous: $\lim_{\epsilon \downarrow 0} F(x + \epsilon) = F(x)$



Bernoulli Distribution

A random variable has a **Bernoulli distribution**

$$f_X(1) = p, \quad (0 \leq p \leq 1) \tag{107}$$

$$f_X(0) = 1 - p \tag{108}$$

$$X \sim Bern(p) \quad \text{or} \quad X \sim Bernoulli(p) \tag{109}$$

- Models experiments with only two possible outcomes
- E.g. coin toss (H vs. T), die comes up six (yes vs. no)
- A random variable with a Bernoulli distribution is called a **Bernoulli trial**

Binomial Distribution

n independent Bernoulli trials (e.g. flipping a coin n times): $X_1, \ldots, X_n \sim Bernoulli(p)$

$X = X_1 + \cdots + X_n$ counts the number of successes in n trials

$$X \sim Bin(n, p) \quad \text{or} \quad X \sim Binomial(n, p) \tag{110}$$

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n \tag{111}$$

- Models the number of successes in n trials
- For $n = 1$ the Binomial distribution is the Bernoulli distribution

Example: Side Effects

Suppose it is known that 5% of adults who take a certain medication experience negative side effects. What is the probability that more than k patients in a random sample of 100 will experience negative side effects?

$$P(X > 1 \text{ patients experience side effects}) = ? \tag{112}$$

$$P(X > 5 \text{ patients experience side effects}) = ? \tag{113}$$

$$P(X > 15 \text{ patients experience side effects}) = ? \tag{114}$$

Binomial Distribution in R

pmf, cdf, and Random generation of a binomial random variable

```r
# pmf
dbinom(3, size=10, prob=0.3)
## [1] 0.2668279

# cdf
pbinom(3, size=10, prob=0.3)
## [1] 0.6496107

# Random generation
rbinom(n=1, size=10, prob=0.3)
## [1] 2

rbinom(n=3, size=10, prob=0.3)
## [1] 0 1 3
```

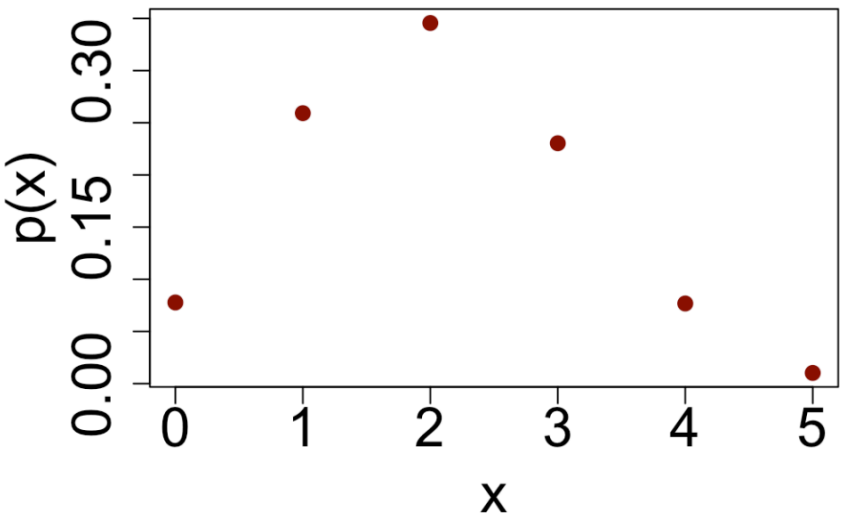
Side Effects Example (continued)

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F_X(1) \tag{115}$$

```r
1 - pbinom(1, size = 100, prob = 0.05)
## [1] 0.9629188

pbinom(1, size = 100, prob = 0.05, lower.tail = FALSE)
## [1] 0.9629188

pbinom(5, size = 100, prob = 0.05, lower.tail = FALSE)
## [1] 0.3840009

pbinom(15, size = 100, prob = 0.05, lower.tail = FALSE)
## [1] 3.705408e-05
```

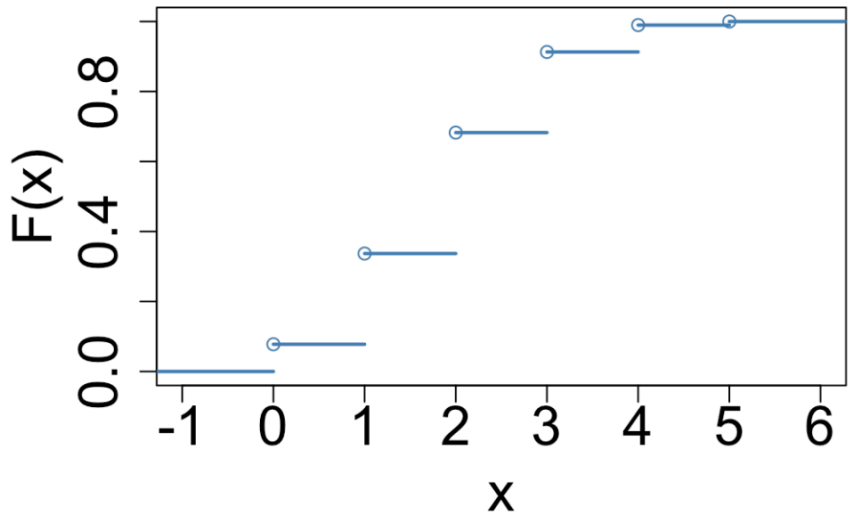
Binomial distribution pmf

```r
par(mar=c(6,8,5,1))

plot(0:5, dbinom(0:5, size=5, prob=0.4), col='red4', type='p', pch=16, cex=1.3, xlab='x', ylab='p(x)',
cex.lab=2, cex.axis=2)
```



Binomial distribution cdf

```
par(mar=c(6,8,5,1))

plot(stepfun(0:5, c(0, pbinom(0:5, size=5, prob=0.4))), pch = 1, lwd=2, col='steelblue', xlab='x', ylab='F(x)',
cex.lab=2, cex.axis=2, main='', verticals = F)
```



Simulating a binomial 3 different ways

Generating n Bernoulli trials

```
Bernoulli_trials_1 = sample(0:1, 10, replace = TRUE, prob=c(0.3, 0.7))
Bernoulli_trials_1

## [1] 1 1 0 1 1 0 1 0 1 0

sum(Bernoulli_trials_1)

## [1] 6
```

Generating n Bernoulli trials using rbinom()

```
Bernoulli_trials_2 = rbinom(10, size = 1, prob=0.3)
Bernoulli_trials_2

## [1] 0 1 0 0 0 0 0 0 0 0

sum(Bernoulli_trials_2)

## [1] 1
```

Directly sampling from the binomial

```
rbinom(1, size=10, prob=0.3)

## [1] 5
```

Geometric Distribution

A discrete random variable X has a **geometric distribution** with parameter p, where $0 < p \leq 1$, if its probability mass function is given by:

$$p_X(k) = P(X = k) = (1-p)^{k-1}p \quad k = 1, 2, \ldots \tag{116}$$

$$X \sim Geo(p) \quad \text{or} \quad X \sim Geometric(p) \tag{117}$$

Models the (discrete) waiting time until an event happens. E.g. number of trials till first heads.

Example: Concert Ticket

You and a friend want to go to a concert, but there's only one ticket left. The salesperson decides to toss a coin until heads appears. In each toss heads appears with probability p, where $0 < p < 1$, independent of each of the previous tosses. If the number of tosses needed is odd, your friend is allowed to buy the ticket; otherwise you can buy it. Would you agree to this arrangement?

Geometric Memoryless Property

$$P(X > n + k | X > k) = P(X > n) \tag{118}$$

The probability it'll take n additional trials if the first k are failures is the same as the probability it'll take n trials at the beginning of the experiment.

Geometric Distribution in R

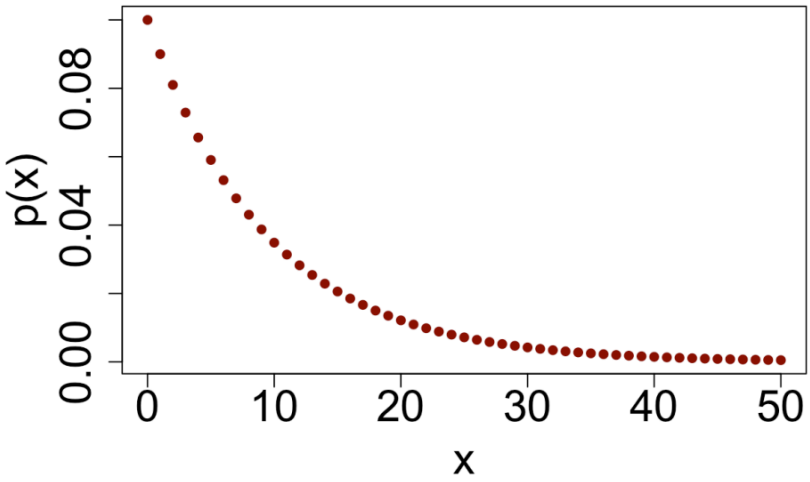
pmf, cdf, and Random generation of a binomial random variable

Warning: The definition of the geometric in R is the number of **failures before the first success**, i.e. $X - 1$.

```
dgeom(x=5, prob = 0.1)
## [1] 0.059049

pgeom(10, prob= 0.1)
## [1] 0.6861894

rgeom(n=1, prob= 0.1)
## [1] 0

rgeom(n=3, prob= 0.1)
## [1] 4 3 4
```

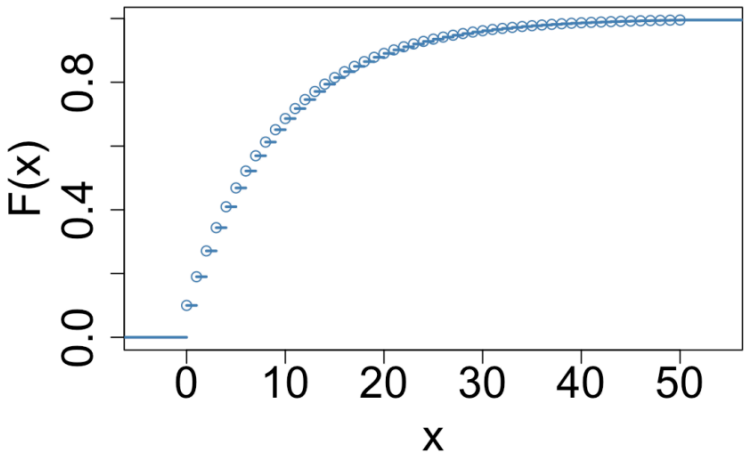
Geometric distribution pmf

```
par(mar=c(6,8,5,1))
plot(0:50, dgeom(0:50, prob=0.1), col='red4', type='p', pch=16, cex=1, xlab='x', ylab='p(x)', cex.lab=2,
cex.axis=2
```



Geometric distribution cdf

```
par(mar=c(6,8,5,1))
plot(stepfun(0:50, c(0, pgeom(0:50, prob=0.1))), pch = 1, lwd=2, col='steelblue', xlab='x', ylab='F(x)',
cex.lab=2, cex.axis=2, m
```



Lecture 4 - Continuous Random Variables

Review from Last Class

- Discrete random variables:
  - Take finite or countable values
  - Completely characterized by their pmf or cdf
- **Bernoulli:** $X = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases}$
- **Binomial:** number of successes in n independent Bernoulli trials with identical probability of success p
  - $X \sim Binomial(n, p), x \in \{0, 1, \ldots, n\}$

- **Geometric distribution:** number of trials until first success in repeated Bernoulli experiments with identical probability of success p
  - $X \sim Geometric(p)$, $x \in \{1, 2, \ldots\}$

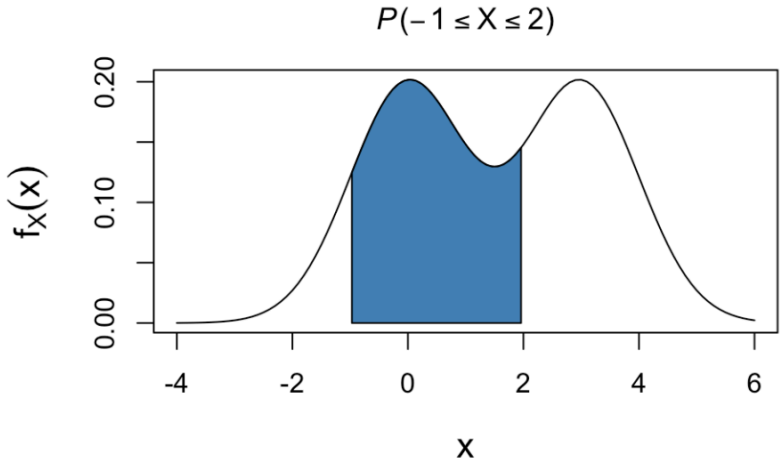
Continuous Random Variables

A continuous random variable is a function $X : \Omega \to \mathbb{R}$ that takes on uncountable infinite values and such that for $a \leq b$:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx \tag{119}$$

for some function $f_X(x) \geq 0,\ \forall x \in \mathbb{R}$ and $\int_{-\infty}^{+\infty} f_X(t)\, dt = 1$

$f_X : \mathbb{R} \to \mathbb{R}$ is called the **probability density function** (or density function) of the random variable X.



Probability Density Function

- More generally, if $S \subset \mathbb{R}$: $P(X \in S) = \int_S f_X(x)\, dx$

Properties of the pdf:

- $P(-\infty < X < +\infty) = \int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt$
- $f_X(x) = \frac{d}{dx} F_X(x) = F'_X(x)$ (Fundamental theorem of calculus)
- $P(x - \frac{\epsilon}{2} < X < x + \frac{\epsilon}{2}) = \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} f_X(t)\, dt \approx \epsilon f_X(x)$
- $f_X(x) = \lim_{\epsilon \to 0} \frac{P(x-\frac{\epsilon}{2} < X < x+\frac{\epsilon}{2})}{\epsilon}$, i.e. represents the density of probability 'mass' at point x
- For a continuous random variable X, $P(X = x) = 0$

Fundamental Theorem of Calculus

Part I: Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined, for all x in $[a, b]$, by:

$$F(x) = \int_a^x f(t)\, dt \tag{120}$$

Then F is uniformly continuous on $[a, b]$ and differentiable on the open interval (a, b), and:

$$F'(x) = f(x) \tag{121}$$

for all x in (a, b).

Part II: Let f be a real-valued function on a closed interval $[a, b]$ and F antiderivative of f in (a, b):

$$F'(x) = f(x) \tag{122}$$

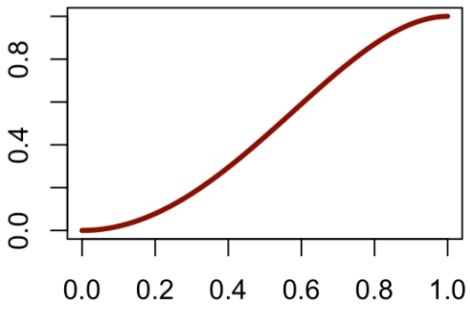
If f is (Riemann) integrable on $[a, b]$ then:

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{123}$$

Example

Let a continuous random variable X be given that takes values in $[0, 1]$, and whose distribution function is given by:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 2x^2 - x^4 & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases} \tag{124}$$

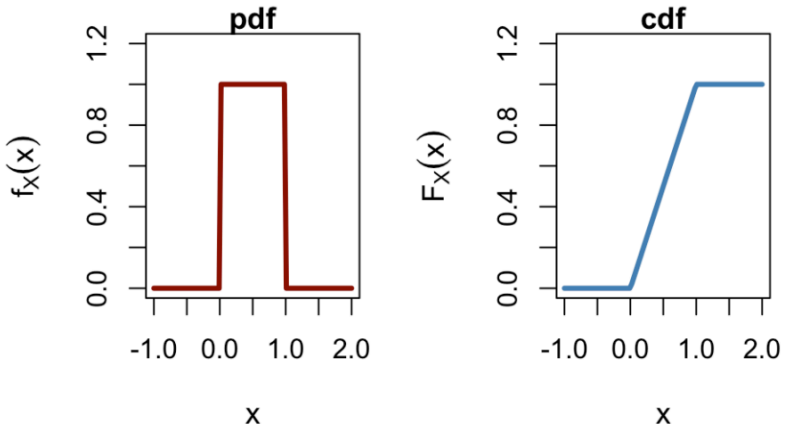


1. Compute $P(\frac{1}{4} \leq X \leq \frac{3}{4})$
2. What is the probability density function of X?
3. Compute $P(\frac{1}{4} \leq X \leq \frac{1}{2} \text{ or } X > \frac{3}{4})$

Uniform Distribution, $X \sim U[a, b]$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{125}$$

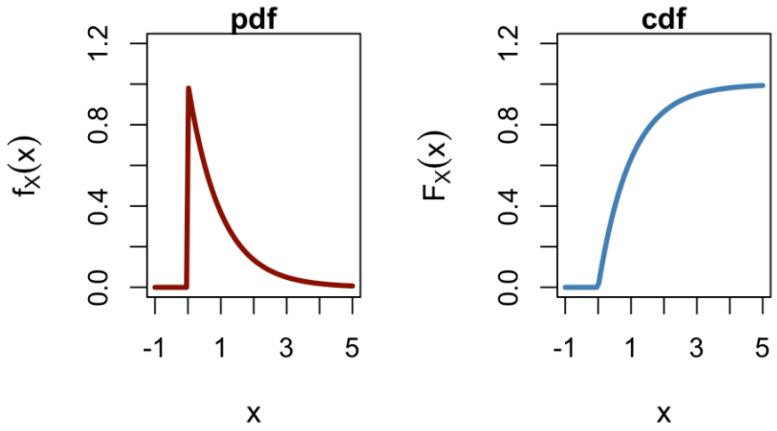
$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b \end{cases} \tag{126}$$



Exponential Distribution, $X \sim Exp(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{127}$$

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{128}$$

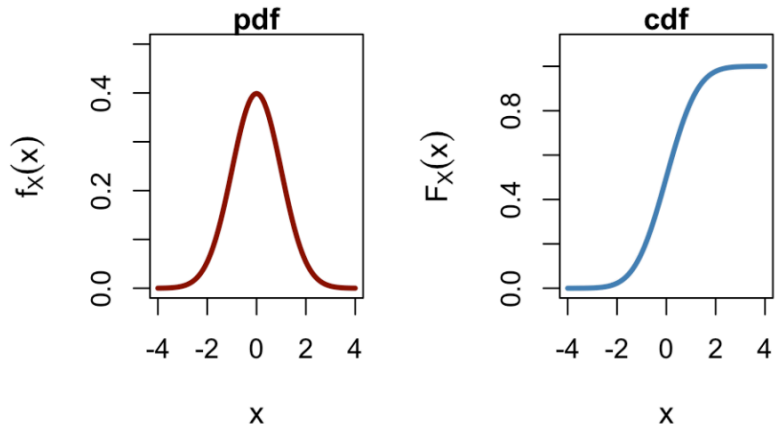


Normal Distribution, $X \sim N(\mu, \sigma^2)$

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{129}$$

$$F(x) = \Phi(x) \tag{130}$$

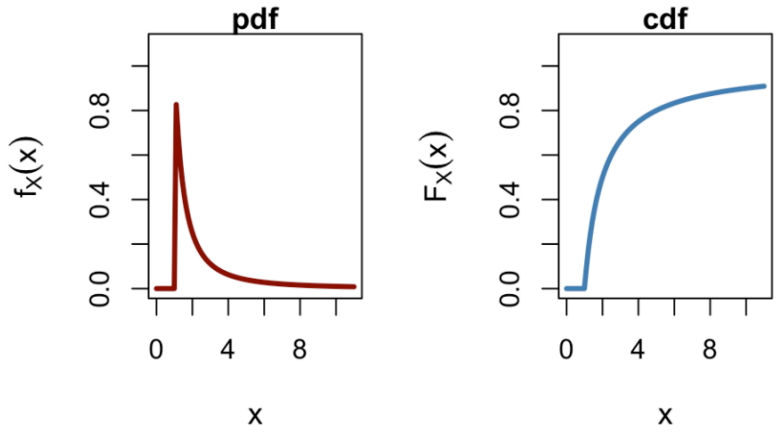
There is no analytical formula for the cdf but it can be numerically computed.



Pareto Distribution, $X \sim Pareto(x_m, \alpha)$

$$f(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & \text{if } x \geq x_m \\ 0 & \text{if } x < x_m \end{cases} \tag{131}$$

$$F(x) = \begin{cases} 0 & \text{if } x < x_m \\ 1 - \left(\frac{x_m}{x}\right)^\alpha & \text{if } x \geq x_m \end{cases} \tag{132}$$

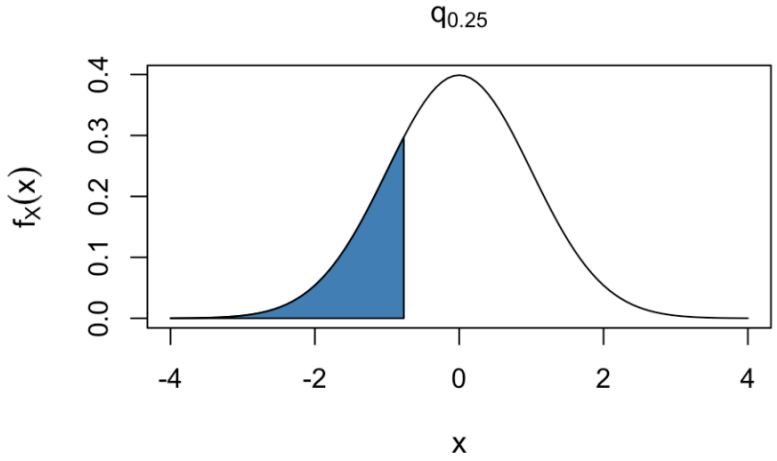


Quantiles

Let X be a continuous random variable and let p be a number between 0 and 1. The p^{th} **quantile** or $100p^{th}$ **percentile** of the distribution of X is the smallest number q_p such that:

$$F(q_p) = P(X \leq q_p) = p \tag{133}$$

The **median** of a distribution is its 50^{th} percentile.



Example 1: Median of Exponential, $X \sim Exp(\lambda)$

$$\frac{1}{2} = F(q_{0.5}) = P(X \leq q_{0.5}) = 1 - e^{-\lambda q_{0.5}} \tag{134}$$

$$q_{0.5} = -\frac{1}{\lambda}\log\left(\frac{1}{2}\right) \tag{135}$$

Example 2: Median of Uniform in $[1, 2] \cup [3, 4]$

$$q_{0.5} = 2 \tag{136}$$

Uniform Distribution in R

pmf, cdf, random generation, and quantile of uniform random variable

```r
dunif(3, min=1, max=5)
## [1] 0.25

punif(3, min=1, max=5)
## [1] 0.5

runif(n=3) #default is uniform[0,1]
## [1] 0.007446826 0.944775617 0.292623820

qunif(0.5, min=1, max=3)
## [1] 2
```

Exponential Distribution in R

pmf, cdf, random generation, and quantile of an exponential random variable

```r
dexp(3, rate = 0.5)
## [1] 0.1115651

pexp(3, rate = 2)
## [1] 0.9975212

rexp(n=5) # default rate is 1
## [1] 4.60679804 0.32611096 0.01889765 1.49367378 0.70605987

c(qexp(p=0.5), -log(1/2))
## [1] 0.6931472 0.6931472
```

Normal Distribution in R

pmf, cdf, random generation, and quantile of a normal random variable

```r
dnorm(x=3, mean = 1, sd=2)
## [1] 0.1209854

dnorm(x=-1, mean = -2, sd=0.5)
## [1] 0.1079819

rnorm(n=5)
## [1] -0.4843401 -1.0643711 1.5258292 2.2244114 0.7464652

qnorm(0.5)
## [1] 0
```

Pareto Distribution in R

pmf, cdf, random generation, and quantile of a Pareto random variable with location x_m and shape α

```
library(EnvStats)

dpareto(x=3, location = 1, shape=2)
## [1] 0.07407407

ppareto(q=3, location = 1, shape=2)
## [1] 0.8888889

rpareto(n=5, location = 1, shape=2)
## [1] 1.471863 1.356510 1.347556 1.011351 2.216670

qpareto(p=0.5, location = 1, shape=2)
## [1] 1.414214
```

Mixtures of Distributions

Example 1: Discrete Mixture

To get to your destination you take a taxi if there is one waiting (probability 1/3) at the stand when you arrive or walk if there is no taxi waiting. A taxi takes you exactly 5 minutes. Walking to your destination takes you exactly 35 minutes. What is the cdf of the time to your destination T?

$$P(T=t) = P(T=t|\text{Taxi})P(\text{Taxi}) + P(T=t|\text{No taxi})P(\text{No taxi}) = \begin{cases} \frac{1}{3} & \text{if } t=5 \\ \frac{2}{3} & \text{if } t=35 \end{cases} \tag{137}$$

$$F_T(t) = \begin{cases} 0 & \text{if } t<5 \\ \frac{1}{3} & \text{if } 5 \leq t < 35 \\ 1 & \text{if } t \geq 35 \end{cases} \tag{138}$$

Example 2: Continuous Mixture

To get to your destination you take a taxi if one is waiting (probability 1/3) at the stand when you arrive or walk if there is no taxi waiting. Walking to your destination takes you an amount of time distributed as $Exp(\lambda_1)$ with $\lambda_1 = 1/35$. A taxi takes you an amount of time distributed as $Exp(\lambda_2)$ with $\lambda_2 = 1/5$. What is the cdf of the time to get to your destination, T?

$$F_T(t) = P(T \leq t) = P(T \leq t|\text{Taxi})P(\text{Taxi}) + P(T \leq t|\text{No taxi})P(\text{No taxi}) \tag{139}$$

$$= \frac{1}{3}(1-e^{-t/5}) + \frac{2}{3}(1-e^{-t/35}) \tag{140}$$

$$f_T(t) = \frac{d}{dt}F_T(t) = F_T'(t) = \frac{1}{3}\left(\frac{1}{5}e^{-t/5}\right) + \frac{2}{3}\left(\frac{1}{35}e^{-t/35}\right) \tag{141}$$

Example 3: Mixed Distribution

To get to your destination you take a taxi if one is waiting (probability 1/3) when you arrive or walk if there is no taxi. Walking to your destination takes you exactly 35 minutes. A taxi takes an amount of time distributed as $Exp(\lambda_2)$ with $\lambda_2 = 1/5$. What is the cdf of the time to your destination T?

$$F_T(t) = P(T \leq t) = P(T \leq t|\text{Taxi})P(\text{Taxi}) + P(T \leq t|\text{No taxi})P(\text{No taxi}) \tag{142}$$

$$P(T \leq t|\text{No taxi}) = \begin{cases} 0 & \text{if } t<35 \\ 1 & \text{if } t \geq 35 \end{cases} \tag{143}$$

$$P(T \leq t|\text{Taxi}) = \begin{cases} 0 & \text{if } t<0 \\ 1-e^{-t/5} & \text{if } t \geq 0 \end{cases} \tag{144}$$

$$F_T(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{2}{3}(1 - e^{-t/5}) & \text{if } 0 \leq t < 35 \\ \frac{1}{3} + \frac{2}{3}(1 - e^{-t/5}) & \text{if } t \geq 35 \end{cases} \tag{145}$$

- How about the probability density function? **There isn't one!**
- This is an example of a **MIXED** random variable
- MIXED random variables are mixtures of discrete and continuous random variables

Discrete, Continuous, and Mixed Random Variables

- **Discrete random variables** have probability mass function but do not have probability density function
- **Continuous random variables** have probability density function but do not have probability mass function
- **Mixed random variables** have **neither** probability mass function **nor** probability density function
- **All types of random variables** (discrete, continuous and mixed distributions) **have cumulative distribution function!!**

Week 5 – Expectation, variance, and transformations of RVs

Last class

- **Continuous random variables:**
  - Take infinite uncountable values
  - Have a probability density function (pdf)
  - Continuous RV are completely characterized by their pdf or cdf
- **Uniform:** $U[a, b]$
- **Exponential:** $Exp[\lambda]$
- **Normal:** $N(\mu, \sigma)$
- **Pareto:** $Pareto(x_m, \alpha)$
- **Mixtures of distributions:**
  - Discrete + Discrete = Discrete
  - Continuous + Continuous = Continuous
  - Discrete + Continuous = Neither discrete nor continuous

For a continuous random variable, it follows from the definition of pdf and the fundamental theorem of calculus that for $a \leq b$:

$$P(a < X \leq b) = \int_a^b f_X(t)dt = F_X(x)|_a^b = F(b) - F(a) \tag{146}$$

But $P(a < X \leq b) = F(b) - F(a)$ is true for any random variable (discrete, continuous or mixed):
Let $A = \{X \leq a\}, B = \{X \leq b\}$. Clearly, $A \subset B$ and $B - A = \{a < X \leq b\}$.

$$P(a < X \leq b) = P(B - A) = P(B) - P(A) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{147}$$

For a continuous RV (but not for a discrete or mixed) this also equals $P(a \leq X \leq b), P(a < X < b), \ldots$

Expectation for discrete RVs

The expected value is a weighted average of the values a random variable takes, weighted by the probability of taking those values. It's the center of 'gravity' where the distribution 'balances'.

For a **Discrete** random variable X, $f_X(x)$ the probability mass function of X, and $\{x_1, x_2, \ldots\}$ is the support of X.

$$E[X] = \sum_{x_i \in \text{supp}(X)} x_i f_X(x_i) \tag{148}$$

The support of a discrete random variable X is the set of points that X takes with non-zero probability:

$$\text{supp}(X) = \{x_i : P(X = x_i) > 0\} \tag{149}$$

Expectation discrete examples: Bernoulli

Bernoulli trial, $X \sim \text{Bernoulli}(p)$

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases} \tag{150}$$

$f_X(0) = p$

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

$$E[X] = 0 \times (1-p) + 1 \times p = p \tag{151}$$

Expectation discrete examples: Binomial

$X \sim \text{Binomial}(n, p)$; models the number of successes in n trials

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n \tag{152}$$

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \\ &= \sum_{k=1}^{n} k \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} = \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = \\ &= np(p + (1-p))^{n-1} = np \end{aligned} \tag{153}$$

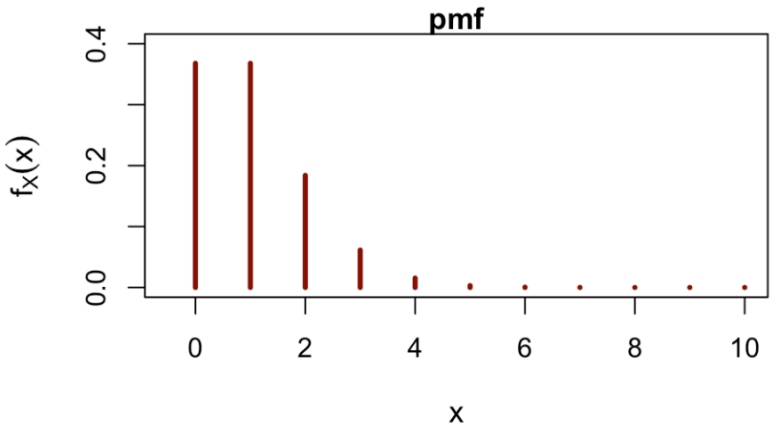
Expectation discrete examples: Poisson

A discrete random variable X is said to have a $\text{Poisson}(\lambda)$ distribution with parameter $\lambda > 0$ if

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \tag{154}$$

for $k = 0, 1, 2, \ldots$

Used to model number of events happening in a period of time e.g. number of mutations per unit length in a DNA strand, number of new patients (incidence rates), number of phone calls/particles arriving in a system, etc.



Expectation examples: Poisson (contd)

$$E[X] = \sum_{k=0}^{\infty} k f_X(k) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \\ = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \\ = \lambda e^{-\lambda} e^{\lambda} = \lambda \tag{155}$$

Expectation for continuous RVs

For a **Continuous** random variable X, with $f_X(x)$ the probability density function of X, the expectation is defined as:

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx \tag{156}$$

NOTE: Expectation may not exist E.g. Cauchy distribution

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < +\infty \tag{157}$$

Expectation examples: continuous with finite support

$X \sim F(x)$

CDF:

$$F(x) = \begin{cases} 0 & if\ x < 0 \\ 2x^2 - x^4 & if\ 0 \le x \le 1 \\ 1 & if\ x \ge 1 \end{cases} \tag{158}$$

PDF:

$$f(x) = F'(x) = \begin{cases} 4x - 4x^3 & if\ 0 \le x \le 1 \\ 0 & otherwise \end{cases} \tag{159}$$

Expectation Calculation:

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 x(4x - 4x^3) dx = 4\left(\frac{x^3}{3} - \frac{x^5}{5}\right)\Big|_0^1 = \frac{8}{15} \tag{160}$$

Expectation examples: continuous with infinite support

$X \sim \text{Exp}[\lambda]$, $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx = \int_{0}^{+\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \int_{0}^{+\infty} t e^{-t} dt \tag{161}$$

(Using $t = \lambda x, dt = \lambda dx$)

$$E[X] = \frac{1}{\lambda} \int_{0}^{+\infty} t e^{-t} dt = \frac{1}{\lambda} \left(-t e^{-t} \Big|_{0}^{+\infty} - \int_{0}^{+\infty} e^{-t} dt \right) = \frac{1}{\lambda} (-e^{-t}) \Big|_{0}^{+\infty} = \frac{1}{\lambda} \tag{162}$$

(integrating by parts)

Expectation for a mixed random variable

For a mixed random variable X with cdf $F(x) = pF_1(x) + (1-p)F_2(x)$ with F_1 continuous and F_2 discrete:

$$E[X] = p \int_{-\infty}^{+\infty} x f_1(x) dx + (1-p) \sum_{x_i} x_i f_2(x_i) \tag{163}$$

where $f_1(x)$ is the density for the continuous component, $f_1(x) = F_1'(x)$, and $f_2(x)$ is the mass function for the discrete component.

Probabilities as expectations

For a set $A \subset \Omega$ the random variable:

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases} \tag{164}$$

is called the indicator function of the set A.

What is the distribution of I_A?

$I_A \sim \text{Bernoulli}(p), p = P(A) \Rightarrow E[I_A] = P(A)$

Allows us to work with random variables (indicator functions) instead of sets and expectations instead of probabilities

Exercise: If $A, B \subset \Omega$, what are $I_A I_B, \min(I_A, I_B), \max(I_A, I_B)$?

Transformations of random variables

Example: $X \sim F(x)$

$$F(x) = \begin{cases} 0 & if\ x < 0 \\ 2x^2 - x^4 & if\ 0 \leq x \leq 1 \\ 1 & if\ x \geq 1 \end{cases} \tag{165}$$

What is the cdf of $Y = -\sqrt{X+1}$?

First, $0 \leq X \leq 1 \iff -\sqrt{2} \leq -\sqrt{X+1} \leq -1$

Transformations of random variables

For $-\sqrt{2} \leq y \leq -1$

$$
\begin{aligned}
F_Y(y) = P(Y \leq y) = \\
= P(-\sqrt{X+1} \leq y) = \\
= P(-y \leq \sqrt{X+1}) = \\
= P(y^2 - 1 \leq X) = \\
= 1 - P(X < y^2 - 1) = 1 - P(X \leq y^2 - 1) = \\
= 1 - F_X(y^2 - 1) = 1 - 2(y^2-1)^2 + (y^2-1)^4
\end{aligned} \tag{166}
$$

Final CDF:

$$
F_Y(y) = \begin{cases} 0 & if\ y < -1 \\ 1 - 2(y^2-1)^2 + (y^2-1)^4 & if\ -\sqrt{2} \leq y \leq -1 \\ 1 & if\ y > -\sqrt{2} \end{cases} \tag{167}
$$

Change-of-variable formula for the expectation

Continuous X:

$$
E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx \tag{168}
$$

Discrete X:

$$
E[g(X)] = \sum_{x_i} g(x_i) f_X(x_i) \tag{169}
$$

Allows us to compute the expectation of $Y = g(X)$ without deriving the pdf or pmf of $g(X)$!!

Change-of-variable formula example

In the previous example computing $E[Y] = \int_{-\sqrt{2}}^{-1} y f_Y(y) dy$ seems to require knowing/deriving $f_Y(y) = F'_Y(y)$ and integrating using the pdf of Y (a big mess)

The change-of-variable formula allows us to use a shortcut:

$$
E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx \tag{170}
$$

Only require us to integrate using the pdf of X!

$$
E[-\sqrt{X+1}] = \int_0^1 -(\sqrt{x+1})(2x^2 - x^4) dx \tag{171}
$$

Using Mathematica: `Integrate[-Sqrt[(1+x)](2x^2-x^4), {x, 0, 1}]`

$$
E[-\sqrt{X+1}] = -\frac{4}{693}(103\sqrt{2} - 40) \tag{172}
$$

Variance

$$
Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \tag{173}
$$

$Var[X]$ is the average squared deviation from the mean. Measure of dispersion/concentration.

Continuous X:

$$
E[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x) dx \tag{174}
$$

Discrete X:

$$E[X^2] = \sum_{x_i} x_i^2 f_X(x_i) \tag{175}$$

Week 6 – Random vectors and independence

Last class

Expectation

- **Continuous:**
  $E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$
- **Discrete:**
  $E[X] = \sum_{x_i} x_i f_X(x_i)$
- **Mixed:**
  $E[X] = w_d \sum_{x_i} x_i f_X(x_i) + w_c \int_{-\infty}^{+\infty} x f_X(x) dx; \quad w_d + w_c = 1$

Variance

- $Var[X] = E[(X - E[X])^2] = E[X^2] - E^2[X]$

Change of variable formula/LOTUS

- **Continuous:**
  $E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$
- **Discrete:**
  $E[g(X)] = \sum_{x_i} g(x_i) P_X(x_i)$

Random vectors

We are typically interested in not one, but multiple related random variables defined on the same space.

- X, Y random variables or (X, Y) a random vector in $\mathbb{R}^2$
  - E.g. weight and height of a randomly selected person
- Geometrically, (X, Y) represents a random point in $\mathbb{R}^2$
- More generally, a random vector in $\mathbb{R}^n$, $\mathbf{X} = (X_1, \ldots, X_n)$
  - E.g. expression levels of n genes for an individuals
- We can have discrete, continuous, and mixed random vectors

Discrete random vectors

A random vector (X, Y) is discrete if it takes a finite or countable number of values
$R_{X,Y} = \{(x_1, y_1), (x_2, y_2), \ldots\}$

Joint probability mass function:

- $P_{X,Y}(x, y) = P(X = x, Y = y) = P(X = x \cap Y = y)$
- $P(X \in A \cap R_X, Y \in B \cap R_Y) = \sum_{x_i \in A} \sum_{y_j \in B} P_{X,Y}(x_i, y_j)$

In general, if $C \subset \mathbb{R}^2$,

- $P((X, Y) \in C) = \sum_{(x_i, y_j) \in C \cap R_{X,Y}} P_{X,Y}(x_i, y_j)$

X and Y are discrete random variables $\iff$ (X, Y) is a discrete random vector

Continuous random vectors

A random vector is continuous if it has a joint probability density function:

- $f_{X,Y}(x, y) \geq 0$
- $\iint_{\mathbb{R}^2} f_{X,Y}(x, y) dx\, dy = 1$
- $P((X, Y) \in C) = \iint_C f_{X,Y}(x, y) dx\, dy$

(X, Y) continuous as a random vector $\Rightarrow$ X and Y continuous as individual random variables

Converse is not true: X and Y are continuous $\not\Rightarrow$ (X, Y) continuous as a random vector (it may not have a density)

Random vectors

The cumulative distribution function (cdf) is defined for both discrete and continuous (and mixture) random vectors:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \begin{cases} \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) ds dt & \text{Continuous} \\ \sum_{x_i \leq x} \sum_{y_j \leq y} P_{X,Y}(x_i, y_j) & \text{Discrete} \end{cases} \tag{176}$$

Just like for random variables, the pdf (continuous), the pmf (discrete), or the cdf (both), completely characterize probabilistically a random vector

For a continuous random vector:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) \tag{177}$$

The distribution (pdf, pmf or cdf) of the component random variables X and Y are called the marginal distribution of X and Y respectively

Marginal distributions

(X, Y) a random vector

$$F_X(x) = \lim_{y \to +\infty} F_{X,Y}(x, y) \quad F_Y(y) = \lim_{x \to +\infty} F_{X,Y}(x, y) \tag{178}$$

For a continuous random vector:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dx \tag{179}$$

For a discrete random vector:

$$P_X(x_i) = \sum_{y_j} P_{X,Y}(x_i, y_j) \quad P_Y(y_j) = \sum_{y_j} P_{X,Y}(x_i, y_j) \tag{180}$$

Example: discrete random vector

Let M and S be the minimum and the sum of two independent rolls of fair 3-faced die.

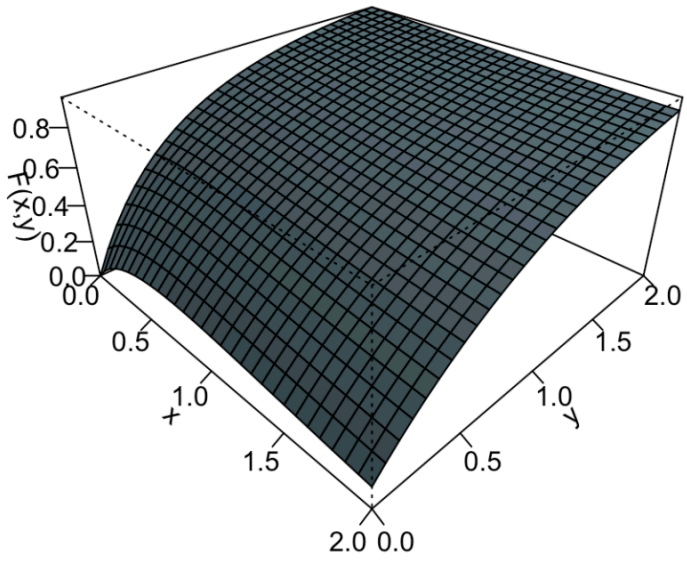
Determine:

- The joint pmf of M and S.
- The marginal pmf of M and of S.

Example: continuous random vector

Suppose that the joint cumulative distribution function of (X, Y) is given by:

$$F_{X,Y}(x,y) = \begin{cases} 1 - e^{-2x} - e^{-y} + e^{-(2x+y)} & \text{if } x > 0, y > 0 \\ 0 & \text{Otherwise} \end{cases} \tag{181}$$



Example: continuous random vector

$$F_{X,Y}(x,y) = \begin{cases} 1 - e^{-2x} - e^{-y} + e^{-(2X+y)} & \text{if } x > 0, y > 0 \\ 0 & \text{Otherwise} \end{cases} \tag{182}$$

1. Determine the joint probability density function of X and Y.
2. Determine the marginal cumulative distribution functions of X and Y.
3. Determine the marginal probability density functions of X and Y.
4. Find out whether X and Y are independent.
5. Determine $\text{Cov}(X, Y)$ and $\rho(X, Y)$

Example: continuous random vector

Suppose that the joint probability density function of X and Y is given:

$$f_{X,Y}(x,y) = \begin{cases} x + cy^2 & if\ 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & Otherwise \end{cases} \tag{183}$$

1. Find the constant *c*.
2. Determine the joint cumulative distribution functions of (X, Y).
3. Determine the marginal probability density functions of X and Y.
4. Find out whether X and Y are independent.
5. Determine \text{Cov}(X,Y) and $\rho(X, Y)$

Independence of random variables

X and Y are independent if for any $A, B \subset \mathbb{R}$, $\{X \in A\}$ and $\{Y \in B\}$ are independent events:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{184}$$

- Equivalent to the factorization of the joint cdf as a product of the marginal cdfs:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \tag{185}$$

- For continuous random vectors, also equivalent to the factorization of the joint pdf as a product of the marginal pdfs:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{186}$$

- For discrete random vectors, also equivalent to the factorization of the joint pmf as a product of the marginal pmfs:

$$P_{X,Y}(x,y) = P_X(x)P_Y(y) \tag{187}$$

- Definition of independence and factorization equivalences extend to multiple random variables $X_1, \dots, X_n$

Propagation of independence

NOTE this is important and not covered in the book

- If X, and Y are independent so are $g(X)$ and $h(Y)$ for $g, h : \mathbb{R} \to \mathbb{R}$
- If $X_1, X_2, \dots, X_n$ are independent so are $h_1(X_1), \dots, h_n(X_n)$, for $h_i : \mathbb{R} \to \mathbb{R}$
- If $X_1, \dots, X_n, Y_1, \dots, Y_m$ are independent then, $g(X_1, \dots, X_n), h(Y_1, \dots, Y_m)$, where $g : \mathbb{R}^n \to \mathbb{R}, h : \mathbb{R}^m \to \mathbb{R}$

Examples:

1. $X_1, X_2, X_3, X_4, Y_1, Y_2$ independent $\Rightarrow Z_1 = \frac{\sin(X_1^2)+e^{X_2}}{X_3^5+1}, Z_2 = \cos(Y_1) - Y_2^3$ are independent
2. $X_1, X_2, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p), Y_1, Y_2, \dots, Y_m \overset{iid}{\sim} \text{Bernoulli}(q)$
   (iid stands for independent identically distributed)

Let $X = X_1 + X_2 + \cdots + X_n, Y = Y_1 + Y_2 + \cdots + Y_m$.

Then $X \sim \text{Binomial}(n,p)$ and $Y \sim \text{Binomial}(m,q)$ and X, Y are independent

Expectation of a random vector

$$E[(X,Y)] = (E[X], E[Y]) \tag{188}$$

Interpretation is analog to that for random variables, 'center' of the two-dimensional distribution, center of mass if we think of probability as mass distributed on the surface of the plane $\mathbb{R}^2$

In general,

$$E[\mathbf{X}] = (E[X_1], \dots, E[X_n]) \tag{189}$$

Example: $X \sim \text{Exp}(\lambda_1), Y \sim \text{Exp}(\lambda_2)$

$$E[(X,Y)] = (E[X], E[Y]) = \left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}\right) \tag{190}$$

Multi-dimensional LOTUS

$\mathbf{X} = (X_1, \dots, X_n)$ a random vector, $g : \mathbb{R}^n \to \mathbb{R}$ a function, then:

$$E[g(\mathbf{X})] = E[g(X_1, \dots, X_n)] = \begin{cases} \int_{\mathbb{R}^n} g(x_1, \dots, x_n) f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \dots dx_n & \text{if } \mathbf{X} \text{ continuous} \\ \sum_{\mathbf{x}_i} g(\mathbf{x}_i) f_{\mathbf{X}}(\mathbf{x}_i) & \text{if } \mathbf{X} \text{ discrete} \end{cases} \tag{191}$$

Consequences:

- **Expectation is linear:**

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n] \tag{192}$$

- $E[XY] = E[X]E[Y] \quad$ for independent random variables X, Y

Example of linearity: $X \sim \text{Binomial}(n, p)$

Then, $X = X_1 + \cdots + X_n$ where $X_1, X_2, \ldots X_n \sim \text{Bernoulli}(p)$

$$E[X] = E[X_1] + \cdots + E[X_n] = \underbrace{p + \cdots + p}_{\text{n times}} = np \tag{193}$$

Covariance

For arbitrary random variables X and Y:

$$Var[X + Y] = Var[X] + Var[Y] + 2E[(X - E[X])(Y - E[Y])] \tag{194}$$

The term:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \tag{195}$$

is called the covariance of X and Y.

It measures how much X and Y co-vary, i.e. vary together.

- $\text{Cov}(X, Y) > 0$ if whenever $X > E[X]$ is likely that also $Y > E[Y]$ and vice versa (and that when $X < E[X]$ also more likely that $Y < E[Y]$)
- $\text{Cov}(X, Y) < 0$ if whenever $X > E[X]$ is likely that $Y < E[Y]$ and vice versa (and that when $X < E[X]$ also more likely that $Y > E[Y]$)
- X, Y independent $\Rightarrow \text{Cov}(X, Y) = 0$
- Converse is not true: $\text{Cov}(X, Y) = 0 \Rightarrow X, Y$ independent

Covariance

Example of additivity of variance for uncorrelated RVs

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

Then, $X = X_1 + \cdots + X_n$ where $X_1, X_2, \ldots X_n \overset{iid}{\sim} \text{Bernoulli}(p)$

$$\begin{aligned} Var[X] = Var[X_1] + \cdots + Var[X_n] + \sum_{i<j} 2\text{Cov}(X_i, X_j) = \\ = \underbrace{p(1-p) + \cdots + p(1-p)}_{\text{n times}} = np(1-p) \end{aligned} \tag{196}$$

Properties of covariance

- Covariance is linear in each of its terms:

$$\text{Cov}(aX + bY, cU + dV) = ac\text{Cov}(X, U) + ad\text{Cov}(X, V) + bc\text{Cov}(Y, U) + bd\text{Cov}(Y, V) \tag{197}$$

$$\text{Cov}(X, X) = Var[X] \quad (\Rightarrow Var[aX] = a^2 Var[X]) \tag{198}$$

- **Cauchy-Schwartz inequality**

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]} \tag{199}$$

- From Cauchy-Schwartz inequality using $X - E[X]$ and $Y - E[Y]$ in place of X and Y, it follows that

$$|\text{Cov}(X, Y)| \leq \sqrt{Var[X]Var[Y]} = \sigma_X \sigma_Y \tag{200}$$

($\sigma_X = \sqrt{Var[X]}, \sigma_Y = \sqrt{Var[Y]}$ are called the standard deviations of X and Y respectively)

- Equivalently,

$$-1 \leq \frac{\mathrm{Cov}(X,Y)}{\sqrt{Var[X]Var[Y]}} \leq 1 \tag{201}$$

Correlation

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{Var[X]Var[Y]}} \quad \text{is the correlation between } X \text{ and } Y \tag{202}$$

- X, Y independent $\Rightarrow \rho(X,Y) = 0$
- Converse is not true: $\rho(X,Y) = 0 \Rightarrow X, Y$ independent
- $\rho(X,Y)$ is a standardized version of $\mathrm{Cov}(X,Y)$
- $\rho(X,Y)$ is unaffected by changes of units:

$$\rho(aX+b, cY+d) = \rho(X,Y) \tag{203}$$

- Covariance it's not invariant: $\mathrm{Cov}(X,Y) = \rho(X,Y)\sigma_X\sigma_Y$, so the larger σ_X and σ_Y, the larger $\mathrm{Cov}(X,Y)$ in absolute value (provided $\rho(X,Y) \neq 0$)

Week 7 – Transformation of RVs, Law of large numbers

Last Class

- Random vector = multiple random variables defined on the same probability space.
- Discrete (have joint pmf) and continuous (have joint pdf) random vectors.
- LOTUS to compute the expectation of a transformed random vector.
- Independence of random variables.
- Covariance and correlation.

Distribution of a sum of discrete RVs

(X, Y) a random vector. What is the distribution of $Z = X + Y$?

- If (X, Y) is discrete then

$$P_Z(z) = \sum_{x_i + y_i = z} P_{X,Y}(x_i, y_i) \tag{204}$$

- Example: Let X and Y be independent random variables with common pmf given by $P_X(0) = \frac{1}{4}, P_X(1) = \frac{1}{2}, P_X(2) = \frac{1}{4}$. Find the pmf of $Z = X + Y$.

Distribution of a sum of continuous RVs

(X, Y) a random vector and $Z = X + Y$

If (X, Y) is continuous then
$f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Y}(x, z-x)dx = \int_{-\infty}^{+\infty} f_{X,Y}(z-y, y)dy$

Example: $X \sim Exp(\lambda)$, $Y \sim U[0,1]$, X and Y independent. Find the cdf and the pdf of $Z = X + Y$.

Bivariate normal distribution

Two random variables X and Y are said to have a **bivariate normal distribution** if their joint pdf is given by:

$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_X}{\sigma_X})^2 + (\frac{y-\mu_Y}{\sigma_Y})^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}]\}$

$\mu = \begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix} = E\begin{pmatrix}X \\ Y\end{pmatrix}$ and

$\Sigma = \begin{pmatrix}Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y)\end{pmatrix} = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}$ is the variance-covariance-matrix.

where $\rho \in (-1,1), \sigma_X > 0, \sigma_Y > 0$.

$(X,Y) \sim N(\mu, \Sigma)$

Sums of normals

- If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ are **independent** then
  $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- If $(X,Y) \sim N(\mu, \Sigma)$ then
  $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$
- If $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ but not bivariate normal then $X + Y$ are not necessarily normal.
- Example: $X \sim N(\mu, \sigma^2), Y = -X. Y \sim N(-\mu, \sigma^2)$ but $X + Y = 0$, a constant random variable.

Sum of independent exponentials

$X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$ then $Z = X_1 + \ldots + X_n \sim Gamma(n, \lambda)$

A continuous random variable has a **gamma distribution** with parameters $\alpha > 0$ and $\lambda > 0$ if its probability density function is given by:

$f(x) = \frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)} I_{[0,+\infty)}(x)$

where

$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt$

for $\alpha > 0$ is the **gamma function**.

The gamma function is a generalization of the factorial: $\Gamma[\alpha + 1] = \alpha\Gamma[\alpha]$ and $\Gamma[n+1] = n!$ for $n = 0, 1, 2, \ldots$

Maximum and minimum

$X_1, \ldots, X_n \sim F(x)$ iid random variables and $U = min(X_1, \ldots, X_n)$,
$V = max(X_1, \ldots, X_n)$

Then:

$F_V(v) = F(v)^n$

(if continuous with density $f(x)$, then $f_V(v) = nf(v)F(v)^{n-1}$

$F_U(u) = 1 - (1 - F(u))^n$

(if continuous with density $f(x)$ then $f_U(u) = nf(u)(1 - F(u))^{n-1}$

Averaging reduces variability

$X_1, X_2 \sim F(x)$ independent

$Var(X_1) = Var(X_2) = \sigma^2$

$Var(\frac{X_1+X_2}{2}) = \frac{1}{2^2}(Var(X_1) + Var(X_2)) = \frac{1}{2^2}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$

Averaging reduces variability

$X_1, X_2, \ldots$ i.i.d with expectation μ and variance σ^2

$\overline{X_n} = \frac{X_1+\ldots+X_n}{n}$

$E[\overline{X_n}] = \frac{1}{n}(E[X_1]+\ldots+E[X_n]) = \frac{1}{n}\underbrace{(\mu+\ldots+\mu)}_{n\ times} = \mu$

$Var[\overline{X_n}] = \frac{1}{n^2}(Var[X_1]+\ldots+Var[X_n]) = \frac{1}{n^2}\underbrace{(\sigma^2+\ldots+\sigma^2)}_{n\ times} = \frac{\sigma^2}{n}$

Markov's inequality

If $U \geq 0$ is a random variable with finite expectation, then, for every $t > 0$:

$P(U \geq t) \leq \frac{E[U]}{t}$

Intuition:

- Let U be the income of a random selected individual from a population. Taking $t = 2E[U]$, Markov's inequality says that $P(U \geq 2E[U]) \leq \frac{1}{2}$. It is impossible for more than half of the population to make at least twice the average income.
- This has to be true because if more than half of the population make more than twice the average income, the average income would have to be higher!
- Taking $t = 3E[U]$, we get $P(U \geq 3E[U]) \leq \frac{1}{3}$. It is impossible for more than one third of the population to make at least 3 times the average income.

Proof

$U \geq 0$ a random variable with finite expectation, $t > 0$:

Define the new random variable $U_t = \begin{cases} 0 & if\ U < t \\ t & if\ U \geq t \end{cases}$

U_t is discrete random variable taking only values 0 and t. So, $E[U_t] = tP(U \geq t)$.

Clearly $U \geq U_t$

Expectations preserve inequalities, so, $E[U] \geq E[U_t] \Rightarrow E[U] \geq tP(U \geq t) \Rightarrow P(U \geq t) \leq \frac{E[U]}{t}$

Chebyshev's inequality

If X is a random variable with mean μ and variance σ^2 , then, for every $t > 0$:

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

or equivalently

- $P(|X - E[X]| < t) \geq 1 - \frac{Var[X]}{t^2}$

or equivalently

- $P(\frac{|X-E[X]|}{\sigma} \geq t) \leq \frac{1}{t^2}$

Proof

X a random variable with mean μ and variance σ^2, $t > 0$

Consider the non-negative random variable $U = (X - \mu)^2$

Apply Markov's inequality to U, with t^2 : $P(U \geq t^2) = P((X - \mu)^2 \geq t^2) \leq \frac{E[(X-\mu)^2]}{t^2} = \frac{Var(X)}{t^2}$

Noticing that $\{|X - \mu| \geq t\} = \{(X - \mu)^2 \geq t^2\}$, we get Chebyshev's inequality:

$P(|X - \mu| \geq t) \leq \frac{Var(X)}{t^2}$

Chebyshev's inequality consequences

Any random variable with finite variance (continuous, discrete, mixed), regardless of its distribution, has most of its probability mass concentrated within a few standard deviations of its mean.

Taking $t = k\sigma, k = 2, 3, 4$ and applying Chebishev's inequality:

$P(|X - E[X]| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$

$P(|X - E[X]| \leq k\sigma) \geq 1 - \frac{1}{k^2}$

For $k = 2, 3, 4$ the bound are $3/4$, $8/9$, and $15/16$ respectively

Weak Law of large numbers (WLLN)

$X_1, X_2, \ldots$ i.i.d with finite expectation μ then:

$\lim_{n\to\infty} P(|\overline{X_n} - \mu| > \epsilon) = 0$

This form of convergence of random variables is called **convergence in probability** and is denoted as
$\overline{X_n} \xrightarrow{P} \mu$

Equivalently:

$\lim_{n\to\infty} P(|\overline{X_n} - \mu| \leq \epsilon) = 1$

The sample mean converges to the true mean

(Weak) Law of large numbers

Proof

$X_1, X_2, \ldots$ i.i.d with finite expectation μ.

To make the proof easier, we will also assume that $\sigma^2 < +\infty$

But the for the WLLN to hold only the expectation needs to exist. The variance may be infinite.

Let $\epsilon > 0$. Apply Chebyshev's inequality to $\overline{X_n}$:

$P(|\overline{X_n} - \mu| \geq \epsilon) \leq \frac{Var(\overline{X_n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$

This implies:

$\lim_{n\to\infty} P(|\overline{X_n} - \mu| \geq \epsilon) = 0$

Strong Law of large numbers (SLLN)

$X_1, X_2, \ldots$ i.i.d with finite expectation μ.

$P(\lim_{n\to\infty} \overline{X_n} = \mu) = 1$

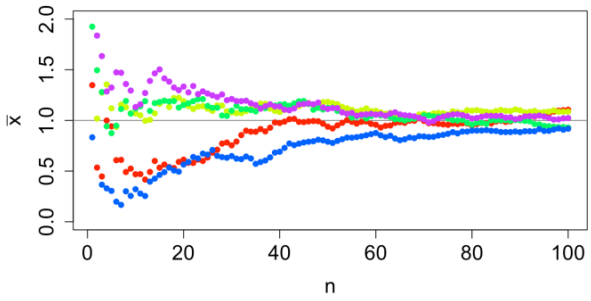
This form of convergence of random variables is called **convergence with probability 1** and is denoted as
$\overline{X_n} \xrightarrow{w.p.1} \mu$
(also called **almost sure convergence**)

The strong law is **stronger** than weak law because convergence with probability 1 is a stronger concept
than convergence in probability: convergence with probability $1 \implies$ convergence in probability (but not the other way around)

Law of large numbers

$X_1, X_2, \ldots, X_n \sim N(1,1)$

Five realizations of $\overline{X_n}$ as a function of n



Week 8 – Central Limit Theorem

Last Class

- **Markov's inequality:** for any RV $U \geq 0$
  $P(U \geq t) \leq \frac{E[U]}{t} \quad for\ any\ t > 0$
- **Chebyshev's inequality:** for any RV $X, Var[X_i] < \infty$;
  $P(|X - E[X]| \geq t) \leq \frac{Var[X]}{t^2} \quad for\ any\ t > 0$
- **Law of large numbers:** $X_1, X_2, \ldots$ i.i.d with finite expectation μ:
  - **Weak:** $\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0$
  - **Strong:** $P(\lim_{n\to\infty} \overline{X}_n = \mu) = 1$

Basic Statistical Model

- Want to learn characteristics of a population:
  - Income of Los Angeles residents
  - Blood pressure of patients form LA county hospital
  - Vaping in among teenagers in California
  - Genotypes at a gene among individuals with Hispanic ancestry
- We model the distribution of the characteristic as a random variable X (e.g. height, blood pressure, vaping (yes vs. no), Genotypes (0,1,2)))
- To learn about X we collect a random sample: $X_1, X_2, \ldots, X_n$ from $F_X(x)$ the distribution of X
- $X_1, X_2, \ldots, X_n$ have the same probability distribution and are mutually independent.
- The distribution $F(x;\theta)$ of X_i is unknown or only partially known

    - E.g. $X_i \sim N(\mu, \sigma^2)$ with known σ^2 but unknown μ
- We conceptualize the observed data $x_1, \ldots, x_n$ as realizations of the random variables $X_1, \ldots, X_n$

Sample statistics

Sample statistics are empirical summaries of the data:

- sample mean: $\overline{x}_n = \frac{x_1 + \ldots + x_n}{n}$
- sample variance: $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x}_n)^2$ (sometimes $\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x}_n)^2$)
- sample minimum: $x_{(1)} = min(x_1, \ldots, x_n)$

We think of these as realizations of the corresponding random variables:

$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$

$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X_n})^2$

$X_{(1)} = min(X_1, \ldots, X_n)$

Central Limit Theorem

$X_1, \ldots, X_n$ i.i.d. $E[X_i] = \mu, Var[X_i] = \sigma^2$ then:

Sum form: $S_n = X_1 + \ldots + X_n$

$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0,1)$

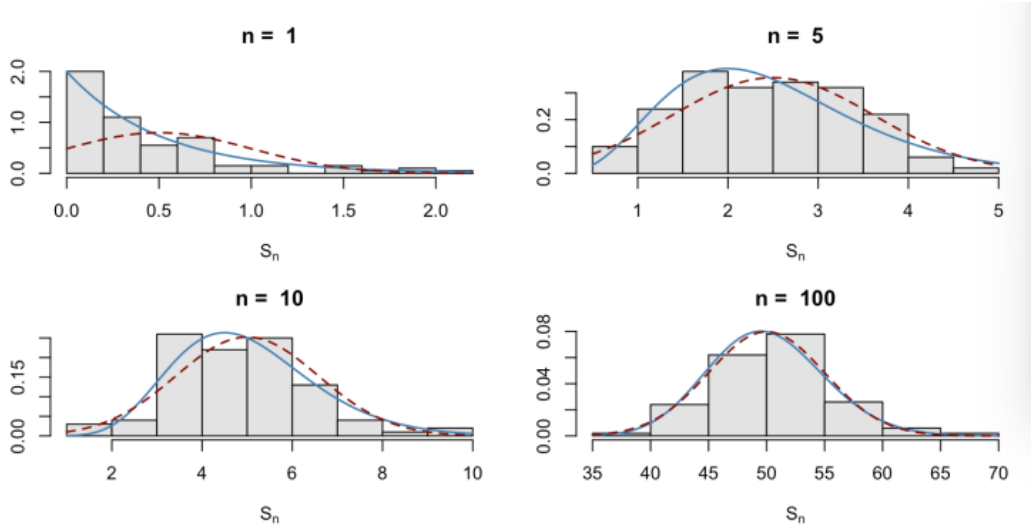
Sample mean form: $\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$

$Z_n = \frac{\overline{X}_n - E[\overline{X}_n]}{\sqrt{Var[\overline{X}_n]}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$

The symbol $\xrightarrow{D}$ denotes **convergence in distribution**. It means that as $n \to \infty$ the cumulative distribution function of $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ gets closer and closer to the cdf of a standard normal.

Formally, if $F_n(x) = F_{Z_n}(x)$ denotes the cdf of Z_n, then $F_n(x) \to \Phi(x)$ for every $x \in \mathbb{R}$, where $\Phi(x)$ is the cdf of a $N(0,1)$ RV.

Sum of n i.i.d. exponentials: $S_n = X_1 + \ldots + X_n, X_i \sim Exp(2)$



Blue line is pdf of S_n. Red line is pdf of a normal $N(\frac{n}{2}, \frac{n}{4})$

Example

A runner attempts to pace a 100m race. Her strides (steps) are independently distributed with mean $\mu = 0.97$ meters and a standard deviation of $\sigma = 0.1$. What is the (approximate) probability that her 100 strides differ from 100 meters by no more than five meters?

Estimation

We want to learn about a distribution in a population

- e.g. proportion of democrat voters in CA, average blood pressure among covid survivors aged $70+$, vaping frequency among young adults in the US, rate of patient ER night admissions in LA county hospitals
- We take a sample of size n from the population and conceptualize it as a random sample $X_1, \ldots, X_n$ from a distribution $F_X(x)$ that is **totally or partially unknown to us**.
- We want to estimate specific characteristics or **parameters** of the underlying distribution:
  - true mean $E[X_i] = \mu$ (e.g. X_i blood pressure)
  - the true variance, $Var[X_i] = \sigma^2$ (e.g. X_i blood pressure)
  - or a probability like $P(X_i = 1)$ (e.g. $1 =$ democrat voter; $0 =$ not democrat voter)
  - or a rate (e.g. expected number of ER patients per hour) ($X_i =$ number of patients within a period of 1-hour)

Example 1: To estimate the unknown mean of a population μ (e.g. blood pressure)

- Natural to use the **sample mean** $\overline{X}_n$ to estimate μ because we know that for large n the sample mean will be close to the true mean.

Example 2. We can model the number of arrivals per hour at an ER unit as a $Poisson(\lambda)$. Suppose we count the arrivals during each on n (non-overlapping) 1-hour intervals to get the random sample $X_1, \ldots, X_n$. Here λ is unknown and we want to estimate it.

- A natural estimate is also the **sample mean** $\overline{X}_n$ because $E[X_i] = \lambda$.
- But using the **sample variance** S_n^2 is also reasonable because $Var[X_i] = \lambda$
- In future classes we'll learn how to choose among different options for estimating a parameter of interest

Estimator vs. estimate

- Estimate: value t that only depends on the dataset $x_1, x_2, \ldots, x_n$, i.e., t is some function of the dataset $t = h(x_1, x_2, \ldots, x_n)$. Example: $t = \overline{x}_n = \frac{x_1 + \ldots + x_n}{n}$
- An *estimate is a number* (or a vector of numbers in multi-parameter estimation problems)
- **Estimator**: Let $t = h(x_1, x_2, \ldots, x_n)$ be an estimate based on the dataset $x_1, x_2, \ldots, x_n$. Then t is a realization of the random variable $T = h(X_1, X_2, \ldots, X_n)$. The random variable T is called an estimator.
- *An estimator is a random variable*
- *An estimate is a realization of random variable*

Sampling distribution

Example: estimating the proportion p of LA teenagers that vape from a sample $Y_1, \ldots Y_n$
$Y_i \sim Bernoulli(p)$.

- Y_i records whether the teenager vapes (yes=1 vs. no=0)
- The sample proportion $\hat{p}_n = \frac{S_n}{n}$ of vapers is a natural estimator of p ($S_n = Y_1 + \ldots + Y_n$ is the total number of vapers in the sample), because by the LLN $\hat{p}_n \to p$

- The sampling distribution is just the standard distribution (pdf, pmf, cdf) of the random variable we call the estimator.
- For example of vapers, the sampling distribution of $\hat{p}_n$ is the distribution of the random variable $\hat{p}_n = \frac{S_n}{n}$

Unbiased Estimators

$X_1, \dots X_n$ a random sample from distribution F, and θ a parameter of interest about F (e.g. mean)

- An estimator T of a parameter θ is unbiased if $E[T] = \theta$
- An unbiased estimator has no systematic tendency to produce estimates that are larger than or smaller than the target parameter θ
- The difference $E[T] - \theta$ is called the bias of the estimator T
- If $bias \neq 0$ the estimator is called biased

Sample Mean and sample variance

- $X_1, \dots X_n$ a random sample with mean $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$
- The sample mean $\overline{X}_n$ is an unbiased estimator of the population mean μ.
  $E[\overline{X}_n] = \mu$
- The sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is an unbiased estimator of the population variance σ^2
  $E[S_n^2] = \sigma^2$

Unbiasedness is not preserved under general transformations

- In general, if T is an unbiased estimator of θ, $g(T)$ is not an unbiased estimate of $g(\theta)$
- Example:
  - $\overline{X}_n$ is unbiased for $E[X_i] = \mu$
  - But $\overline{X}_n^2$ is not unbiased for $E[X_i]^2 = \mu^2$, Because by Jensen's inequality $E[\overline{X}_n^2] > E[\overline{X}_n]^2 = \mu^2$
- **Jensen's inequality:** if $g(t)$ is a convex function, then $E[g(T)] \geq g(E[T])$. Equality holds only if $g(x) = at + b$
- If g linear $g(t) = at + b, E[g(T)] = E[aT + b] = aE[T] + b = g(E[T])$ so $g(T)$ is unbiased for $g(\theta)$

Week 9 – Unbiased estimation - Efficiency, MSE

Last Class-CLT

$X_1, \dots, X_n \overset{i.i.d.}{\sim} X\ E[X_i] = \mu, Var[X_i] = \sigma^2$ then:

Sum form: $S_n = X_1 + \dots + X_n$

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0,1) \tag{205}$$

Sample mean form: $\overline{X_n} = \frac{X_1 + \dots + X_n}{n}$

$$Z_n = \frac{\overline{X_n} - E[\overline{X_n}]}{\sqrt{Var[\overline{X_n}]}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{D} N(0,1) \tag{206}$$

The symbol $\xrightarrow{D}$ denotes **convergence in distribution**. It means that as $n \to \infty$ the cumulative distribution function of $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma}$ gets closer and closer to the cdf of a standard normal.

Formally, if $F_n(x) = F_{Z_n}(x)$ denotes the cdf of Z_n, then $F_n(x) \to \Phi(x)$ for every $x \in \mathbb{R}$, where $\Phi(x)$ is the cdf of a $N(0,1)$ RV.

Last Class-Basic statistical model

- Want to learn characteristics of a population:
  - Income of Los Angeles residents
  - Blood pressure of patients from LA county hospital
  - Vaping among teenagers in California
  - Genotypes at a gene among individuals with Hispanic ancestry
- We model the distribution of the characteristic as a random variable X (e.g. height, blood pressure, vaping (yes vs. no), Genotypes (0,1,2))
- To learn about X we collect a random sample: $X_1, X_2, \ldots, X_n$ from $F_X(x)$, the distribution of X
- $X_1, X_2, \ldots, X_n$ have the same probability distribution and are mutually independent.
- The distribution $F_X(x)$ of X_i is unknown or only partially known.
- E.g. $X_i \sim N(\mu, \sigma^2)$ with known σ^2 but unknown μ

Choosing among unbiased estimators. Example

- $X_1, \ldots, X_n \overset{i.i.d}{\sim} U[0, \theta]$
- $f_X(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$
- $E[X_i] = \frac{\theta}{2}$, so $E[\overline{X_n}] = \frac{\theta}{2}$
- Then $\hat{\theta} = 2\overline{X_n}$ is an unbiased estimator of θ

Intuitively a natural estimate for theta could also be based on $T = \max\{X_1, \ldots, X_n\}$

- $f_T(t) = n\frac{t^{n-1}}{\theta^n} I_{[0,\theta]}(t)$
- $E[T] = \frac{n}{n+1}\theta$ so $\tilde{\theta} = \frac{n+1}{n} T$ is an unbiased estimator of θ

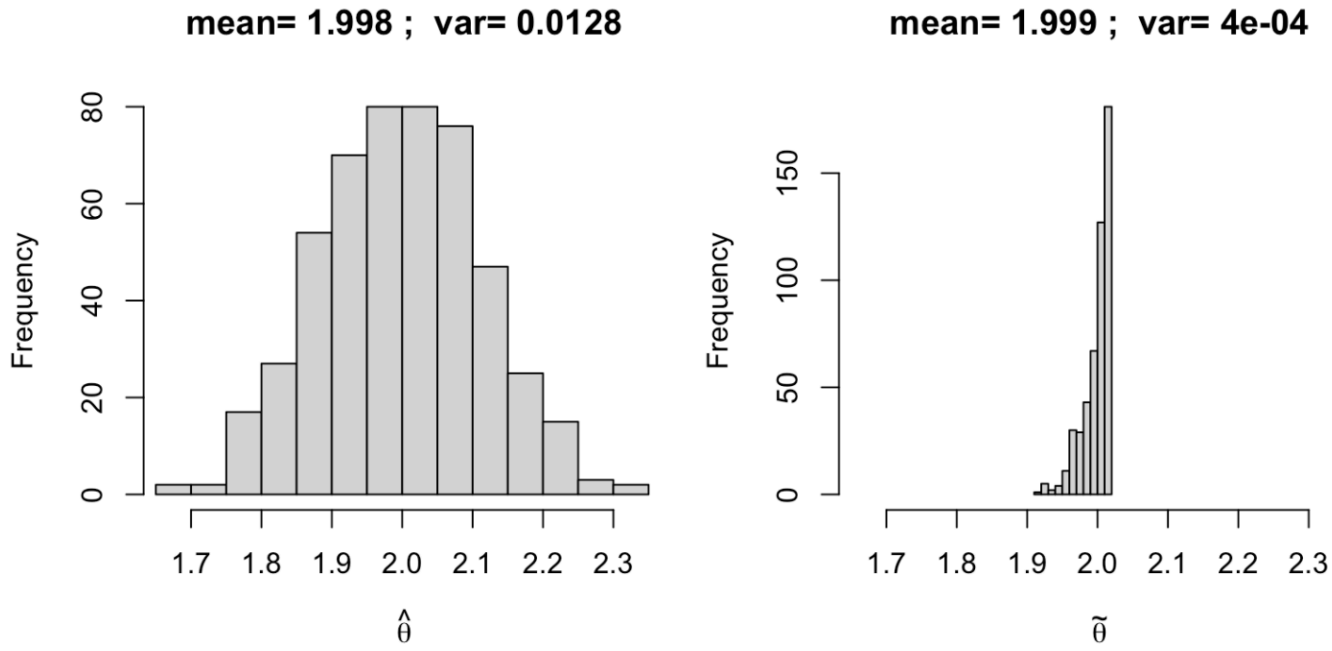
So, both $\hat{\theta}$ and $\tilde{\theta}$ are unbiased estimators of θ, which one is better?

Choosing among unbiased estimators-Example.

Let's perform a simulation to see how the two estimators behave:

```
set.seed(101)
theta = 2
n = 50
nsims = 500
theta_hat = theta_tilde = numeric(nsims)

for (i in 1:nsims){
    x = runif(n=n, max=theta)
    theta_hat[i] = 2 * mean(x)
    theta_tilde[i] = ((n+1)/n) * max(x)
}
```

Choosing among unbiased estimators-Example



Choosing among unbiased estimators-Example

- We see empirically that $\tilde{\theta}$ is much less variable than $\hat{\theta}$
- But also theoretically, $Var[\tilde{\theta}] = \frac{\theta^2}{n(n+2)}$ and $Var[\hat{\theta}] = \frac{\theta^2}{3n}$
- So, $Var[\tilde{\theta}] < Var[\hat{\theta}]$, for $n \geq 2$ and goes much faster to zero!
- Additionally, $\hat{\theta}$ can take values way over the true θ
- So, $\tilde{\theta}$ is a much better estimate of θ than $\hat{\theta}$

Mean square error

- Although **unbiasedness** is a desirable property, unbiased estimators do not always exist
- Even when they exist, requiring unbiasedness maybe too stringent (i.e. there can be other good estimators that are biased)
- A general performance of an estimator can be judged by the way it spreads around the parameter to be estimated:
  - If T is an estimator for a parameter θ, the **mean squared error** of T is the number: $MSE(T) = E[(T-\theta)^2]$
- It's easy to show that $MSE(T) = Var(T) + Bias(T)^2$ where $Bias(T) = E[T] - \theta$
- A **biased estimator** with a small bias could be more useful than an **unbiased estimator** with a large variance.
- Better to use $MSE(T)$ to choose between estimators
- When $Bias(T) = 0, MSE(T) = Var(T)$

Mean square error - example

$X_1, \ldots, X_n \overset{i.i.d}{\sim} Poisson(\mu)$

Two candidate estimators for $p_0 = P(X_i = 0) = e^{-\mu}$:

- $\widetilde{p}_0 = \frac{\text{number of } X_i = 0}{n}$

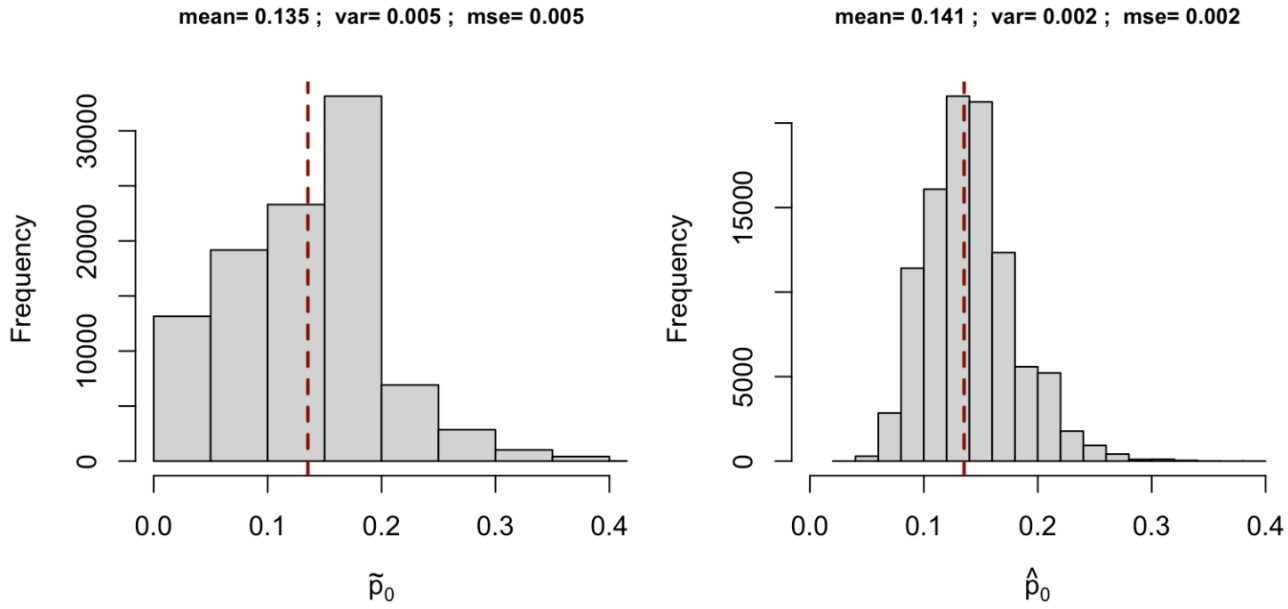
- $\widehat{p}_0 = e^{-\overline{X_n}}$

$\widetilde{p}_0$ is unbiased but $\widehat{p}_0$ is not? Which one is better?

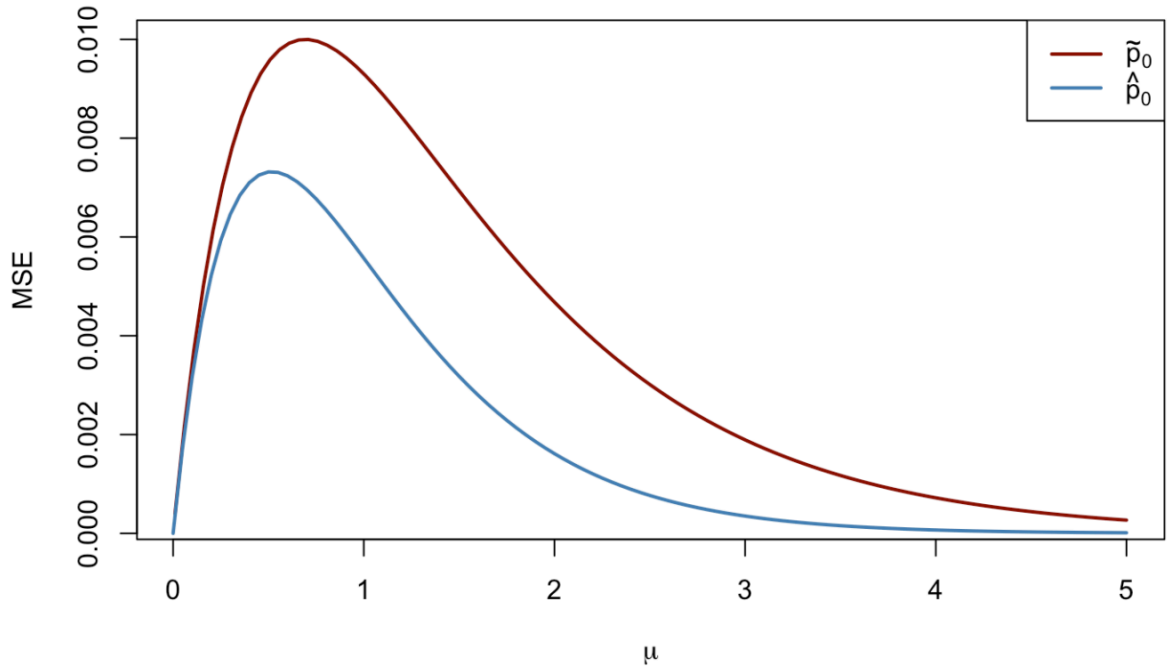
Mean square error - example

Simulation to see how the two estimators behave

- True $p_0 = e^{-2} \approx 0.135$



Mean square error - example



Week 10 – Maximum Likelihood Estimation

Review - Unbiased Estimators

$X_1, \ldots, X_n$ a random sample from distribution F, and a parameter of interest about F (e.g. mean)

- An estimator T of a parameter θ is **unbiased** if $E[T] = \theta$
- An unbiased estimator has no systematic tendency to produce estimates that are larger than or smaller than the target parameter θ
- $\overline{X_n}$ is an unbiased estimator of the population mean μ.
- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ is an unbiased estimator of the population variance σ^2
- Unbiasedness is not preserved under transformations.

Last Class- Choosing among unbiased estimators. Example

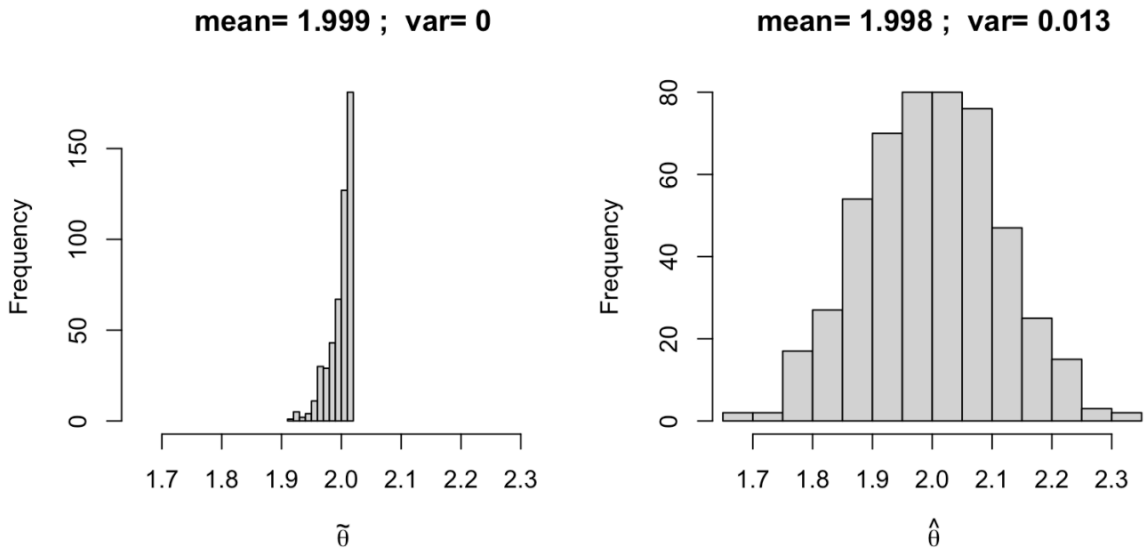
- $X_1, \ldots, X_n \overset{i.i.d}{\sim} U[0, \theta]$
- $f_X(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$
- $E[X_i] = \frac{\theta}{2}$, so $E[\overline{X_n}] = \frac{\theta}{2}$
- Then $\tilde{\theta} = 2\overline{X_n}$ is an unbiased estimator for θ

Intuitively a natural estimator for θ would also be based on $T = \max\{X_1, \ldots, X_n\}$

- $f_T(t) = n\frac{t^{n-1}}{\theta^n} I_{[0,\theta]}(t)$
- $E[T] = \frac{n}{n+1}\theta$ so $\hat{\theta} = \frac{n+1}{n}T$ is an unbiased estimator of θ

So, both $\tilde{\theta}$ and $\hat{\theta}$ are unbiased estimators of θ, which one is better?

Last Class - Choosing among unbiased estimators-Example



Last Class - Choosing among unbiased estimators example.

- We see empirically that $\tilde{\theta}$ is much less variable than $\hat{\theta}$

- But also theoretically, $Var[\tilde{\theta}] = \frac{\theta^2}{n(n+2)}$ and $Var[\hat{\theta}] = \frac{\theta^2}{3n}$
- So, $Var[\tilde{\theta}] < Var[\hat{\theta}]$ for $n \geq 2$ and goes much faster to zero!
- Additionally, $\hat{\theta}$ can take values way over the true θ
- So, $\tilde{\theta}$ is a much better estimate of θ than $\hat{\theta}$

Last Class- Mean square error

- Although **unbiasedness** is a desirable property, unbiased estimators do not always exist
- Even when they exist, requiring unbiasedness maybe too stringent (i.e. there can be other good estimators that are biased)
- A general performance of an estimator can be judged by the way it spreads around the parameter to be estimated:
  - If T is an estimator for a parameter θ, the **mean squared error** of T is the number: $MSE(T) = E[(T-\theta)^2]$
- It's easy to show that $MSE(T) = Var(T) + Bias(T)^2$ where $Bias(T) = E[T] - \theta$
- A **biased estimator** with a small bias could be more useful than an **unbiased estimator** with a large variance.
- Better to use $MSE(T)$ to choose between estimators
- When $Bias(T) = 0$, $MSE(T) = Var(T)$

Last Class- Mean square error example

$X_1, \ldots, X_n \overset{i.i.d}{\sim} Poisson(\mu)$

Two candidate estimators for $p_0 = P(X_i = 0) = e^{-\mu}$:

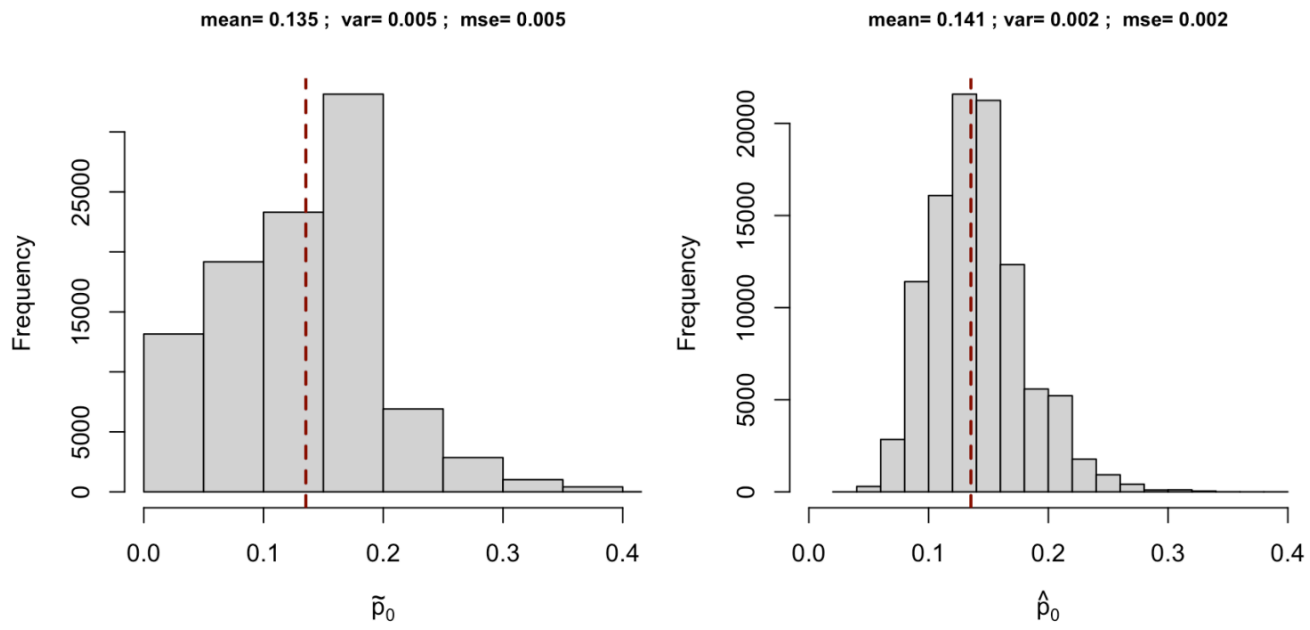
- $\widetilde{p}_0 = \frac{\text{number of } X_i = 0}{n}$
- $\widehat{p}_0 = e^{-\overline{X_n}}$

$\widetilde{p}_0$ is unbiased but $\widehat{p}_0$ is not? Which one is better?

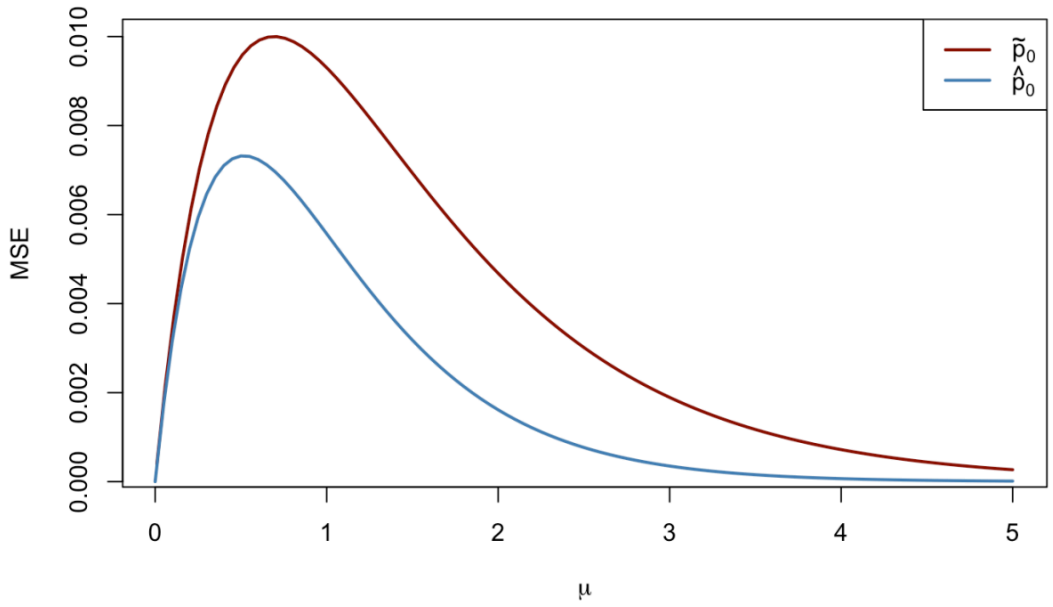
Last Class - Mean square error example

Simulation to see how the two estimators behave

- True $p_0 = e^{-2} \approx 0.135$



Last Class - Mean square error example



Maximum likelihood estimation

Previous examples of estimators were sample analogs of population parameters (e.g. sample mean and sample variance) or intuitive (e.g. estimator based on $\max\{X_1, \ldots, X_n\}$ for $U[0, \theta]$)

Maximum likelihood provides a very general way to construct estimators with good properties. The idea is intuitive:

- Choose as the estimator the value of θ that **maximizes the chances to produce the observed data.**

Example: An infusion pump is a medical device that delivers fluids - typically medications, nutrients, or saline - into a patient's body in controlled amounts over time. We track time to failure (in hours) for five identical infusion pumps used in a hospital.

- Observed data in hours: $120, 85, 200, 60, 150$

- Assuming independent and identically distributed lifetimes following an $Exponential(\lambda)$ **model.**
- Compute the maximum likelihood estimate of λ.

MLE - exponential example

Data: $t_1 = 120, t_2 = 85, t_3 = 200, t_4 = 60, t_5 = 150$

Model: i.i.d data with density $f(t|\lambda) = \lambda e^{-\lambda t}$

Likelihood:

$L(\lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda T_i}$
$= (\lambda e^{-\lambda \cdot 120})(\lambda e^{-\lambda \cdot 85})(\lambda e^{-\lambda \cdot 200})(\lambda e^{-\lambda \cdot 60})(\lambda e^{-\lambda \cdot 150})$
$= \lambda^5 exp[-\lambda(120 + 85 + 200 + 60 + 150)] = \lambda^5 exp(-615\lambda)$

log-likelihood:

$l(\lambda) = log\ L(\lambda) = 5\ log\ \lambda - 615\ \lambda$

MLE - exponential example

Setting the score to zero to find the candidate maxima of the log-likelihood:

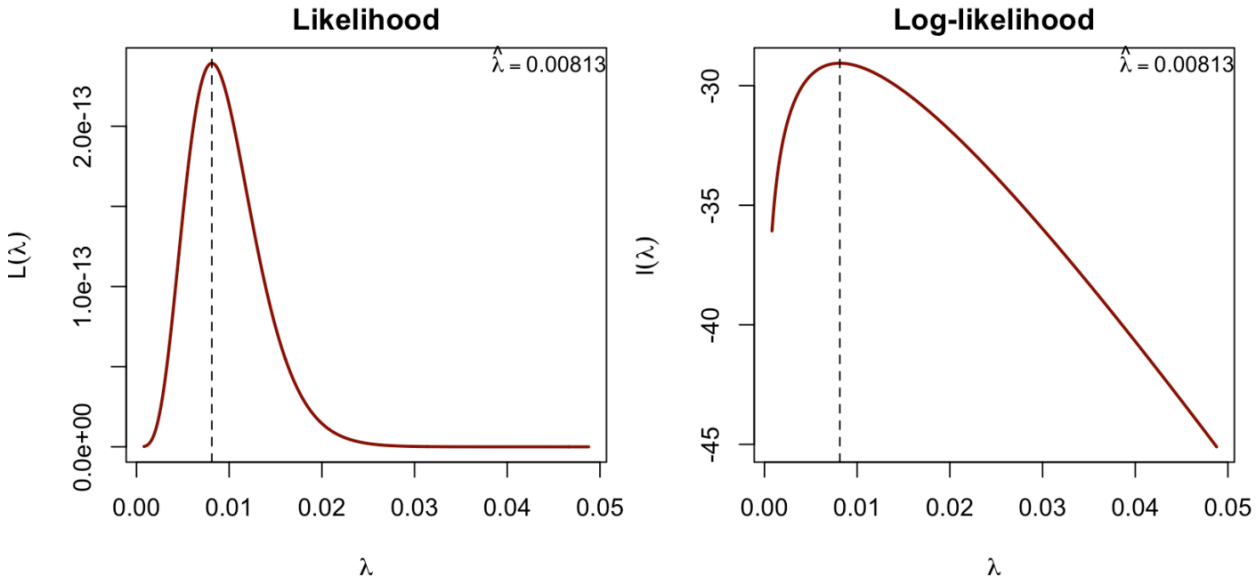
$\frac{dl}{d\lambda} = \frac{5}{\lambda} - 615$

$\frac{5}{\lambda} - 615 = 0 \Rightarrow \hat{\lambda} = \frac{5}{615} = 0.00813$

Checking 2^{nd} derivative:

$\frac{d^2 l}{d\lambda^2} = -\frac{5}{\lambda^2} < 0$

Maximum likelihood estimation – exponential example



Maximum likelihood estimation - exponential example

Let $T_1, \ldots, T_n \overset{i.i.d}{\sim} Exp(\lambda)$ with density $f(t; \lambda) = \lambda e^{-\lambda t}$ for $t > 0$. The likelihood and log-likelihood are

$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda T_i} = \lambda^n exp\left(-\lambda \sum_{i=1}^{n} T_i\right)$

$l(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} T_i$

Differentiate and set to zero:

$\frac{dl}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} T_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i}$

MLE - a more complex example Number of cycles until pregnancy:

Number of cycles until pregnancy:

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

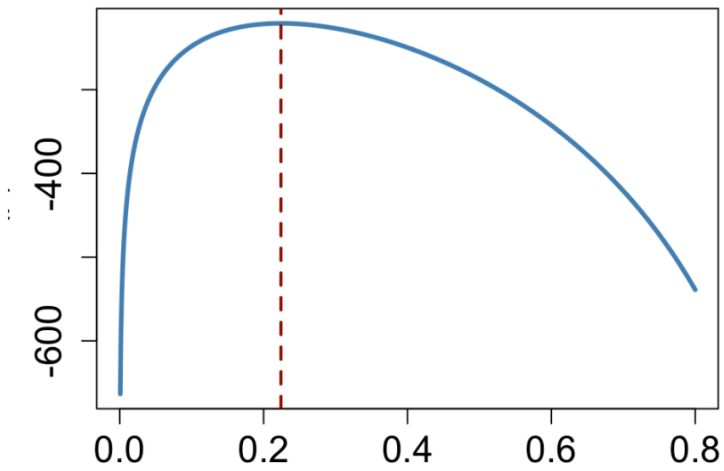
MLE - a more complex example

- Want to estimate $p = P(X = 1)$, where X_i is the number of cycles up to pregnancy among the i^{th} smoker
- We know that $S = \frac{\text{number of } X_i = 1}{n}$ is an unbiased estimator for p
- This yields the estimate $\widetilde{p} = \frac{29}{100} = 0.29$
- But it uses only a fraction of the data discarding the rest. Maybe there is a better estimator using all the data?

MLE - a more complex example

- Let X_i denote the number of cycles up to pregnancy for the i^{th} smoker. We can assume the $X_i \sim Geometric(p)$, and the X_i are independent. There are 415 smokers
- The probability of observing the data in the first row of the table is then given by:
  $L(p) = P(X_1 = x_1|p) \times P(X_2 = x_2|p) \times \ldots P(X_{100} = x_{100}|p)$
  $= p^{29} \cdot (p(1-p))^{16} \ldots ((1-p)^{11}p)^3 \cdot ((1-p)^{12})^7$
  $= p^{93}(1-p)^{322}$
- We want to find $0 \leq p \leq 1$ that makes $L(p)$ as large as possible, i.e. find the value of p that maximizes $L(p)$
- Equivalent to finding the maximizer of $l(p) = \log(L(p))$

MLE - a more complex example



$l(p) = 93\log(p) + 322\log(1-p)$

$0 = l'(p) = \frac{93}{p} - \frac{322}{1-p} \iff p = \frac{93}{415}$

$\hat{p} = \frac{93}{415} = 0.224$ is called the maximum likelihood estimate of p.

Maximum likelihood estimation

- The **maximum likelihood estimate** of θ is the value $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ that maximizes the log-likelihood function $l(\theta)$.
- The corresponding random variable $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is called the **maximum likelihood estimator** of θ.
- (We often use the same notation, $\hat{\theta}$ for both the estimate and the estimator. We understand which one we are referring to based on context)

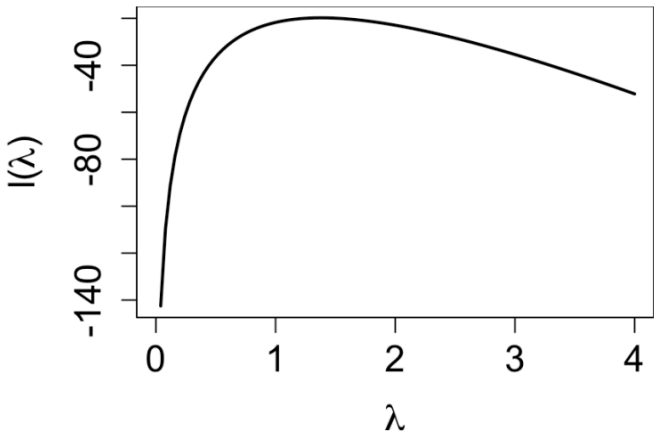
MLE with no analytical solution

Dataset: $x_1, \ldots, x_n$ modeled as a realization of random variables $X_1, \ldots, X_n$.

Where the pdf of X_i is a mixture of exponentials:

$X_i \sim Exp(\lambda)$ with probability $\frac{1}{2}$ and $Exp(2\lambda)$ with probability $\frac{1}{2}$

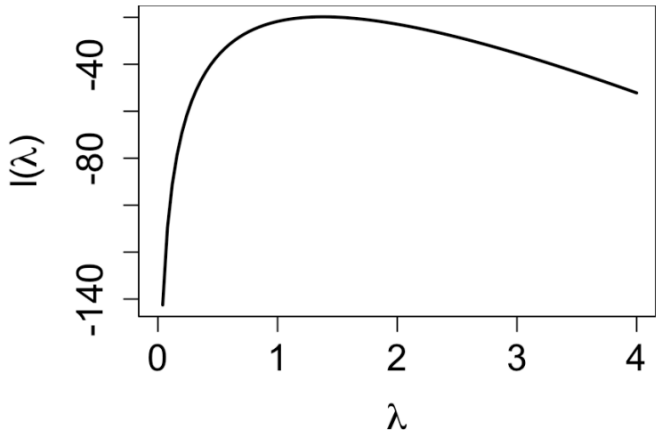
$f(x) = \frac{1}{2}\lambda e^{-\lambda x} + \frac{1}{2}2\lambda e^{-2\lambda x}$



MLE with no analytical solution

```r
set.seed(101)
n = 50

rate = 1
mix = rbinom(n=n, size=1, prob=1/2)
dataset = mix*rexp(n, rate=rate) + (1-mix)*rexp(n, rate=2*rate)

negloglik = function(x, data) {
    -sum(log(0.5*dexp(data, rate=x) + 0.5*dexp(data, rate=2*x)))
}

optimize(f=negloglik, interval=c(0.01, 10), data=dataset)

## $minimum
## [1] 1.37906

## $objective
## [1] 19.78651
```

MLE with no analytical solution



Properties of Maximum likelihood estimators

The maximum likelihood estimator (MLE) is **consistent**:
$\hat{\theta}_n \xrightarrow{P} \theta \text{ as } n \to +\infty.$

- i.e. as the sample size gets larger the MLE gets closer and closer to the true parameter θ.
- Consistency is much stronger that asymptotic unbiasedness.

Invariance:

- If $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Asymptotic minimum variance

- No other consistent estimator has lower asymptotic mean squared error.

Asymptotic normality

- In most situations of interest (but not always, e.g. $U[0, \theta]$) the sampling distribution of the MLE gets closer and closer to a **normal distribution** as $n \to \infty$.

With few exceptions MLE's are generally **biased estimators**.

Asymptotic variance of the MLE

Let $X_1, \ldots, X_n \sim f(x;\theta)$ and the log-likelihood function: $l(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta)$

The **Fisher's information** of θ (for one sample) is defined as:

$$I(\theta) = -E\left[\frac{\partial^2 l(\theta, X)}{\partial \theta^2}\right] \tag{207}$$

If $f(x;\theta)$ is 'well behaved' (e.g. its support does not depend on θ)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right) \tag{208}$$

- $\frac{1}{nI(\theta)}$ is the asymptotic variance of $\hat{\theta}$

The standard error of the MLE is $se(\hat{\theta}) = \frac{1}{\sqrt{nI(\theta)}}$

We can estimate the standard error of the MLE by $\widehat{se}(\hat{\theta}) = \frac{1}{\sqrt{nI(\hat{\theta})}}$

Week 11 – Confidence Intervals

Last Class - Maximum Likelihood

Maximum likelihood is a general method for estimating parameters of interest

Idea: choose parameter the maximizes the probability of observing the data actually observed

- The probability of observing the data actually observed as a function of parameter = **likelihood**
- Previous examples of estimators were sample analogs of population parameters (e.g. sample mean and sample variance) or intuitive (e.g. estimator based on $\max_{1 \leq i \leq n} X_i$ for $U[0, \theta]$)

Maximum likelihood can be used in cases when there is no 'natural' estimator like sample mean or variance

Maximum likelihood estimation

- The **maximum likelihood estimate** of θ is the value $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ that maximizes the log-likelihood function $l(\theta)$.
- The corresponding random variable $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is called the **maximum likelihood estimator** of θ.
- (We often use the same notation, $\hat{\theta}$ for both the estimate and the estimator. We understand which one we are referring to based on context)

Properties of Maximum likelihood estimators

The **maximum likelihood estimator (MLE) is consistent**:
$\hat{\theta}_n \to \theta$ as $n \to +\infty$

- i.e. as the sample size gets larger the MLE gets closer and closer to the true parameter θ.
- Consistency is **stronger than asymptotic unbiasedness** (which MLEs also have)

Invariance:

- If $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
- E.g. MLE of θ^2 is $\hat{\theta}^2$

Asymptotic minimum variance

- Lowest asymptotic mean squared error.

Asymptotic normality

- The distribution of the MLE gets closer and closer to a **normal distribution** as $n \to \infty$.
- MLE's are generally **biased estimators** - But bias disappears as sample size increases (asymptotically unbiased).

Asymptotic variance of the MLE

Let $X_1, \ldots, X_n \sim f(x; \theta)$ and the log-likelihood function: $l(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$

The **Fisher's information** of θ (for one sample) is defined as:

$$I(\theta) = -E\left[\frac{\partial^2 l(\theta, X)}{\partial \theta^2}\right] \tag{209}$$

If $f(x; \theta)$ is 'well behaved' (e.g. its support does not depend on θ):

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right) \tag{210}$$

- $\frac{1}{nI(\theta)}$ is the asymptotic variance of $\hat{\theta}$

The standard error of the MLE is $se(\hat{\theta}) = \frac{1}{\sqrt{nI(\theta)}}$

We can estimate the standard error of the MLE by $\widehat{se}(\hat{\theta}) = \frac{1}{\sqrt{nI(\hat{\theta})}}$

Confidence intervals

So far we've considered what are called **point estimators**.

- A point estimator by itself not that useful; need some measure of how variable the estimator is
- To supplement point estimators with information about their variability we can also report their SD, called the **standard error (SE)**
- A better option is to report a whole interval of plausible values for the parameter of interest, i.e. a **confidence interval**.

Based on Chebysheb's inequality, if T is an estimator for θ, and σ_T is the standard deviation of T:

$P(|T - \theta| \geq 2\sigma_T) \leq \frac{3}{4}$

$P(\theta \in (T - 2\sigma_T, T + 2\sigma_T)) \geq \frac{3}{4}$

$(T - 2\sigma_T, T + 2\sigma_T)$ is an **interval estimator** for θ

Confidence intervals

- Dataset $x_1, \ldots, x_n$ modeled as a realization of random variables $X_1, \ldots, X_n$.
- θ the parameter of interest (e.g. mean, variance, probability, etc.), and $0 < \alpha < 1$
- A $(1 - \alpha)$-level interval estimator for θ is a a pair of sample statistics $L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that:
  - $P(L_n < \theta < U_n) = 1 - \alpha$ for every value of θ
- The interval (l_n, u_n) where $l_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$, is called a $100(1 - \alpha)\%$ **confidence interval** for θ.
- The number α is called the **confidence level**.

- Often, we only require $P(L_n < \theta < U_n) > 1 - \alpha$ or $P(L_n < \theta < U_n) \approx 1 - \alpha$
    - These are **conservative** and **approximate** interval estimators respectively

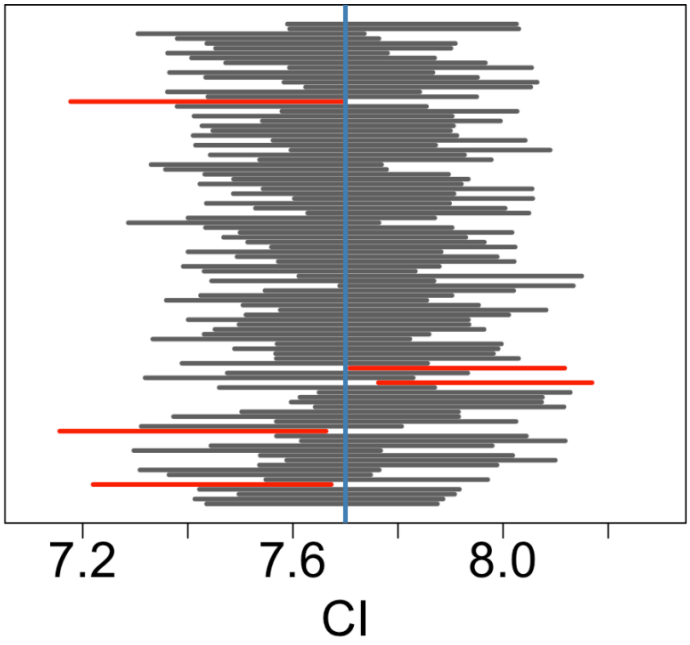
Interval estimator for the normal mean - variance known

$X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, with σ^2 known. We want to construct an interval estimator for μ.

- $\overline{X_n}$ is an unbiased estimator for μ. It is also consistent and it's the MLE (with all its properties)
- We know that $\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
- Then, $P\left(-z_{\alpha/2} \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$
- Equivalently, $P\left(\mu \in \left(\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$
- $\left(\overline{x_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is a $100(1-\alpha)\%$ confidence interval for μ
- Common confidence levels are 95% and 99%

Interval estimator for the normal mean - variance known

One hundred 95% confidence intervals for $\mu = 7.7$

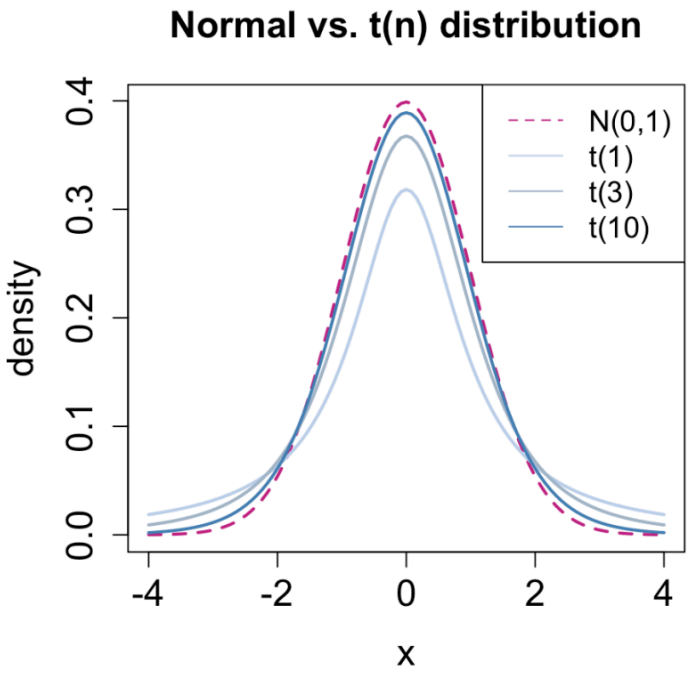


Confidence interval for the normal mean - variance unknown

$X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ with σ^2 unknown. We want an interval estimator for μ.

- In most situations the σ^2 is not known, so we cannot use the interval estimator above
- Idea: estimate σ^2 by the sample variance S_n^2 (and σ by S_n)
- $\frac{\overline{X_n} - \mu}{S_n/\sqrt{n}}$ is not normal; it has a **t-distribution** $n-1$ **degrees of freedom**, $t(n-1)$
- The pdf of a $t(n)$ distribution is $f(x) = k_n\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$ for $-\infty < x < +\infty$, where k_n is a constant.

t-distribution



Confidence interval for the normal mean - variance unknown

We know that $\frac{\overline{X_n}-\mu}{S_n/\sqrt{n}} \sim t(n-1)$

So, $P\left(\mu \in \left(\overline{X_n} - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X_n} + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right)\right) = 1-\alpha$

$\left(\overline{x_n} - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{x_n} + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right)$ is a $100(1-\alpha)\%$ confidence interval for μ

Asymptotic interval estimator for the mean

$X_1, \ldots, X_n \overset{i.i.d}{\sim} F$ with $E[X_i] = \mu$. We want an interval estimator for μ.

- Now we don't know the distribution of X_i (it could be normal or anything else)
- The CLT states that: $\frac{\sqrt{n}(\overline{X_n}-\mu)}{\sigma} \xrightarrow{D} N(0,1)$
- A variant of the CLT applies if we replace σ with the estimate S_n: $\frac{\sqrt{n}(\overline{X_n}-\mu)}{S_n} \xrightarrow{D} N(0,1)$
- So, $P\left(\mu \in \left(\overline{X_n} - z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right)\right) \approx 1-\alpha$
- $\left(\overline{X_n} - z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right)$ is an **approximate** $100(1-\alpha)\%$ interval estimator for μ

Confidence interval for the mean - example

Let $x_1, \ldots, x_n$ be a dataset modeled as a realization of i.i.d random variables $X_1, \ldots, X_n$ with $E[X_i] = \mu$

Which interval for μ would you choose in each of these scenarios?

- $X_i \sim N(\mu, 2)$
- $X_i \sim N(\mu, \sigma^2)$
- $X_i \sim F$

Confidence interval for a proportion

$X_1, \ldots, X_n \overset{i.i.d}{\sim} Bernoulli(p)$ or equivalently $X = X_1 + \cdots + X_n \sim Binomial(n, p)$

- $\hat{p} = \overline{X_n} = \frac{X}{n}$ is the MLE of p
- $Var(\hat{p}) = \frac{p(1-p)}{n}$; $SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
- A natural estimator of $SD(\hat{p})$ is $\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; standard error of $\hat{p}$
- By the CLT we have $\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \approx N(0,1)$
- By a variant of the CLT we have $\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \approx N(0,1)$
- So, $\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ is an approximate $100(1-\alpha)\%$ interval estimator for p
- Can also write as: $(\hat{p} - z_{\alpha/2}\hat{se}(\hat{p}), \hat{p} + z_{\alpha/2}\hat{se}(\hat{p}))$

Confidence interval for a proportion

Example: $n = 1,000$ poll showed that 557 voters intend to vote for candidate A and 443 for candidate B.

p proportion of all voters that intend to vote for A
$\hat{p} = 557/1000 = 0.557 = 55.7\%$

$\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{0.557(1-0.557)/1000)} = 0.0157$

95% CI for p: $(\hat{p} - 1.96 \times \hat{se}(\hat{p}), \hat{p} + 1.96 \times \hat{se}(\hat{p})) = (0.526, 0.588)$

Margin of error $= 1.96 \times \hat{se}(\hat{p}) = 0.03$

"This poll is accurate to plus or minus 3% points 19 times out of 20"

Asymptotic variance of the MLE

Let $X_1, \ldots, X_n \sim f(x;\theta)$ and the log-likelihood function: $l(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta)$

The **Fisher's information** of θ (for one observation) is defined as:
$I_1(\theta) = -E\left[\frac{\partial^2 l_1(\theta,X)}{\partial\theta^2}\right]$

where $l_1(\theta) = \log f(X;\theta)$ is the log-likelihood for one observation

If $f(x;\theta)$ is 'well behaved' (e.g. its support does not depend on θ)

- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta)}\right)$
- $\frac{1}{nI_1(\theta)}$ is the asymptotic variance of $\hat{\theta}$
- The standard error of the MLE is $se(\hat{\theta}) = \frac{1}{\sqrt{nI_1(\theta)}}$
- We can estimate the standard error of the MLE by $\widehat{se}(\hat{\theta}) = \frac{1}{\sqrt{nI_1(\hat{\theta})}}$

Asymptotic confidence intervals based on MLES

Estimate of the standard error (i.e. $sd(\hat{\theta})$):
$\hat{se}(\hat{\theta}) = \frac{1}{\sqrt{nI_1(\hat{\theta})}}$

Based on the asymptotic normality of the MLE an approximate $100(1-\alpha)\%$ confidence interval can be computed as:

$$\left(\hat{\theta} - z_{\alpha/2}\hat{se}(\hat{\theta}), \hat{\theta} + z_{\alpha/2}\hat{se}(\hat{\theta})\right) = \left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI_1(\hat{\theta})}}, \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI_1(\hat{\theta})}}\right)$$

Asymptotic confidence intervals based on MLES

Example: $X_1, \ldots, X_n \sim Geometric(p)$; $P(X = x) = p(1-p)^{x-1}$

$L(p) = \prod_{i=1}^n p(1-p)^{x_i - 1} = p^n(1-p)^{\sum_{i=1}^n X_i - n}$

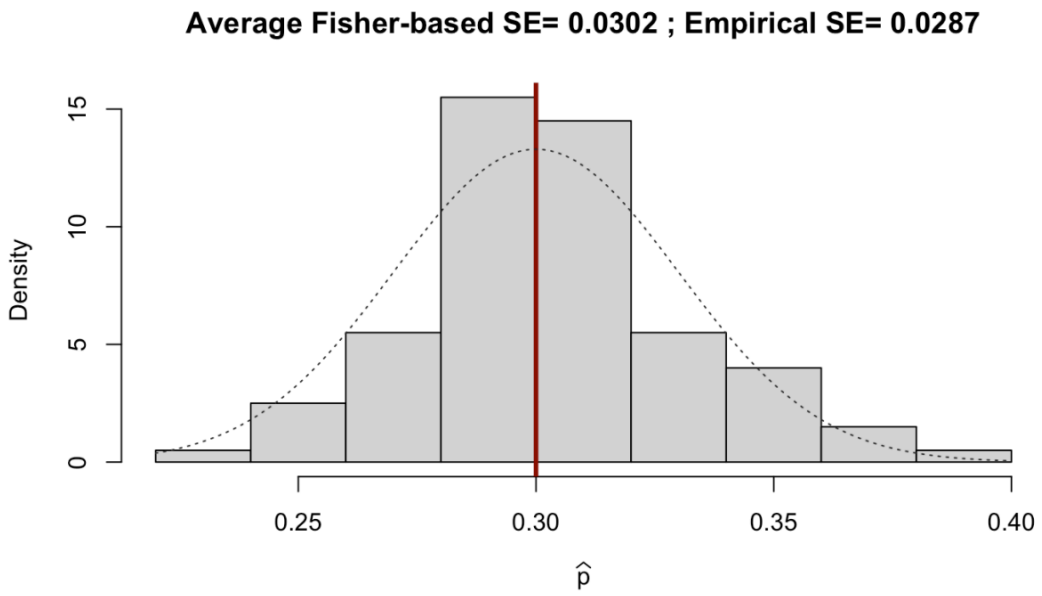
$l(p) = n\log(p) + \left(\sum_{i=1}^n X_i - n\right)\log(1-p)$

$l'(p) = \frac{n}{p} - \frac{\sum_{i=1}^n X_i - n}{1-p} = 0 \implies p = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\overline{X}_n}$

So, MLE of p is $\hat{p} = \frac{1}{\overline{X}_n}$

Asymptotic confidence intervals based on MLES

```
set.seed(1909)
nsims = 100; n = 70; p = 0.3
p_hat = replicate(nsims, {X = rgeom(n, p) + 1; 1/mean(X)})
```



Asymptotic confidence intervals based on MLEs

$l_1''(p) = -\frac{1}{p^2} - \frac{X-1}{(1-p)^2}$ ($l_1(p)$ is log-likelihood for a single observation X)

Fisher information for one observation:

$I_1(p) = -E[l_1''(p)] = \frac{1}{p^2} + \frac{\frac{1}{p}-1}{(1-p)^2} = \frac{1}{p^2(1-p)}$

$se(\widehat{p}) = \sqrt{(nI_1(p))^{-1}} = \frac{p\sqrt{1-p}}{\sqrt{n}}$

$\widehat{se}(\widehat{p}) = \frac{\widehat{p}\sqrt{1-\widehat{p}}}{\sqrt{n}}$

$\left(\widehat{p} - z_{\alpha/2}\frac{\widehat{p}\sqrt{1-\widehat{p}}}{\sqrt{n}}, \widehat{p} + z_{\alpha/2}\frac{\widehat{p}\sqrt{1-\widehat{p}}}{\sqrt{n}}\right)$

Asymptotic confidence intervals based on MLEs

Example $X_1, \ldots, X_{70} \sim Geometric(p), \overline{X} = 3.714$

$\widehat{p} = 0.269$

$\widehat{se}(\widehat{p}) = \frac{\widehat{p}\sqrt{1-\widehat{p}}}{\sqrt{70}} = 0.0275$

$\left(\widehat{p} - z_{\alpha/2}\widehat{se}(\widehat{p}), \widehat{p} + z_{\alpha/2}\widehat{se}(\widehat{p})\right) =$

$= (0.269 - 1.96 \times 0.0275, \quad 0.269 + 1.96 \times 0.0275) = (0.220, 0.318)$