Chapter 2: Probability

- Why Statistics, Probability?
- Review syllabus

Sec 2.3: Set Theory

- Set: a collection of distinct mathematical objects. (Will use S to denote the set of all elements under consideration)
- Empty set $\emptyset$: A set that contains no element
- Subset: Set A is a subset of B, denoted as $A \subset B$ if every element of A is also in B. So $A \subset A$?

Question: Are these set or not?

$$A = \{1, 3, 5, 6\};\ B = \{x, e^x, \sin x\};\ C = \{2, 5, 6, 8, 2\};\ D = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$$

Operations of sets: $S = \{0, \cdots, 9\}, A = \{1, 2, 3, 4, 5\}, B = \{3, 5\}, C = \{5, 7, 9\}$

- Union: $A \cup B = \{x \in A \text{ or } x \in B\} = \{1, 2, 3, 4, 5\}$
- Intersection: $A \cap B = \{x \in A \text{ and } x \in B\} = \{3, 5\} \Rightarrow$ A and B are disjoint if $A \cap B = \emptyset$
- Complement: $\bar{A} = \{x \in S, x \notin A\} = \{0, 6, 7, 8, 9\}$

Common Laws for Sets [Prove if time allows]

Distributive law:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

DeMorgan's law: $\overline{(A \cap B)} = \bar{A} \cup \bar{B}, \qquad \overline{(A \cup B)} = \bar{A} \cap \bar{B}$

Set Theory Based Probability

- Sample space: The set of all possible sample points (possible outcomes)
- Event: a set of outcomes of an experiment (a subset of the sample space)
- Simple event: event that can't be decomposed – denoted by $E_1, E_2, \ldots,$
- Compound event: event that consists more than one sample point.
- Discrete sample space: S is finite or countable.

Axioms of Probability

A probability P is a real-valued function defined on subsets of S

1. $0 \leq P(A) \leq 1$, for all events $A \subset S$
2. $P(S) = 1$
3. If $A_1, A_2, \cdots$ form a sequence of pairwise disjoint sets (i.e. $A_i \cap A_j = \emptyset$ all i, j), then

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

If the sets are finite in number (and still disjoint), we have:

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

Basic Properties of Probability

- If $A \subset B$, then $P(A) \leq P(B)$
- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$

Proof of $A \subset B$, then $P(A) \leq P(B)$.

- First observe that $B = A \cup (B \setminus A)$ and that $A \cap (B \setminus A) = \emptyset$.
- Then, by Axiom 3:
$$P(B) = P(A) + P(B \setminus A)$$
- By Axiom 1: $P(B \setminus A) \geq 0$. Therefore:
$$P(B) = P(A) + P(B \setminus A) \geq P(A) + 0 = P(A)$$

Sec 2.5: Simple event method to Calculate Probability of an Event

Simple Event Method

- Define the sample space S and simple events (sample points) E_i
- Assign probability to simple events: Mapping $E_i \mapsto P(E_i)$
- Define the event of interest A as a specific collection of sample points:

$$A = \cup_{i=1}^{n} E_i$$

- Add up probabilities of samples to get $P(A)$ using Axiom 3:

$$P(A) = \sum_{i=1}^{n} P(E_i)$$

- If X is finite and each simple event has the same probability (they are "equiprobable"), then (prove it!)

$$P(A) = \frac{|A|}{|S|},$$

where $|\cdot|$ denotes the cardinality of a set (number of elements).

Example 1

- Select two job candidates from five who are labeled 1 to 5.
  - Simple events take the form $E_{xy} = \{x, y\}$ for candidates $x, y = 1, ..., 5$ (order does not matter)
  - $S = \cup_{x,y=1,x<y}^{5} E_{xy}$ and $|S| = \frac{5\times 4}{2} = 10$ (don't double-count!)
  - Event A: $\{1, 4\}$ or $\{1, 5\}$ are selected.

$$P(A) = \frac{|A|}{|S|} = \frac{2}{10} = .2$$

[revisit in Sec 2.6 using $|S| = \binom{n}{k} = \frac{5!}{2!3!} = 10$ where $n = 5, k = 2$]

- Select two job candidates for two distinct positions (president/VP) from the same five.
  - Simple events take the form $E_{xy} = (x, y)$ for candidates $x, y = 1, ..., 5$ (ordered pairs, $(x, y) \neq (y, x)$!)
  - $X = \cup_{x,y=1,x\neq y}^{5} E_{xy}$ and $|S| = 5 \times 4 = 20$
  - Event A: $(1, 4)$ or $(1, 5)$ are selected.

$$P(A) = \frac{|A|}{|S|} = \frac{2}{20} = .1$$

Therefore we need to learn how to count efficiently... see 2.6!

- Multiplication principle: Ordered pairs — first entry can be selected in m ways and the second entry is selected in n ways, then the ordered pair can be selected in mn ways. **e.g., rolling 2 dice:** 6×6
- Permutation: An ordered arrangement of n distinct objects $\Rightarrow$ $n!$ ways
- Partial permutation: Ordered arrangements into k-tuple $(a_1, a_2, ..., a_k)$ of n distinct objects: $P_k^n = \dfrac{n!}{(n-k)!} = n(n-1)(n-2)\cdots(n-k+1)$
- Combination: Pick k items from n distinct objects (the order does not matter) $\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$

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

  properties: $\binom{n}{0} = \binom{n}{n} = 1$ and $\binom{n}{k} = \binom{n}{n-k}$

- k-partition of n: Partition n distinct objects into k distinct groups with $n_1, \cdots, n_k$ objects:
  $\binom{n}{n_1, \cdots, n_k} = \dfrac{n!}{n_1! \cdots n_k!}$, $n_1 + n_2 + .. + n_k = n$,

  - Multinomial coefficients for $(x_1 + x_2 + ... + x_k)^n$

Proofs:

Permutation of n objects:

- Build an ordered n-tuple $(a_1, a_2, ..., a_n)$: choose a_1 then choose a_2 then $\cdots$
- There are $n \times (n-1) \times (n-2) \times \cdots 2 \times 1$ permutations (ordered, without replacement)
- $n! = n \times (n-1) \times (n-2) \times \cdots 2 \times 1$

Permutations of k objects taken from n *distinct* objects

- Ordered arrangements into k-tuple $(a_1, a_2, ..., a_k)$ of n distinct objects (*without replacement*)
- There are $n \times (n-1) \times (n-2) \times \cdots \times (n-(k-1))$ permutations
- $P_k^n = \dfrac{n!}{(n-k)!} = n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)$

  Example: From 10 persons, choose a president, a vice president, a secretary and a treasurer for a club (ordering matters!). How many possible results?

  $$P_4^{10}$$

Proof – Continued.

Combinations

- **Task:** count the number of **unordered** sets of k-element sets $\{a_1, a_2, ..., a_k\}$ chosen from n objects (*without replacement*)
- **Observe:** there are $k!$ different orderings (permutations) for each k-element sets $\{a_1, a_2, ..., a_k\}$.
- **Answer:** Therefore, the number of **unordered** k-element sets $\{a_1, a_2, ..., a_k\}$ chosen from n objects is:

$$C_k^n = \frac{P_k^n}{k!} = \frac{n!}{k!\,(n-k)!} = \binom{n}{k} = \binom{n}{n-k}$$

Example: Choose 5 students from 20 students; no order: $C_5^{20} = \binom{20}{5}$

Example: There are 10 balls, of which 3 are basketballs, and 7 are soccer balls. If we want to put them in a row, how many possible arrangements?

Suffices to distribute 3 balls (or 7) in 10 slots in a row: $C_3^{10} = \binom{10}{3} = \binom{10}{7}$ (ordering between same type of balls is irrelevant)

□

Example 2

- Toss a fair die for 6 times. What is the probability that the numbers are $1, 2, 3, 4, 5, 6$ in any order?
- Choose 3 numbers from $0 \sim 9$. What's the probability that $1, 2, 3$ are chosen? What's the probability that 0 is chosen? What's the probability that 5 is not chosen?
- **Birthday Problem(s)** Consider the experiment, called the **birthday problem**, where our task is to determine the probability that in a group of k people there are at least two people who have the same birthday (assume no leap years). Show the following (surprising?) fact: it takes $k = 23$ people to have at least two people with the same birthday with probability 50% or more.

Solution:

- Consider first the (easier to calculate) event

$$A_k = \{\text{the first } k \text{ BDs are all distinct}$$

- We want to calculate for $k = 23$

$$p_k = P(\bar{A}_k) = 1 - P(A_k)$$

Then

$$P(A_k) = \frac{|A_k|}{|S|} = \frac{365 \cdot 364 \ldots \cdot (365 - (k-1))}{365^k} = \frac{365!}{365^k (365-k)!} = \frac{P_k^{365}}{365^k},$$

- thus,

$$p_k = P(\bar{A}_k) = 1 - P(A_k) = 1 - \frac{365!}{365^k (365-k)!}$$

- For $k = 23$ we get (use $\frac{365!}{365^{23} 342!} \approx .4927$)

$$p_{23} = 1 - \frac{365!}{365^{23}(365-23)!} \approx 1 - .4927 = .5073$$

Sec 2.7: Conditional Probability and Independence

Definition 1 (Conditional Probability and Independence)

- Conditional Probability: If $P(B) > 0$, the conditional probability of A given B is

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

- Independence: Events A and B are independent if

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

Equivalent condition:

- If $P(B) > 0$, $P(A|B) = P(A)$
- If $P(A) > 0$, $P(B|A) = P(B)$

Example 3

- Two events, $P(A) = 0.5$, $P(B) = 0.3$, $P(A \cap B) = 0.1$. Find $P(A|B)$, $P(B|A)$, $P(A|A \cup B)$.
- Toss two dice, probability of seeing one on the second if we know the number on the first is odd.
- Suppose $A \subset B$ and $P(A) > 0$. Are A and B independent?
- If $P(A) > 0$, $P(B) > 0$, and $P(A) < P(A|B)$. Show that $P(B) < P(B|A)$.

Definition 2

Events $A_1, \cdots, A_n$ are mutually independent if

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} A_{i_j}, \qquad \text{for any subsets } A_{i_1}, \ldots, A_{i_k}, \quad k \leq n,$$

Remark 1

Pairwise independence does NOT imply mutual independence: (Bernstein's example)

- X and Y are two independent tosses of a fair coin
- Let Z be equal to 1 if exactly one of those coin tosses resulted in "heads", and 0 otherwise. Then

$$(X, Y, Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4 \\ (0,1,1) & \text{with probability } 1/4 \\ (1,0,1) & \text{with probability } 1/4 \\ (1,1,0) & \text{with probability } 1/4 \end{cases}$$

Sec 2.8: Laws of Probability I

- Multiplicative law: $P(A \cap B) = P(A)P(B|A)$
- Additive law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Corollary: $P(A) = 1 - P(\bar{A})$.

Example 4

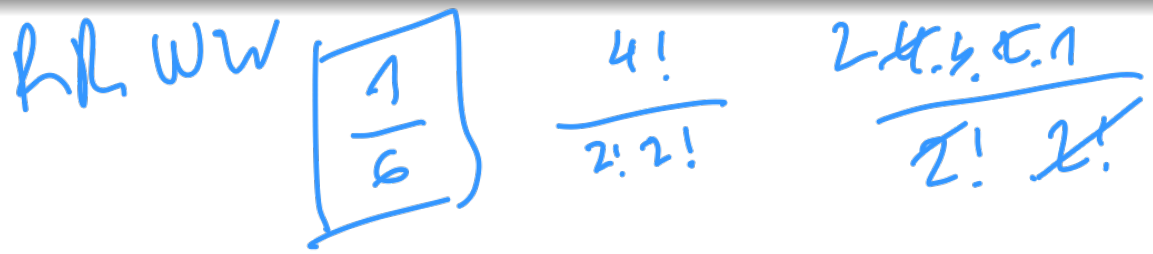
- Two events A and B with $P(A) = 0.2$, $P(B) = 0.3$, and $P(A \cup B) = 0.4$. Find $P(A \cup B)$, $P(\bar{A} \cup \bar{B})$, $P(\bar{A} \cap \bar{B})$, $P(\bar{A} \mid B)$.
- A and B are independent with $P(A) = 0.5$ and $P(B) = 0.2$. What is $P(A \cup B)$?

Example 5

Suppose an urn contains 8 red balls and 4 white balls.

- Draw two balls without replacement. What is the probability that both are white?
- Draw fall balls without replacement. What is the probability that the first two draws are red and the next two draws are white?

Counting or use conditional probability (W_1, W_2, etc.) Answer: $1/11$; $(8/12) \cdot (7/11) \cdot (4/10) \cdot (3/9) = 28/495$.



Theorem 1 (Suppose $A_1, A_2, \ldots, A_n$ are events, then we have)

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

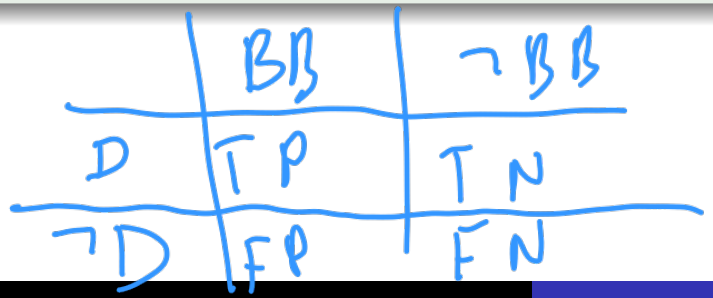
Case 1: 2 heads :

Sec 2.9: Event-composition Method

- Identify sample space, sample point, and event of interest, denoted by A
- Express the event of interest as composition of several events
- Use law of probability to find $P(A)$

Example 6

- Toss three fair coins. What is the probability of seeing ≥ 2 heads?
- 10 students apply for a graduate program. Randomly select 2. What is the probability student A is chosen? The probability both A and B are chosen? The probability that at least one of A and B are chosen? The probability that exactly one of A or B are chosen?
- Randomly toss three balls red, blue, and green to three boxes with color red, blue, and green. What is the probability that there are no color matches? What is the probability that there is exactly one color match?



A partition $\{B_1, \cdots, B_N\}$ of the sample space S is a collection of subsets $B_1, \cdots, B_N$ such that

$$S = \bigcup_{i=1}^{N} B_i \qquad \text{and} \quad B_i \cap B_j = \emptyset, \quad i \neq j$$

- Multiplicative law: $P(A \cap B) = P(A)P(B|A)$
- Additive law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Law of total probability: Let $\{B_1, \cdots, B_N\}$ be a partition of S, then

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

- Bayes' law: $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$, or

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}$$

when $\{B_1, \cdots, B_N\}$ is a partition of sample space S

Example 7 (Witness Reliability)

After a robbery the thief jumped into a taxi and disappeared. An eyewitness on the crime scene is telling the police that the cab is yellow. Is this testimony worth something?

Solution:

The assistant DA makes a Bayesian analysis of the situation:

- After some research, DA comes up with following info
  - In that city 70% of taxis are black and 30% of taxis are yellow.
  - Eyewitness are not always reliable and from past experience, it is expected that an eyewitness is 80% accurate – identify the color of a taxi accurately (yellow or black) 8 out 10 times
- Equipped with this information assistant. DA defines adequate events:
  - $true = Y \quad \Rightarrow$ color of the taxi was actually yellow
  - $true = B \quad \Rightarrow$ color of the taxi was actually black
  - $report = Y \quad \Rightarrow$ eyewitness identified the taxi as yellow
  - $report = B \quad \Rightarrow$ eyewitness identified the taxi as black

Goal: compute $P(true = Y | report = Y)$

Solution – Continued

Using Bayes formula

But $$P(true = Y|report = Y) = \frac{P(report = Y|true = Y) \cdot P(true = Y)}{P(report = Y)}$$

- $P(true = Y) = .3$: the prior probability
- $P(report = Y|true = Y) = .8$: the accuracy of witness testimony
- Now compute $P(report = Y)$. DA argues that this depends on the actual color of the taxi, so using conditional probability

$$\begin{aligned} P(report = Y) &= P(report = Y|true = Y) \cdot P(true = Y) \\ &\quad + P(report = Y|true = B) \cdot P(true = B) \\ &= 0.8 \times 0.3 + 0.2 \times 0.7 = 0.38 \end{aligned}$$

Therefore:

$$P(\text{true} = Y|\text{report} = Y) = \frac{P(report = Y|true = Y)P(true = Y)}{P(report = Y)} = \frac{.24}{.38} = 0.63$$

Example 8 (The Monty's Hall Problem)

At a game show, the host hides a prize (say \$1 million) behind one of three doors and nothing of much value behind the other two doors (in the usual story two goats). The contestant picks one of three doors, let us say door 1, and then the host opens one of the remaining doors, let us say he opens door 3 which reveals a goat. The contestant is then given the choice to either switch to door 2 or keep door 1. What should he do?

Solution:

We will argue that he should switch to door 2 since there is a greater probability to find the prize behind door 2 than behind door 1.

- The trick is to carefully make assumptions precise
- The \$1 million is put randomly behind any door $\Rightarrow$, the contestant, upon choosing a door, has probability $1/3$ to find the prize
- The host show knows behind which doors the prize is and always opens an empty door. If he has two empty doors he can open any of the two.

There are many ways to find the solution and we will present two:

Solution – Continued

Solution 1: Name two goats G1, G2. The following arrangements are possible

Door 1	Door 2	Door 3
P	G1	G2
P	G2	G1
G1	P	G2
G2	P	G1
G1	G2	P
G2	G1	P

- The prize are arranged randomly $\Rightarrow$ assume all 6 arrangements have equal probabilities 1/6.
- WLOG, assume the contestant opens door 1. In the 1st case, the contestant has the prize behind his door and the host will open either door 2 or door 3 in which case the contestant will lose if he switches door. The second case is similar.
- In the 3rd case the host will open door 3 and the contestant will win if he switches door.
- The fourth, fifth and sixth case are similar. So in 2 out of 6 case you win by keeping your door while in 4 out of 6 cases you win by switching. Hence the probability to win if you switch is 2/3. So you should switch.

Solution – Continued

Solution 2: Bayesian approach. WLOG, assume the contestant has chosen door 1. Then

$$\begin{aligned} P(\text{keep and win}) &= P(\text{prize door 1}|\text{host door 2}) \cdot P(\text{host door 2}) \\ &\quad + P(\text{prize door 1}|\text{host door 3}) \cdot P(\text{host door 3}) \end{aligned}$$

But

$$P(\text{ prize door 1 }|\text{host door 2}) = \frac{P(\text{ host door 2}|\text{ prize door 1})P(\text{prize door 1})}{P(\text{host door 2})}$$

Therefore

$$\begin{aligned} P(\text{keep and win}) &= P(\text{ host door 2}|\text{ prize door 1})P(\text{prize door 1}) \\ &\quad + P(\text{ host door 3}|\text{ prize door 1})P(\text{prize door 1}) \\ &= \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{3}. \end{aligned}$$

Example 9

Joe chooses a 6-side or a 8-side die randomly, rolls it, and tells you the number. Given the number, what is the probability that he rolls a 6-side die?

Example 10 (Medical Testing)

You are given the following information. (a) In random testing, you tested positive for a disease. b) This test has 5% false positive test rate and 10% false negative test rate; c) In the population at large, one person in 1,000 has the disease. What is the probability that you have the disease?

- A : event one person has the disease; B : event test results are positive.
- Need to compute $P(A|B) = ??$
- Explain the meaning of false positive/negative
- Explain the meaning of the probability!!

Example 11 (Defective Fuse)

5 production line (with equal capacity) produce fuses with 2% defective rate. Products in each line were put into a box of 100 fuses. The boxes are then randomly mixed and sold to the customers. In March, line 1 produces fuses with defective rate 5%. Assume a customer purchased a box of fuses manufactured in March. He randomly tested three fuses, and one is defective. What is the probability that they are produced by line 1.

- A: The box bought by the customer was manufactured from line 1
- B: One of three tested fuses is defective

Solution:

Need to find $P(A|B)$:

- Note that 2% defective rate, means EVERY fuse has 2% chance being defective !!
- Bayes' Law: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$
- Clearly, $P(A) = 1/5$. And $P(B|A) = \binom{3}{1} \cdot 0.05 \cdot 0.95^2 \simeq 0.1354$
- Also, $P(\bar{A}) = 4/5$. And $P(B|\bar{A}) = \binom{3}{1} \cdot 0.02 \cdot 0.98^2 \simeq 0.0576$

Solution – Continued

- Hence

$$P(B) = P(B|A)\cdot P(A)+P(B|\bar{A})\cdot P(\bar{A}) \simeq 0.2\cdot 0.1354+0.8\cdot 0.0576 \simeq 0.0732$$

- Finally: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)} \simeq \dfrac{0.1354 \cdot 0.2}{0.0732} \simeq 0.37$

Example 12

5 bowls are labeled 1,2,3,4,5. Bowl i contains i red balls and $5 - i$ blue balls. Randomly select a bowl and choose two balls randomly without replacement.

- What is the probability that both balls are red?
- What is the probability that bowl 3 was selected if both balls are red?

Sec 2.11: Random Variable

- Events that are identified by numbers are called numerical events.
- A random variable is a real-valued function for which the domain is a sample space

$$X: \quad S \to \mathcal{R}$$

Examples:

- Number on a die.
- Numbers of heads when toss 4 coins.
- Stock price.
- Life span of car battery

Sec 3.1-3.2: Discrete r.v.s and Their Probability Distributions

Definition 3 (Discrete Random Variable)

A random variable is discrete if it can only take a countable number of values.

Example: We have 10 balls: 5 red and 5 blue. Choose 2 balls randomly. X: Number of red balls.

- Sample with replacement.
- Sample without replacement.

Definition 4 (Probability Distribution of Random Variable)

X is a **discrete** random variable with range E. $X : S \to E$. The probability distribution (also called probability function) of X is defined as

$$p(x) = P(X = x) \qquad \text{for all } x \in E$$

ie, $p(x)$ = sum of probabilities of all sample points in S that are assigned value x.

Example:

- Toss a fair coin 4 times. X = number of heads
- Choose two balls from 3 red and 3 blue. X = number of red balls.

Theorem 2

For any probability function $p = p(x)$ we have

$$0 \leq p(x) \leq 1\,, \quad \textit{for all possible values } x \in E \textit{ of the random variable } X$$

and and

$$\sum_{x \in E} p(x) = 1$$

Example 13

Uniform distribution (e.g. rolling a FAIR die for $n = 6$)

$$p(x) = \begin{cases} \frac{1}{n}\,, & x \in E = \{1, 2, ...n\} \\ 0, \text{otherwise} \end{cases}$$

Definition 5 (Expectation or Expected Value)

Let X be a discrete random variable with probability distribution $p(x)$.

$$\mathbb{E}[X] \doteq \sum_{x \in E} x\, p(x)\,.$$

Example: 10 balls with 5 red and 5 blue. Randomly choose 5. $X =$ number of red balls. Find $\mathbb{E}[X]$

Expectation of Function of Random Variable

- Let $g : E \to F$ be a function on the range (value space) of X.
- $g(X)$ is a random variable that takes value on F.

$$\mathbb{E}[g(X)] = \sum_{x \in E} g(x)p(x)$$

Definition 6 (Variance and Standard Variation)

X is a random variable with mean $\mathbb{E}(X) = \mu$. The variance of X is

$$\sigma^2 = Var(X) = \mathbb{E}[(X - \mu)^2]$$

The standard deviation is the positive square root of $Var(X)$ – usually denoted as σ

$Var(X_1) = \frac{1}{6}(1 - \frac{7}{2})^2 + \frac{1}{6}(2 - \frac{7}{2})^2 + \frac{1}{6}(3 - \frac{7}{2})^2 + \ldots + \frac{1}{6}(6 - 7/2)^2$

Example: coin toss. Head: $+1$, Tail: -1.

Properties of Variance and Expectation

- $\mathbb{E}[c] = c$
- $\mathbb{E}[cg(X)] = c\mathbb{E}[g(X)]$
- $\mathbb{E}[g_1(X) + g_2(X)] = \mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)]$
- $Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \to \mu^2$
- $Var[cX + b] = c^2 Var[X]$

Example 14 (Distribution with Finite Expectation but Infinite Variance)

$$P[X = \sqrt{n}] = \frac{1}{n(n+1)}, \quad n = 1, 2, \ldots$$

Why Expected value & Variance?

- Will explore this through examples (next)
- They are important because they can provide **guarantees** for "tail" probabilities (that may be hard to compute otherwise).

  See the probabilistic inequalities (we discuss them in Sec 3.11/4.11 and in Chapter 7):

Theorem 3

Markov Inequality: *If* $X \geq 0$, *then for all* $a > 0$,

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

Chebyshev Inequality: *For any random variable* X, *and any* $\epsilon > 0$,

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

Proof:

Markov inequality:

$$\begin{aligned} E[X] &= \sum_{\alpha} \alpha P(X=\alpha) = \sum_{\alpha \geq a} \alpha P(X=\alpha) + \sum_{\alpha < a} \alpha P(X=\alpha) \\ &\geq \sum_{\alpha \geq a} \alpha P(X=\alpha) \geq a P(X \geq a) \end{aligned}$$

Chebyshev inequality: Apply Markov's inequality to RV: $Y = (X - E[X])^2$

- $E[Y] = \text{var}(X)$
- $P(|X - E[X]| \geq \epsilon) = P(|X - E[X]|^2 \geq \epsilon^2) = P(Y \geq \epsilon^2) \leq \frac{E[Y]}{\epsilon^2} = \frac{\text{var}(X)}{\epsilon^2}$

Sec 3.4 - 3.8: Common Discrete Random Viables and Their PDFs

- Discrete Uniform
- Bernoulli (coin tossing)
- Binomial
- Geometric
- Poisson
- Hypergeometric, negative binomial, etc

Definition 7 (Discrete Uniform)

Defined for $k = 1, 2, \ldots, n$:

$$p(k) = \begin{cases} \frac{1}{n}, & k \in E = \{1, 2, \ldots, n\} \\ 0, & \text{otherwise} \end{cases}$$

- Expected value (mean): $\mathbb{E}[X] = \frac{n+1}{2}$
- Variance: $\mathrm{Var}(X) = \frac{n^2-1}{12}$

Definition 8 (Bernoulli)

Defined for $k = 0, 1$:

$$p(k) = p^k (1-p)^{1-k}$$

- $p : 0 \leq p \leq 1$ (single) parameter of the model.
- $P(X = 1) = p, \quad P(X = 0) = 1 - p.$
- Expected value (mean): $\mathbb{E}[X] = p$
- Variance: $\mathrm{Var}(X) = p(1-p)$

$\begin{cases} P(X=0) = 1-p \\ P(X=1) = p \end{cases}$

$\mu = 0(1-p) + 1 \cdot p = p$

$\sigma^2 = E[X^2] - \mu^2 = [0^2(1-p) + 1^2 p] - p^2 = p - p^2$

Definition 9 (Binomial Distribution)

- The probability of success of each trial is p.
- A binomial distribution $B(n, p)$ with parameters n and p is the number of successes in a sequence of n independent trials.
- The probability distribution of $B(n, p)$ is

probability landing h/t
k: event get exact k times h/t

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

Example 15

Each coin has probability $p = 0.1$ landing on Head. Toss 10 such coins. What are the probabilities of seeing 0, 1, and 2 heads?

Theorem 4

Let X be a binomial distribution $B(n, p)$, then

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

$$E[X] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} = np \sum_{k-1}^{n-1} \binom{n-1}{k-1} p^{n-1-(k-1)} (1-p)^{n-1-(k-1)}$$

$$= np \cdot (p + (1-p))^{n-1}$$

Example 16 ("Post-processing" internet traffic)

n packets go through a router every day. With probability p, each packet is selected/stored for "post-processing/monitoring". The process is randomized to remain unpredictable. Then clearly we have a binomial distribution at play:

$$X: \ \#\text{ packets selected} \sim B(n,p)$$

Decision under uncertainty: With a fixed amount of memory and traffic, how often can I sample the incoming packets?

- each packet 10KB; storage 10^9KB
- An initial answer:

$$\underbrace{\mathbb{E}(X)}_{\text{average \# of selected packets}} \cdot 10 = 10^9, \quad \mathbb{E}(X) = np$$

- However in reality, I need to somehow ensure that $10X \leq 10^9$ which is not guaranteed by $\mathbb{E}(X)$ (being an average!).
- A better question: what is the highest p for which I can guarantee $10X \leq 10^9$ with a mutually agreed high probability?

$$P(X \leq 10^8) \leq 95\% \ ??$$

Markov inequality (very crude for Binomials-see Sec 3.11!)

$$P(X > 10^8) \leq \frac{\mathbb{E}(X)}{10^8} = \frac{np}{10^8}$$

Therefore I need $p \approx 5\% \cdot 10^8/n$

$\frac{\mathbb{E}(X)}{10^8} = 5\% \ (=) \ \frac{np}{10^8} = \frac{5}{100}$

Definition 10 (Geometric Distribution)

- Assume the probability of seeing an Head is p when toss a coin. Let $X =$ number of tosses needed to see the first Head.

$$P(X = 5) = (1 - p)^4 p$$

- A random variable X is a geometric probability distribution if and only if

$$p(n) = (1 - p)^{n-1} p$$

Theorem 5

Let X be a geometric distribution with parameter p, then

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

Proof.

Use derivative □

Theorem 6 (Memoryless Property)

Let X be a geometric distribution, then

$$P(X > m + n | X > m) = P(X > n)$$

Given that the first success has not yet occurred, the probability of the number of additional trials does not depend on how many failures have been observed.

Poisson Distribution

What is the distribution of λ, the number of particles emitted by a radioactive material in a second?

- Divide a second into N intervals – small enough s.t. there is at most one particles emitted
- Assume each interval is independent, and has probability p of emitting one particle
- Then we have a binomial distribution:

$$f(x) = \binom{N}{x} p^x (1-p)^{N-x}, \qquad x \text{ is number of particles emitted}$$

- Let $\lambda = Np$. Then

$$\begin{aligned}
\lim_{N\to\infty} f(x) &= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-x+1)}{x!} \left(\frac{\lambda}{N}\right)^x \left(1-\frac{\lambda}{N}\right)^{N-x} \\
&= \frac{\lambda^x}{x!} \lim_{N\to\infty} \left(1-\frac{\lambda}{N}\right)^N \left(1-\frac{\lambda}{N}\right)^{-x} (1-\frac{1}{N})\cdots\left(1-\frac{x-1}{N}\right) \\
&= \frac{\lambda^x}{x!} e^{-x}
\end{aligned}$$

Definition 11 (Poisson Distribution)

A random variable X is said to have a Poisson distribution if and only if

$$p(n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

Remark 2

- Poisson distribution is the limit of binomial distribution
- Poisson distribution is used to express the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.
- Example: number of clients per hour; number of transactions per second; number of accidents per week...

Theorem 7

Let X be a Poisson distribution with parameter λ, then

$$\mathbb{E}[X] = \lambda, \quad Var(X) = \lambda\,.$$

Example 17

If X is a Poisson distribution with $p(0) = p(1)$. What is $p(2)$?

Example 18 (Approximate Binomial Distribution with Poisson)

Assume X has a binomial distribution with $n = 20, p = 0.1$.

- Exact value $P(X \leq 3) \sim 0.867$
- Y has Poisson distribution with $\lambda = np = 2$. And $P(Y \leq 3) \sim 0.857$

Example 19

- Industrial accidents occur according to a Poisson process, average 3 per month
- During the last two months, 10 occurred. Does it indicate an increase in the average accidents per month?

Analysis:

- Number of accidents per TWO months, X, also has Poisson distribution, but with $\lambda^* = 2 \cdot 3 = 6$
- Mean: $\mu = \lambda^* = 6$; Standard deviation: $\sigma = \sqrt{6} \sim 2.45$
- Empirical rule $\Rightarrow$ Y take values in interval $\mu \pm 2\sigma$ with high probability
- $\mu + 2\sigma = 10.9 \Rightarrow$ Observed # is WITHIN the interval, but close to boundary
- Not highly improbable, but may be enough to warrant an investigation.
- Use the table ...

Definition 12 (Hypergeometric Distribution)

- There are N balls: r red, $N-r$ white. Choose n from N without replacement. Denote $X =$ number of red balls.
- X is said to be a hypergeometric distribution with parameters N, r, and n if and only if

$$p(k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}.$$

Theorem 8

Let X be a hypergeometric distribution with parameters N, r, n. Then

$$\mathbb{E}[X] = \frac{nr}{N}, \quad Var(X) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

$$\frac{2^{k-1}}{2^k - 1} =$$

Example 20

- An industrial product is shipped in boxes of 20
- Sample to minimize the number of defectives shipped to customers: test 5 out of a box and reject the box if more than one is defective
- If a box has 4 defectives, what is the probability it will be rejected?

It has hypergeometric distribution with $N = 20, n = 5, r = 4$

$$\begin{aligned} P(\text{reject}) &= P(X \geq 2) = p(2) + p(3) + p(4) = 1 - p(0) - p(1) \\ &= 1 - \frac{\binom{4}{0}\binom{16}{5}}{\binom{20}{5}} - \frac{\binom{4}{1}\binom{16}{4}}{\binom{20}{5}} \simeq 0.2487 \end{aligned}$$

Remark 3

Hypergeometric distribution is very close to binomial distribution when N is very large, but r/N held constant: (let $p = r/N$)

$$\lim_{N\to\infty} \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} = \binom{n}{k} p^k (1-p)^{n-k}$$

Why?!

Definition 13 (Negative Binomial Distribution)

- Toss coins, p is the probability of getting Head for each toss. $X =$ number of tosses needed for the r-th head:

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

- A random variable X is said to have a negative binomial probability distribution if and only if

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

Theorem 9

If X is a random variable with a negative binomial distribution, then

$$\mathbb{E}[X] = \frac{r}{p}, \quad Var[X] = rqp^{-2}$$

Example 21

If X is a binomial distribution $B(20, 0.5)$. What is the probability that $5 < X < 15$? Use the table.

Definition 14 (Joint Probability Function for Discrete Random Variables)

Let X_1 and X_2 be two discrete random variables. The joint probability function for X_1 and X_2 is given by

$$p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

If $p(x_1, x_2)$ is a joint probability function, then $p(x_1, x_2) \geq 0$ for all x_1, x_2, and

$$\sum_{x_1, x_2} p(x_1, x_2) = 1\,.$$

Example 22

Toss two fair coins. $H = 1, T = 0$ for each coin. Define $X_1 =$ number on the first coin, and $X_2 =$ number on the 2nd coin. What is the joint probability function of (X_1, X_2)?

Example 23

Toss two fair coins. $H = 1, T = 0$ for each coin. Define $X_1 =$ number on the first coin, and $X_2 =$ total number of heads. What is the joint probability function of (X_1, X_2)?

Definition 15 (Joint Distribution Function)

Let X_1 and X_2 be two discrete random variables. The joint distribution function for X_1 and X_2 is defined as

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$$

Theorem 10

The joint distribution function $F(x_1, x_2)$ has the following properties:

- $F(-\infty, -\infty) = F(-\infty, x_2) = F(-\infty, x_1) = 0$
- $F(\infty, \infty) = 1$
- *If $\hat{x}_1 > x_1$ and $\hat{x}_2 > x_2$, then*

$$F(\hat{x}_1, \hat{x}_2) - F(\hat{x}_1, x_2) - F(x_1, \hat{x}_2) + F(x_1, x_2) \geq 0$$

Proof.

Part 3:

$$F(\hat{x}_1, \hat{x}_2) - F(\hat{x}_1, x_2) - F(x_1, \hat{x}_2) + F(x_1, x_2) = P(x_1 < X_1 \leq \hat{x}_1, x_2 < X_2 \leq \hat{x}_2)$$

□

Definition 16 (Marginal Probability Function)

Let X_1 and X_2 be jointly discrete random variables with probability function $p(x_1, x_2)$. The marginal probability function of X_1 and X_2, respectively, are given by

$$p_1(x_1) = \sum_{x_2} p(x_1, x_2), \quad p_2(x_2) = \sum_{x_1} p(x_1, x_2)$$

Example 24

- Toss two fair coins. $H = 1$, $T = 0$. And $X_1 =$ number on the first coin, $X_2 =$ number of the sum. What is the joint probability function?
- If we only observe X_2, what is the distribution of X_2? Does it equal to the marginal from above?

Definition 17 (Conditional Probability Function)

If X_1 and X_2 are jointly discrete random variables with joint probability function $p(x_1, x_2)$ and marginal probability functions $p_1(x_1)$, $p_2(x_2)$, respectively, then the conditional discrete probability function of X_1 given X_2 is

$$p(x_1|x_2) = P(X_1 = x_1 \,|\, X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} = \frac{p(x_1, x_2)}{p_2(x_2)}$$

Marginal + conditional $\Rightarrow$ joint: $p(x_1, x_2) = p_2(x_2) \cdot p(x_1|x_2)$

Definition 18 (Independent)

Let X_1 have distribution function $F_1(x_1)$, X_2 have distribution function $F_2(x_2)$, and X_1, X_2 have the joint distribution function $F(x_1, x_2)$. Then X_1 and X_2 are independent if and only if

$$F(x_1, x_2) = F(x_1)F(x_2)$$

for every pair of x_1, x_2.

Recall that $F(a_1, a_2) = P(X_1 \leq a_1, X_2 \leq a_2) = \sum_{x_1 \leq a_1} \sum_{x_2 \leq a_2} p(x_1, x_2)$

Theorem 11

If X_1 and X_2 are jointly discrete random variables with joint probability function $p(x_1, x_2)$ and marginal probability functions $p_1(x_1)$, $p_2(x_2)$, respectively, then X_1 and X_2 are independent if and only if

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \qquad \text{for every pair of } x_1, x_2.$$

Example 25 (Mixture Models)

Denote X as the number of customers that arrive in a store tomorrow. If the day is rainy X is Poisson with the expected (mean) number of customers per hour to be 20. If the day is sunny X is also Poisson with the expected number of customers to be 40. The probability for rain tomorrow is 30%. **Question:** a:) What is the probability the store will have more than 35 customers?

- [Recall that if $Y \sim \text{Poisson}(\lambda)$, then $P(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ and $E(Y) = \lambda$]
- Define **Bernoulli** random variable R for a rainy day ($R = 1$) or not rainy ($R = 0$). We have: $P(R = 1) = .3$ and $P(R = 0) = .7$.
- $X \sim \text{Poisson}(\lambda)$, $\lambda = 20$ customers ($R = 1$) or $\lambda = 40$ ($R = 0$). Thus

$$P(X = k|R = 1) = e^{-20}\frac{20^k}{k!} \quad \text{and} \quad P(X = k|R = 0) = e^{-40}\frac{40^k}{k!}$$

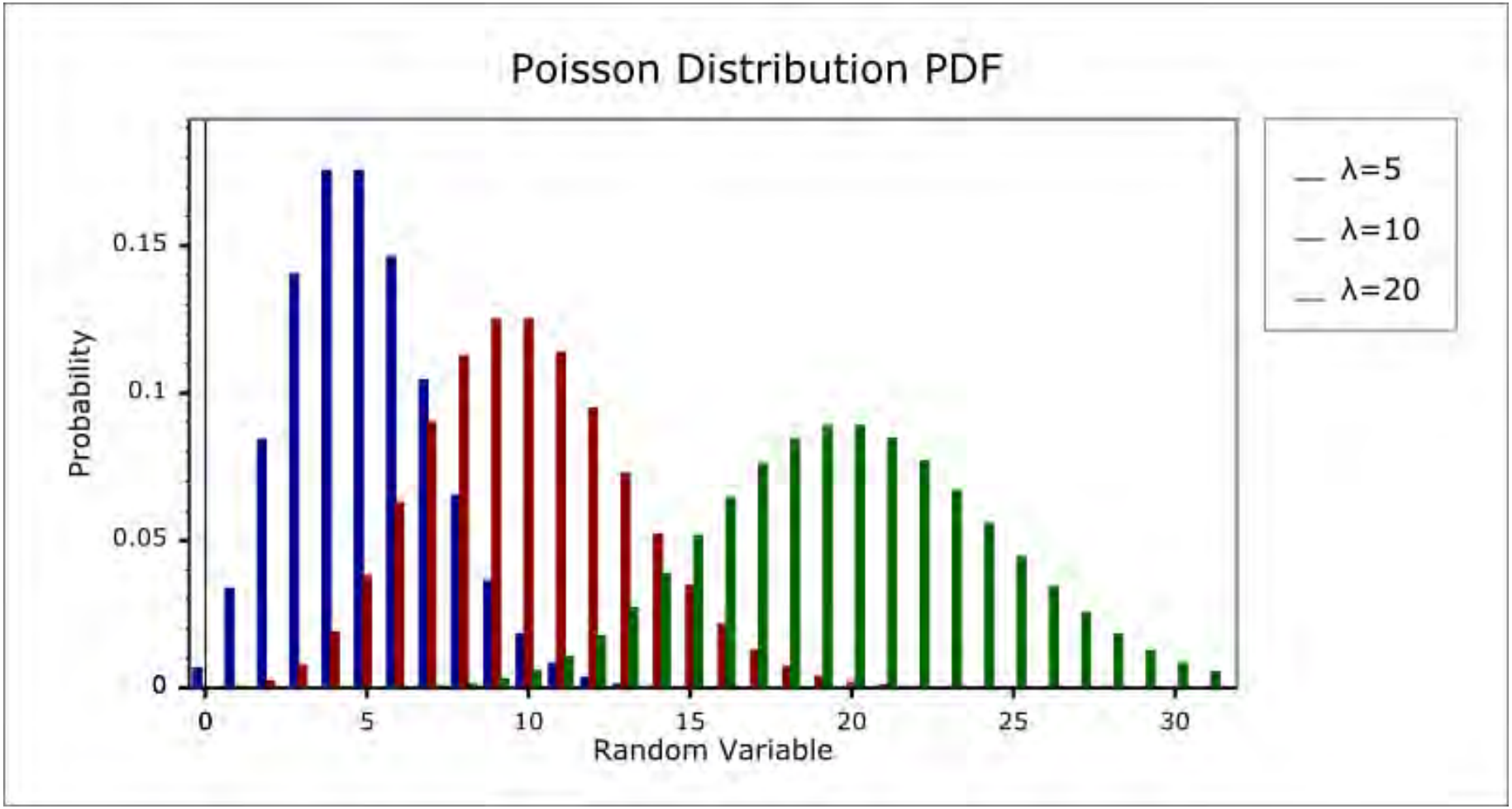
- **Marginal** in k:

$$\begin{aligned} P(X = k) &= P(X = k, R = 1) + P(X = k, R = 0) \\ &= P(X = k|R = 1)P(R = 1) + P(X = k|R = 0)P(R = 0) \end{aligned}$$

- Therefore we get the Poisson mixture model (see next slide for a visual)

$$P(X = k) = .3 \times e^{-20}\frac{20^k}{k!} + .7 \times e^{-40}\frac{40^k}{k!}$$

Mixture of Poisson distributions



Example – Continued

- $P(X > 35) = \sum_{k=36}^{\infty} \left[.3 \times e^{-20} \frac{20^k}{k!} + .7 \times e^{-40} \frac{40^k}{k!} \right]$

Question b:) What is the expected number of customers?

$$\mathbf{E}(X) = \sum_k k \times P(X = k) = \sum_{k=0}^{\infty} k \left[.3 \times e^{-20} \frac{20^k}{k!} + .7 \times e^{-40} \frac{40^k}{k!} \right]$$

Also we can use that if $Y \sim \text{Poisson}(\lambda)$, then $P(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}$ and $E(Y) = \lambda$:

$$\mathbf{E}(X) = .3 \sum_{k=0}^{\infty} k e^{-20} \frac{20^k}{k!} + .7 \sum_{k=0}^{\infty} k e^{-40} \frac{40^k}{k!} = .3 \times 20 + .7 \times 40 = 34$$

Definition 19 (Expectations of Multivariate Random Variables)

Let $g(X_1, \cdots, X_n)$ be a function of random variables.

- If $X_1, \cdots, X_n$ are discrete random variables with a probability function $p(x_1, \cdots, x_n)$, then

$$\mathbb{E}[g(X_1, \cdots, X_n)] = \sum_{x_1, \cdots, x_n} g(x_1, \cdots, x_n) p(x_1, \cdots, x_n)$$

- If $X_1, \cdots, X_n$ are continuous random variables with a probability density function $f(x_1, \cdots, x_n)$, then

$$\mathbb{E}[g(X_1, \cdots, X_n)] = \int_{\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_n) f(x_1, \cdots, x_n) \mathrm{d}x_1 \cdots \mathbf{d}x_n \,.$$

Example 26

Toss two dices. X_1= number shown on top of die 1, and X_2 = number shown on top of die 2. Find the expectation of $X_1 + X_2$, and the expectation of $X_1 X_2$.

Sec 5.6: Special Theorems (discrete)

Theorem 12

$\mathbb{E}[c] = c$

Theorem 13

Let $g(X_1, X_2)$ be a function of random variables, then

$$\mathbb{E}[cg(X_1, X_2)] = c\mathbb{E}[g(X_1, X_2)]\,.$$

Theorem 14

Let $g_1(X_1, X_2), \cdots, g_n(X_1, X_2)$ be functions of random variables, then

$$\mathbb{E}\left[\sum_{i=1}^{n} g_i(X_1, X_2)\right] = \sum_{i=1}^{n} \mathbb{E}[g_i(X_1, X_2)]\,.$$

Theorem 15

Let X_1, X_2 be independent random variables and g_1, g_2 be two functions. Then

$$\mathbb{E}[g_1(X_1)g_2(X_2)] = \mathbb{E}[g_1(X_1)] \cdot \mathbb{E}[g_2(X_2)]$$

Example 27

Toss two dices. X_1= number shown on top of die 1, and X_2 = number shown on top of die 2. Find the expectation of $X_1 + X_2$, and the expectation of $X_1 X_2$.

Example 28

Toss two dices. X_1= number shown on top of die 1, and X_2 = number shown on top of die 2. Find the expectation of $X_1 + X_2$, and the expectation of $X_1 X_2$.

Sec 5.7: Covariance (discrete)

Definition 20 (Covariance Function)

Random variables X_1, X_2 have mean μ_1, μ_2 and standard deviation σ_1, σ_2. The covariance function is defined as

$$Cov(X_1, X_2) = \mathbb{E}\left[(X_1 - \mu_1) \cdot (X_2 - \mu_2)\right]$$

Correlation coefficient is $\rho = \dfrac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}$

Theorem 16

$$Cov(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \cdot \mathbb{E}[X_2]\,.$$

If X_1 and X_2 are independent, then $Cov(X_1, X_2) = 0$ and $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$

Remark 4

- $Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)$
- If $Cov(X_1, X_2) = 0$, we say X_1, X_2 are uncorrelated. linearly

Theorem 17

$$Cov^2(X_1, X_2) = \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)]^2 \leq \mathbb{E}[(X_1 - \mu_1)^2] \cdot \mathbb{E}[(X_2 - \mu_2)^2]$$
$$= Var(X_1)Var(X_2)$$

Hence, $-1 \leq \rho(X_1, X_2) \leq 1$.

Theorem 18

Let $X_1, \cdots X_n$ *and* $Y_1, \cdots, Y_m$ *be sequences of random variables. Let*

$$U = \sum a_i X_i, \quad V = \sum b_j Y_j \, .$$

- $\mathbb{E}[U] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$
- $Var(U) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j Cov(X_i, X_j)$
- $Cov(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$

Proof:

Show the case of two variables.

Example 29

A box has r red balls and $N-r$ blue balls. A random sample of n balls is drawn without replacement. Let X be the number of red balls drawn. Find X's mean and variance (X has a hypergeometric distribution)

Solution:

- Label the balls as $1, 2, \ldots, n$, and define

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th draw is a red ball} \\ 0, & \text{otherwise} \end{cases}$$

- Clearly, $P(X_k = 1) = \dfrac{r}{N}$ for $k = 1, \ldots, n$ because for k-th ball, r out of N choices are red

$$\mathbb{E}[X_k] = \frac{r}{N}, \quad \mathbb{E}[X_k^2] = \frac{r}{N}, \quad Var(X_k) = \mathbb{E}[X_k^2] - \mathbb{E}[X_k]^2 = \frac{r}{N}\left(1 - \frac{r}{N}\right)$$

- When $j \neq k$, $\binom{r}{2}$ out of $\binom{N}{2}$ choices lead to red j- and k-th ball. Hence

$$P(X_j = 1, X_k = 1) = \frac{\binom{r}{2}}{\binom{N}{2}} = \frac{r(r-1)}{N(N-1)}$$

Solution – Continued

- Covariance, when $i \neq j$

$$\begin{aligned} Cov(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \cdot \mathbb{E}[X_j] = \frac{r(r-1)}{N(N-1)} - \left(\frac{r}{N}\right)^2 \\ &= -\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1} \end{aligned}$$

$C_r^2 \cdot C$

- Clearly, $X = \sum_{i=1}^{n} X_i \qquad \Rightarrow \mathbb{E}[X] = n\frac{r}{N}$

Finally:

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j < n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} \frac{r}{N}\left(1 - \frac{r}{N}\right) + 2 \sum_{1 \leq i < j < n} \left[-\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1}\right] \\ &= n \cdot \frac{r}{N}\left(1 - \frac{r}{N}\right) - n(n-1) \cdot \left[\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1}\right] \\ &= n\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-n}{N-1} \end{aligned}$$

Example 30 (Binomial revisited)

- Let $X_i\,, i = 1\ldots, n$, Independent, Identically distributed Bernoulli with parameter p:

$$P(X_i = 1) = p\,, \quad P(X_i = 0) = 1 - p = q$$

[Such random variables (not necess. Bernoulli) are called IID]

- Consider the new random variable: $X = X_1 + X_2 + \cdots + X_n$.
- Observe $X = \#$ "successes", if 1=success $\Rightarrow$ X is Binomial $B(n, p)$!
- Easy calculation of $E(X)$ (using also Theorem 13):

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} E(X_i) = np$$

- For $Var(X)$ (use quadratic formula, covariance and independence):

$$Var(X) = Var(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} Var(X_i)$$

- Recall for Bernoulli: $Var(X_i) = pq$. Thus we **easily** get:

$$Var(X) = Var(X_1 + X_2 + \cdots + X_n) = \sum_{i=1^n} Var(X_i) = npq$$

Another interesting observation: variance reduction

- Let as consider X_i, $i = 1, \ldots, n$ to be IID with finite variance

$$Var(X_1) = Var(X_2) = \ldots = Var(X_n) = \sigma^2 < \infty$$

- Consider the average (aka sample mean)

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Then, using that $Var(cY) = c^2 Var(Y)$ we get

$$Var(\bar{X}_n) = \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{n}{n^2} Var(X_1)$$

Therefore variance is reduced through averaging:

$$Var(\bar{X}_n) = \frac{Var(X_1)}{n}$$

Sec 5.9: Multinomial probability distribution

Background

- From Chapter 3, we know that a binomial r.v. results from an experiment consisting of n trials with two possible outcomes per trial: **S or F**
- In the case where the number of possible outcomes per trial is more than two, one needs another concept:
- Recall the binomial random experiment:
  1. The experiment consists of n identical and independent trials.
  2. Each trial results in one of two outcomes (concerned with the r.v. X of interest). We call one outcome a success S and the other a failure F.
  3. **Bernoulli building blocks:** The probability of success on a single trial is equal to p and remains the same from trial to trial. The probability of a failure is equal to $q = 1 - p$.
  4. The random variable of *interest* is:
     $Y = \#$ of successes observed during the n trials.

Multinomial Experiment

Is a generalization of the binomial experiment, and possesses the following properties:

1. The experiment consists of n identical trials.
2. The trials are independent.
3. The outcome of each trial falls into one of k classes.
4. The probability that the outcome of a single trial falls into class i is p_i, $i = 1, 2, \ldots, k$ and remains the same from trial to trial.

   Notice that $p_1 + p_2 + \cdots + p_k = 1$.
5. The r.v.'s of interest are $Y_1, Y_2, \ldots, Y_k$, where

   $Y_i = \#$ **of trials for which the outcome falls into class** i**.**
6. Notice $Y_1 + Y_2 + \cdots + Y_k = n$.

Definition 21 (Multinomial probability distribution)

Assume that $p_1, p_2, \ldots, p_k$ are such that $\sum_{i=1}^{k} p_i = 1$, and $p_i > 0$ for $i = 1, 2, \ldots, k$. The r.v.'s $Y_1, Y_2, \ldots, Y_k$ are said to have a *multinomial distribution* with parameters n and $p_1, p_2, \ldots, p_k$ if the joint probability function of $Y_1, Y_2, \ldots, Y_k$ is given by

$$p(y_1, y_2, \ldots, y_k) = \frac{n!}{y_1!\, y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k},$$

where for each i, $y_i = 0, 1, \ldots, n$ and $\sum_{i=1}^{k} y_i = n$.

Remark 5

- Note that $k = 2$ provides the binomial experiment/distribution.
- Notation: $\binom{n}{y_1, y_2, \ldots, y_k} = \frac{n!}{y_1! y_2! \cdots y_k!}$
- Notation $Y = (Y_1, Y_2, ..., Y_k) \sim \text{Multi}(n, p_1, p_2, ..., p_k)$
- Recall Multinomial Theorem:

$$(p_1 + ... + p_k)^n = \sum_{y_1=0}^{n} \sum_{y_2=0}^{n} \cdots \sum_{y_k=0, \sum_{i=1}^{k} y_i = n}^{n} \frac{n!}{y_1!\, y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$$

Example 31 (Multinomial)

We roll a fair die 100 times. Find the probability that we get 22 ones and 17 fives.

Solution:

- Classes: $1, 2, 3, 4, 5, 6$ – We do not need all of them!
- Better Classes: $1, 5$, everything else
- Probabilities to be in each class: $1/6 + 1/6 + 4/6$
- $Y_i = \#$ of trials for which the outcome is $i = 1, 5$, $X =$everything else
- Therefore $(Y_1, Y_5, X) \sim \text{Multi}(100, 1/6, 1/6, 2/3)$
- Solution: 61=100-22-17

$$P(Y_1 = 22, Y_5 = 17) = P(Y_1 = 22, Y_5 = 17, X = 61) =$$
$$= \frac{100!}{22!\,17!\,61!}\left(\frac{1}{6}\right)^{22}\left(\frac{1}{6}\right)^{17}\left(\frac{2}{3}\right)^{61} \approx .0037$$

Theorem 19

Assume $Y = (Y_1, Y_2, ..., Y_k) \sim Multi(n, p_1, p_2, \ldots, p_k)$. *Then*

1. $E(Y_i) = np_i$,
2. $V(Y_i) = np_iq_i = np_i(1 - p_i)$ *where* $q_i = 1 - p_i$.
3. $Cov(Y_i, Y_j) = -np_ip_j$ *if* $i \neq j$. *(negative covariance-we discuss this point later)*

Example 32

Suppose that a fair die is rolled 9 times. Let Y_i be the number of trials for which number i appears.

- Contrast to the previous example. Different random variables **of interest**!
- What is the probability that 1 appears three times, 2 and 3 twice each, 4 and 5 once each, and 6 not at all?
- Find $E(Y_i)$, $V(Y_i)$ and $Cov(Y_i, Y_j)$ where $i, j = 1, \ldots, 6$.

Definition 22 (Distribution Function)

X is a random variable. The distribution function (also called cumulative distribution function, aka *CDF*) of X, denoted by $F(x)$, is

$$F(x) = P(X \leq x) \qquad -\infty < x < \infty$$

Example 33

X has binomial distribution with $n = 2, p = \frac{1}{2}$. Find $F(x)$

- The probability function is $p(x) = \binom{2}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{2-x}, \qquad x = 0, 1, 2$
- $p(0) = 1/4,\ p(1) = 1/2,\ p(2) = 1/4$
- So, the distribution function is as follows:

$$F(x) = P(X \leq x) = \begin{cases} 0, & \text{if } x < 0 \\ 1/4, & \text{if } 0 \leq x < 1 \\ 3/4, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } 2 \leq x \end{cases}$$

Distribution functions for discrete random variables are always step-functions!

Theorem 20 (Properties of CDFs)

If $F(x)$ is a distribution function, then

- $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$
- $F(\infty) = \lim_{x \to \infty} F(x) = 1$
- $F(x)$ *is nondecreasing.*

Definition 23 (Continuous Random Variable)

A random variable X with CDF $F(x)$ is said to be continuous if $F(x)$ is continuous on $\mathbb{R}$.

Remark 6

- Probability of intervals: $P(X \in (a, b]) = F(b) - F(a)$
- Probability of point: $P(X = b) = F(b) - \lim_{x \to b^-} F(x)$
- If X is a continuous random variable, then $P(X = a) = 0$ for all $a \in \mathbb{R}$.

Definition 24

Let $F(x)$ be the CDF for X. If $F(x)$ is differentiable at x, then

$$f(x) = F'(x)$$

is the probability density function (or simply density function – aka PDF) of X.

Example 34

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \le x \le 1 \\ 1 & \text{if } 1 < x \end{cases} \quad \text{and} \quad F(y) = \begin{cases} 0, & \text{if } y < 0 \\ y^2, & \text{if } 0 \le y \le 1 \\ 1 & \text{if } 1 < y \end{cases}$$

Theorem 21 (Properties of Density Function)

If $f(x)$ is a probability density function of a continuous random variable, then

- $f(x) \ge 0$
- $\int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1$

Theorem 22

If a continuous random variable X has density function $f(x)$ for all $x \in (a, b)$, then

$$P(X \in [a, b]) = P(X \in (a, b)) = \int_a^b f(x)\mathrm{d}x\,.$$

So,

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

if $f(x)$ exists everywhere. Note that $P(X = c) = \int_c^c f(x)\mathrm{d}x = 0$.

Definition 25 (Percentile)

p-th percentile ϕ_p of X is defined by

$$\phi_p = \inf_{\phi} \{F(\phi) \geq p\}.$$

- Note such a ϕ_p always exists, by the completeness of $\mathbb{R}$
- For a continuous r.v. $F(\phi_p) = p$
- If $p = 0.5$, $\phi_{0.5}$ is called the median of X.

Example 35

$$f(x) = cx, \quad 0 \leq x \leq 3$$

- Find c such that $f(x)$ is a density function of X
- Find $P(X \in (1, 2))$
- Find the median and 0.75 percentile of X

Sec 4.3: Expected Value and Variance

Definition 26 (Expected Value)

If X is a continuous random variable with density function $f(x)$, then

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\mathrm{d}x \qquad \text{and} \quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\mathrm{d}x\,.$$

Variance

$$Var(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2$$

Theorem 23 (Properties of Expected Values)

Let X be a continuous random variable, then

- $\mathbb{E}[c] = c, \qquad \mathbb{E}[cg(X)] = c\mathbb{E}[g(X)]$
- $\mathbb{E}[g_1(X) + g_2(X)] = \mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)]$

Example 36

- X is a continuous random variable with density function $\frac{1}{2}x$ for $x \in [0, 2]$. Find the expectation and variance of X.
- Example: $f(y) = cy^2$ for $0 \le y \le 4$. Find c such that $f(y)$ is a density function. Find $P(1 \le Y \le 2)$.

Definition 27 (Uniform Distribution)

A random variable $X \sim U(\theta_1, \theta_2)$ is said to be a uniform random variable on the interval (θ_1, θ_2) if it has the probability density function

$$f(x) = \frac{1}{\theta_2 - \theta_1}$$

for $\theta_1 \leq x \leq \theta_2$ and 0 elsewhere.

Theorem 24

If $X \sim U(\theta_1, \theta_2)$, then

$$\mathbb{E}[X] = \frac{\theta_1 + \theta_2}{2}, \quad Var(X) = \frac{(\theta_1 - \theta_2)^2}{12}.$$

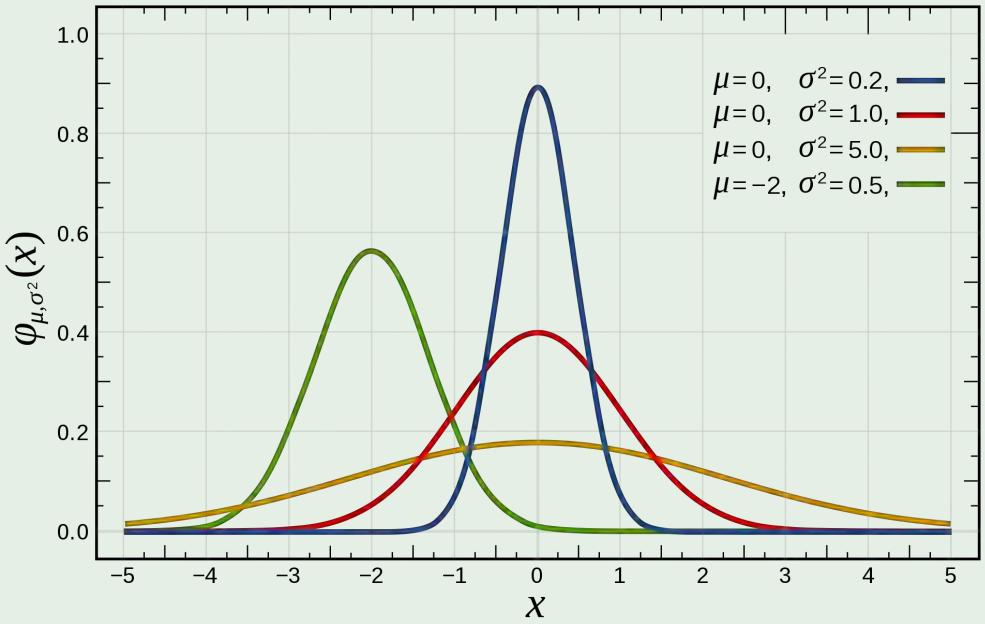
Proof.

□

Definition 28 (Normal Distribution)

A random variable X is said to have a normal distribution if and only if for $\sigma > 0$ and $\mu \in \mathbb{R}$, the density function of X is

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



Theorem 25

If X is a normal distribution with parameters μ and σ, then

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

Example 37 (Standard Normal & calculation of $P(Z > z)$)

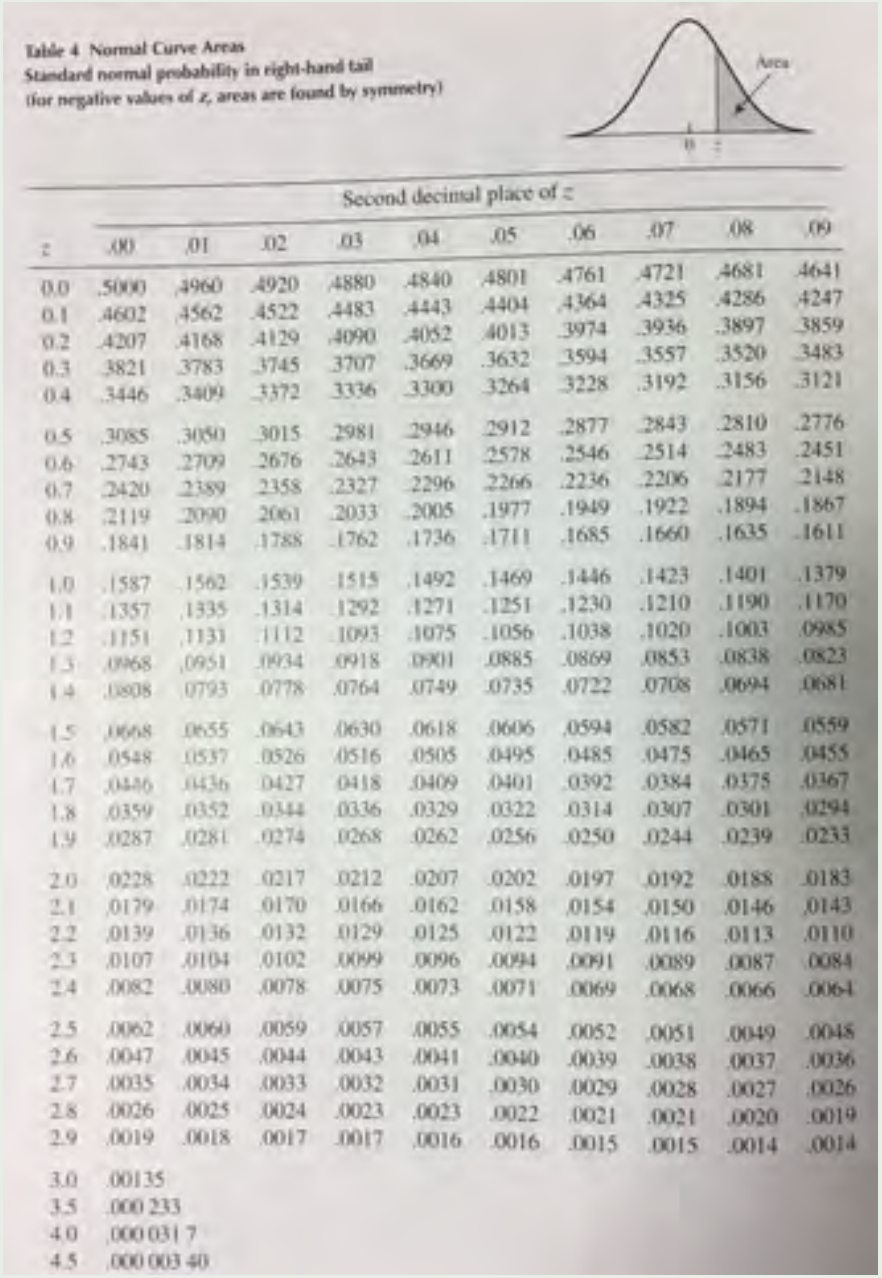


Table 4 Normal Curve Areas
Standard normal probability in right-hand tail
(for negative values of z, areas are found by symmetry)

	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000 233									
4.0	.000 031 7									
4.5	.000 003 40									

Question: Let $Z \sim N(0,1)$, find $P(Z > 2)$, $P(Z > 3)$, $P(-1 < Z < 1)$

Solution: Check the table:

- $P(Z > z) = 1 - F(z)$
- $P(Z > 2) = 0.0228$
- $P(Z > 3) = 0.00135$
- $P(-1 < Z < 1) = 1 - 2 \times 0.1587 = 0.6826$
- Usually we call it "standard" normal: $Z \sim N(0,1)$ (see next slide)

Theorem 26

Assume $X \sim N(\mu, \sigma)$. *Then* $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$

Proof.

- Let $z = \dfrac{x-\mu}{\sigma}$ any $x \in \mathbb{R}$.
- $F(z) = P(Z < z) = P(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{|y-\mu|^2}{2\sigma^2}} dy$
- By the change of variables $w = \frac{y-\mu}{\sigma}$ we have that

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{|y-\mu|^2}{2\sigma^2}} dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{|y-\mu|^2}{2\sigma^2}} d\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{|w|^2}{2}} dw$$

- Therefore:

$$F(z) = P(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{|w|^2}{2}} dw$$

and thus: $Z \sim N(0,1)$

□

Example 38

The weekly production in a factory depends on the number of orders from customers which can vary significantly.

- The number of orders in a factory during a specific week is normally distributed random variable with mean 50. If the standard deviation is known to equal to 10, then what can be said about the probability that the orders will be between 40 and 60 items?
- What is the probability that the number of items ordered **exceeds** 30% over the mean?

Solution:

- Let $X = \#(\text{orders in a factory during a specific week of the year})$. Then $X \sim \mathcal{N}(50, 10^2)$. Denote by F the CDF of X, the probability in question:

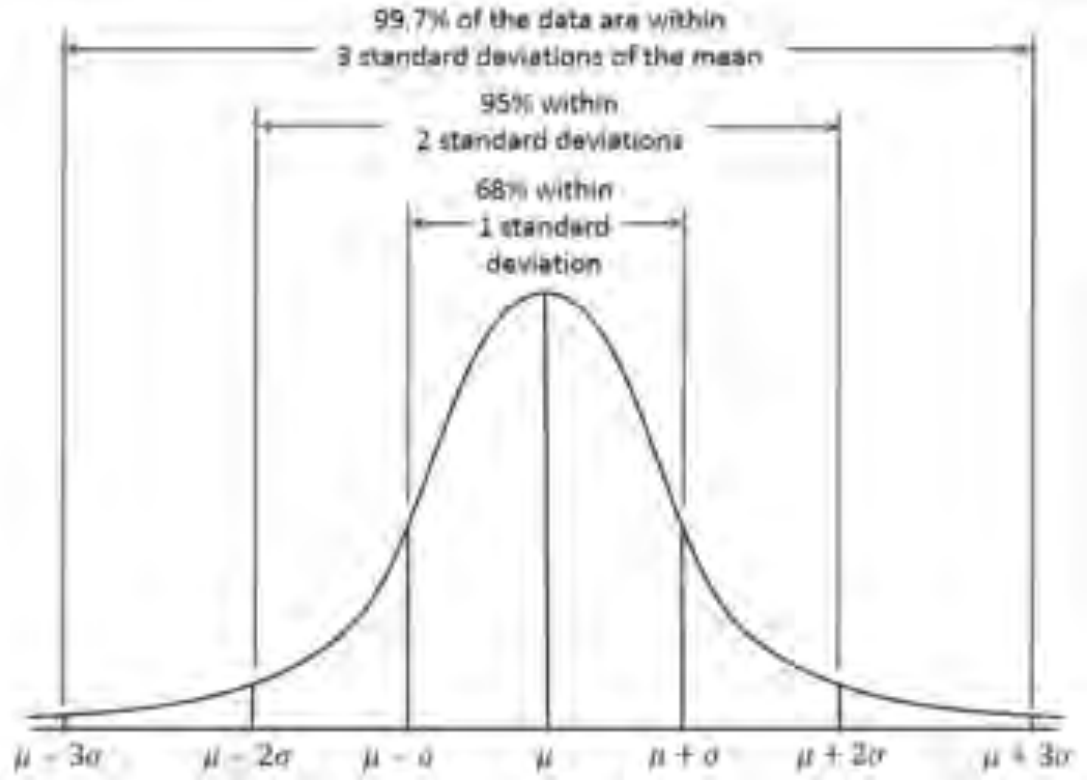
$$\begin{aligned} \mathbf{P}(40 < X < 60) &= \mathbf{P}\left(\frac{40-50}{10} < Z < \frac{60-50}{10}\right) = \mathbf{P}\left(-1 < Z < 1\right) \\ &= F(1) - F(-1) = 1 - 2P(Z > 1) \\ &= 1 - 2 \cdot 0.1587 = 0.6826 \end{aligned}$$

  where $Z = \frac{X-50}{10}$.
- $P(X > 50 + 0.3 \times 50) = \mathbf{P}(X > 65) = \mathbf{P}(Z > \frac{65-50}{10}) = P(Z > 1.5) = .0668$

Confidence intervals: "X is within k standard deviations from μ"

Confidence Interval

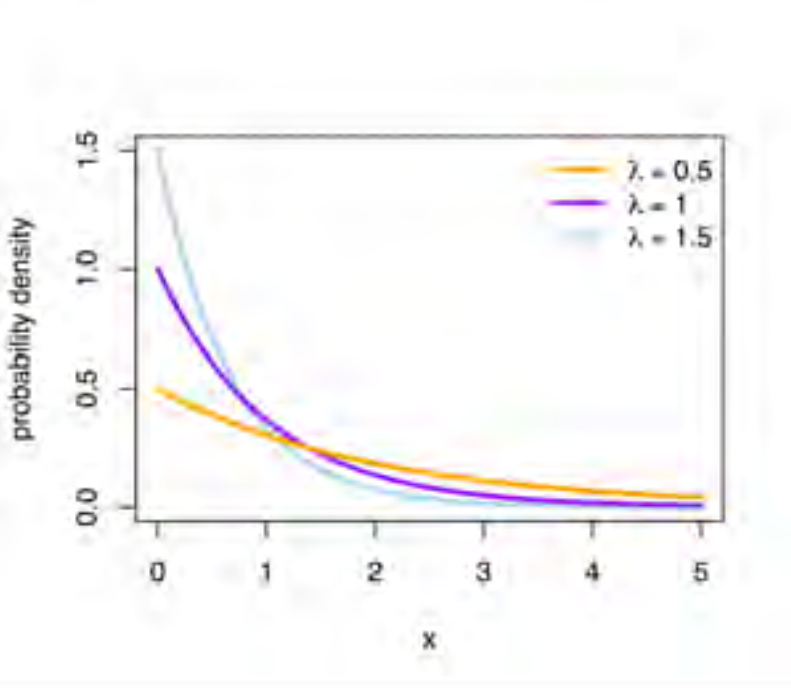
- Based on $X \sim N(\mu, \sigma) \iff Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- Use standard normal: $P(-2 < Z < 2) \approx 95\%$ (from Tables) to calculate:
- $P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < \frac{X-\mu}{\sigma} < 2) = P(-2 < Z < 2) \approx 95\%$



Definition 29 (Exponential distribution)

A random variable $X \sim \text{Exp}(\lambda)$ is said to be an exponential random variable with rate λ if it has the probability density function

$$f(x) = \lambda e^{-\lambda x}, \qquad \text{for } x > 0$$



Theorem 27

Let $X \sim Exp(\lambda)$, then $P(X > a + b | X > a) = P(X > b)$. *ie, exponential distribution is memoryless. (Continuous version of geometric distribution)*

Convergence of Geometric Random Variable

Assume a car engine has some problem st probability of failure in any time interval with length t is λt. Let T be the time of first failure., claim that $T \sim Exp(\lambda)$

- Divide t into n sub-intervals of equal length $\Delta t = \frac{t}{n}$
- Recall that: Geometric distribution X has $p(n) = (1-p)^{n-1}p$. Then,

$$P(T > t) = \left(1 - \lambda \frac{t}{n}\right)^n \quad \Rightarrow \lim_{n\to\infty} P(T > t) = e^{-\lambda t}$$

- So

$$P(T \leq t) = 1 - e^{-\lambda t} \quad \Rightarrow \text{density function } f(t) = \lambda e^{-\lambda t}$$

Remark 7

- In exponential distribution, $\lambda \cdot \Delta t$ is the probability something happens in Δt duration
- **Poisson point process:** run an exponential random Exp(λ) variable repeatedly, independently. Let $N(a, b)$ be the number of arrivals during the interval (a, b), then (by Poisson distribution with parameter $\lambda(b-a)$)

$$P(N(a,b) = n) = \frac{(\lambda(b-a))^n}{n!} e^{-\lambda(b-a)}.$$

Theorem 28

If $X \sim Exp(\lambda)$, *then*

$$\mathbb{E}[X] = \lambda^{-1}, \quad Var(X) = \lambda^{-2}.$$

Example 39

- If Y has an exponential distribution and $P(Y > 2) = 0.0821$. Find $\beta = \mathbb{E}[Y]$ and $P(Y \leq 1.7)$. $P(Y \leq t) = 1 - e^{-\lambda t}$
- Let X be an exponential distribution with mean β. Let Y be the ceiling of X. Show that Y is a geometric random variable.

  Check Y's values, and probability of taking each value!

Example 40

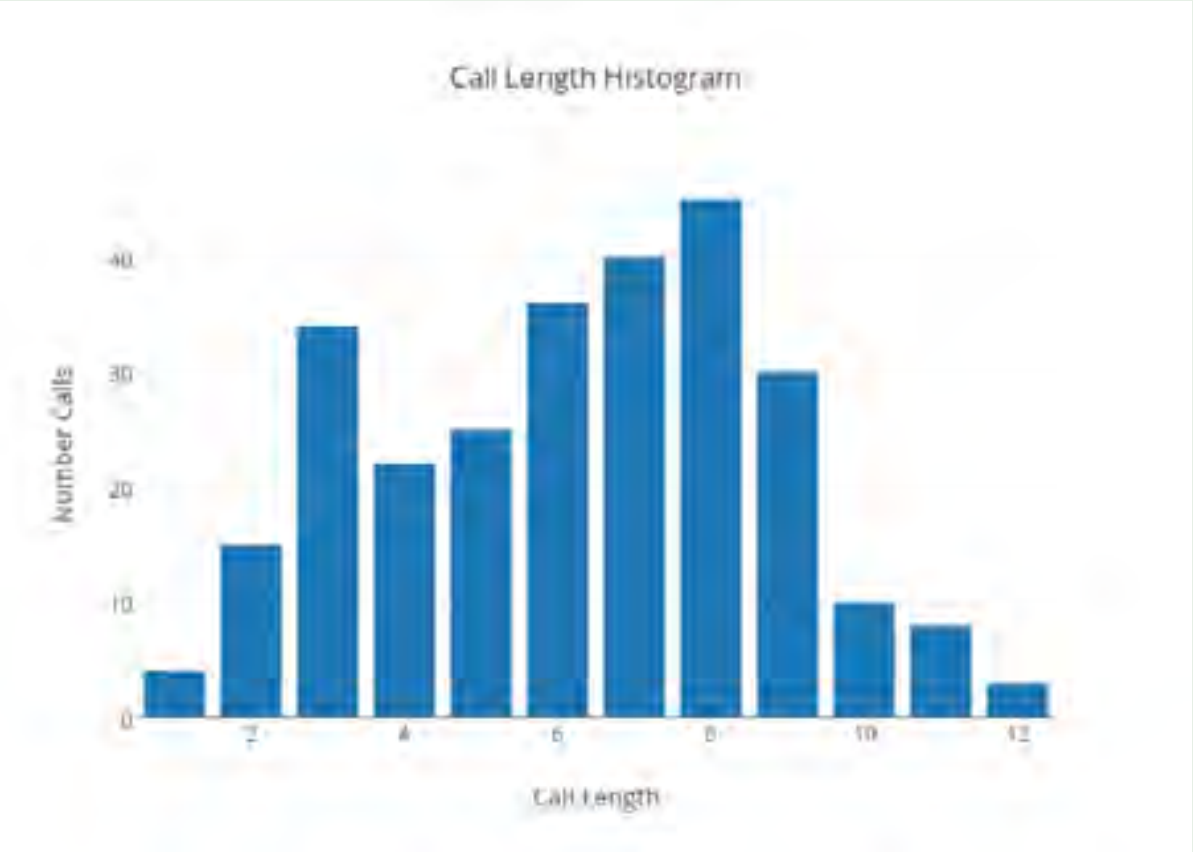
- Length X of a phone call in minutes at a call center modeled by an exponential distribution.
- From available data: Average length of phone call is 10 min:

$$E(X) = 10 = \lambda^{-1} = \beta, \quad \text{[occasionally we use } \beta = \lambda^{-1} \text{ notation]}$$

- Probability a call takes more than 8 minutes: $P(X > 8)$.
- *Recall* $P(X \leq t) = 1 - e^{-\lambda t}$

Remark 8

- Sometimes data need more flexible distributions
- Data from a call center: not exponential, see Gamma distribution next



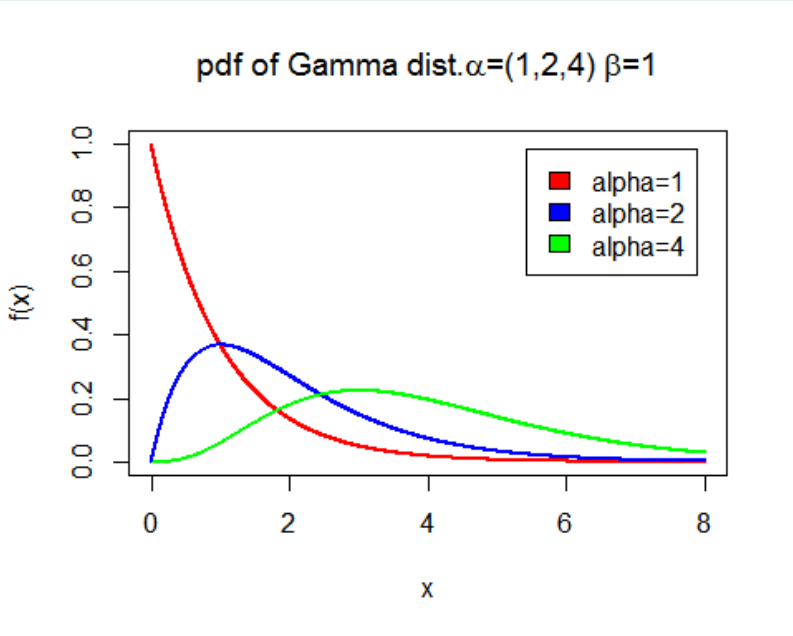
Sec 4.6b: Gamma Distribution

Definition 30 (Gamma distribution)

A random variable X is said to be a Gamma distribution with parameters $\alpha, \beta > 0$ if and only if the density function is

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad x > 0\,,$$

Written as $X \sim \Gamma(\alpha, \beta)$



where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}\mathrm{d}x, \qquad \Gamma(\alpha+1) = \alpha\Gamma(\alpha)\,.$$

Theorem 29

If $X \in \Gamma(\alpha, \beta)$, *then* $\mathbb{E}[X] = \alpha\beta, \quad Var(X) = \alpha\beta^2\,.$

Proof.

Only do the expectation. □

Remark 9

- α is the *shape parameter*, and β is the *scale parameter*
- If $\alpha = 1$, gamma becomes exponential distribution with $\lambda = \frac{1}{\beta}$
- $\Gamma(n, \beta)$ is the distribution of sum of n i.i.d exponential distributions $\text{Exp}(\beta^{-1})$.
- More generally, the sum of independent Gamma variables $\Gamma(\alpha_i, \beta)$ also has Gamma distribution $\Gamma(\sum_i \alpha_i, \beta)$:

$$\begin{cases} X_i \sim \Gamma(\alpha_i, \beta) \\ \{X_i\} \text{ are mutually independent} \end{cases} \Rightarrow \sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i}^{n} \alpha_i, \beta\right)$$

Definition 31 (Chi-square Distribution)

Let v be a positive integer. A random variable Y is said to have a chi-square distribution with v degree of freedom if and only if Y is a Gamma distribution with parameters $\alpha = v/2$ and $\beta = 2$.

Example 41

Suppose a random variable Y has density function $f(y) = ky^3 e^{-y/2}$ for $y > 0$. Find k and identify the random variable.

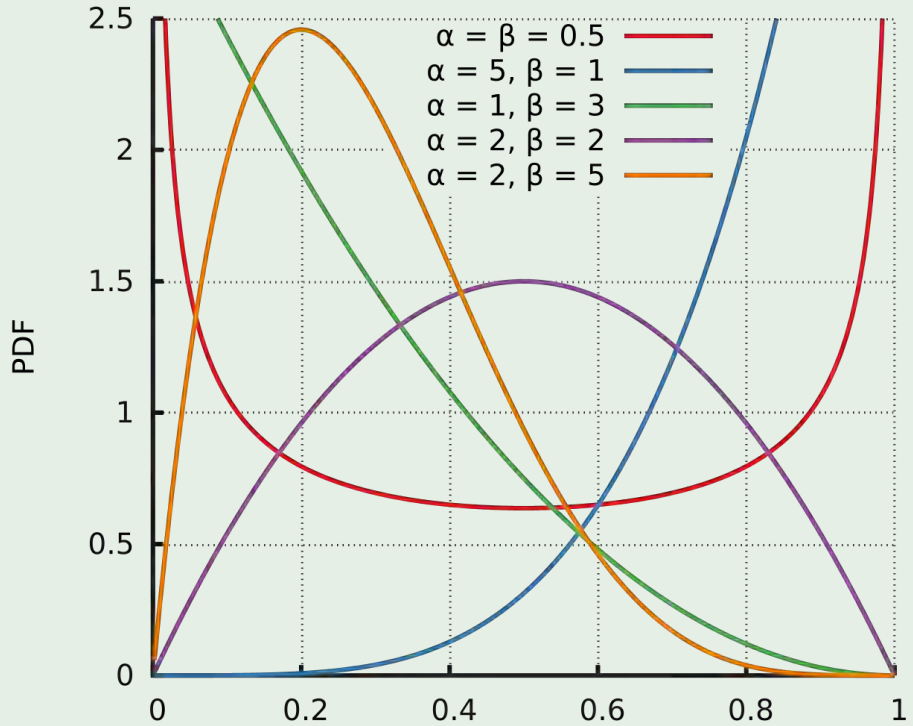
Definition 32 (Beta distribution)

A random variable is said to be a Beta distribution with parameters $\alpha, \beta > 0$ if and only if the density function is

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad x > 0\,,$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\,.$$



Important comments for Beta, Gamma and much more

Remark 10

- The denominators in both the Gamma and the Beta

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad x > 0\,,$$

and

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < x < 1\,,$$

are constants selected so that the $f's$ are pdfs, namely $\int f(x)dx = 1$ (since they are already non-negative).

- In general we can define a pdf for any function $g = g(x)$ such that $g(x) \geq 0$ and $\int g(x)dx < \infty$ as follows:

$$f(x) = \frac{g(x)}{\int g(y)dy}\,, \quad \textbf{and write:} \quad f(x) \propto g(x)$$

The constant $\int g(y)dy$ is called "normalizing" since it turns f into a pdf. In statistical physics it is called the "partition function".

Theorem 30

If X is a beta distribution with parameters α and β, then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Example 42

EX: The weekly repair cost Y for a machine has a probability density function given by $f(y) = 3(1-y)^2, 0 < y < 1$ with measurement in hundreds of dollars. How much should we budget each week such that the actual cost will exceed the budget only 10% of the time?

Solution:

- Y has Beta distribution $B(1,3)$
- Let the budget be u. Then the probability that $Y \leq u$ is

$$P(Y \leq u) = \int_0^u 3(1-y)^2 dy = -(1-y)^3\Big|_0^u = 1-(1-u)^3$$

- Need

$$P(Y \leq u) = 1-0.1 \quad \Rightarrow (1-u)^3 = 0.1 \quad \Rightarrow u = 1-\sqrt[3]{0.1} \simeq 0.54$$

Example 43

Assume a machine produces defective and good parts but we don't know what proportion of parts are defective. If we sampled N parts. The distribution of defective parts is

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

Sec. 4.9+3.10: Moments and Moment Generating Functions

Moment Generating Function

- The k-th moment of X is defined as $\mu'_k = \mathbb{E}[Y^k]$
- The k-th central moment is defined as $\mu_k = \mathbb{E}[(Y-\mu)^k]$
- The moment generating function of X is given by: $m_X(t) = \mathbb{E}[e^{tX}]$

The m.g.f. is said to exist if $\exists$ a constant $b > 0$ st $m_X(t)$ is finite for $|t| < b$.

Remark 11

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-\infty}^{\infty} \left(1 + tx + \frac{t^2x^2}{2!} + \cdots\right) f(x)dx \\ &= \int_{-\infty}^{\infty} f(x)dx + t\int_{-\infty}^{\infty} xf(x)dx + \frac{t^2}{2!}\int_{-\infty}^{\infty} x^2 f(x)dx + \cdots \\ &= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots \end{aligned}$$

Power series in t

Theorem 31 (Moments can be computed as derivatives)

If $m(t)$ exists for X, then: $\mathbb{E}[X^k] = \left.\frac{d^k m(t)}{dt^k}\right|_{t=0}$

Example 44 (Binomial Distribution)

Moment generating function: $m(t) = (pe^t + 1 - p)^n$

- Mean: $\mu = \left.\frac{dm}{dt}\right|_{t=0} = \left. n(pe^t + 1 - p)^{n-1} pe^t \right|_{t=0} = np$
- Variance

$$\begin{aligned}
\left.\frac{d^2 m}{dt^2}\right|_{t=0} &= \left(n(n-1)(pe^t + 1 - p)^{n-2} pe^t pe^t + n(pe^t + 1 - p)^{n-1} pe^t \right)\Big|_{t=0} \\
&= n(n-1)p^2 + np \\
\Rightarrow \quad &\text{Variance} = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1-p)
\end{aligned}$$

Moment Generating Functions for Discrete Random Variables

$A = 1 + qe^t + (qe^t)^2 + \dots$ $\quad A \cdot qe^t = qe^t + (qe^t)^2 + \dots +$

$A - Aqe^t = 1 \quad A = \frac{1}{qe^t - 1}$

Definition 33

For **discrete random variables** X with probability distribution function $P(X = x) = p(x)$ we define:

$$m_X(t) = \mathbb{E}[e^{tX}] = \sum_{\text{all } x} e^{tx} p(x)$$

$(1-p)e^t - 1$

$e^t - pe^t - 1$

Example 45 (MGF for Bernoulli Distribution)

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

$$m(t) = \mathbb{E}[e^{tX}] = \sum_{x=0}^{1} e^{tx} p(x) = e^{t \cdot 0} p(0) + e^{t \cdot 1} p(1) = e^t p + (1 - p)$$

Example 46 (MGF for Poisson Distribution)

$$\begin{aligned} m(t) &= \mathbb{E}[e^{tX}] = \sum_{x=0}^{\infty} e^{tx} p(x) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} \cdot e^{\lambda e^t} = e^{\lambda(e^t - 1)} \end{aligned}$$

Example 47 (MGF for Binomial Distribution)

$$\begin{aligned} m(t) &= \mathbb{E}[e^{tX}] = \sum_{x=0}^{n} e^{tx} p(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (pe^t + 1 - p)^n \end{aligned}$$

However: there is a MUCH simpler way: use that $X = X_1 + \ldots + X_n$ of Bernoullis + independence:

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(X_1 + \ldots X_n)}] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] = (pe^t + 1 - p)^n$$

Example 48 (Moment Generating Function for Gamma Distribution)

$$
\begin{aligned}
m_X(t) &= \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \left(\frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \right) \mathrm{d}x \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} \exp\left[-x \left(\frac{1}{\beta} - t \right) \right] dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} \exp\left[\frac{-x}{\beta/(1-\beta t)} \right] dx
\end{aligned}
$$

- The integrand is part of distribution $\Gamma(\alpha, \frac{\beta}{1-\beta t})$, if $1 - \beta t > 0 \Leftrightarrow t < 1/\beta$
- So, when $t < 1/\beta$, the integral equals $\left(\frac{\beta}{1-\beta t} \right)^\alpha \Gamma(\alpha)$
- Hence,

$$
m_X(t) = \left(\frac{\beta}{1-\beta t} \right)^\alpha \Gamma(\alpha) \cdot \frac{1}{\beta^\alpha \Gamma(\alpha)} = \frac{1}{(1-\beta t)^\alpha}, \qquad \text{for } t < \frac{1}{\beta}
$$

Example 49 (Moment Generating Function)

- Gamma distribution $m(t) = (1 - \beta t)^{-\alpha}$
- Normal distribution $m(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$
- Exponential distribution $m(t) = (1 - t/\lambda)^{-1}$
- Uniform distribution $U(a, b)$: $m(t) = \dfrac{e^{tb} - e^{ta}}{t(b - a)}$

Theorem 32

Let X be a random variable and $g(X)$ be a function of X. The moment-generating function for $g(X)$ is

$$\mathbb{E}[e^{tg(X)}] = \int e^{tg(x)} f(x) \mathrm{d}x \,.$$

Theorem 33

If $X \geq 0$ is a discrete random variable that takes only non-negative integer values, then

$$p(n) = \frac{1}{n!} \frac{d^n[m(\log t)]}{dt^n}\bigg|_{t=0}$$

Proof.

- Recall moment generating function: $m(t) = \mathbb{E}[e^{tX}] = \sum_{x=0}^{\infty} e^{tx} p(x)$
- Let $t = \log u$. Then $e^t = u$ and $m(\log u) = \sum_{x=0}^{\infty} u^x p(x)$

$$\frac{d\,[m(\log u)]}{du} = \sum_{x=1}^{\infty} x u^{x-1} p(x)$$

$$\Rightarrow \quad \frac{d^n\,[m(\log u)]}{du^n} = \sum_{x=n}^{\infty} x(x-1)\cdots(x-n+1) u^{x-n} p(x)$$

$$\Rightarrow \quad \frac{d^n\,[m(\log u)]}{du^n}\bigg|_{u=0} = n!\, p(n) \quad \Rightarrow \quad \boxed{p(n) = \frac{1}{n!} \frac{d^n[m(\log t)]}{dt^n}\bigg|_{t=0}}$$

□

Example 50

- Moment generating function for Poisson distribution is $m(t) = e^{\lambda(e^t - 1)}$
- Then

$$\begin{aligned} p(n) &= \frac{1}{n!} \frac{d^n[m(\log t)]}{dt^n}\bigg|_{t=0} = \frac{1}{n!} \frac{d^n[e^{\lambda(t-1)}]}{dt^n}\bigg|_{t=0} \\ &= \frac{\lambda^n}{n!} e^{-\lambda} \end{aligned}$$

Sec 4.10+3.11: Probabilistic Guarantees, Part I

Key ideas

- predict random variables with "confidence" by using a few moments (**moments are deterministic quantities!**)
- not necessary to know the entire pdf/cdf of the r.v.; this is in contrast to the normal distribution confidence intervals.

Theorem 34

Markov Inequality: *If $X \geq 0$, then for all $a > 0$,*

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

Chebyshev Inequality I: *For any random variable X, and any $\epsilon > 0$,*

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

Chebyshev Inequality II: *Let X be a random variable with mean $\mu = E[X]$ and finite variance $\sigma^2 var(X)$, then for any $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad [\text{follows from I for } \epsilon = k\sigma]$$

Proof for Discrete/Continuous r.v/'s.

- Markov inequality for discrete r.v.'s [$X \geq 0$!]

$$\begin{aligned} E[X] &= \sum_x xP(X=x) = \sum_{x \geq a} xP(X=x) + \sum_{0 \leq x < a} xP(X=x) \\ &\geq \sum_{x \geq a} xP(X=x) \geq aP(X \geq a) \end{aligned}$$

- For continuous r.v.s [use $X \geq 0$!]:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx = \int_a^{\infty} xf(x)dx + \int_0^a xf(x)dx \\ &\geq \int_a^{\infty} xf(x)dx \geq aP(X \geq a) \end{aligned}$$

- Chebyshev inequality: Apply Markov's inequality to r.v.: $Y = (X - E[X])^2$
  - $E[Y] = \text{var}(X)$
  - $P(|X - E[X]| \geq \epsilon) = P(|X - E[X]|^2 \geq \epsilon^2) = P(Y \geq \epsilon^2) \leq \frac{E[Y]}{\epsilon^2} = \frac{\text{var}(X)}{\epsilon^2}$

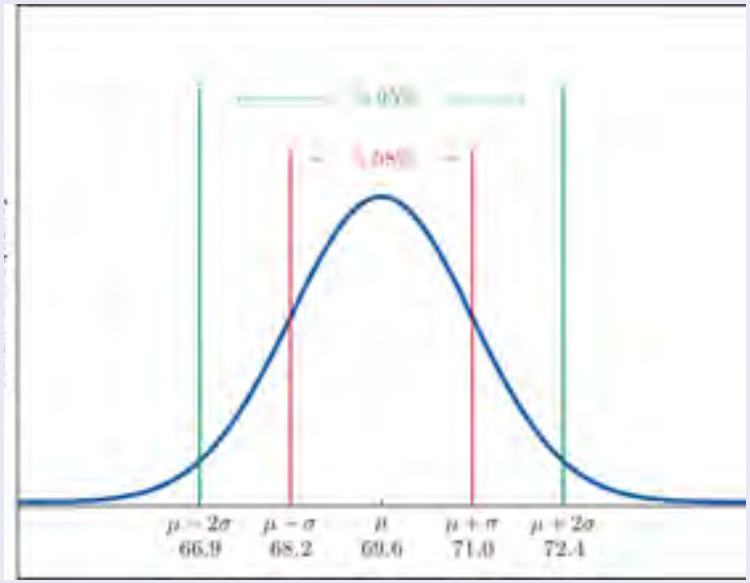
Similar proof for discrete random variables. □

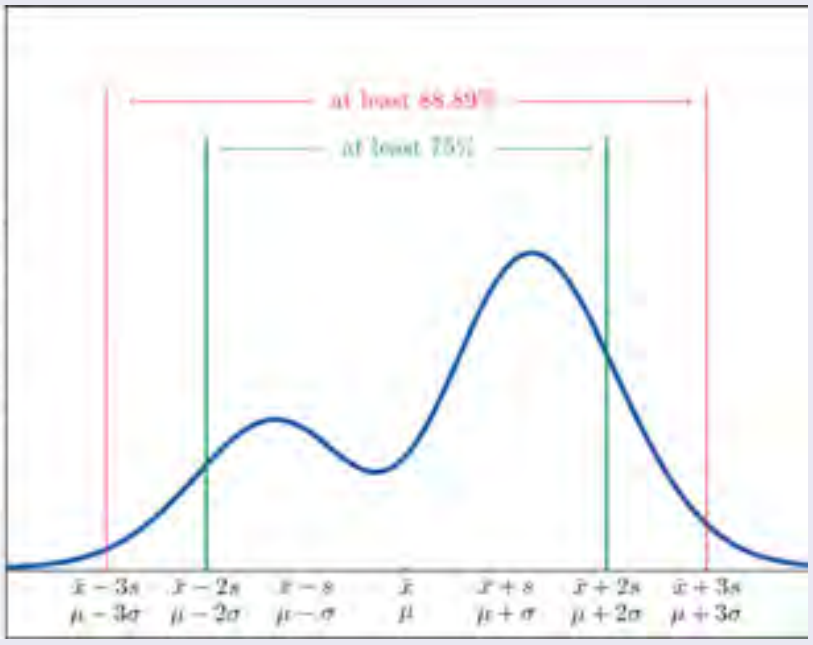
Confidence intervals II: "X is within k standard deviations from μ"

- Left: If we now $X \sim N(\mu, \sigma)$, then

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < \frac{X - \mu}{\sigma} < 2) = P(-2 < Z < 2) \approx 95\%$$

- Right: If we do not know the distribution of X but we know mean, variance – use Chebyshev inequality
- Observe the trade-offs between knowledge and confidence





Example 51 (Distribution-agnostic)

A certain brand of cereal has an average shelf life of 7 months. Assume that based on the current technology used in the manufacturing and packaging processes the standard deviation from its' average shelf life is $\sigma = 1$ month.

- Use Chebyshev's Theorem to determine whether it is likely (probability of occurrence $> 66\%$) that the contents of a cereal box will stay fresh anywhere between 5 and 9 months
- What is the desired standard deviation if the percentage in (a) needs to improve to at least 99%?

Solution:

- Let the random variable X be the shelf life of the cereal. 5 months is 2 standard deviations to the left of the mean; 9 months is 2 standard deviation to the right of the mean. By Chebyshev

$$P(-2\sigma \leq X - \mu \leq 2\sigma) \geq 3/4 = 75\%$$

- Use Chebyshev's Theorem, $1 - \frac{1}{k^2} = .99$, thus $k = 10$. Then we need $k\sigma \leq 2$, thus $\sigma = .2$ months or less.

$P(Y \geq 355) = P(Y - 150 \geq 505)$

Example 52

$\leq P(|Y - E[Y]| \geq k \cdot \delta)$

For one midterm exam, the class average is 80, and standard deviation is 5 [estimated from data!]. If randomly choose a student, what is the probability that $X \leq 60$?

Solution:

$$\begin{aligned} P(X \leq 60) &= P(X - 80 \leq -20) \\ &\leq P(|X - 80| \geq 20) = P(|X - E[X]| \geq 4 \times 5) \\ &\leq \frac{1}{4^2} = 0.0625 \end{aligned}$$

Note:

- we got this result w/o knowing the pdf of X!
- Data based-guarantees

Example 53

Assuming that a particular exam has 100 problems, each of which is a multiple choice problem with 5 choices. What is the chance of getting above 80 if some student randomly guessed all the answers? What about the chance of pass?

Solution:

Let X be the total score. Then,

- $X \sim B(100, p = 0.2)$. So $\mu = 100 \times 0.2 = 20$, and variance $= np(1-p) = 100 \times 0.2 \times 0.8 = 16$
- Standard deviation: $\sigma = \sqrt{16} = 4$

$$\begin{aligned} P(X > 80) &= P(X - 20 > 60) \leq P(|X - 20| > 60) = P(|X - 20| > 4 \times 15) \\ &\leq \frac{1}{15^2} \simeq 0.0044 \end{aligned}$$

Similarly

$$\begin{aligned} P(X > 60) &= P(X - 20 > 40) \leq P(|X - 20| > 40) = P(|X - 20| > 4 \times 10) \\ &\leq \frac{1}{10^2} \simeq 0.01 \end{aligned}$$

Solution – Continued

Actual probabilities: $X \sim B(100, 0.2)$

- $P(X > 80) = \sum_{n=81}^{100} \binom{100}{n} (0.2)^n (0.8)^{100-n} \simeq 4.8938 \times 10^{-39}$
- $P(X \geq 60) = \sum_{n=60}^{100} \binom{100}{n} (0.2)^n (0.8)^{100-n} \simeq 2.5158 \times 10^{-18}$

Example 54 (Gamma Distribution)

If X has Gamma distribution: $X \sim \Gamma(\alpha, \beta)$ with $\alpha = 16, \beta = 2$. Question: What is $P(X \geq 64)$?

Solution:

Estimate using Chebyshev's inequality:

- Expectation: $\mu = \alpha \cdot \beta = 32$
- Variance: $\sigma^2 = \alpha\beta^2 = 64 \quad \Rightarrow \sigma = 8$
- $P(X \geq 64) = P(|X - 32| \geq 32) = P(|X - 32| \geq 4 \cdot \sigma) \leq 1/4^2 \simeq 0.0625$
  [We used that $X \geq 0$ in the first equality]

Note:

- the result is only upper bound which is usually not too tight because we did not use the Gamma pdf!
- The result is a crude but guaranteed upper bound applicable to ALL r.v.'s with the same μ, σ^2!
- For comparison – actual result using that $X \sim \Gamma(\alpha, \beta)$ with $\alpha = 16, \beta = 2$:

$$P(X \geq 64) \approx .00066 \ll 0.0625$$

Definition 34 (Joint Distribution Function)

Let X_1 and X_2 be two random variables (discrete or continuous). The joint distribution function for X_1 and X_2 is defined as

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$$

Example 55

$X_1 \sim U(0,1)$, $X_2 \sim U(0,2)$. What is the joint distribution function?

Solution:

- Assume X_1, X_2 are independent.
- Since $0 \leq X_1 \leq 1,\ 0 \leq X_2 \leq 2$, only need to consider (x_1, x_2) with $0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 2$:

$$P(X_1 \leq x_1) = x_1 \ \text{ if } 0 \leq x_1 \leq 1 \qquad P(X_2 \leq x_2) = \frac{x_2}{2} \ \text{ if } 0 \leq x_2 \leq 2$$

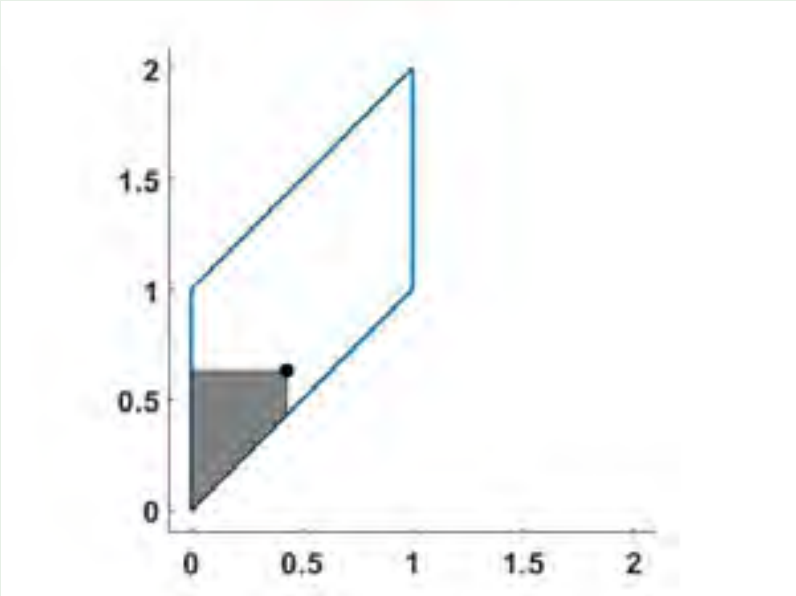
So

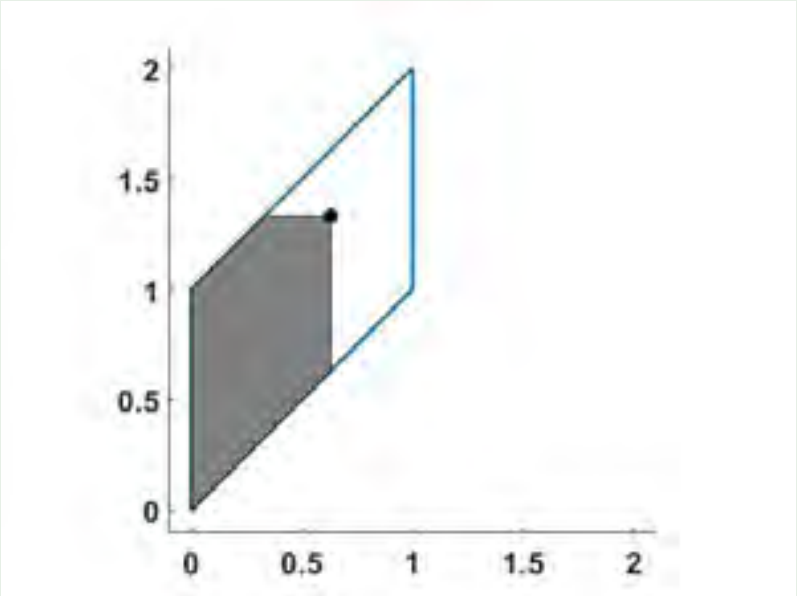
$$F(x_1, x_2) = \frac{x_1 x_2}{2} \qquad \text{if } 0 \leq x_1 \leq 1 \ \text{ and } \ 0 \leq x_2 \leq 2$$

Example 56 (Uniform Distributions)

$X_1 \sim U(0,1), X_2 \sim U(X_1, 1+X_1)$. What is the joint distribution function?

Solution: Only consider (x_1, x_2) with $0 \le x_1 \le 1,\ x_1 \le x_2 \le 1 + x_1$.





- Left: $0 \le x_1 \le 1, \quad x_1 \le x_2 \le 1$

$$F(x_1, x_2) = \text{Shaded area} = \frac{x_2 + (x_2 - x_1)}{2} x_1$$

- Right: $0 \le x_1 \le 1, \quad 1 \le x_2 \le 1 + x_1$

$$F(x_1, x_2) = \text{Shaded area} = 1 \cdot (x_2 - 1) + \frac{1 + (x_2 - x_1)}{2}[x_1 - (x_2 - 1)]$$

Theorem 35

The joint distribution function $F(x_1, x_2)$ has the following properties:

- $F(-\infty, -\infty) = F(-\infty, x_2) = F(-\infty, x_1) = 0$
- $F(\infty, \infty) = 1$
- *If $\hat{x}_1 > x_1$ and $\hat{x}_2 > x_2$, then*

$$F(\hat{x}_1, \hat{x}_2) - F(\hat{x}_1, x_2) - F(x_1, \hat{x}_2) + F(x_1, x_2) \geq 0$$

Proof.

Part 3:

$$F(\hat{x}_1, \hat{x}_2) - F(\hat{x}_1, x_2) - F(x_1, \hat{x}_2) + F(x_1, x_2) = P(x_1 < X_1 \leq \hat{x}_1, x_2 < X_2 \leq \hat{x}_2)$$

□

Definition 35

Let X_1 and X_2 be two continuous random variables with a joint distribution function $F(x_1, x_2)$. If there exists a nonnegative function $f(x_1, x_2)$ such that

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2) \mathrm{d}t_1 \mathrm{d}t_2$$

then $f(x_1, x_2)$ is called the joint probability density function.

Theorem 36

The joint probability density function $f(x_1, x_2)$ for continuous random variables X_1, X_2 has the following properties

- $f(x_1, x_2) \geq 0$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2 = 1$

Remark 12

Integral is the area under the curve:

$$P(a_1 \leq X_1 \leq a_2, b_1 \leq X_2 \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f(x_1, x_2)\, dx_1\, dx_2$$

Example 57 (Oil Storage)

- Y_1: available capacity at the start of the week, Y_2: oil sold during the week
- Assume the joint density function is

$$f(y_1, y_2) = 3y_1, \quad 0 \leq y_2 \leq y_1 \leq 1$$

- Find $P(0 \leq Y_1 \leq 0.5, Y_2 \leq 0.25)$

Solution:

$$P(0 \leq Y_1 \leq 0.5, Y_2 \leq 0.25) = \int_{\Omega} 3y_1 \, dy_1 dy_2$$

where the domain is the intersection of $0 \leq y_2 \leq y_1 \leq 1$ and $0 \leq Y_1 \leq 0.5, Y_2 \leq 0.25$. Therefore

$$P(0 \leq Y_1 \leq 0.5, Y_2 \leq 0.25) = \int_0^{0.25} \left[\int_{y_2}^{0.5} 3y_1 \, dy_1 \right] dy_2$$

$$= \int_0^{1/4} dy_2 \left(\frac{3}{2} y_1^2 \right) \Big|_{y_2}^{0.5} = \int_0^{1/4} \left[\frac{3}{8} - \frac{3}{2} y_2^2 \right] dy_2 = \left(\frac{3}{8} y_2 - \frac{1}{2} y_2^3 \right) \Big|_0^{1/4} = \frac{11}{128}$$

Example 58

X_1, X_2 are independent $U(0,1)$. Calculate $P(X_1 - X_2 > 0.5)$ and $P(X_1 X_2 < 0.5)$.

Solution:

- Joint distribution function:

$$F(x_1, x_2) = x_1 \cdot x_2 \qquad \text{when } 0 \leq x_1, x_2 \leq 1$$

  elsewhere does NOT matter

- Clearly, the joint probability density function is:

$$f(x_1, x_2) = 1, \qquad \text{for } \ 0 \leq x_1, x_2 \leq 1$$

  and zero elsewhere.

- The probability is simply the area of region that satisfies the condition.

Definition 36 (Marginal Density Function)

Let X_1 and X_2 be jointly continuous random variables with joint density function $f(x_1, x_2)$. The marginal density function of X_1 and X_2, respectively, are given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_2, \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_1$$

Example 59

Let $f(x_1, x_2) = \begin{cases} 2x_1, & 0 \le x_1 \le 1, 0 \le x_2 \le 1 \\ 0, & \text{elsewhere} \end{cases}$ Find the marginal density functions for X_1 and X_2.

Solution:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 = \int_0^1 2x_1 dx_2 = 2x_1, \quad 0 \le x_1 \le 1$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1 = \int_0^1 2x_1 dx_1 = 1, \quad 0 \le x_2 \le 1$$

Example 60 (Oil Storage)

Assume the joint density for continuous random variables Y_1, Y_2 is

$$f(y_1, y_2) = \begin{cases} 3y_1, & 0 \le y_2 \le y_1 \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the marginal density functions for Y_1 and Y_2.

Solution:

Marginal density:

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_2 = \int_0^{y_1} 3y_1 dy_2 = \boxed{3y_1^2}, \qquad 0 \le y_1 \le 1$$

and

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_1 = \int_{y_2}^{1} 3y_1 dy_1 = \frac{3}{2}y_1^2\Big|_{y_2}^{1} = \boxed{\frac{3}{2}(1 - y_2^2)}, \ 0 \le y_2 \le 1$$

Example 61 (Uniform Distributions)

Let $X_1 \sim U(0,1)$, $X_2 \sim U(X_1, X_1+1)$. Find the joint pdf, and marginal pdf for X_1, X_2.

Solution:

- Joint pdf: $f(x_1,x_2) = 1$, for $0 \le x_1 \le 1,\ x_1 \le x_2 \le 1+x_1$
- Marginal density function of X_1,

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_2 = \int_{x_1}^{1+x_1} 1\,dx_2 = \boxed{1}, \qquad \text{for } 0 \le x_1 \le 1$$

- Marginal density function of X_2

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_1 = \begin{cases} \int_0^{x_2} 1\,dx_1 = \boxed{x_2}, & \text{if } 0 \le x_2 \le 1 \\ \int_{x_2-1}^{1} 1\,dx_1 = \boxed{2-x_2}, & \text{if } 1 \le x_2 \le 2 \end{cases}$$

Definition 37 (Conditional Probability Function)

If X_1 and X_2 are jointly discrete random variables with joint probability function $p(x_1, x_2)$ and marginal probability functions $p_1(x_1)$, $p_2(x_2)$, respectively, then the conditional discrete probability function of X_1 given X_2 is

$$p(x_1|x_2) = P(X_1 = x_1 \,|\, X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} = \frac{p(x_1, x_2)}{p_2(x_2)}$$

Definition 38 (Conditional Distribution Function)

If X_1 and X_2 are continuous random variables, then the conditional distribution function of X_1 given $X_2 = x_2$ is

$$F(x_1|x_2) = P([X_1 \leq x_1 \,|\, X_2 = x_2)\,.$$

- So $F(x_1) = \int_{-\infty}^{\infty} F(x_1|x_2) f_2(x_2) dx_2$
- Also

$$F(x_1) = \int_{-\infty}^{x_1} f_1(t_1) dt_1 = \int_{-\infty}^{x_1} \left[\int_{-\infty}^{\infty} f(t_1, x_2) dx_2 \right] dt_1$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{x_1} f(t_1, x_2) dt_1 \, dx_2$$

Therefore: $F(x_1|x_2) f_2(x_2) = \int_{-\infty}^{x_1} f(t_1, x_2) dt_1$

$$\Rightarrow F(x_1|x_2) = \int_{-\infty}^{x_1} \frac{f(t_1, x_2)}{f_2(x_2)} dt_1$$

Definition 39 (Conditional Density)

Let X_1 and X_2 be jointly continuous random variables with joint density function $f(x_1, x_2)$ and marginal density function $f_1(x_1)$, $f_2(x_2)$, respectively. The conditional density function of X_1 given $X_2 = x_2$ is

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}. \qquad \text{Similarly,} \quad f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

Example 62 (Oil Storage)

- Y_1: available capacity at start of the week, Y_2: oil sold during the week
- Assume the joint density function is

$$f(y_1, y_2) = 3y_1, \quad 0 \le y_2 \le y_1 \le 1$$

- Find $f(y_1|y_2 = 0.5), f(y_2|y_1 = 1)$ and $P(Y_2 \le 0.5|Y_1 = 1)$

Solution:

From previous calculations, the marginal density:

$$f_1(y_1) = 3y_1^2, \ \ 0 \le y_1 \le 1; \qquad f_2(y_2) = \frac{3}{2}(1 - y_2^2), \ \ 0 \le y_2 \le 1$$

Hence,

- $f(y_1|y_2 = 0.5) = \dfrac{f(y_1, 0.5)}{f_2(0.5)} = \dfrac{8}{3}y_1, \quad$ when $0.5 \le y_1 \le 1$
- $f(y_2|y_1 = 1) = \dfrac{f(y_1 = 1, y_2)}{f_1(1)} = 1, \quad$ when $0 \le y_2 \le 1$
- $P(Y_2 \le 0.5|Y_1 = 1) = \displaystyle\int_{-\infty}^{0.5} f(y_2|y_1 = 1)dy_2 = \int_0^{0.5} 1 dy_2 = 0.5$

Sec 5.4: Independent Random Variables (continuous)

Definition 40 (Independent)

Let X_1 have distribution function $F_1(x_1)$, X_2 have distribution function $F_2(x_2)$, and X_1, X_2 have the joint distribution function $F(x_1, x_2)$. Then X_1 and X_2 are independent if and only if

$$F(x_1, x_2) = F(x_1)F(x_2)$$

for every pair of x_1, x_2.

Theorem 37

If X_1 and X_2 are jointly discrete random variables with joint probability function $p(x_1, x_2)$ and marginal probability functions $p_1(x_1)$, $p_2(x_2)$, respectively, then X_1 and X_2 are independent if and only if

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \qquad \text{for every pair of } x_1, x_2.$$

Theorem 38

Let X_1 and X_2 be jointly continuous random variables with joint density function $f(x_1, x_2)$ and marginal density function $f_1(x_1)$, $f_2(x_2)$, respectively, then X_1 and X_2 are independent if and only if

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \qquad \text{for every pair of } x_1, x_2.$$

Example 63

- $f(x_1, x_2) = 6x_1^2 x_2$ for $0 \leq x_1, x_2 \leq 1 \longrightarrow$ independent
- Oil storage problem: X_1 storage at the beginning of the week, X_2 oil sold during the week. Joint distribution $p(x_1, x_2) = 3x_1$ for $0 \leq x_2 \leq x_1 \leq 1 \longrightarrow$ dependent
- $X_1 \sim U(0,1),\ X_2 \sim U(X_1, 1 + X_1) \longrightarrow$ dependent

Theorem 39 (Independence Check)

Let X_1 and X_2 have a joint density function $f(x_1, x_2)$ which is positive if and only if $a \leq x_1 \leq b$ and $c \leq x_2 \leq d$ for some constants a, b, c, d. Then X_1 and X_2 are independent if and only if

$$f(x_1, x_2) = h(x_1)g(x_2)$$

for certain nonnegative functions g and h.

Example 64

X_1, X_2 have a joint density function:

$$f(x_1, x_2) = \begin{cases} 2x_1, & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, \\ 0, & \text{elsewhere} \end{cases}$$

Are X_1, X_2 independent?

Sec 5.5: Expected Value (continuous)

Definition 41 (Expectations of Multivariate Random Variables)

Let $g(X_1, \cdots, X_n)$ be a function of random variables.

- If $X_1, \cdots, X_n$ are discrete random variables with a probability function $p(x_1, \cdots, x_n)$, then

$$\mathbb{E}[g(X_1, \cdots, X_n)] = \sum_{x_1, \cdots, x_n} g(x_1, \cdots, x_n) p(x_1, \cdots, x_n)$$

- If $X_1, \cdots, X_n$ are continuous random variables with a probability density function $f(x_1, \cdots, x_n)$, then

$$\mathbb{E}[g(X_1, \cdots, X_n)] = \int_{\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_n) f(x_1, \cdots, x_n) \mathrm{d}x_1 \cdots \mathbf{d}x_n \, .$$

Example 65

Let X_1, X_2 have a joint density function given by

$$f(x_1, x_2) = \begin{cases} 2x_1 & 0 \leq x_1, x_2 \leq 1 \\ 0 & \text{elsewhere} \end{cases} \, .$$

Find $\mathbb{E}[X_1 X_2]$. Find $Var(X_1)$. Find $\mathbb{E}[X_1 - X_2]$.

Example 66

Toss two dices. X_1= number shown on top of die 1, and X_2 = number shown on top of die 2. Find the expectation of $X_1 + X_2$, and the expectation of $X_1 X_2$.

Theorem 40

$\mathbb{E}[c] = c$

Theorem 41

Let $g(X_1, X_2)$ be a function of random variables, then

$$\mathbb{E}[cg(X_1, X_2)] = c\mathbb{E}[g(X_1, X_2)]\,.$$

Theorem 42

Let $g_1(X_1, X_2), \cdots, g_n(X_1, X_2)$ be functions of random variables, then

$$\mathbb{E}\left[\sum_{i=1}^{n} g_i(X_1, X_2)\right] = \sum_{i=1}^{n} \mathbb{E}[g_i(X_1, X_2)]\,.$$

Example 67

Toss two dices. X_1= number shown on top of die 1, and X_2 = number shown on top of die 2. Find the expectation of $X_1 + X_2$, and the expectation of $X_1 X_2$.

Theorem 43

Let X_1, X_2 be independent random variables and g_1, g_2 be two functions. Then

$$\mathbb{E}[g_1(X_1)g_2(X_2)] = \mathbb{E}[g_1(X_1)] \cdot \mathbb{E}[g_2(X_2)]$$

Example 68

Toss two dices. X_1= number shown on top of die 1, and X_2 = number shown on top of die 2. Find the expectation of $X_1 + X_2$, and the expectation of X_1X_2.

Example 69

Find $\mathbb{E}[X_1X_2]$, where X_1, X_2 have a joint density function

$$f(x_1, x_2) = \begin{cases} 2x_1 & 0 \leq x_1, x_2 \leq 1 \\ 0 & \text{elsewhere} \end{cases}.$$

Example 70

Let X_1, X_2 have joint density function

$$f(x_1, x_2) = \begin{cases} e^{-(x_1+x_2)} & x_1, x_2 > 0 \\ 0 & \textit{elsewhere} \end{cases}$$

Find $\mathbb{E}[X_1 + X_2]$, $Var[X_1 + X_2]$, $P(X_1 - X_2 > 3)$, $Var[X_1 - X_2]$.

Definition 42 (Covariance Function)

Random variables X_1, X_2 have mean μ_1, μ_2 and standard deviation σ_1, σ_2. The covariance function is defined as

$$Cov(X_1, X_2) = \mathbb{E}\left[(X_1 - \mu_1) \cdot (X_2 - \mu_2)\right]$$

Correlation coefficient is $\rho = \dfrac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}$

Theorem 44

$$Cov(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \cdot \mathbb{E}[X_2]\,.$$

If X_1 and X_2 are independent, then $Cov(X_1, X_2) = 0$.

Remark 13

$Cov(X_1, X_2)$ measures how X_1 change when X_2 changes. If $Cov(X_1, X_2) = 0$, we say X_1, X_2 are uncorrelated.

Example 71 (Oil Storage)

Find $Cov(X_1, X_2)$, where X_1 and X_2 have joint distribution

$$f(x_1, x_2) = 3x_1, \quad 0 \le x_2 \le x_1 \le 1\,.$$

Solution:

The marginal density functions:

$$f_1(y_1) = 3y_1^2, \;\; 0 \le y_1 \le 1; \qquad f_2(y_2) = \frac{3}{2}(1 - y_2^2), \;\; 0 \le y_2 \le 1$$

Example 72 (Uniform Distributions)

Given $X_1 \sim U(0,1)$, $X_2 \sim U(X_1, X_1 + 1)$, find $Cov(X_1, X_2)$.

Solution:

- Joint pdf: $f(x_1, x_2) = 1$, for $0 \le x_1 \le 1,\ x_1 \le x_2 \le 1 + x_1$
- From previous calculations, the marginal density function of X_1,

$$f_1(x_1) = 1, \qquad \text{for } 0 \le x_1 \le 1$$

- Marginal density function of X_2

$$f_2(x_2) = \begin{cases} \int_0^{x_2} 1\, dx_1 = x_2, & \text{if } 0 \le x_2 \le 1 \\ \int_{x_2-1}^{1} 1\, dx_1 = 2 - x_2, & \text{if } 1 \le x_2 \le 2 \end{cases}$$

Example 73 (Uniform Distributions)

$X_1 \sim U(0,1), X_2 = U(X_1, 1+X_1)$. Calculate the variance of $X_1 - X_2$.

Solution:

$$Var(X_1 - X_2) = Var(X_1) + Var(X_2) - 2Cov(X_1, X_2)$$

Theorem 45

$$Cov^2(X_1, X_2) = \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)]^2 \leq \mathbb{E}[(X_1 - \mu_1)^2] \cdot \mathbb{E}[(X_2 - \mu_2)^2]$$
$$= Var(X_1) Var(X_2)$$

Hence, $-1 \leq \rho(X_1, X_2) \leq 1$. *Furthermore,*

- $\rho = 1 \quad \Leftrightarrow \quad Y = aX + b$ *with* $a > 0$
- $\rho = -1 \quad \Leftrightarrow \quad Y = aX + b$ *with* $a < 0$

Example 74

$X_1 \sim U(-1,1)$ and $X_2 = \cos X_1$. Are X_1 and X_2 independent? What is $Cov(X_1, X_2)$?

Sec 5.8: Linear Functions of Random Variables (continuous)

Theorem 46

Let $X_1, \cdots X_n$ and $Y_1, \cdots, Y_m$ be sequences of random variables. Let

$$U = \sum a_i X_i, \quad V = \sum b_j Y_j \, .$$

- $\mathbb{E}[U] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$
- $Var(U) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j Cov(X_i, X_j)$
- $Cov(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$

Proof:

Show the case of two variables.

Example 75 (Oil Storage)

Random variables X_1, X_2 has the joint density function

$$f(x_1, x_2) = 3x_1, \qquad \text{if } 0 \leq x_2 \leq x_1$$

Find marginal distribution, conditional distribution given that $x_1 = 0.5$, expectation $\mathbb{E}[X_1 X_2]$, $\mathbb{E}[X_1 - X_2]$, covariance, correlation coefficient, and variance of $X_1 - X_2$.

Example 76

A box has r red balls and $N-r$ blue balls. A random sample of n balls is drawn without replacement. Let X be the number of red balls drawn. Find X's mean and variance (X has a hypergeometric distribution)

Solution:

- Label the balls as $1, 2, \ldots, n$, and define

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th draw is a red ball} \\ 0, & \text{otherwise} \end{cases}$$

- Clearly, $P(X_k = 1) = \dfrac{r}{N}$ for $k = 1, \ldots, n$ because for k-th ball, r out of N choices are red

$$\mathbb{E}[X_k] = \frac{r}{N}, \quad \mathbb{E}[X_k^2] = \frac{r}{N}, \quad Var(X_k) = \mathbb{E}[X_k^2] - \mathbb{E}[X_k]^2 = \frac{r}{N}\left(1 - \frac{r}{N}\right)$$

- When $j \neq k$, $\binom{r}{2}$ out of $\binom{N}{2}$ choices lead to red j- and k-th ball. Hence

$$P(X_j = 1, X_k = 1) = \frac{\binom{r}{2}}{\binom{N}{2}} = \frac{r(r-1)}{N(N-1)}$$

Solution – Continued

- Covariance, when $i \neq j$

$$\begin{aligned} Cov(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \cdot \mathbb{E}[X_j] = \frac{r(r-1)}{N(N-1)} - \left(\frac{r}{N}\right)^2 \\ &= -\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1} \end{aligned}$$

- Clearly, $X = \sum_{i=1}^{n} X_i \qquad \Rightarrow \mathbb{E}[X] = n\frac{r}{N}$

Finally:

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j < n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} \frac{r}{N}\left(1 - \frac{r}{N}\right) + 2 \sum_{1 \leq i < j < n} \left[-\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1}\right] \\ &= n \cdot \frac{r}{N}\left(1 - \frac{r}{N}\right) - n(n-1) \cdot \left[\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{1}{N-1}\right] \\ &= n\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-n}{N-1} \end{aligned}$$

Law of Large Numbers + Finite Sample Guarantees

$X_1, \cdots, X_n, \cdots$ be independent random variables with mean μ and variance σ^2.

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Claim: $\bar{X}_n$ converges to μ in probability, ie,

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

Proof:

- Clearly, $\mathbb{E}[\bar{X}] = \mu$
- Variance:

$$Var(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n^2} V(X_i) + \frac{2}{n^2} \sum_{1\le i<j\le n} Cov(X_i, X_j) = \frac{\sigma^2}{n} .$$

- Use Chebyshev's inequality:

$$P(|\bar{X}_n - \mu| > \epsilon) \le \frac{\sigma^2}{\epsilon^2 n} \qquad \to 0 \quad \text{as } n \to \infty$$

Law of Large Numbers+Finite Sample Guarantees

We proved the Law of Large Numbers: $\bar{X}_n$ converges to μ in probability, ie,

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

- "Modes of Convergence": Convergence in probability (there are others!)
- Universality Theorem: no matter the pdf of X_i's!
- Finite Sample Guarantees: On the way to the proof we obtained:

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

  - n : can be interpreted as the number of IID samples from X_1
  - For finite n we can estimate how close we are to the mean μ
  - Rate of convergence: $\frac{\sigma^2}{\epsilon^2 n}$

Bivariate Normal Distribution

- Joint density function:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-Q/2}$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\beta\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]$$

- Mean: $\mathbb{E}[X_i] = \mu_i,\ i = 1, 2$
- Variances: $Var(X_i) = \sigma_i,\ i = 1, 2,$ and $Cov(X_1, X_2) = \rho\sigma_1\sigma_2.$
- ρ : Correlation coefficient
- For Gaussians, $\rho = 0 \quad \Rightarrow \quad$ independence:

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)$$

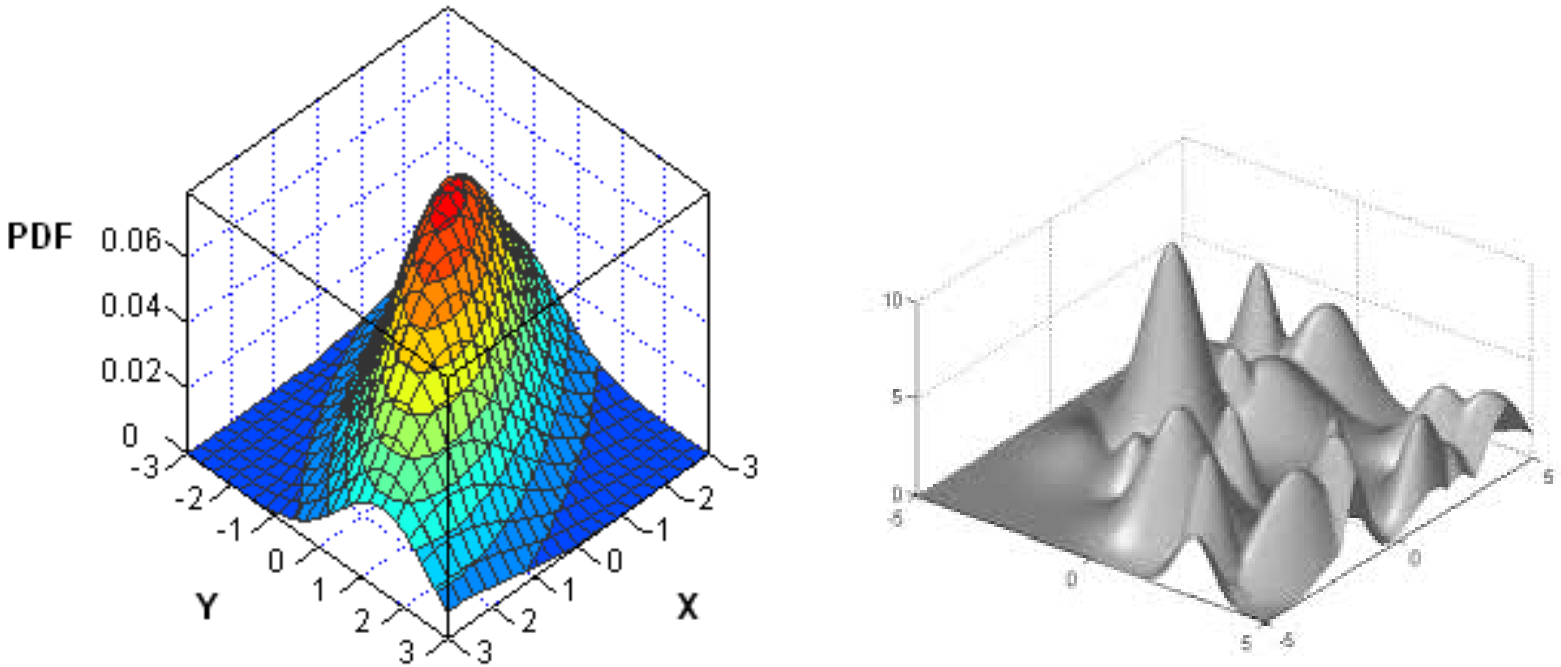


Figure: (a) Bivariate Gaussian, (b) Mixture model of bivariate Gaussians

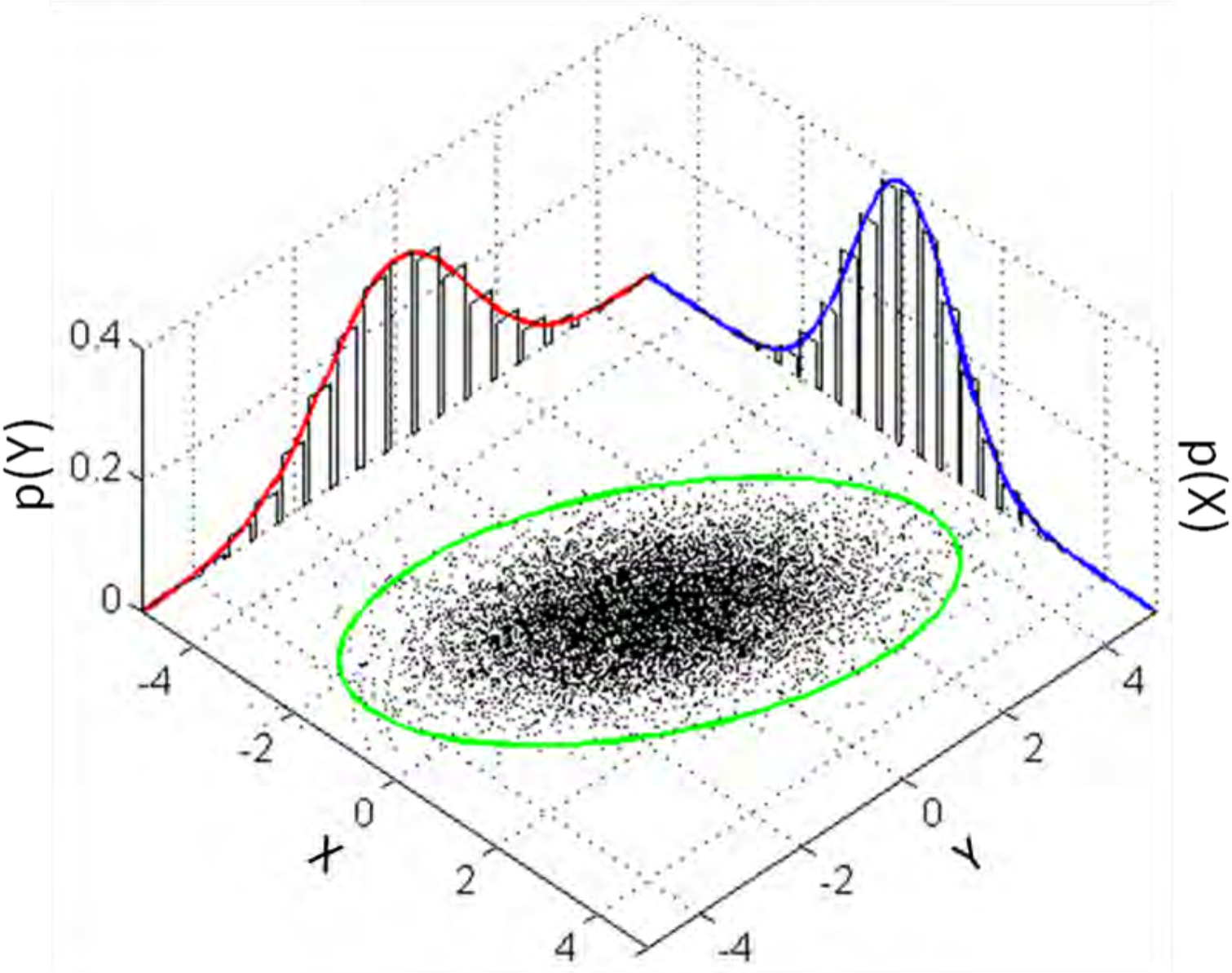


Figure: Marginals, conditionals, etc

Definition 43 (Conditional Expectation)

Let X_1 and X_2 be two random variables.

$$\mathbb{E}[g(X_1) \,|\, X_2 = x_2] = \int_{-\infty}^{\infty} g(x_1) f(x_1|x_2) \mathrm{d}x_1 \qquad \text{if continuous}$$

and

$$\mathbb{E}[g(X_1) \,|\, X_2 = x_2] = \sum_{x_1} g(x_1) p(x_1|x_2) \qquad \text{if discrete}$$

Remark 14

- Because $\mathbb{E}[X_1|X_2]$ is a function of the r.v. X_2, it is itself a random variable.
- And as such, $\mathbb{E}[X_1|X_2]$ has a mean and a variance: see next two Theorems
- Practical use: Conditioning
  - Fixes some of the random variables to simplify calculations
  - Injects known info (and reduces uncertainty/variance)

Example 77 (Oil Storage – Conditional Expectation)

Y_1: available capacity at the start of the week, Y_2: oil sold during the week. Assume the joint density for continuous random variables Y_1, Y_2 is

$$f(y_1, y_2) = \begin{cases} 3y_1, & 0 \le y_2 \le y_1 \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Predict expected capacity given 50% sale of inventory: compute $\mathbb{E}[Y_1 \mid Y_2 = 0.5]$

Solution:

- From previous calculations, the marginal density functions are:

$$f_1(y_1) = 3y_1^2, \ \ 0 \le y_1 \le 1; \qquad f_2(y_2) = \frac{3}{2}(1 - y_2^2), \ \ 0 \le y_2 \le 1$$

and

$$f(y_1|y_2 = 0.5) = \frac{f(y_1, 0.5)}{f_2(0.5)} = \frac{8}{3}y_1, \quad \text{when} \quad 0.5 \le y_1 \le 1$$

- Thus:

$$\mathbb{E}[Y_1 \mid Y_2 = 0.5] \quad = \quad \int_{-\infty}^{\infty} y_1 f(y_1|y_2 = 0.5) dy_1 = \int_{0.5}^{1} y_1 \frac{8}{3} y_1 dy_1 = \frac{7}{9}$$

Example 78 (Uniform Distributions)

Given that $X_1 = U(0,1)$, $X_2 = U(X_1, 1+X_1)$. Calculate $\mathbb{E}[X_1 \mid X_2 = 0.5]$

Theorem 47 (Law of Total Expectation)

$$\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1|X_2]]$$

Proof.

$$\begin{aligned}
\mathbb{E}[X_1] &= \int_{-\infty}^{\infty} x_1 f_1(x_1) dx_1 = \int_{-\infty}^{\infty} x_1 \left[\int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \right] dx_1 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1|x_2) f_2(x_2) dx_2 dx_1 \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x_1 f(x_1|x_2) dx_1 \right] f_2(x_2) dx_2 \\
&= \int_{-\infty}^{\infty} \mathbb{E}[X_1|X_2 = x_2] \, f_2(x_2) dx_2 \\
&= \mathbb{E}[\mathbb{E}[X_1|X_2]]
\end{aligned}$$

□

Example 79 (Total Expectation)

Suppose that a company has determined that the number of jobs per week, N, varies from week to week and has a Poisson distribution with mean λ. The number of hourse to complete each job, Y_i, is Gamma distributed with parameters α and β. The total time[a] to complete all jobs in a week is

$$T = \sum_{i=1}^{N} Y_i$$

[a] Note that T is the sum of a **random number** N of **random** variables Y_i

Solution: Use conditioning to "divide and conquer"

1. What is $\mathbb{E}(T|N=n)$? (Idea: fix one of the random variables)
   $\mathbb{E}(T|N=n) = \mathbb{E}(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = \sum_{i=1}^{n} \alpha\beta = n\alpha\beta.$
2. What is $\mathbb{E}(T)$? (Use Total Expectation)
   $\mathbb{E}(T) = [\mathbb{E}(T|N)] = \mathbb{E}(N\alpha\beta) = \lambda\alpha\beta.$

Example 80

There are 10 red balls, and 6 blue balls. Randomly draw a ball. If red, toss a coin 5 times, if blue, toss a coin 6 times. What is the expected number of heads?

Solution:

Let $X=$ number of heads, and $Y=$ number of times of coin toss:

- $\mathbb{E}[Y] = 5 \cdot \frac{10}{16} + 6 \cdot \frac{6}{16} = \frac{86}{16}$
- For a given Y, X has binomial distribution $B\left(y, \frac{1}{2}\right)$. So
  $\mathbb{E}[X|Y=y] = \frac{y}{2}$
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}\left[\frac{Y}{2}\right] = \frac{86}{32} = \frac{43}{16}$

Conditional Variance

$$Var(X_1|X_2=x_2) = \mathbb{E}[X_1^2|X_2=x_2] - \mathbb{E}[X_1|X_2=x_2]^2$$

Replace the density function in the variance with conditional density function!

Remark 15

Both $\mathbb{E}[X_1|X_2=x_2]$ and $Var(X_1|X_2=x_2)$ are random variables.

Theorem 48 (Conditioning & Variance Reduction)

$$Var(X_1) = \mathbb{E}[Var(X_1|X_2)] + Var\left(\mathbb{E}[X_1|X_2]\right)$$

Conditioning reduces variance (randomness!):

$$Var(X_1) \geq Var\left(\mathbb{E}[X_1|X_2]\right)$$

Proof.

$$\begin{aligned}
Var(X_1) &= \mathbb{E}[X_1^2] - \{\mathbb{E}[X_1]\}^2 = \mathbb{E}\left[\mathbb{E}[X_1^2|X_2]\right] - \{\mathbb{E}\left[\mathbb{E}[X_1|X_2]\right]\}^2 \\
&= \left(\mathbb{E}\left[\mathbb{E}[X_1^2|X_2]\right] - \mathbb{E}\left[\{\mathbb{E}[X_1|X_2]\}^2\right]\right) \\
&\quad + \left(\mathbb{E}\left[\{\mathbb{E}[X_1|X_2]\}^2\right] - \{\mathbb{E}\left[\mathbb{E}[X_1|X_2]\right]\}^2\right) \\
&= \mathbb{E}\left[\mathbb{E}[X_1^2|X_2] - \{\mathbb{E}[X_1|X_2]\}^2\right] + Var\left(\mathbb{E}[X_1|X_2]\right) \\
&= \mathbb{E}[Var(X_1|X_2)] + Var\left(\mathbb{E}[X_1|X_2]\right)
\end{aligned}$$

□

Example 81

There are 10 red balls, and 6 blue balls. Randomly draw a ball. If red, toss a coin 5 times, if blue, toss a coin 6 times. What is the variance for the number of heads?

Solution:

Let $X=$ number of heads, and $Y=$ number of times of coin toss. Then for a given Y, X has binomial distribution $B\left(y, \frac{1}{2}\right)$. $\left(\mathbb{E}[Y]=\frac{86}{16}\right)$

- So $Var(X|Y)=npq=y\cdot\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{y}{4}$, $\mathbb{E}[X|Y]=np=y\cdot\frac{1}{2}=\frac{y}{2}$
- Hence,

$$\begin{aligned} Var(X) &= \mathbb{E}[Var(X|Y)]+Var(\mathbb{E}[X|Y])=\mathbb{E}\left[\frac{Y}{4}\right]+Var\left(\frac{Y}{2}\right) \\ &= \frac{1}{4}\mathbb{E}[Y]+\left(\frac{1}{2}\right)^2 Var(Y) \end{aligned}$$

- And $Var(Y)=\mathbb{E}[Y^2]-\mathbb{E}[Y]^2=\left(5^2\cdot\frac{10}{16}+6^2\cdot\frac{6}{16}-\left(\frac{86}{16}\right)^2\right)=\frac{15}{64}$
- Finally, $Var(X)=\frac{1}{4}\cdot\frac{86}{16}+\frac{1}{4}\cdot\frac{15}{64}=\frac{359}{256}$

Motivation: a "Netflix" example

- Consider outcomes of some complex process, e.g. predicting if one customer (out of n) will watch a proposed movie:

$$X_i = 1 \quad \text{prediction is wrong} \qquad X_i = 0 \quad \text{prediction is correct}$$

- $\bar{X}_n = \frac{1}{n}\sum_i X_i$: observed error rate
- Build a simple data-based prediction model: $X_i \sim$ Bernoulli with mean p.
- Use the estimator $p \approx \bar{X}_n = \frac{1}{n}\sum_i X_i$. Good idea? Quantify your answer!
- Estimate the confidence interval

$$P(\bar{X}_n - \epsilon \leq p \leq \bar{X}_n + \epsilon) = P(|\bar{X}_n - p| < \epsilon), \qquad \text{Or equivalent estimate: } P(|\bar{X}_n -$$

Ideas

1. LLN: Use Chebyshev Inequality: $P(|\bar{X}_n - p| \geq \epsilon) \leq \dfrac{Var(\bar{X}_n)}{\epsilon^2} = ???$
2. CLT: if I knew the **distribution of** $\bar{X}_n$ (for $n >> 1$), I could compute anything, including

$$P(|\bar{X}_n - p| \geq \epsilon)$$

Law of Large Numbers+Finite Sample Guarantees

$X_1, \cdots, X_n, \cdots$ be independent random variables with mean μ and variance σ^2.

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Claim: $\bar{X}_n$ converges to μ in probability, ie,

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

Proof:

- Clearly, $\mathbb{E}[\bar{X}_n] = \mu$
- Variance – **from Sec 5.8**:

$$Var(\bar{X}_n) = \sum_{i=1}^{n} \frac{1}{n^2} V(X_i) + \frac{2}{n^2} \sum_{1\le i<j\le n} Cov(X_i, X_j) = \frac{\sigma^2}{n} .$$

- Use Chebyshev's inequality: $P(|X - E[X]| \ge \epsilon) \le \frac{Var(\mathbf{X})}{\epsilon^2}$

$$P(|\bar{X}_n - \mu| > \epsilon) \le \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n} \quad \to \quad 0 \quad \text{as } n \to \infty$$

Law of Large Numbers+Finite Sample Guarantees

We proved the Law of Large Numbers: $\bar{X}_n$ converges to μ in probability, ie,

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

- **"Modes of Convergence":** Convergence in probability (there are others!)
- **Universality** Theorem: no matter the pdf of X_i's!
- **Finite Sample Guarantees:** On the way to the proof we obtained:

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

  - n : can be interpreted as the number of IID samples from X_1
  - For finite n we can estimate how close we are to the mean μ
  - Rate of convergence: $\frac{\sigma^2}{\epsilon^2 n}$

"Netflix" example – continued

- Chebyshev Inequality/LLN: (recall p is unknown!)

$$P(|\bar{X}_n - p| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

- Hoeffding: (recall $0 \leq X_1 \leq 1$)

$$P(|\mu_N - p| \geq \epsilon) \leq 2e^{-2N\epsilon^2/(b-a)^2} = 2e^{-2N\epsilon^2}$$

- Hoeffding much tighter than Chebyshev, e.g. $n = 100$, $\epsilon = .2$
  Hoeffding= .00067, Chebyshev= .0625

Introduction

- Statistic: A function of the observable random variables in a sample $\{X_1, ..., X_n\}$:

$$\text{Sample mean: } \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{Sample variance: } S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$$

Recall: $\bar{X}_n$ as an estimator for μ (recall "netflix example"):

- (LLN) Infinite limit of samples: $\lim_{n\to\infty} \bar{X}_n = \mu$ "in probability", i.e.

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

- (Chebyshev's inequality) For any finite samples n:

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

- **Can we do better?** What does that mean really? To be explored....

Theorem 49 (Central Limit Theorem)

Let $X_1, \cdots, X_n$ be i.i.d random variables with $\mathbb{E}[X_i] = \mu$ and $V(X_i) = \sigma^2$. Define

$$U_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\,.$$

*Then the distribution of U_n converges to the standard **normal** distribution:*

$$\lim_{n\to\infty} P(U_n \le u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t = P(Z \le u)\,.$$

Remark 16

Practically: If $X_1, \cdots, X_n$ be i.i.d with $\mathbb{E}[X_i] = \mu$ and $V(X_i) = \sigma^2$

- for $n >> 1$, $\quad U_n \approx Z \sim \mathcal{N}(0,1), \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- We can compute probabilities of $\bar{X}_n$ (approximately) as

$$P(\bar{X}_n \le a) = P\left(\underbrace{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}_{=U_n} \le \frac{a-\mu}{\sigma/\sqrt{n}}\right) \approx P(Z \le \frac{a-\mu}{\sigma/\sqrt{n}})$$

Summary of (some) Limit Theorems

- (Strong) Law of Large Numbers:

$$\bar{X}_n = \frac{S_n}{n}, \quad S_n = \sum_{i=1}^{n}$$

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \bar{X}_n = \mu = 1$$

- Central Limit Theorem:

$$\lim_{n\to\infty} P\left(U_n := \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz$$

where $\Phi = \Phi(x)$ is the CDF of the standard normal $\mathcal{N}(0, 1)$.

Example 82 (How to use CLT in practice?)

Test scores for all high school seniors in one state has an average of 60 and variance 64. What is the probability a random sample of 100 students from one high school has average scores less than 58?

Solution:

Denote $\bar{X}_n$ as the mean of a random sample of $n = 100$ scores from a population with $\mu = 60, \sigma^2 = 64$. Need to estimate $P(\bar{X}_n \leq 58)$.

- $U_n = \dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ has a distribution that can be approximated by $N(0,1)$.
- So, because $n = 100 >> 1$ we can approximate ($\approx$) using CLT

$$P(\bar{X} \leq 58) = P\left(\frac{\bar{X} - 60}{8/\sqrt{100}} \leq \frac{58 - 60}{0.8}\right) \approx P(Z \leq -2.5) \simeq 0.0062$$

Normal Confidence Intervals

- Practically estimate a (95%) Confidence Interval:

$$P\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx P(-1.96 \leq Z \leq 1.96)$$

  Recall that $P(-1.96 \leq Z \leq 1.96) = 95\%$.

- In [STAT516] What if σ^2 not known (most common!)? Use the statistical estimator

$$\widehat{\sigma^2}_n = \frac{1}{n-1} = \sum_{i=i}^{n}(X_i - \hat{X}_n)^2$$

Berry-Esseen Theorem: Confidence Intervals revisited

- Rigorous CLT statement:

$$P\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) \to P(-1.96 \leq Z \leq .1.96) = .95$$

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

- How fast? **How many n's need** for good approximation?
- Berry Esséen Inequality (rate of convergence!):
  For $U_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}$ we have:

$$\sup_z \left| P(U_n \leq x) - \Phi(x) \right| \leq \frac{33}{4} \frac{E|X_1 - \mu|^3}{\sqrt{n}\sigma^3}$$

- Answers the question: **how big n should be to use** CLT approximations such as

$$P\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx P(-1.96 \leq Z \leq .1.96) = .95$$

- Depends on the "skewness" terms in the bound above

Example 83

Service times for customers at a certain counter are independent random variables with mean 1.5 minutes and standard deviation 1.0. Approximate the probability that 100 customers can be served in less than 2 hours.

Solution:

Let X_i denote the service time for i-th customer. Then

$$\begin{aligned} P\left(\sum_{i=1}^{100} X_i \leq 120\right) &= P\left(\bar{X}_n \leq \frac{120}{100}\right) = P(\bar{X}_n \leq 1.20) \\ &= P\left(\frac{\bar{X}_n - 1.50}{1/\sqrt{100}} \leq \frac{1.20 - 1.50}{1/\sqrt{100}}\right) \simeq P(Z \leq -3) \\ &\simeq 0.0013 \end{aligned}$$

where $Z = \dfrac{\bar{X}_n - 1.50}{1/\sqrt{100}}$ has a distribution that is approximately $N(0,1)$

Example 84 (Fairness)

Test scores for all high school seniors in one state has an average of 60 and variance 64. A random sample of 100 students from one high school has mean 58. Is this evidence that this high school is inferior?

Solution:

Denote $\bar{X}_n$ as the mean of a random sample of $n = 100$ scores from a population with $\mu = 60, \sigma^2 = 64$. Need to estimate $P(\bar{X}_n \leq 58)$.

- $U_n = \dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ has a distribution that can be approximated by $N(0,1)$.
- So, because $n = 100 >> 1$ we can approximate ($\approx$) using CLT

$$P(\bar{X} \leq 58) = P\left(\frac{\bar{X} - 60}{8/\sqrt{100}} \leq \frac{58 - 60}{0.8}\right) \approx P(Z \leq -2.5) \simeq 0.0062$$

The probability is very low. Therefore, it is unlikely that the sample from this school can be regarded as a random sample from that general population.

Summary

We really use something like this provided n is large enough (as quantified...)

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx P(-1.96 \leq Z \leq .1.96) = .95$$

Can answer different questions:

- Compute probabilities using $Z \sim N(0, 1)$
- For fixed tolerances find how big n or should be
- How big σ, etc...

Sec 7.5:Normal Approximation of Binomial Distribution

Background

- Let X_i be Bernoulli random variable, i.e.

$$P(X_i = 1) = p, \qquad P(X_i = 0) = q = 1 - p$$

- Easy to see that $\mathbb{E}[X_i] = p, \quad V(X_i) = p(1-p)$
- Obviously, $X = \sum_{i=1}^{n} X_i$ has Binomial distribution $B(n, p)$
- When n is large, by the Central Limit Theorem, we have

$$B(n, p) \approx N(np, np(1-p)).$$

Example 85

Assume $X \sim B(25, 0.4)$. Find the probability $P(X = 8)$.

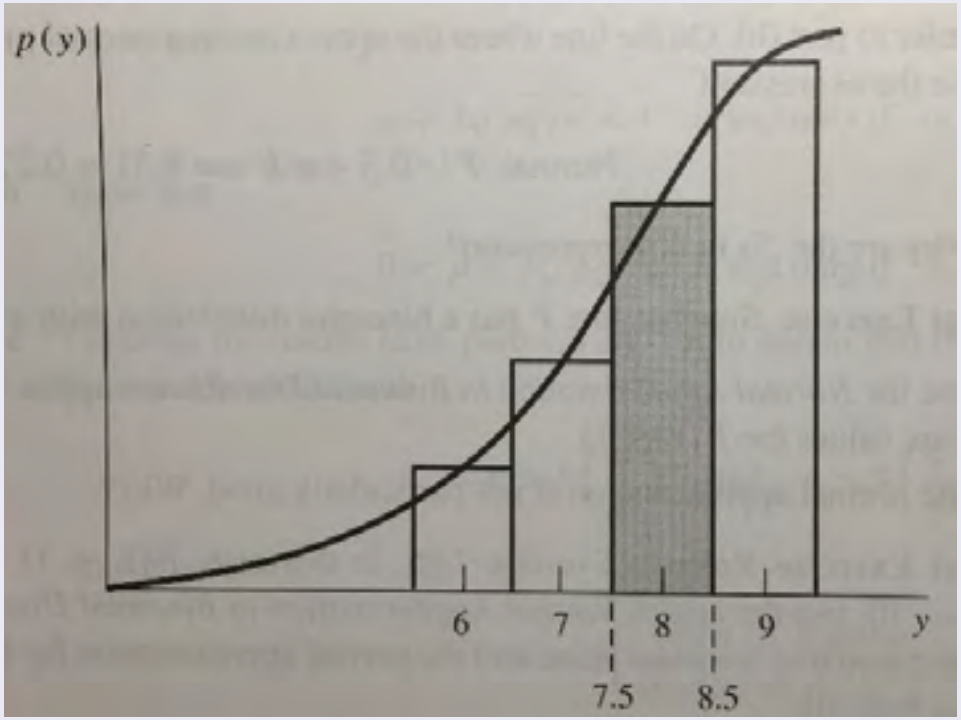
Solution:

- Use formula for $B(n, p) \rightarrow$ factorial involved!
- Let $W \sim N(np, np(1-p))$ and $Z = \dfrac{W - np}{\sqrt{np(1-p)}} \sim N(0, 1)$

Solution – Continued

- Then $P(X = 8) = P(X \leq 8) - P(X \leq 7)$, and normal approximation

$$P(X \leq 8) \quad \simeq \quad P(W \leq 8.5)$$



$$\begin{aligned} P(X = 8) &\simeq P(7.5 \leq W \leq 8.5) = P\left(\frac{7.5 - 10}{\sqrt{6}} \leq \frac{W - 10}{\sqrt{6}} \leq \frac{8.5 - 10}{\sqrt{6}}\right) \\ &= P(-1.02 \leq Z \leq -0.61) = 0.2709 - 0.1539 = 0.1170 \end{aligned}$$

Example 86

A biased coin leads to $p[H] = 0.55$. Toss it 100 times. What is the probability of getting at least 50 Heads? (use .5 rule as previous example, OK if you don't results always approximate!)

Solution: $n = 100,\ p = 0.55 \quad \Rightarrow \mu = 55$

$$P(X \geq 50) \simeq P(W \geq 49.5) \simeq P\left(Z \geq \frac{49.5 - 55}{\sqrt{100 \times 0.55 \times 0.45}}\right)$$

Remark 17

Rule of thumb: normal approximation of binomial is valid if

- $0 < p - 3\sqrt{\frac{pq}{n}}, \quad$ and $\quad p + 3\sqrt{\frac{pq}{n}} < 1$
- Or $\quad n > 9\frac{\max\{p, q\}}{\min\{p, q\}}$.

Remark 18 (Other Approximations)

- Poisson distribution Pois(λ) is approximated by $N(\lambda, \lambda)$ if λ is large.
- Gamma distribution $\Gamma(\alpha, \beta)$ is approximated by $N(\alpha\beta, \alpha\beta^2)$ if α is large.

Theorem 50 (Moment Generating Function)

If X and Y have the same moment generating function for all t, then X and Y have the same probability distribution.

Example 87

Let $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Proof:

- $N(\mu, \sigma^2)$ has moment generating function $m_X(t) = Ee^{tX} = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$
- On the other hand:

$$m_Z(t) = \mathbb{E}[e^{tZ}] = \mathbb{E}\left[e^{\frac{t}{\sigma}(X-\mu)}\right] = e^{-t\mu/\sigma}\mathbb{E}\left[e^{\frac{tX}{\sigma}}\right] = e^{-t\mu/\sigma} m_X(t/\sigma)$$

Using that $m_X(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$ we have

$$m_Z(t) = e^{-t\mu/\sigma} m_X(t/\sigma) = e^{-t\mu/\sigma} e^{\frac{t^2}{2}} e^{+t\mu/\sigma} = e^{\frac{t^2}{2}}$$

- Thus Z has the moment generating function as $N(0,1)$. By the Theorem above $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Theorem 51

Let $U = X_1 + \cdots + X_n$ be the sum of independent random variables. Assume the moment generating functions are $m_{X_1}(t), \cdots, m_{X_n}(t)$, then

$$m_U(t) = m_{X_1}(t) \cdots m_{X_n}(t)\,.$$

Proof.

□

Proof of CLT

$$CLT: \quad U_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad \to \quad Z \sim \mathcal{N}(0,1)\,.$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

- Re-write:

$$U_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i\,, \quad Z_i = \frac{X_i - \mu}{\sigma}\,.$$

- Notice $E(Z_i) = 0, Var(Z_i) = 1$. Plan: work on MGF space and show that

$$\lim_{n\to\infty} m_{U_n}(t) = m_Z(t) = e^{t^2/2}$$

- $m_{\sum_i Z_i}(t) = \prod_i m_{Z_i}(t) = m_i(t)^n$
- $m_{U_n}(t) = m_{\sum_i Z_i}(t/\sqrt{n}) = \left[m_1(\frac{t}{\sqrt{n}})\right]^n$
- Use Taylor's Theorem:

$$m_{Z_1}(t) = m_{Z_1}(0) + m'_{Z_1}(0)t + \ldots$$

Proof of CLT – Continued.

- Use Taylor's Theorem:

$$m_{Z_1}(t) = m_{Z_1}(0) + m'_{Z_1}(0)t + \ldots$$

- Better: use Taylor's Theorem with remainder:

$$m_{Z_1}(t) = m_{Z_1}(0) + m'_{Z_1}(0)t + m''_{Z_1}(\xi)t^2/2\,, \quad \text{where} \quad 0 \le \xi \le t$$

- Use that $m_{Z_1}(0) = Ee^{0Z_1} = 1$ and $m'_{Z_1}(0) = E(Z_1) = 0$
- Then

$$m_{U_n}(t) = \left[m_1(\frac{t}{\sqrt{n}})\right]^n = \left[1 + \frac{m''_{Z_1}(\xi_n)}{2}\left(\frac{t}{\sqrt{n}}\right)^2\right]^n = \left[1 + m''_{Z_1}(\xi_n)\frac{t^2/2}{n}\right]^n$$

  where $0 \le \xi_n \le t/\sqrt{n} \implies \lim_{n\to\infty} \xi_n = 0$.
- Use that $m''_{Z_1}(\xi_n) \to m''_{Z_1}(0) = Var(Z_1) = 1$ and that $\lim_n (1 + b/n)^n \to e^b$:

$$\lim_{n\to\infty} m_{U_n}(t) = e^{t^2/2} = m_Z(t) \quad \text{CLT proved!}$$

□

Poisson Distribution as a limit of Binomials

- Binomial $B(N,p)$: $P(X=x)=\binom{N}{x}p^x(1-p)^{N-x}, x=0,1,2,...,N$
- Let $\lambda = Np$. Then we can do two proofs:
- Space of MGFs:

$$m_X(t)=(pe^t+1-p)^N=\left(\frac{\lambda}{N}e^t+1-\frac{\lambda}{N}\right)^N \to e^{\lambda(e^t-1)} = \text{MGF of Poisson}$$

(use L' Hôpital's rule)

- Space of probabilities:

$$\begin{aligned}\lim_{N\to\infty} P(X=x) &= \lim_{N\to\infty}\frac{N(N-1)\cdots(N-x+1)}{x!}\left(\frac{\lambda}{N}\right)^x\left(1-\frac{\lambda}{N}\right)^{N-x}\\ &= \frac{\lambda^x}{x!}\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-x}(1-\frac{1}{N})\cdots\left(1-\frac{x-1}{N}\right)\\ &= \frac{\lambda^x}{x!}e^{-\lambda}\end{aligned}$$

Sec 7.2: Sampling Distributions

Introduction:

- Statistic: A function of the observable random variables in a sample and known constants

Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, **Sample variance:** $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

Theorem 52

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is normally distributed with mean $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$

Proof:

Use next Theorem for $a_i = 1/n\,, \mu_i = \mu\,, \sigma_i = \sigma$

Contrast to the Central Limit Theorem. CLT need ∞ many data but requires minimal assumptions

Theorem 53 (Sum of Independent Gaussian Random Variables)

Let $X_1, \cdots, X_N$ be independent normal distribution $N(\mu_i, \sigma_i^2)$, $i = 1, \cdots, N$, then

$$U = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Proof.

- X_i has MGF: $m_{X_i}(t) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right)$
- $a_i X_i$ has MGF: $m_{a_i X_i}(t) = \mathbb{E}\left[e^{t a_i X_i}\right] = m_{X_i}(a_i t) = \exp\left(\mu_i a_i t + \frac{a_i^2 \sigma_i^2 t^2}{2}\right)$

Hence,

$$\begin{aligned} m_U(t) &= m_{a_1 X_1}(t) \times \cdots \times m_{a_n X_n}(t) \\ &= \exp\left(\mu_1 a_1 t + \frac{a_1^2 \sigma_1^2 t^2}{2}\right) \times \cdots \times \exp\left(\mu_n a_n t + \frac{a_n^2 \sigma_n^2 t^2}{2}\right) \\ &= \exp\left(t \sum_{i=1}^{n} a_i \mu_i + \frac{t^2}{2} \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) \end{aligned}$$

ie, U has normal distribution with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$ □

Example 88

$X_1, X_2, \ldots, X_9$ are sampled from a normal distribution with $\mu = 10$ and $\sigma = 1.0$. What is the probability that the sample mean is within 0.3 of the true mean?

Solution:

- Let $n = 9$. From previous theorem, sample mean has distribution

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\mu, \left(\frac{1}{\sqrt{n}}\right)^2\right)$$

- Let $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}(\bar{X} - \mu) \quad \Rightarrow Z \sim N(0,1)$. [always convert to Z!]
- Then

$$P(|\bar{X} - \mu| \leq 0.3) = P\left(\left|\frac{\bar{X} - \mu}{1/\sqrt{n}}\right| \leq \frac{0.3}{1/\sqrt{n}}\right) = P(|Z| \leq 0.3\sqrt{n})$$

- From tables:

$$P\left(|Z| \leq 0.9\right) = 1 - 2P(Z > 0.9) = 1 - 2 \times 0.1841 = 0.6318$$

Example 89

How many observations are needed if we wish the sample mean to be within 0.3 of the true mean with probability 0.95? Assume the observations from distribution $N(\mu, 1)$.

Solution:

- Let $Z \sim N(0,1)$. From *last* example, we want:

$$P(|Z| \leq 0.3\sqrt{n}) = 0.95 \quad \Rightarrow \quad P(Z > 0.3\sqrt{n}) = \frac{1 - 0.95}{2} = 0.025$$

- By checking the table, we have:

$$P(Z > 1.96) = 0.025 \quad \Rightarrow 1.96 = 0.3\sqrt{n}$$

- Hence

$$n = \left(\frac{1.96}{0.3}\right)^2 = 42.68$$

- 43 observations are needed.

Theorem 54 (χ^2)

Let $X_1, \cdots, X_n$ be independent normal distribution $N(\mu_i, \sigma_i^2)$, $i = 1, \cdots, n$. And $Z_i = \dfrac{X_i - \mu_i}{\sigma_i}$. Then $\sum_{i=1}^{n} Z_i^2$ has a chi-square distribution with N degrees of freedom.

Proof.

- $Z_i \sim N(0,1) \quad \Rightarrow Z_i^2$ has χ^2 distribution with 1 degree of freedom
- MGF: $m_{Z_i^2}(t) = (1-2t)^{-1/2}$
- $V = \sum_{i=1}^{n} Z_i^2$ has MGF:

$$m_V(t) = \prod_{i=1}^{n} (1-2t)^{-1/2} = (1-2t)^{-n/2}$$

By checking MGF tables:

V has χ^2 distribution with n degrees of freedom.

□

Queueing Theory example

Example 90

Goal: Model waiting times of n customers at a queue:

- Y_1, time until 1st arrival; ...; Y_i: time between $(i-1)$-th and i-th arrival
- Assume, $Y_1, \ldots, Y_n$ are independent, have exponential distribution. Density function: (θ **is the average time between arrivals**)

$$f_{Y_i}(y_i) = \begin{cases} \frac{1}{\theta} e^{-\frac{y_i}{\theta}}, & y_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Find pdf for $U = Y_1 + \cdots + Y_n$: total waiting time of the n customers

Solution:

- Moment generating function for exponential RV: $m_{Y_i}(t) = (1-\theta t)^{-1}$
- MGF for U : $m_U(t) = \prod_{i=1}^{n} (1-\theta t)^{-1} = (1-\theta t)^{-n}$
- It is the MGF for gamma-distributed RV with $\alpha = n, \beta = \theta$. Therefore U has gamma distribution with density function

$$f_U(u) = \begin{cases} \frac{1}{\Gamma(n)\theta^n} \left(u^{n-1} e^{-u/\theta}\right), & u > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Example 91

6 observations $Z_1, \ldots, Z_6$ are from a standard normal population $N(0,1)$. Find b s.t.

$$P\left(\sum_{i=1}^{6} Z_i^2 \leq b\right) = 0.95\,.$$

Solution:

Use previous Theorem:
$\sum_{i=1}^{6} Z_i^2$ has χ^2 distribution with 6 degrees of freedom.
Next, use the tables in the book:

$$P\left(\sum_{i=1}^{6} Z_i^2 > 12.5916\right) = .05 \quad \Leftrightarrow P\left(\sum_{i=1}^{6} Z_i^2 \leq 12.5916\right) = 0.95$$

Hence, $b = 12.5916$ is the 0.95 quantile (95th percentile).

Theorem 55

If $X_i \sim N(\mu, \sigma^2)$, *then*

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

is a chi-square distribution with $n-1$ *degrees of freedom . In addition,* $\bar{X}$ *and* S^2 *are independent.*

Remark 19

Estimate of the variance from data: Observations $X_1, \ldots, X_n$ from $N(0, \sigma^2)$.

Then for $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, we have

- $\mathbb{E}[S^2] = \sigma^2$
- $Var(S^2) = \frac{2\sigma^4}{n-1}$

Proof: for n=2

When $n = 2$, $\bar{X} = \frac{1}{2}(X_1 + X_2)$, and

$$S^2 = \frac{1}{2-1}\sum_{i=1}^{2}(X_i - \bar{X})^2 = \left[X_1 - \frac{1}{2}(X_1 + X_2)\right]^2 + \left[X_2 - \frac{1}{2}(X_1 + X_2)\right]^2$$

$$= \left[\frac{1}{2}(X_1 - X_2)\right]^2 + \left[\frac{1}{2}(X_2 - X_1)\right]^2 = \frac{(X_1 - X_2)^2}{2}$$

So when $n = 2$, we have $\frac{(n-1)S^2}{\sigma^2} = \left(\frac{X_1 - X_2}{\sqrt{2\sigma^2}}\right)^2$

Let $Z = \frac{X_1 - X_2}{\sqrt{2\sigma^2}}$

- It has normal distribution, with mean $\mu = \mathbb{E}\left[\frac{X_1 - X_2}{\sqrt{2\sigma^2}}\right] = 0$
- Variance $= \left(\frac{1}{\sqrt{2\sigma^2}}\right)^2 Var(X_1) + \left(\frac{-1}{\sqrt{2\sigma^2}}\right)^2 Var(X_2) = 1$
- Hence, $Z \sim N(0, 1)$, and Z^2 has χ^2 distribution with 1 degree of freedom.

Example 92

16 observations from a standard normal population. Find b_1, b_2 such that

$$P(b_1 \leq S^2 \leq b_2) = 0.9\,.$$

Solution:

$$\begin{aligned} P(b_1 \leq S^2 \leq b_2) &= P\left(\frac{(n-1)b_1}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b_2}{\sigma^2}\right) \\ &= P\left(15b_1 \leq Z \leq 15b_2\right) \qquad \text{b/c } n = 16, \sigma = 1 \end{aligned}$$

where $Z = \dfrac{(n-1)S^2}{\sigma^2}$ has a χ^2 distribution with 15 degrees of freedom.

- Then check the table: cut off 0.05 from top and 0.05 from bottom

$$15b_1 = 7.26, \Rightarrow b_1 = 0.484; \quad 15b_2 = 25.00, \Rightarrow b_2 = 1.667\,.$$

Definition 44

Let Z be a standard normal random variable and let W be a chi-square distributed variable with v degrees of freedom. If Z and W are independent, then

$$T = \frac{Z}{\sqrt{W/v}}$$

is a student t-distribution with v degrees of freedom.

Definition 45

Let W_1 and W_2 be two independent chi-square distributed random variables with v_1 and v_2 degrees of freedom. Then

$$F = \frac{W_1/v_1}{W_2/v_2}$$

is said to have an F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom.

Introduction

- Main Topic: find the distributions of **functions of random variables**
- Various methods:
  - **Method 1:** Distribution Functions (find the CDF)
  - **Method 2:** Transformations (similar to Method 1 but with "formula")
  - **Method 3:** Use MGF (most powerful & easy to use)
- Random number generators (aka random variables on a computer)

Method 1: Distribution Function Method

Let U be a function of $X_1, \cdots, X_N$. Find the probability density function of U?

- Find region $U \leq u$, and $F_U(u) = P(U \leq u)$ by integrating joint density function
- Find density function $f_U(u)$ by differentiating distribution function of U.

Example 93

Random variable X has density function $f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$ Find the density function of $U = 3X + 1$.

Solution:

- Distribution function for U:

$$F_U(u) = P(U \leq u) = P(3X + 1 \leq u) = P\left(X \leq \frac{u-1}{3}\right)$$

$$\Rightarrow F_U(u) = \begin{cases} 0 & \text{if } u \leq 1 \\ \int_0^{\frac{u-1}{3}} 2x dx = \left(\frac{u-1}{3}\right)^2 & \text{if } 1 \leq u \leq 4 \\ 1 & \text{if } u \geq 4 \end{cases}$$

- Density function for U:

$$f_U(u) = \frac{d\, F_U(u)}{du} = \begin{cases} \frac{2(u-1)}{9}, & \text{if } 1 \leq u \leq 4 \\ 0 & \text{elsewhere} \end{cases}$$

Example 94 (Convolution – **for most tasks: much easier using MGFs!**)

X_1 and X_2 are independent random variables with density function $f_1(x)$ and $f_2(x)$. Find density function of $X_1 + X_2$

Solution:

Let $U = X_1 + X_2$. Then

$$\begin{aligned} F_U(u) &= P(U = X_1 + X_2 \leq u) = \int_{x_1+x_2 \leq u} f(x_1, x_2)\, dx_1\, dx_2 \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{u-x_2} f_1(x_1) dx_1 \right] f_2(x_2) dx_2 = \int_{-\infty}^{\infty} F_1(u - x_2) f_2(x_2) dx_2 \end{aligned}$$

So:

$$\begin{aligned} f_U(u) &= \frac{dF_U(u)}{du} = \int_{-\infty}^{\infty} \frac{dF_1(u - x_2)}{du} f_2(x_2) dx_2 = \int_{-\infty}^{\infty} f_1(u - x_2) f_2(x_2) dx_2 \\ &= \int_{-\infty}^{\infty} f_1(u - x) f_2(x) dx \end{aligned}$$

Example 95

Independent random variables $X_1, X_2 \sim U(0,1)$. Find density function of $X_1 + X_2$ (much easier using MGFs!)

Solution:

- $f_1(x_1) = f_2(x_2) = 1$, for $0 \leq x_1, x_2 \leq 1$
- Clearly, $f_U(u) = 0$ when $u < 0$ or $u > 2$.
- For $0 \leq u \leq 2$, we have:

$$f_U(u) = \int_{-\infty}^{\infty} f_1(u-x) f_2(x) dx$$

Need to figure out the integration domain

- Need: $0 \leq x \leq 1$ and $0 \leq u - x \leq 1 \Leftrightarrow u - 1 \leq x \leq u$
- When $0 \leq u \leq 1$, the domain is $0 \leq x \leq u \quad \Rightarrow f_U(u) = \int_0^u 1 \cdot dx = u$
- If $1 \leq u \leq 2$, the domain is $u - 1 \leq x \leq 1 \quad \Rightarrow f_U(u) = \int_{u-1}^1 dx = 2 - u$
- In summary:

$$f_U(u) = \begin{cases} u, & 0 \leq u \leq 1 \\ 2-u, & 1 \leq u \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

Example 96

Consider $U = Y^2$. We have

$$F_U(u) = P(U \leq u) = P(Y^2 \leq u) = P(-\sqrt{u} \leq Y \leq \sqrt{u}) = F_Y(\sqrt{u}) - F_Y(-\sqrt{u})$$

Taking derivatives.

$$f_U(u) = \frac{1}{2\sqrt{u}} \left[f_Y(\sqrt{u}) + f_Y(-\sqrt{u}) \right], \quad u > 0$$

Example: X is a random variable with density function

$$f(x) = \begin{cases} \frac{x+1}{2} & -1 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the density function for $U = X^2$.

Solution: Obviously, $0 \leq U \leq 1$. Then

$$f_U(u) = \begin{cases} \frac{1}{2\sqrt{u}} \left(\frac{\sqrt{u}+1}{2} + \frac{-\sqrt{u}+1}{2} \right) = \frac{1}{2\sqrt{u}}, & 0 \leq u \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Method 2: Transformations $[U = h(X), \quad X = h^{-1}(U)]$

- X is a random variable and h is an **increasing** function.
- Define a new random variable $U = h(X)$, then

$$P(U \leq u) = P\left(X \leq h^{-1}(u)\right) \quad \Rightarrow F_U(u) = F_X\left(h^{-1}(u)\right)$$

- Density function: $f_U(u) = f_X\left(h^{-1}(u)\right)\dfrac{\mathrm{d}h^{-1}}{\mathrm{d}u}$

Example 97

X has density function $f_X = 2x$ for $0 \leq x \leq 1$. Find the density function for $U = 3X + 1$.

Solution:

- $u = h(x) = 3x + 1 \quad \Rightarrow h^{-1}(u) = \frac{u-1}{3}$
- Hence, $f_U(u) = f_X\left(h^{-1}(u)\right)\dfrac{\mathrm{d}h^{-1}}{\mathrm{d}u} = \left(2 \cdot \dfrac{u-1}{3}\right) \cdot \dfrac{1}{3} = \dfrac{2(u-1)}{9}$, for $1 \leq u \leq 4$

Remark 20

If h is **decreasing**, then similar calculation follows.

$$f_U(u) = -f_X\left(h^{-1}(u)\right)\frac{\mathrm{d}h^{-1}}{\mathrm{d}u}$$

Example 98

$f_X = 2x$ for $0 \leq x \leq 1$. $U = -3X + 2$. What is the density function for U?

Theorem 56 (Method of Transformation)

X is a random variable with density function $f_X(x)$, and $U = h(X)$ where $h(x)$ is monotone (either decreasing or increasing) for all values of x such that $f_X(x) > 0$. Then compute the density function of U as

- *Find $x = h^{-1}(u)$, and evaluate $\dfrac{\mathrm{d}h^{-1}}{\mathrm{d}u}$*
- *Then, $f_U(u) = f_X\left(h^{-1}(u)\right)\left|\dfrac{\mathrm{d}h^{-1}}{\mathrm{d}u}\right|$*

Remark 21

Case of multiple variables MGF method is (typically) easier :

- If a function of X_1, X_2, eg, $U = X_1 + X_2$
- Treat one variable as a constant, say X_1, and find the joint distribution (or density function) of U, X_1.
- Then calculate the marginal distribution (or density) of U.

Example 99

Random variables X_1, X_2 have joint density function: $f(x_1, x_2) = e^{-(x_1+x_2)}$ for $0 \leq x_1, x_2$. Find the density function for $U = X_1 + X_2$

Solution:

- Treat $X_1 = x_1$ as constant. Transformation is: $U = x_1 + X_2$: $X_2 \longrightarrow U$
- That is $u = x_1 + x_2 \quad \Rightarrow x_2 = u - x_1 = h^{-1}(u)$
- Denote $g(x_1, u)$ as the joint density for X_1, U. The domain will be

$$\{0 \leq x_1, \quad \text{and} \quad 0 \leq x_2 = u - x_1\} \quad \Rightarrow 0 \leq x_1 \leq u$$

- Joint density function:

$$g(x_1, u) = \begin{cases} f\left(x_1, h^{-1}(u)\right) \left|\frac{\mathrm{d}h^{-1}}{\mathrm{d}u}\right| = e^{-(x_1+u-x_1)} \cdot 1 = e^{-u}, & 0 \leq x_1 \leq u \\ 0 & \text{elsewhere} \end{cases}$$

- Density function for U:

$$f_U(u) = \int_{-\infty}^{\infty} g(x_1, u) dx_1 = \begin{cases} \int_0^u e^{-u} \mathrm{d}x_1 = ue^{-u}, & 0 \leq u \\ 0 & \text{elsewhere} \end{cases}$$

Example 100

Random variables X_1, X_2 have joint density: $f(x_1, x_2) = 2(1 - x_1)$ for $0 \le x_1, x_2 \le 1$. Let $U = X_1 X_2$. Find the density function of U.

Solution:

- Treat $X_1 = x_1$ as constant. Transformation: $U = x_1 X_2$: $X_2 \longrightarrow U$
- That is $u = x_1 x_2 \quad \Rightarrow x_2 = u/x_1 = h^{-1}(u)$
- Denote $g(x_1, u)$ as the joint density for X_1, U. The domain will be

$$\left\{0 < x_1 \le 1, \quad \text{and} \quad 0 \le x_2 = \frac{u}{x_1} \le 1\right\} \qquad \Rightarrow 0 \le u \le x_1 \le 1$$

- Joint density function:

$$g(x_1, u) = \begin{cases} f\left(x_1, h^{-1}(u)\right) \left|\frac{\mathrm{d}h^{-1}}{\mathrm{d}u}\right| = 2(1 - x_1) \cdot \frac{1}{x_1}, & 0 \le u \le x_1 \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

- Density function for U:

$$f_U(u) = \int_{-\infty}^{\infty} g(x_1, u) dx_1 = \begin{cases} \int_u^1 \frac{2(1-x_1)}{x_1} \mathrm{d}x_1 = 2(u - \ln u - 1), & 0 \le u \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Theorem 57 (Moment Generating Function)

If X and Y have the same moment generating function for all t, then X and Y have the same probability distribution.

Example 101

Let $X \sim N(\mu, \sigma^2)$, then $\quad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Proof:

- $N(\mu, \sigma^2)$ has moment generating function $m(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$
- So $(X - \mu)$ has moment generating function: $m_{X-\mu}(t) = e^{\frac{1}{2}\sigma^2 t^2}$. Then

$$m_Z(t) = \mathbb{E}[e^{tZ}] = \mathbb{E}\left[e^{\frac{t}{\sigma}(X-\mu)}\right] = m_{X-\mu}\left(\frac{t}{\sigma}\right) = e^{\frac{\sigma^2}{2}\left(\frac{t}{\sigma}\right)^2} = e^{\frac{t^2}{2}}$$

Same as moment generating function for $N(0,1)$ random variables.

Theorem 58

Let $U = X_1 + \cdots + X_n$ be the sum of independent random variables. Assume the moment generating functions are $m_{X_1}(t), \cdots, m_{X_n}(t)$, then

$$m_U(t) = m_{X_1}(t) \cdots m_{X_n}(t).$$

Proof.

□

Example 102

Let $X \sim N(0,1)$. Find the distribution of X^2.

Solution:

Moment generating function for X^2:

$$
\begin{aligned}
m_{X^2}(t) &= \mathbb{E}\left[e^{tX^2}\right] = \int_{-\infty}^{\infty} e^{tx^2} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} e^{tx^2} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}(1-2t)}\mathrm{d}x \qquad \text{exists if } 1-2t>0 \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}(1-2t)^{-1/2}} \exp\left(\frac{-x^2}{2(1-2t)^{-1}}\right)\mathrm{d}x \\
&= \frac{1}{\sqrt{1-2t}} = (1-2t)^{-1/2}
\end{aligned}
$$

Hence, X^2 has gamma distribution with $\alpha = 1/2, \beta = 2$, aka χ^2 with 1 degree of freedom. Its density function ($U = X^2$):

$$
f_U(u) = \begin{cases} \frac{u^{-1/2}e^{-u/2}}{\Gamma(1/2)2^{1/2}}, & u \geq 0 \\ 0, & \text{elsewhere} \end{cases}
$$

Theorem 59 (Sum of Independent Gaussian Random Variables)

Let $X_1, \cdots, X_N$ be independent normal distribution $N(\mu_i, \sigma_i^2)$, $i = 1, \cdots, N$, then

$$U = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{N} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Proof.

- X_i has MGF: $m_{X_i}(t) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right)$
- $a_i X_i$ has MGF: $m_{a_i X_i}(t) = \mathbb{E}\left[e^{t a_i X_i}\right] = m_{X_i}(a_i t) = \exp\left(\mu_i a_i t + \frac{a_i^2 \sigma_i^2 t^2}{2}\right)$

Hence,

$$\begin{aligned} m_U(t) &= m_{a_1 X_1}(t) \times \cdots \times m_{a_n X_n}(t) \\ &= \exp\left(\mu_1 a_1 t + \frac{a_1^2 \sigma_1^2 t^2}{2}\right) \times \cdots \times \exp\left(\mu_n a_n t + \frac{a_n^2 \sigma_n^2 t^2}{2}\right) \\ &= \exp\left(t \sum_{i=1}^{n} a_i \mu_i + \frac{t^2}{2} \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) \end{aligned}$$

ie, U has normal distribution with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$ □

Capstone Example: Random Number Generators

Question: how to generate random variables on computer?
Computer uses only 0 and 1's, it is deterministic! Use Number Theory(!). For example:

- First generate a (pseudo) random integer between 0 and 2^{31} as follows:

$$x_{n+1} = (1103515245X_n + 12345) \mathrm{mod} 2^{31}$$

- Convert to uniform in $[0, 1]$ random variable: $U_n = X_n/2^{31}$.

Remark 22

- Historically, the first pseudo-random number sampling were developed for Monte Carlo simulations for Statistical Physics in the Manhattan Project. First published by **John von Neumann** in 1951 (next slide)
- Famous generators: Mersenne twister, Xorshift family, Ziggurat,...
- A "truly horrible" generator: RANDU (google it!)
- Function RAND in all programming languages.

Various Techniques Used in Connection With Random Digits

By John von Neumann

Summary written by George E. Forsythe

In manual computing methods today random numbers are probably being satisfactorily obtained from tables. When random numbers are to be used in fast machines, numbers will usually be needed faster. More significant is the fact that, because longer sequences will be used, one is likely to have more elaborate requirements about what constitutes a satisfactory sequence of random numbers. There are two classes of questions

control call for these numbers as needed. The real objection to this procedure is the practical need for checking computations. If we suspect that a calculation is wrong, almost any reasonable check involves repeating something done before. At that point the introduction of new random numbers would be intolerable. I think that the direct use of a physical supply of random digits is absolutely inacceptable for this reason and for this

Transform method for random number simulation

- Let X be a continuous random variable with distribution function F
- For CDF $F_X(x)$ of X:

$$F_X(x) = \mathbb{P}[X \leq x]$$

- However, $0 \leq F_X(x) \leq 1$ (a CDF!) and for any $0 \leq c \leq 1$, $\mathbb{P}[U \leq c] = c$. Then:

$$F_X(x) = \mathbb{P}[U \leq F(x)] = \mathbb{P}[F^{-1}(U) \leq x]$$

provided F_X has an inverse (argument still works if not!).

- Suggested (pseudo) Algorithm:
  - Generate $U \sim U(0,1)$
  - Return $X = F^{-1}(U)$

Example 103

If U_n is the pseudorandom uniform in $[0,1]$:

$$X_n = a + (b-a)U_n \quad \sim \quad U(a,b)$$

$$X_n = -\lambda \ln(U_n) \quad \sim \quad Exp(\lambda)$$

Theorem 60

Random variable X is uniform in $[0,1]$. Show that the random variable $Y = -\ln X$ is exponential with parameter 1.

Proof:

Method of distributions: Let F_X is the distribution of $X \sim \mathrm{Unif}(0,1)$, that is

$$F_X(x) = \begin{cases} 0 & , \quad x < 0 \\ x & , \quad 0 \le x < 1 \\ 1 & , \quad x \ge 1 \end{cases}$$

- Compute the distribution of Y,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(-\ln X \le y) = 1 - P(X \le e^{-y}) = 1 - F_X(e^{-y}) \\ &= 1 - e^{-y}, \quad y \ge 0 \end{aligned}$$

- For $y < 0$, $F_Y(y) = 0$. Therefore indeed $Y \sim \mathrm{Exp}(\lambda = 1)$.
- Can also compute the pdf of Y by differentiating F_Y:

$$f_Y(y) = \begin{cases} e^{-y} & , \quad y \ge 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Proof – Continued.

- Or, we can get the same result, if use the method of Transformations as the function $y = h(x) = -\ln x$ is decreasing
- Recall the formula

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right| = 1 \times e^{-y}.$$

with

$$h^{-1}(y) = e^{-y}, \quad y \in (0, -\infty)$$

Chapter: Probability Inequalities: Confidence Intervals and Monte Carlo Integration

The slides are self-contained. For further reference see:

- Book on "All of Statistics", by Wasserman, Chapter 4.
- Book on "Probability and Computing: Randomized Algorithms, and Probabilistic Analysis" by Mitzenmacher and Upfal (2017), Chapter 4

Recall: Probabilistic Guarantees

Part I: Chebyshev's Inequality

Key ideas:

- predict random variables with "confidence" by using a few moments; moments are deterministic quantities!)
- not necessary to know the entire pdf/cdf of the r.v.; this is in contrast to the normal distribution confidence intervals.

Theorem 61

Markov Inequality: *If* $X \geq 0$, *then for all* $a > 0$,

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

Chebyshev Inequality I: *For any random variable* X, *and any* $\epsilon > 0$,

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

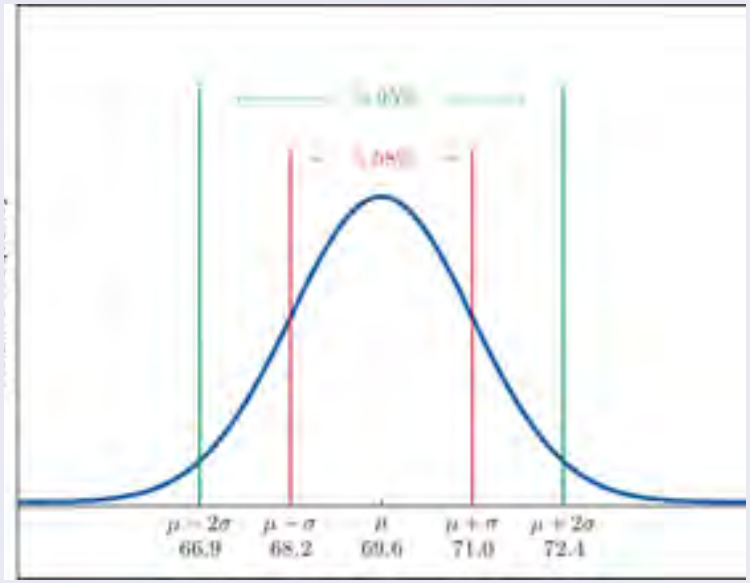
Chebyshev Inequality II: *Let* X *be a random variable with mean* $\mu = E[X]$ *and finite variance* $\sigma^2 var(X)$, *then for any* $k > 0$,

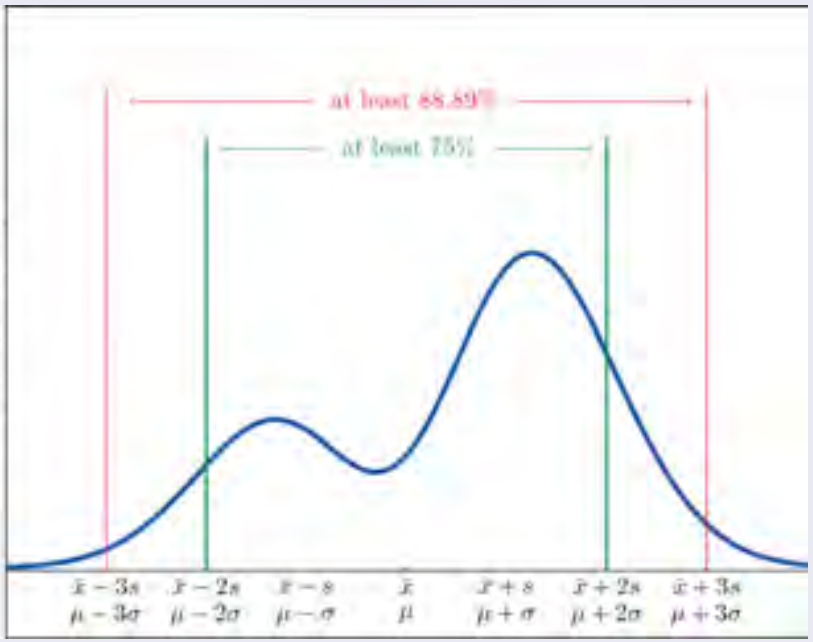
$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \textit{[follows from I for } \epsilon = k\sigma\textit{]}$$

- Left: If we know $X \sim N(\mu, \sigma)$, then

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < \frac{X - \mu}{\sigma} < 2) = P(-2 < Z < 2) \approx 95\%$$

- Right: If we do not know the distribution of X but we know mean, variance – **use Chebyshev**
- Observe the trade-offs between knowledge and confidence





Recall Chebyshev+ Law of Large

$X_1, \cdots, X_n, \cdots$ be independent random variables with mean μ and variance σ^2.

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Claim: $\bar{X}_n$ converges to μ in probability, ie,

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

Proof:

- Clearly, $\mathbb{E}[\bar{X}_n] = \mu$
- Variance:

$$Var(\bar{X}_n) = \sum_{i=1}^{n} \frac{1}{n^2} V(X_i) + \frac{2}{n^2} \sum_{1\le i<j\le n} Cov(X_i, X_j) = \frac{\sigma^2}{n} .$$

- Use Chebyshev's inequality: $P(|X - E[X]| \ge \epsilon) \le \frac{\text{var}(X)}{\epsilon^2}$

$$P(|\bar{X}_n - \mu| > \epsilon) \le \frac{\sigma^2}{\epsilon^2 n} \quad \to \quad 0 \quad \text{as } n \to \infty$$

Confidence Intervals via Chernoff bounds

Key question in Statistical Learning:

- How much data do we need to "learn" with a given confidence?
- Be "data-driven": use minimal assumptions of the pdf's of the random variables

Further References:

- Book on "All of Statistics", by Wasserman, Chapter 4.
- Book on "Probability and Computing: Randomized Algorithms, and Probabilistic Analysis" by Mitzenmacher and Upfal (2017), Chapter 4
- Book on "Understanding Machine Learning: From Theory to Algorithms" (2014) by Shai Shalev-Shwartz and Shai Ben-David

Moment Generating Functions and Tail Events

MGF of random variable X:

$$m_X(t) = \mathbb{E}[e^{tX}]$$

Taylor expansion of the MGF into moments: $m_X^{(n)}(0) = \mathbb{E}X^n$

- **Chernoff Bounds:** For any $a \in \mathbb{R}$ and $t > 0$,

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = \frac{m_X(t)}{e^{ta}}$$

- Minimizing over all $t > 0$ yields the *Chernoff bound*:

$$P(X \geq a) \leq \inf_{t>0} \left\{ \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \right\}$$

  Note that inf = min when the minimum is attained

- Same bounds for $P(X \leq a)$:

$$P(X \leq a) \leq \inf_{t<0} \left\{ \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \right\}$$

- Knowing the MGF of $X \mapsto$ tail behavior of random variable X.

MGF of random variable X:

$$M_X(t) = \mathbb{E}[e^{cX}]$$

Taylor expansion of the MGF into moments: $M_X^{(n)}(0) = \mathbb{E}X^n$

- For any $a \in \mathbb{R}$ and $c > 0$,

$$P(X \geq a) = P(e^{cX} \geq e^{ca}) \leq \frac{\mathbb{E}[e^{cX}]}{e^{ca}}$$

- Minimizing over all $c > 0$ yields the *Chernoff bound*:

$$P(X \geq a) \leq \inf_{c>0} \left\{ \frac{\mathbb{E}[e^{cX}]}{e^{ca}} \right\}$$

  or is it more convenient to use the (worse) bound:

$$\log P(X \geq a) \leq \inf_{c>0} \left\{ \log \mathbb{E}[e^{cX}] - ca \right\}$$

- Same bounds for $P(X \leq a)$:

$$\log P(X \leq a) \leq \inf_{c<0} \left\{ \log \mathbb{E}[e^{cX}] - ca \right\}$$

- Knowing the MGF of $X \mapsto$ tail behavior of random variable X.

Example 104 (Chernoff Bound)

$$P(X \geq a) \leq \inf_{t>0} \left\{ \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \right\} = \inf_{t>0} \left\{ m_X(t) e^{-ta} \right\}$$

Standard normal $Z \sim N(0,1)$, $m_Z(t) = e^{t^2/2}$. Pick in Chernoff bound $t = a$:

$$P(Z \geq a) \leq e^{-a^2/2} = .011\,, \quad \text{for } a = 3$$

Compare to Chebyshev and bounds and z-tables:

- Chernoff bound relates directly to the pdf
- Chernoff: 97% much better than Chebyshev 89%
- However z-table is "exact" answer: 99.865%

Chernoff gets even better the more "rare" the event ($a > 3$). Truly shine for sample means $\bar{X}_N = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} X_i$

- Tails of sample mean $\bar{X}_N$ with finite data:

$$P(\bar{X}_N \geq a) \leq \inf_{t>0} \left\{ \frac{\prod_{i=1}^{N} \mathbb{E}[e^{tX_i/N}]}{e^{ta}} \right\} = \inf_{t>0} \left\{ \frac{\prod_{i=1}^{N} m_{X_i}(t/N)}{e^{ta}} \right\}$$

Example – Continued

- Assume $X_i \sim N(\mu, \sigma^2)$ iid. Then Chernoff gives:

$$P(\bar{X}_N \geq a) \leq \inf_{t>0} \left\{ \frac{\prod_{i=1}^{N} m_{X_i}(t/N)}{e^{ta}} \right\} = \inf_{t>0} \{ e^{\frac{t^2}{2\sigma^2 N} + (\mu - a)t} \}$$

- The min is at $t^* = (a - \mu)\sigma^2 N$. For $a = \mu + \epsilon$ bounds becomes:

$$P(\bar{X}_N \geq \mu + \epsilon) \leq e^{-\epsilon^2 \sigma^2 N/2}$$

- Similarly:

$$P(\bar{X}_N \leq \mu - \epsilon) \leq e^{-\epsilon^2 \sigma^2 N/2}$$

- Chernoff (exponential decay in N):

$$P(|\bar{X}_N - \mu| \geq \epsilon) = P(\bar{X}_N \geq \mu + \epsilon) + P(\bar{X}_N \leq \mu - \epsilon) \leq 2e^{-\epsilon^2 \sigma^2 N/2}$$

- Compare to Chebyshev (polynomial decay in N):

$$P(|\bar{X}_N - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 N}$$

Chernoff bounds & Concentration Inequalities

- Not always possible to calculate explicitly or numerically the MGF:

$$m_{X_i}(t) = Ee^{tX_i}$$

- Usually we don't fully know the pdf of X! So we cannot calculate the MGF.
- Maybe we can bound it?
- **Example:** X_i are independent Bernoulli with parameter p:

$$m_{X_i}(t) = Ee^{tX_i} = 1 + p(e^t - 1) \leq e^{p(e^t - 1)}$$

can obtain various bounds... for example if p is unknown:

$$m_{X_i}(t) = Ee^{tX_i} = 1 + p(e^t - 1) \leq e^t$$

Concentration Inequalities

Chernoff bounds:

$$P(X \geq a) \leq \inf_{t>0} \left\{ \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \right\}$$

- Not always possible to calculate the MGF $\quad m_X(m) = Ee^{mX}$
- One way is to estimate it using **concentration inequalities:**

$$\mathbb{E}[e^{tX}] \leq \Psi(t)\,, \quad \text{all } t > 0\,.$$

- Then

$$P(X \geq a) \leq \inf_{t>0} \left\{ \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \right\} \leq \inf_{t>0} \left\{ \frac{\Psi(t)}{e^{ta}} \right\}$$

- Simplest non-trivial example–key tool in Statistical Learning: **Hoeffding inequality**.

  Key question: How much data do we need to "learn" with a given confidence?

Hoeffding's Lemma and Bound

MGF Bound

Let X be a random variable such that

- $a \leq X \leq b, \quad$ and $\quad EX = 0$ (centered)
- Then we have the Hoeffding **MGF bound**:

$$m_X(t) = \mathbb{E}[e^{tX}] \leq e^{t^2(b-a)^2/8} := \Psi(t)$$

Hoeffding Inequality for sample means

Let $X_1, ..., X_N$ IID random variables,

- $EX_i = \mu$, $a \leq X_i \leq b$.
- No other assumptions on X_i's!. Then:

$$P(|\bar{X}_N - \mu| \geq \epsilon) \leq 2e^{-2N\epsilon^2/(b-a)^2}$$

For a single random variable X we get a Chebyshev analogue:

$$P(|X - \mu| \geq \epsilon) \leq 2e^{-2\epsilon^2/(b-a)^2}$$

Proof of Hoeffding Inequality from MGF bound.

$$P(|\bar{X}_N - \mu| \geq \epsilon) \leq 2e^{-2N\epsilon^2/(b-a)^2}$$

1. Chernoff bound: tails of **sample mean** $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$:

$$P(\bar{X}_N - \mu \geq \epsilon) \leq \inf_{t>0}\left\{\frac{\prod_{i=1}^{N}\mathbb{E}[e^{t(X_i-\mu)/N}]}{e^{t\epsilon}}\right\}$$

2. Hoeffding bound: $\mathbb{E}[e^{t(X_i-\mu)/N}] \leq e^{\frac{t^2(b-a)^2}{8N^2}}$
3. $\inf_{t>0}\left\{e^{-t\epsilon}\prod_{i=1}^{N} e^{\frac{t^2(b-a)^2}{8N^2}}\right\} = \inf_{t>0}\left\{e^{-t\epsilon+\frac{t^2(b-a)^2}{8N}}\right\}$
4. Minimum at $t_{\min} = \frac{4N\epsilon}{(b-a)^2}$
5. Bound: $e^{-\frac{4N\epsilon^2}{(b-a)^2}+\frac{16N^2\epsilon^2}{(b-a)^4}\frac{(b-a)^2}{8N}} = e^{-\frac{2N\epsilon^2}{(b-a)^2}}$
6. Other side of tail same bound $\mapsto$ factor 2.

Discussion of the Results

- Control of rare/extreme events happening at the "tail":

$$P(|\bar{X}_N - \mu| \geq \epsilon) \leq 2e^{-2N\epsilon^2/(b-a)2}$$

- Minimal knowledge regarding the random variables
- No uncertainties related to asymptotics: is N adequate? what is the error? etc.
- Comparison to Chebyshev:

$$P(|\bar{X}_N - \mu| > \epsilon) \leq \frac{Var(X_1)}{\epsilon^2 N}$$

- No variance?
- Literature on Concentration Inequalities: Bennett, Bernstein, McDiarmid, Talagrand, etc.

Proof of Hoeffding Bound–optional

$$\mathbb{E}[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

1. Since $EX = 0$, if $a = 0$ then $b = 0$. Then, without loss of generality assume $a < 0$, $b > 0$.
2. Convexity, $f(x) = e^{tx}$ implies: $f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b)$
3. Let $\lambda = \frac{b-x}{b-a}$, $x = \lambda a + (1-\lambda)b$; then by convexity:

$$f(x) = e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

4. Use $EX = 0$ to obtain (take expectations):

$$Ee^{tX} \leq \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} := \Psi_1(t)$$

This is also a concentration inequality! But a morte versatile version is coming up:

Proof of Hoeffding Bound (cont'd).

$$\mathbb{E}[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

1. A less accurate concentration inequality (but easier to use):

$$Ee^{tX} \leq \Psi_1(t) \leq e^{\frac{t^2(b-a)^2}{8}} := \Psi(t)$$

2. To prove that use the following:
   - $\log \Psi_1(t) = \log\left(\frac{a}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}\right) := \phi(t(b-a))$
   - where $\phi(t) = -\theta t + \log(1-\theta+\theta e^t)$, and $\theta = -a/(b-a) > 0$.
   - Using Taylor's Thm show $\phi(t) \leq t^2/8$, i.e.

$$\phi(t(b-a)) \leq t^2(b-a)^2/8$$

□

Application: Estimating a parameter

Example 105

Evaluate the probability a gene mutation occurs in a population. Given a DNA sample a lab test can determine if it carries the mutation. However the test is expensive, so we want to get an answer with few samples.

Math formulation:

$X_i = 1$ mutation detected in test $X_i = 0$ mutation not detected

Sample mean = estimated probability of mutation, inferred from N samples:

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i := \tilde{p}$$

But the answer should also be quantifiably reliable:

Definition 46

$1-\gamma$ confidence interval for a parameter p: $[\tilde{p}-\delta, \tilde{p}+\delta]$ $(\delta > 0)$

$$P(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-\gamma$$

or equivalently $P(p \notin [\tilde{p}-\delta, \tilde{p}+\delta]) \leq \gamma$

We prefer: $\gamma, \delta > 0$ as small as possible (obv. $0 \leq \gamma \leq 1$).

Back to a "Netflix"-style problem: parameter estimation

- Consider outcomes of some complex process, e.g. predicting if one customer (out of N) will watch a proposed movie:

$$X_i = 1 \quad \text{prediction is wrong} \quad X_i = 0 \quad \text{prediction is correct}$$

- Build a simple data-based prediction model: $X_i \sim$ is Bernoulli with unknown parameter p; note that p=error probability.
- **Parameter Estimation:**

$$p \approx \tilde{p} = \bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- **Sample mean** $= \bar{X}_N =$ observed error rate in N customers
- How **confident** are we in predicting p using the estimator $\tilde{p} = \bar{X}_N$? Will we need more data to reach a level of **guarantees** we are comfortable with?
- Estimate $P(|\bar{X}_N - p| \geq \epsilon)$, equiv. **confidence interval** $P(|\bar{X}_N - p| < \epsilon)$.

Compare Probabilistic Inequalities

1. Chebyshev Inequality:

$$P(|\bar{X}_N - p| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_N)}{\epsilon^2} = \frac{p(1-p)}{N\epsilon^2} \leq \frac{1}{4N\epsilon^2}$$

Recall p is unknown, we have to use that $p(1-p) \leq 1/4$!

2. Hoeffding: (Recall $a = 0 \leq X_i \leq 1 = b$)

$$P(|\bar{X}_N - p| \geq \epsilon) \leq 2e^{-2N\epsilon^2/(b-a)^2} = 2e^{-2N\epsilon^2}$$

3. Hoeffding much tighter for the same data size N: for instance,

$$N = 10^5\,, \quad \epsilon = .01 = 1\% \text{ error}$$

Then

- Hoeffding$= 4 \cdot 10^{-9}$
- Chebyshev$= 2.5 \cdot 10^{-2}$

Exercises

1. Getting additional data has in many applications requires additional cost. How much data N do we need (at least!) in the previous example to **guarantee** error $\epsilon = 1\%$ with probability 99.99%? Compare different N's obtained by Chebyshev and Hoeffding: N_{Cheb} vs. N_{Hoeff}.
2. Revisit the gene mutation example using the Hoeffding inequality.
3. Do we get a tighter bound in the previous example if we use the Chernoff inequality instead of Hoeffding?

Application of Chebyshev on Monte Carlo Integration

Task: Numerically approximate a high-dimensional integral:

$$I(g) := \int_{\mathbb{R}^d} g(y)dy$$

- Standard integration approximation $I_N(g)$ with N **grid points** yield an approximation error

$$|I(g) - I_N(g)| \leq O(N^{-1/d}) \sim \frac{1}{N^{\frac{1}{d}}} \tag{1}$$

Curse of dimensionality: in (1),

$$O(N^{-1/d}) \sim \frac{1}{N^{\frac{1}{d}}} \sim 1\,, \quad \text{if } d >> 1$$

Namely we get a **very** slow rate of convergence in N

Main Idea behind Monte Carlo Integration

Task: Approximate a high-dimensional integral: $I(g) := \int_{\mathbb{R}^d} g(y)dy$

1. For "any" (see later) probability measure P on $\mathbb{R}^d$ with density $p = p(y), y \in \mathbb{R}^d$, rewrite $I(g)$:

$$I(g) = \int_{\mathbb{R}^d} g(y)dy = \int_{\mathbb{R}^d} \frac{g(y)}{p(y)} p(y)dy = E_P \frac{g(Y)}{p(Y)},$$

   where Y is a random variable with distribution p.

2. Probabilistic formulation $\Longrightarrow$ take N iid samples $y_1, ..., y_N$. Build the **statistical estimator**

$$J_N(g) := \frac{1}{N} \sum_{i=1}^{N} \frac{g(y_i)}{p(y_i)} \approx E_P \frac{g(Y)}{p(Y)} = I(g)$$

3. By the Law of Large Numbers (need finite variance-see next slide!):

$$J_N(g) := \frac{1}{N} \sum_{i=1}^{N} \frac{g(y_i)}{p(y_i)} \to I(g) \quad \text{in probability}$$

4. But practically... is $J_N(g)$ a good approximation of $I(g)$?

Monte Carlo Integration+Chebyshev's Inequality

Main question: Is $J_N(g)$ a good approximation of $I(g)$?

- First note that $J_N(g)$ is unbiased: $E_P J_N(g) = I(g)$
- By Chebyshev (using $E_P J_N(g) = I(g)$) we have:

$$P(|J_N(g) - I(g)| > \epsilon) \leq \frac{Var(J_N(g))}{\epsilon^2}$$

- Variance calculation (similar to LLN, here use $E_P J_N(g) = I(g)$)):

$$Var(J_N(g)) = E\left(J_N(g) - I(g)\right)^2 = \frac{1}{N^2} E\left(\sum_{i=1}^{N}\left(\frac{g(Y_i)}{p(Y_i)} - I(g)\right)\right)^2$$

- Using independence, we get

$$Var(J_N(g)) = \frac{1}{N^2}\sum_{i=1}^{N} Var\left(\frac{g(Y_i)}{p(Y_i)}\right) = \frac{1}{N} Var\left(\frac{g(Y)}{p(Y)}\right)$$

- Back to Chebyshev:

$$P(|J_N(g) - I(g)| > \epsilon) \leq \frac{Var(J_N(g))}{\epsilon^2} = \frac{1}{N\epsilon^2} \cdot Var\left(\frac{g(Y)}{p(Y)}\right) \qquad (2)$$

Monte Carlo Integration+Chebyshev $\Longrightarrow$ Importance Sampling

- Remarks on MC Integration + CoD: at first glance (2) independent of dimension d, in contrast to (1)
- However the variance term can become large, depending on dimension, choice of p. But at least we may have a way forward - in contrast to (1):
- Importance sampling idea: minimize the bound of (2) $Var\left(\frac{g(Y)}{p(Y)}\right)$ over all densities p to get the tightest "guarantee"–for the same "data cost" N:

$$p^*(y) = \text{argmin}_p Var\left(\frac{g(Y)}{p(Y)}\right) = \frac{|g(y)|}{\int |g(y)| dy}$$

  Not practical (answer includes the question!) but at least a good start: p^* depends on our target; should be a density that "looks like" g.
- Lot's of current research on IS methods...
- Add examples: Robert-Casella book (Example 3.11 pg 93/94) and rare events example from Glynn's book