Math Tripos Part IA: Probability

Michael Li

February 7, 2020

Basic concepts
Classical probability, equally likely outcomes. Combinatorial analysis, permutations and combinations. Stirling's formula (asymptotics for $\log n!$ proved). [3]

Axiomatic approach
Axioms (countable case). Probability spaces. Inclusion-exclusion formula. Continuity and subadditivity of probability measures. Independence. Binomial, Poisson and geometric distributions. Relation between Poisson and binomial distributions. Conditional probability, Bayes's formula. Examples, including Simpson's paradox. [5]

Discrete random variables
Expectation. Functions of a random variable, indicator function, variance, standard deviation. Covariance, independence of random variables. Generating functions: sums of independent random variables, random sum formula, moments.

Conditional expectation. Random walks: gambler's ruin, recurrence relations. Difference equations and their solution. Mean time to absorption. Branching processes: generating functions and extinction probability. Combinatorial applications of generating functions. [7]

Continuous random variables
Distributions and density functions. Expectations; expectation of a function of a random variable. Uniform, normal and exponential random variables. Memoryless property of exponential distribution. Joint distributions: transformation of random variables (including Jacobians), examples. Simulation: generating continuous random variables, independent normal random variables. Geometrical probability: Bertrand's paradox, Buffon's needle. Correlation coefficient, bivariate normal random variables. [6]

Inequalities and limits
Markov's inequality, Chebyshev's inequality. Weak law of large numbers. Convexity: Jensen's inequality for general random variables, AM/GM inequality.

Moment generating functions and statement (no proof) of continuity theorem. Statement of central limit theorem and sketch of proof. Examples, including sampling. [3]

Contents

1 Classical Probability

Definition. *Classical probability* applies when there are a finite number of equally likely outcomes.

1.1 Diverse Notions of Probability

Example. (i) The probability that a fair coin will land heads is $\frac{1}{2}$. [Classical Probability]

(ii) The probability that a drawing pin will land "point up" is 0.62 [Frequentist Probability]

(iii) The probability that a large earthquake will happen in the next 30 years in San Andreas Fault is about 21%. [Subjective Probability]

1.2 Sample Spaces and events

Definition. Consider an experiment that has a random outcome. Set of all possible outcomes is the *sample space.*, Ω. Define $\omega \in \Omega$ as an *observation.*

1.2.1 Sets and Probability

Definition. A subset of Ω, A, is called an *event.* We define A^c or $\bar{A}$ to be the *complement* of A. For two sets $A, B \in \Omega$, we have $A \cup B$ means "A or B", while $A \cap B$ means "A and B". If $A \cap B = \emptyset$, then we say A and B are *mutually exclusive.*

Let $\Omega = \{\omega_1, \omega_2, \cdots, \omega_N\}$ of equally likely outcomes, then for $A \in \Omega$, then

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

2 Combinatorial Analysis

2.1 Counting

Theorem (Fundamental rule of counting)**.** r multiple choices are to be made in sequence. There are m_1 choices for first choice, m_2 for second choice.$\cdots m_r$ choices for rth choice, then the total number of possible choices are $m_1 * m_2 * m_3 \cdots * m_r$.

2.2 Sampling with replacement or without

Definition. Let $N = \{1, 2, \cdots, n\}$ be a list and $X = \{1, 2, \cdots, x\}$ be items. Create function $N \to X$ with $f(i) =$item at the ith list position.

(i) (Sampling with replacement) After choosing an item it is put back and can be chosen again. "any function".

(ii) (Sampling without replacement) After choosing an item we set it aside. "Injective function" $x \geq n$.

(iii) (Sampling with replacement) We require each item has to be chosen at least once. So $n \geq x$.

Example. $N = \{a, b, c\}$, $X = \{p, q, r, s\}$. How many injective functions are there from $N \to X$? There are $4 \times 3 \times 2 = 24$ injective functions.

Example (The Birthday Problem)**.** How many people are needed in a room so that with probability $> \frac{1}{2}$ two people have the same birthday? We calculate its complement event (in r people, everyone has different birthdays):

$$P^c = 1 \cdot (1 - \frac{1}{365})(1 - \frac{2}{365})(1 - \frac{3}{365}) \cdots (1 - \frac{r-1}{365}) = \frac{365 * 364 * \cdots * (365 - r + 1)}{365^r}$$

So the probability we want is:

$$P = 1 - \frac{365 \cdot 364 \cdots (365 - r + 1)}{365^r}$$

2.3 Sampling with or without regard to ordering

Do labels given to list positions on items matter?

(i) Leave list alone.

(ii) Sort list ascending/descending so that list labelling doesn't matter.

(iii) Renumber each item by the number of the draw in which it was first seen, so (2,5,2) and (5,4,5) becomes (1,2,1). Labelling on items doesn't matter.

(iv) Both (ii) and (iii) we have $(2, 5, 2), (8, 5, 5) \to (1, 1, 2)$.

2.4 Conclusions

For a list of x items picked n times:

(i) (Sampling with replacement+ordering) We have x^n ways.

(ii) (Sampling without replacement+no ordering) We have $x^{\underline{n}} = x * (x - 1) * \cdots * (x - n + 1)$ ways. [Bar indicates "*x to the falling*"]

(iii) (Sampling without replacement+ no ordering) This is just the binomial $\binom{x}{n}$.

(iv) (sampling with replacement +no ordering) Using the ball and stick model, we list the x items and put "$n - 1$" bars to separate it into n lists as we only care how many times an item is chosen. Now we have $\binom{n+x-1}{n-1}$.

3 Stirling's Formula

3.1 Multinomial Coefficient

Definition. Suppose we fill successive positions in a list of length n, with replacement, from a set of x items. The number of ways possible so that item $i \in X$ is used n_i times, $i = 1, 2, \cdots, x$ is

$$\binom{n}{n_1, n_2, \cdots, n_x} = \binom{n}{n_1}\binom{n-n_1}{n_2} \cdots \binom{n-n_1 \cdots - n_{x-1}}{n_x} = \frac{n!}{n_1! n_2! n_3! \cdots n_x!}$$

Theorem (Stirling's Formula)**.** As $n \to \infty$

$$\log \frac{n!e^{-n}}{n^{n+\frac{1}{2}}} = \log \sqrt{2\pi} + O(\frac{1}{n})$$

Proof. The proof is non-examinable. Yay! □

Corollary.

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

Lemma. The weaker version of this is $\log n! \sim n \log n$

Proof.

$$\log n! = \sum_{k=1}^{n} \log k$$

$$\int_1^n \log x \, dx \leq \sum_{k=1}^{n} \log k \leq \int_1^{n+1} \log x \, dx$$

$$n \log n - n \leq \log n! \leq (n+1)\log(n+1) - (n+1)$$

Divide this by $n \log n$ and take the limit $n \to \infty$ then we can see the result using the sandwich theorem. □

Example. Suppose we toss a coin $2n$ times. What is the probability of equal number of heads and tails? The probability is

$$\frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{(n!)^2 2^{2n}} \sim \frac{1}{\sqrt{n\pi}}$$

4 Axioms of probability

Definition (Probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{F}, P)$. Ω is the *sample space*, $\mathcal{F}$ is a collection of subsets of Ω, and $P : \mathcal{F} \to [0, 1]$ is the *probability measure*,with $\mathcal{F}$ and $\mathcal{P}$ satisfying:

$\mathcal{F}$

(i) $\emptyset, \Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$.

(iii) $A_1, \cdots, \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

$\mathcal{P}$

(i) $0 \leq \mathcal{P}(A) \leq 1$ for all $A \in \mathcal{F}$

(ii) $\mathcal{P}(\Omega) = 1$

(iii) For a disjoint countable collection of events $A_1, A_2, \cdots$:

$$\mathcal{P}\left(\bigcup_i A_i\right) = \sum_i \mathcal{P}(A_i).$$

We say $\mathcal{P}(A)$ is the probability of the event A.

Definition (Probability distribution)**.** Let $\Omega = \{\omega_1, \omega_2, \cdots\}$. Choose $\{p_1, p_2, \cdots, \}$ such that $\sum_{i=1}^{\infty} p_i = 1$. Let $p(\omega_i) = p_i$. Then define

$$\mathcal{P}(A) = \sum_{\omega_i \in A} p(\omega_i).$$

This $\mathcal{P}(A)$ satisfies the above axioms, and $p_1, p_2, \cdots$ is the *probability distribution*

Theorem.

(i) $\mathcal{P}(\emptyset) = 0$

(ii) $\mathcal{P}(A^C) = 1 - \mathcal{P}(A)$

(iii) $A \subseteq B \Rightarrow \mathcal{P}(A) \leq \mathcal{P}(B)$

(iv) $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$.

Proof.

(i) Ω and $\emptyset$ are disjoint. So $\mathcal{P}(\Omega) + \mathcal{P}(\emptyset) = \mathcal{P}(\Omega \cup \emptyset) = \mathcal{P}(\Omega)$. So $\mathcal{P}(\emptyset) = 0$.

(ii) $\mathcal{P}(\Omega) = 1 = \mathcal{P}(A) + \mathcal{P}(A^C)$ since A and A^C are disjoint.

(iii) Write $B = A \cup (B \cap A^C)$. Then $\mathcal{P}(B) = \mathcal{P}(A) + \mathcal{P}(B \cap A^C) \geq \mathcal{P}(A)$.

(iv) $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B \cap A^C)$. We also know that $\mathcal{P}(B) = \mathcal{P}(A \cap B) + \mathcal{P}(B \cap A^C)$. Then the result follows.

□

From above, we know that $\mathcal{P}(A \cup B) \leq \mathcal{P}(A) + \mathcal{P}(B)$. So we say that $\mathcal{P}$ is a *subadditive* function. Also, $\mathcal{P}(A \cap B) + \mathcal{P}(A \cup B) \leq \mathcal{P}(A) + \mathcal{P}(B)$ (in fact both sides are equal!). We say $\mathcal{P}$ is *submodular*.

Definition (Limit of events)**.** A sequence of events $A_1, A_2, \cdots$ is *increasing*/ *decreasing* if $A_1 \subseteq A_2 \cdots / A_1 \supseteq A_2 \cdots$. Then we define the *limit* as

$$\lim_{n\to\infty} A_n = \bigcup_1^\infty A_n \quad \text{(increasing)} \quad \text{or} \quad \bigcap_1^\infty A_n \quad \text{(decreasing)}$$

Theorem. If $A_1, A_2, \cdots$ is increasing or decreasing, then

$$\lim_{n\to\infty} \mathcal{P}(A_n) = \mathcal{P}\left(\lim_{n\to\infty} A_n\right).$$

Proof. Take $B_1 = A_1$, $B_2 = A_2 \setminus A_1$. In general,

$$B_n = A_n \setminus \bigcup_1^{n-1} A_i.$$

Then

$$\bigcup_1^n B_i = \bigcup_1^n A_i, \quad \bigcup_1^\infty B_i = \bigcup_i^\infty A_n.$$

$$\begin{aligned}\mathcal{P}(\lim A_n) = \mathcal{P}\left(\bigcup_1^\infty A_i\right) = \mathcal{P}\left(\bigcup_1^\infty B_i\right) &= \lim_{n\to\infty} \sum_{i=1}^n \mathcal{P}(B_i) \text{ (Axiom III)} \\ &= \lim_{n\to\infty} \mathcal{P}\left(\bigcup_1^n A_i\right) = \lim_{n\to\infty} \mathcal{P}(A_n).\end{aligned}$$

and the decreasing case is proven similarly (or apply above to A_i^C). □

4.1 Boole's inequality

Theorem (Boole's inequality)**.** For any $A_1, A_2, \cdots$,

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

Proof. The axiom states a similar formula that only holds for disjoint sets. So we need a clever trick to make them disjoint. We define

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_i = A_i \setminus \bigcup_{k=1}^{i-1} A_k.$$

So we know that $\bigcup B_i = \bigcup A_i$. But the B_i are disjoint. So our Axiom (iii) gives

$$\mathcal{P}\left(\bigcup_i A_i\right) = \mathcal{P}\left(\bigcup_i B_i\right) = \sum_i \mathcal{P}\left(B_i\right) \leq \sum_i \mathcal{P}\left(A_i\right).$$

□

Example. We have $\mathbb{N}$ cardinality of biased coins. Let $A_k = [k\text{th toss head}]$ and $\mathcal{P}(A_k) = p_k$. Assume $\sum_1^\infty p_k < \infty$. Let's calculate the probability of infinite heads.

The event "there is at least one more head after the ith coin toss" is $\bigcup_{k=i}^\infty A_k$. There are infinitely many heads iff there are unboundedly many, so whatever i is, there is still at least more head after the ith toss. So the probability required is

$$\mathcal{P}\left(\bigcap_{i=1}^{\infty}\bigcup_{k=i}^{\infty} A_k\right) \leq \lim_{i\to\infty} \mathcal{P}\left(\bigcup_{k=i}^{\infty} A_k\right) = \lim_{i\to\infty} \sum_{k=i}^{\infty} p_k \to 0$$

4.2 Inclusion-exclusion formula

Theorem (Inclusion-exclusion formula)**.**

$$\mathcal{P}\left(\bigcup_i^n A_i\right) = \sum_{S\subset 1,2,,\cdots,m} (-1)^{|S|-1} P(\bigcap_{j\in S} B_j)$$

Proof. Perform induction on n. $n = 2$ is proven above. Then

$$\mathcal{P}(A_1 \cup A_2 \cup \cdots A_n) = \mathcal{P}(A_1) + \mathcal{P}(A_2 \cup \cdots \cup A_n) - \mathcal{P}\left(\bigcup_{i=2}^{n}(A_1 \cap A_i)\right).$$

Then we can apply the induction hypothesis for $n-1$, and expand the mess. □

Example. Let $1, 2, \cdots, n$ be randomly permuted to $\pi(1), \pi(2), \cdots, \pi(n)$. If $i \neq \pi(i)$ for all i, we say we have a *derangement*. Let $A_i = [i = \pi(i)]$. Then

$$\begin{aligned}
\mathcal{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_k \mathcal{P}(A_k) - \sum_{k_1<k_2} \mathcal{P}(A_{k_1} \cap A_{k_2}) + \cdots \\
&= n \cdot \frac{1}{n} - \binom{n}{2}\frac{1}{n}\frac{1}{n-1} + \binom{n}{3}\frac{1}{n}\frac{1}{n-1}\frac{1}{n-2} + \cdots \\
&= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!} \to e^{-1}
\end{aligned}$$

So the probability of derangement is $1 - \mathcal{P}(\bigcup A_k) \approx 1 - e^{-1} \approx 0.632$.

5 Independence

5.1 Bonferroni's Inequalities

Theorem. For any events $A_1, A_2 \cdots, A_n$, $a \leq r \leq n$:

$$P(\bigcup_1^n A_i) \geq \sum_1^n P(A_i) - \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) + \cdots (-1)^{r-1} \sum_{i_1<i_2\cdots<i_r} P(A_{i_1}A_{i_2}\cdots A_{i_r})$$

if r is even. If r is odd, the inequality sign is leq.

Proof. Proof by induction on n. □

5.2 Independence

Definition. Two events A and B are *independent* if: $P(A \cap B) = P(A)P(B)$. Otherwise we say it is *dependent*.

It is easy to show thatif A is independent of B, then A is independent with B^c and A^c is independent with B^c.

Example. Roll two fair dice. Let A_1 be the event that the first die is odd. Let $A_2 =$ [sum is odd]. The event probabilities are as follows:

Event	Probability
A_1	$1/2$
A_2	$1/2$
$A_1 \cap A_2$	$1/4$

Then we say A_1 and A_2 are independent.

5.2.1 Independent Experiments

Let $\Omega_1 = \{\alpha_1, \cdots,\}$ with distribution $\{p_1, p_2 \cdots\}$ and $\Omega_2 = \{\beta_1, \beta_2, \cdots\}$ with probability $\{q_1, q_2, \cdots\}$. and $P((\alpha_i, \beta_j)) = p_i q_j$. Then we have:

$$P(A \cap B) = \sum_{\alpha_i \in A, \beta_j \in B} p_i q_j = \sum_{\alpha_i \in A} p_i \sum_{\beta_j \in B} q_j = P(A)P(B)$$

5.2.2 Independence of Multiple events

Events $A_1, A_2, \cdots A_n$ are said to be *independent* or *mutually independent* if $P(A_{i_1} \cap A_{i_2} \cdots A_{i_r}) = P(A_{i_1}) \cdots P(A_{i_r})$ for all $i_1 \cdots i_r \in \{1, 2, \cdots, n\}$.

5.3 Important Discrete Distributions

Definition. Toss coin and $\Omega = \{H, T\}$ and $p \in [0, 1]$, we have the *Bernoulli distribution* $B(1, p)$.

Definition. Toss coin n times with probability of heads $p \in [0, 1]$ and we have the *Binomial distribution*, denoted $B(n, p)$.

Definition. Toss a coin with $p \in [0, 1]$. The probability that first head is after k tails is $= (1 - p)^k p$ and forms a probability space. It is the *Geometric distribution* and is *memoryless*, as tails we got in the past doesn't give us information about future heads.

Definition. From an urn with containing n_1 red balls and n_2 black balls, we pick n balls with n_k red balls. The probability of this (shown below) forms a probability space and is called the*hypergeometric distribution.*

$$\mathcal{P}(n_k \text{ red}) = \frac{\binom{n_1}{n_k}\binom{n_2}{n-n_k}}{\binom{n_1+n_2}{n}}.$$

Definition. The Poisson distribution, denoted $P(\lambda)$, we have:

$$p_k = \frac{\lambda^k}{k!}e^{-\lambda}$$

Theorem (Poisson approximation to binomial)**.** Suppose $n \to \infty$ and $p \to 0$ such that $np \to \lambda$. Then the binomial distribution $B(n,p)$ tends to Poisson $P(\lambda)$.

Proof.

$$p_k = \binom{n}{k}p^k(1-p)^{n-k} = \frac{1}{k!}\frac{n\cdots(n-k+1)}{n^k}(np)^k(1-\frac{pn}{n})^{n-k} \to \frac{1}{k!}\lambda^k e^{-\lambda}$$

□

6 Conditional Probability

6.1 Basic Principles

Definition. Suppose B is a event where $P(B) > 0$. For any event $A \subseteq \Omega$ the conditional probability of A and B is:

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

If two events are independent then $P(A|B) = P(A)$.

Example. A_i=[Player i gets royal flush]

$P(A_1) = 1.539 * 10^{-6}$ while $P(A_2|A_1) = 1.959 * 10^{-6}$, so "good hands attract". Also, if event A attracts event B, then B attracts A also (same for repel, easily checked).

6.2 Properties of Conditional Probability

Theorem.

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

(ii) $P(A \cap B \cap C) = P(A|(B \cap C)P(B|C)P(C)$

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

(iv) The function $P(x|B)$ restricted to subsets of B is a probability function (or measure).

Proof. 1,2,3 are trivial. For 4: $A \subseteq B$, $P(A|B) = \frac{P(A \cap B)}{P(B)} \leq 1$. One could see that if A_i are disjoint, we have $P(\cup_i A_i|B) = \sum_i P(A_i|B)$. □

6.3 Law of Total Probability

Definition. Suppose $\{B_i\}_{i=1}^{\infty}$ is a disjoint set of events that partition the sample space Ω ($\cup_{i=1}^{\infty}\{B_i\} = \Omega$). For any event A, we have the law of total probability:

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

6.4 Bayes' Formula

Theorem (Bayes' Formula)**.** Suppose $\{B\}_{i=1}^{\infty}$ is a partition of the sample space and A is an event such that $P(A) > 0$. Then for any B_j such that $P(B_j) > 0$:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

Example. I have two children, and (i) one is a boy, or (ii) one is a boy born on Tuesday. What is the probability that both of them are boys?

(i) $\mathcal{P}(BB|BB \cup BG) = \frac{1/4}{1/4+2/4} = \frac{1}{3}$.

(ii) Let B^* denote a boy born on a Tuesday, and B a boy not born on a Tuesday:

$$\mathcal{P}(B^*B^* \cup B^*B|BB^* \cup B^*B^* \cup B^*G) = \frac{\frac{1}{14}\cdot\frac{1}{14} + \frac{2\cdot1}{14}\cdot\frac{6}{14}}{\frac{1}{14}\cdot\frac{1}{14} + \frac{2\cdot1}{14}\cdot\frac{6}{14} + 2\cdot\frac{1}{14}\cdot\frac{1}{2}} = \frac{13}{27}.$$

7 Discrete random variables

7.1 Discrete random variables

Definition. A *random variable* X taking values in a set Ω_X is a function $X : \Omega \to \Omega_X$. Ω_X is usually numbers, eg. $\mathbb{R}$ or $\mathbb{N}$.

Definition. A random variable is *discrete* if Ω_X is finite or countably infinite.

Definition. A *discrete uniform distribution* is a distribution with finite number of outcomes, and each outcome is equally likely, such as the value after throwing a die.

7.2 Expectation and variance

Definition (Expectation)**.** The *expectation* (or *mean*) of a real-valued X is equal to

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} p_\omega X(\omega).$$

provided this is absolutely convergent. Otherwise, we say the expectation doesn't exist (or it could be infinite). Alternatively,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\Omega_X}\sum_{\omega:X(\omega)=x} p_\omega X(\omega)\\ &= \sum_{x\in\Omega_X} x\sum_{\omega:X(\omega)=x} p_\omega\\ &= \sum_{x\in\Omega_X} xP(X=x).\end{aligned}$$

Example (Petersburg paradox)**.** Play a game where we keep tossing a coin until you get a tail. If you get a tail on the ith round, I pay you $\$2^i$. The expected value is

$$\mathbb{E}[X] = \frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 4 + \frac{1}{8}\cdot 8 + \cdots = \infty.$$

Example. (i) Poisson $P(\lambda)$. Let $X \sim P(\lambda)$. Then

$$\mathbb{E}[X] = \sum_{r=0}^{\infty} rP(X=r) = \sum_{r=0}^{\infty} \frac{r\lambda^r e^{-\lambda}}{r!} = \sum_{r=1}^{\infty} \lambda \frac{\lambda^{r-1}e^{-\lambda}}{(r-1)!} = \lambda \sum_{r=0}^{\infty} \frac{\lambda^r e^{-\lambda}}{r!} = \lambda$$

Theorem.

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$ and $\mathbb{E}[X] = 0$ iff $\mathcal{P}(X=0) = 1$.

(ii) If a and b are constants, then $\mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. This is true even if Y and X are not independent.

(iii) $\mathbb{E}[X]$ is a constant that minimizes $\mathbb{E}[(X-c)^2]$ over c.

Proof.

(i) $X \geq 0$ means that $X(\omega) \geq 0$ for all ω. Then

$$\mathbb{E}[X] = \sum_{\omega} p_\omega X(\omega) \geq 0.$$

If exists ω so $X(\omega) > 0$ and $p_\omega > 0$, then $\mathbb{E}[X] > 0$. So $X(\omega) = 0$ for all ω.

(ii)

$$\mathbb{E}[aX+bY] = \sum_{\omega} (aX(\omega)+bY(\omega))p_\omega = a\sum_{\omega} p_\omega X(\omega) + b\sum_{\omega} p_\omega Y(\omega) = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

(iii)

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mathbb{E}[X]+\mathbb{E}[X]-c)^2] \\ &= \mathbb{E}[(X-\mathbb{E}[X])^2 + 2(\mathbb{E}[X]-c)(X-\mathbb{E}[X]) + (\mathbb{E}[X]-c)^2] \\ &= \mathbb{E}(X-\mathbb{E}[X])^2 + 0 + (\mathbb{E}[X]-c)^2.\end{aligned}$$

This is clearly minimized when $c = \mathbb{E}[X]$.

□

Theorem. For any random variables $X_1, X_2, \cdots X_n$, for which the following expectations exist,

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

Proof.

$$\sum_{\omega} p(\omega)[X_1(\omega) + \cdots + X_n(\omega)] = \sum_{\omega} p(\omega)X_1(\omega) + \cdots + \sum_{\omega} p(\omega)X_n(\omega).$$

□

Definition. The *variance*, which is the square of *standard deviation*, of a random variable X is defined as

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

Theorem.

(i) $\operatorname{Var} X \geq 0$. If $\operatorname{Var} X = 0$, then $\mathcal{P}(X = \mathbb{E}[X]) = 1$.

(ii) $\operatorname{Var}(a+bX) = b^2 \operatorname{Var}(X)$. This can be proved by expanding the definition and using the linearity of the expected value.

(iii) $\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, also proven by expanding the definition.

Example.

$$\begin{aligned}\mathbb{E}[X] = \sum_0^\infty rpq^r = pq\sum_0^\infty rq^{r-1} = pq\sum_0^\infty \frac{\mathrm{d}}{\mathrm{d}q}q^r = pq\frac{\mathrm{d}}{\mathrm{d}q}\sum_0^\infty q^r \\ = pq\frac{\mathrm{d}}{\mathrm{d}q}\frac{1}{1-q} = \frac{pq}{(1-q)^2} = \frac{q}{p}.\end{aligned}$$

Then

$$\mathbb{E}[X(X-1)] = \sum_0^\infty r(r-1)pq^r = pq^2\sum_0^\infty r(r-1)q^{r-2} = pq^2\frac{\mathrm{d}^2}{\mathrm{d}q^2}\frac{1}{1-q} = \frac{2pq^2}{(1-q)^3}$$

So the variance is

$$\operatorname{Var}(X) = \frac{2pq^2}{(1-q)^3} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q}{p^2}.$$

7.3 Indicator random variables

Definition (Indicator function)**.** The *indicator function* or *indicator variable* $I[A]$ (or I_A) of an event $A \subseteq \Omega$ is

$$I[A](\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

We have the following properties (trivial to prove):

Proposition.

(i) $\mathbb{E}[I[A]] = \sum_\omega p(\omega)I[A](\omega) = \mathcal{P}(A)$.
(ii) $I[A^C] = 1 - I[A]$.
(iii) $I[A \cap B] = I[A]I[B]$.
(iv) $I[A \cup B] = I[A] + I[B] - I[A]I[B]$.
(v) $I[A]^2 = I[A]$.
(vi) $\mathbb{E}[I[A]I[B]] = P(A \cap B)$

Example. Let $2n$ people (n husbands and n wives, with $n > 2$) sit alternate man-woman around the table randomly. Let $N =$ number of couples sitting next to each other.

Let $A_i = [\mathrm{i}th couples it stogether]$. So $N = \sum_{i=1}^n I[A_i]$.. Then

$$\mathbb{E}[N] = E\left[\sum I[A_i]\right] = \sum_1^n \mathbb{E}\big[I[A_i]\big] = n\mathbb{E}\big[I[A_1]\big] = n\mathcal{P}(A_i) = n \cdot \frac{2}{n} = 2.$$

We can use these to prove the inclusion-exclusion formula:

Theorem (Inclusion-exclusion formula)**.**

$$\mathcal{P}\left(\bigcup_i^n A_i\right) = \sum_{S\subset 1,2,,\cdots,m} (-1)^{|S|-1} P(\bigcap_{j\in S} B_j)$$

Proof. Let I_j be the indicator function for A_j. Write

$$S_r = \sum_{i_1<i_2<\cdots<i_r} I_{i_1} I_{i_2} \cdots I_{i_r}, \quad \& \quad s_r = \mathbb{E}[S_r] = \sum_{i_1<\cdots<i_r} \mathcal{P}(A_{i_1} \cap \cdots \cap A_{i_r}).$$

Then

$$1 - \prod_{j=1}^n (1 - I_j) = S_1 - S_2 + S_3 \cdots + (-1)^{n-1} S_n.$$

And

$$\mathcal{P}\left(\bigcup_1^n A_j\right) = \mathbb{E}\left[1 - \prod_1^n (1 - I_j)\right] = s_1 - s_2 + s_3 - \cdots + (-1)^{n-1} s_n.$$

□

7.4 Independent random variables

Definition (Independent random variables)**.** Let $X_1, X_2, \cdots, X_n$ be discrete random variables. They are *independent* iff for any $x_1, x_2, \cdots, x_n$,

$$\mathcal{P}(X_1 = x_1, \cdots, X_n = x_n) = \mathcal{P}(X_1 = x_1) \cdots \mathcal{P}(X_n = x_n).$$

Theorem. If $X_1, \cdots, X_n$ are independent random variables, and $f_1, \cdots, f_n$ are functions $\mathbb{R} \to \mathbb{R}$, then $f_1(X_1), \cdots, f_n(X_n)$ are independent random variables.

Proof. Note that there can be many different x_i for which $f_i(x_i) = y_i$. When finding $\mathcal{P}(f_i(x_i) = y_i)$, we need to sum over all x_i such that $f_i(x_i) = f_i$. Then

$$\begin{aligned}
&\mathcal{P}(f_1(X_1) = y_1, \cdots f_n(X_n) = y_n) \\
&= \sum_{\substack{x_1:f_1(x_1)=y_1 \\ \vdots \\ x_n:f_n(x_n)=y_n}} \mathcal{P}(X_1 = x_1, \cdots, X_n = x_n) = \sum_{\substack{x_1:f_1(x_1)=y_1 \\ \vdots \\ x_n:f_n(x_n)=y_n}} \prod_{i=1}^n \mathcal{P}(X_i = x_i) \\
&= \prod_{i=1}^n \sum_{x_i:f_i(x_i)=y_i} \mathcal{P}(X_i = x_i) = \prod_{i=1}^n \mathcal{P}(f_i(x_i) = y_i).
\end{aligned}$$

□

Theorem. If $X_1, \cdots, X_n$ are independent random variables and all the following expectations exists, then

$$\mathbb{E}\left[\prod X_i\right] = \prod \mathbb{E}[X_i].$$

Proof. Write R_i for the range of X_i (or Ω_{X_i})

$$\begin{aligned}\mathbb{E}\left[\prod_1^n X_i\right] &= \sum_{x_1\in R_1}\cdots\sum_{x_n\in R_n} x_1x_2\cdots x_n \times \mathcal{P}(X_1=x_1,\cdots,X_n=x_n)\\ &= \prod_{i=1}^n \sum_{x_i\in R_i} x_i\mathcal{P}(X_i=x_i) = \prod_{i=1}^n \mathbb{E}[X_i].\end{aligned}$$

□

Theorem. If $X_1, X_2, \cdots X_n$ are independent random variables, then

$$\operatorname{Var}\left(\sum X_i\right) = \sum \operatorname{Var}(X_i).$$

Proof.

$$\begin{aligned}\operatorname{Var}\left(\sum X_i\right) &= \mathbb{E}\left[\left(\sum X_i\right)^2\right] - \left(\mathbb{E}\left[\sum X_i\right]\right)^2\\ &= \mathbb{E}\left[\sum X_i^2 + \sum_{i\neq j} X_iX_j\right] - \left(\sum \mathbb{E}[X_i]\right)^2\\ &= \sum \mathbb{E}[X_i^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j] - \sum(\mathbb{E}[X_i])^2 - \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\\ &= \sum \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2.\end{aligned}$$

□

Example. Let X_i be iid $B(1,p)$, ie. $\mathcal{P}(1) = p$ and $\mathcal{P}(0) = 1-p$. Then $Y = X_1 + X_2 + \cdots + X_n \sim B(n,p)$.

Since $\operatorname{Var}(X_i) = \mathbb{E}[X_i]^2 - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p)$, we have $\operatorname{Var}(Y) = np(1-p)$.

7.5 Inequalities

Here we prove a lot of different inequalities which may be useful for certain calculations. In particular, Chebyshev's inequality will allow us to prove the weak law of large numbers.

Definition (Convex function)**.** A function $f : (a,b) \to \mathbb{R}$ is *convex* if for all $x_1, x_2 \in (a,b)$ and $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$,

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geq f(\lambda_1 x_1 + \lambda_2 x_2).$$

It is *strictly convex* if the inequality above is strict (except when $x_1 = x_2$ or λ_1 or $\lambda_2 = 0$. A function is *concave* if $-f$ is convex.

Proposition. If f is differentiable and $f''(x) \geq 0$ for all $x \in (a,b)$, then it is convex. It is strictly convex if $f''(x) > 0$.

Theorem (Jensen's inequality)**.** If $f : (a, b) \to \mathbb{R}$ is convex, then

$$\sum_{i=1}^{n} p_i f(x_i) \geq f\left(\sum_{i=1}^{n} p_i x_i\right)$$

for all $p_1, p_2, \cdots, p_n$ such that $p_i \geq 0$ and $\sum p_i = 1$, and $x_i \in (a, b)$.

This says that $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ (where $\mathcal{P}(X = x_i) = p_i$).

If f is strictly convex, then equalities hold only if all x_i are equal, ie. X takes only one possible value.

Proof. Induct on n. It is true for $n = 2$ by the definition of convexity. Then

$$\begin{aligned} f(p_1x_1 + \cdots + p_nx_n) &= f\left(p_1x_1 + (p_2 + \cdots + p_n)\frac{p_2x_2 + \cdots + l_nx_n}{p_2 + \cdots + p_n}\right) \\ &\leq p_1 f(x_1) + (p_2 + \cdots p_n) f\left(\frac{p_2x_2 + \cdots + p_nx_n}{p_2 + \cdots + p_n}\right). \\ &\leq p_1 f_(x_1) + (p_2 + \cdots + p_n)\left[\frac{p_2}{(\,)} f(x_2) + \cdots + \frac{p_n}{(\,)} f(x_n)\right] \\ &= p_1 f(x_1) + \cdots + p_n(x_n). \end{aligned}$$

where the $(\,)$ is $p_2 + \cdots + p_n$.

Strictly convex case is proved with sign $<$ by definition of strict convexity. □

Corollary (AM-GM inequality)**.** Given $x_1, \cdots, x_n$ positive reals, then

$$\left(\prod x_i\right)^{1/n} \leq \frac{1}{n}\sum x_i.$$

Proof. Take $f(x) = -\log x$. This is concave since its second derivative is $x^{-2} > 0$.

Take $\mathcal{P}(x = x_i) = 1/n$. Then

$$\mathbb{E}[f(x)] = \frac{1}{n}\sum -\log x_i = -\log \text{GM} \;\;\&\;\; f(\mathbb{E}[X]) = -\log\frac{1}{n}\sum x_i = -\log \text{AM}$$

Since $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$, AM $\geq$ GM. Since $-\log x$ is strictly convex, AM $=$ GM only if all x_i are equal. □

Theorem (Cauchy-Schwarz inequality)**.** For any two random variables X, Y,

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Proof. If $Y = 0$, then both sides are 0. Otherwise, $\mathbb{E}[Y^2] > 0$. Let

$$w = X - Y \cdot \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}.$$

Then

$$\begin{aligned} \mathbb{E}[w^2] = \mathbb{E}\left[X^2 - 2XY\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]} + Y^2\frac{(\mathbb{E}[XY])^2}{(\mathbb{E}[Y^2])^2}\right] &= \mathbb{E}[X^2] - 2\frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} + \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} \\ &= \mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} \end{aligned}$$

Since $\mathbb{E}[w^2] \geq 0$, the Cauchy-Schwarz inequality follows. □

Theorem (Markov inequality)**.** If X is a random variable with $\mathbb{E}|X| < \infty$ and $\varepsilon > 0$, then

$$\mathcal{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}|X|}{\varepsilon}.$$

Proof. We have, using the property of expected value of indicators:

$$I[\{|X| \geq \varepsilon\}] \leq \frac{|X|}{\varepsilon}. \Rightarrow \mathcal{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}|X|}{\varepsilon}.$$

□

Theorem (Chebyshev inequality)**.** If X is a random variable with $\mathbb{E}[X^2] < \infty$ and $\varepsilon > 0$, then

$$\mathcal{P}((|X - \mathbb{E}[x])| \geq \epsilon) \leq \frac{\operatorname{Var} X}{\varepsilon^2}.$$

Proof. We have, similarly as above:

$$I[\{|X - \mathbb{E}[X]| \geq \varepsilon\}] \leq \frac{(x - \mathbb{E}[X])^2}{\varepsilon^2}.$$

□

7.5.1 Weak law of large numbers

Theorem. Let $X_1, X_2, \cdots$ be iid random variables, with mean μ and $\operatorname{Var} \sigma^2$.Let $S_n = \sum_{i=1}^{n} X_i$. Then for all $\varepsilon > 0$,

$$\mathcal{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$. We say, $\frac{S_n}{n}$ tends to μ (in probability), or $\frac{S_n}{n} \to_p \mu$.

Proof. By Chebyshev,

$$\begin{aligned}\mathcal{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2}{\varepsilon^2} = \frac{1}{n^2}\frac{\mathbb{E}(S_n - n\mu)^2}{\varepsilon^2} &= \frac{1}{n^2\varepsilon^2}\operatorname{Var}(S_n) \\ &= \frac{n}{n^2\varepsilon^2}\operatorname{Var}(X_1) = \frac{\sigma^2}{n\varepsilon^2} \to 0\end{aligned}$$

□

Theorem (Strong Law of Large Numbers)**.** We say $\frac{S_n}{n}$ tends almost surely to μ:

$$\mathcal{P}\left(\frac{S_n}{n} \to \mu \text{ as } n \to \infty\right) = 1.$$

7.6 Covariance and correlation

Definition (Covariance)**.** Given two random variables X, Y, the *covariance* is

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Proposition.

(i) $\operatorname{cov}(X, c) = 0$ for constant c.
(ii) $\operatorname{cov}(X + c, Y) = \operatorname{cov}(X, Y)$.
(iii) $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.
(iv) $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
(v) $\operatorname{cov}(X, X) = \operatorname{Var}(X)$.
(vi) $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{cov}(X, Y)$.
(vii) X, Y independent $\Rightarrow \operatorname{cov}(X, Y) = 0$.

$\operatorname{cov}(X, Y) = 0$ does *NOT* imply that they are independent.

Definition (Correlation coefficient)**.** The *correlation coefficient* of X and Y is

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \leq 1.$$

For ≤ 1. Apply Cauchy-Schwarz to $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$. □

7.7 Conditional distribution and expectation

Definition. Let X and Y be random variables with joint distribution $\mathcal{P}(X = x, Y = y)$. Then the *marginal distribution* (or simply *distribution*) of X is

$$\mathcal{P}(X = x) = \sum_{y \in \Omega_u} \mathcal{P}(X = x, Y = y).$$

The *conditional distribution* of X given Y is

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

The *conditional expectation* is the mean of that distribution, summed over all $x \in \Omega_X$.

Example. Consider a dice roll. Let $Y = 1$ denote an even roll and $Y = 0$ denote an odd roll. Let X be the value of the roll. Then $\mathbb{E}[X|Y] = 3 + Y$, ie 4 if even, 3 if odd.

Theorem. If X and Y are independent, then

$$\mathbb{E}[X|Y] = \mathbb{E}[X]$$

Proof.

$$\mathbb{E}[X|Y = y] = \sum_x x\mathcal{P}(X = x|Y = y) = \sum_x x\mathcal{P}(X = x) = \mathbb{E}[X]$$

□

Theorem (Tower property of conditional expectation)**.**

$$\mathbb{E}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}_X[X],$$

Proof.

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_X[X|Y]] &= \sum_y \mathcal{P}(Y = y)\mathbb{E}[X|Y = y] \\
&= \sum_y \mathcal{P}(Y = y) \sum_x x\mathcal{P}(X = x|Y = y) = \sum_x \sum_y x\mathcal{P}(X = x, Y = y) \\
&= \sum_x x \sum_y \mathcal{P}(X = x, Y = y) = \sum_x x\mathcal{P}(X = x) = \mathbb{E}[X].
\end{aligned}$$

□

We sum $\mathbb{E}[X]$ over the correct subscripts. This is also called the law of total expectation. Suppose $A_1, A_2, \cdots, A_n$ is a partition of Ω. Then

$$\mathbb{E}[X] = \sum_{i:\mathcal{P}(A_i)>0} \mathbb{E}[X|A_i]\mathcal{P}(A_i).$$

8 Probability generating functions

Definition. Consider a random variable X, taking value $0, 1, 2, \cdots$. The *probability generating function* of X or of the distribution (Pr, $r = 0, 1, \cdots$, $Pr = P(X = r)$) is

$$p(z) = E[z^X] = \sum_{r=0}^{\infty} P(X = r)z^r = \sum_{r=0}^{\infty} p_r z^r$$

Thus $p(z)$ is a polynomial or a power series that converges for $|z| \leq 1$. We usually write $p_X(z)$ when we are using its p.g.f.

Theorem. The distribution of X is uniquely determined by the p.g.f $p(z)$.

Proof. We find p_0 from $p_0 = p(0)$. We know that we can differentiate $p(z)$ term by term for $|z| \leq 1$. Thus:

$$p'(0) = p_1$$

Repeated differentiation gives the rest of the values. □

Theorem.

$$E[X] = \lim_{x\to 1} p'(z)$$

Proof. We first prove $\geq$. we have

$$p'(z) = \sum_{r=1}^{\infty} rp_r z^{r-1} \leq \sum_{r=1}^{\infty} rp_r = E[X]$$

Now we prove $\leq$. Choose $\epsilon \geq 0$. Let N be large enough that $\sum_{r=1}^{N} rp_r \geq E[X] - \epsilon$. Then:

$$E[X] - \epsilon \leq \sum_{r=1}^{N} rp_r = \lim_{z\to 1} \sum_{r=1}^{N} rp_r z^{r-1} \leq \lim_{z\to 1} \sum_{r=1}^{\infty} rp_r z^{r-1} = \lim_{z\to 1} p'(z)$$

Since this is true for all $\epsilon \geq 0$, we have $E[X] \leq \lim_{z\to 1} p'(z)$. □

Similarly (with a similar proof), we have:

Theorem.

$$E[X(X-1)] = \lim_{z\to 1} p''(z)$$

Theorem. Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with p.g.fs $p_1(z), p_2(z), \cdots, p_n(z)$. Then the p.g.f of $X_1 + X_2 + \cdots X_n$ is: $p_1(z)p_2(z)\cdots p_n(z)$.

Proof.

$$E[z^{X_1+X_2+\cdots+X_n}] = E[z^{X_1}]E[z^{X_2}]\cdots E[z^{X_n}] = p_1(z)p_2(z)\cdots p_n(z)$$

□

Example. Consider the Poisson distribution. Then

$$p_r = \mathcal{P}(X = r) = \frac{1}{r!}\lambda^r e^{-\lambda}.$$

Then

$$p(z) = \mathbb{E}[z^X] = \sum_0^\infty z^r \frac{1}{r!}\lambda^r e^{-\lambda} = e^{\lambda z}e^{-\lambda} = e^{\lambda(z-1)}.$$

We can have a sanity check: $p(1) = 1$, which makes sense, since $p(1)$ is the sum of probabilities.

We have

$$\mathbb{E}[X] = \left.\frac{\mathrm{d}}{\mathrm{d}z}e^{\lambda(z-1)}\right|_{z=1} = \lambda,$$

and

$$\mathbb{E}[X(X-1)] = \left.\frac{\mathrm{d}^2}{\mathrm{d}x^2}e^{\lambda(z-1)}\right|_{z=1} = \lambda^2$$

So

$$\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Example. Let $X_1, X_2, \cdots, X_n$ be iid with pgf $p(z) = \mathbb{E}z^X$. Let N be a random variable independent of X_i with pgf $h(z)$. What is the pgf of $S_n = X_1 + \cdots + X_N$?

$$\begin{aligned}
\mathbb{E}[z^{S_n}] = \mathbb{E}[z^{X_1+\cdots+X_N}] = \mathbb{E}_N[\underbrace{\mathbb{E}_{X_i}[z^{X_1+\ldots+X_N}|N]}_{\text{assuming fixed } N}] &= \sum_{n=0}^\infty \mathcal{P}(N=n)\mathbb{E}[z^{X_1+X_2+\cdots+X_n}] \\
&= \sum_{n=0}^\infty \mathcal{P}(N=n)\mathbb{E}[z^{X_1}]\mathbb{E}[z^{X_2}]\cdots\mathbb{E}[z^{X_n}] \\
&= \sum_{n=0}^\infty \mathcal{P}(N=n)(\mathbb{E}[z^{X_1}])^n \\
&= \sum_{n=0}^\infty \mathcal{P}(N=n)p(z)^n = h(p(z))
\end{aligned}$$

Then

$$\mathbb{E}[S_N] = \left.\frac{\mathrm{d}}{\mathrm{d}z}h(p(z))\right|_{z=1} = h'(p(1))p'(1) = \mathbb{E}[N]\mathbb{E}[X_1]$$

To calculate the variance, use the fact that

$$\mathbb{E}[S_n(S_n-1)] = \left.\frac{\mathrm{d}^2}{\mathrm{d}z^2}h(p(z))\right|_{z=1}.$$

Then we can find that

$$\operatorname{Var}(S_N) = \mathbb{E}[N]\operatorname{Var}(X_1) + \mathbb{E}[X_1^2]\operatorname{Var}(N).$$

Note. The mean and variance can also be found quickly using towered expectation directly.

9 Biology Stuff

9.1 Branching processes

Branching processes are used to model reproduction. Consider $X_0, X_1, \cdots$, where X_n is the number of individuals in the nth generation. We assume $X_0 = 1$ and:

(i) Each individual lives for 1 and produces k offspring with probability p_k.

(ii) Suppose all offspring behave independently. Then

$$X_{n+1} = Y_1^n + Y_2^n + \cdots + Y_{X_n}^n,$$

where Y_i^n are iid random variables, which is the same as X_1.

Let $F(z)$ be the pgf of Y_i^n. Then let

$$F(z) = E[z^{Y_i^n}] = E[z^{X_1}] = \sum_{k=0}^{\infty} p_k z^k. \quad \& \quad F_n(z) = E[z^{X_n}].$$

Theorem.

$$F_{n+1}(z) = F_n(F(z)) = F(F(F(\cdots F(z)\cdots))) = F(F_n(z)).$$

Proof.

$$\begin{aligned} F_{n+1}(z) = \mathbb{E}[z^{X_{n+1}}] = \mathbb{E}[\mathbb{E}[z^{X_{n+1}}|X_n]] &= \sum_{k=0}^{\infty} \mathcal{P}(X_n = k)\mathbb{E}[z^{Y_1^n+\cdots+Y_k^n}|X_n = k] \\ &= \sum_{k=0}^{\infty} \mathcal{P}(X_n = k)\mathbb{E}[z^{Y_1}]\cdots\mathbb{E}[z^{Y_k}] = \sum_{k=0}^{\infty} \mathcal{P}(X_n = k)(\mathbb{E}[z^{X_1}])^k \\ &= \sum_{k=0}^{\infty} \mathcal{P}(X_n = k)F(z)^k = F_n(F(z)) \end{aligned}$$

□

Theorem. Suppose $\mathbb{E}[X_1] = \sum kp_k = \mu$ and $\operatorname{Var}(X_1) = \mathbb{E}[(X-\mu)^2] = \sum(k-\mu)^2 p_k < \infty$. Then $\mathbb{E}[X_n] = \mu^n$, and $\operatorname{Var} X_n = n\sigma^2$ if $\mu = 1$. If not:

$$\operatorname{Var} X_n = \frac{\sigma^2\mu^{n-1}(\mu^n - 1)}{\mu - 1},$$

Proof.

$$\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}[X_n|X_{n-1}]] = \mathbb{E}[\mu X_{n-1}] = \mu\mathbb{E}[X_{n-1}]$$

Then by induction, $\mathbb{E}[X_n] = \mu^n$ (since $X_0 = 1$). To calculate the variance, we have:

$$\begin{aligned} \mathbb{E}[X_n^2] = \mathbb{E}[\mathbb{E}[X_n^2|X_{n-1}]] = \mathbb{E}[\operatorname{Var}(X_n) + (\mathbb{E}[X_n])^2|X_{n-1}] &= \mathbb{E}[X_{n-1}\operatorname{Var}(X_1) + \mathbb{E}[(\mu X_{n-1})^2]] \\ &= \mathbb{E}[X_{n-1}\sigma^2 + \mathbb{E}[(\mu X_{n-1})^2]] \\ &= \sigma^2\mu^{n-1} + \mu^2\mathbb{E}[X_{n-1}^2] \end{aligned}$$

So

$$\begin{aligned} &\operatorname{Var} X_n \\ &= \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2 && = \mu^4\operatorname{Var}(X_{n-2}) + \sigma^2(\mu^{n-1} + \mu^n) \\ &= \mu^2(\mathbb{E}[X_{n-1}^2] - \mathbb{E}[X_{n-1}]^2) + \sigma^2\mu^{n-1} && = \mu^{2(n-1)}\operatorname{Var}(X_1) + \sigma^2(\mu^{n-1} + \cdots + \mu^{2n-3}) \\ &= \mu^2\operatorname{Var}(X_{n-1}) + \sigma^2\mu^{n-1} && = \sigma^2\mu^{n-1}(1 + \mu + \cdots + \mu^{n-1}). \end{aligned}$$

Of course, we can obtain this using the pgf. □

Extinction probability

Let A_n be the event $X_n = 0$, ie extinction has occurred by the nth generation. Let q be the probability that extinction eventually occurs. Then $A = \bigcup_{n=1}^{\infty} A_n =$ [extinction eventually occurs]. Since $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, we know that

$$q = \mathcal{P}(A) = \lim_{n\to\infty} \mathcal{P}(A_n) = \lim_{n\to\infty} \mathcal{P}(X_n = 0). \quad \& \quad \mathcal{P}(X_n = 0) = F_n(0),$$

since $F_n(0) = \sum \mathcal{P}(X_n = k) z^k$. We know that

$$F(q) = F\left(\lim_{n\to\infty} F_n(0)\right) = \lim_{n\to\infty} F(F_n(0)) = \lim_{n\to\infty} F_{n+1}(0) = q.$$

Theorem. The probability of extinction q is the smallest root to the equation $q = F(q)$. Suppose $\mu = \mathbb{E}[X_1]$, then if $\mu \leq 1$, then $q = 1$; if $\mu > 1$, then $q < 1$.

Proof. Let α be the smallest root. Note that $0 \leq \alpha \Rightarrow F(0) \leq F(\alpha) = \alpha$ since F is increasing. Hence $F(F(0)) \leq \alpha$. Continuing inductively, $F_n(0) \leq \alpha$ for all n. So

$$q = \lim_{n\to\infty} F_n(0) \leq \alpha \Rightarrow q = \alpha.$$

To show that $q = 1$ when $\mu \leq 1$, we know that $F'(z), F''(z) \geq 0$ for $z \in (0, 1)$. So F is convex increasing. Since $F'(1) = \mu \leq 1$, it must approach $(1, 1)$ from above the $F = z$ line. So $z = 1$ is the only root. □

9.2 Random walk and gambler's ruin

Definition. Let $X_1, \cdots, X_n$ be iid random variables such that $X_n = +1$ with chance p, and -1 otherwise. Let $S_n = X_0 + \cdots + X_n$. Then $(S_0, S_1, \cdots, S_n)$ is a *1-dimensional random walk*. If $p = q = \frac{1}{2}$, we say it is a *symmetric random walk*.

Example. A gambler starts with $\$z$, with $z < a$, and plays a game in which he wins $\$1$ or loses $\$1$ at each turn with probabilities p and q respectively. What are the proabbility p_z, q_z that he hits $a, 0$ before $0, a$ respectively if he starts on z?

He either wins his first game, with probability p, or loses with q. Then

$$p_z = qp_{z-1} + pp_{z+1} \text{where } 0 < z < a, p_0 = 0, p_a = 1$$

Try $p_z = t^z$ gives $pt^2 - t + q = (pt - q)(t - 1) = 0$,.If $p \neq q$, then

$$p_z = A1^z + B\left(\frac{q}{p}\right)^z.$$

Since $p_z = 0$, $A = -B$. Since $p_a = 1$, we obtain (and similarly for q_z)

$$p_z = \frac{1 - (q/p)^z}{1 - (q/p)^a}. \qquad q_z = \frac{(q/p)^z - (q/p)^a}{1 - (q/p)^a}$$

if $p \neq q$.If $p = q$

$$p_z = \frac{z}{a} \qquad q_z = \frac{a - z}{z}$$

Since $p_z + q_z = 1$, we know that the game will eventually end.

Duration of the game

Let $D_z =$ expected time until the random walk hits 0 or a, starting from z. We know there are ways that this can happen, so the value is finite, thus bounded. Then:

$$\begin{aligned} D_z = \mathbb{E}[\mathbb{E}[\text{duration}|X_1]] &= p\mathbb{E}[\text{duration}|X_1 = 1] + q\mathbb{E}[\text{duration}|X_1 = -1] \\ &= p(1 + D_{z+1}) + q(1 - D_{z-1}) \end{aligned}$$

So

$$D_z = 1 + pD_{z+1} + qD_{z-1},$$

subject to $D_0 = D_a = 0$. We know how to solve this (assuming $p \neq q$):

$$D_z = \frac{z}{q-p} - \frac{a}{q-p} \cdot \frac{1-(q/p)^z}{1-(q/p)^a}.$$

If $p = q$, then we find a new particular solution and get

$$D_z = z(a - z).$$

Using generating functions

Let $U_{z,n} = \mathcal{P}(\text{random walk absorbed at 0 at } n|\text{start in } z)$.We have the following conditions: $U_{0,0} = 1$; $U_{z,0=0}$ for $0 < z \leq a$; $U_{0,n} = U_{a,n} = 0$ for $n > 0$. Then:

$$U_z(s) = \sum_{n=0}^{\infty} U_{z,n} s^n.$$

We know that (Multiply by s^{n+1} and sum on $n = 0, 1, \cdots$)

$$U_{z,n+1} = pU_{z+1,n} + qU_{z-1,n}. \Rightarrow U_z(s) = psU_{z+1}(s) + qsU_{z-1}(s).$$

We try $U_z(s) = [\lambda(s)]^z$. Then

$$\lambda(s) = ps\lambda(s)^2 + s. \Rightarrow \lambda_1(s), \lambda_2(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2ps}.$$

So

$$U_z(s) = A(s)\lambda_1(s)^z + B(s)\lambda_z(s)^2.$$

Since $U_0(s) = 1$ and $U_a(s) = 0$, we know that

$$A(s) + B(s) = 1 \quad \& \quad A(s)\lambda_1(s)^a + B(s)\lambda_2(s)^a = 0.$$

Then we find that (remember $\lambda_1(s)\lambda_2(s) = \frac{q}{p}$):

$$U_z(s) = \frac{\lambda_1(s)^a\lambda_2(s)^z - \lambda_2(s)^a\lambda_1(s)^z}{\lambda_1(s)^a - \lambda_2(s)^a} = \left(\frac{q}{p}\right)^z \cdot \frac{\lambda_1(s)^{a-z} - \lambda_2(s)^{a-z}}{\lambda_1(s)^a - \lambda_2(s)^a}.$$

We see that $U_z(1) = q_z$. We can apply the same method to find the generating function for absorption at a, say $V_z(s)$. Then the generating function for the duration is $U + V$. Hence the expected duration is $D_z = U_z'(1) + V_z'(1)$.

10 Continuous random variables

10.1 Continuous random variables

Definition. A *continuous random variable* X is a real function $X : \Omega \to \mathbb{R}$ such that

$$\mathcal{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x, \quad f \text{ satisfies } \ f \geq 0 \ \ \& \ \ \int_{-\infty}^{\infty} f(x) = 1.$$

where f is the *probability density function.* Informally, (very important)

$$\mathcal{P}(X \in [x, x+\delta x]) = \int_x^{x+\delta x} f(z)\,\mathrm{d}z \approx f(x)\delta x.$$

Definition. The *cumulative distribution function* (or simply *distribution function*) of any random variable X is an increasing function ($F(x) \to 1$ as $x \to \infty$):

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

In the case of continuous random variables, we have

$$f(x) = \int_{\infty}^{x} f(z)\,\mathrm{d}z.$$

Then F is continuous and differentiable. $F'(x) = f(x)$ whenever F is differentiable.

Note that we always have

$$\mathcal{P}(a < x \leq b) = F(b) - F(a).$$

This will be equal to $\int_a^b f(x)\,\mathrm{d}x$ in the case of continuous random variables.

10.2 Certain important distributions

Definition. The *uniform distribution* on $[a, b]$ satisfies:

$$f(x) = \frac{1}{b-a}. \quad \& \quad F(x) = \int_a^x f(z)\,\mathrm{d}z = \frac{x-a}{b-a}$$

for $a \leq x \leq b$.If X is a uniform distribution on $[a, b]$, we write $X \sim U[a, b]$.

Definition. The *exponential random variable with parameter* λ satisfies:

$$f(x) = \lambda e^{-\lambda x} \quad \& \quad F(x) = 1 - e^{-\lambda x}$$

for $x \geq 0$. We write $X \sim \mathcal{E}(\lambda)$.

Proposition. The exponential random variable, similar to the geometric distribution, is *memoryless*, ie.

$$\mathcal{P}(X \geq x + z | X \geq x) = \mathcal{P}(X \geq z).$$

Proof.

$$\mathcal{P}(X \geq x+z | X \geq x) = \frac{\mathcal{P}(X \geq x+z)}{\mathcal{P}(X \geq x)} = \frac{\int_{x+z}^{\infty} f(u)\,\mathrm{d}u}{\int_x^{\infty} f(u)\,\mathrm{d}u} = \frac{e^{-\lambda(x+z)}}{e^{-\lambda x}}$$
$$= e^{-\lambda z} = \mathcal{P}(X \geq z).$$

□

10.3 Distribution of a function of a random variable

Definition. For a continuous random variable X, with pdf $f(x)$ and $h(x)$ is a continuous strictly increasing function with $h^{-1}(x)$ differentiable, then $y = h(x)$ is a continuous random variable with pdf:

$$f_Y(y) = f(h^{-1}(y))\frac{d}{dy}h^{-1}(y)$$

Proof.

$$P(Y \leq y) = F_Y(y) = P(h(y) \leq y) = P(X \leq h^{-1}(y)) = F(h^{-1}(y))$$

$$F'_y(y) = f_Y(y) = f(h^{-1}(y))\frac{d}{dx}h^{-1}(y)$$

□

Example. Let $X \sim U[0,1]$. Let $Y = -\log X$. Then

$$\mathcal{P}(Y \leq y) = \mathcal{P}(-\log X \leq y) = \mathcal{P}(X \geq e^{-y}) = (1 - e^{-y}).$$

Theorem. $U \sim U[0,1]$. For any strictly increasing distribution function F the random variable $X = F^{-1}(U)$ has cdf $F(x)$. (Proof trivial)

10.4 Expectation

Definition. The *expectation* of a continuous random variable is:

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

provided not both of $\int_0^\infty xf(x)\,\mathrm{d}x$ and $\int_{-\infty}^0 xf(x)\,\mathrm{d}x$ are infinite.

Theorem. $E[X] = \int_0^\infty P(X \geq x)\,dx - \int_0^\infty P(X \leq -x)\,dx$.

Proof.

$$\int_0^\infty P(X \geq x)\,\mathrm{d}x = \int_0^\infty \int_x^\infty f(y)\;\mathrm{d}y\;\mathrm{d}x = \int_0^\infty \int_0^\infty I[y \geq x]f(y)\,\mathrm{d}y\,\mathrm{d}x$$

$$= \int_0^\infty \int_0^y \mathrm{d}x\, f(y)\;\mathrm{d}y = \int_0^\infty yf(y)\;\mathrm{d}y$$

similarly ,we prove the other side and result follows. □

Definition. Exactly same as the discrete case ($\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (E[X])^2$.), we have:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - \left(\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x\right)^2.$$

10.5 Stochastic ordering of random variables

Definition. We sat $X \geq_{sr} Y$ (*stochastic ordering*) if $P(X > t) \geq P(Y > t) \forall t$.

Remark. We can deduce from this that $E(X) \geq E(Y)$ using the lemma above. So stochastic ordering is stronger than expectation ordering,

11 Jointly distributed random variables

Two random variables X, Y have *joint distribution functions*:

$$F(x,y) = P(X \leq x, Y \leq y) \text{ where } F : \mathbb{R}^2 \to [0,1]$$

The *marginal distribution* of X is $F_X(x) = P(X \leq x) = P(X \leq x, Y < \infty) = F(x,\infty) = \lim_{y\to\infty} F(x,y)$. We say $X_1, X_2, \cdots, X_n$ are *jointly distributed continuous random variables* and have *joint pdf* if for any set $A \subseteq \mathbb{R}^n$:

$$P((x_1, x_2, \cdots, x_n) \in A) = \int\int\cdots\int f(x_1, \cdots, x_n)dx_1 \cdots dx_n$$

where $f(x_1, \cdots, x_n) = 0$ and $\int_{\mathbb{R}^n} f(x_1, \cdots, x_n)dx_1 \cdots dx_n = 1$

Theorem. If X, Y are jointly continuous random variables then they are individually continuous random variables.

Proof.

$$P(X \in A) = \int_{x\in A}\int_{-\infty}^{\infty} f(x,y)dydx = \int_{x\in A} f_X(x)dx$$

where $f_X(x)$ is the (*marginal*) pdf of X. □

11.1 Independence of Continuous Random Variables

The Variables $X_1, \cdots, X_n$ are independent if:

$$P(X_1 \in A_1, X_2 \in A_2 \cdots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2)\cdots P(X_n \in A_n)$$

Remark.

(i) This is equivalent to saying that the joint distribution function is the product of the marginal distribution functions, or that the joint pdf is the product of the marginal densities.

(ii) The theorems that hold for discrete independent random variables hold for continuous random variables.

11.2 Geometric Probability

11.2.1 Bertrand's paradox

Suppose we draw a random chord in a circle. What is the probability that the length of the chord is greater than the length of the side of an inscribed equilateral triangle?

(i) We randomly pick two end points over the circumference independently. If we draw the inscribed triangle with the vertex at one end point, the length of the chord is longer than a side of the triangle if the other end point lies between the two other end points of the triangle. This happens with probability $1/3$.

(ii) wlog the chord is horizontal, on the lower side of the circle. The mid-point is uniformly distributed along the vertical radius. Since the side of the triangle bisects the radius, the probability of getting a long line is $1/2$.

(iii) The mid point of the chord is distributed uniformly across the circle. Then you get a long line iff the mid-point lies in the smaller circle inscribed in the triangle with half radius.

11.2.2 Buffon's Needle

Example. A needle of length l is tossed randomly onto a floor with parallel lines L apart, $l \leq L$. What is $P(A)$, where A is the event that the needle intersects a line?

Proof. If we assume that the distance to the next line, X, and angle with the parallel line, θ is uniformly distributed, we have:

$$f(x, \theta) = \frac{1}{L\pi}$$

Then, if $X \leq L \sin\theta$, it would work. So: (we can estimate π)

$$P = \int_0^\pi \frac{L\sin\theta}{L}\frac{\mathrm{d}\theta}{\pi} = \frac{2l}{\pi L}$$

□

12 Normal Distribution

Definition. The pdf of normal distribution $N(\mu, \sigma^2)$ is defined as:

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

Where $-\infty < x < \infty$, μ the mean, and σ^2 the variance. We denote the cdf of *standard normal distribution*, a normal distribution $N(0, 1)$, as $\Phi(x)$ and its pdf as $\varphi(x)$.

We can check that it is indeed a probability distribution (through polar variable exchange), with mean μ and variance σ^2 easily checked by basic integration.

12.1 Mean, Median, and Mode

Definition. Given a p.df, we say $\hat{x}$ is a mode if $f(\hat{x}) \geq f(x)$ for all x, and is a *median* if:

$$\int_{-\infty}^{\hat{x}} f(x)\mathrm{d}x = \frac{1}{2} \quad \& \quad P(X \leq \hat{x}) = \frac{1}{2} = P(X \geq \hat{x})$$

The *sample mean* is defined as the mean of a random sample taken from a distribution.

12.2 Order Statistics

let $Y_1 \cdots, Y_n$ be the values of $X_1 \cdots X_n$ in increasing order. These are called the order Statistics, with common notation $X_{(i)} = Y_i$.

The *sample median* is $Y_{[\frac{n+1}{2}]}$ if n is odd or any value in $[Y_{\frac{n}{2}}, Y_{\frac{n}{2}+1}]$ if n is even. Then, if we have the maximum of the values is Y_n and $X_1 to X_n$ has cdf F and pdf f, then:

$$P(Y_n < y) = F(y)^n$$

And its pdf is:

$$g(y) = nF(y)^{n-1}f(y)$$

Similarly for the smallest Y_1, we have cdf $1-(1-F(y))^n$ and pdf $n(1-F(y))^{n-1}f(y)$. The joint cdc of Y_1, Y_n can also be calculated to be $G(y_1.y_n) = F(y_n)^n - (F(y_n) - F(y_1))^n$. Then its joint pdf is:

$$g(y_1, g_n) = n(n-1)(F(y_n) - F(y_1))^{n-2}f(y_1)f(y_n)$$

13 Transformation of Random Variables

13.1 Transformation of RV

Suppose $X_1, \cdots, X_n$ random variables with pdf f and get:

$$y_1 = r_1(X_1, \cdots, X_n)$$

$$y_2 = r_2(X_1, \cdots, X_n)$$

$$\vdots$$

$$y_n = r_n(X_1, \cdots, X_n)$$

Let $R \subseteq R^n$ such that $P((X_1 \cdots X_n) \in R) = 1$. Suppose S is the image of R and map $R \to S$ is a bijection:

$$X_1 = r_1(y_1, \cdots, y_n)$$

$$X_2 = r_2(y_1, \cdots, y_n)$$

$$\vdots$$

$$X_n = r_n(y_1, \cdots, y_n)$$

Assume $\frac{\partial s_i}{\partial y_j}$ exists and is eontinuous at every point in S. Define the *Jacobian determinant*:

$$J = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \frac{\partial s_1}{\partial y_2} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \frac{\partial s_n}{\partial y_2} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

Take $A \subseteq R$ and $B = r(A)$: $P((X_1, \cdots, X_n) \in A) = P((Y_1, \cdots, Y_n) \in B)$ with $Y_1 \to Y_n$ having density:

$$g(y_1 \cdots, y_n) = f(s_1(y_1, \cdots, y_n), \cdots, s_n(y_1, \cdots, y_n))|J|$$

Example. Suppose (X, Y) has density

$$f(x, y) = \begin{cases} 4xy & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We see that X and Y are independent, with each having a density $f(x) = 2x$.

Define $U = X/Y, V = XY$. We will later show that they are not independent.

Then we have $X = \sqrt{UV}$ and $Y = \sqrt{V/U}$.

The Jacobian is

$$\det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} = \det \begin{pmatrix} \frac{1}{2}\sqrt{v/u} & \frac{1}{2}\sqrt{u/v} \\ -\frac{1}{2}\sqrt{v/u^3} & \frac{1}{2}\sqrt{1/uv} \end{pmatrix} = \frac{1}{2u}$$

Note that this can be more easily found by considering

$$\det \begin{pmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial u/\partial y \end{pmatrix} = 2u$$

and then inverting the matrix. So

$$g(u,v) = 2\sqrt{uv}\sqrt{\frac{v}{u}}\frac{1}{2u} = \frac{2v}{u},$$

if (u,v) is in the image S, 0 otherwise. So

$$g(u,v) = \frac{2v}{y} I[(u,v) \in S].$$

Since this is not separable, we know that U and V are not independent.

In the linear case: $(y_1 \cdots, y_n) = A(x_1, \cdots, x_n) = AX$
$X = A^{-1}Y$, so $|J| = \frac{1}{\det A}$. Then we have:

$$g(y_1 \cdots y_n) = \frac{1}{|\det A|} f(A^{-1}y)$$

13.2 Convolution

Example. Suppose X_1, X_2 have joint pdf $f(x_1, x_2)$ and we want to calculate the pdf of $X_1 + x_2$. Let $Y = X_1 + X_2$, and $Z = X_2$. Then $X_1 = Y - Z$ and $X_2 = Z$.

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

So $|J| = 1$. The joint distribution of Y and Z is $g(y,z) = f(y-z, z)$. The marginal density of Y is:

$$\begin{aligned} g(y) &= \int_{-\infty}^{\infty} f(y-z, z)dz \\ &= \int_{-\infty}^{\infty} f(z, y-z)dz \end{aligned}$$

If X_1 and X_2 are independent with pdf f_1 and f_2, then :

$$g(y) = \int_{-\infty}^{\infty} f_1(x) f_2(y-x) \mathrm{d}x$$

This is called the convolution of f_1 and f_2.

13.3 Cauchy Distribution

Definition. The *Cauchy Distrbution* has pdf:

$$f(x) = \frac{1}{\pi(1+x^2)}$$

We can check that it is a density function, with area under the curve is 1. However, $E(X)$ is undefined. The mode and median are both 0.

If X and Y are two independent Cauchy distributions, then the pdf of $Z = X + Y$ is:

$$f_Z(z) = \frac{\frac{1}{2}}{\pi(1+\frac{z}{2})^2}$$

So $\frac{1}{2}Z$ has the Cauchy distribution. So the mean of n Cauchy random variables will have the Cauchy distribution.

14 Moment Generating Function

14.1 What happens if the mapping is not $1-1$?

Suppose X has pdf f. What is the pdf of $Y = |X|$.

$$P(|X| \in (a,b)) = \int_a^b (f(x) + f(-x))\mathrm{d}x \Rightarrow f_Y(x) = f(x) + f(-x)$$

Example. Suppose $X_1, X_2, \cdots, X_n$ iid random variables. What is the joint distribution of order statistics $Y_1, \cdots, Y_n$, say G? It is (there are $n!$ ways of ordering them)

$$g(y_1, \cdots, y_n) = n! f(y_1) \cdots f(y_n) \qquad y_1 \leq y_2 \cdots \leq y_n$$

14.2 Minimum of exponentials is exponential

Example. Suppose X_1, X_2 are iid with $\epsilon(\lambda)$ and $\epsilon(\mu)$ respectively.

$$Y = \min(X_1, X_2)$$

$$P(Y \geq t) = P(X \geq t, X_2 \geq t) = e^{-\lambda t} e^{-\mu t} = e^{-(\lambda+\mu)t}$$

So $Y \sim \epsilon(\lambda + \mu)$.

Example. $X_1, \cdots, X_n$ iid $\sim \epsilon(\lambda)$. let $Y_1, \cdots, Y_n$ be order statistics. Then define $Z_i = Y_i - Y_{i-1}$. So we have:

$$Z = AY$$

where $A = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & -1 & 1 \end{pmatrix}$ Define the pdf of $Z_1, \cdots, Z_n$ to be h. Then:

$$\begin{aligned} h(z_1, \cdots, z_n) = g(y_1, \cdots, y_n) \cdot 1 = n! f(y_1) \cdots f(y_n) &= n! \lambda^n e^{-\lambda(y_1 + \cdots + y_n)} \\ &= n! \lambda^n e^{-\lambda(nz_1 + \cdots z_n)} \\ &= \prod_{i=1}^{n} (\lambda i) e^{(-\lambda i) z_{n+1-i}} \end{aligned}$$

Since h is expressed as a product of n density functions, we have

$$Z_i \sim \mathcal{E}((n+1-i)\lambda).$$

with all Z_i independent.

14.3 Moment generating functions

Definition. The *moment generating function* of a random variable X is defined by:

$$m(\theta) = E[e^{\theta x}]$$

for those θ such that $m(\theta)$ is finite. It is computed as:

$$m(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x$$

Theorem. The moment generating function determines the distribution of X, provided $m(\theta)$ is finite for all θ in some interval containing the origin.

Definition. $E[X^r]$ is called the "rth *moment* of X".

Theorem. The rth moment of X is the coefficient of $\frac{\theta^r}{r!}$ in the power series expansion of $m(\theta)$, equivalently, the rth derivative evaluated at $\theta = 0$, i.e. $m^{(r)}(0)$.

14.4 Gamma distribution

Let $S_n = X_1 + \cdots + X_n$. X_i iid $\mathcal{E}(\lambda)$. The moment generating function of S_n is:

$$E[e^{\theta(X_1+\cdots+X_n)}] = (\frac{\lambda}{\lambda-\theta})^n$$

Definition. The gamma distribution, defined $\Gamma(n,\lambda)$, with parameters $n \in Z^+$ and $\lambda > 0$, satisfies:

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} \quad \& \quad mgf = E[e^{\theta Y}] = (\frac{\lambda}{\lambda-\theta})^n$$

So $S_n \sim \Gamma(n,\lambda)$.

14.5 More on the normal distribution

14.5.1 Moment generating function

Suppose $X \sim N(\mu, \sigma^2)$. The mgf is found as follows:

$$\begin{aligned} E[e^{\theta X}] = \int_{-\infty}^{\infty} e^{\theta x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\sigma^2(x-\mu)^2}\,\mathrm{d}x. &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\theta(\mu+\sigma z)} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= e^{\theta\mu+\frac{1}{2}\theta^2\sigma^2} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\theta\sigma)^2}}_{\text{pdf of } N(\sigma\theta,1)}\,\mathrm{d}z \\ &= e^{\theta\mu+\frac{1}{2}\theta^2\sigma^2}. \end{aligned}$$

14.5.2 Functions of normal random variables

Theorem. Suppose X, Y are independent random variables with $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim (\mu_2, \sigma_2^2)$. Then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ and $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

Proof.

$$\mathbb{E}[e^{\theta(X+Y)}] = \mathbb{E}[e^{\theta X}] \cdot \mathbb{E}[e^{\theta Y}] = e^{\mu_1\theta+\frac{1}{2}\sigma_1^2\theta^2} \cdot e^{\mu_2\theta+\frac{1}{2}\sigma_2^2\theta_2} = e^{(\mu_1+\mu_2)\theta+\frac{1}{2}(\sigma_1^2+\sigma_2^2)\theta^2}$$

$$\mathbb{E}[e^{\theta(aX)}] = \mathbb{E}[e^{(\theta a)X}] = e^{\mu(a\theta)+\frac{1}{2}\sigma^2(a\theta)^2} = e^{(a\mu)\theta+\frac{1}{2}(a^2\sigma^2)\theta^2}$$

□

14.5.3 Bounds on tail probabilities

Suppose $X \sim N(0,1)$. Write $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for its pdf. It would be very difficult to find a closed form for its cdf, but we can find an upper bound for it:

$$\mathcal{P}(X \geq x) = \int_x^\infty \phi(t)\, \mathrm{d}t \leq \int_x^\infty \left(1 + \frac{1}{t^2}\right) \phi(t)\, \mathrm{d}t = \frac{1}{x}\phi(x)$$

To see the last step works, simply differentiate the result.So

$$\mathcal{P}(X \geq x) \leq \frac{1}{x}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \Rightarrow \log \mathcal{P}(X \geq x) \sim -\frac{1}{2}x^2.$$

14.6 Multivariate normal

Let $X_1, \cdots, X_n$ be iid $N(0,1)$. Then their joint density is ($\mathbf{x} = (x_1, \cdots, x_n)^T$)

$$g(x_1, \cdots, x_n) = \prod_{i=1}^n \phi(x_i) = \prod_1^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_1^n x_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}},$$

Suppose we are interested in $\mathbf{Z} = \boldsymbol{\mu} + A\mathbf{X}$, where A is an invertible $n \times n$ matrix. We can think of this as n measurements $\mathbf{Z}$ that are affected by underlying standard-normal factors $\mathbf{X}$. Then

$$\mathbf{X} = A^{-1}(\mathbf{Z} - \boldsymbol{\mu}) \quad \& \quad |J| = |\det(A^{-1})| = \frac{1}{\det A}$$

So

$$\begin{aligned} f(z_1, \cdots, z_n) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\det A} \exp\left[-\frac{1}{2}\left((A^{-1}(\mathbf{z} - \boldsymbol{\mu}))^T (A^{-1}(\mathbf{z} - \boldsymbol{\mu}))\right)\right] \\ &= \frac{1}{(2\pi)^{n/2} \det A} \exp\left[-\frac{1}{2}(\mathbf{z} - \mu)^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})\right] \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})\right]. \end{aligned}$$

where $\Sigma = AA^T$ and $\Sigma^{-1} = (A^{-1})^T A^{-1}$. We say

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \sim MUN(\boldsymbol{\mu}, \Sigma) \text{ or } N(\boldsymbol{\mu}, \Sigma).$$

This is the multivariate normal.

Then $\mathrm{cov}(Z_i, Z_j) = \mathbb{E}[(Z_i - \mu_i)(z_j - \mu_j)]$, which is the i, jth entry of Σ, since

$$\begin{aligned} \mathbb{E}[(Z - \mu)(Z - \mu)^T] = \mathbb{E}[AX(AX)^T] = \mathbb{E}(AXX^TA^T) = A\mathbb{E}[XX^T]A^T = AIA^T \\ = AA^T = \Sigma \end{aligned}$$

We have $\Sigma = \sigma^2$ when $n = 1$.

Now suppose $Z_1, \cdots, Z_n$ have covariances 0. Then it is the diagonal matrix $\Sigma = (\sigma_1^2, \cdots, \sigma_n^2)$. Then

$$f(z_1, \cdots, z_n) = \prod_1^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2\sigma_i^2}(z_i - \mu_i)^2}.$$

Then $Z_1, \cdots, Z_n$ are independent, with $Z_i \sim N(\mu_i, \sigma_i^2)$. Note that it is only true for normal distributions that $\text{cov} = 0 \Rightarrow$ independent. The moment generating function is

$$m(\boldsymbol{\theta}) = \mathbb{E}[e^{\boldsymbol{\theta}^T \mathbf{X}}] = \mathbb{E}[e^{\theta_1 X_1 + \cdots + \theta_n X_n}].$$

14.6.1 Bivariate normal

This is the special case of the multivariate normal when $n = 2$. Since there aren't too many terms, we can actually write them out.

The *bivariate normal* has

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \delta_2^2. \end{pmatrix}$$

Then

$$\text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\,\text{Var}(X_2)}} = \frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_1} = \rho.$$

Then

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{pmatrix}$$

The joint mgf of the bivariate normal is

$$m(\theta_1, \theta_2) = e^{\theta_1\mu_1 + \theta_2\mu_2 + \frac{1}{2}(\theta_1\sigma_1^2 + 2\theta_1\theta_2\rho\sigma_1\sigma_2 + \sigma_1\sigma_2^2)}.$$

15 Central Limit Theorem

15.1 Central Limit Theorem

Suppose $X_1 \cdots X_n$ are iid random variables with mean μ and variance σ?2. Let $S_n = X_1 + \cdots + X_n$.

$$\text{Var}(\frac{S_n}{\sqrt{n}}) = \text{Var}(\frac{S_n - n\mu}{\sqrt{n}}) = \sigma^2$$

Theorem. Let $X_1, \cdots, X_n$ be iid random variables with $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2 < \infty$. Define $S_n = X_1 + \cdots + X_n$. Then for all (a, b) such that $-\infty < a \leq b < \infty$:

$$\lim_{n\to\infty} P(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \mathrm{d}x$$

Then we write:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to_D N(0, 1)$$

Theorem (Continuity Theorem)**.** If random variables $X_1, \cdots$ have m.g.f.s $m_1(\theta), \cdots$ and $m_i(\theta) \to m(\theta)$ as $i \to \infty$, pointwise for every θ, then $X_n \to_D$ the random variable having m.g.f $m(\theta)$.

Proof. Let $\mu = 0, \sigma^2 = 1$. The m.g.f. of X_i is, as $n \to \infty$, $N(0, 1)$:

$$m_{X_i}(\theta) = 1 + \frac{1}{2}\theta^2 + \frac{1}{3!}E[X_i^3] + \cdots$$

The m.g.f of $\frac{S_n}{\sqrt{n}}$:

$$E[e^{\theta\frac{S_n}{\sqrt{n}}}] = E[e^{\frac{\theta}{\sqrt{n}}X_i}]^n = (1 + \frac{1}{2}\theta^2 + \frac{1}{3!}E[X_i^3] + \cdots)^n \to e^{\frac{1}{2}\theta^2}$$

□

15.2 Normal approximation to the binomial

If $S_n \sim B(n,p)$ so that $X_i = 1$ and 0 with probabilities p and $1-p$, respectively, then:

$$\frac{S_n - np}{\sqrt{npq}} \to_D N(0,1)$$

Example. An unknown fraction of the electorate, p, got Labour. We desire to find p with error not exceeding 0.005. How large the sample size be?

Let the fraction of labour votes in the sample be $p' = \frac{S_n}{n}$, with $S_n = X_1 + \cdots + X_n$ and $X_i \sim B(1,p)$. We are never 100% certain that it could be within this range, but we can be sure up to 95%. Then:

$$P(|p'-p| \leq 0.005) = P(|S_n - np| \leq 0.005n) = P(\frac{|S_n - np|}{\sqrt{npq}} \leq \frac{0.005n}{\sqrt{npq}})$$

We choose n such that:

$$\frac{0.005n}{\sqrt{npq}} \geq \mathcal{P}hi^{-1}(0.975) = 1.96$$

Although we don't know p. But $pq \leq \frac{1}{4}$. So we need:

$$n \geq 38416$$

If we replace 0.005 by 0.03, we get 1068, which is the usual sample size for polls.

15.3 Estimating π with Buffon's Needle

Example. A needle of length l is tossed at random onto a floor marked with parallel liens a distance L apart, where $l \leq L$. Recall that p =the probability that the needle intersects one of the parallel lines $= \frac{2l}{\pi L}$. N is # of times the point hits line in n trials.

$$N \sim B(n,p), \text{ so}, N \approx (np, np(1-p))$$

$$\hat{\pi} = \frac{2l}{\frac{N}{n}L} = \frac{\frac{\pi 2L}{\pi L}}{E[\frac{N}{n}] + (\frac{N}{n} - \frac{E[N]}{n})} = \frac{\pi p}{p + (p'-p)} = \pi(1 - \frac{p'-p}{p} + \cdots)$$

Since $p' \sim N(p, \frac{p(1-p)}{n})$. Then:

$$\hat{\pi} - \pi \sim N(0, \frac{\pi^2 p(1-p)}{np^2})$$

We have the smallest variance when we take p as large as possible, so $l = L$. Then:

$$\hat{\pi} - \pi \sim N(0, \frac{(\pi-2)\pi^2}{2n})$$

If we want it to be correct to third decimal point with 95% confidence, then $|\hat{\pi} - \pi| < 0.001$, then we have $n \geq 2.16 * 10^7$.

16 Summary of distributions

16.1 Discrete distributions

Distribution	PMF	Mean	Variance	PGF
Bernoulli	$p^k(1-p)^{1-k}$	p	$p(1-p)$	$q+pz$
Binomial	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(q+pz)^n$
Geometric	$(1-p)^k p$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$\frac{1-p}{1-pz}$
Poisson	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ	$e^{\lambda(z-1)}$

16.2 Continuous distributions

Distribution	PDF	CDF	Mean	Variance	PGF
Uniform	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{1}{12}(b-a)^2$	$\frac{e^{\theta b}-e^{\theta a}}{\theta(b-a)}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	/	μ	σ^2	$e^{\theta\mu+\frac{1}{2}\theta^2\sigma^2}$
Exponential	$\lambda e^{-\lambda x}$	$1-e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-\theta}$
Cauchy	$\frac{1}{\pi(1+x^2)}$	/	undefined	undefined	undefined
Gamma	$\frac{\lambda^n x^{n-1}e^{-\lambda x}}{(n-1)!}$	/	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-\theta}\right)^n$
Multivariate normal	$\frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})\right]$	/	$\boldsymbol{\mu}$	Σ	/