Part IA — Probability

Theorems with proof

Based on lectures by R. Weber

Notes taken by Dexter Chua

Lent 2015

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

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

0 Introduction

1 Classical probability

1.1 Classical probability

1.2 Counting

Theorem (Fundamental rule of counting)**.** Suppose we have to make r multiple choices in sequence. There are m_1 possibilities for the first choice, m_2 possibilities for the second etc. Then the total number of choices is $m_1 \times m_2 \times \cdots m_r$.

1.3 Stirling's formula

Proposition. $\log n! \sim n \log n$

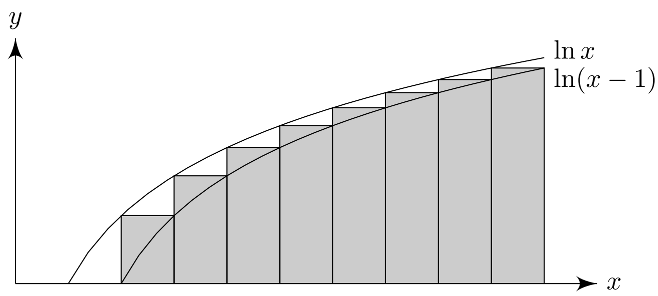
Proof. Note that

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

Now we claim that

$$\int_1^n \log x \; \mathrm{d}x \leq \sum_1^n \log k \leq \int_1^{n+1} \log x \; \mathrm{d}x.$$

This is true by considering the diagram:



We actually evaluate the integral to obtain

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

Divide both sides by $n \log n$ and let $n \to \infty$. Both sides tend to 1. So

$$\frac{\log n!}{n \log n} \to 1. \qquad \square$$

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

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

Corollary.

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

Proof. (non-examinable) Define

$$d_n = \log\left(\frac{n!e^n}{n^{n+1/2}}\right) = \log n! - (n+1/2)\log n + n$$

Then

$$d_n - d_{n+1} = (n+1/2)\log\left(\frac{n+1}{n}\right) - 1.$$

Write $t = 1/(2n+1)$. Then

$$d_n - d_{n+1} = \frac{1}{2t}\log\left(\frac{1+t}{1-t}\right) - 1.$$

We can simplifying by noting that

$$\begin{aligned}
\log(1+t) - t &= -\frac{1}{2}t^2 + \frac{1}{3}t^3 - \frac{1}{4}t^4 + \cdots\\
\log(1-t) + t &= -\frac{1}{2}t^2 - \frac{1}{3}t^3 - \frac{1}{4}t^4 - \cdots
\end{aligned}$$

Then if we subtract the equations and divide by $2t$, we obtain

$$\begin{aligned}
d_n - d_{n+1} &= \frac{1}{3}t^2 + \frac{1}{5}t^4 + \frac{1}{7}t^6 + \cdots\\
&< \frac{1}{3}t^2 + \frac{1}{3}t^4 + \frac{1}{3}t^6 + \cdots\\
&= \frac{1}{3}\frac{t^2}{1-t^2}\\
&= \frac{1}{3}\frac{1}{(2n+1)^2 - 1}\\
&= \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right)
\end{aligned}$$

By summing these bounds, we know that

$$d_1 - d_n < \frac{1}{12}\left(1 - \frac{1}{n}\right)$$

Then we know that d_n is bounded below by d_1+ something, and is decreasing since $d_n - d_{n+1}$ is positive. So it converges to a limit A. We know A is a lower bound for d_n since (d_n) is decreasing.

Suppose $m > n$. Then $d_n - d_m < \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{12}$. So taking the limit as $m \to \infty$, we obtain an upper bound for d_n: $d_n < A + 1/(12n)$. Hence we know that

$$A < d_n < A + \frac{1}{12n}.$$

However, all these results are useless if we don't know what A is. To find A, we have a small detour to prove a formula:

Take $I_n = \int_0^{\pi/2} \sin^n \theta \;\mathrm{d}\theta$. This is decreasing for increasing n as $\sin^n \theta$ gets smaller. We also know that

$$\begin{aligned}
I_n &= \int_0^{\pi/2} \sin^n \theta \;\mathrm{d}\theta\\
&= \left[-\cos\theta \sin^{n-1}\theta\right]_0^{\pi/2} + \int_0^{\pi/2} (n-1)\cos^2\theta\sin^{n-2}\theta \;\mathrm{d}\theta\\
&= 0 + \int_0^{\pi/2}(n-1)(1-\sin^2\theta)\sin^{n-2}\theta\;\mathrm{d}\theta\\
&= (n-1)(I_{n-2} - I_n)
\end{aligned}$$

So

$$I_n = \frac{n-1}{n} I_{n-2}.$$

We can directly evaluate the integral to obtain $I_0 = \pi/2$, $I_1 = 1$. Then

$$\begin{aligned}
I_{2n} &= \frac{1}{2}\cdot\frac{3}{4}\cdots\frac{2n-1}{2n}\pi/2 = \frac{(2n)!}{(2^n n!)^2}\frac{\pi}{2}\\
I_{2n+1} &= \frac{2}{3}\cdot\frac{4}{5}\cdots\frac{2n}{2n+1} = \frac{(2^n n!)^2}{(2n+1)!}
\end{aligned}$$

So using the fact that I_n is decreasing, we know that

$$1 \leq \frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} = 1 + \frac{1}{2n} \to 1.$$

Using the approximation $n! \sim n^{n+1/2}e^{-n+A}$, where A is the limit we want to find, we can approximate

$$\frac{I_{2n}}{I_{2n+1}} = \pi(2n+1)\left[\frac{((2n)!)^2}{2^{4n+1}(n!)^4}\right] \sim \pi(2n+1)\frac{1}{ne^{2A}} \to \frac{2\pi}{e^{2A}}.$$

Since the last expression is equal to 1, we know that $A = \log\sqrt{2\pi}$. Hooray for magic! $\square$

Proposition (non-examinable)**.** We use the $1/12n$ term from the proof above to get a better approximation:

$$\sqrt{2\pi}n^{n+1/2}e^{-n+\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi}n^{n+1/2}e^{-n+\frac{1}{12n}}.$$

2 Axioms of probability

2.1 Axioms and definitions

Theorem.

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

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

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

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

Proof.

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

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

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

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

Theorem. If $A_1, A_2, \cdots$ is increasing or decreasing, then
$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\lim_{n\to\infty} A_n\right).$$

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

Then
$$\begin{aligned}
\mathbb{P}(\lim A_n) &= \mathbb{P}\left(\bigcup_1^\infty A_i\right)\\
&= \mathbb{P}\left(\bigcup_1^\infty B_i\right)\\
&= \sum_1^\infty \mathbb{P}(B_i) \text{ (Axiom III)}\\
&= \lim_{n\to\infty} \sum_{i=1}^n \mathbb{P}(B_i)\\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_1^n A_i\right)\\
&= \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

and the decreasing case is proven similarly (or we can simply apply the above to A_i^C). $\square$

2.2 Inequalities and formulae

Theorem (Boole's inequality). For any $A_1, A_2, \cdots$,
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Proof. Our third axiom states a similar formula that only holds for disjoint sets. So we need a (not so) clever trick to make them disjoint. We define
$$\begin{aligned} B_1 &= A_1\\ B_2 &= A_2 \setminus A_1\\ B_i &= A_i \setminus \bigcup_{k=1}^{i-1} A_k. \end{aligned}$$

So we know that
$$\bigcup B_i = \bigcup A_i.$$

But the B_i are disjoint. So our Axiom (iii) gives
$$\mathbb{P}\left(\bigcup_i A_i\right) = \mathbb{P}\left(\bigcup_i B_i\right) = \sum_i \mathbb{P}\left(B_i\right) \leq \sum_i \mathbb{P}\left(A_i\right).$$

Where the last inequality follows from (iii) of the theorem above. $\square$

Theorem (Inclusion-exclusion formula).
$$\mathbb{P}\left(\bigcup_i^n A_i\right) = \sum_1^n \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} \cap A_{j_2}) + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n-1}\mathbb{P}(A_1 \cap \cdots \cap A_n).$$

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

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

Then we can apply the induction hypothesis for $n - 1$, and expand the mess. The details are very similar to that in IA Numbers and Sets. $\square$

Theorem (Bonferroni's inequalities). For any events $A_1, A_2, \cdots, A_n$ and $1 \leq r \leq n$, if r is odd, then
$$\mathbb{P}\left(\bigcup_1^n A_i\right) \leq \sum_{i_1} \mathbb{P}(A_{i_1}) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} A_{i_2}) + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} A_{i_2} A_{i_3}) + \cdots + \sum_{i_1 < i_2 < \cdots < i_r} \mathbb{P}(A_{i_1} A_{i_2} A_{i_3} \cdots A_{i_r}).$$

If r is even, then

$$\mathbb{P}\left(\bigcup_1^n A_i\right) \geq \sum_{i_1} \mathbb{P}(A_{i_1}) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} A_{i_2}) + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} A_{i_2} A_{i_3}) + \cdots$$
$$- \sum_{i_1 < i_2 < \cdots < i_r} \mathbb{P}(A_{i_1} A_{i_2} A_{i_3} \cdots A_{i_r}).$$

Proof. Easy induction on n. □

2.3 Independence

Proposition. If A and B are independent, then A and B^C are independent.

Proof.

$$\begin{aligned}
\mathbb{P}(A \cap B^C) &= \mathbb{P}(A) - \mathbb{P}(A \cap B)\\
&= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B)\\
&= \mathbb{P}(A)(1 - \mathbb{P}(B))\\
&= \mathbb{P}(A)\mathbb{P}(B^C)
\end{aligned}$$

□

2.4 Important discrete distributions

Theorem (Poisson approximation to binomial)**.** Suppose $n \to \infty$ and $p \to 0$ such that $np = \lambda$. Then

$$q_k = \binom{n}{k} p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Proof.

$$\begin{aligned}
q_k &= \binom{n}{k} p^k (1-p)^{n-k}\\
&= \frac{1}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} (np)^k \left(1 - \frac{np}{n}\right)^{n-k}\\
&\to \frac{1}{k!} \lambda^k e^{-\lambda}
\end{aligned}$$

since $(1 - a/n)^n \to e^{-a}$. □

2.5 Conditional probability

Theorem.

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

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

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

(iv) The function $\mathbb{P}(\,\cdot \mid B)$ restricted to subsets of B is a probability function (or measure).

Proof. Proofs of (i), (ii) and (iii) are trivial. So we only prove (iv). To prove this, we have to check the axioms.

(i) Let $A \subseteq B$. Then $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} \leq 1$.

(ii) $\mathbb{P}(B \mid B) = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$.

(iii) Let A_i be disjoint events that are subsets of B. Then

$$\begin{aligned}\mathbb{P}\left(\left.\bigcup_i A_i\right| B\right) &= \frac{\mathbb{P}(\bigcup_i A_i \cap B)}{\mathbb{P}(B)}\\ &= \frac{\mathbb{P}\left(\bigcup_i A_i\right)}{\mathbb{P}(B)}\\ &= \sum \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)}\\ &= \sum \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}\\ &= \sum \mathbb{P}(A_i \mid B).\end{aligned}$$ □

Proposition. If B_i is a partition of the sample space, and A is any event, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Theorem (Bayes' formula)**.** Suppose B_i is a partition of the sample space, and A and B_i all have non-zero probability. Then for any B_i,

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

Note that the denominator is simply $\mathbb{P}(A)$ written in a fancy way.

3 Discrete random variables

3.1 Discrete random variables

Theorem.

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.

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

(iii) If a and b are constants, then $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$.

(iv) If X and Y are random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This is true even if X and Y are not independent.

(v) $\mathbb{E}[X]$ 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

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

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

(iii)

$$\mathbb{E}[a + bX] = \sum_\omega (a + bX(\omega))p_\omega = a + b\sum_\omega p_\omega = a + b\ \mathbb{E}[X].$$

(iv)

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

(v)

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

This is clearly minimized when $c = \mathbb{E}[X]$. Note that we obtained the zero in the middle because $\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. □

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.

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

Theorem.

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

(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.

Proposition.

- $\mathbb{E}[I[A]] = \sum_\omega p(\omega) I[A](\omega) = \mathbb{P}(A)$.
- $I[A^C] = 1 - I[A]$.
- $I[A \cap B] = I[A]I[B]$.
- $I[A \cup B] = I[A] + I[B] - I[A]I[B]$.
- $I[A]^2 = I[A]$.

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

$$\begin{aligned}\mathbb{P}\left(\bigcup_i^n A_i\right) &= \sum_1^n \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} \cap A_{j_2}) + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots \\ &\quad + (-1)^{n-1}\mathbb{P}(A_1 \cap \cdots \cap A_n).\end{aligned}$$

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

$$S_r = \sum_{i_1 < i_2 < \cdots < i_r} I_{i_1} I_{i_2} \cdots I_{i_r},$$

and

$$s_r = \mathbb{E}[S_r] = \sum_{i_1 < \cdots < i_r} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_r}).$$

Then

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

So

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

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 given a particular y_i, there can be many different x_i for which $f_i(x_i) = y_i$. When finding $\mathbb{P}(f_i(x_i) = y_i)$, we need to sum over all x_i such that

$f_i(x_i) = f_i$. Then

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

Note that the switch from the second to third line is valid since they both expand to the same mess. □

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. Then

$$\begin{aligned}
\mathbb{E}\left[\prod_1^n X_i\right] &= \sum_{x_1 \in R_1} \cdots \sum_{x_n \in R_n} x_1 x_2 \cdots x_n \times \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) \\
&= \prod_{i=1}^n \sum_{x_i \in R_i} x_i \mathbb{P}(X_i = x_i) \\
&= \prod_{i=1}^n \mathbb{E}[X_i].
\end{aligned}$$

□

Corollary. Let $X_1, \cdots X_n$ be independent random variables, and $f_1, f_2, \cdots f_n$ are functions $\mathbb{R} \to \mathbb{R}$. Then

$$\mathbb{E}\left[\prod f_i(x_i)\right] = \prod \mathbb{E}[f_i(x_i)].$$

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.

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

Corollary. Let $X_1, X_2, \cdots X_n$ be independent identically distributed random variables (iid rvs). Then

$$\operatorname{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\operatorname{var}(X_1).$$

Proof.

$$\begin{aligned}
\operatorname{var}\left(\frac{1}{n}\sum X_i\right) &= \frac{1}{n^2}\operatorname{var}\left(\sum X_i\right) \\
&= \frac{1}{n^2}\sum \operatorname{var}(X_i) \\
&= \frac{1}{n^2} n \operatorname{var}(X_1) \\
&= \frac{1}{n}\operatorname{var}(X_1)
\end{aligned}$$

$\square$

3.2 Inequalities

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 $\mathbb{P}(X = x_i) = p_i$).

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

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

$$\begin{aligned}
f(p_1x_1 + \cdots + p_nx_n) &= f\left(p_1x_1 + (p_2 + \cdots + p_n)\frac{p_2x_2 + \cdots + l_nx_n}{p_2 + \cdots + p_n}\right) \\
&\leq p_1f(x_1) + (p_2 + \cdots p_n)f\left(\frac{p_2x_2 + \cdots + p_nx_n}{p_2 + \cdots + p_n}\right). \\
&\leq p_1f(x_1) + (p_2 + \cdots + p_n)\left[\frac{p_2}{(\,)}f(x_2) + \cdots + \frac{p_n}{(\,)}f(x_n)\right] \\
&= p_1f(x_1) + \cdots + p_n(x_n).
\end{aligned}$$

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

Strictly convex case is proved with $\leq$ replaced by $<$ 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 convex since its second derivative is $x^{-2} > 0$.

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

$$\mathbb{E}[f(x)] = \frac{1}{n}\sum -\log x_i = -\log \mathrm{GM}$$

and

$$f(\mathbb{E}[x]) = -\log\frac{1}{n}\sum x_i = -\log \mathrm{AM}$$

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

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

Proof. We make use of the indicator function. We have

$$I[|X| \geq \varepsilon] \leq \frac{|X|}{\varepsilon}.$$

This is proved by exhaustion: if $|X| \geq \varepsilon$, then LHS $= 1$ and RHS ≥ 1; If $|X| < \varepsilon$, then LHS $= 0$ and RHS is non-negative.

Take the expected value to obtain

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

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

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

Proof. Again, we have

$$I[\{|X| \geq \varepsilon\}] \leq \frac{x^2}{\varepsilon^2}.$$

Then take the expected value and the result follows. $\square$

3.3 Weak law of large numbers

Theorem (Weak law of large numbers)**.** 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 $\varepsilon > 0$,

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

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,

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

Theorem (Strong law of large numbers)**.**

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

We say

$$\frac{S_n}{n} \to_{\text{as}} \mu,$$

where “as” means “almost surely”.

3.4 Multiple random variables

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) If X, Y are independent, $\operatorname{cov}(X, Y) = 0$.

Proposition. $|\operatorname{corr}(X, Y)| \leq 1$.

Proof. Apply Cauchy-Schwarz to $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$. $\square$

Theorem. If X and Y are independent, then

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

Proof.

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \sum_x x\mathbb{P}(X = x \mid Y = y) \\ &= \sum_x x\mathbb{P}(X = x) \\ &= \mathbb{E}[X] \end{aligned}$$

$\square$

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

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

where the subscripts indicate what variable the expectation is taken over.

Proof.

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

□

3.5 Probability generating functions

Theorem. The distribution of X is uniquely determined by its probability generating function.

Proof. By definition, $p_0 = p(0)$, $p_1 = p'(0)$ etc. (where p' is the derivative of p). In general,

$$\left.\frac{\mathrm{d}^i}{\mathrm{d}z^i} p(z)\right|_{z=0} = i!p_i.$$

So we can recover $(p_0, p_1, \cdots)$ from $p(z)$. □

Theorem (Abel's lemma)**.**

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

If $p'(z)$ is continuous, then simply $\mathbb{E}[X] = p'(1)$.

Proof. For $z < 1$, we have

$$p'(z) = \sum_1^\infty r p_r z^{r-1} \leq \sum_1^\infty r p_r = \mathbb{E}[X].$$

So we must have

$$\lim_{z \to 1} p'(z) \leq \mathbb{E}[X].$$

On the other hand, for any ε, if we pick N large, then

$$\sum_1^N r p_r \geq \mathbb{E}[X] - \varepsilon.$$

So

$$\mathbb{E}[X] - \varepsilon \leq \sum_1^N r p_r = \lim_{z \to 1} \sum_1^N r p_r z^{r-1} \leq \lim_{z \to 1} \sum_1^\infty r p_r z^{r-1} = \lim_{z \to 1} p'(z).$$

So $\mathbb{E}[X] \leq \lim_{z \to 1} p'(z)$. So the result follows □

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

Proof. Same as above. □

Theorem. Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with pgfs $p_1, p_2, \cdots, p_n$. Then the pgf of $X_1 + X_2 + \cdots + X_n$ is $p_1(z)p_2(z)\cdots p_n(z)$.

Proof.
$$\mathbb{E}[z^{X_1+\cdots+X_n}] = \mathbb{E}[z^{X_1}\cdots z^{X_n}] = \mathbb{E}[z^{X_1}]\cdots\mathbb{E}[z^{X_n}] = p_1(z)\cdots p_n(z).$$
□

4 Interesting problems

4.1 Branching processes

Theorem.

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

Proof.

$$\begin{aligned}
F_{n+1}(z) &= \mathbb{E}[z^{X_{n+1}}] \\
&= \mathbb{E}[\mathbb{E}[z^{X_{n+1}} \mid X_n]] \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n = k)\mathbb{E}[z^{X_{n+1}} \mid X_n = k] \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n = k)\mathbb{E}[z^{Y_1^n + \cdots + Y_k^n} \mid X_n = k] \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n = k)\mathbb{E}[z^{Y_1}]\mathbb{E}[z^{Y_2}]\cdots\mathbb{E}[z^{Y_n}] \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n = k)(\mathbb{E}[z^{X_1}])^k \\
&= \sum_{k=0} \mathbb{P}(X_n = k)F(z)^k \\
&= F_n(F(z))
\end{aligned}$$

□

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

$$\mathbb{E}[X_n] = \mu^n, \quad \operatorname{var} X_n = \sigma^2\mu^{n-1}(1 + \mu + \mu^2 + \cdots + \mu^{n-1}).$$

Proof.

$$\begin{aligned}
\mathbb{E}[X_n] &= \mathbb{E}[\mathbb{E}[X_n \mid X_{n-1}]] \\
&= \mathbb{E}[\mu X_{n-1}] \\
&= \mu\mathbb{E}[X_{n-1}]
\end{aligned}$$

Then by induction, $\mathbb{E}[X_n] = \mu^n$ (since $X_0 = 1$).

To calculate the variance, note that

$$\operatorname{var}(X_n) = \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2$$

and hence

$$\mathbb{E}[X_n^2] = \operatorname{var}(X_n) + (\mathbb{E}[X])^2$$

We then calculate

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

So

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

Of course, we can also obtain this using the probability generating function as well. □

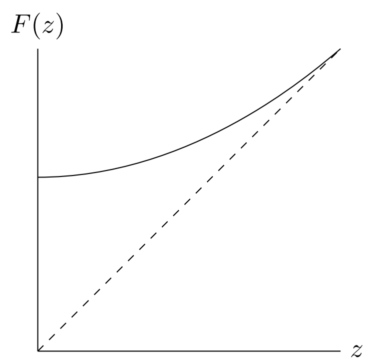
Theorem. The probability of extinction q is the smallest root to the equation $q = F(q)$. Write $\mu = \mathbb{E}[X_1]$. Then if $\mu \leq 1$, then $q = 1$; if $\mu > 1$, then $q < 1$.

Proof. To show that it is the smallest root, let α be the smallest root. Then note that $0 \leq \alpha \Rightarrow F(0) \leq F(\alpha) = \alpha$ since F is increasing (proof: write the function out!). Hence $F(F(0)) \leq \alpha$. Continuing inductively, $F_n(0) \leq \alpha$ for all n. So

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

So $q = \alpha$.

To show that $q = 1$ when $\mu \leq 1$, we show that $q = 1$ is the only root. We know that $F'(z), F''(z) \geq 0$ for $z \in (0, 1)$ (proof: write it out again!). So F is increasing and convex. Since $F'(1) = \mu \leq 1$, it must approach $(1, 1)$ from above the $F = z$ line. So it must look like this:



So $z = 1$ is the only root. □

4.2 Random walk and gambler's ruin

5 Continuous random variables

5.1 Continuous random variables

Proposition. The exponential random variable is *memoryless*, i.e.

$$\mathbb{P}(X \geq x + z \mid X \geq x) = \mathbb{P}(X \geq z).$$

This means that, say if X measures the lifetime of a light bulb, knowing it has already lasted for 3 hours does not give any information about how much longer it will last.

Proof.

$$\begin{aligned}\mathbb{P}(X \geq x + z \mid X \geq x) &= \frac{\mathbb{P}(X \geq x + z)}{\mathbb{P}(X \geq x)}\\ &= \frac{\int_{x+z}^{\infty} f(u)\;\mathrm{d}u}{\int_x^{\infty} f(u)\;\mathrm{d}u}\\ &= \frac{e^{-\lambda(x+z)}}{e^{-\lambda x}}\\ &= e^{-\lambda z}\\ &= \mathbb{P}(X \geq z).\end{aligned}$$

□

Theorem. If X is a continuous random variable, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x)\;\mathrm{d}x - \int_0^\infty \mathbb{P}(X \leq -x)\;\mathrm{d}x.$$

Proof.

$$\begin{aligned}\int_0^\infty \mathbb{P}(X \geq x)\;\mathrm{d}x &= \int_0^\infty \int_x^\infty f(y)\;\mathrm{d}y\;\mathrm{d}x\\ &= \int_0^\infty \int_0^\infty I[y \geq x] f(y)\;\mathrm{d}y\;\mathrm{d}x\\ &= \int_0^\infty \left(\int_0^\infty I[x \leq y]\;\mathrm{d}x\right) f(y)\;\mathrm{d}y\\ &= \int_0^\infty y f(y)\;\mathrm{d}y.\end{aligned}$$

We can similarly show that $\int_0^\infty \mathbb{P}(X \leq -x)\;\mathrm{d}x = -\int_{-\infty}^0 y f(y)\;\mathrm{d}y$. □

5.2 Stochastic ordering and inspection paradox

5.3 Jointly distributed random variables

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

Proof. We prove this by showing that X has a density function.

We know that

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(X \in A, Y \in (-\infty, +\infty))\\ &= \int_{x \in A} \int_{-\infty}^{\infty} f(x, y)\;\mathrm{d}y\;\mathrm{d}x\\ &= \int_{x \in A} f_X(x)\;\mathrm{d}x\end{aligned}$$

So

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\;\mathrm{d}y$$

is the (marginal) pdf of X. □

Proposition. For independent continuous random variables X_i,

(i) $\mathbb{E}[\prod X_i] = \prod \mathbb{E}[X_i]$

(ii) $\operatorname{var}(\sum X_i) = \sum \operatorname{var}(X_i)$

5.4 Geometric probability

5.5 The normal distribution

Proposition.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\;\mathrm{d}x = 1.$$

Proof. Substitute $z = \frac{(x-\mu)}{\sigma}$. Then

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\;\mathrm{d}z.$$

Then

$$\begin{aligned}I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\;\mathrm{d}x \int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\;\mathrm{d}y\\ &= \int_0^{\infty} \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} r\;\mathrm{d}r\;\mathrm{d}\theta\\ &= 1.\end{aligned}$$

□

Proposition. $\mathbb{E}[X] = \mu$.

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/2\sigma^2}\;\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu) e^{-(x-\mu)^2/2\sigma^2}\;\mathrm{d}x + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \mu e^{-(x-\mu)^2/2\sigma^2}\;\mathrm{d}x.\end{aligned}$$

The first term is antisymmetric about μ and gives 0. The second is just μ times the integral we did above. So we get μ. □

Proposition. $\text{var}(X) = \sigma^2$.

Proof. We have $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Substitute $Z = \frac{X-\mu}{\sigma}$. Then $\mathbb{E}[Z] = 0$, $\mathbb{E}[Z^2] = \frac{1}{\sigma^2}\mathbb{E}[X^2]$.

Then

$$\begin{aligned}\text{var}(Z) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,\mathrm{d}z\\ &= \left[-\frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\,\mathrm{d}z\\ &= 0 + 1\\ &= 1\end{aligned}$$

So $\text{var}\, X = \sigma^2$. □

5.6 Transformation of random variables

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

$$f_Y(y) = f_X(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y).$$

Proof.

$$\begin{aligned}F_Y(y) &= \mathbb{P}(Y \leq y)\\ &= \mathbb{P}(h(X) \leq y)\\ &= \mathbb{P}(X \leq h^{-1}(y))\\ &= F(h^{-1}(y))\end{aligned}$$

Take the derivative with respect to y to obtain

$$f_Y(y) = F'_Y(y) = f(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y).$$ □

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

Proof.

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$ □

Proposition. $(Y_1, \cdots, Y_n)$ has density

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

if $(y_1, \cdots, y_n) \in S$, 0 otherwise.

5.7 Moment generating functions

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

Theorem. The rth moment X is the coefficient of $\frac{\theta^r}{r!}$ in the power series expansion of $m(\theta)$, and is

$$\mathbb{E}[X^r] = \frac{\mathrm{d}^n}{\mathrm{d}\theta^n} m(\theta)\bigg|_{\theta=0} = m^{(n)}(0).$$

Proof. We have

$$e^{\theta X} = 1 + \theta X + \frac{\theta^2}{2!}X^2 + \cdots .$$

So

$$m(\theta) = \mathbb{E}[e^{\theta X}] = 1 + \theta\mathbb{E}[X] + \frac{\theta^2}{2!}\mathbb{E}[X^2] + \cdots \qquad \square$$

Theorem. If X and Y are independent random variables with moment generating functions $m_X(\theta), m_Y(\theta)$, then $X + Y$ has mgf $m_{X+Y}(\theta) = m_X(\theta)m_Y(\theta)$.

Proof.

$$\mathbb{E}[e^{\theta(X+Y)}] = \mathbb{E}[e^{\theta X}e^{\theta Y}] = \mathbb{E}[e^{\theta X}]\mathbb{E}[e^{\theta Y}] = m_X(\theta)m_Y(\theta). \qquad \square$$

6 More distributions

6.1 Cauchy distribution

Proposition. The mean of the Cauchy distribution is undefined, while $\mathbb{E}[X^2] = \infty$.

Proof.

$$\mathbb{E}[X] = \int_0^\infty \frac{x}{\pi(1+x^2)}\,\mathrm{d}x + \int_{-\infty}^0 \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \infty - \infty$$

which is undefined, but $\mathbb{E}[X^2] = \infty + \infty = \infty$. □

6.2 Gamma distribution

6.3 Beta distribution*

6.4 More on the normal distribution

Proposition. The moment generating function of $N(\mu, \sigma^2)$ is

$$\mathbb{E}[e^{\theta X}] = \exp\left(\theta\mu + \frac{1}{2}\theta^2\sigma^2\right).$$

Proof.

$$\mathbb{E}[e^{\theta X}] = \int_{-\infty}^\infty e^{\theta x}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\,\mathrm{d}x.$$

Substitute $z = \frac{x-\mu}{\sigma}$. Then

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

□

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

(i) $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(ii) $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

Proof.

(i)

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

which is the mgf of $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(ii)

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

□

6.5 Multivariate normal

7 Central limit theorem

Theorem (Central limit theorem)**.** Let $X_1, X_2, \cdots$ be iid random variables with $\mathbb{E}[X_i] = \mu$, $\mathrm{var}(X_i) = \sigma^2 < \infty$. Define

$$S_n = X_1 + \cdots + X_n.$$

Then for all finite intervals (a, b),

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

Note that the final term is the pdf of a standard normal. We say

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

Theorem (Continuity theorem)**.** If the random variables $X_1, X_2, \cdots$ have mgf's $m_1(\theta), m_2(\theta), \cdots$ and $m_n(\theta) \to m(\theta)$ as $n \to \infty$ for all θ, then $X_n \to_D$ the random variable with mgf $m(\theta)$.

Proof. wlog, assume $\mu = 0, \sigma^2 = 1$ (otherwise replace X_i with $\frac{X_i - \mu}{\sigma}$).
Then

$$\begin{aligned} m_{X_i}(\theta) = \mathbb{E}[e^{\theta X_i}] &= 1 + \theta\mathbb{E}[X_i] + \frac{\theta^2}{2!}\mathbb{E}[X_i^2] + \cdots \\ &= 1 + \frac{1}{2}\theta^2 + \frac{1}{3!}\theta^3\mathbb{E}[X_i^3] + \cdots \end{aligned}$$

Now consider $S_n/\sqrt{n}$. Then

$$\begin{aligned} \mathbb{E}[e^{\theta S_n/\sqrt{n}}] &= \mathbb{E}[e^{\theta(X_1 + \ldots + X_n)/\sqrt{n}}] \\ &= \mathbb{E}[e^{\theta X_1/\sqrt{n}}] \cdots \mathbb{E}[e^{\theta X_n/\sqrt{n}}] \\ &= \left(\mathbb{E}[e^{\theta X_1/\sqrt{n}}]\right)^n \\ &= \left(1 + \frac{1}{2}\theta^2\frac{1}{n} + \frac{1}{3!}\theta^3\mathbb{E}[X^3]\frac{1}{n^{3/2}} + \cdots\right)^n \\ &\to e^{\frac{1}{2}\theta^2} \end{aligned}$$

as $n \to \infty$ since $(1 + a/n)^n \to e^a$. And this is the mgf of the standard normal. So the result follows from the continuity theorem. $\square$

8 Summary of distributions

8.1 Discrete distributions

8.2 Continuous distributions