Here we are assuming that each outcome is equally likely to happen, which is the case in (fair) dice rolls and coin flips.

Example. Suppose r digits are drawn at random from a table of random digits from 0 to 9. What is the probability that

(i) No digit exceeds k;

(ii) The largest digit drawn is k?

The sample space is $\Omega = \{(a_1, a_2, \cdots, a_r) : 0 \leq a_i \leq 9\}$. Then $|\Omega| = 10^r$.

Let $A_k = [\text{no digit exceeds } k] = \{(a_1, \cdots, a_r) : 0 \leq a_i \leq k\}$. Then $|A_k| = (k+1)^r$. So

$$P(A_k) = \frac{(k+1)^r}{10^r}.$$

Now let $B_k = [\text{largest digit drawn is } k]$. We can find this by finding all outcomes in which no digits exceed k, and subtract it by the number of outcomes in which no digit exceeds $k-1$. So $|B_k| = |A_k| - |A_{k-1}|$ and

$$P(B_k) = \frac{(k+1)^r - k^r}{10^r}.$$

1.2 Counting

To find probabilities, we often need to *count* things. For example, in our example above, we had to count the number of elements in B_k.

Example. A menu has 6 starters, 7 mains and 6 desserts. How many possible meals combinations are there? Clearly $6 \times 7 \times 6 = 252$.

Here we are using the fundamental rule of 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 $m_1 \times m_2 \times \cdots m_r$.

Example. How many ways can $1, 2, \cdots, n$ be ordered? The first choice has n possibilities, the second has $n-1$ possibilities etc. So there are $n \times (n-1) \times \cdots \times 1 = n!$.

Sampling with or without replacement

Suppose we have to pick n items from a total of x items. We can model this as follows: Let $N = \{1, 2, \cdots, n\}$ be the list. Let $X = \{1, 2, \cdots, x\}$ be the items. Then each way of picking the items is a function $f : N \to X$ with $f(i) = $ item at the ith position.

Definition (Sampling with replacement)**.** When we *sample with replacement*, after choosing at item, it is put back and can be chosen again. Then *any* sampling function f satisfies sampling with replacement.

Definition (Sampling without replacement)**.** When we *sample without replacement*, after choosing an item, we kill it with fire and cannot choose it again. Then f must be an injective function, and clearly we must have $x \geq n$.

- Re-number each item by the number of the draw on which it was first seen. For example, we can rename $(2, 5, 2)$ and $(5, 4, 5)$ both as $(1, 2, 1)$. This happens if the labelling of items doesn't matter.
- Both of above. So we can rename $(2, 5, 2)$ and $(8, 5, 5)$ both as $(1, 1, 2)$.

Total number of cases

Combining these four possibilities with whether we have replacement, no replacement, or "everything has to be chosen at least once", we have 12 possible cases of counting. The most important ones are:

- Replacement + with ordering: the number of ways is x^n.
- Without replacement + with ordering: the number of ways is $x_{(n)} = x^{\underline{n}} = x(x-1)\cdots(x-n+1)$.
- Without replacement + without order: we only care which items get selected. The number of ways is $\binom{x}{n} = C_n^x = x_{(n)}/n!$.
- Replacement + without ordering: we only care how many times the item got chosen. This is equivalent to partitioning n into $n_1 + n_2 + \cdots + n_k$. Say $n = 6$ and $k = 3$. We can write a particular partition as

  $$** \mid * \mid ***$$

  So we have $n + k - 1$ symbols and $k - 1$ of them are bars. So the number of ways is $\binom{n+k-1}{k-1}$.

Multinomial coefficient

Suppose that we have to pick n items, and each item can either be an apple or an orange. The number of ways of picking such that k apples are chosen is, by definition, $\binom{n}{k}$.

In general, suppose we have to fill successive positions in a list of length n, with replacement, from a set of k items. The number of ways of doing so such that item i is picked n_i times is defined to be the *multinomial coefficient* $\binom{n}{n_1, n_2, \cdots, n_k}$.

Definition (Multinomial coefficient)**.** A *multinomial coefficient* is

$$\binom{n}{n_1, n_2, \cdots, n_k} = \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1\cdots-n_{k-1}}{n_k} = \frac{n!}{n_1!n_2!\cdots n_k!}.$$

It is the number of ways to distribute n items into k positions, in which the ith position has n_i items.

Example. We know that

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \cdots + y^n.$$

If we have a trinomial, then

$$(x+y+z)^n = \sum_{n_1, n_2, n_3}\binom{n}{n_1, n_2, n_3}x^{n_1}y^{n_2}z^{n_3}.$$

Example. How many ways can we deal 52 cards to 4 player, each with a hand of 13? The total number of ways is

$$\binom{52}{13, 13, 13, 13} = \frac{52!}{(13!)^4} = 53644737765488792839237440000 = 5.36 \times 10^{28}.$$

While computers are still capable of calculating that, what if we tried more ~~power~~ cards? Suppose each person has n cards. Then the number of ways is

$$\frac{(4n)!}{(n!)^4},$$

which is *huge*. We can use Stirling's Formula to approximate it:

1.3 Stirling's formula

Before we state and prove Stirling's formula, we prove a weaker (but examinable) version:

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

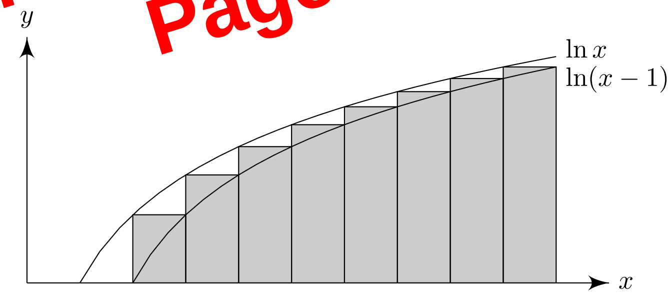
Proof. Note that

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

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

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

Now we prove Stirling's Formula:

(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. ☐

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

The next theorem is better expressed in terms of *limits*.

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

$$\lim_{n\to\infty} A_n = \bigcup_1^\infty A_n.$$

Similarly, if they are *decreasing*, i.e. $A_1 \supseteq A_2 \cdots$, then

$$\lim_{n\to\infty} A_n = \bigcap_1^\infty A_n.$$

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

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

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

Recall that, from inclusion exclusion,

$$\mathbb{P}(A\cup B\cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(AB) - \mathbb{P}(BC) - \mathbb{P}(AC) + \mathbb{P}(ABC),$$

where $\mathbb{P}(AB)$ is a shorthand for $\mathbb{P}(A\cap B)$. If we only take the first three terms, then we get Boole's inequality

$$\mathbb{P}(A\cup B\cup C)\leq \mathbb{P}(A)+\mathbb{P}(B)+\mathbb{P}(C).$$

In general

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

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

If r is even, then

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

Proof. Easy induction on n. □

Example. Let $\Omega = \{1, 2, \cdots, m\}$ and $1\leq j, k\leq m$. Write $A_k = \{1, 2, \cdots, k\}$. Then

$$A_k\cap A_j = \{1, 2, \cdots, \min(j,k)\} = A_{\min(j,k)}$$

and

$$A_k\cup A_j = \{1, 2, \cdots, \max(j,k)\} = A_{\max(j,k)}.$$

We also have $\mathbb{P}(A_k) = k/m$.

Now let $1\leq x_1, \cdots, x_n\leq m$ be some numbers. Then Bonferroni's inequality says

$$\mathbb{P}\left(\bigcup A_{x_i}\right)\geq \sum \mathbb{P}(A_{x_i}) - \sum_{i<j}\mathbb{P}(A_{x_i}\cap A_{x_j}).$$

So

$$\max\{x_1, x_2, \cdots, x_n\}\geq \sum x_i - \sum_{i_1<i_2}\min\{x_1, x_2\}.$$

Example. Let A_{ij} be the event that i and j roll the same. We roll 4 dice. Then

$$\mathbb{P}(A_{12} \cap A_{13}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = \mathbb{P}(A_{12})\mathbb{P}(A_{13}).$$

But

$$\mathbb{P}(A_{12} \cap A_{13} \cap A_{23}) = \frac{1}{36} \neq \mathbb{P}(A_{12})\mathbb{P}(A_{13})\mathbb{P}(A_{23}).$$

So they are not mutually independent.

We can also apply this concept to experiments. Suppose we model two independent experiments with $\Omega_1 = \{\alpha_1, \alpha_2, \cdots\}$ and $\Omega_2 = \{\beta_1, \beta_2, \cdots\}$ with probabilities $\mathbb{P}(\alpha_i) = p_i$ and $\mathbb{P}(\beta_i) = q_i$. Further suppose that these two experiments are independent, i.e.

$$\mathbb{P}((\alpha_i, \beta_j)) = p_i q_j$$

for all i, j. Then we can have a new sample space $\Omega = \Omega_1 \times \Omega_2$.

Now suppose $A \subseteq \Omega_1$ and $B \subseteq \Omega_2$ are results (i.e. events) of the two experiments. We can view them as subspaces of Ω by rewriting them as $A \times \Omega_2$ and $\Omega_1 \times B$. Then the probability

$$\mathbb{P}(A \cap B) = \sum_{\alpha_i \in A, \beta_i \in B} p_i q_i = \sum_{\alpha_i \in A} p_i \sum_{\beta_i \in B} q_i = \mathbb{P}(A)\mathbb{P}(B).$$

So we say the two experiments are "independent" even though the term usually refers to different events in the same experiment. We can generalize this to n independent experiments, or even countably many infinite experiments.

2.4 Important discrete distributions

We're now going to quickly go through a few important discrete probability distributions. By *discrete* we mean the sample space is countable. The sample space is $\Omega = \{\omega_1, \omega_2, \cdots\}$ and $p_i = \mathbb{P}(\{\omega_i\})$.

Definition (Bernoulli distribution)**.** Suppose we toss a coin. $\Omega = \{H, T\}$ and $p \in [0, 1]$. The *Bernoulli distribution*, denoted $B(1, p)$ has

$$\mathbb{P}(H) = p; \quad \mathbb{P}(T) = 1 - p.$$

Definition (Binomial distribution)**.** Suppose we toss a coin n times, each with probability p of getting heads. Then

$$\mathbb{P}(HHTT \cdots T) = pp(1 - p) \cdots (1 - p).$$

So

$$\mathbb{P}(\text{two heads}) = \binom{n}{2} p^2 (1 - p)^{n-2}.$$

In general,

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

We call this the *binomial distribution* and write it as $B(n, p)$.

Note that if A and B are independent, then

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

Example. In a game of poker, let A_i = [player i gets royal flush]. Then

$$\mathbb{P}(A_1) = 1.539 \times 10^{-6}.$$

and

$$\mathbb{P}(A_2 \mid A_1) = 1.969 \times 10^{-6}.$$

It is significantly bigger, albeit still incredibly tiny. So we say "good hands attract".

If $\mathbb{P}(A \mid B) > \mathbb{P}(A)$, then we say that B attracts A. Since

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} > \mathbb{P}(A) \Leftrightarrow \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} > \mathbb{P}(B),$$

A attracts B if and only if B attracts A. We can also say A repels B if A attracts B^C.

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 $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \le 1$.

(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

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

□

(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 $\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. □

An easy generalization of (iv) above is

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 (Variance and standard deviation)**.** The *variance* of a random variable X is defined as

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

The *standard deviation* is the square root of the variance, $\sqrt{\mathrm{var}(X)}$.

We also have

$$\begin{aligned}
\mathbb{E}[N^2] &= \mathbb{E}\left[\left(\sum I[A_i]\right)^2\right]\\
&= \mathbb{E}\left[\sum_i I[A_i]^2 + 2\sum_{i<j} I[A_i]I[A_j]\right]\\
&= n\mathbb{E}\big[I[A_i]\big] + n(n-1)\mathbb{E}\big[I[A_1]I[A_2]\big]
\end{aligned}$$

We have $\mathbb{E}[I[A_1]I[A_2]] = \mathbb{P}(A_1 \cap A_2) = \frac{2}{n}\left(\frac{1}{n-1}\frac{1}{n-1} + \frac{n-2}{n-1}\frac{2}{n-1}\right)$. Plugging in, we ultimately obtain $\operatorname{var}(N) = \frac{2(n-2)}{n-1}$.

In fact, as $n \to \infty$, $N \sim P(2)$.

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

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. 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

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

We can extend the idea of independence to random variables. Two random variables are independent if the value of the first does not affect the value of the second.

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

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

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. 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. □

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

We have

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

and

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

So

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

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.

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

Example. Let $X \sim B(n, p)$. Then

$$p(z) = \sum_{r=0}^{n} \mathbb{P}(X = r)z^r = \sum \binom{n}{r} p^r(1-p)^{n-r}z^r = (pz + (1-p))^n = (pz+q)^n.$$

So $p(z)$ is the product of n copies of $pz + q$. But $pz + q$ is the pgf of $Y \sim B(1, p)$.

This shows that $X = Y_1 + Y_2 + \cdots + Y_n$ (which we already knew), i.e. a binomial distribution is the sum of Bernoulli trials.

Example. If X and Y are independent Poisson random variables with parameters λ, μ respectively, then

$$\mathbb{E}[t^{X+Y}] = \mathbb{E}[t^X]\mathbb{E}[t^Y] = e^{\lambda(t-1)}e^{\mu(t-1)} = e^{(\lambda+\mu)(t-1)}$$

So $X + Y \sim \mathbb{P}(\lambda + \mu)$.

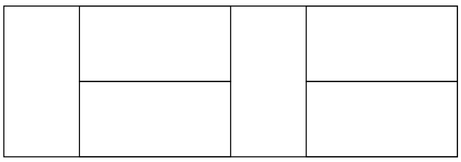
We can also do it directly:

$$\mathbb{P}(X + Y = r) = \sum_{i=0}^{r} \mathbb{P}(X = i, Y = r - i) = \sum_{i=0}^{r} \mathbb{P}(X = i)\mathbb{P}(X = r - i),$$

but is much more complicated.

We can use pgf-like functions to obtain some combinatorial results.

Example. Suppose we want to tile a $2 \times n$ bathroom by 2×1 tiles. One way to do it is



In general, we get the following result:

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

This condition "strictly increasing" is needed for the inverse to exist. If it is not, we can define

$$F^{-1}(u) = \inf\{x : F(x) \geq u, 0 < u < 1\},$$

and the same result holds.

This can also be done for discrete random variables $\mathbb{P}(X = x_i) = p_i$ by letting

$$X = x_j \text{ if } \sum_{i=0}^{j-1} p_i \leq U < \sum_{i=0}^{j} p_i,$$

for $U \sim U[0, 1]$.

Multiple random variables

Suppose $X_1, X_2, \cdots, X_n$ are random variables with joint pdf f. Let

$$\begin{aligned} Y_1 &= r_1(X_1, \cdots, X_n)\\ Y_2 &= r_2(X_1, \cdots, X_n)\\ &\vdots\\ Y_n &= r_n(X_1, \cdots, X_n). \end{aligned}$$

For example, we might have $Y_1 = \frac{X_1}{X_1+X_2}$ and $Y_2 = X_1 + X_2$.

Let $R \subseteq \mathbb{R}^n$ such that $\mathbb{P}((X_1, \cdots, X_n) \in R) = 1$, i.e. R is the set of all values (X_i) can take.

Suppose S is the image of R under the above transformation, and the map $R \to S$ is bijective. Then there exists an inverse function

$$\begin{aligned} X_1 &= s_1(Y_1, \cdots, Y_n)\\ X_2 &= s_2(Y_1, \cdots, Y_n)\\ &\vdots\\ X_n &= s_n(Y_1, \cdots, Y_n). \end{aligned}$$

For example, if X_1, X_2 refers to the coordinates of a random point in Cartesian coordinates, Y_1, Y_2 might be the coordinates in polar coordinates.

Definition (Jacobian determinant)**.** Suppose $\frac{\partial s_i}{\partial y_j}$ exists and is continuous at every point $(y_1, \cdots, y_n) \in S$. Then the *Jacobian determinant* is

$$J = \frac{\partial(s_1, \cdots, s_n)}{\partial(y_1, \cdots, y_n)} = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

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

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = A\mathbf{X}$$

Then $|J| = 1/|\det A| = 1$. Then

$$g(y, z) = f(y - z, z)$$

So

$$g_Y(y) = \int_{-\infty}^{\infty} f(y - z, z)\;\mathrm{d}z = \int_{-\infty}^{\infty} f(z, y - z)\;\mathrm{d}z.$$

If X_1 and X_2 are independent, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Then

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

Non-injective transformations

We previously discussed transformation of random variables by injective maps. What if the mapping is not? There is no simple formula for that, and we have to work out each case individually.

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

We use our definition. We have

$$\mathbb{P}(|X| \in x(a, b)) = \int_a^b f(x) + \int_{-b}^{-a} f(x)\;\mathrm{d}x = \int_a^b (f(x) + f(-x))\;\mathrm{d}x.$$

So

$$f_Y(x) = f(x) + f(-x),$$

which makes sense, since getting $|X| = x$ is equivalent to getting $X = x$ or $X = -x$.

Example. Suppose $X_1 \sim \mathcal{E}(\lambda), X_2 \sim \mathcal{E}(\mu)$ are independent random variables. Let $Y = \min(X_1, X_2)$. Then

$$\begin{aligned} \mathbb{P}(Y \geq t) &= \mathbb{P}(X_1 \geq t, X_2 \geq t) \\ &= \mathbb{P}(X_1 \geq t)\mathbb{P}(X_2 \geq t) \\ &= e^{-\lambda t} e^{-\mu t} \\ &= e^{-(\lambda + \mu)t}. \end{aligned}$$

So $Y \sim \mathcal{E}(\lambda + \mu)$.

Given random variables, not only can we ask for the minimum of the variables, but also ask for, say, the second-smallest one. In general, we define the *order statistics* as follows:

Definition (Order statistics)**.** Suppose that $X_1, \cdots, X_n$ are some random variables, and $Y_1, \cdots, Y_n$ is $X_1, \cdots, X_n$ arranged in increasing order, i.e. $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. This is the *order statistics*.

We sometimes write $Y_i = X_{(i)}$.