6.875: Cryptography and Cryptanalysis September 17, 2025

Background on probability theory

Instructor: Yael Kalai **TAs:** Aparna Gupte and Andrew Huang

Contents

1 Basic theory

1.1 Definitions and background

Probability is always defined over a set of possible outcomes Ω of a (discrete) random experiment. For example, the possible outcomes for a single coin flip are heads (H) or tails (T), and so we have $\Omega = \{H, T\}$. We call Ω the *sample space* of the experiment. On the other hand, if our random experiment consists of flipping a coin n times independently, then

$$\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \{H, T\}\}.$$

A **probability distribution** over Ω is a function $p : \Omega \to \mathbb{R}_{\geq 0}$ such that $\sum_{x \in \Omega} p(x) = 1$. That is, the sum of all p_i over all possible outcomes in Ω is 1.

An **event** is any subset $A \subseteq \Omega$. The probability of an event A is

$$\Pr_p[A] = \sum_{x \in A} p(x).$$

In words, we can define the probability of an event in a uniform distribution as

$$\Pr[\text{event happens}] = \frac{\text{number of ways it can happen}}{\text{total number of outcomes}}$$

Note: We will often just write Pr instead of $\Pr_p$ when the distribution p is clear from context.

Two events $A, B \subseteq \Omega$ are said to be **independent**, if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$.

Let look at an example involving coin tosses.

Example 1: Tossing a coin

Recall the experiment of tossing a coin n times: $\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \{H, T\}\}$.
The event that the first flip is heads is represented as the set

$$A_{1,H} = \{(H, \omega_2, \ldots, \omega_n) : \omega_i \in \{H, T\}\},$$

and similarly the event that the first flip is tails is

$$A_{1,T} = \{(T, \omega_2, \ldots, \omega_n) : \omega_i \in \{H, T\}\}.$$

Example 2: Tossing a coin (continued...)

We can define the events $A_{i,H}$ for the i-th flip to be heads, and $A_{i,T}$ for tails. If we assume that the coin is **fair**[1] and each toss is **independent**, then we have that for fixed $\omega_1, \cdots \omega_n$:

$$\begin{aligned} p((\omega_1, \ldots, \omega_n)) &= \Pr\left[A_{1,\omega_1} \cap \ldots \cap A_{n,\omega_n}\right] \\ &= \Pr\left[A_{1,\omega_1}\right] \ldots \Pr\left[A_{n,\omega_n}\right] \quad \texttt{// by independence} \\ &= \frac{1}{2} \cdot \ldots \cdot \frac{1}{2} \quad \texttt{// by fairness} \\ &= \frac{1}{2^n}. \end{aligned}$$

[1] Formally, this means that $p(H) = p(T)$.

A (real-valued) **random variable** is a function $X : \Omega \to \mathbb{R}$. In Example 1, the number of heads is a random variable represented by the function

$$X((\omega_1, \ldots, \omega_n)) = \sum_{i=1}^{n} \omega_i$$

Two discrete real-valued random variables X, Y are said to be **independent** if

$$\Pr\left[X = x, Y = y\right] = \Pr\left[X = x\right] \cdot \Pr\left[Y = y\right],$$

for any $x, y \in \mathbb{R}$.
The random variables $X_1, \ldots, X_n$ are said to be **(jointly) independent** if

$$\Pr\left[X_1 = x_1, \ldots, X_n = x_n\right] = \prod_{i=1}^{n} \Pr\left[X_i = x_i\right]$$

for any $x_1, \ldots, x_n$.

Note: $X_1, \ldots, X_n$ can be pairwise independent without being jointly independent!

Example 3: Jointly-independent random variables

Let X_i be the random variable that is 1 if the i-th fair coin landed heads and 0 otherwise.
That is,

$$X_i = \begin{cases} 1, & \text{if } i\text{th coin lands } H. \\ 0, & \text{if } i\text{th coin lands } T. \end{cases}$$

The random variables $X_1, \ldots, X_n$ are jointly independent.

Example 4: Pairwise-independent but not jointly-independent random variables [1]

Let $\Omega = \{1, 2, 3, 4\}$. Suppose that the distribution $p(i) = \frac{1}{4}$, for $i \in \{1, 2, 3, 4\}$.
As a sanity-check, observe that:

$$\sum_{i \in \Omega} p(i) = 1.$$

Define the events $A = \{1, 2\}$, $B = \{1, 3\}$, and $C = \{2, 3\}$. A, B, C are pairwise-independent but **not jointly-independent**. To see why,

$$\Pr[A \cap B] = \Pr[\{1\}] = p(1) = \frac{1}{4} = \Pr[A] \cdot \Pr[B]$$
$$\Pr[A \cap C] = \Pr[\{2\}] = p(2) = \frac{1}{4} = \Pr[A] \cdot \Pr[C]$$
$$\Pr[B \cap C] = \Pr[\{3\}] = p(3) = \frac{1}{4} = \Pr[B] \cdot \Pr[C]$$

However,

$$\Pr[A \cap B \cap C] = \Pr[\emptyset] = 0 \neq \Pr[A] \cdot \Pr[B] \cdot \Pr[C]$$

[1] Example taken from https://math.stackexchange.com/questions/1783225/ which cites Jacod and Protter [1].

1.2 Law of total probability

The law of total probability states that if we have events $A_1, A_2, \ldots, A_n$ which partition the sample space (i.e., Ω is a disjoint union of these events), and B is any event, then

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B \cap A_i].$$

Note: the law of total probability is also valid if we have a countably infinite partition into events $A_1, A_2, \ldots, A_n, \ldots$, in which case

$$\Pr[B] = \sum_{i=1}^{\infty} \Pr[B \cap A_i].$$

1.3 Conditional probability.

Conditioning on an event means assuming examining what happens given another event occurs. For example, we can condition on the probability of it raining tomorrow *given* that it's raining today.
Formally, the probability of event A **conditioned** on event B is defined as

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

In words, this formula captures the probability that both events happen, subject to B happening. The intuition is that we focus only on the part of our sample space Ω in which B happens (i.e. the subset

$B \subset \Omega$). We can then think of B as our new sample space. Then, the part of A that matters is only the intersection $A \cap B$. However, if we make B our new sample space, then we need to *normalize* by the probability that B happens (in Ω), because B doesn't necessarily have the full probability mass within Ω. This is where the denominator $\Pr[B]$ comes from in the formula.

To see that this is indeed a valid probability, note that

$$\Pr[A|B] + \Pr[\overline{A}|B] = \frac{\Pr[A \cap B] + \Pr[\overline{A} \cap B]}{\Pr[B]} = 1$$

From the above, we get **Bayes' rule**, one of the most important formulas in probability:

$$\Pr[A|B] = \frac{\Pr[B|A] \Pr[A]}{\Pr[B]}.$$

Additionally, using conditional probability, we can obtain a formula for the probability of an intersection of events, even if the events are not independent!

$$\Pr[A_1 \cap A_2 \cdots \cap A_n] = \Pr[A_1] \Pr[A_2|A_1] \ldots \Pr[A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}]$$

The above is often called the "**chain rule**" of conditional probability.

1.4 Union bound.

Consider events $A_1, \ldots, A_n \subseteq \Omega$, where Ω is a sample space. Then we have

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i]$$

The proof is quite natural. We have

$$\begin{aligned}\Pr\left[\bigcup_{i=1}^{n} A_i\right] &= \sum_{\omega \in \cup_{i=1}^{n} A_i} p(\omega) \\ &\leq \sum_{i=1}^{n} \sum_{\omega \in A_i} p(\omega) = \sum_{i=1}^{n} \Pr[A_i].\end{aligned}$$

This technique is commonly used when we want to provide a bound on the probability that at least one event happens, from a family of events. We upper bound this probability by the sum of the probabilities of the individual events. This is tight when all the A_i are disjoint.

1.5 Expectation.

For a discrete real-valued random variable X taking possible values $x_1, \ldots, x_n$, the expectation is defined as

$$\mathbb{E}[X] = \sum_{i=1}^{n} \Pr[X = x_i] x_i.$$

You can think of the expectation as being the "average" value that X will take on.

Example 5: Expectation of a coin toss

Let X be the random variable that is 1 if a fair coin landed heads and 0 otherwise.

$$\mathbb{E}[X] = \left(\frac{1}{2}\right) \cdot 1 + \left(\frac{1}{2}\right) \cdot 0 = \frac{1}{2}.$$

Linearity of Expectation
Given random variables $X_1, ..., X_n$ and $X = \sum_{i=1}^{n} X_i$, we have

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

In words, the expected value of the sum of random variables is equal to the sum of the expected values. A very important takeaway from this result is that it holds even if the random variables are **not** independent. This will be used frequently when we have to find the expected value of a sum of random variables when they might not be independent.

Example 6: Linearity of Expectation for a coin toss

Let X_i be the random variable that is 1 if the i-th fair coin landed heads and 0 otherwise.
Let $X = \sum_{i=1}^{n} X_i$.

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \frac{n}{2}.$$

Which matches our intuition that, on average, the number of times that the coin comes up heads is roughly half the number of trials.

Multiplicativity of expectation under independence
Another cool property of expectation is that the expectation of a product of independent variables is the product of individual expectations:

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right].$$

To see this, it is easiest to start manipulating the right side. Suppose X can take values in S and Y can take values in T, and let $W = \{xy : x \in S, y \in T\}$. Then we have

$$\begin{aligned}
\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] &= \sum_{x\in S}\sum_{y\in T} \Pr\left[X = x\right]\Pr\left[Y = y\right] xy \\
&= \sum_{x\in S}\sum_{y\in T} \Pr\left[X = x, Y = y\right] xy \\
&= \sum_{a\in W}\sum_{(x,y)\in S\times T : xy = a} \Pr\left[X = x, Y = y\right] a \\
&= \sum_{a\in W} \Pr\left[XY = a\right] a = \mathbb{E}\left[XY\right].
\end{aligned}$$

1.6 Variance.

For a discrete real-valued random variable X, the *variance* is defined as

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

Intuitively, the variance captures how far the random variable is from its expectation in a squared, expected sense. Note that this can be alternatively expressed as

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

Linearity of variance under pairwise independence.
An important property of the variance is that it is additive when the summands are pairwise independent random variables. That is, if $X_1, \ldots, X_n$ are pairwise independent random variables, we have

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}\left[X_i\right]$$

To see this, note that

$$\begin{aligned}
\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\sum_{i=1}^{n} \mathbb{E}\left[X_i\right]\right)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] + 2\sum_{i<j} \mathbb{E}\left[X_i X_j\right] - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]^2 - 2\sum_{i<j} \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]^2 \\
&= \sum_{i=1}^{n} \mathbf{Var}\left[X_i\right]
\end{aligned}$$

where we used the fact that $\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$ for independent X, Y.

1.7 Useful approximations

When dealing with binomial coefficients or factorials, it may be useful to use one of the following approximations:

Binomial bound

$$\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$$

Stirling's approximation for factorials

$$n! = \Theta\left(\sqrt{2\pi n} \cdot \frac{n}{e}\right)^n$$

1.8 Entropy and min-entropy

Entropy is a concept that shows up in multiple contexts, including thermodynamics, chemistry, economics, and information theory. We will focus on how entropy is used in information theory and cryptography. In this context, entropy can be thought of as a measure of how "random" or "uncertain" a random variable looks, i.e. the average level of uncertainty or information we have about the outcomes of a random variable or probability distribution.
Suppose $P = (p_1, \ldots, p_n)$ is the probability mass function of a probability distribution. The **entropy** or **Shannon entropy** of P is defined as

$$H(P) = -\sum_i p_i \log p_i.$$

This is the most basic definition of entropy, which characterizes to some extent how much information P contains. If P is concentrated on a single element, i.e. $p_1 = 0$, then it has no entropy, which captures the

intuition that we should already know with certainty the outcome of a sample from P. On the other hand, a fair toin coss (e.g. $P = (\frac{1}{2}, \frac{1}{2})$) has an entropy of 1, which means there is one bit of uncertainty about the outcome (either heads or tails).
One can generalize the notion of Shannon entropy to similar notions of randomness which we will use in this course; one such notion is **min-entropy**. The min-entropy of a distribution P characterizes how difficult it is to guess the most likely output of P and is defined as

$$H_{\min}(P) := -\log p_{\max}.$$

Min-entropy often shows up in cryptography and randomenss extraction. To contrast the two notions of entropy, observe that while the distribution $P = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ has Shannon entropy $1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} = \frac{7}{4}$, it only has min-entropy 1.

2 Concentration inequalities

Concentration inequalities are tools that allow us to bound the probability with which a random variable can be far from its expectation. There is a vast number of concentration inequalities corresponding to the different assumptions on the random variable.
For example, if a random variable is a sum of many independent random variables, intuitively it seems very (exponentially in the number of summands) unlikely for all individual random variables in the sum to conspire to bring the value of the sum away from its expectation. As we'll see below, in such a situation we in fact have theorems saying that deviating from the expectation is exponentially unlikely, as one intuitively expects.

2.1 Markov's inequality.

Let Y be a discrete random variable taking non-negative values in the set S. Then for any $a > 0$,

$$\Pr[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a}$$

A nice feature of this inequality is that it only depends on the expectation of the random variable.
Proof.

$$\mathbb{E}[Y] = \sum_{y \in S} y \Pr[Y = y] = \sum_{y \in S, y < a} y \Pr[Y = y] + \sum_{y \in S, y \geq a} y \Pr[Y = y]$$

$$\geq \sum_{y \in S, y \geq a} y \Pr[Y = y] \geq \sum_{y \in S, y \geq a} a \Pr[Y = y] = a \Pr[Y \geq a] \qquad \square$$

This is tight when Y is a with probability 1. Markov's inequality is important because it ties the probability of a random variable being greater than some threshold to the expected value of the random variable. What's not obvious though is that it can also be extended to prove much more powerful inequalities.

2.2 Chebyshev's inequality.

Let X be a random variable with expected value μ and strictly positive variance σ^2. Then for all real $k > 0$:

$$\Pr[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$$

What this is saying is that the probability that X is a distance from the mean is related directly to the variance and inversely to the squared distance. In general Chebyshev's inequality provides us with a stronger bound than Markov's inequality because we utilize the variance of the random variable.
Proof. Since $(X - \mu)^2$ is a nonnegative random variable, by Markov's inequality we get

$$\Pr[(X - \mu)^2 \geq k^2] \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2}$$

$$\Pr[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2} \qquad \square$$

2.3 Chernoff bounds.

Suppose $X_1, \ldots, X_n$ are independent random variables taking values in $\{0, 1\}$. Let X denote their sum and let $\mu = \mathbb{E}[X]$ denote the sum's expected value. Then we have the following inequalities:

- $\Pr[X > (1 + \beta)\mu] < \left(\frac{e^\beta}{(1+\beta)^{1+\beta}}\right)^\mu$, for $\beta > 0$

- $\Pr[X < (1-\beta)\mu] < \left(\frac{e^{-\beta}}{(1-\beta)^{1-\beta}}\right)^{\mu}$, for $0 < \beta < 1$

Sometimes, it suffices to use the following looser but often more convenient bounds:

- $\Pr[X > (1+\beta)\mu] < e^{-\beta^2\mu/3}$, for $0 < \beta < 1$
- $\Pr[X > (1+\beta)\mu] < e^{-\beta\mu/3}$, for $\beta > 1$
- $\Pr[X < (1-\beta)\mu] < e^{-\beta^2\mu/2}$, for $0 < \beta < 1$

Chernoff bounds allow us to show even tighter concentration than Chebyshev or Markov bounds because we can use the fact that the random variables exhibit full mutual independence. Note that this is a stronger assumption than pairwise independence! There are groups of random variables which are all pairwise independent but which are *not* mutually independent.

3 Example problems

3.1 Problems pertaining to random variables.

1. For each of the following distributions, compute their expectation and variance:
   (1) Uniform in $\{1, 2, \ldots, n\}$,
   (2) Bernoulli[1] with success probability p.

   Solution.

   (1) We have

$$\mathbb{E}[X] = \sum_{i=1}^{n} \frac{1}{n} i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

   and

$$\begin{aligned} \mathbf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 &= \sum_{i=1}^{n} \frac{1}{n} i^2 - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{1}{n} \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12} \end{aligned}$$

   (2)

$$\mathbb{E}[X] = p \cdot 1 + (1-p) \cdot 0 = p$$

$$\mathbf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p \cdot 1^2 + (1-p) \cdot 0^2 - p^2 = p(1-p)$$

2. Suppose Alice flips 6 fair coins. What is the probability that result is three heads and three tails? Suppose furthermore that Alice has to pay \$1 to flip 6 coins. What is the expected number of dollars she must pay until she sees the result of three heads and three tails?

   Solution. The probability space can be represented as $\Omega = \{(a_1, \ldots, a_6) : a_i \in \{0, 1\}\}$. The event of getting three heads is then $A = \{(a_1, \ldots, a_6) : \sum_{i=1}^{6} a_i = 3\}$. Since every possible choice $(a_1, \ldots, a_6)$ has the same probability $\frac{1}{2^6}$, we have

$$\Pr[A] = \frac{|A|}{2^6} = \frac{\binom{6}{3}}{2^6} = \frac{5}{16}$$

   For the second part, we're in the following general situation: we have a Bernoulli (i.e., $\{0, 1\}$) random variable X such that $\Pr[X = 1] = p$, and we sample independent copies $X_1, X_2, \ldots$ of X. We want to know what is the expected time $\mathbb{E}[T]$ such that $X_T = 1$ for the first time. Well, we have

$$\Pr[T = t] = \Pr[X_1 = 0, \ldots, X_{t-1} = 0, X_t = 1] = (1-p)^{t-1} p$$

[1] The Bernoulli distribution is defined by the random variable X where $X = 1$ with probability p and $X = 0$ with probability $1 - p$.

and hence

$$\begin{aligned}\mathbb{E}[T] &= \sum_{t=1}^{\infty} \Pr[T=t]\, t = \sum_{t=1}^{\infty} (1-p)^{t-1} pt \\ &= p \sum_{t=1}^{\infty} (1-p)^{t-1} t \\ &= p \left(\sum_{t=1}^{\infty} (1-p)^{t-1} + \sum_{t=2}^{\infty} (1-p)^{t-1} + \dots \right) \\ &= p \left(\frac{1}{p} + (1-p)\frac{1}{p} + (1-p)^2 \frac{1}{p} + \dots \right) \\ &= 1 + (1-p) + (1-p)^2 + \dots = \frac{1}{p}.\end{aligned}$$

So, we get a very neat result: the expected number of independent trials until a Bernoulli random variable with probability of being 1 equal to p is 1 is $\frac{1}{p}$.

Applying this to our case, the expected number of dollars will be $\frac{16}{5}$.

This calculation can be simplified using the following identity which holds whenever T ranges over the natural numbers:

$$\mathbb{E}[T] = \sum_{t=0}^{\infty} \Pr[T > t]$$

3. Alice flips a fair coin n times, and so does Bob. Show that the probability that they get the same number of heads is $\binom{2n}{n}/4^n$. Use your argument to verify the identity

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$$

Solution. Let our probability space be $\Omega = \{(a_1, \dots, a_n, b_1, \dots, b_n) : a_i \in \{0,1\}, b_i \in \{0,1\}\}$, where $a_i = 1$ if the i-th flip of Alice was heads and 0 otherwise, and $b_i = 1$ if the i-th flip of Bob was *tails*, and 0 otherwise. Note that we encode heads and tails in opposite ways for Alice and Bob.

Then note that the event that they flipped the same number of heads is

$$\begin{aligned}A &= \left\{ (a_1, \dots, a_n, b_1, \dots, b_n) : \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} (1 - b_i) \right\} \\ &= \left\{ (a_1, \dots, a_n, b_1, \dots, b_n) : \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i = n \right\}\end{aligned}$$

which immediately tells us that $\Pr[A] = \frac{\binom{2n}{n}}{2^{2n}}$ as wanted.

Now, note that we could have computed the same probability with a different probability space: namely, the one where we encode heads and tails in the same way. Here $\Omega = \{(a_1, \dots, a_n, b_1, \dots, b_n) : a_i \in \{0,1\}, b_i \in \{0,1\}\}$, where $a_i = 1$ if the i-th flip of Alice was heads and 0 otherwise, and $b_i = 1$ if the i-th flip of Bob was heads, and 0 otherwise. Now we have

$$A = \left\{ (a_1, \dots, a_n, b_1, \dots, b_n) : \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i \right\}$$

We can calculate the probability by considering all the different possible numbers of heads that the two players can have (we're using the law of total probability here):

$$\begin{aligned}
\Pr[A] &= \sum_{k=0}^{n} \Pr\left[A \cap \sum_{i=1}^{n} a_i = k\right] \\
&= \sum_{k=0}^{n} \Pr\left[\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = k\right] \\
&= \sum_{k=0}^{n} \Pr\left[\sum_{i=1}^{n} a_i = k\right] \Pr\left[\sum_{i=1}^{n} b_i = k\right] \\
&= \sum_{k=0}^{n} \frac{\binom{n}{k}}{2^n} \frac{\binom{n}{k}}{2^n} \\
&= \frac{\sum_{k=0}^{n} \binom{n}{k}^2}{4^n}.
\end{aligned}$$

Comparing the two expressions, we get the desired identity.

3.2 Problems pertaining to concentration bounds

1. Let's say that we flip a biased coin that lands heads with probability $\frac{1}{3}$ a total of n times. Use Chernoff bounds to determine a value of n such that the probability of getting more than half of the flips heads is less than $\frac{1}{1000}$.

   Solution. Let X_i be a random variable that is 1 if the i-th flip landed heads and 0 otherwise. If we denote $X = \sum_{i=1}^{n} X_i$, we want to find the smallest n such that $\Pr[X > \frac{n}{2}] < \frac{1}{1000}$.

   Note that $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{3} = \frac{n}{3}$. Applying Chernoff bounds from the previous section with $\beta = \frac{1}{2}$ we get

$$\begin{aligned}
&\Pr[X > \frac{3}{2}\mu] < e^{-(1/2)^2 \mu/3} \\
\Leftrightarrow &\Pr[X > \frac{n}{2}] < e^{-n/36}
\end{aligned}$$

   So for $e^{-n/36} < 1/1000 \Leftrightarrow n > 36 \log 1000 \approx 250$ we have the required bound.

2. Bar the bear decides he wants to manage beehives in his old age. He's just received k bees that he wants to allocate to his n beehives. Since Bar is old, he often loses count when trying to allocate the bees to beehives. He decides to just allocate the bees randomly to his hives. That is, for each bee, he chooses a beehive uniformly at random. Help Bar prove that his strategy yields an approximately uniform distribution of bees with high probability.

   (a) Let X_i be the number of bees in the i-th beehive. Compute $E[X_i]$.

   Solution. Let Y_{ji} be 1 if the j-th bee is allocated to the i-th beehive, and 0 otherwise. We have $E[Y_{ji}] = \Pr[j\text{-th bee is put into } i\text{-th beehive}] = 1/n$. Then $X_i = \sum_{j=1}^{k} Y_{ji}$, so $E[X_i] = \sum_{j=1}^{k} E[Y_{ji}] = \sum_{j=1}^{k} 1/n = k/n$.

   (b) Show that X_i and X_j are not independent.

   Solution. We see that $\Pr[X_i = k \cap X_j = k] = 0$. However, $\Pr[X_i = k] \Pr[X_j = k] = (1/n)^{2k}$. Thus, X_i and X_j are not independent.

(c) Let $M = \max(X_1, X_2, \ldots, X_n)$. Show $\Pr[M \geq 2k/n] \leq ne^{-k/(3n)}$.

Solution. The idea is to use Chernoff bounds to show that $\Pr[X_i \geq 2k/n]$ is small and then use the union bound to bound the probability that any of the X_i variables is greater than $2k/n$. Recall that $X_i = \sum_{j=1}^{k} Y_{ji}$. We have $\Pr[X_i \geq (1+\delta)E[X_i]] \leq e^{-\delta^2 E[X_i]/3}$ by Chernoff. Thus, we get $\Pr[X_i \geq 2k/n] \leq e^{-k/(3n)}$, and by union bound $\Pr[M \geq 2k/n] \leq \sum_{i=1}^{n} \Pr[X_i \geq 2k/n] \leq \sum_{i=1}^{n} e^{-k/(3n)} = ne^{-k/(3n)}$.

3.3 The Coupon Collector problem.

Suppose there are n different kinds of coupons, and we want to collect at least one coupon from every kind. We start out with nothing, and at each step, we get a new random coupon, equally likely to be any of the n kinds, and independent of the previous coupons. This is known as the *coupon collector's problem.*

- What is the expected time T when we're done collecting?
- What is the variance of T?
- Use Chebyshev's inequality to bound the probability that T deviates far from its expectation.

Solution. Let T_i be the random variable equal to the first time we have i different kinds of coupons. Then, we can break the total time to collect all kinds of coupons T_n into the phases between getting a new kind of coupon:

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}[T_1 + (T_2 - T_1) + \ldots + (T_n - T_{n-1})] \\ &= \mathbb{E}[T_1] + \mathbb{E}[T_2 - T_1] + \ldots + \mathbb{E}[T_n - T_{n-1}].\end{aligned}$$

Now let's think about the random variable $T_{k+1} - T_k$: it is the time it takes us to get a $k+1$-th coupon given that we already have k coupons. No matter what kinds of coupons we have already, the probability that we get a new coupon is $\frac{n-k}{n}$ in each step independently.
This is identical to the earlier problem where we had a Bernoulli random variable X such that $\Pr[X = 1] = p$, and we showed that the expected time until it becomes 1 for the first time is $\frac{1}{p}$. Thus, $\mathbb{E}[T_{k+1} - T_k] = \frac{n}{n-k}$, and

$$\begin{aligned}\mathbb{E}[T_n] &= 1 + \frac{n}{n-1} + \ldots + \frac{n}{1} \\ &= n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{1}\right) = nH_n\end{aligned}$$

where H_n is the n-th harmonic number. It is known that $H_n = \Theta(\log n)$ (which can be proved using an integral among other methods), hence $\mathbb{E}[T] = \Theta(n \log n)$.
For the variance, note that the random variables $T_{k+1} - T_k$ are independent. Indeed, if $k > l$, we have

$$\Pr[T_{k+1} - T_k = t_k, T_{l+1} - T_l = t_l] = \Pr\left[T_{k+1} - T_k = t_k \;\middle|\; T_{l+1} - T_l = t_l\right] \Pr[T_{l+1} - T_l = t_l]$$

Now, note that conditioning on $T_{l+1} - T_l = t_l$ has no effect on the probability that $T_{k+1} - T_k = t_k$, since the future coupons we get are independent of the past. Hence the above is

$$= \Pr[T_{k+1} - T_k = t_k] \Pr[T_{l+1} - T_l = t_l]$$

which shows that the random variables are indeed independent. This means that

$$\begin{aligned}\mathbf{Var}[T_n] &= \mathbf{Var}[T_1 + (T_2 - T_1) + \ldots + (T_n - T_{n-1})] \\ &= \mathbf{Var}[T_1] + \mathbf{Var}[T_2 - T_1] + \ldots + \mathbf{Var}[T_n - T_{n-1}]\end{aligned}$$

Now we're faced with the general task of computing the variance of the random variable T which is the first time that a Bernoulli random variable X with $\Pr[X = 1] = p$ becomes 1. We have

$$\Pr[T = t] = (1-p)^{t-1}p$$

and as we saw earlier, $\mathbb{E}[T] = \frac{1}{p}$. It remains to compute

$$\begin{aligned}\mathbb{E}\left[T^2\right] &= \sum_{t=1}^{\infty} \Pr[T = t]\, t^2 \\ &= \sum_{t=1}^{\infty} (1-p)^{t-1} p t^2 \\ &= p \sum_{t=1}^{\infty} (1-p)^{t-1} t^2\end{aligned}$$

We could compute this sum by decomposing it into simpler sums in a clever way. But here's a useful (and more principled) trick for computing sums like this: consider the function $f(x) = \frac{1}{1-x}$ for $|x| < 1$. Then we have the power series expansion

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots = \sum_{n=0}^{\infty} x^n$$

Differentiating both sides, we have

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + \ldots = \sum_{t=0}^{\infty} (t+1) x^t$$

and differentiating again,

$$\frac{2}{(1-x)^3} = 2 + 6x + 12x^2 + \ldots = \sum_{t=0}^{\infty} (t+1)(t+2) x^t$$

Using this, we have

$$\begin{aligned}\sum_{t=1}^{\infty} (1-p)^{t-1} t^2 &= \sum_{t=1}^{\infty} (1-p)^{t-1} t(t+1) - \sum_{t=1}^{\infty} (1-p)^{t-1} t \\ &= \sum_{t=0}^{\infty} (1-p)^{t} (t+1)(t+2) - \sum_{t=0}^{\infty} (1-p)^{t} (t+1) \\ &= \frac{2}{p^3} - \frac{1}{p^2}\end{aligned}$$

and so

$$\begin{aligned}\mathbf{Var}[T] &= \mathbb{E}\left[T^2\right] - \mathbb{E}[T]^2 \\ &= \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}\end{aligned}$$

which implies that

$$\mathbf{Var}[T_n] = \sum_{k=1}^{n} \frac{1 - \frac{n-k}{n}}{\left(\frac{n-k}{n}\right)^2} = \sum_{k=1}^{n} \frac{nk}{(n-k)^2} \leq n^2 \sum_{l=1}^{\infty} \frac{1}{l^2} \leq 2n^2.$$

Thus, by Chebyshev,

$$\Pr[|T_n - \mathbb{E}[T_n]| \geq cn] \leq \frac{2}{c^2}.$$

References

[1] Jean Jacod and Philip Protter. *Probability essentials*. Springer Science & Business Media, 2012.