Brief Treatise On Discrete Probability

John-Michael Laurel [1] jm-laurel@berkeley.edu

(1) Random Variables and Distributions

A random variable (r.v.) X, formally defined as

$$X : \Omega \mapsto \mathbb{R}$$

where Ω is the *outcome space*, is said to follow a *named distribution* if for all $x \in X$ we have

$$\sum_x \mathbb{P}(X = x) = 1$$

and $\mathbb{P}(X = x)$ is a well constructed *probability mass function* (p.m.f.). We write

$$X \sim \textsf{namedDistribution}(\,\cdot\,)$$

to say that X is distributed "**namedDistribution**" where " $\cdot$ " is the distribution's parameter(s). For a discrete r.v., it's understood that the range of X is countable.

(2) Measures of Distributions

When a random variable (r.v.) follows a named distribution with known parameters, we have the ability to extract insight into that distribution. In particular, its *expectation* and *standard deviation*; both common statistical *measures*.

(2.1) Expectation

The expectation (or mean) of X is a sum of the values it takes, weighted by their probabilities

$$\mathbb{E}(X) = \sum_x x\,\mathbb{P}(X = x)$$

, we interpret expectation as the long-run average of an experiment.

[1] University of California, Berkeley (July 2019)
This document assumes understanding of basic probabilistic knowledge such as partitioning, independence, conditioning, Bayes' rule, joint distributions, et cetera, combinatorics, and set theory. Special thanks to Ella Hiesmayr for providing feedback on this document.

(2.1.1) Properties of Expectation

Expectation is linear. Explicitly for $\alpha_i, \beta_i \in \mathbb{R}$,

$$\mathbb{E}\left[\sum_{i=1}^{m}(\alpha_i X_i + \beta_i)\right] = \sum_{i=1}^{m} \alpha_i\,\mathbb{E}(X_i) + \sum_{i=1}^{m}\beta_i$$

whether the X_i's are independent or otherwise. If the X_i's are independent, then the following holds

$$\mathbb{E}\left(\prod_{i=1}^{m}\alpha_i X_i\right) = \prod_{i=1}^{m}\alpha_i\,\mathbb{E}(X_i)$$

One can easily substitute the X in $\mathbb{E}(X)$ in §(2.1) for a function of X_i's and the above properties sustain. For clarity, $\mathbb{E}\big[g(X_1, \ldots, X_m)\big]$ is precisely

$$\sum_{x_1,\ldots,x_m} g(X_1, \ldots, X_m)\mathbb{P}(X_1 = x_1, \ldots, X_m = x_m)$$

(2.1.2) Tail Sum Formula

Another method of computing expectation; by considering the tail probabilities for $X \geq 0$. The following derivation uses indicators, see §(3.2.1) for reference. Suppose $X \in \{0, 1, \ldots, n\}$ is a count, then

$$X = \sum_{j=1}^{n} \mathbb{1}_{A_j}$$

where A_j is the event $X \geq j$. Applying expectation,

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{j=1}^{n}\mathbb{1}_{A_j}\right) = \sum_{j=1}^{n}\mathbb{E}(\mathbb{1}_{A_j})$$

this quickly leads to the *Tail Sum Formula*

$$\mathbb{E}(X) = \sum_{j=1}^{n}\mathbb{P}(X \geq j)$$

□

Tail sum formula employs itself when computing $\mathbb{E}(Y)$ for $Y \sim \textsf{Geometric}(p)$, see §(3.5).

(2.2) Variance and Standard Deviation

The *variance* (or spread) denoted σ^2 is the average squared difference of X from its mean

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

and has computational form

$$\sigma^2(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$

The *standard deviation* denoted σ is related to variance in the following way

$$\sigma(X) = \sqrt{\sigma^2(X)}$$

(2.2.1) Properties of Variance

If $\{X_i\}_{1\leq i\leq m}$ are independent r.v.'s and $\alpha_i, \beta_i \in \mathbb{R}$, then

$$\sigma^2\left[\sum_{i=1}^{m}(\alpha_i X_i + \beta_i)\right] = \sum_{i=1}^{m}\alpha_i^2\,\sigma^2(X_i)$$

(2.2.2) Properties of Standard Deviation

Given it's close relationship to variance, standard deviation is also in-variate to shifting. Consider the same setting as in §(2.2.1), without loss of generality and for $i = 1$, we have

$$\sigma(\alpha_1 X_1 + \beta_1) = |\alpha_1|\sigma(X_1)$$

The above cements the fact that standard deviation is the square root of variance, see §(2.2).

(3) Named Distributions

This section highlights discrete distributions; their properties and relationships to one another.

(3.1) Uniform Distribution

If $X \sim \textsf{Uniform}\big(\{a, a+1, \dots, b\}\big)$

$$\mathbb{P}(X = x) = \frac{1}{n}, \quad \forall x \in X$$

That is, the chance of getting any x is the same. $\mathbb{E}(X) = \frac{a+b}{2}$, $\sigma^2(X) = \frac{(b-a+1)^2-1}{12}$ e.g. rolling a fair n-sided

(3.2) Bernoulli Distribution

If $X \sim \textsf{Bernoulli}(p)$, then X is defined on $\{1, 0\}$ (success or failure) with probability p and $1-p$ respectively. $\mathbb{E}(X) = p$ and $\sigma^2(X) = p(1-p)$. e.g. flipping a p coin

(3.2.1) Indicator Random Variables

An *indicator* r.v. for event A is defined as

$$\mathbb{1}_A = \begin{cases} 1, & \text{if event } A \text{ happens} \\ 0, & \text{otherwise} \end{cases}$$

and notably $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$. This special application of Bernoulli, where $p = \mathbb{P}(A)$ proves quite powerful when considering counts of events.

(3.3) Binomial Distribution

A generalization of one Bernoulli trial; in particular the sum of n *independent and identically distributed* (i.i.d.) $\textsf{Bernoulli}(p)$ r.v.'s. That is to say if [2] $\{Y_j\}_{1\leq j\leq n} \sim \textsf{Bernoulli}(p)$, then

$$X = \sum_{j=1}^{n} Y_j \sim \textsf{Binomial}(n, p)$$

and for the event $X = k$ successes

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

$\mathbb{E}(X) = np$ and $\sigma^2(X) = np(1-p)$ e.g. a sequence of independent p coin flips

(3.3.1) Mode of Binomial

The *mode* of a distribution is the value(s) with highest probability. The histogram of Binomial is strictly increasing before reaching a maximum and strictly decreasing thereafter. If $k = \lfloor np + p \rfloor$, then the mode of Binomial is defined to be

$$\text{mode} = \begin{cases} k & \text{for } np + p \notin \{0, 1, \dots\} \\ k-1, k & \text{for } np + p \in \{1, 2, \dots\} \end{cases}$$

[2] notation is equivalent to $Y_j \sim \textsf{Bernoulli}(p)$ for $1 \leq j \leq n$, which is equivalent to $Y_1, \dots, Y_n \sim \textsf{Bernoulli}(p)$

(3.4) Multinomial Distribution

A generalization of **Binomial**(n, p), where instead of two categories (success or failure), we have k categories. Define $X_i = n_i$ to be the number of occasions of category i, $1 \leq i \leq k$, where $\sum_{i=1}^{k} n_i = N$ and $\mathbb{P}(X_i = n_i) = p_i$. We are then interested in the joint p.m.f.

$$\mathbb{P}\left(\bigcap_{i=1}^{k} X_i = n_i\right) = \binom{N}{n_1, \ldots, n_k} \prod_{i=1}^{k} p_i^{n_i}$$

Just as in Binomial, the probabilities of getting an element from each of the k categories sum to unity. i.e. $\sum_{i=1}^{k} p_i = 1$. To say the joint distribution of the X_i's is distributed multinomial, we write

$$(X_1, \ldots, X_n) \sim \textsf{Multinomial}(N, \vec{p})$$

Where $\vec{p}$ is a probability vector[3]. It can be easily shown that the marginal distribution of any $X_i \sim$ **Binomial**(N, p_i), reinforcing our intuition. e.g. Finding the probability of getting 1 A, 3 B's, 5 C's, 15 D's, and 10 F's from a class, where

Letter Grade Count

grade	A	B	C	D	F
frequency	15	22	10	32	21

(3.5) Geometric Distribution

Another extension of Bernoulli, where we yield success on the k^{th} trial, implying exactly $k-1$ failures before that. For $X \sim$ **Geometric**(p), we have

$$\mathbb{P}(X = k) = (1-p)^{k-1} p$$

One could also ask the chance that success happens after k trials, which is logically equivalent to not succeeding in the first k runs

$$\mathbb{P}(X > k) = (1-p)^{k}$$

We interpret the geometric distribution as describing the number of trials up to and including the first success. $\mathbb{E}(X) = \frac{1}{p}$ and $\sigma^2(X) = \frac{1-p}{p^2}$ e.g. tossing a p coin until you get a head

[3] a *probability vector* is one whose entries sum to unity

(3.6) Negative Binomial Distribution

A generalization of geometric, where instead we wait for the r^{th} success. Naturally, if the r^{th} success happens on the k^{th} trial, this implies that in the first $k-1$ trials we've had exactly $r-1$ successes. For $T_r \sim$ **NegativeBinomial**(r, p), where T_r denotes the number of trials until the r^{th} success, we have

$$\mathbb{P}(T_r = k) = \binom{k-1}{r-1} p^{r-1} (1-p)^{k-r} \times p$$

$\mathbb{E}(T_r) = \frac{r}{p}$ and $\sigma^2(X) = \frac{r(1-p)}{p^2}$ e.g. tossing a p coin until you get the r^{th} head

(3.7) Hypergeometric Distribution

The analog to Binomial where trials are dependent. Suppose you have a population N with G good elements and B bad elements. You collect a sample of $n \leq N$ elements without replacement and wish to know the chance of getting g good elements. For $X \sim$ **Hypergeometric**(n, N, G),

$$\mathbb{P}(X = g) = \frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$$

$\mathbb{E}(X) = n\left(\frac{G}{N}\right)$, $\sigma^2(X) = n\left(\frac{G}{N}\right)\left(\frac{B}{N}\right)\left(\frac{N-n}{n-1}\right)$ e.g. chance of getting 3 aces in a hand of 13 cards dealt from a standard deck

(3.8) Poisson Distribution

A limit of Binomial where $np \to \mu$ as $n \to \infty$ and $p \to 0$. In words, we have many trials and the event of success is rare. Via consecutive probability ratios and for $X \sim$ **Poisson**(μ), we derive

$$\mathbb{P}(X = k) = \mathbb{P}(0) \prod_{i=1}^{k} R(i) = e^{-\mu} \frac{\mu^k}{k!}$$

, where $R(i) = \frac{\mathbb{P}(i)}{\mathbb{P}(i-1)}$. Poisson may be a limit of Binomial, but nonetheless is a distribution in its own right. $\mathbb{E}(X) = \sigma^2(X) = \mu$, (intuitively) this makes sense when you consider $np(1-p)$ as $p \to 0$. If $\{X_j\}_{1 \leq j \leq n} \sim$ **Poisson**(μ_j), then $\sum_{j=1}^{n} X_j \sim$ **Poisson**$(\sum_{j=1}^{n} \mu_j)$ e.g. Twitter notifications

(4) Normal Approximation to Binomial

The continuous analog to Binomial is the Normal distribution. We approximate Binomial by Normal. Suppose $X \sim \textbf{Binomial}(n,p)$ and we are interested in $\mathbb{P}(a \le X \le b)$. We first standardize X by performing a linear change in scale, in particular

$$X^* = \frac{X - \mu_X}{\sigma_X}$$

we then approximate $\mathbb{P}(a \le X \le b)$ by

$$\Phi\left(\frac{b + \frac{1}{2} - \mu_X}{\sigma_X}\right) - \Phi\left(\frac{a - \frac{1}{2} - \mu_X}{\sigma_X}\right)$$

where $\pm\frac{1}{2}$ are continuity corrections (since $X \in \mathbb{N}$) and $\Phi(z) = \int_{-\infty}^{z} f_X(x)\,dx$ [4]. Use approximation when $n \ge 20$, $\sigma_X > 3$, and p not close to 0 or 1.

(5) Central Limit Theorem (c.l.t)

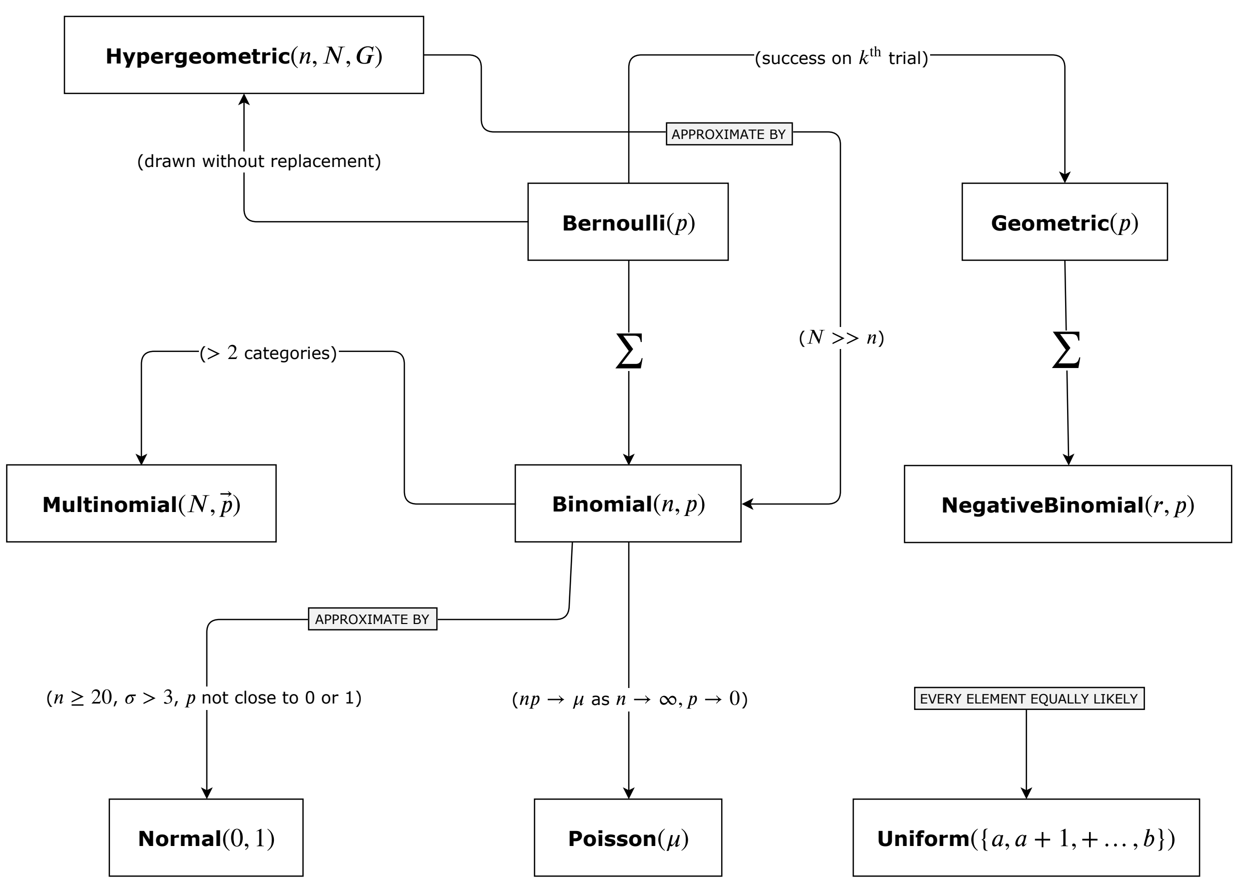
A powerful theorem. Let $\{X_j\}_{1 \le j \le n}$ be i.i.d. with mean μ_X and standard deviation σ_X for $1 \le j \le n$. Define $S_n = \sum_{j=1}^{n} X_j$. Then

$$\mathbb{P}\left(a < \frac{S_n - n\mu_X}{\sqrt{n}\ \sigma_X} \le b\right) \approx \Phi(b^*) - \Phi(a^*)$$

as $n \to \infty$. This statement holds regardless of what the X_j's are distributed. By rule of thumb, apply C.L.T when $n \ge 25$ and $\mathbb{E}(S_n) \pm 3\sigma(S_n) \in$ {possible values of S_n}

(6) Distributions and Their Relationships

This section provides a diagrammatic overview of the distributions showcased in §(3)–§(4) and their relationships to one another. See schematic below.



[4]here $f_X(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$

(7) Bounds on Tail Probabilities

We have the ability to bound probabilities at a distributions tail. The two methods we show are Markov's Inequality and Chebyshev's Inequality.

(7.1) Markov's Inequality

For $X \geq 0$ and constant α, an upper bound on the probability that X is at least α is

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

this bound is interesting when $\mathbb{E}(X) < \alpha$.

Proof. Suppose $Z = \alpha \mathbb{1}_{X \geq \alpha}$. It's clear $\mathbb{E}(Z) \leq \mathbb{E}(X)$. Then,

$$\mathbb{E}(Z) = \alpha \mathbb{P}(X \geq \alpha)$$

which implies,

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}(X)}{\alpha}$$

□

(7.2) Chebyshev's Inequality

Given X, its expectation μ_X, and standard deviation σ_X, the probability that X lies beyond k standard deviations from its mean is bounded by

$$\mathbb{P}(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

equivalently for $k\sigma_X = \alpha$,

$$\mathbb{P}(|X - \mu_X| \geq \alpha) = \mathbb{P}\left[(X - \mu_X)^2 \geq \alpha^2\right]$$

$$\leq \frac{\sigma_X^2}{\alpha^2}$$

In general, Chebyshev gives a tighter upper bound compared to Markov.

Proof. Let $X = (Y - \mu_Y)^2$, then

$$\mathbb{P}(|Y - \mu_Y| \geq \alpha^2) = \mathbb{P}(X \geq \alpha)$$

$$\leq \frac{\mathbb{E}(X)}{\alpha^2} = \frac{\sigma_X^2}{\alpha^2}$$

□

(8) (Strong) Law of Large Numbers

Let $\{X_j\}_{\leq j \leq n}$ be independent trials with finite mean $\mathbb{E}(X_j) = \mu$ and variance $\sigma^2(X_j) = \sigma^2$, $1 \leq j \leq n$, then

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\frac{1}{n}\sum_{j=1}^{n} X_j - \mu\right| < \epsilon\right) = 1$$

for small $\epsilon > 0$. In words, the *sample average* converges to the true mean of X_j, with probability one. For example, suppose $X_j \sim$ **Bernoulli**(p) for $1 \leq j \leq n$, then

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to p$$

as $n \to \infty$.

(9) Craps Principle

Suppose that in a two person game the probability of person A winning is p_A, probability of person B winning is p_B, and the probability of a draw is p_D, hence $p_A + p_B + p_D = 1$. The probability of person A winning eventually is a proportion of the relative frequencies of person A winning and person A and B winning.

$$\mathbb{P}(A \text{ wins eventually}) = \frac{p_A}{p_A + p_B}$$

Proof. The probability that person A wins eventually can be partitioned into person A winning in the first game, a draw in the first game, then person A winning, two draws, then person A winning, and so on. Hence, $\mathbb{P}(A \text{ wins eventually})$ is

$$p_A + p_D p_A + p_D^2 p_A + \ldots$$

$$= p_A(1 + p_D + p_D^2 + \ldots)$$

$$= \frac{p_A}{1 - p_D}$$

$$= \frac{p_A}{p_A + p_B}$$

□

(10) Techniques and Important Problems

This section and its succeeding ones are meant to be a compendium of interesting (discrete probability) problems. It is broken into two parts: methods and problems. Methods will ensue for the remainder of this section and capture important techniques to consider when solving problems, along with sprinkling some mathematical tools that pop up in probability. For part two on problems, it shall be broken in this manner: problem statement, problem solution, and problem remarks.

(10.1) Counting

Knowing how to count is important. The *Fundamental Principle of Counting* is where we start. Suppose that in a k stage experiment, regardless of the first $i-1$ stages $2 \leq i \leq k$, there is exactly n_i choices for the i^{th} stage. Then the number of possible sequences (or paths) one can take is

$$\prod_{i=1}^{k} n_i$$

Suppose $\mathcal{S} = \{1, 2, \ldots, n\}$, the number of subsets (including $\mathcal{S}$ and $\varnothing$) in $\mathcal{S}$ is clearly 2^n. This follows immediately from the product above. Another conceptualization of the quantity 2^n is revealed in the following equality

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

, where $\binom{n}{k}$ is the *binomial coefficient.* It is the number of ways of selecting a subset of size $k \leq n$ from n. We say $\binom{n}{k}$ gives the number of *combinations* of choosing k elements from a set of size n. It possesses symmetric properties and is defined below

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{(n-k)!k!}$$

This gives rise to another counting construction; a *permutation*, which is the number of distinct orderings of k elements from a set of n. Its intimate connection to combinations is revealed below [5]

$$(n)_k = \prod_{i=0}^{k-1}(n-i) = \binom{n}{k}k!$$

Logically, count all combinations of choosing k elements from a set of n elements and there are $k!$ ways of ordering the k length sequence.

[5] $(n)_k$ is read "n order k."

(10.2) Method of Indicators

The method of indicators makes use of the fact (sometimes referred to as the *fundamental bridge*):

$$\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$$

(10.2.1) Variance of a Sum of Dependent Indicators

Let $\{\mathbb{1}_{A_j}\}_{1 \leq j \leq n}$ be exchangeable indicators for event A_j and $S_n = \sum_{j=1}^{n} \mathbb{1}_{A_j}$. Expectation of S_n works as expected,

$$\mathbb{E}(S_n) = \mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}_{A_j}\right) = \sum_{j=1}^{n} \mathbb{E}(\mathbb{1}_{A_j}) = n\mathbb{E}(\mathbb{1}_{A_j})$$

by the fundamental bridge,

$$\mathbb{E}(S_n) = n\mathbb{P}(A_j)$$

Recall $\sigma^2(S_n) = \mathbb{E}(S_n^2) - \mathbb{E}^2(S_n)$. Focusing on the first term,

$$\begin{aligned}\mathbb{E}(S_n^2) &= \mathbb{E}\left[\left(\sum_{j=1}^{n} \mathbb{1}_{A_j}\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{j=1}^{n} \mathbb{1}_{A_j}\right)\left(\sum_{j=1}^{n} \mathbb{1}_{A_j}\right)\right]\end{aligned}$$

, let's pause and dissect the product on the inside of the expectation

$$(\mathbb{1}_{A_1} + \mathbb{1}_{A_2} + \ldots + \mathbb{1}_{A_n})(\mathbb{1}_{A_1} + \mathbb{1}_{A_2} + \ldots + \mathbb{1}_{A_n})$$

There are two types of terms in this expansion

1. like terms (on diagonal): $\mathbb{1}_{A_i}\mathbb{1}_{A_j}$ where $i = j$
2. cross terms (off diagonal): $\mathbb{1}_{A_i}\mathbb{1}_{A_j}$ where $i \neq j$

For the like terms,

$$\mathbb{1}_{A_i}\mathbb{1}_{A_i} = \mathbb{1}_{A_i}^2 = \mathbb{1}_{A_i}$$

and there are clearly n of them. Consider the cross terms

$$\mathbb{1}_{A_i}\mathbb{1}_{A_j}$$

, where $i \neq j$. Their product will be 1 if and only if $\mathbb{1}_{A_i} = 1$ and $\mathbb{1}_{A_j} = 1$. Crucially, this means that the product of these indicators is itself an indicator of the two events A_i and A_j, that is

$$\mathbb{1}_{A_i}\mathbb{1}_{A_j} = \mathbb{1}_{A_i \cap A_j}$$

The product's expansion will have n^2 terms in total, hence the number of cross terms is $n^2 - n$. Another way to see that we have $n(n-1)$ cross terms is counting the number of available indices such that $i \neq j$ for $1 \leq i, j \leq n$. We have n choices for the first index and $(n-1)$ choices for the next. Equivalently we have $\binom{n}{2}$ ways of selecting the indices and 2 ways to order (i, j), $i \neq j$ for $1 \leq i, j \leq n$. Now we realize $\mathbb{E}(S_n^2)$ as

$$\mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}_{A_j} + \sum_{i \neq j} \mathbb{1}_{A_i \cap A_j}\right)$$

by linearity and what we counted previously, the above is equivalent to

$$n\mathbb{E}(\mathbb{1}_{A_j}) + n(n-1)\mathbb{E}(\mathbb{1}_{A_i \cap A_j})$$

, and again by the fundamental bridge,

$$n\mathbb{P}(A_j) + n(n-1)\mathbb{P}(A_i \cap A_j)$$

We now have every component to construct the variance of S_n,

$$\underbrace{n\mathbb{P}(A_j) + n(n-1)\mathbb{P}(A_i \cap A_j)}_{\mathbb{E}(S_n^2)} - \underbrace{\big(n\mathbb{P}(A_j)\big)^2}_{\mathbb{E}^2(S_n)}$$

That's it. Here's a matrix

$$\mathbf{A} = \{a_{ij} = \mathbb{1}_{A_i}\mathbb{1}_{A_j}\} \in \{0,1\}^{n \times n}$$

capturing the like (on diagonal) and cross terms (off diagonal):

$$\begin{bmatrix} \mathbb{1}_{A_1}\mathbb{1}_{A_1} & \mathbb{1}_{A_1}\mathbb{1}_{A_2} & \cdots & \mathbb{1}_{A_1}\mathbb{1}_{A_n} \\ \mathbb{1}_{A_2}\mathbb{1}_{A_1} & \mathbb{1}_{A_2}\mathbb{1}_{A_2} & \cdots & \mathbb{1}_{A_2}\mathbb{1}_{A_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}_{A_n}\mathbb{1}_{A_1} & \mathbb{1}_{A_n}\mathbb{1}_{A_2} & \cdots & \mathbb{1}_{A_n}\mathbb{1}_{A_n} \end{bmatrix}$$

(11) Birthday Problem

Suppose there are n students in a class. What is the probability that at least two students in the class have the same birthday? Give an expression, that is numerically simple to evaluate, to estimate this probability.

(11.1) Solution: Birthday Problem

Define A to be the event that ≥ 2 students share a the same birthday. We're interested in computing $\mathbb{P}(A)$ and it's clear that it's best to consider A^c. That is,

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

$\mathbb{P}(A^c)$ is the chance that no one shares the same birthday, this is

$$\frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365-(i-1)}{365}$$

expressing the above using product notation,

$$\mathbb{P}(A) = 1 - \prod_{i=1}^{n}\left(1 - \frac{i-1}{365}\right)$$

we now mend the product into an approximation, one that is tangibly easy to compute for all n

$$\begin{aligned}
\prod_{i=1}^{n}\left(1 - \frac{i-1}{365}\right) &= \prod_{i=0}^{n-1}\left(1 - \frac{i}{365}\right) \\
&= \exp\left[\log \prod_{i=0}^{n-1}\left(1 - \frac{i}{365}\right)\right] \\
&= \exp\left[\sum_{i=0}^{n-1} \log\left(1 - \frac{i}{365}\right)\right] \\
&\approx \exp\left[\sum_{i=0}^{n-1} -\frac{i}{365}\right] \\
&= \exp\left[-\frac{1}{365}\sum_{i=0}^{n-1} i\right] \\
&= \exp\left[-\frac{1}{365}\frac{n(n-1)}{2}\right]
\end{aligned}$$

hence,

$$\mathbb{P}(A) \approx 1 - \exp\left[-\frac{1}{365}\frac{n(n-1)}{2}\right]$$

□

(11.2) Remarks: Birthday Problem

Notice $\mathbb{P}(A)$ quickly approaches 1 as n becomes larger and larger. It's graph is a sigmoid, leveling off near 1 as $n \to \infty$ and for $n = 23$, $\mathbb{P}(A) = 0.5$.

One might ask about the product ending with $\frac{365-(i-1)}{365}$. Try computing $\mathbb{P}(A^c)$ for small $n \geq 2$, things should become clear. It also reinforces two edge cases: the chance that one person does not share the same birthday in a class of one, that probability being 1 and the case where there are more than 365 students, in which case it is necessarily true that at least two students share the same birthday and $\mathbb{P}(A^c) = 0$.

Regarding the math, recall the following: $\exp\big[\log(x)\big] = x$, $\log(xy) = \log(x) + \log(y)$, and $\log(1+x) \approx x$ for small x, `#complement #independence #approximation`

(12) Elevator Problem

A building has 10 floors above the basement. If 12 people get into an elevator at the basement, and each chooses a floor at random to get out, independently of the others, at how many floors do you expect the elevator to make a stop to let out one or more of these 12 people?

(12.1) Solution: Elevator Problem

We're interested in the number of distinct elevator buttons pressed by the 12 passengers at the basement floor. This begs the question: "Was that button pressed?" Leading us to construct the indicator

$$\mathbb{1}_{A_j} = \begin{cases} 1, & \text{if the } j^{\text{th}} \text{ button is pressed} \\ 0, & \text{otherwise} \end{cases}$$

since there are 10 floors,

$$X = \sum_{j=1}^{10} \mathbb{1}_{A_j}$$

quantifying our objective,

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{j=1}^{10} \mathbb{1}_{A_j}\right) = \sum_{j=1}^{10} \mathbb{E}(\mathbb{1}_{A_j}) = 10\mathbb{P}(A_j)$$

$\mathbb{P}(A_j)$ is sticky to consider, so let's take a look at its complement: the chance that the j^{th} elevator button is *not* pressed,

$$\mathbb{P}(A_j) = 1 - \mathbb{P}(A_j^c)$$

$\mathbb{P}(A_j^c)$ is easy to compute; it's simply the chance that every passenger does not push button j. By independence of passengers pushing elevator buttons,

$$\mathbb{P}(A_j^c) = \underbrace{\frac{9}{10} \times \frac{9}{10} \times \cdots \times \frac{9}{10}}_{\text{12 times}}$$

stitching everything together,

$$\mathbb{E}(X) = 10\left[1 - \left(\frac{9}{10}\right)^{12}\right] \approx 7.2$$

□

(12.2) Remarks: Elevator Problem

One might ask why $\mathbb{P}(A_j)$ is a sticky quantity to pursue. Simply consider (without loss of generality) the event A_1, the event where the first button is pushed. This can happen in many ways: 1 of the 12 pushes it, 2 of the 12 pushes it, and so forth. Clearly it's more efficient to consider the complement of A_1. `#indicators #expectation #complement #independence`

(13) Counting Pairs of Socks Problem

Suppose you take $k \leq 2n$ socks out of a drawer. Let M be the number of pairs of socks that appear in your pile of k. Compute $\mathbb{E}(M)$ and $\sigma^2(M)$.

(13.1) Solution: Counting Pairs of Socks

Clearly M is a count; we can express M as a sum of indicators:

$$M = \sum_{j=1}^{n} \mathbb{1}_{A_j}$$

, where

$$\mathbb{1}_{A_j} = \begin{cases} 1, & \text{if the } j^{\text{th}} \text{ pair of socks appears} \\ 0, & \text{otherwise} \end{cases}$$

Computing $\mathbb{E}(M)$, we have

$$\mathbb{E}(M) = \mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}_{A_j}\right) = \sum_{j=1}^{n} \mathbb{E}(\mathbb{1}_{A_j})$$

we know that $\mathbb{E}(\mathbb{1}_{A_j}) = \mathbb{P}(A_j)$, in words we're looking to compute the chance that the j^{th} pair of socks appear. Since it's implied that we are drawing without replacement, that probability is

$$\mathbb{P}(A_j) = \frac{\binom{2}{2}\binom{2n-2}{k-2}}{\binom{2n}{k}}$$

by linearity of expectation,

$$\mathbb{E}(M) = n \times \frac{\binom{2}{2}\binom{2n-2}{k-2}}{\binom{2n}{k}}$$

Now for variance, we focus on $\mathbb{E}(M^2)$, which we know to be

$$n\mathbb{E}(\mathbb{1}_{A_j}) + n(n-1)\mathbb{E}(\mathbb{1}_{A_i}\mathbb{1}_{A_j})$$

the first term is cake (since we already computed). The second term is

$$\mathbb{E}(\mathbb{1}_{A_i}\mathbb{1}_{A_j}) = \mathbb{P}(A_iA_j)$$

, that is we're interested in computing the probability that both pair i and j appear in our sample k. Similarly to $\mathbb{P}(A_j)$, we use hypergeometric

$$\mathbb{P}(A_iA_j) = \frac{\binom{4}{4}\binom{2n-4}{k-4}}{\binom{2n}{k}}$$

We finally have everything to compute variance, $\sigma^2(M)$ is

$$n\frac{\binom{2}{2}\binom{2n-2}{k-2}}{\binom{2n}{k}} + n(n-1)\frac{\binom{4}{4}\binom{2n-4}{k-4}}{\binom{2n}{k}}$$

$$-\left[n\frac{\binom{2}{2}\binom{2n-2}{k-2}}{\binom{2n}{k}}\right]^2$$

$\square$

(13.2) Remarks: Counting Pairs of Socks

It's pretty obvious that there's a dependency on the n indicators. If you're wondering why suppose you take $k = 2$ socks out of the drawer and we know that for some j, $\mathbb{1}_j = 1$, then it's necessarily true that the remaining $n-1$ indicators are zero. Another way to fashion this notion is supposing that you take one more than half of the socks, that is $k = n + 1$. Then it's necessarily true that at least one of the indicators turn on. So let's say that you have 4 socks in the drawer and you pick out 3 of them, then one of the indicators must be on. In summary,

$$\mathbb{P}(A_i A_j) \neq \mathbb{P}(A_i)\mathbb{P}(A_j)$$

#dependentIndicators #expectation #variance #hypergeometric

(14) Sampling With and Without Replacement

A box contains tickets marked $1, 2, \ldots, n$. Two tickets are drawn from the box. Find the probabilities of the following events:

a) the first ticket drawn is 1 and the second 2

b) the numbers on the two tickets are consecutive integers e.g. $i, i+1$ for $1 \leq i \leq n-1$

c) the second number is drawn is bigger than the first number drawn

with and without replacement.

(14.1) Solution: Sampling With and Without Replacement

Let T_i be the ticket appearing for the i^{th} draw $i = 1, 2$.

a) Each $\omega \in \Omega$, looks like (i, j), $1 \leq i, j \leq n$. Since each pair is equally likely and $|\Omega| = n^2$, we have

$$\mathbb{P}(T_1 = 1, T_2 = 2) = \frac{1}{n^2}$$

b) The pairs of tickets satisfying $T_1 = i$ and $T_2 = i+1$ are $\{(1,2), (2,3), \ldots, (n-1, n)\}$, with each pair being equally likely and mutually exclusive we have for $1 \leq i \leq n-1$

$$\mathbb{P}(T_1 = i, T_2 = i+1) = \frac{n-1}{n^2}$$

c) We're looking to compute $\mathbb{P}(T_1 < T_2)$, if $T_1 = 1$, then we can choose any of the $n-1$ tickets such that $T_2 > T_1$, for $T_1 = 2$, we then have $n-2$ choices, and so on. In general, if $T_1 = i$, then there are $n-i$ choices for selecting T_2, such that $T_1 < T_2$. Since each event $T_1 = i$ is mutually exclusive, we're looking at

$$\sum_{i=1}^{n-1} \mathbb{P}(T_1 = i)\mathbb{P}(T_2 > i \,|\, T_1 = i)$$

It's clear that the first factor in the sum is $\frac{1}{n}$ and the second $\frac{n-i}{n}$. After a little math, we see

$$\mathbb{P}(T_1 < T_2) = \frac{n-1}{2n}$$

Sampling without replacement changes the cardinality of Ω, in particular $|\Omega| = n(n-1)$; since we have n choices for T_1 and $n-1$ choices for T_2. So for parts a) and b) we simply exchange the denominator with $n(n-1)$. For part c), the sum becomes

$$\sum_{i=1}^{n-1} \frac{1}{n} \cdot \frac{n-i}{n-1}$$

and again after a bit of math, we have

$$\mathbb{P}(T_1 < T_2) = \frac{1}{2}$$

$\square$

(14.2) Remarks: Sampling With and Without Replacement

You could make a simple counting argument for part c). For all $\omega \in \Omega$, of which (with replacement) there are n^2 of them , we're searching for (i, j) such that $i < j$. Observe,

$$\begin{matrix} (1,1) & (\mathbf{1},\mathbf{2}) & \cdots & (\mathbf{1},\mathbf{n}) \\ (2,1) & (2,2) & \cdots & (\mathbf{2},\mathbf{n}) \\ \vdots & \vdots & \ddots & \vdots \\ (n,1) & (n,2) & \cdots & (n,n) \end{matrix}$$

in other words, we need to count the elements above the main diagonal and there are clearly $\frac{n(n-1)}{2}$ of them. Hence,

$$\mathbb{P}(T_1 < T_2) = \frac{\frac{n(n-1)}{2}}{n^2} = \frac{n-1}{2n}$$

as expected from before. Sampling without replacement plays out similarly.