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 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 independence, conditioning, Bayes' rule, joint distributions, et cetera, counting methods (combinatorics), and set theory.

(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*

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

$\square$

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

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

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

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

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

(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 $\sum_i^k n_i = N$ 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

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

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 and a special case of Binomial 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 until 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 $p \to 0$. 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 $\mathbb{P}(a \leq X \leq b)$. We first standardize X by performing a linear change in scale, in particular

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

we then approximate $\mathbb{P}(a \leq X \leq b)$ 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 $n \geq 20$, $\sigma_X > 3$, and p not close to 0 or 1.

(5) CENTRAL LIMIT THEOREM (C.L.T)

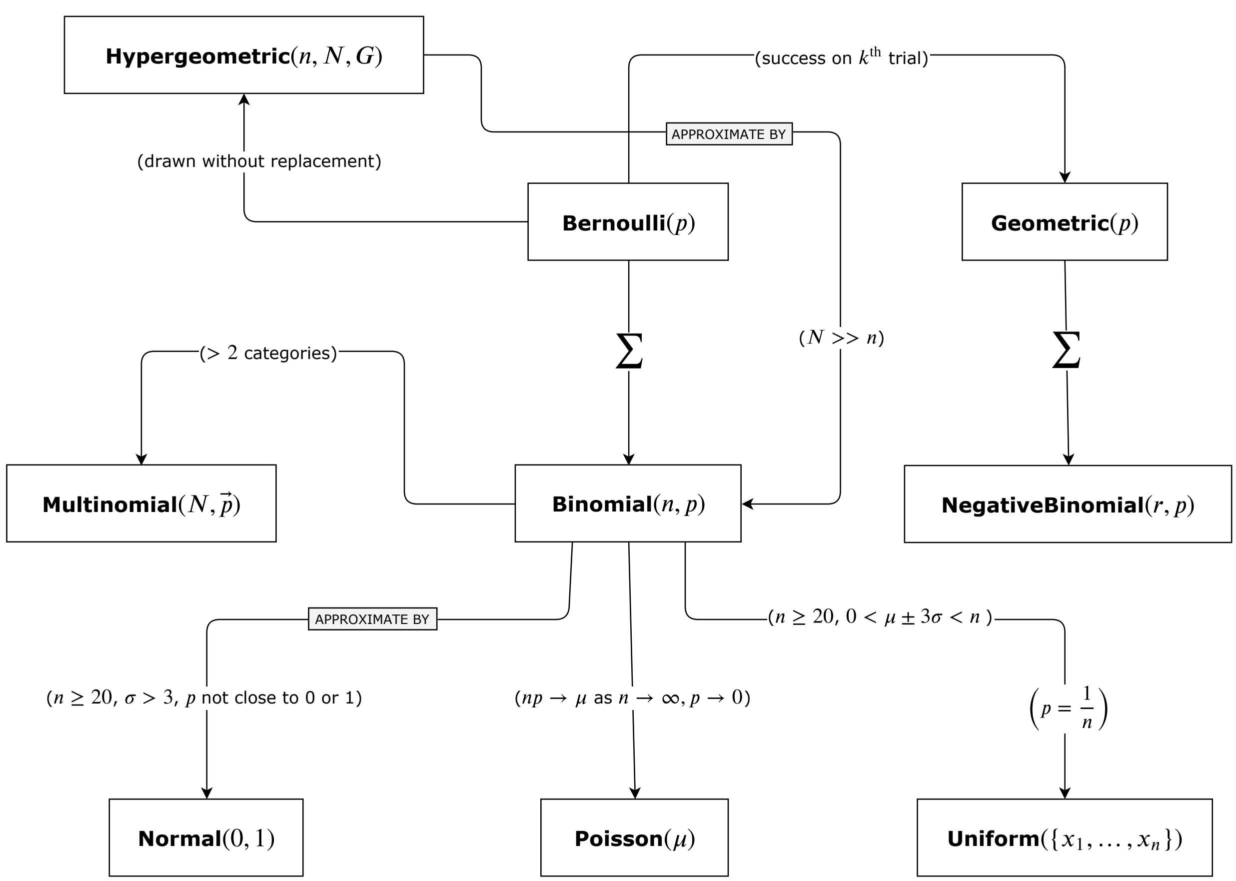
A powerful theorem. Let $\{X_j\}_{1 \leq j \leq n}$ be i.i.d. with mean μ_X and standard deviation σ_X for $1 \leq j \leq n$. Define $S_n = \sum_{j=1}^{n} X_j$. Then

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

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 $n \geq 25$ and $\mathbb{E}(S_n) \pm 3\sigma(S_n) \in \{\text{possible values of } X\}$

(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 $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$