


---

# Brief Treatise On Discrete Probability

John-Michael Laurel  [jm-laurel@berkeley.edu](mailto:jm-laurel@berkeley.edu)

---

## (1) RANDOM VARIABLES AND DISTRIBUTIONS

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

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

where  $\Omega$  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

$$X \sim \text{namedDistribution}(\cdot)$$

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.

---

<sup>1</sup>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, 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,  $\mathbb{E}[g(X_1, \dots, X_m)]$  is precisely

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

## (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  $X \in \{0, 1, \dots, n\}$  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  $Y \sim \text{Geometric}(p)$ , see §(3.5). □

## (2.2) VARIANCE AND STANDARD DEVIATION

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

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

and has computational form

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

The *standard deviation* denoted  $\sigma$  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  $X \sim \mathbf{Uniform}(\{a, a+1, \dots, b\})$

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

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  $X \sim \mathbf{Bernoulli}(p)$ , 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.)  $\mathbf{Bernoulli}(p)$  r.v.'s. That is to say if  $\{Y_j\}_{1 \leq j \leq n} \sim \mathbf{Bernoulli}(p)$ , then

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

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

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

<sup>2</sup>notation is equivalent to  $Y_j \sim \mathbf{Bernoulli}(p)$  for  $1 \leq j \leq n$ , which is equivalent to  $Y_1, \dots, Y_n \sim \mathbf{Bernoulli}(p)$

## (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.

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

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

$$(X_1, \dots, X_n) \sim \mathbf{Multinomial}(N, \vec{p})$$

Where  $\vec{p}$  is a probability vector<sup>3</sup>. It can be easily shown that the marginal distribution of any  $X_i \sim \mathbf{Binomial}(N, p_i)$ , 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  $k^{\text{th}}$  trial, implying exactly  $k - 1$  failures before that. For  $X \sim \mathbf{Geometric}(p)$ , 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

## (3.6) NEGATIVE BINOMIAL DISTRIBUTION

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

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

$\mathbb{E}(X) = n \left(\frac{G}{N}\right)$ ,  $\sigma^2(X) = n \left(\frac{G}{N}\right) \left(\frac{B}{N}\right) \left(\frac{N-n}{n-1}\right)$  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  $np \rightarrow \mu$  as  $n \rightarrow \infty$  and  $p \rightarrow 0$ . In words, we have many trials and the event of success is rare. Via consecutive probability ratios and for  $X \sim \mathbf{Poisson}(\mu)$ , 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  $p \rightarrow 0$ . e.g. Twitter notifications

<sup>3</sup>a probability vector is one whose entries sum to unity

## (4) NORMAL APPROXIMATION TO BINOMIAL

The continuous analog to Binomial is the Normal distribution. We approximate Binomial by Normal. Suppose  $X \sim \mathbf{Binomial}(n, p)$  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  $\Phi(z) = \int_{-\infty}^z f_X(x) dx$ <sup>4</sup>. Use approximation when  $n \geq 20$ ,  $\sigma_X > 3$ , and  $p$  not close to 0 or 1.

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

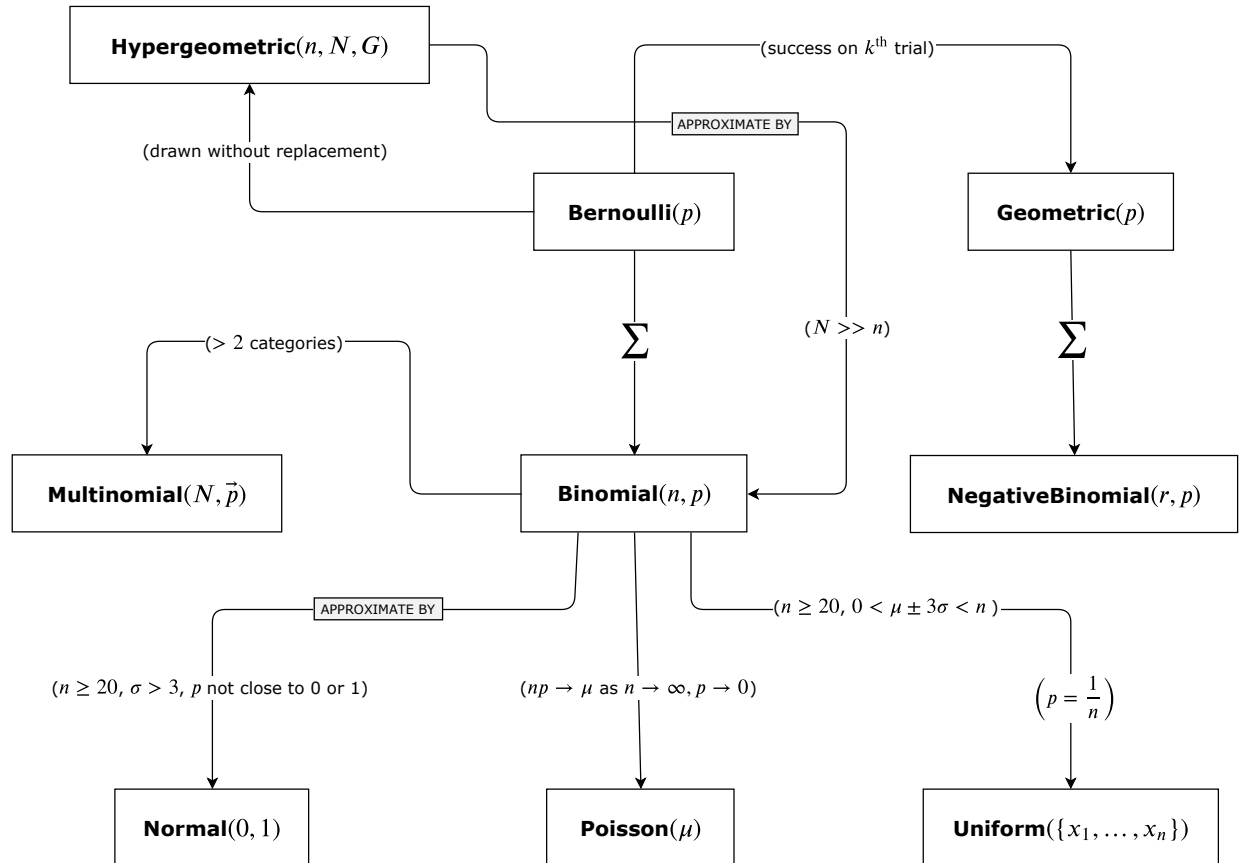
A powerful theorem. Let  $\{X_j\}_{1 \leq j \leq n}$  be i.i.d. with mean  $\mu_X$  and standard deviation  $\sigma_X$  for  $1 \leq j \leq n$ . Define  $S_n = \sum_{j=1}^n X_j$ . Then

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

as  $n \rightarrow \infty$ . 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  $\mathbb{E}(S_n) \pm 3\sigma(S_n) \in \{\text{possible values of } S_n\}$

## (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.



<sup>4</sup>here  $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

## (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  $\alpha$ , an upper bound on the probability that  $X$  is at least  $\alpha$  is

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

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

## (7.2) CHEBYSHEV'S INEQUALITY

Given  $X$ , its expectation  $\mu_X$ , and standard deviation  $\sigma_X$ , 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  $k = c/\sigma_X$ , where  $c$  is constant

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

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

## (8) 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

$$\begin{aligned} & p_A + p_D p_A + p_D^2 p_A + \dots \\ &= p_A(1 + p_D + p_D^2 + \dots) \\ &= p_A \left( \sum_{k=0}^{\infty} p_D^k \right) \\ &= \frac{p_A}{1 - p_D} \\ &= \frac{p_A}{p_A + p_B} \end{aligned}$$

□

## (9) 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,

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

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

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

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  $n^2 - n$ . 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  $S_n$ ,

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

That's it.