Wednesday, October 4 **Fall 2023**

Section 4: Discrete Distributions and Expectation

Srihari Ganesh

Based on section note formatting template by Rachel Li and Ginnie Ma '23

1 Summary

1.1 Random variables

Definition 1 (Random variable)**.** A **random variable** X is a function $X : S \to \mathbb{R}$ that maps the sample space to the real line. We usually omit the function notation and just say talk about the value of X.

Definition 2 (Support)**.** The **support** of a random variable X is the image of the sample space S. Precisely, the support is $\{x : \text{ there exists } \omega \in S \text{ such that } X(\omega) = x\}$.

Definition 3 (Probability mass function, PMF)**.** The **probability mass function (PMF)** is a function that takes a real number, $x \in \mathbb{R}$, as input and outputs the probability that the random variable takes on that value, $P(X = x)$. So the PMF is a function with domain $\mathbb{R}$ and codomain $[0, 1]$.

- You should address every possible value of x when defining the function.
  - For every x in the support of X, $P(X = x) > 0$.
  - For every x not in the support of X, $P(X = x) = 0$.
- The probabilities should be valid:

$$\sum_{x \in \mathbb{R}} P(X = x) = 1,$$
$$P(X = x) \in [0, 1] \text{ for all } x \in \mathbb{R}.$$

Definition 4 (Cumulative density function, CDF)**.** The **cumulative density function (CDF)** is a function that takes in a real number, $x \in \mathbb{R}$, and outputs the probability that the random variable is less than or equal to x, $P(X \leq x)$. So the CDF is a *non-decreasing* function (either increasing or a flat line) with domain $\mathbb{R}$ and codomain $[0, 1]$.

- You should again address every possible value of x, both in and not in the support.
- A valid CDF should be non-decreasing, with

$$\lim_{x \to 0} P(X \leq x) = 0,$$
$$\lim_{x \to 1} P(X \leq x) = 1.$$

- If the random variable has a bounded support (basically, there are some values that are too small or too big to be possible), then
  - If x is smaller than every value in the support, $P(X \leq x) = 0$.
  - If x is larger than every value in the support, $P(X \leq x) = 1$.

Notation 5 (i.i.d.)**.** We often have some random variables $X_1, \ldots, X_n$ that are **independent and identically distributed**, which we will abbreviate with the acronym **i.i.d.**

Notation 6 ("distributed as")**.** For a named distribution like the Binomial, we notate X being distributed as $\text{Bin}(n, p)$ with $X \sim \text{Bin}(n, p)$.

☣ **7.** We CANNOT set $X = \text{Bin}(n, p)$. Random variables cannot equal named distributions, they just follow the pattern given by the named distribution. As Joe says, the named distributions are a blueprint, the random variable is a house.

Remark 8 (finding the distribution of an r.v.)**.** If you are asked the distribution of a random variable (r.v.), you can give either the named distribution (with parameters defined), the PMF, or the CDF. Here's a general workflow:

1. Define the support of your r.v.
2. See if the random variable matches the story of any of the named distributions we have discussed. To see if an r.v. matches a distribution, some things to check are
   - For which named distributions is the support of your r.v. possible?
   - Are there draws/samples/trials? If so, are they independent?
   - If there is sampling, is it done with or without replacement?
3. If you can match a named distribution, what are the parameters? Are those parameters allowed for that named distribution?
4. If you can't match a named distribution, calculate the PMF using the information you checked about sampling and your counting skills.

1.2 Discrete distributions

You can find things like the support, PMF, CDF, and expectation in the table of distributions at the end of the textbook or start of the midterm handout. We'll focus on the stories and connections between distributions. For these discrete random variables (except for the Poisson), you should develop comfort with calculating their PMFs from scratch.

1.2.1 Bernoulli

Story: We run a trial with probability p of success. Let the random variable X be 1 if the trial succeeds or 0 if the trial fails. Then $X \sim \text{Bern}(p)$.

Connections:

- For $X \sim \text{Bern}(p)$, $1 - X \sim \text{Bern}(1 - p)$.
- For $X \sim \text{Bern}(p)$, $X^2 = X$, so $X^2 \sim \text{Bern}(p)$. If you're wondering why, check the support!

1.2.2 Binomial

Story: We run n *independent* trials, each with an *equal* probability p of success. Let X be the number of successful trials. Then $X \sim \text{Bin}(n, p)$.

Connections:

- For *n independent and identically distributed* Bernoulli random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Bern(p)$,

$$\sum_{i=1}^{n} X_i \sim \text{Bin}(n, p).$$

- This means $\text{Bern}(p)$ is equivalent to $\text{Bin}(1, p)$.
- For independent random variables $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$,

$$X + Y \sim \text{Bin}(n + m, p).$$

1.2.3 Hypergeometric

Story:

- *Capture/recapture elk*: There are N elk in the forest. In the past, we captured and tagged m of the elk. We now recapture n of the elk, where every set of n is equally likely and elk are sampled without replacement. Let X be the number of tagged elk among our n recaptured elk. Then $X \sim \text{HGeom}(m, N - m, n)$.
- *White and black balls in an urn*: There are w white balls and b black balls in a urn. We draw n balls from the urn without replacement, where each set of n balls is equally likely to be drawn. Let X be the number of white balls in our sample. Then $X \sim \text{HGeom}(w, b, n)$.

Connections:

- Notice the comparison between the Binomial and the Hypergeometric: using the urn story, if we sampled *with* replacement our random variable would be distributed $\text{Bin}(n, \frac{w}{w+b})$.

1.2.4 Geometric/First Success

Story: Suppose we're running independent Bernoulli trials with probability p of success. We stop running trials once one succeeds. Let X be the number of failed trials before (and *not* including) the first successful trial. Then $X \sim \text{Geom}(p)$.

Connections:

- The First Success distribution is essentially the same as the Geometric, but we include the first successful trial as part of our count. So it always holds that for $X \sim \text{Geom}(p)$, we have $X + 1 \sim \text{FS}(p)$.
- Note that the Geometric/First Success distributions have infinite supports, while the Binomial has a fixed number of trials. This is a quick way to tell them apart.

1.2.5 Negative Binomial

Story: Suppose we're running independent Bernoulli trials with probability p of success. We stop running trials after the r^{th} success. Let X be the number of failed trials before the r^{th} success (not including any of the successes in that count). Then $X \sim \text{NBin}(r, p)$.

Connections

- For *independent and identically distributed* $X_1, X_2, \ldots, X_r \overset{i.i.d.}{\sim} \text{Geom}(p)$, we get $\sum_{i=1}^{r} X_i \sim \text{NBin}(r, p)$.
- This means $\text{NBin}(1, p)$ is equivalent to $\text{Geom}(p)$.

1.2.6 Poisson

Story: There's no exact story to derive a Poisson. The only situation in which you'll have to come up with the Poisson on your own is in approximation, and that is quite rare.

Approximate story: Say there are many rare events $A_1, A_2, \ldots, A_n$ (so n large and $P(A_i) =<< 1$, which stands for much smaller than 1) which are nearly independent (which doesn't have a rigorous definition). Then if we let $\lambda = \sum_{i=1}^{n} P(A_i)$, $X = \sum_{i=1}^{n} I(A_i)$ is approximately distributed $\text{Pois}(\lambda)$.

Connections:

- As you can see in the approximate story, you can use the Poisson to count the number of independent/weakly-dependent rare events that occur.
- Suppose $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ with X, Y independent. Then $X + Y \sim \text{Pois}(\lambda + \mu)$.
- **Chicken-Egg**: suppose a chicken lays N eggs, with $N \sim \text{Pois}(\lambda)$. Suppose each egg has a probability p of hatching, with each egg's hatching being independent, and let X be the number of eggs that hatch and Y be the number of eggs that don't hatch.
  - X and Y are independent. X and Y are very conditionally independent given N since $N = X + Y$.
  - $X \sim \text{Pois}(\lambda p)$, $Y \sim \text{Pois}(\lambda(1-p))$.
  - $X|N = n \sim \text{Bin}(n, p)$.

1.3 Expectation

Definition 9 (Expectation)**.** The **expectation** of a random variable X with support A is the weighted average of its possible values, where we weight based on the probability of X taking on each value in its support. It is formally defined as

$$E(X) = \sum_{x \in A} xP(X = x).$$

Result 10 (Linearity)**. Linearity** states that for any random variables X, Y (which can be dependent!) and real number c,

$$E(X + Y) = E(X) + E(Y),$$

$$E(cX) = cE(X).$$

Result 11 (LOTUS)**.** The **law of the unconscious statistician (LOTUS)** states that the expectation of any function of a random variable, $g(X)$, can be found by

$$E(g(X)) = \sum_{x \in A} g(x)P(X = x).$$

For example, if we want to find $E(X^2)$, we simply swap x^2 in for x in the expectation formula to get $E(X^2) = \sum_{x \in A} x^2 P(X = x)$. Note that the probabilities here don't change, only what goes in front.

1.3.1 Indicator Random Variables

Definition 12 (Indicators)**.** An **indicator random variable** converts an event into a Bernoulli random variable. For an event A with $P(A) = p$, the corresponding indicator random variable $I(A) \sim \text{Bern}(p)$. This random variable is defined such that $I(A) = 1$ if A occurs and $I(A) = 0$ if A^c occurs. You might see other equivalent notation like I_A or I, just be clear about which event your indicator random variable corresponds to.

Result 13 (Fundamental bridge)**.** The **fundamental bridge** (vocab which is not used outside of Stat 110) gives that

$$E(I(A)) = P(A).$$

We use this result a lot to calculate expectations of random variables that can be expressed as the sum of indicators. This is nice because the indicators can be dependent, but linearity allows us to break the expectations apart!

Remark 14 (Using indicators)**.** A very common workflow to calculate an expectation is to write

1. Write the random variable as the sum of indicators, $X = \sum_i I(A_i)$, where each A_i is an event.
2. Apply linearity, $E(X) = \sum_i E(I(A_i))$.
3. Use the fundamental bridge, $E(X) = \sum_i P(A_i)$.

1.4 Variance

Definition 15 (Variance)**.** The **variance** of a random variable X is

$$Var(X) = E((X - E(X))^2).$$

Here basically all of the facts you have to know about variance:

- You'll usually use this equivalent formula instead:

$$Var(X) = E(X^2) - (E(X))^2$$

- $Var(X) = 0$ if X takes on a certain value with probability 1; for example, if $P(X = 3) = 1$. $Var(X) > 0$ otherwise.
- For a scalar $c \in \mathbb{R}$ and a random variable X,

$$Var(cX) = c^2 Var(X)$$

$$Var(X + c) = Var(X).$$

- For *independent* random variables X and Y

$$Var(X + Y) = Var(X) + Var(Y).$$

  For *dependent* random variables X and Y,

$$Var(X + Y) \neq Var(X) + Var(Y).$$

1.5 Handy math facts

- You are expected to know how to find the sum of an infinite geometric series: if $|x| < 1$,

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

otherwise the sum does not exist (it diverges). For finite geometric series (and any $x \neq 1$),

$$\sum_{n=0}^{\infty} x^n = \frac{1-x^n}{1-x}.$$

- You are also expected to be familiar with some e^x approximations, but you usually won't be asked to approximate without prompting. The Taylor series of e^x is

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

The compound interest formula also gives

$$e^x = \lim_{n\to\infty} (1 + \frac{x}{n})^n.$$

2 Practice Problems

1. **The difference between "=" and "$\sim$".** Every day, I flip a fair coin and eat breakfast if it lands heads. Let B_1 be the event that I ate breakfast yesterday, and let B_2 be the event that I ate breakfast today.

   For each of the following questions, you have three options:

   - They are equal, $=$,
   - they only share the same distribution, $\sim$,
   - or neither.

   Note that random variables that are equal always share the same distribution - pick the most specific option.

   (a) What is the relationship between $I(B_1)$ and $I(B_2)$?

   > **Solution**
   > $I(B_1)$ and $I(B_2)$ are both distribution Bern$(1/2)$.

   (b) What is the relationship between $I(B_1)$ and $I(B_1^c)$?

   > **Solution**
   > $I(B_1)$ and $I(B_1^c)$ are both distribution Bern$(1/2)$.

   (c) What is the relationship between $I(B_1)$ and $1 - I(B_1^c)$?

   > **Solution**
   > $I(B_1) = 1 - I(B_1^c)$.

2. **Category errors.** Let X_1, X_2 be random variables, let A_1, A_2 be events, and let $p_1 = 0.4, p_2 = 2$. For each pair, identify which are category errors (usually only one, but could be neither/both!).

$$P(A_1) = p_1 \text{ vs. } P(X_1) = p_1$$
$$P(A_1 = 3) = p_1 \text{ vs. } P(X_1 = 3) = p_1$$
$$P(I(A_1) = 1) = p_1 \text{ vs. } P(I(X_1) = 3) = p_1$$
$$A_1 + A_2 \text{ vs. } X_1 + X_2$$
$$A_1 \cap A_2 \text{ vs. } X_1 \cap X_2$$
$$P(A_2) = p_2 \text{ vs. } P(X_2) = p_2$$
$$P(p_1) = p_2 \text{ vs. } P(p_2) = p_1$$
$$A_1 \sim \text{HGeom}(n, r, p) \text{ vs. } X_1 \sim \text{Bin(n, r, p)}$$

   > **Solution**
   > We'll go through every category error and why:
   > - $P(X_1) = p_1$: $P(X_1)$ is the problem. We can't take the probability of a random variable.
   > - $P(A_1 = 3) = p_1$: $A_1 = 3$ is the problem. A_1 is an event (so a set), so it can't equal a number like 3.

- $P(I(X_1) = 3) = p_1$: $I(X_1)$ is the problem. We can take the indicator of an event (like $I(A_1)$), but the indicator of a random variable is not well-defined.
- $A_1 + A_2$: adding events is the problem. A_1 and A_2 are both sets and we need to do things like intersections or unions to combine them.
- $X_1 \cap X_2$: intersecting random variables is the problem. X_1 and X_2 can be thought of as functions or numbers; either way, we can only take the intersection of sets, not random variables.
- Both $P(A_2) = p_2$ and $P(X_2) = p_2$ are category errors. Both are invalid since $p_2 \notin [0, 1]$, so p_2 cannot be the value of a probability. $P(X_2)$ is additionally problematic because we can't take the probability of a random variable.
- Both $P(p_1) = p_2$ and $P(p_2) = p_1$ are category errors. Both are invalid since we can't take the probability of a number, only events. $P(p_1) = p_2$ is additionally problematic because $p_2 \notin [0, 1]$.
- Both $A_1 \sim \text{HGeom}(n, r, p)$ and $X_1 \sim \text{Bin}(\text{n}, \text{r}, \text{p})$ are category errors. A_1 is an event, not a random variable, so it cannot follow a distribution. While X_1 is a random variable, the Binomial distribution only takes two parameters, so $\text{Bin}(n, r, p)$ is not well-defined.

3. **Rearranging probabilities into known PMFs/CDFs.** It's generally desirable to reuse known results. For every problem below, rewrite the probability in terms the PMF of a named distribution; you don't have to solve beyond that. You may define new random variables and write your probability in terms of the new r.v.s as long as those r.v.s also have named distributions.

   (a) Suppose $X \sim \text{Bin}(n, p)$. What is $P(X \geq 1)$? Can you write this without a sum?

   Solution
   The support of a Binomial is $\{0, \ldots, n\}$. Then

   $$P(X \geq 1) = \sum_{k=1}^{n} P(X = k) = 1 - P(X = 0),$$

   where the second step comes by complementary counting.

   (b) Suppose $Y \sim \mathcal{FS}(p)$. What is $P(3 < (Y + 1)^2 \leq 9)$?

   Solution
   Let's first simplify the inequality:

   $$\begin{aligned} P(2 < (Y + 1)^2 \leq 9) &= P(\sqrt{2} < Y + 1 \leq 3) \\ &= P(\sqrt{2} - 1 < Y \leq 2). \end{aligned}$$

   The support of a First Success r.v. is $\{1, 2, \ldots, \}$. So since the $\sqrt{2} - 1 < 1$, we actually have that

   $$\begin{aligned} P(\sqrt{2} - 1 < Y \leq 2) &= P(Y \leq 2) \\ &= P(Y = 1) + P(Y = 2). \end{aligned}$$

(c) Suppose $X, Y \overset{i.i.d.}{\sim} \text{Pois}(\lambda)$. What is $P(X+Y=16)$?

> **Solution**
> Since X and Y are independent Poissons, $X+Y \sim \text{Pois}(\lambda+\lambda)$. So if we define $Z = X+Y$, then $Z \sim Pois(2\lambda)$ and we get
>
> $$P(X+Y=16) = P(Z=16).$$

4. **Calculating expectations**

(a) **[indicators]** *(based on Will Nickols' Stat 111 section notes)* There are n students in my section and k practice problems to do. For each problem, I draw a student's name out of a hat to explain their solution, putting the student's name back into the hat (so sampling students with replacement).

i. What is the expected number of students get selected **at least** once?

> **Solution**
> Let X be the number of students selected at least once. We can write this as $X = \sum_{i=1}^{n} I(A_i)$, where A_i is the event that student i gets selected at least once. So
>
> $$E(X) = \sum_{i=1}^{n} E(I(A_i)) = \sum_{i=1}^{n} P(A_i) = nP(A_1).$$
>
> by linearity, the fundamental bridge, and symmetry, in that order.
> Now see that $P(A_1) = 1 - P(A_1^c)$, where A_1^c is the event that student 1 never gets picked. So
>
> $$P(A_1) = 1 - P(A_1^c) = 1 - \left(\frac{n-1}{n}\right)^k.$$
>
> This makes the final answer
>
> $$E(X) = n\left[1 - \left(\frac{n-1}{n}\right)^k\right].$$

ii. What is the expected number of students who get selected **exactly** once? You may leave your answer as a sum.

> **Solution**
> Let Y be the number of students who get selected exactly once. Let $Y = \sum_{j=1}^{k} I(B_j)$, where B_j is the event that the student who solves the j-th did not solve any of the previous problems. In this way, we only count a student for the first time that they solve a problem, so each student is only accounted in one indicator.

From here we proceed similarly to part (ii):

$$\begin{aligned} E(Y) &= \sum_{j=1}^{k} E(I(B_j)) \\ &= \sum_{j=1}^{k} P(B_j) \\ &= \sum_{j=1}^{k} \left[1 - \left(\frac{n-1}{n}\right)^{j-1}\right] \\ &= k - \sum_{j=1}^{k} \left(\frac{n-1}{n}\right)^{j-1}. \end{aligned}$$

(b) Suppose X is a random variable such that $E(X^2) = a$ and $Var(X^2) = b$. What is $E(X^4)$ in terms of a and b?

Solution
Note that

$$\begin{aligned} Var(X^2) &= E((X^2)^2) - (E(X^2))^2 \\ &= E(X^4) - (E(X^2))^2. \end{aligned}$$

So

$$\begin{aligned} E(X^4) &= Var(X^2) + (E(X^2))^2 \\ &= b + a^2. \end{aligned}$$

(c) **[challenge problem]** Let $X \sim \text{Geom}(p)$ with $q = 1 - p$, let A_X be the event that X is even, and $Y = I(A_X)X/2$. Find $E(Y)$ exactly by pattern-matching the LOTUS expression for $E(Y)$ to the expectation of a (different) Geometric r.v.

Solution
Let's use the notation that even if the subscript is not a random variable, A_k is the event that k is event; so $I(A_8) = 1$ always, $I(A_{13}) = 0$ always, etc. Applying LOTUS gives us

$$\begin{aligned} E(Y) &= E(I(A_X)X/2) \\ &= \sum_{k=0}^{\infty} I(A_k)\frac{k}{2}P(X = k) \\ &= \sum_{k \in \{0,2,4,6,\ldots\}} \frac{k}{2}P(X = k). \end{aligned}$$

We can rewrite this as a normal sum by setting $k = 2i$ and summing over i, so

$$E(Y) = \sum_{i=0}^{\infty} iP(X = 2i)$$

Plugging in the PMF,

$$E(Y) = \sum_{i=0}^{\infty} ipq^{2i}.$$

This looks similar to the expectation for a Geom$(1 - q^2)$ random variable. Then

$$\begin{aligned} E(Y) &= \sum_{i=0}^{\infty} ipq^{2i} \\ &= \frac{p}{1-q^2}\sum_{i=0}^{\infty} i(1-q^2)(q^2)^i. \end{aligned}$$

Using the expectation of a Geometric, the sum above is $\sum_{i=0}^{\infty} i(1-q^2)(q^2)^i = \frac{q^2}{1-q^2}$. This gives us

$$E(Y) = \frac{p}{1-q^2}\frac{q^2}{1-q^2} = \frac{pq^2}{(1-q^2)^2}.$$