# BINOMIAL DISTRIBUTION, LLN, CLT AND NORMAL DISTRIBUTION

Contents

## 1. Reading

- Read sections 2.8 and 3.1 [**?**]

## 2. Learning objectives

- Familiarity with expectations and variances
- Familiarity with the binomial distribution and using it as a model for data.
- The Law of large numbers
- Understanding of continuous probability distributions and what a probability density is.
- Central limit theorem and Normal distribution

## 3. Expectation and variance

### 3.1. Sample averages and expectation.

- Usually it is difficult to obtain the full distribution of a random variable from data and it may not even be that relevant for the questions we are asking. Instead, we would like to summarize properties of a random variable from **statistics**. The simplest example is the **sample mean** or **sample average**. If we have many samples $Y_1, Y_2, \ldots, Y_n$ of a random variable $Y$ (e.g. answers to a survey question), the sample mean is defined as

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

More generally, we might look at the average of some function of a random variable

$$\overline{f(Y)} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)$$

- These sample means can be computed from the probability distribution. Suppose each $Y_i$ can take on outcomes $Y = 1, 2, 3, \ldots, m$. If $n$ is large, then the fraction of samples for which $Y_1 = y$ will be $P(Y_1 = y)$, thus the sample mean converges to the true mean:

$$\overline{Y} = \frac{1}{n}\sum_{i} Y_i = \frac{1}{n}\sum_{y=1}^{m} y n_i = \sum_{y=1}^{m} y\frac{n_i}{n} \approx \sum_{y=1}^{m} yP(Y = y)$$

- Sometimes we write $P(Y_i = y_i)$ and sometimes we write The expression on the right is the definition of the mean, or **expectation** denoted $E[Y]$.
- In many cases we would like to measure deviations from the mean. For this we look at the **variance**,

$$\text{var}(Y) = E[(Y - E[Y])^2]$$

Another way to write this is

$$\text{var}(Y) = E[Y^2] - 2E[Y]^2 + E[Y]^2 = E[Y^2] - E[Y]^2$$

We can of course estimate this from a large number of samples, although as we will later see the obvious way of doing this by replacing the expectations with sample averages is not optimal. The square root of the variance is called the **standard deviation**.

**Example 1** (Mean and variance of Bernoulli random variable)**.** *Let $Y$ be a Bernoulli random variable with parameter $q$. We will use the convention that $Y = 1$ with probability $q$.*

*Question: What is $E[Y]$ and $\text{var}(Y)$?*

*Solution: Using the definitions above*

$$E[Y] = P(Y = 0) \times 0 + P(Y = 1) \times 1 = q$$

*similarly you should be able to see that* $\text{var}(Y) = q(1-q)$. *In the python notebook we test the formula* $E[Y] = q$.

### 3.2. Conditional expectation.

- We define the conditional expectation as the expectation of the conditional variable; that is,

$$\mathbb{E}[X|Y = y] = \sum_x xP(X|Y = y)$$

With samples

$$\{(x_1, y_1), \dots, (x_n, y_n)\}$$

we could obtain this from a sample by taking the sample mean of $X$ among just the samples where $y_i = y$; that is, let

$$n_1 = \text{number of samples where } y_i = y$$

then

$$\mathbb{E}[X|Y = y] \approx \frac{1}{n_1}\sum_{i=1}^{n} 1_{\{y_i = y\}} x_i$$

where $1_{\{y_i=y\}}$ is the **indicator function**

$$1_{\{y_i=y\}} = \begin{cases} 0 & \text{if } y_i \neq y \\ 1 & \text{if } y_i = y \end{cases}$$

- As we already noted, the conditional probabilities can tell us whether two variables are independent. That is, $P(X|Y) = P(X)$ if and only if $X$ and $Y$ are independent. If $X$ and $Y$ are independent, then $E[X|Y = y] = E[X]$ for all $y$ but the converse is false: **it is possible that this is true but $X$ and $Y$ are not independent!** We will say about this later.

**Example 2** (Computing conditional from probability model ). *Consider the pair of random variables* $(Y_A, Y_B)$ *defined by the probability distribution we saw in week 1:*

$$P(Y_A, Y_B) = \begin{cases} 1/2 & \textit{if } Y_A = 0 \textit{ and } Y_B = 0 \\ 1/8 & \textit{if } Y_A = 0 \textit{ and } Y_B = 1 \\ 1/8 & \textit{if } Y_A = 1 \textit{ and } Y_B = 0 \\ 1/4 & \textit{if } Y_A = 1 \textit{ and } Y_B = 1 \end{cases} \tag{1}$$

*Question: Compute* $E[Y_A|Y_B = 1]$

*Solution: We can obtain the conditional distribution of* $Y_A$ *as*

$$P(Y_A = 1|Y_B = 1) = \frac{P(Y_A = 1, Y_B = 1)}{P(Y_B = 1)} = \frac{1/4}{3/8} = \frac{2}{3}$$

*Note that his means* $P(Y_A = 0|Y_B = 1) = 1/3$ *and so the conditional distribution of* $Y_A$ *is*

$$Y_A|(Y_B = 1) \sim \text{Bernoulli}(2/3)$$

*which means*

$$E[Y_A|Y_B = 1] = \frac{2}{3}.$$

**Example 3** (Computing conditional expectation from data ). *Consider the following data containing children's test scores and some other information. Let* $Y$ *be the test score and* $X$ *be a binary variable representing whether the mother graduated high school.*

*Question: Compute* $E[Y|X = 0]$ *and* $E[Y]$. *Based on this, do you think it's likely that* $X$ *and* $Y$ *are independent?*

*Solution: See Python notebook.*

### 3.3. Some additional properties of expectation.

- Previously, I introduced the notation of expectation and we saw how to compute expectations in Python. Now we cover some addition properties.
  (1) **Linearity:** For two random variables $X$ and $Y$,

$$E[X + Y] = E[X] + E[Y]$$

*Proof.* We have

$$\begin{aligned} E[X + Y] &= \sum_y \sum_x (x + y)P(X = x, Y = y) \\ &= \sum_x \sum_y xP(X = x, Y = y) + \sum_x \sum_y yP(X = x, Y = y) \\ &= \sum_x x\left(\sum_y P(X = x, Y = y)\right) + \sum_y y\left(\sum_x P(X = x, Y = y)\right) \\ &= E[X] + E[Y] \end{aligned}$$

□

(2) **Multiplication by a constant:** If $a$ is a constant (meaning it is not random), then

$$E[aX] = aE[X]$$

*Proof.* Left as an exercise. □

(3) **Factoring for independent variables:** If $X$ and $Y$ are independent, the

$$E[XY] = E[X]E[Y]$$

*Proof.* Using independence, we have

$$\begin{aligned} E[XY] &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xP(X = x)yP(Y = y) \\ &= \left(\sum_x xP(X = x)\right)\left(\sum_y yP(X = x)\right) = E[X]E[Y] \end{aligned}$$

□

(4) **Tower property:** Let $X$ and $Y$ be two random variables.

$$E[E[X|Y]] = E[X]$$

*Proof.* Left as an exercise. □

**Example 4** (Calculating probabilities)**.** *Consider the model probability model for a variable $X$ (similar to Equation 4)*

$$P(X) = \begin{cases} 1/2 & \text{if } X = 1 \\ 1/8 & \text{if } X = 2 \\ 1/8 & \text{if } X = 3 \\ 1/4 & \text{if } X = 4 \end{cases}$$

*and define*

$$Y = X \cdot Z, \quad Z = \text{Geometric}(1/2).$$

*In this example we consider the joint distribution $(Y, X)$.*
*Question: Using simulations of this model, illustrate properties of expectation in Python.*

*Solution: See Python notebook.*

## 4. Binomial Distribution

- Suppose

$$Y_i \sim \text{Bernoulli}(q), \quad i = 1, \dots, N$$

are independent. We will use the convention that $Y_i = 1$ with probability $q$. Let

$$Y = \sum_{i=1}^{N} y_i$$

Then we say $Y$ follows **binomial distribution** and write

$$Y \sim \text{Binomial}(N, q)$$

**Example 5** (Calculating probabilities)**.** *Let $N = 3$ and $k = 2$.*

*Question: Calculate $P(Y = 2)$?*

*Solution: There are 3 possible sequences that give $Y = 2$.*

$$(1, 0, 1), (1, 1, 0), (0, 1, 1)$$

*The probability that we see any particular one of these is $(1 - q)q^2$. For example,*

$$\begin{aligned} P(y_1 = 1, y_2 = 0, y_3 = 1) &= P(y_1 = 1)P(y_2 = 0)P(y_3 = 1) \\ &= q(1 - q)q = q^2(1 - q). \end{aligned}$$

*Therefore the chance to observe $Y = 2$ with $N = 3$ is*

$$P(Y = 2) = P((1, 0, 1)) + P((1, 1, 0)) + P((0, 1, 1)) = 3q^2(1 - q).$$

Note that the binomial distribution has two parameters, $N$ and $q$, representing the number of flips and probability of success respectively.

- Now let's think about what the probability distribution will look like. The chance to find any **particular** configuration of $k$ ones is

$$q^k(1 - q)^{N-k}$$

because they are independent.

However, we need to account for the fact that there are many configurations with $k$ ones. In general, there are

$$\binom{N}{k} = \frac{N!}{k!(N - k)!} = \frac{N \times (N - 1) \times (N - 2) \times \cdots \times (N - k + 1)}{k \times (k - 1) \times (k - 2) \times \cdots \times 1}$$

way to have $k$ ones among $N$ samples. [1]

This implies

$$P(Y=k)=\binom{N}{k}q^k(1-q)^{N-k}.$$

A graph of this function looks like a bell curve when $N$ is large – we will see this in Python. Let's explore binomial random variables in python.

**Example 6** (Expectation of binomial)**.** *Let $Y$ be a binomial random variable.*

*Question: What are $E[Y]$ and $\text{var}(Y)$?*

*Solution:*

$$E[Y]=\sum_{k=1}^{N}kP(Y=k)=\sum_{k=1}^{N}k\binom{N}{k}q^k(1-q)^{N-k}=\cdots.$$

*A much easier way is to use the definition of a Binomial random variable and exceptions*

$$\begin{aligned}E[Y]&=E\left[\sum_{j=1}^{N}X_i\right]\\&\underset{(1)}{=}\sum_{j=1}^{N}E\left[X_i\right]=Nq\end{aligned}$$

*where we are using the fact that averages are additive (property (1)). Similarly,*

$$\begin{aligned}E[Y^2]&=E\left[\left(\sum_{j=1}^{N}X_i\right)^2\right]=E\left[\sum_{i=1}\sum_{j=1}^{N}X_iX_j\right]\\&\underset{(1)}{=}\sum_{i=1}\sum_{j=1}^{N}E[X_iX_j]\underset{(3)}{=}\sum_{i=1}^{N}\sum_{j\neq i}^{N}q^2+Nq(1-q)+Nq^2\\&=N(N-1)q^2+Nq(1-q)+Nq^2=Nq(1-q)+N^2q^2\end{aligned}$$

*Therefore*

$$\text{var}(Y)=E[Y^2]-E[Y]^2=Nq(1-q)$$

To summarize what we learned in Example 6

$$E[Y]=qN \qquad \text{var}(Y)=Nq(1-q).$$

- The important observation that the mean grows much faster with $N$ than the variance is also captured by the coefficient of variation:

$$\text{CV}=\frac{\sqrt{\text{var}(Y)}}{\mathbb{E}[Y]}=\sqrt{\frac{(1-q)}{q}\frac{1}{N}}.$$

The idea is that we are measuring the variation *relative* to the average. This is relevant for many applications where we only care about the relative deviations.

**Example 7** (Example: election modeling)**.** *Consider a model of votes in an election involving two candidate. Let $q$ be the fraction of people in the population who support candidate one and suppose $N$ people vote at the election (you can assume $N$ is much less than the total number of people in the population, as voter turnout is low). Then the number of people, $M$, who vote for the first candidate can be modeled as*

$$M\sim\text{Binomial}(N,q)$$

*Think about the assumption we are making when we use this model.*

*Question: Suppose there is a city in which a fraction $q=0.51$ of people support a candidate for city council. If $N=1000$ people turnout for the election, what is the chance that the actual vote share, $\phi=M/N$, differs from the actual fraction of support throughout the population by more than 1%?*

*Solution: See the Python notebook.*

- You should be able to see that $E[\phi]=q$. What about the variance?

$$\text{var}(\phi)=\text{var}(Y/N)=\frac{1}{N^2}\text{var}(Y)=\frac{q(1-q)}{N} \tag{2}$$

Notice that this will tend towards zero as $N\to\infty$. Meanwhile, $E[\phi]$ has no dependence on $N$. This is a very important property, as it allows us to determine $q$ by approximating $E[\phi]$ with the sample mean.

[1] To the formula above, let $C_{N,k}$ denote the number of sequences with $k$ ones. We can break $C_{N,k}$ up into the number of terms for which 1 appears as the first element of the sequence and those for which zero is the first. If one comes first, we have $N-1$ remaining slots to place $k-1$ ones, thus there are $C_{N-1,k-1}$ of these sequences. Similarly, if zero comes first, we have $N-1$ slots but now all $k$ ones to place, thus there are $C_{N-1,k}$ of these. It follows that

$$C_{N,k}=C_{N-1,k-1}+C_{N-1,k}.$$

Notice that the quantity $C_{N,k}$ will be smallest when $k=1$ or $k=N$, since in these cases there is only one way to configure the sequence: $C_{N,1}=C_{N,N}=1$. You can solve the recursion to obtain the formula.

- You should recognize that the assumption of independence is very important here. In the exercises you will see an instance where this may break down for an election model. It is a bit contrived, but contrived examples, which we sometimes refer to as **toy models**, can be very helpful when it comes to building our intuition.

## 5. Law of large numbers

- The binomial distribution illustrates a very basically principle that we have already used a number of times: When we sum over a large number of independent random variables and divide by the total number, the result is close to the mean. This is the Law of Large Numbers (LLN).

  **Theorem 1.** *Let $X_i$ be independent and identically distributed and set*

$$S_N = \sum_{i=1}^{N} X_i.$$

  *If $E[X_i] < \infty$, then $S_N/N \to E[X_i]$.*

  This is not very precise, since we should really be specific about what it means for a random number to converge to something, but for our purposes it will suffice to think of this as saying that for large enough $N$, $S_N/N$ will not differ from $E[X_i]$ very much.

## 6. Mathematical detour: Continuous random variables

- Here we will talk about random variables which take on continuous values. This is useful because if we think of something like the fraction of a vote that one candidate got in an election, for a very large number of votes this should be close to a continuous variable which takes values on $[0, 1]$.
- A uniform random variable, denoted

$$Y \sim \text{Uniform}(a, b) \tag{3}$$

  has an equal chance of taking any number in the interval $[a, b]$ (we assume $a < b$). Let $L = b - a$. This is distinct from other distributions we have encountered in that it is a **continuous distribution**, rather than discrete, so we need to think slightly differently about how to define the probability distribution.
- For the uniform distribution,

$$P(y_1 \leq Y \leq y_2) = \frac{y_2 - y_1}{L} \tag{4}$$

  for $a < y_1 < y_2 < b$. That is, the chance for $Y$ to fall in any interval is simply the length of that interval. This insures that that the probability of $Y$ being somewhere in $[a, b]$ is one: $P(a \leq Y \leq b) = 1$. Note that as $y_2 \to y_1$, $P(y_1 \leq Y \leq y_2) \to 0$. This tells us that the chance for $Y$ to take any specific value is 0. Indeed, there are simply two many number (uncountably many) in any interval to assign positive probability to each. For continuous variables, it is sometimes useful to work with the density, $f(y)$ (we will use lower case letters for density and uppercase for probability distributions). $f(y)$ is the the probability per unit $Y$, meaning that if we look in a small interval

$$f(y)dy = P(y \leq Y \leq y + dy) = \frac{dy}{L}. \tag{5}$$

  Thus, for uniform distribution the density is $1/L$ if $y \in [a, b]$ and 0 otherwise.

**Example 8** (Condition with continuous random variables)**.** *If $Y$ is uniform on $[0, 1]$.*

*Question: What is the density of $Y|Y < 1/2$? That is, what is the density of the of $Y$ among samples for which we know $Y < 1/2$. Check the answer with simulations.*

*Solution: We can start with the definition of density*

$$P(y_1 < Y < y_2 | Y < 1/2) = \frac{P(y_1 < Y < y_2, Y < 1/2)}{P(Y < 1/2)} \tag{6}$$

*What is the think on the top? We will assume $y_1 > 0$ and $y_2 < 1/2$, then the numerator is $y_2 - y_1$, since $Y < 1/2$. The key here is that if $Y \in [y_1, y_2]$ $Y < 1/2$ is automatically true, so the chance that BOTH of these things are true in a sample is the chance that the more restrictive one is true.*

*The denominator is $P(Y < 1/2) = 1/2$. This means*

$$P(y_1 < Y < y_2 | Y < 1/2) = 2(y_2 - y_1) \tag{7}$$

*This means the density is*

$$f(y|Y < 1/2) = 2 \tag{8}$$

*Thus*

$$Y|(Y < 1/2) \sim \text{Uniform}(0, 1/2) \tag{9}$$

*We also check this with simulations in the Python notebook.*

### 6.1. Cumulative density function.

- Sometimes it is useful to characterize a continuous distribution not by the density, but by the **cumulative distribution function (CDF)**, defined as

$$F(y) = P(Y < y). \tag{10}$$

- The **median** is the value $y_m$ for which $F(y_m) = 1/2$.

**Example 9** (median of a Uniform distribution)**.** *To better understand density and CDF, imagine a student says they will arrive at my office between noon and 3. Let $Y$ represent the time a student arrives, which we will model as a Uniform random variable. Then the density is $f(y) = 1/3$ which has units 1/hours. We can think of $f$ as the rate at which the CDF increases – that is, it is the velocity of probability.*