BASIC STATISTICAL MODELS

ETHAN LEVIEN

Contents

1. Binomial Distribution

A situation that often arrises is that we take many, say N, independent samples from a Bernoulli distribution. Now let Y be the number of 1s. Symbolically,

$$Y = \sum_{i=1}^{N} y_i, \quad y_i \sim \text{Bernoulli}(q). \tag{1}$$

Then Y follows **binomial distribution**:

$$Y \sim \text{Binomial}(N, q) \tag{2}$$

The binomial distribution has two parameters, N and p. Now let's think about the probability distribution. The chance to find any particular configuration of k ones is $q^k(1-q)^{N-k}$ because they are independent. For example

$$P(y_1 = 1, y_2 = 0, y_3 = 1) = P(y_1 = 1)P(y_2 = 0)P(y_3 = 1) \tag{3}$$

$$= q(1-q)q = q^2(1-q). \tag{4}$$

However, there are many configurations with k ones, in-fact there are

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}, \tag{5}$$

and therefore

$$P(Y = k) = \binom{N}{k} q^k (1-q)^{N-k}. \tag{6}$$

The binomial distribution has a mean and variance

$$\mathbb{E}[Y] = qN \qquad \text{var}(Y) = Nq(1-q). \tag{7}$$

These formulas come from the fact that for sums of independent variables, the variance and expectation sum.

An important feature of the Bernoulli random variables is that the mean grows much faster in N than the standard deviation. This means that when N is very large, the deviations from the average will become very small relative to the mean. An important measure of variation relative to the mean is the coefficient of variation

$$\text{CV} = \frac{\sqrt{\text{var}(Y)}}{\mathbb{E}[Y]}. \tag{8}$$

Binomial samples can be generated in numpy with

```
> y = np.random.binomial(n,p,n_samples)
```

Date: April 2022.

Often we are interested not in Y, but the fraction $\phi = Y/N$. For example, we might be interested in the vote share in an election. You should be able to see that $\mathbb{E}[\phi] = q$. What about the variance?

$$\text{var}(\phi) = \text{var}(Y/N) = \frac{1}{N^2}\text{var}(Y) = \frac{q(1-q)}{N} \tag{9}$$

Notice that this will tend towards zero as $N \to \infty$.

Example 1. *Coefficient of variation*

Exercise 1: *Generating binomial samples*

Exercise 2: *Binomial election modeling*

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

Exercise 3: *More election modeling*

2. Uniform distribution and probability density (optional)

A uniform random variable, denoted

$$Y \sim \text{Uniform}(a, b) \tag{10}$$

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. For the uniform distribution,

$$P(y_1 \le Y \le y_2) = \frac{y_2 - y_1}{L} \tag{11}$$

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: $P(a \le Y \le b) = 1$. Note that as $y_2 \to y_1$, $P(y_1 \le Y \le y_2) \to 0$. 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

$$f(y)dy = P(y \le Y \le y + dy) = \frac{dy}{L}. \tag{12}$$

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

2.1. Joint density and conditional density.

Conditioning works for probability density just as it does for probability distributions.

Example 2. *Conditioning with continuous variables*

2.2. 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

$$F(y) = P(Y < y). \tag{13}$$

What is the CDF of the uniform distribution? The **median** is the value y_m for which $F(y_m) = 1/2$. What is the 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.

3. Normal distribution and the central limit theorem

In the previous example, we say that if we take the average of many Bernoulli random variables, we get a histogram that looks a lot like a “bell curve" with a standard deviation was proportional to $1/\sqrt{n}$.

It turns out this is true when we add up *any* sequence of independent and independent distributed random variables which are not too pathological (actually it is also true for many sequences of random variables which are not independent). Since the “bell curve" arrises in the limit where we sum or average many random variables,

$$Y = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{14}$$

it makes sense to approximate it with a continuous variable. We call this a Normal random variable

$$Y \sim \text{Normal}(\mu, \sigma) \tag{15}$$

can take on any number, positive or negative, decimal or integer. We can generate Normal random variables in python with

```
> np.random.normal(0,1)
```

The Normal distribution is defined by the Gaussian

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{-(y-\mu)^2}{2\sigma}}. \tag{16}$$

This is the classic bell curve shown in Figure 1. The probability distribution for the Normal distribution is defined by the area under this curve. We I discuss in the previous section can think of f as the probability *per unit of the random variable*, e.g. probability/feet.

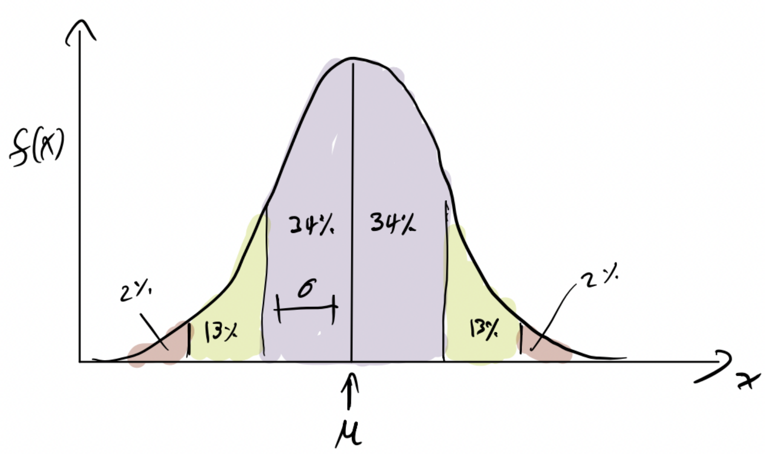

Figure 1. Probabilities in the Normal distribution

Suppose that

$$Y \sim \text{Normal}(5, 2) \tag{17}$$

what is (approximately) $P(Y > 7)$? Note that $5 + 2 = 7$, so this is asking how likely it is that a Normal variable is greater than 1 standard deviation above the mean. This about $13.5 + 2 = 15.5\%$

Example 3. *Comparing histograms*

Example 4. *Working with Normal random variables*

Exercise 4: *Hemoglobin levels*

4. Transformations of random variables

Now consider

$$X \sim \mathrm{Normal}(\mu_x, \sigma_x) \tag{18}$$

and let

$$Y = aX + b \tag{19}$$

We are just multiplying and shifting everything. Think about what this does to the histogram (and try it in Python). Hopefully you can convince yourself that Y should also be Normal, but what are the mean and variance? Taking the average of both sides,

$$\mathbb{E}[Y] = a\mu + b \tag{20}$$

and

$$\mathrm{var}(Y) = \mathrm{var}(aX) + \mathrm{var}(b) \tag{21}$$

Form the formula for variance, we know $\mathrm{var}(aX) = a^2\mathrm{var}(X)$. Also, $\mathrm{var}(b) = 0$ So

$$Y \sim \mathrm{Normal}(a\mu_x + b, |a|\sigma_x). \tag{22}$$

4.1. Standardizing.

We can transform any random variable into a so-called standard normal

$$Z \sim \mathrm{Normal}(0, 1). \tag{23}$$

For defined above,

$$Z = \frac{X - \mu_x}{\sigma_x} \tag{24}$$

Then $a = 1/\sigma_x$ and $b = -\mu_x/\sigma_x$. Plugging into Equation (22) yields a standard Normal. **Transforming X to a standard Normal is equivalent to measuring X in units of standard deviations.** For example, if we make a histogram of X, all this transformation does is change the X axis to units of standard deviations from the mean.

5. Linear regression model

We now introduce the concept of regression modeling. A very broad class of models in statistics for the relationship between two variables X and Y is a regression model:

$$Y = f(X) + \epsilon \tag{25}$$

where f is a deterministic function; that is, if we evaluate f at a particular number, we get something that is not random. The term ϵ represents some source of noise **independent of X**, and is typically modeled with a Normal distribution

$$\epsilon \sim \mathrm{Normal}(0, \sigma_\epsilon). \tag{26}$$

In other words, it represents things other than X which may influence Y. Regardless of how X is distributed, for any given values $X = x$, Y must have a Normal distribution:

$$Y|(X = x) \sim \mathrm{Normal}(f(x), \sigma_\epsilon). \tag{27}$$

Of particular interest (due to its simplicity) is the case

$$f(x) = ax + b \tag{28}$$

which is the subject of this class. That is, we are interested in the model

$$Y = aX + b + \epsilon \tag{29}$$

where

$$\epsilon \sim \mathrm{Normal}(0, \sigma_\epsilon). \tag{30}$$

5.1. Working with regression. In Equation (29), X could be anything, but let's suppose X is drawn from a Normal distribution. This gives us the model

$$Y = aX + b + \epsilon \tag{31}$$

$$X \sim \mathrm{Normal}(\mu_x, \sigma_x) \tag{32}$$

(technically this is not a regression model anymore because we specify the distribution of X). Now that we have two random variables, we can ask questions like, what is the joint distribution? what are the conditional distributions? What are the Marginal distributions? Using properties of Normal random variables, we get

$$Y \sim \mathrm{Normal}(a\mu + b, |a|\sigma_x + \sigma_\epsilon) \tag{33}$$

This is the distribution of Y regardless of X; that is, it is the distribution we would get if we randomly sampled Y values ignoring what the value of X is. What is another name for this? This distribution can be understood visually [DRAW DIAGRAM ON BOARD].

X and Y are not independent. This can be seen visually. **Note: even though X and Y are not independent, we don't say they effect each other. To see why this terminology is problematic note that Y has no "effect" on X (it is determined by X). However, what is $X|Y = y$?**

Example 5. *conditioning with conditinous variables*

Exercise 5: *Conditioning in the regression model*

Exercise 6: *Kid's test scores*

5.2. Covariance and correlations coefficient. For a model like Equation (31), we know that X and Y are not independent, but it would be nice to have some way of quantifying how much Y depends on X and in which direction the relationship goes. To understand how this can be done, we start by considering the covariance:

$$\mathrm{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{34}$$

What is $\mathrm{cov}(X, X)$?

$$\mathrm{cov}(X, X) = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}\left[X^2 - \mathbb{E}[X]^2\right] = \mathrm{var}(X). \tag{35}$$

If X and Y are independent, the covariance will be zero, it is possible for the covariance to be zero without the variables being independent.

Just like standard deviation and mean, estimates of covariance are obtained by replacing $\mathbb{E}$ with the sample average: That is, if we have samples $(x_1, y_1), (x_2, y_2), \ldots,$

$$\mathbb{E}[XY] \to \frac{1}{N}\sum_i x_i y_i \tag{36}$$

In python, you can compute the covariance

```
> np.cov(x,y)[0,1]
```

The reason for the $[0, 1]$ is that the covariance function in numpy actually computes a 2D array (a Matrix), where the off diagonal entries are the covariance.

Example 6. *Zero covariance without independence*

However, **in a linear model** like Equation (29) we can show that

$$\mathrm{cov}(X, Y) = a\sigma_x^2 \tag{37}$$

This captures one aspect of the relation: if $\mathrm{cov}(X, Y)$ is positive then there is a positive association, for example. However, does this reflect how "related" X and Y are? It depends! One issue is that this seems not to depend on σ_ϵ. If σ_ϵ is HUGE (much larger than $|a|\sigma_x$), then $\mathrm{cov}(X, Y)$ could be arbitrarily large.

An alternative metric is

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_y \sigma_x} \tag{38}$$

What will this correspond to in the Linear model? We know from Equation (37) that

$$\rho(X, Y) = a\frac{\sigma_x}{\sigma_y} \tag{39}$$

and from before

$$\rho(X, Y) = \frac{a\sigma_x}{|a|\sigma_x + \sigma_\epsilon} \tag{40}$$

Exercise 7: *Covariance and correlation*

6. Additional exercises

The following exercise will help you practice the process of learning about a distribution by playing with simulations. You will also learn about the Beta distribution, which we will use later on.

Exercise 8: *Learning about a new distribution using simulations*

Just as it is important to understand where the Normal distributions comes from (this is what the central limit theorem tells us), it is important to understand what processes give rise to distributions which are not Normal.

Exercise 9: *Model of income*