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

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

Binomial samples can be generated in numpy with

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

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

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

Example 1. *Coefficient of variation*

Exercise 1: *Generating binomial samples*

Exercise 2: *Binomial election modeling*

Date: April 2022.

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

A uniform random variable, denoted

$$Y \sim \text{Uniform}(a, b) \tag{9}$$

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,

$$P(y_1 \leq Y \leq y_2) = \frac{y_2 - y_1}{L} \tag{10}$$

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

$$f(y)dy = P(y \leq Y \leq y + dy) = \frac{dy}{L}. \tag{11}$$

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. As an example, consider

$$Y \sim \text{Uniform}(0, 1) \tag{12}$$

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)
```

We want to describe this random variable in terms of a probability distribution, but just as for the uniform distribution, the probability for Y to equal any given value of Y is zero. For example, the chance that someone is six feet tall plus h inches is going to decrease as h gets very small, and the chance that someone is *exactly* 6 feet tall is zero (although it won't appear that way in data to to imprecision in our measurements). Therefore, instead of defining our model in terms of the chance that someone is exactly Y feet tall, we define it in terms of a density. For the Normal distribution, the density is

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

This is the classic bell curve shown in Figure 1. Again, we can think of f as the probability *per unit of the random variable*, e.g. probability/feet.

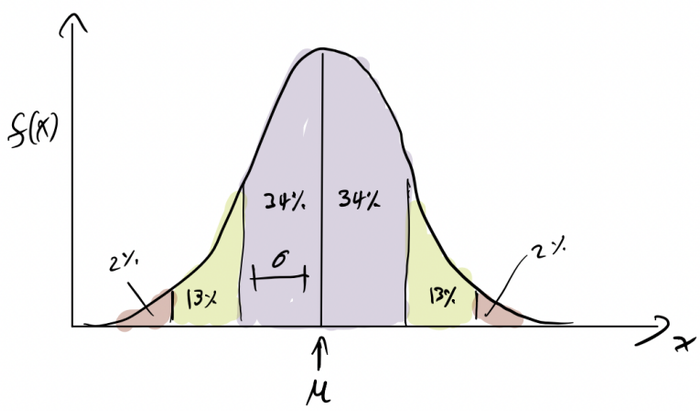

FIGURE 1. Probabilities in the Normal distribution

Exercise 4: *Comparing histograms*

Example 3. *Working with Normal random variables*

Exercise 5: *Hemoglobin levels*

3.1. **Properties of Normal random variables.** Going forward, you will need to know the following properties of Normally distributed random variables. Let

$$Y_1 \sim \text{Normal}(0,1), \quad Y_2 \sim \text{Normal}(0,1) \tag{17}$$

Then

$$aY_1 + b \sim \text{Normal}(b, |a|) \tag{18}$$

$$aY_1 + bY_2 \sim \text{Normal}(0, \sqrt{a^2 + b^2}) \tag{19}$$

We can derive more general relationships from these identities:

$$Y \sim \text{Normal}(\mu, \sigma) \tag{20}$$

then

$$aY + b \sim \text{Normal}(a\mu + b, |a|\sigma). \tag{21}$$

For example, if I say

$$Y \sim \text{Normal}(1,3) \tag{22}$$

you should recognize that

$$-4Y + 1 \sim \text{Normal}(-3, 4 \times 3). \tag{23}$$

3.2. **The central limit theorem.** We now return to the connection to sums of random variables and the Binomial distribution. The **central limit theorem** tells us that for a set of random variables $y_1, \ldots, y_N$ with each have $\mathbb{E}[y_i] = \mu_y$ and $\mathrm{var}(y_i) = \sigma_y^2$,

$$Y = \sum_i^N y_i \tag{24}$$

is approximately Normal with mean μ and variance $\sigma/\sqrt{n}$. In other words Normal random variables emerge when we add up many small sources of randomness.

4. Additional exercises

The following exercise will help you practice the process of learning about a distribution by playing with simulations.

Exercise 6: *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 7: *Model of income*