4452 Mathematical Modeling Lecture 14: Discrete and Continuous Probability

Introduction

If you have taken mathematical statistics, then you have seen all this material before. I will simply collect all the major ideas from probability that we will require.

First of all, there is the idea of *populations* and *samples.* A population is the set of all the things you are measuring (the standard example being something like the height of all men aged 21). A sample is a subset of this population. Typically, people will work with samples, since they cannot measure the entire population. This variable you are measuring can either be discrete or continuous, and is called a *random variable.* Typically it is denoted by a capital letter, such as X.

If you measure something, you can get a distribution. A distribution can be used to produce a histogram, which is a plot of the frequency of the data. For example, say you are interested in knowing something about the size of lab rats used in university experiments.

The population would be all such rats, which would be (virtually) impossible for you to collect data on. I suppose you could, if you made everyone who works with such rats send you their data–but is that a good use of everyone's time? Do you really want to analyze all that data?

Instead, let's say you decide to work with a sample of lab rats. You pick two universities in your vicinity, and spend an afternoon and evening weighing lab rats. The weight of lab rats over your sample space would be a continuous random variable $X \in (0, \infty)$. You collect a healthy amount of data, which I won't reproduce here (I will create a sample of the data in *Mathematica* so we can create some plots). How do you want to represent the data? A table would certainly contain the data, but a table would tell the person looking at it virtually nothing. A better way would be in a *histogram.* A histogram takes data sets, and plots the frequency of occurrence of data in data ranges as a bar graph. The data ranges are typically called *classes. Mathematica* can do this easily:

```
<< Graphics`Graphics`
Histogram[data, HistogramCategories -> 50, Frame -> True,
   FrameLabel -> {"Weight of Rat", "Frequency"}]
```

where "data" is a fictitious data set I created. The histograms are included in Fig. 1. The second histogram in Fig. 1 has smaller class sizes, or a larger number of classes. Therefore, there are fewer numbers from the data set which inhabit each class. This is why the numbers on the vertical axis are smaller. If we added up the number for the height of each class, we would get the total number of measurements we recorded.

The histograms give a much better way of representing data than a data table. From the histograms, we can see that the majority of rats seem to weigh around 942 units (I have no idea what a rat weighs. . . would 942 g seem reasonable?)

Choosing the classes is important. If you make the data range for each class small enough, you will get one measured data point in each class, which is about as useful as a table of values (not very). If you choose the class size too large, important variations in the data can be obscured. Whenever you use a histogram, you should not simply use the default class size–you should spend some time figuring out what best represents the data. It is your job as researcher/reporter/investigator to find the best way to easily convey the information you have learned.

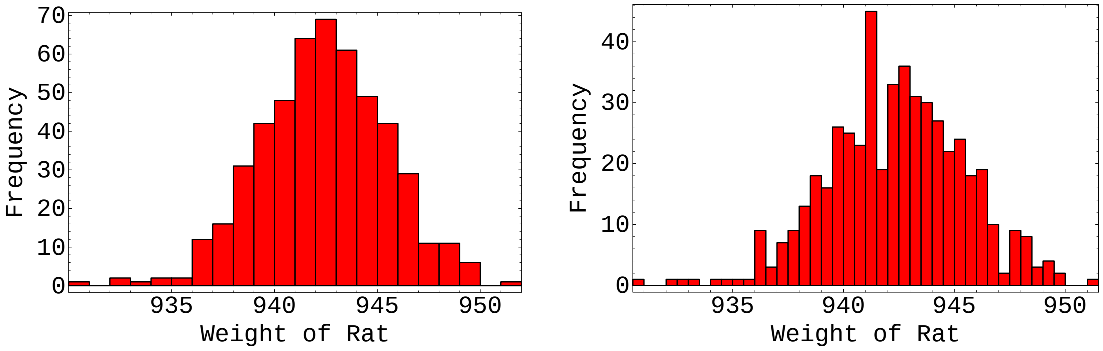

Figure 1: Two histograms representing the data that was (supposedly) collected by weighing lab rats.

We could scale our histogram so that the heights of the bars (which was the frequency) approximates the probability of a rat having a weight in that class. In this case, the area under all of the bars will be one. You should note that this is the approximate probability based on a rat taken from our sample, which hopefully is a good representation of the general population of rats which our sample models.

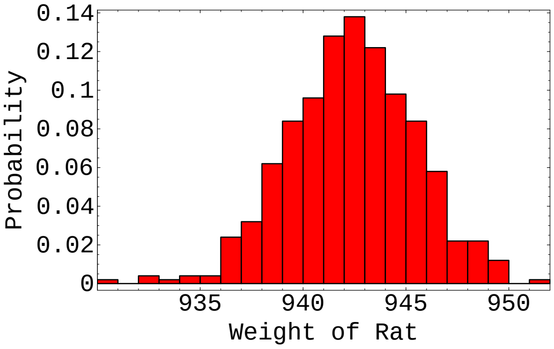

Figure 2: A scaled histogram for which the height of the bars represents the approximate probability that the weight of a rate will fall in that class. For example, the probability that a rat will have a weight between 940 and 941 g is about 0.11.

There is an underlying functional distribution contained in the histogram in Fig. 2. If we had an infinite number of rats in our sample size, and decreased the class size, we could draw a smooth, continuous curve, $y = f(x)$, through the top of the bars in the histogram. This curve is called a *probability density function*, and would have the property that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Let's turn our attention to the important concepts in discrete and continuous probability distributions.

Probability Distributions

Discrete

A discrete random variable is a variable that can assume a countable number of values. That is, the values of the random variable come from the integers. Examples of such a variable would be the number of people waiting in line at a bank, the number of defective compact discs in a group of 100 drawn from a production line, or the number of robberies in a police district in a year.

The probability distribution associated with a discrete random variable $X \in \{x_1, x_2, x_3, \ldots\}$ is denoted p_i and is the probability that X equals x_i, $Pr\{X = x_i\} = p_i$. The probability distribution must satisfy

$$\sum_i p_i = 1 \quad \text{and} \quad 0 \leq p_i \leq 1 \, \forall \, i.$$

The *mean*, or *average*, or *expected value* of X is given by

$$E(X) = \sum_i x_i p_i.$$

The *variance* measures the extent which X deviates from the mean (the average square distance from the sample mean), and is given by

$$V(X) = \sum_i (x_i - EX)^2 p_i$$

The probability that $a \leq X \leq b$ is given by

$$Pr(a \leq X \leq b) = \sum_{i=a}^{b} p_i.$$

Example Consider tossing a coin three times and recording the number of heads that occurs. The discrete random variable X is the number of heads. It can take on the values $X \in \{0, 1, 2, 3\}$.

I did this experiment 10 times, and my data are contained in Table 1 and Fig. 3. Since the data is recorded as frequency data, I want to let *Mathematica* know this when I create my histograms. I also want the class label to be centered under the class. I do this with the command

```
Histogram[data2, FrequencyData -> True, Frame -> True,
  FrameLabel -> {"Number of Heads", "Frequency"},
  FrameTicks -> {{{0.5, "0"}, {1.5, "1"}, {2.5, "2"}, {3.5, "3"}},Automatic, {}, {}}]
```

We can calculate the expected value and variance of the sample,

$$E(X) = \sum_{i=1}^{4} x_i p_i = 0.9,$$

$$V(X) = \sum_{i=1}^{4} (x_i - EX)^2 p_i = 0.49.$$

This does not agree with what we would expect if we had an equal likelihood of getting a head or tail with each toss, which would be a mean of 1.5 heads. This is because I have only done 10 trials, which is not very many. Also, the coin I used may be unbalanced, resulting in more tails than heads. Before you could say the coin is unbalanced, you would need to run many more trials. *Mathematica* can help you do this, if you wish to pursue it. I have included the results for a large number of trials in Fig. 4 for the coin toss game with three tosses and four tosses.

i	Number of Heads, $x_i = i - 1$	Frequency	Probability, p_i
1	0	3	0.3
2	1	5	0.5
3	2	2	0.2
4	3	0	0.0

Table 1: The frequency and probability distribution for tossing a coin three times and recording the number of heads. The experiment was performed $n = 10$ times.

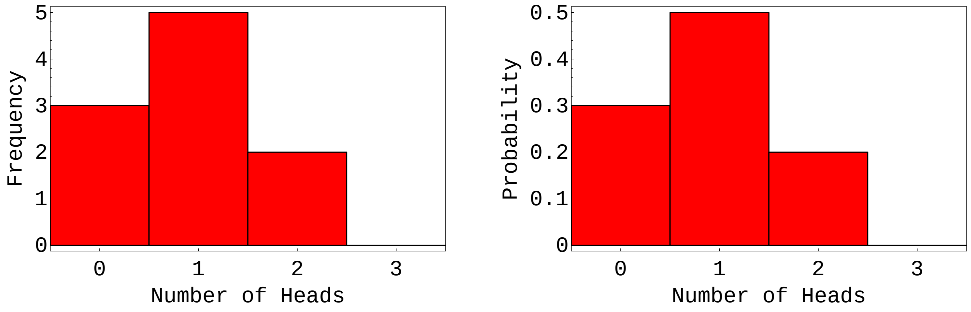

Figure 3: The frequency (left) and probability (right) distributions for tossing a coin three times and recording the number of heads. The experiment was performed $n = 10$ times.

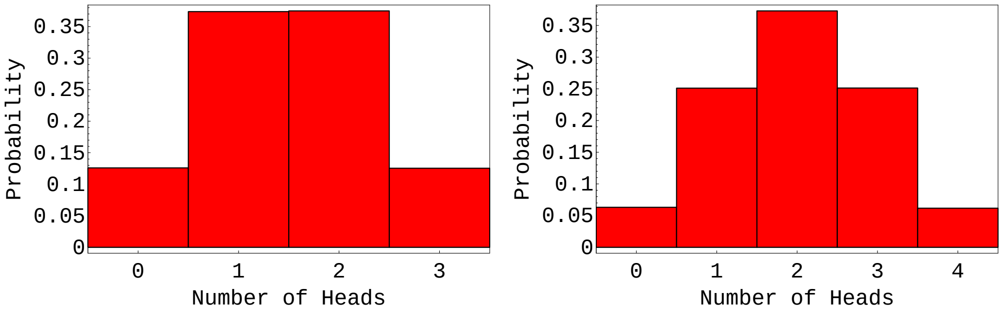

Figure 4: The probability distributions for tossing a coin three times and recording the number of heads (left) and tossing a coin four times and recording the number of heads (right). The experiment was performed $n = 100000$ times in each case.

Continuous

A continuous random variable is a variable that can assume an infinitely large number of values corresponding to the points on a line interval. That is, the values of the random variable come from the reals. Examples of such a variable would be the weight of lab rats, the length of time it takes to answer a phone at an office, the height of trees in a forest.

The probability distribution for a continuous random variable is given by $F(x)$ where

$$F(x) = Pr(X \le x).$$

The probability density function associated with a continuous random variable $X \in D$ is denoted $f(x) = F'(x)$. Typically, $D = (-\infty, \infty)$ or $(0, \infty)$. The probability distribution must satisfy

$$\int_D f(x) = 1 \quad \text{and} \quad f(x) \ge 0 \,\forall\, x \in D.$$

The *mean*, or *average*, or *expected value* of X is given by

$$E(X) = \int_D x f(x)\, dx.$$

The *variance* measures the extent which X deviates from the mean, and is given by

$$V(X) = \int_D (x - EX)^2 f(x)\, dx.$$

The probability that $a \le X \le b$ is given by

$$Pr(a \le X \le b) = \int_a^b f(x)\, dx.$$

Example Consider the distribution function given by

$$f(x) = 4x^2 e^{-2x}, \;\; D = \{x \ge 0\}. \tag{1}$$

This distribution function is used in quantum chemistry, to describe the location of an electron around the nucleus in a hydrogen atom.

The average value of this distribution is given by

$$E(X) = \int_D x f(x)\, dx = \int_0^\infty 4x^3 e^{-2x}\, dx = \frac{3}{2}.$$

The variance is given by

$$V(X) = \int_D (x - E(X))^2 f(x)\, dx = \int_0^\infty 4\left(x - \frac{3}{2}\right)^2 x^2 e^{-2x}\, dx = \frac{3}{4}.$$

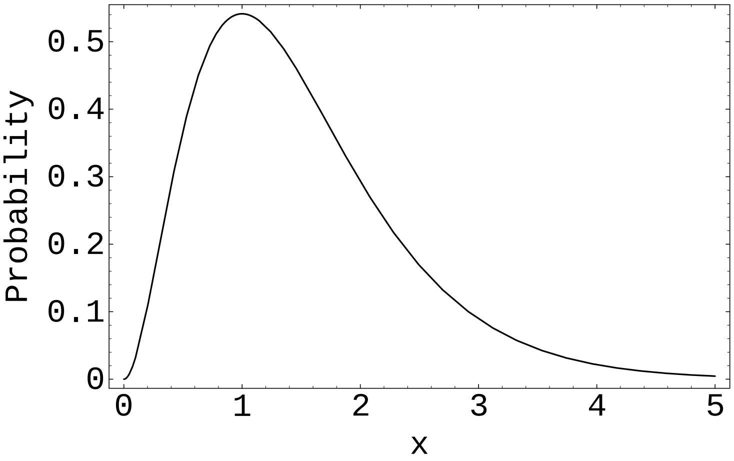

Figure 5: The probability distribution in Eq. (1) .

Expectation Values

If we want the expected value of some quantity which depends on the random variable X, we can calculate it in the following manner:

$$E(A(X)) = \int_D A(x) f(x)\, dx.$$

The variance is defined as

$$\begin{aligned} V(X) &= \int_D (x - E(X))^2 f(x)\, dx \\ &= E((X - E(X))^2) \end{aligned}$$

For random variables X and Y, we can say that the operation E is *linear* if

$$E(aX + bY) = aE(X) + bE(Y).$$

In this case, we can write the variance as

$$\begin{aligned} V(X) &= E(X - E(X))^2) \\ &= E(X^2 + E(X)^2 - 2XE(X)) \\ &= E(X^2) + E(E(X)^2) - 2E(XE(X)) \\ &= E(X^2) + E(X)^2 - 2E(X)E(X) \\ &= E(X^2) + E(X)^2 - 2E(X)^2 \\ &= E(X)^2 - E(X)^2 \end{aligned}$$

Useful Distributions

Distributions which occur frequently in the simulations are the uniform, the normal, and the binomial distribution. The following shows how to generate these distributions in *Mathematica*.

The Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{a-b} & a < x < b, \\ 0 & \text{elsewhere.} \end{cases}$$

```
a = 2
b = 3
f[x_] = If[x > b, 0, UnitStep[x - a]]
```

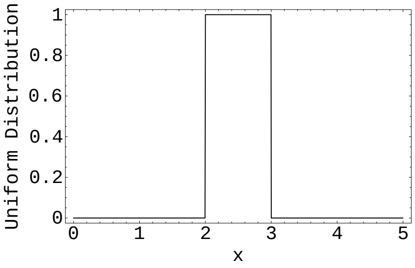

Figure 6: The uniform probability distribution with $a = 2$, $b = 3$.

The Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\,\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\,\sigma^2}}$$

```
<< Statistics‘NormalDistribution‘
ndist = NormalDistribution[12, Pi]
f[x_] = PDF[ndist, x]
```

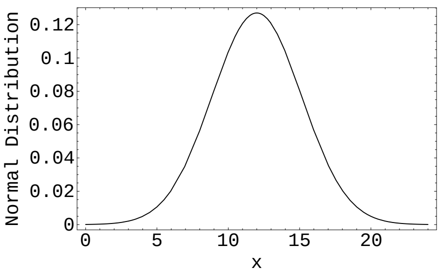

Figure 7: The normal probability distribution with $\mu = 12$, $\sigma = \pi$.

The Binomial Distribution

$$f(k) = \frac{n!}{(n-k)!\,k!} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, 3, \ldots, n.$$

where n is the number of trials and p is the probability of success of each trial.

```
<< Statistics‘DiscreteDistributions‘
bdist = BinomialDistribution[34, 0.3]
f[k_] = PDF[bdist, k]
list = Table[{k, f[k]}, {k, 0, 34}]
```

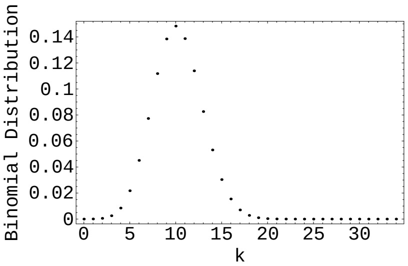

Figure 8: The binomial distribution with $n = 34$, $p = 0.3$.

References

[1] W. Mendenhall, Introduction to Probability and Statistics, Prindle, Weber, & Schmidt (Boston) 1983.

[2] D. D. Mooney & R. J. Swift, A Course In Mathematical Modeling, The Mathematical Association of America, 1999.

[3] M. Meerschaert, Mathematical Modelling, 2nd ed., Academic Press (San Diego) 1999.