FIT2086 Studio 2
Probability Distributions and Random Variables

Daniel F. Schmidt
Based in part on material prepared by David L. Dowe

August 3, 2021

Contents

1 Introduction **2**

2 Binomial distribution **2**

3 Gaussian Distribution **3**

4 Poisson Distribution **4**

5 The Uniform Distribution **5**

6 Using R and Simulation to Explore the Weak Law of Large Numbers **7**

7 Appendix I: Standard Normal Distribution Table **9**

1 Introduction

These studio notes introduce you to some problems and ideas regarding random variables and probability distributions. You should complete the questions using either R, or by hand as appropriate.

For each of the distributions we will be looking at there are four associated functions in R: the density function, the (cumulative) distribution function, the quantile function and a function to generate realisations of random variables. For example, for the binomial distribution they are:

- `dbinom()`: probability function;
- `pbinom()`: cumulative distribution function;
- `qbinom()`: the quantile function;
- `rbinom()`: the function to generate random variables.

Use the R command `?dbinom` to bring up the help information for all these functions.

2 Binomial distribution

The binomial distribution is given by:

$$\mathbb{P}(m \,|\, \theta) = \binom{n}{m} \theta^m (1-\theta)^{n-m} \tag{1}$$

It models the number of successes, m, occuring in a sequence of n trials, with the probability of a success for each trial being θ. That is, if $X_1, \ldots, X_n \sim \text{Be}(\theta)$ (i.e., $X_1, \ldots, X_n$ are Bernoulli variates) then $m = \sum_{i=1}^{n} X_i$.

1. In equation (1), how should you interpret $(1-\theta)^{n-m}$?
2. How should you interpret θ^m?
3. What is $\binom{n}{m}$?
4. How can you interpret $\theta^m(1-\theta)^{n-m}$?
5. What is
$$\sum_{k=i}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \,?$$
6. Using equation (1), compute the probability of obtaining at least two heads in four coin tosses if $\theta = 1/2$ (a "fair" coin).
7. We denote a binomial distribution with success probability θ and number of trials n by $\text{Bin}(n, \theta)$. Which of the following are (potentially) binomial experiments? Why or why not?
   (a) US Presidential elections?
   (b) Shuttle launches?
   (c) Football matches?
   (d) Depth of the Yarra river at a random point?
   (e) Rolls of a six-sided die?

8. Suppose that you sample a sequence of 10 Bernoulli random variables, each with success probability of $\theta = 1/2$. What is the probability of:
   (a) Seeing the sequence $(0, 0, 1, 0, 1, 1, 0, 0, 1, 1)$?
   (b) Seeing the sequence $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$?
   (c) Seeing five or more ones in total? *(hint: use the* `pbinom()` *function)*

3 Gaussian Distribution

The Gaussian (normal) distribution is one of the most important distributions in data science. It is defined for continuous data. The probability density function for a normal is

$$p(x \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2}$$

where μ is the mean, σ^2 is the variance, and σ is therefore the standard deviation. Recall that we use $N(\mu, \sigma^2)$ to denote a normal distribution.

1. Answer the following questions regarding the normal distribution. (Use the `pnorm`() function (i.e., the normal cumulative distribution function) in R as appropriate.)
   (a) What is the probability that a random variable from $N(\mu, \sigma^2)$ lies within one standard deviation of μ?
   (b) What is the probability that a random variable from $N(0, 1)$ is greater than 2?
   (c) What is the probability that a random variable from $N(0, 4^2)$ is greater than 2?
2. Which of the following could be normally distributed?
   (a) A coin toss?
   (b) A dice roll?
   (c) The height of adults?
   (d) Number of phone calls received by a call centre in one hour?
   (e) Measurement error when measuring the velocity of a car.
3. The standard normal, also known as the unit normal, is N(0,1). Any normal variate $X \sim N(\mu, \sigma^2)$ can be converted into a unit normal ("standardised") by using

   $$Z = (X - \mu)/\sigma$$

   that is, we take a value from $X \sim N(\mu, \sigma^2)$, find its distance from the mean and convert that distance into standard deviation units. This allows us to use Z (called a "z-score") to look up probabilities using a single table for the standard normal.

   Use a z-table (Appendix I of this document) to find probabilities for $X \sim N(3, 16)$ distribution:
   (a) $\mathbb{P}(X < 5)$?
   (b) $\mathbb{P}(X > -4)$?
   (c) $\mathbb{P}(2 < X < 7)$?

   (See Example 5.5a, pp. 172-173 of Ross; (b) is a minor variation).

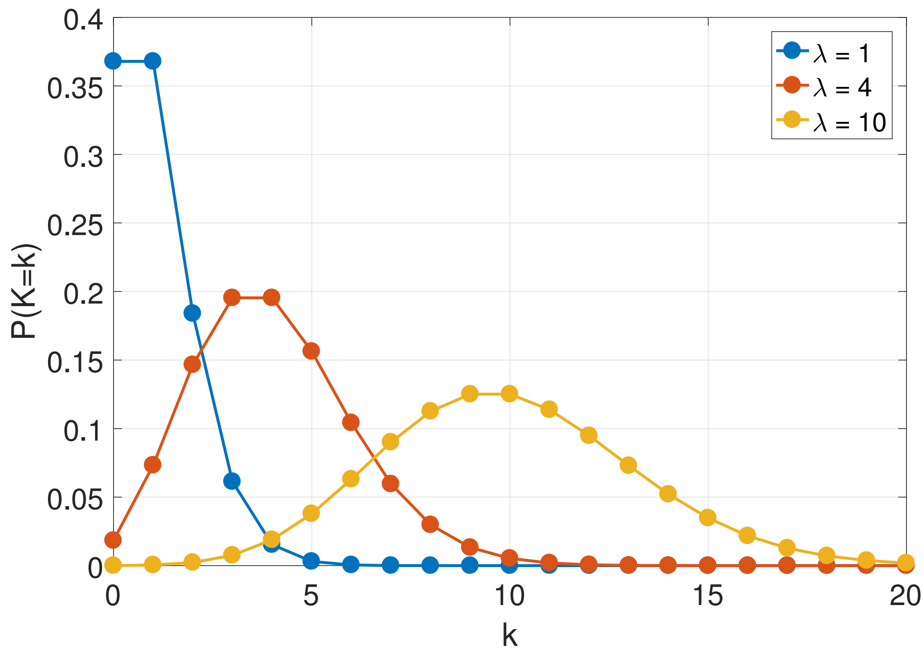

Figure 1: Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

4 Poisson Distribution

The Poisson distribution is used to model count (non-negative integer) data. The Poisson probability distribution for rate λ is given by

$$\mathbb{P}(X = k \mid \lambda) \equiv p(k \mid \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

where $k!$ is the factorial function.

1. What are the expected values of the three Poisson distributions shown in Figure 1?

2. Use the `exp()` and `factorial()` functions in R to compute the following probabilities for $\lambda = 4$:
   (a) $\mathbb{P}(X = 1)$;
   (b) $\mathbb{P}(X = 2)$;
   (c) $\mathbb{P}(X < 2)$.

3. Use the `ppois()` and `dpois()` functions in R to compute the following for $\lambda = 4$:
   (a) $\mathbb{P}(X = 1)$;
   (b) $\mathbb{P}(X = 2)$;
   (c) $\mathbb{P}(X < 2)$;
   (d) $\mathbb{P}(X > 5)$.

4. Which of the following are likely Poisson processes? If not Poisson, what might be a better distribution to use to model them?
   (a) The number of people shopping at a market during the day?
   (b) Meteorites striking land versus water?
   (c) The number of heart attacks in a month?
   (d) Populations of cities?
   (e) Average weights of women?
   (f) Number of workplace accidents per week?

5. The Poisson distribution has the following useful property. If

$$X_1 \sim \text{Poi}(\lambda_1),\ X_2 \sim \text{Poi}(\lambda_2),\ \ldots,\ X_m \sim \text{Poi}(\lambda_m),$$

   then

$$\sum_{i=1}^{m} X_i \sim \text{Poi}\left(\sum_{i=1}^{m} \lambda_i\right). \tag{3}$$

   Suppose the average number of heart attack patients seen by a hospital is 6 per week. Using a Poisson model, and property (3), answer the following questions:
   (a) What is the probability the hospital will have 2 or fewer heart attack patients in a week?
   (b) What is the probability it will see one heart attack patient on any given day (assuming they are independent of day of the week)?
   (c) What is the probability it will see *at least* one heart attack patient on any given day (assuming they are independent of day of the week)?

5 The Uniform Distribution

The uniform distribution is an interesting distribution as it can be used to model continuous, discrete numerical and categorical data. A continous RV X is said to follow a uniform distribution $U(a, b)$, with $-\infty < a < b < \infty$, if

$$p(x \,|\, a, b) = 1/(b - a),\ x \in [a, b] \tag{4}$$

A discrete random variable X over the set $\mathcal{X}$ is said to follow a uniform distribution if

$$\mathbb{P}(X = x) = 1/|\mathcal{X}|$$

where $|\mathcal{X}|$ denotes the number of items in the set $\mathcal{X}$. Clearly, a discrete uniform distribution is only defined if the number of different items the RV X can assume is finite.

1. For $X \sim U(a, b)$, with $a, b > 0$, what is:
   (a) $\mathbb{P}(X > 2b)$?
   (b) $\mathbb{P}(X < (a + b)/2)$?
   (c) $\mathbb{P}(X \in [(a + 3b)/4, b])$?

   (d) $\mathbb{P}(X \in [0, a])$?

2. Which of the following are uniformly distributed?

   (a) A coin toss?

   (b) A roll of a six-sided dice?

   (c) Heights of adult female humans?

   (d) Daily temperature in Belgrade, Serbia?

3. If $X \sim U(0, b)$, what is $\mathbb{E}[X]$ (i.e., the mean of the uniform distribution)?

4. If X is the outcome of rolling a six-sided die, and assuming that the die is fair (i.e., not biased to any side), what is:

   - The average value of X?
   - If Y is the outcome of rolling the die again, what is the average value of $X + Y$ (i.e., the sum of two six-sided dice rolls)?

6 Using R and Simulation to Explore the Weak Law of Large Numbers

Please complete this exercise out of your studio. One of the main advantages of using R and having access to a computer is the ability to explore complicated problems, and gain an understanding of basic ideas, by using simulation. R can generate realisations from most of the common random variables we will be studying. For example, let's consider the binomial distribution. The function `rbinom(n, size, prob)` generates `n` realisations from a Bin(`size`, `prob`) distribution. For example, type

$$\texttt{rbinom}(10, 5, 0.25)$$

which will generate 10 realisations from a $\text{Bin}(5, 0.25)$ distribution. An important special case is `rbinom(n,size=1,prob)`, which is the Bernoulli distribution. Note, a little confusingly in R, the `rbinom()` function uses the argument `n` to denote the *number of random variables to generate*, and `size` to denote the n parameter in the usual definition of the Binomial distribution. For example, type

$$\texttt{x} = \texttt{rbinom}(25, 1, 0.75)$$

which will store a sequence of 25 random `0`s and `1`s in `x`, with the probability of a `1` being 0.75. Lets now look at how we can use simulation to explore a few different aspects of probability.

For this question will write some simulation code to explore the convergence of the sample mean $\bar{x}$ to the population mean for several different distributions. By convergence, we mean that as the sample size n of a sample $\mathbf{x} = (x_1, \ldots, x_n)$ grows, the sample mean $\bar{x}$ will get closer and closer to the "true" mean of the distribution from which the sample comes. This question will introduce you to the power of using computer simulations to explore and understand statistical problems that has only become available to students and researchers in the last 20 to 30 years.

1. Let's begin by creating an R function that takes a vector of n samples, and returns the running mean for all $j = 1, \ldots, n$; the running mean is given by

   $$\bar{x}_j = \frac{1}{j} \sum_{i=1}^{j} x_i.$$

   That is, our function should return the vector $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$. This can be efficiently implemented so that the function only takes $O(n)$ (that is, linear in the number of samples) time steps to run. Use the code in `studio2.skel.R` as a basis to get started.

2. Now that we our function from step (1), we can use it to explore convergence of the sample means for several different distributions. To do this, generate $n = 1,000$ random samples from a normal distribution with standard deviation $\sigma = 1$, and a mean of zero. Remember that we can use the notation $X_i \sim N(0, 1)$, $i = 1, \ldots, n$ as shorthand for this. You can do this using the code

   ```
   x = rnorm(n, mean = 0, sd = 1)
   ```

   Then plot the running means from this sample on the y-axis, against the number of samples (1 through to n) on the x-axis. Also plot the population mean as a function of the number of samples (i.e., a constant line) to get a reference. Try this for several samples – they should all exhibit convergence behaviour to the constant line. You can use the `plot()` function to plot the first line, then the `lines()` function to add additional plots over the top – see the skeleton R file `studio2.skel.R` for Studio 2 for examples.

3. Repeat step (2), but this time using data from a $1,000$ Bernoulli variables with $\theta = 1/2$, and also with $\theta = 0.9$, using the `rbinom()` function as described above. Overlay these curves on the same plot, using the `ylim` $=$ `c(0, 1)` option to set the y-axis to $(0, 1)$ (i.e., the allowable parameter space for a Bernoulli). How do the two convergence curves differ? Try several examples to see the difference in the curves. Why are they different in behaviour?

7 Appendix I: Standard Normal Distribution Table

$\|z\|$	$\mathbb{P}(Z < -\|z\|)$	$\mathbb{P}(Z < \|z\|)$	$\|z\|$	$\mathbb{P}(Z < -\|z\|)$	$\mathbb{P}(Z < \|z\|)$
0.000	0.500000	0.500000	2.047	0.020353	0.979647
0.093	0.462943	0.537057	2.140	0.016196	0.983804
0.186	0.426204	0.573796	2.233	0.012789	0.987211
0.279	0.390096	0.609904	2.326	0.010020	0.989980
0.372	0.354912	0.645088	2.419	0.007790	0.992210
0.465	0.320924	0.679076	2.512	0.006009	0.993991
0.558	0.288375	0.711625	2.605	0.004598	0.995402
0.651	0.257471	0.742529	2.698	0.003491	0.996509
0.744	0.228382	0.771618	2.791	0.002630	0.997370
0.837	0.201237	0.798763	2.884	0.001965	0.998035
0.930	0.176125	0.823875	2.977	0.001457	0.998543
1.023	0.153093	0.846907	3.070	0.001071	0.998929
1.116	0.132151	0.867849	3.163	0.000781	0.999219
1.209	0.113273	0.886727	3.256	0.000565	0.999435
1.302	0.096403	0.903597	3.349	0.000406	0.999594
1.395	0.081455	0.918545	3.442	0.000289	0.999711
1.488	0.068326	0.931674	3.535	0.000204	0.999796
1.581	0.056894	0.943106	3.628	0.000143	0.999857
1.674	0.047024	0.952976	3.721	0.000099	0.999901
1.767	0.038577	0.961423	3.814	0.000068	0.999932
1.860	0.031410	0.968590	3.907	0.000047	0.999953
1.953	0.025381	0.974619	> 4.000	< 0.000032	> 0.999968

Table 1: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0, 1)$