FIT2086 Studio 2
Probability Distributions and Random Variables

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

August 1, 2017

Contents

1 Introduction **2**

2 Binomial distribution **2**

3 Gaussian Distribution **3**

4 Poisson Distribution **5**

5 The Uniform Distribution **7**

6 Using R and Simulation to Explore Probability **9**
6.1 Weak Law of Large Numbers 9
6.2 Calculating Probabilites of Complex Events 11

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 (see Ross Section 5.1) 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 $(1-\theta)^{n-m}$?
   A: As the probability of seeing $(n-m)$ failures; i.e., the probability of our random event not occuring $(n-m)$ times.

2. How should you interpret θ^m?
   A: As the probability of seeing m successes.

3. What is $\binom{n}{m}$?
   A: The number of different ways to order m successes in $(n-m)$ failures.

4. How can you interpret $\theta^m(1-\theta)^{n-m}$?
   A: It is the probability of seeing m successes and $(n-m)$ failures in a *particular* ordering, chosen from $\binom{n}{m}$ possible orderings.

5. What is
   $$\sum_{k=i}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \,?$$
   A: It is the probability that we will see i or more successes from n Bernoulli trials, irrespective of ordering.

6. Using equation (1), compute the probability of obtaining at least two heads in four coin tosses if $\theta = 1/2$ (a "fair" coin).

A: From the above formula, we have

$$\begin{aligned}\mathbb{P}(m=2\,|\,\theta=1/2,n=4) &= \binom{4}{2}\frac{1}{2}^2\frac{1}{2}^2+\binom{4}{3}\frac{1}{2}^3\frac{1}{2}+\binom{4}{4}\frac{1}{2}^4\frac{1}{2}^0\\ &= (6+4+1)\frac{1}{2}^4=11/16\end{aligned}$$

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? **No, unless there is only two candidates.**

   (b) Shuttle launches? **Yes, if we treat a successful launch as a success.**

   (c) Football matches? **No, as draws give this three possible outcomes – though we could redefine the outcomes so that we are interested in either win vs. draw/lose (did they win?) or lose vs win/draw (did they lose?)**

   (d) Depth of the Yarra river at a random point? **No, this is a continuous measurement.**

   (e) Rolls of a six-sided die? **No, there are six possible outcomes.**

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 $(0,0,1,0,1,1,0,0,1,1)$?
   A: With probability $\theta = 1/2$, all sequences have equal probability of $\theta^n = \frac{1}{2}^{10} = 1/1024$.

   (b) Seeing the sequence $(0,0,0,0,0,0,0,0,0,0)$?
   A: Same as above – $1/1024$.

   (c) Seeing five or more ones in total? *(hint: use the* `pbinom()` *function)*
   A: `1 - pbinom(4,10,1/2)` from $\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4)$.

3 Gaussian Distribution

The Gaussian (normal) distribution (see Ross, Section 5.5) is one of the most important distributions in data science. It is defined for continuous data. The probability density function for a normal is

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

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?

A: **Simple answer**: approximately ≈ 0.6825, from the rules regarding normal distributions (see Lecture 2).
Long answer. This is equivalent to finding

$$\mathbb{P}(\mu - \sigma < X < \mu + \sigma) = \mathbb{P}(X < \mu + \sigma) - \mathbb{P}(X < \mu - \sigma)$$

if $X \sim N(\mu, \sigma^2)$, which is equivalent to

$$\int_{\mu-\sigma}^{\mu+\sigma} p(x \,|\, \mu, \sigma^2)dx = \int_{-\infty}^{\mu+\sigma} p(x \,|\, \mu, \sigma^2)dx - \int_{-\infty}^{\mu-\sigma} p(x \,|\, \mu, \sigma^2)dx. \tag{3}$$

In R, we can find that using

$$\texttt{pnorm}(\texttt{mu} + \texttt{sigma}, \texttt{mu}, \texttt{sigma}) - \texttt{pnorm}(\texttt{mu} - \texttt{sigma}, \texttt{mu}, \texttt{sigma})$$

which, if you try for different values of `mu` and `sigma` you will see is always ≈ 0.6825. Why is this? It is because of the self-similarity property of normals. As all normal distributions are scaled and shifted versions of unit normal $N(0, 1)$, we can calculate the probabilities for any $N(\mu, \sigma^2)$ from the $N(0, 1)$ by rescaling our numbers of interest to z-scores. A z-score is a standardised difference from the mean, and the z-score for $X \sim N(\mu, \sigma^2)$ is

$$Z_X = (X - \mu)/\sigma.$$

So, if we are calculating probabilities of the form

$$\mathbb{P}(X < \mu + k\sigma)$$

for an $N(\mu, \sigma^2)$ distribution, the appropriate z-score will be

$$Z_{\mu+k\sigma} = ((\mu + k\sigma) - \mu)/\sigma = k.$$

which does not depend on μ and σ. So the z-score tells you how far away, in standard deviation units the quantity is from the mean. So it is clear that, irrespective of the values of μ and σ, the probability (3) is equivalent to

$$\int_{\mu-\sigma}^{\mu+\sigma} p(x \,|\, \mu, \sigma^2)dx = \int_{-1}^{1} p(x \,|\, 0, 1)dx$$

which is ≈ 0.6825. This means that if we generated many samples from $N(\mu, \sigma^2)$, then approximately 68.25% of them would have values between $\mu - \sigma$ and $\mu + \sigma$.

(b) What is the probability that a random variable from $N(0, 1)$ is greater than 2?
A: `1 - pnorm(2,0,1)` ≈ 0.0227

(c) What is the probability that a random variable from $N(0, 4)$ is greater than 2?
A: `1 - pnorm(2,0,4)` ≈ 0.308. As an aside, from the above discussion on z-scores this is the same as `1 - pnorm((2-0)/4,0,1)` as $Z_2 = (2 - 0)/4$.

2. Which of the following could be normally distributed?

(a) A coin toss? **No, discrete binary variable**.

(b) A dice roll? **No, discrete with small number of distinct values (six)**.

(c) The height of adults? **Yes**.

(d) Number of phone calls received by a call centre in one hour? **No, discrete – though could be approximately normal if number of calls is very large.**

(e) Measurement error when measuring the velocity of a car. **Yes.**

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 (downloadable from Moodle) 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).

4 Poisson Distribution

The Poisson distribution (see Ross, Section 5.2) 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. For the three Poisson distributions shown in Figure 1, what are their mean values?

   A: 1, 4 and 10 for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$, respectively.

2. Use the `exp()` and `factorial()` functions in R to compute the following probabilities for $\lambda = 4$:

   (a) $\mathbb{P}(X = 1)$;
   A: `4^1 * exp(-4) / factorial(1)`

   (b) $\mathbb{P}(X = 2)$;
   A: `4^2 * exp(-4) / factorial(2)`

   (c) $\mathbb{P}(X < 2)$.
   A: `4^0 * exp(-4) / factorial(0) + 4^1 * exp(-4) / factorial(1)`

3. Use the `ppois()` and `dpois()` functions in R to compute the following for $\lambda = 4$:

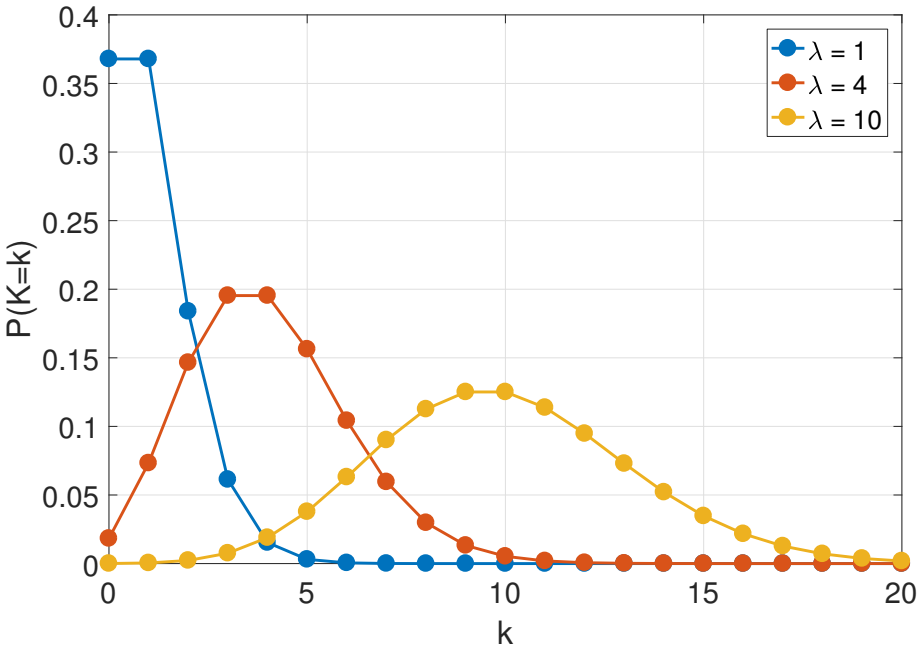

Figure 1: Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

(a) $\mathbb{P}(X = 1)$;
A: `dpois(1,4)`

(b) $\mathbb{P}(X = 2)$;
A: `dpois(2,4)`

(c) $\mathbb{P}(X < 2)$;
A: Either `dpois(0,4) + dpois(1,4)` or `ppois(1,4)` (using the cdf)

(d) $\mathbb{P}(X > 5)$.
A: `1 - ppois(5,4)`, i.e., $1 - \mathbb{P}(X \leq 5)$.

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? **Yes.**

   (b) Meteorites striking land versus water? **No, is a binary event, so binomial is appropriate.**

   (c) The number of heart attacks in a month? **Yes.**

   (d) Populations of cities? **Maybe – even though discrete, normal might be a better approximation.**

   (e) Average weights of women? **No, more likely to be normally distributed.**

(f) Number of workplace accidents per week? **Yes.**

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

$$\sum_{i=1}^{m} X_i \sim \text{Poi}\left(\sum_{i=1}^{m} \lambda_i\right). \tag{4}$$

Suppose the average number of heart attack patients seen by a hospital is 6 per week. Using a Poisson model, and property (4), answer the following questions:

(a) What is the probability the hospital will have 2 or fewer heart attack patients in a week?
A: `ppois(2,6)`

(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)?

A: This answer is a little trickier. From the above we know that the number of heart attacks in a week, say X, is equal to the sum of the number of heart attacks in each day of the week, i.e.,

$$X = X_1 + X_2 + X_3 + \cdots + X_7,$$

where X_i is the i-th day in the week. We are told that the rate of heart attacks on an individual day is independent of what day it is, which implies that the X_i are identically distributed. From property (4), and the fact that X is Poisson distributed, we know that

$$X = \sum_{i=1}^{7} X_i \sim \text{Poi}\left(\sum_{i=1}^{7} \lambda_i\right)$$

so that $\lambda = \sum_{i=1}^{7} \lambda_i = 6$, and $\lambda_1 = \lambda_2 = \cdots = \lambda_7$, which implies that $\lambda_i = 6/7$ as each day is identical. Then, the probability of seeing on heart attack on any given day is simply

$$\mathbb{P}(X = 1 \,|\, \lambda = 6/7) \approx 0.3637$$

(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)?
A: Using the above logic $\mathbb{P}(X \geq 1 \,|\, \lambda = 6/7) = 1 - \exp(-6/7) \approx 0.5756$.

5 The Uniform Distribution

The uniform distribution (see Ross, Section 5.4) 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

$$p(x \,|\, a, b) = 1/(b - a),\ x \in [a, b] \tag{5}$$

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) $\mathbb{P}(X > 2b)$?
   A: The RV X is only defined on $[a, b]$, so this probability is zero as $2b > b$.

   (b) $\mathbb{P}(X < (a + b)/2)$?
   A: The point $(a + b)/2$ is halfway between a and b, so the probability is $1/2$.

   (c) $\mathbb{P}(X \in [(a + 3b)/4, b])$?
   A: The point $(a + 3b)/4$ is three-quarters of the way between a and b, so the probability from $(a + 3b)/4$ to b is $1/4$.

   (d) $\mathbb{P}(X \in [0, a])$?
   A: The upper limit of the interval is a, which is the lower limit of the range on which X is defined, so the probability is zero.

   Long answer: We can also answer (b) and (c) by finding the cdf, i.e.,

   $$P(X < x)$$

   for the uniform distribution, where $x \in [a, b]$. This is given by

   $$\begin{aligned} P(X < x) &= \int_a^x p(x' \mid a, b) dx' \\ &= \frac{1}{b-a} \int_a^x (1) dx' \\ &= \frac{1}{b-a} [x']_a^x \\ &= \frac{x-a}{b-a} \end{aligned}$$

   Now to answer (b) calculate $P(X < (a + b)/2)$ and simplify to get $1/2$; to answer (c) calculate $1 - P(X < (a + 3b)/4)$ and simplify to get $1/4$.

2. Which of the following are uniformly distributed?

   (a) A coin toss? **Yes, if the coin is fair.**

   (b) A roll of a six-sided dice? **Yes, if the dice is fair.**

   (c) Heights of adult female humans? **No, likely to be normal.**

   (d) Daily temperature in Belgrade, Serbia? **No, unlikely to be uniformly distributed.**

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

A: We use the formula for expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^b x p(x \mid 0, b) dx \\
&= \int_0^b x \frac{1}{b} dx \\
&= \frac{1}{b} \int_0^b x dx \\
&= \frac{1}{b} \left[\frac{x^2}{2} \right]_0^b \\
&= \frac{b^2}{2b} = \frac{b}{2}
\end{aligned}$$

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?
     A: Each of the six sides has equal probability $1/6$, so
     $$\mathbb{E}[X] = (1/6)1 + (1/6)2 + (1/6)3 + (1/6)4 + (1/6)5 + (1/6)6 = 3.5$$
   - 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)?
     A: By independence $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 7$.

6 Using R and Simulation to Explore Probability

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.

6.1 Weak Law of Large Numbers

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. While for some of the simpler problems we will explore we can derive exact formulae, the real power of computer simulation approaches becomes apparent when dealing with very complex statistics that are mathematically intractable (i.e., very difficult).

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.

   A: See the file `studio2_solns.R` for code answers.

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 for Studio 2 for examples.

   A: See the file `studio2_solns.R` for code answers.

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?

   A: See the file `studio2_solns.R` for code answers. The difference between the curves for the two sequences of Bernoulli variables is obvious if you plot the two curves a number of times – the curve for data generated by $\theta = 0.9$ will be seen to generally fluctuate around $\theta = 0.9$ significantly less than the curve for $\theta = 0.5$, and will appear to converge quicker on the true value on average.

6.2 Calculating Probabilites of Complex Events

A powerful aspect of using a computer is to compute the probabilities of complex events. For example, imagine $X \sim N(0,1)$, and we want to know

$$\mathbb{P}\left(\log|X| > \sqrt{|X|}/4\right)?$$

In the past, to compute these probabilities would have either required complex mathematical manipulation, or simply been impossible. Now, by repeatedly generating random realisations we can find out how many times the event has occured, and use this as an estimate of the probability. Further, by the Weak Law of Large Numbers, we are (often) guaranteed that the larger the number of realisations we take, the more accurate our estimate will be. This approach is called the Monte Carlo method (after the famous casino in Monaco!).

For example, to answer the above question we can use the following code

```
x = rnorm(1e3, 0, 1)
mean(log(abs(x)) > sqrt(abs(x))/4)
```

The first line generates $1,000$ realisations from $N(0,1)$. The second line calculates the proportion of these samples that satisfy our condition. If you re-run this code several times, you will get a slightly different answer every time. This is because you are using random numbers to estimate the probability. If you increase the sample from `1e3` to `1e5`, the probability will vary less from run to run.

Using this idea, write a loop that repeatedly generates a sequence of 0s and 1s from 10 Bernoulli variables with $\theta = 1/2$, i.e., $X_1, \ldots, X_{10} \sim \text{Be}(1/2)$. Then, use this loop to estimate how likely it is to:

1. See a sequence which starts a 1 and ends with a 0 *(see if you can work out the probability of this without simulation)*

   A: See the file `studio2_solns.R` for code answers. You can also get the exact probability without simulation by noticing you are interested in all the patterns of 10 Bernoulli variables that look like:

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

   where X is either a 0 or a 1. There are 8 X's, so there are $2^8 = 256$ sequences that match this pattern. The probability is then $256/1024 = 1/4$.

2. See *at least one* sequence of five zeroes or five ones in succession? for example $(0,1,1,1,1,1,0,1,0,0)$

3. See five or more alternations (changes from 1 to 0, or 0 to 1)? For example: $(0,0,1,1,0,1,0,0,0,1)$

   A: See the file `studio2_solns.R` for code answers to the above two questions.

See the provided R skeleton file to get started on answering these questions.