**Reminder: To receive full credit for your homework, you should include as part of your solution a brief description of the problem that suffices to put the calculations you present or the results you provide in context.**

1. List the four conditions that indicate when the binomial distribution is an appropriate probability model for an application.

   Solution:

   **Binary outcomes:** two possible outcomes ("success", the outcome we count, and "failure", the one we don't).

   **Independent trials:** outcomes of the trials are mutually independent.

   $n$ **is fixed:** there is a fixed, predetermined sample size.

   **Same value of** $p$**:** the success probability on a single trial is the same for all trials.

2. For each of these examples, indicate why the random variable $X$ does *not* have a binomial distribution. (In other words, list which condition(s) from the previous problem are not met.)

   (a) A cage contains nine mice, three of which are female. $X$ is the number of female mice in a simple random sample of four mice from the cage.

   Solution: There is sampling without replacement, so the trials are not independent. This is especially important when the sample size is a substantial fraction of the population size.

   (b) In a botany experiment, a student plants forty seeds in separate pots. All experimental conditions are as identical as possible except that twenty of the pots are watered every day while the other twenty are watered every other day. $X$ is the total number of seeds that have germinated after one week.

   Solution: The probability of germination will be different in each of the two treatments. The numbers of seeds that germinate in each of the two treatment groups separately may be modeled as binomial, however.

   (c) In a particular region, previous studies have shown that thirty percent of the mushrooms are of a specific species of interest. A biologist randomly samples mushrooms (of all species) from this region until she has collected twenty specimens of the species of interest. $X$ is the total number of mushroom specimens sampled.

   Solution: There is not a fixed sample size, $n$. The possible values of the random variable are 20, 21, .... This cannot be binomial.

3. Exercise 3.16 (page 113). In the U.S., 42% of the people have blood type A. (In the class, by the way, 38% of students who knew were type A.) In a random sample of four people, let $Y$ be the number with blood type A. (This shows that even though there are *four* possible blood types, A, B, AB, and O, you can still use the binomial distribution by considering just A and *not A*.)

   Solution: I will show the R code for these computations.

```
R> dbinom(0:2, 4, 0.42)

[1] 0.1131650 0.3277882 0.3560458

R> pbinom(2, 4, 0.42)

[1] 0.7969989

R> pbinom(2, 4, 0.42) - dbinom(0, 4, 0.42)

[1] 0.6838339
```

   We get the same answers using the formula $\Pr\{Y = j\} = {}_nC_j p^j (1-p)^{n-j}$.

4. Exercise 3.18 (page 113).

   Solution: Shells of a species of land snail are either streaked (60%) or pallid (40%). In a sample of ten land snails, the probability that the sample proportion is 50%, 60%, or 70% are the probabilities that there are exactly 5, 6, or 7 streaked snails respectively. Using R we find this.

```
R> dbinom(5:7, 10, 0.6)

[1] 0.2006581 0.2508227 0.2149908
```

We get the same answers using the formula $\Pr\{Y = j\} = {}_nC_j p^j (1-p)^{n-j}$.

5. Exercise 3.24 (page 118).

Solution: An experiment involves sacrificing 310 pregnant mice and counting the number of living embryos. We are asked to fit a binomial distribution and compare the expected and observed frequencies. Here is how to do this in R.

```
R> dead <- 0:9
R> observed <- c(136, 103, 50, 13, 6, 1, 1, 0, 0, 0)
R> totalDead <- sum(dead * observed)
R> totalDead

[1] 277

R> totalMice <- 9 * 310
R> totalMice

[1] 2790

R> phat <- totalDead/totalMice
R> phat

[1] 0.09928315

R> expected <- dbinom(0:9, 9, phat) * 310
R> rbind(0:9, observed, round(expected, 1))

         [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
            0    1  2.0  3.0  4.0  5.0    6    7    8     9
observed  136  103 50.0 13.0  6.0  1.0    1    0    0     0
          121  120 52.9 13.6  2.2  0.2    0    0    0     0
```

The expression `sum(dead*observed)` calculated $(0 \times 136) + (1 \times 103) + \cdots + (9 \times 0)$. The expected count for each outcome is $n$ times the binomial probability of the outcome. The function `rbind` takes vectors of the same length and binds them together as rows of a matrix. You can improve the output by adding a label to the rounded values.

```
R> rbind(0:9, observed, expected = round(expected, 1))

         [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
            0    1  2.0  3.0  4.0  5.0    6    7    8     9
observed  136  103 50.0 13.0  6.0  1.0    1    0    0     0
expected  121  120 52.9 13.6  2.2  0.2    0    0    0     0
```

6. Exercises 3.30 and 3.31 (pages 119–120). In Exercise 3.31(b), it should read, "*How* large must $n$ be ...".

Solution: In Exercise 3.30, we have a sample size of 50 patients and know that a drug has a 1% chance of causing kidney damage. The probability that none of the patients experience kidney damage is

$$\Pr\{Y = 0 | n = 50\} = (0.99)^{50} \doteq 0.605$$

In the followup exercise we are asked to find the probability of at least one patient getting kidney damage in a sample size of 100.

$$\Pr\{Y \geq 1 | n = 100\} = 1 - \Pr\{Y = 0 | n = 100\} = 1 - (0.99)^{100} \doteq 0.634$$

Finally, we are asked to determine how large $n$ would need to be to make the probability greater than 95%, $\Pr\{Y \geq 1|n\} = 1 - (0.99)^n > 0.95$. We solve this by taking logarithms and doing algebra.

$$\begin{aligned} 1 - (0.99)^n &> 0.95 \\ (0.99)^n &< 0.05 \\ n \ln 0.99 &< \ln 0.05 \\ n &> \frac{\ln 0.05}{\ln 0.99} \\ n &> 298.1 \end{aligned}$$

7. Use the normal table on the inside cover of your textbook to complete Exercise 4.3 (page 133). Then use the R function `pnorm` to do the same thing. (Include the R commands in your solution.)

   Solution: By hand, we need to standardize by subtracting the mean and dividing by the standard deviation. With R we use `pnorm`.

   (a) $\Pr\{Y < 1500\} = \Pr\{(Y - 1400)/1500 < (1500 - 1400)/100\} = \Pr\{Z < 1\} \doteq 0.8413$.
   In R, `pnorm(1500,1400,100) = 0.8413447`.

   (b) $\Pr\{1325 < Y < 1500\} = \Pr\{(1325 - 1400)/100 < (Y - 1400)/100 < (1500 - 1400)/100\} = \Pr\{-0.75 < Z < 1\} \doteq 0.8413 - 0.2266 = 0.6147$.
   In R, `pnorm(1500,1400,100) - pnorm(1325,1400,100) = 0.6147174`.

   (c) $\Pr\{Y > 1325\} = \Pr\{(Y - 1400)/100 > (1325 - 1400)/100\} = \Pr\{Z > -0.75\} = \Pr\{Z < 0.75\} \doteq 0.7734$.
   In R, `1 - pnorm(1325,1400,100) = 0.7733726`.

   (d) $\Pr\{Y > 1475\} = \Pr\{(Y - 1400)/100 > (1475 - 1400)/100\} = \Pr\{Z > 0.75\} = \Pr\{Z < -0.75\} \doteq 0.2266$.
   In R, `1 - pnorm(1475,1400,100) = 0.2266274`.

   (e) $\Pr\{1475 < Y < 1600\} = \Pr\{(1475 - 1400)/100 < (Y - 1400)/100 < (1600 - 1400)/100\} = \Pr\{0.75 < Z < 2\} \doteq 0.9772 - 0.7734 = 0.2038$.
   In R, `pnorm(1600,1400,100) - pnorm(1475,1400,100) = 0.2038772`.

   (f) $\Pr\{1200 < Y < 1325\} = \Pr\{(1200 - 1400)/100 < (Y - 1400)/100 < (1325 - 1400)/100\} = \Pr\{-2 < Z < -0.75\} \doteq 0.2266 - 0.0228 = 0.2038$.
   In R, `pnorm(1325,1400,100) - pnorm(1200,1400,100) = 0.2038772`.

8. For the normal curve described in the previous problem, use the normal table in your textbook to find the 10th, 90th, and 95th percentiles. Then use the R function `qnorm` to do the same thing. (Include the R commands in your solution.)

   Solution: The 10th percentile is the location where the area to the left is 0.1000. The nearest z-score from the table is $z = -1.28$. (The area is actually 0.1003, but this is as accurate as the table allows.) So, the 10th percentile is 1.28 standard deviations below the mean, $1400 - 128 = 1272$.

   The 90th percentile is the location where the area to the left is 0.9000. The nearest z-score from the table is $z = 1.28$, which makes sense because the normal distribution is symmetric around 0— cutting off the bottom 90 percent is the same as cutting off the top 10 percent. So, the 90th percentile is 1.28 standard deviations above the mean, $1400 + 128 = 1528$.

   The 95th percentile is the location where the area to the left is 0.9500. The two nearest areas are 0.9495 and 0.9505 which correspond to $z = 1.64$ and $z = 1.65$. We can average and say $z = 1.645$. So, the 95th percentile is 1.645 standard deviations above the mean, $1400 - 164.5 = 1564.5$.

   Here is the solution in R.

```
R> qnorm(c(0.1, 0.9, 0.95), 1400, 100)

[1] 1271.845 1528.155 1564.485
```