Lab 5: The Binomial and Poisson Distributions

For this lab, it will be helpful to have a copy of the knitted version of this document to answer the questions as much of it is written using mathematical notation that may be difficult to read when the document is not knitted.

Review: Computing Probability

When each outcome in the sample space is *equally likely*, and there are only *finite* number of outcomes in the sample space:

$$P(A) = \frac{\#\text{ in event } A}{\#\text{ in sample space } \Omega}$$

However, there are cases where

- the outcomes in the sample space are NOT equally likely, and/or
- you cannot count the number of outcomes inside the sample space and the event.

In these cases, you have to use the properties of probability, law of total probability, definition of conditional probability, Bayes theorems, etc. Binomial distribution and Poisson distribution (and many, many others) offer you off-the-shelf formulae to compute probabilities assuming that some specific conditions are satisfied.

Lab Goals

1. The purpose of this lab is to explore the following distributions:

- Binomial distribution
- Poisson distribution

2. You will be asked to compute probabilities of events using the pmf of these distributions both

- By hand using the formula
- Using R functions

Emphasis is on identifying random events that can be modeled using these distributions and learning how to calculate probabilities using these distributions by hand and in R.

Some Notation

Your solutions to the problems below must include the formula used for each calculation. Here is some helpful notation you can copy, edit, and paste as needed. This notation is **not** R code. The notation between dollar signs in the Rmd file is in something called `latex` which lets you write mathematical expressions nicely. Without latex you'd get, for example, (x^2+y)/z, whereas with latex you get $\frac{x^2+y}{z}$. Remember that code chunks are for R code only, not `latex`.

1. For a binomial random variable, X, the pmf is given by $P(X = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$ for $k = 0, 1, \ldots, n$.
2. For a Poisson random variable X, the pmf is given by $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$.

Some R Code

1. $e \approx 2.7182818$ is an important constant in mathematics. `exp()` is the exponential function in R, and

```
exp(1)
```

returns the approximate value of e. e^3 (e to the 3rd power) is

```
exp(3)
```

We can verify whether

```
exp(1)^3
```

is indeed equal to `exp(3)`

2. `factorial()` is the R function for computing factorials, so for instance, $3! = 3 \times 2 \times 1$ is

```
factorial(3)
```

Note that we define $0! = 1$ (and of course, $1! = 1$)

```
factorial(0)
```

3. In R the following operators can be used

```
2+2  # + for addition
3-2  # - for subtraction
3*2  # * for multiplication
4/4  # / for division
2^3  # ^ for exponents
```

4. Keep in mind that R follows the order of operations. So, when evaluating 3 + 6*8. Multiplication will be performed before addition.

```
3+6*8
```

```
## [1] 51
```

To have the addition evaluated first, you must use parenthesis.

```
(3+6)*8
```

```
## [1] 72
```

R will evaluate an expression in the following order:

- 1) Parenthesis
- 2) Exponents
- 3) Multiplication and/or Division
- 4) Addition and/or Subtraction

5. Here is some code you can copy, paste, and edit in code chunks to calculate probabilities in this document for the binomial and Poisson distributions. DO NOT CHANGE THESE CODE CHUNKS TO `eval=TRUE`. Variables in them have not been defined, and it will not run.

a) To calculate $P(X = k)$ for a binomial distribution use the following code where you have specified `n`,`k`, and `p`.

```
P_k = choose(n,k)*(p^k)*(1-p)^(n-k)
```

b) To calculate P(X = k) for the Poisson distribution use the following code where you have defined `lambda` and `k`.

```
P_k = (exp(-lambda)*lambda^(k))/factorial(k)
```

Important Concepts

I flip two fair coins, so

- Experiment: Flipping two fair coins
- Outcome: Four possible outcomes: (H,H),(H,T),(T,H),(T,T)
- Sample space: {(H,H),(H,T),(T,H),(T,T)}
- Event: "One head, one tail": {(H,T),(T,H)}
- **Random Variable**: Summarize the outcomes/events. For instance, X = Number of head, such that
  - $X = 0$ represents {(T,T)} (or event "two tails")
  - $X = 1$ represents {(H,T),(T,H)} (or event "one head, one tail")
  - $X = 2$ represents {(H,H)} (or event "two tails")
- **Probability distribution**: Intuitively, assign probabilities to the random variable X taking a specific value such that they sum up to 1.
  - $P(X = 0) = 1/4$
  - $P(X = 1) = 1/2$
  - $P(X = 2) = 1/4$
- **Probability mass function**: the function to do the probability assignment. The following is the pmf of binomial distribution.
  - $P(X = k) = \binom{2}{k} 0.5^k (1 - 0.5)^k$ where $k \in \{0, 1, 2\}$

Properties of Random Variables

For any random variable X,

1. $E(cX) = cE(X)$ for any constant, c
2. $E(X + c) = E(X) + c$ for any constant, c
3. $Var(cX) = c^2 Var(X)$. Note, this one makes more sense if you think about the standard deviation, which is the square root of the variance. It says that the standard deviation of cX is |c| times the standard deviation of X.
4. $Var(X + c) = Var(X)$

These rules actually make sense if you think about an example. Suppose we are playing a board game in which **you roll a fair die**. Let X represent the value rolled.

Problem 1

a) Compute $E(X)$, $Var(X)$, $SD(X)$

b) Suppose in this game, you go forward 2X spaces when you roll X. On average how many spaces forward do you go? (Which of the properties did you use?)

c) For the same rules as above, what is the standard deviation of the number of spaces you move ahead? Your answer can be expressed in terms of "SD(X)," the standard deviation of X. (Which of the properties did you use?)

d) Now suppose the rules are different. If we roll X, then we go forward X+4 spaces. What is the expected number of spaces that you go forward?

e) Using the rules in (d), what is the standard deviation for the number of spaces you go forward?

More properties:

1. For any random variables X and Y,

   $E(X+Y) = E(X) + E(Y)$

   Suppose you roll two dice and X and Y are random variables representing the values. What is the expected sum of the two dice?

2. (IMPORTANT) For any random variables X and Y, the equality

   $Var(X+Y) = Var(X) + Var(Y)$

   is **NOT** true in general.

Binomial Distribution

A binomial random variable, X~Binomial(n,p), is characterized by the following (p.p. 147 of *OpenIntro Statistics*, 3rd Edition):

1. The trials are *independent* of each other.
2. The number of trials, n, is *fixed* in advance.
3. Each trial outcome can be classified as a *success* or *failure*.
4. The probability of a success, p, is the *same* for each trial.
5. The random variable X = Number of "successes" out of n "trials"

Some examples:

1. X = Number of hits in n "at bats"
2. X = Number of cars with defective airbags out of n manufactured
3. X = Number of days you wake up on time in a week (n =?)

The probability mass function for $X \sim \text{Binomial}(n, p)$ is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k \in \{0, 1, 2, ..., n\}$. Here, $\binom{n}{k}$ is the number of ways to choose k items from n and is defined as

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

[Optional] Verify that $\binom{n}{k} = \binom{n}{n-k}$

The mean and variance of a binomial distribution are

$$E(X) = \mu = np$$

$$Var(X) = \sigma^2 = np(1-p)$$

Problem 2

The proportion of Norwegians in their early 40s who drink at least 1 cup of coffee per day is about p = 0.894. Suppose that we take a random sample of 50 Norwegians from this age group. Let X = the number of Norwegians in the sample that drink at least 1 cup of coffee a day. In your answers to the questions below, include the correct notation for the probability being asked for in terms of the random variable X.

[Optional] Describe the sample space of this experiment. How many outcomes are there in the sample space? Are the outcomes equally likely?

a) Why do you choose binomial distribution to model X? (The five conditions to check.)

b) What is the expected value of the number of Norwegians that drink at least 1 cup of coffee a day in the sample?

c) What is the standard deviation of the number of Norwegians that drink at least 1 cup of coffee a day in the sample?

For questions 1.(d)-(f), calculate the required probabilities using R as only a calculator. Include all formulas used.

d) What is the probability that exactly 48 Norwegians in the sample drink at least 1 cup of coffee a day?

[Optional] "$X = 48$" is a summary of the outcomes in the event "exactly 48 Norwegians in the sample drink at least 1 cup of coffee a day". Describe the outcomes inside the event.

e) What is the probability that exactly 2 Norwegians in the sample do not drink at least 1 cup of coffee a day?

f) What is the probability that more than 2 Norwegians in the sample do not drink at least 1 cup of coffee a day?

g) The following simulates 10,000 observations from a Binomial(50,.894) distribution. `numCoffee` is the simulated observations. The last three lines of code use the simulated values to approximate the answers to (b)-(d).

```
set.seed(1)
num_simulations = 10000
numCoffee = rbinom(num_simulations, size=50, p=0.894)
# check b
mean(numCoffee)
```

```
## [1] 44.6968
```

```
# check c
sd(numCoffee)
```

```
## [1] 2.194801
```

```
# check d
mean(numCoffee == 48)
```

```
## [1] 0.0644
```

Problem 3

R has built in functions to calculate probabilities and quantiles for the binomial distribution.

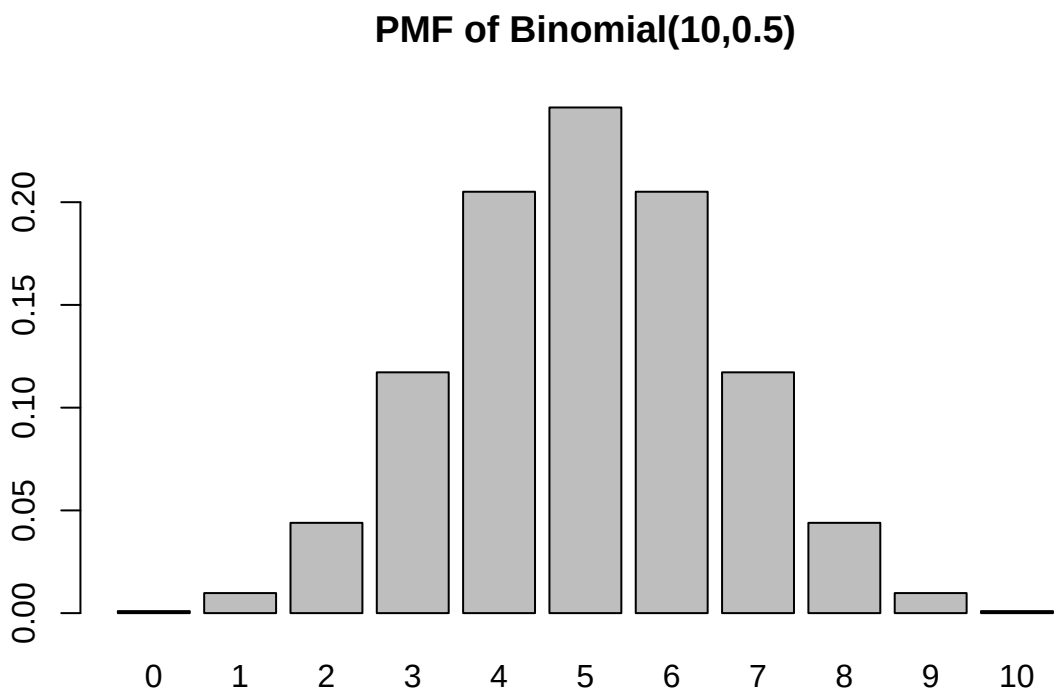
The `dbinom(k,n,p)` is the probability mass function for the binomial distribution, i.e. $P(X = k) =$ `dbinom(k,n,p)`. For this function, the following need to be specified:

- `k`, the number of successes associated with the desired probability

- `n`, the number of trials
- `p`, the probability of success for a single trial

The height of each bar in the following illustration is $P(X = k)$ for X~Binomial(10,.5) and k= 0,..,10.

```
barplot(dbinom(0:10,10,.5),names.arg = 0:10, main="PMF of Binomial(10,0.5)")
```



The `pbinom(k,n,p)` function in R is the probability distribution function for the binomial distribution, i.e. $P(X \leq k)$ =`pbinom(k,n,p)`. Here `k`, `n`, and `p` are as specified above. [**Notice the less than or equal to sign**]

The `pbinom(k,n,p)` function can also be used to find the upper tail probabilities, i.e. $P(X > k)$, by using the optional parameter `lower.tail=FALSE`. Thus, $P(X > k)$ = `pbinom(k,n,p,lower.tail=FALSE)` [**Notice the less than sign**]

For 3. (a)-(c), repeat the probability calculations for 2. (d)-2.(f) using these R functions.

The `qbinom(p_quant,n,p)` function in R is the quantile function for the binomial distribution. It finds the smallest value of k such that $P(X \leq k) \geq$ `p_quant`. Here `k`, `n`, and `p` are as specified above. `p_quant` is such that the 100 $\times$ `p_quant` percentile of the binomial distribution is given by `qbinom(p_quant,n,p)`.For instance, specifying `p_quant = .5` will return the median of the binomial distribution.

d) Find the 25th and 75th percentiles of the binomial distribution specified in problem 2.

e) Find the smallest number such that the probability of the number of Norwegians that drink at least 1 cup of coffee a day in the sample is greater than this number is less than or equal to 10%.

For now you have learned four functions for binomial distributions:

- `rbinom(num_simulations, n, p)`: generates `num_simulations` number of random numbers from the binomial distribution [in Problem 2. (g)]
- `dbinom(k, n, p)` $= P(X = k)$
- `pbinom(k, n, p, lower.tail = TRUE)` $= P(X \leq k)$ [`pbinom(k, n, p, lower.tail = FALSE)` $= P(X > k)$]
- `qbinom(p_quant, n, p)` finds the smallest value of k such that $P(X \leq k) \geq$ `p_quant`

You will soon find that for each distribution there are four similar functions in R with prefix `r` (random number generator), `d` (density), `p` (cumulative distribution function), `q` (quantile function).

[Optional] Review of `for` loop: write code to compute $P(X \leq 14)$ where $X \sim \text{Binomial}(20, 1/3)$.

```
n <- 20
k <- 14
p <- 1/3
sum_prob <- 0
# compute P(X=0)+P(X=1)+...+P(X=14)
for (i in 0:k) {
  sum_prob <- sum_prob + choose(n,i)*(p^i)*(1-p)^(n-i)
}
# print out the result
sum_prob
pbinom(k, n, p)
```

Optional: Bernoulli Distribution and Binomial Distribution

Assume that $X_1 \sim \text{Bernoulli}(p)$, $X_2 \sim \text{Bernoulli}(p)$, ..., $X_n \sim \text{Bernoulli}(p)$ and for each $i \in \{1, 2, \ldots, n\}$, X_i is independent of each other.

1. Justify that $X = X_1 + X_2 + \cdots + X_n$ is a bionomial random variable. (Five conditions to check)
2. Bernoulli distribution is a special case of binomial distribution when $n = 1$ (the number of trials). Plug $n = 1$ in the formula of pmf, mean, variance and standard deviation for *binomial distrution*, to obtain their counterparts for *Bernoulli distribution*.

You might notice that

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = np$$

where $E(X_i) = p$ for each $i \in \{1, 2, \ldots, n\}$. This follows naturally from the property of mean.

However, for the variance,

$$Var(X) = Var(X_1 + X_2 + \cdots + X_n) = np(1 - p)$$

Because $Var(X_i) = p(1 - p)$ for each $i \in \{1, 2, \ldots, n\}$, you might *guess* that

$$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$$

in this case. You are right! In general, when Y_1, Y_2, ..., Y_n are **INDEPENDENT** of each other, then the following equality holds

$$Var(Y_1 + Y_2 + \cdots + Y_n) = Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n)$$

(however, the above equality is not true in general).

The definition of independent *random variables* is different from the definition of independent *events*. We will leave it to future lectures/labs.

Poisson Distribution

A poisson random variable, X~Poisson(λ), is characterized by the following:

1. X = Number of rare events that occur over a fixed amount of time or space
2. Events are independent
3. The maximum number of events that occur is not fixed

Some examples:

1. X = Number of tornadoes in a particular area over a year
2. X = Number of raindrops that fall on a particular square inch of roof during a one-second interval of time
3. X = Number of people that arrive at a train station during a 5 minute interval of time

For X~Poisson(λ) and any $k \in \{0, 1, 2, ...\}$, the following probability mass function defines P(X=k).

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

The mean and variance of a Poisson distribution are

$$E(X) = \mu = \lambda$$

$$Var(X) = \sigma^2 = \lambda$$

Problem 4

In the summer months, northern Minnesota has a thriving mosquito population. When the average person steps outside in the summer, the number of mosquito bites he receives every hour follows a Poisson(10) distribution if he has not applied mosquito repellant. If he has applied mosquito repellant, the number of mosquito bites he receives every hour has a Poisson(3) distribution. Assume a randomly chosen Minnesotan has just stepped outside without applying mosquito repellant. Let

$X =$ number of mosquito bites this person receives in the first hour.

$Y =$ number of mosquito bites the person receives in the second hour.

[Optional] Describe the sample space of this experiment. Can you count the number of outcomes in the sample space? Are all outcomes equally likely?

[Optional] Why do you choose Poisson distribution to model X?

a) What is the probability in the first hour outside this individual does not get bit by a mosquito?

b) What is $P(10 < X \leq 12)$?

[Optinoal] "$10 < X \leq 12$" is another summary of the outcomes in the event. Describe in words what the outcomes in this event look like.

c) If this Minnesotan plans to stay outside for 2 hours in total and he applies mosquito repellant after the first hour, how many mosquito bites can he expect to receive?

d) After spending two hours outdoors, this Minnesotan will treat his mosquito bites with a topical ointment. For complete relief from the allergic reaction caused by the mosquito bites, the ointment must be applied to every mosquito bite 3 times. How many times in total will he expect to apply the ointment after the 2 hours to alleviate the allergic reaction of every bite?

Problem 5

R also has built in functions to calculate probabilities and quantiles for the Poisson distribution.

The `dpois`(k, λ) is the probability mass function for the Poisson distribution, i.e. $P(X = k) =$ `dpois`(k, λ). For this function, the following need to be specified:

- `k`, the number of successes associated with the desired probability
- λ, the mean and variance of the Poisson distribution

[Optional] Modify the code in Problem 3 to generate a barplot with the height of each bar being $P(X = k)$ for $X \sim$ Poisson(10). You can plot from $X = 0$ to $X = 50$

The `ppois`(k, λ) function in R is the cumulative probability distribution function for the Poisson distribution, i.e. $P(X \leq k) =$ `ppois`(k, λ). Here `k` and λ are as specified above.

Similar to the `pbinom()` function, the `ppois`(k, λ) function can also be used to find the upper tail probabilities, i.e. $P(X > k)$, by using the optional parameter `lower.tail=FALSE`. Thus, $P(X > k) =$ `ppois`(k,λ,`lower.tail=FALSE`)

For 5. (a)-(b), repeat the probability calculations for 4. (a)-(b) using these R functions.

The `qpois`(p_quant,λ) function in R is the quantile function for the Poisson distribution. It finds the smallest value of k such that $P(X \leq k) \geq$ `p_quant`. Here `k` and λ are as specified above. `p_quant` is such that the $100 \times$ `pquant` percentile of the Poisson distribution is given by `qpois`(p_quant,λ).

c) Find the median of the first Poisson distribution specified in problem 4.

[Optional] There's a function `rpois()` to generate random numbers that follow the Poisson distributions. Perform Monte Carlo simulations to approximate the answers to (a) and (b)

```
set.seed(1)
num_simulations = 1e6
num_bites_1st_hour <- rpois(num_simulations, lambda = 10)
# check a: P(X=0)=4.540e-05
mean(num_bites_1st_hour == 0)
```

```
## [1] 5.1e-05
```

```
# check b: P(10<X<=12)=0.2085
mean(num_bites_1st_hour > 10 & num_bites_1st_hour <= 12)
```

```
## [1] 0.208432
```

[Optional] Compare the R functions `rpois`, `dpois`, `ppois` and `qpois`.