Week 9: Discrete Probability Distributions

In the lecture this week, we covered discrete probability distributions. In this live R, we're going to explore these concepts further. We'll start by loading the tidyverse:

```
library(tidyverse)
```

Imagine you are a doctor working in geriatric medicine in Colombia. You have noticed that several of your patients are presenting with symptoms of Alzheimer's disease. While AD is not unfamiliar to you as a geriatric practitioner, you are seeing more patients with AD symptoms than you would expect given your previous experience. You want to know whether the prevalence of AD in this region is on track with prevalence rates from the general population. You know that the prevalence of AD in the overall population (starting at age 65) is around 4%.

Because our outcome (*presence of disease*) is binomial, we can model expected probabilities using a binomial distribution. In this example, our random variable is:

$$X = \begin{cases} 0 \; if \; AD \; absent \\ 1 \; if \; AD \; present \end{cases}$$

1 Describing our Random Variable

Before you investigate your patients further, you can compute some descriptive data about your random variable, X. One thing that you can do is to calculate the **expected value E(X) and standard deviation SD(X)**. These values give you an idea of the long-term mean and standard deviation of your data given a certain probability of success. In other words, if you were to see a set of n new patients over and over, you would expect to diagnose $E(X)$ with AD. The $SD(X)$ helps to account for the differences in $E(X)$ that may occur due to random sampling across trials. Sometimes you'll diagnose exactly 4% of your patients, but other times, you might diagnose 5%, and other times 3%. $SD(X)$ tells you, on average, how much difference in $E(X)$ you see across trials.

To calculate the long-term mean, we use the formula

$$E(X) = n \times p$$

To calculate the standard deviation, we use the formula

$$SD(X) = \sqrt{n \times p \times (1-p)}$$

Let's compute $E(X)$ and $SD(X)$ for our random variable. Imagine you have 100 patients. Rather than plugging in the numbers by hand, you may find it easier to assign them outside of the formula, then plug the assigned values into the formula:

```
n <- 100
p <- .04

n*p
```

```
## [1] 4
```

```
sqrt(n*p*(1-p))
```

```
## [1] 1.959592
```

This tells us that for every 100 patients that you see, you would expect to diagnose 4 $\pm$ 1.96 of them with AD.

2 Probability Mass Function in Practice

8 of your 100 patients are presenting with symptoms of AD. We want to know the probability of having 8 patients with a diagnosis of AD, given an expected prevalence of 4%. We can calculate this using the Probability Mass Function.

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

where:

$k = number\ of\ successes$

$n = number\ of\ trials$

$p = P(success)$

$q = P(failure)$

Our total `successes` in this case would be 8. We have 100 patients total. We know the global probability is 4%. We can plug these values into the PMF:

$$P(X = 8) = \binom{100}{8} \times .04^8 \times (1 - .04)^{100-8}$$

We'll start by doing it by hand in R. We can use the `factorial()` function to compute the factorials within the *combination* portion of the equation.

```
k <- 8

step1 <- factorial(n)/(factorial(k)*factorial(n-k))
step2 <- p^k
step3 <- (1-p)^(n-k)

step1*step2*step3
```

```
## [1] 0.02852013
```

If you've done it properly, you will get the same value using the `dbinom` function:

```
dbinom(x=k, size = n, p = p)
```

```
## [1] 0.02852013
```

This tells us that the probability of 8 out of 100 patients having AD is only 2.9%, given the population prevalence. Very unlikely indeed.

Conveniently, the `dbinom` function allows you to add multiple values for k, so you can compute the probability of the number of patients having AD being the specific values you list:

```
dbinom(x=0:k, size = n, p = p)
```

```
## [1] 0.01687032 0.07029300 0.14497931 0.19733295 0.19938850 0.15951080 0.10523282
## [8] 0.05888027 0.02852013
```

We can visualise the discrete probability distribution by using the above output in a bar plot. Remember, the `dbinom` function gives us the probability of each outcome given a certain probability level. In other words, given that the probability of `success` is 4%, what is the likelihood that 0 of 8 trials will result in success? What is the probability that 1 of 8 trials will result in `success`?...and so on.

First, let's produce the dataset with the probabilities (`pK`) linked to each outcome (`k`)

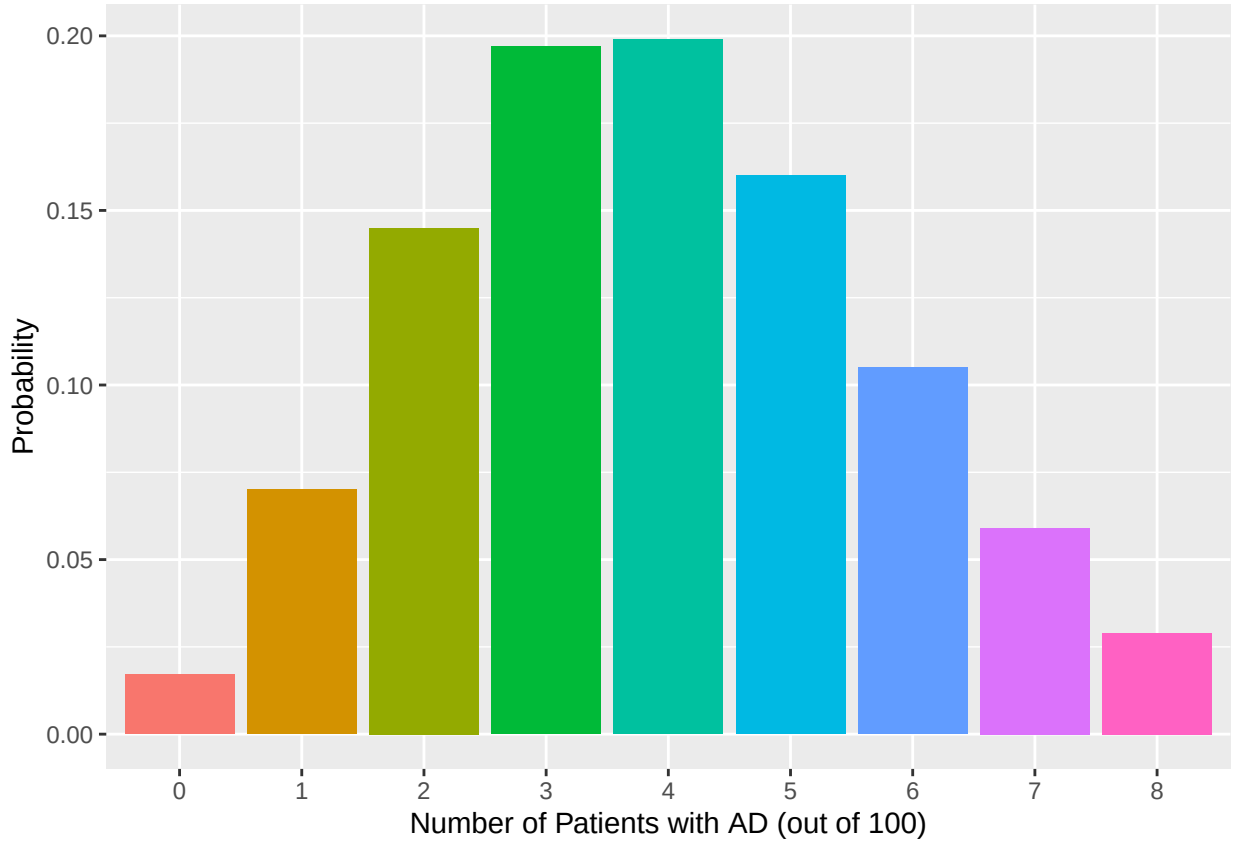
```
discDat <- data.frame(k=as.factor(0:k),
                      pK=round(dbinom(x=0:k, size = n, p = p), 3))
discDat
```

```
##   k    pK
## 1 0 0.017
## 2 1 0.070
## 3 2 0.145
## 4 3 0.197
## 5 4 0.199
## 6 5 0.160
## 7 6 0.105
## 8 7 0.059
## 9 8 0.029
```

We can pass this data to a bar plot. This way, our bar plot shows us the probability that 0-8 of 100 patients would have AD, given the prevalence in the population.

Note that we use `geom_col` here, as we are giving `ggplot` the heights of each bar in the `pK` variable, rather than having `ggplot` calculate the heights based on raw data (as we have seen with `geom_bar`).

```
ggplot(discDat, aes(x=k, y=pK, fill = k)) + geom_col() +
  theme(legend.position = 'none') +
  labs(x = 'Number of Patients with AD (out of 100)',
       y = 'Probability')
```



Here, you can see that 4 patients out of 100 is the most likely (which makes sense given a population prevalence of 4%). This lines up with $E(X)$. It is much less likely that exactly 8 of your 100 patients have AD.

3 Cumulative Distribution Function

The PMF allows us to compute the probability that *exactly* 8 of our 100 patients have AD. But perhaps we want to calculate the probability of a range of outcomes. In this case, we can use the cumulative distribution function:

$$P(X \leq x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$$

This looks complex, but really, it's just summing the outcome of the PMF across all values *less than or equal to* the value of interest.

We could do this using the `sum` function on the output of `dbinom()`. We could also use the `pbinom()` function. This function applies the cdf of the binomial distribution to k. It can take, as arguments, the same arguments we used with `dbinom()`. However, depending on our range of interest, we'll have to change the values of the arguments slightly.

```
sum(dbinom(0:k, n, p))
```

```
## [1] 0.9810081
```

```r
pbinom(k, n, p)
```

```
## [1] 0.9810081
```

If we use $k = 8$, as before, it gives us the probability that the total number of your current patients with AD will fall between 0-8. In other words, given that you have 100 patients total, the likelihood that you have between 0-8 patients with AD is equal to 98%.

```r
pbinom(0:4, n, p)
```

```
## [1] 0.01687032 0.08716332 0.23214262 0.42947557 0.62886407
```

Let's say you want to know what the probability is that more than 4 (the expected amount) of your 100 patients have AD. You could sum the specified range using `dbinom`:

```r
sum(dbinom(5:100, n, p))
```

```
## [1] 0.3711359
```

But, with the understanding that all outcomes sum to a probability of 1, you can also take 1 minus the cumulative value produced by `pbinom`:

```r
1-pbinom(4, n, p)
```

```
## [1] 0.3711359
```

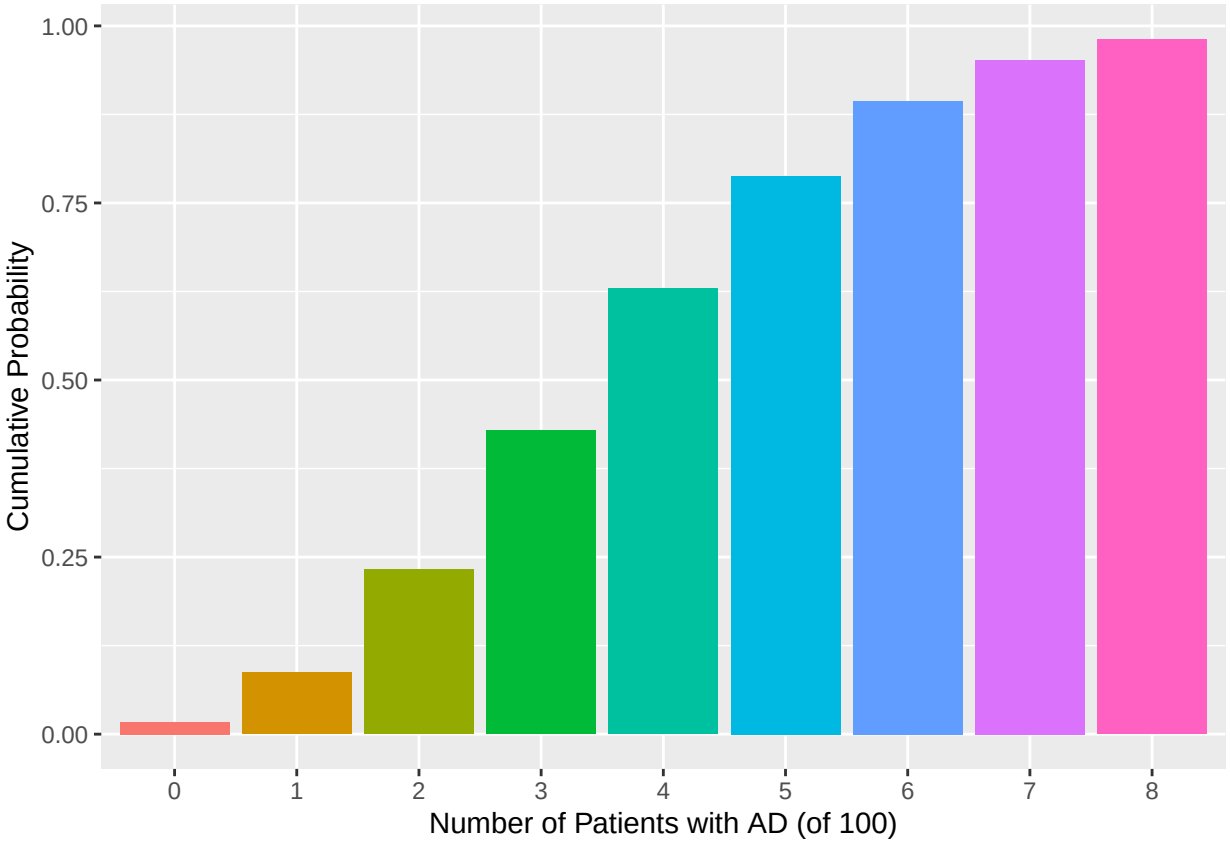
In other words, the probability of having more than 4 patients with AD is equal to 1 minus the probability of having 0-4 patients with AD. In our case, the probability of having more than 4 patients with AD is ~37%

Like before, we can also produce a plot from our results. First, let's add a cumulative probability column to our dataframe:

```r
discDat$cuPk <- round(pbinom(0:k, n, p), 3)
discDat
```

```
##   k    pK  cuPk
## 1 0 0.017 0.017
## 2 1 0.070 0.087
## 3 2 0.145 0.232
## 4 3 0.197 0.429
## 5 4 0.199 0.629
## 6 5 0.160 0.788
## 7 6 0.105 0.894
## 8 7 0.059 0.952
## 9 8 0.029 0.981
```

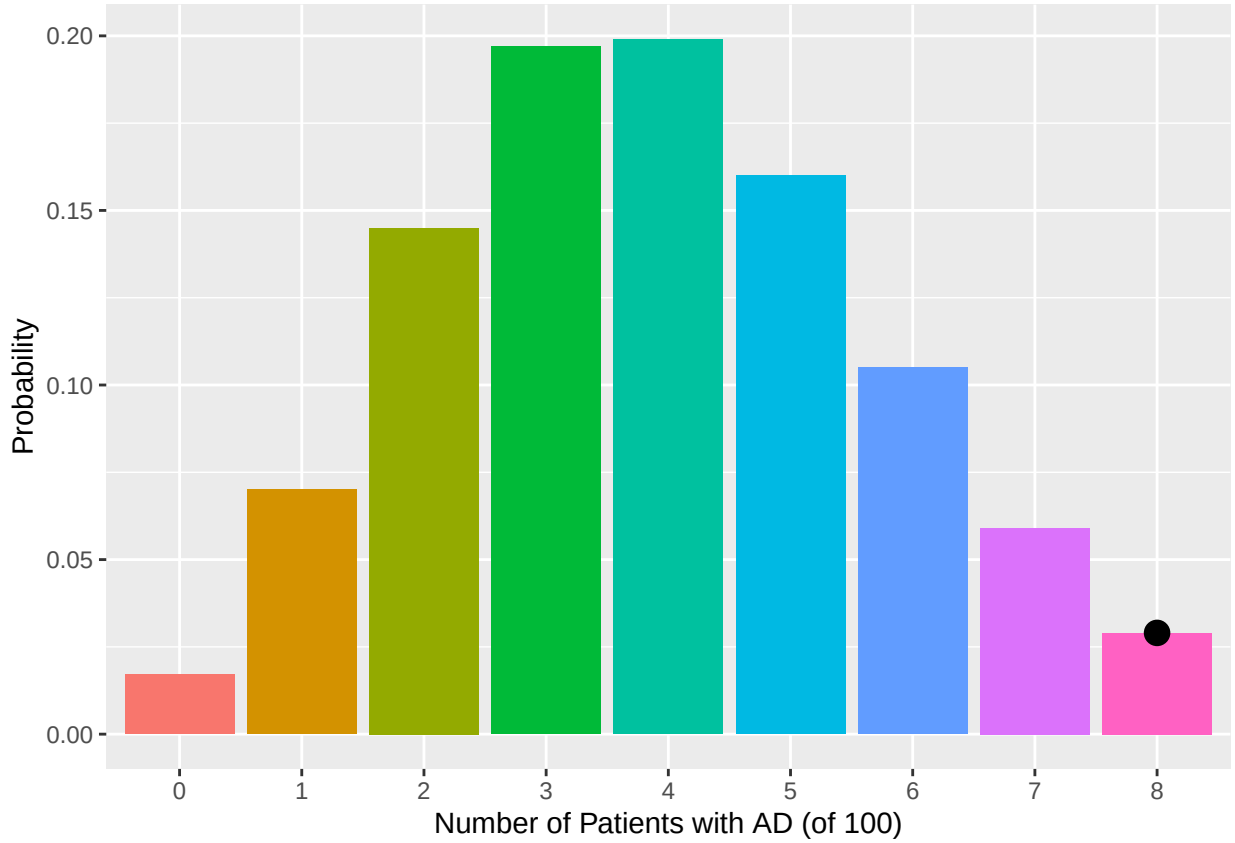
```r
ggplot(discDat, aes(x = k, y = cuPk, fill = k)) + geom_col() +
  theme(legend.position = 'none') +
  labs(x = 'Number of Patients with AD (of 100)',
       y = 'Cumulative Probability')
```



4 Comparing our Results to the Expected Probability

Now, let's compare your patient data to the binomial frequency distribution we've created based on p of 4% and n of 100. We can use this to judge how well our sample's data matches up with what we would expect, given the data of the overall population. To do this, we can plot the expected frequency distribution we produced using `dbinom` and see how our sample (black point) matches up with the expectations.

```
ggplot(discDat, aes(x=k, y=pK, fill = k)) + geom_col() +
  theme(legend.position = 'none') +
  labs(x = 'Number of Patients with AD (of 100)',
       y = 'Probability') +
  annotate(geom='point', x = '8', y = discDat$pK[discDat$k=='8'], size = 4)
```



Upon further investigation, you find that 6 of the 8 patients with AD are members of the same family, and you've had reports that other family members are showing some mild symptoms as well. You compute the likelihood that 6 or more members of a 15-person family has AD, given a global precedence of 4%:

```
1-pbinom(5, 15, .04)
```

```
## [1] 1.499059e-05
```

```
myP <- 1-pbinom(5, 15, .04)
myP |> format(scientific = F)
```

```
## [1] "0.00001499059"
```

This probability value is incredibly small. After seeing this, you have doubts that the rate of AD in this family is the same as the general population.

Your patient data are really unlikely, given what is expected from the population. You might think "I would expect 4 patients with AD, but instead, I have 8, and 6 are from the same family! This is very unlikely. . . something must be different about this family that's causing the higher rate of AD." In fact, this example is based on a true story. An extended family group in Colombia has a genetic mutation that causes them to develop AD at a much higher rate (and at a much younger age) than the general population (see here for more info).

There is a key piece from today's example to take through the remainder of your time in DapR: We've got an expected value for an outcome variable, and we can investigate the likelihood of our data, given that expected value is true. **This is the basis for null-hypothesis statistics testing!** If our data are extremely unlikely, there's probably another explanation for why we have the results that we do (e.g. there is a factor within our group that makes them different from the overall population).

5 Example Write-up

In this analysis, we investigated the prevalence of Alzheimer's disease (AD) in 100 patients at a geriatric medical practice in Colombia. Global precedence of AD in patients aged 65+ is ~4%. Given our sample of 100 patients, we would expect AD to present in around 4. Indeed, the probability of at least 4 of 100 patients in our practice being diagnosed with AD is 63%. However, in our patient sample, 8 of 100 patients were diagnosed with AD. Given a global precedence of 4%, the probability of 8 patients out of 100 presenting with AD is only 2.85%. Additionally, 6 of the 8 AD patients are members of the same 15-person family. The probability of 6 or more family members developing AD is less than .001. The unlikely nature of these results suggests there is an external factor increasing the precedence of AD in these patients. See Figure 1 for a visualisation of the discrete probability distribution associated with the expected probability (4%) and the observation number (100).

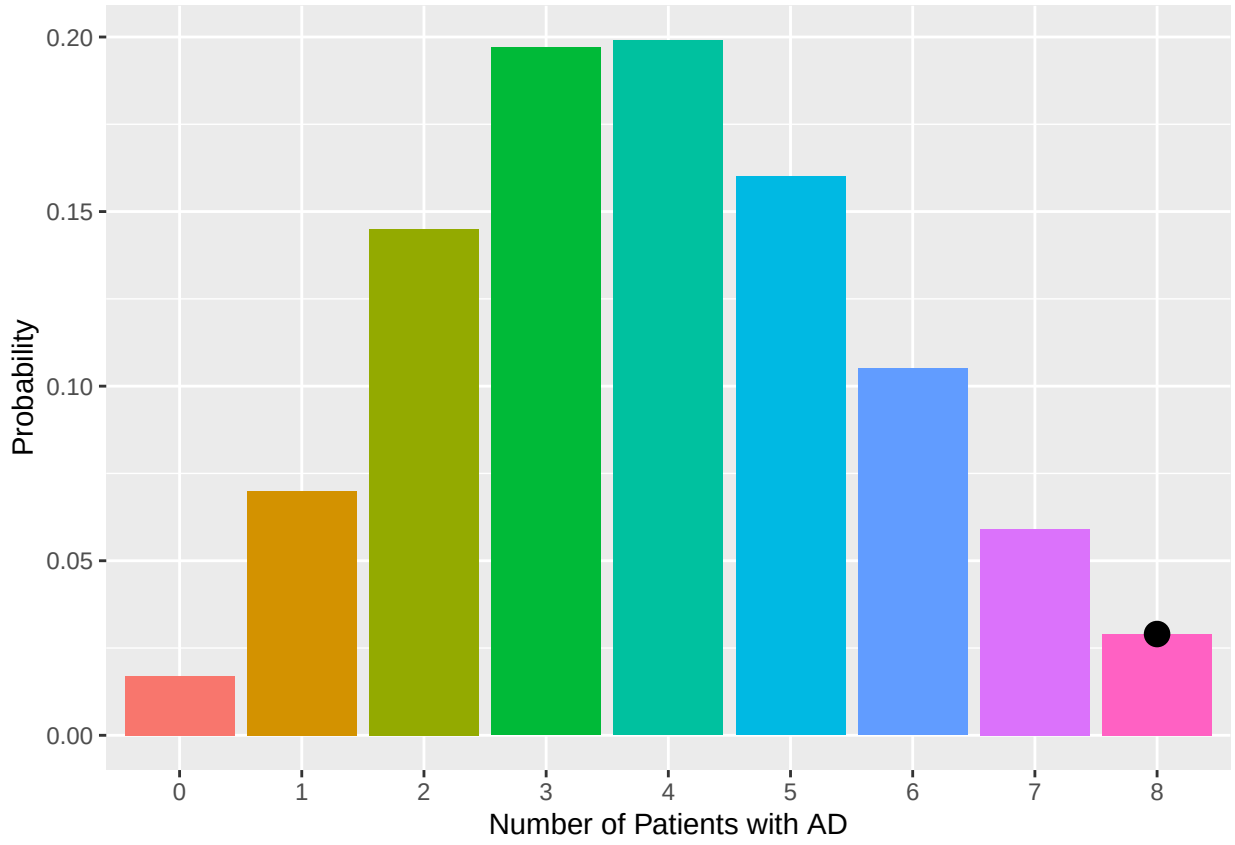


Figure 1: Probability of AD Diagnosis across 100 observations