# Lecture 2: The Bernoulli/Binomial distribution

## Homework

In urban areas of Monte Negro Municipality, Western Amazon, Brazil, 4% of dogs are infected with *Rickettsia*. 80% of serological tests detect *Rickettsia* when present. The test's false positive rate is 17% (i.e., *Rickettsia* is detected but not present). A randomly sampled dog has tested positive for *Rickettsia*. What is the probability that the dog is indeed infected?

To solve this problem, it helps if we create a contingency table based off of what we know:

|  | Disease |  |
|---|---|---|
| Test result | Rickettsia | No Rickettsia |
| $T = +$ | 0.80 | 0.17 |
| $T = -$ | 0.20 | 0.83 |

and then write out the equation.

$$P(Infect|+) = \frac{P(+|Infect) \times P(Infect)}{P(+)}$$

Adding the prior in, we can get the joint and marginal probabilities:

```
(posDisease <- 0.04 * 0.8)
```

```
[1] 0.032
```

```
(negDisease <- 0.04 * 0.2)
```

```
[1] 0.008
```

```
(posNoDisease <- 0.96 * 0.17)
```

```
[1] 0.1632
```

```
(negNoDisease <- 0.96 * 0.83)
```

```
[1] 0.7968
```

Filling this information into the contingency table, we have:

|  | Disease |  |  |
|---|---|---|---|
| Test result | Rickettsia | No Rickettsia | Marginal |
| $T = +$ | 0.032 | 0.1632 | 0.1952 |
| $T = -$ | 0.008 | 0.7968 | 0.848 |
| Marginal | 0.04 | 0.96 | 1.0 |

The posterior probability of a dog being infected with *Rickettsia* given the dog has tested positive is the likelihood of a positive test result when the disease is present (0.8) times the prior probability of the disease prevelance (0.04).

- This is cell 1,1 of our matrix.

We then divide that result by the marginal probability of a positive test result (row 1).

So, the model is as follows:

$$P(Infect|+) = \frac{P(+|Infect) \times P(Infect)}{P(+)} = \frac{0.8 \times 0.04}{0.8 \times 0.04 + 0.17 \times 0.96}$$

$$P(Infect|+) = \frac{0.032}{0.032 + 0.1632} = \frac{0.032}{0.1952} \approx 0.164$$

# Introduction to likelihood with the Bernoulli distribution

## Bernoulli distribution: two possible outcomes

$$p + q = 1.$$

Often $p$ is the parameter of interest:

$$1 - q = p$$

or $q$ given 1 - $p$ = $q$. We can rewrite the equation with respect to $p$ as

$$p + (1 - p) = 1.$$

So now let's think about how to apply a Bernoulli to a "real world" situation. Take the Melissa blue butterfly, *Lyceides melissa.* This handsome insect occasionally feeds on feral alfalfa along roadsides.

Assume the probability that this butterfly is present in any given randomly chosen alfalfa patch is 0.02.

- The probability that it is absent must be $1 - 0.02 = 0.98$.

Further, let's assume that there are 349 patches in Nevada that we can sample. For each, there is a 2% chance of finding *L. melissa.* Thus,

$$Y \sim \text{Bernoulli}(p)$$

$Y$ is a random variable drawn from a Bernoulli distribution with parameter $p$ where $p = 0.02$.

So, the probability that it is present in one particular patch and absent in the next particular patch is

$$p \times q = 0.02 \times 0.98 = 0.0196.$$

. If absent in two sampled patches then

$$q \times q = 0.96^2 = 0.9604.$$

If present in 10 sampled patches then

$$p^{10} = 1.024 \times 10^{-17}.$$

What is the probability that *L. melissa* occurs in any 10 of the 349 patches? How many different ways are there to get 10 gaps?

$$349 \times 348 \times \ldots \times 340 = 2.353647 \times 10^{25}.$$

That's a pretty big number. But it's not accounting for the fact that there are multiple ways to choose the same 10 patches. For example, maybe you sample patch 1 before patch 2, or maybe you sample patch 2 before you sample patch 1.

How many unique ways are there of getting 10 patches? Divide the above by 10!.

In fancy speak

$$\frac{N!}{X!(N-X)!}$$

where

$$N = 349$$
$$X = 10$$

this can be described as 349 choose 10.

This is a binomial coefficient,

$$\binom{N}{X} \qquad \text{N choose X.}$$

Good for us, but all we've figured out is the number of unique combinations. Probability of *melissa* in exactly 10 patches:

$$\text{prob in 10 } \times \text{prob absent from 339 } \times \text{unique combinations of 10}$$
$$0.02^{10} \times 0.98^{339} \times \binom{349}{10} = 0.07045$$

. Small, but not nearly as small as 10 specific gaps.

Probabilities are the cornerstone of the likelihood framework... because they *are* the likelihood framework.

Many Bernoulli trials = Binomial random variables (thus, we are sampling from a binomial distribution).

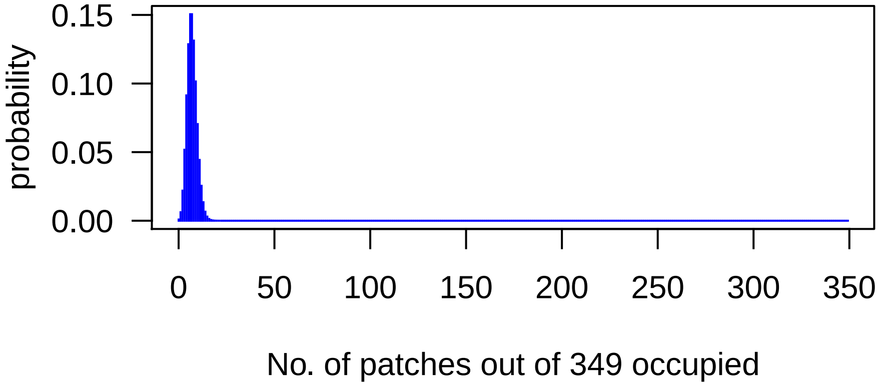
$$Y \sim \text{Binomial}(N, p)$$

The probability of obtaining $Y$ successful outcomes in $N$ independent Bernoulli trials where the probability of success for any event is $p$. If $N = 1$, binomial random variable $Y$ is equivalent to a Bernoulli random variable.

Let's think about, given that $p = 0.02$, what the probability is that we would observe various numbers of occupied patches.

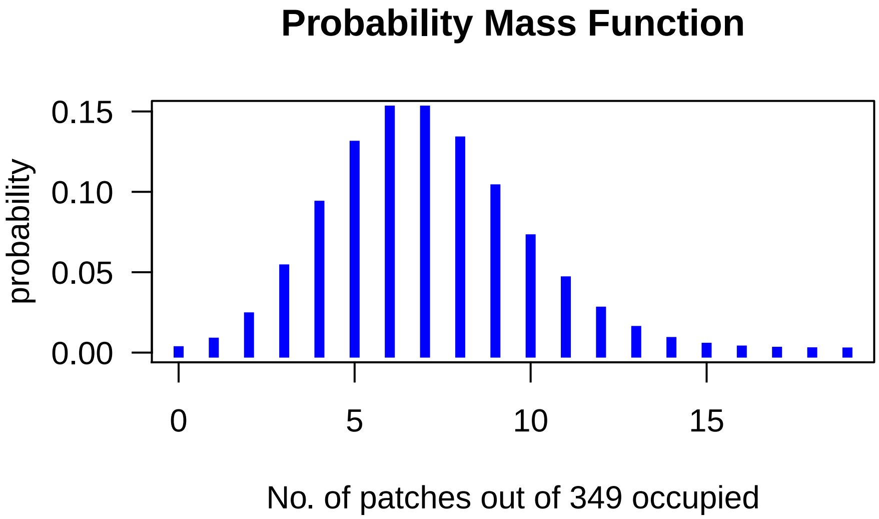
```
p <- 0.02
q <- 1 - p
the.probs <- numeric()#create an empty numeric vector
obs  <- 0:349
for (i in 1:length(obs)){  #starts a loop - each time i will increase by 1
#calculate the probability it is in exactly obs patches
  the.probs[i] <- p^obs[i] * q^(349-obs[i]) * choose(349,obs[i])
}

plot(0:349,the.probs,type="h",lend=2,xlab="No. of patches out of 349 occupied",
  ylab="probability", las=1, col="blue")
```



Let's zoom in on the interesting bits:

```
plot(0:349, the.probs, type="h", lend=2, xlim=c(0,19), lwd=5, las=1,
  xlab="No. of patches out of 349 occupied", ylab="probability",
  main="Probability Mass Function", col="blue")
```



These plots are what are called a *Probability Mass Function*, or *PMF*. A quick search on the wikis shows that the PMF is expressed as

$$\binom{N}{X} \times p^X \times (1-p)^{N-X}.$$

If we were to sum all these probabilities, they would sum to 1.

We can also generate the same plot using the `dbinom` function in R.

Using this PMF, we can calculate the probability (for example) of finding the butterfly in 0 to 5 patches, which would be 0.3006996.

```
sum(dbinom(0:5,349,0.02))
```

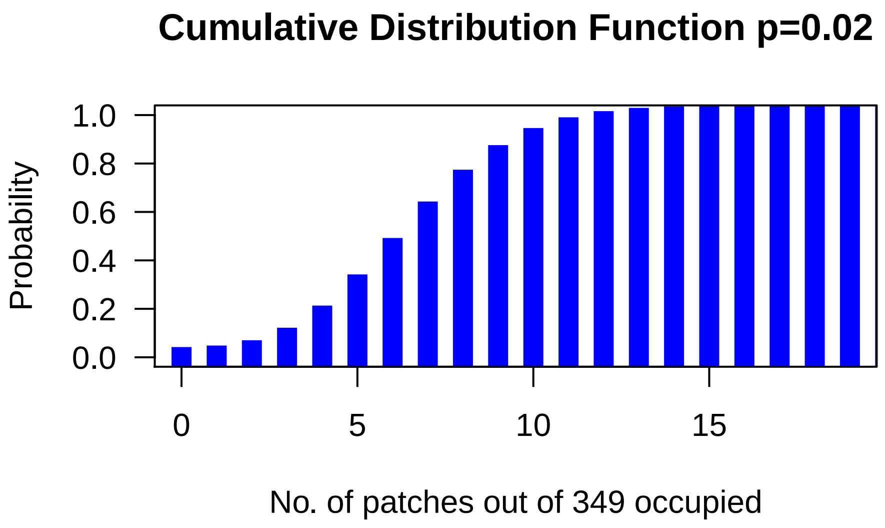
The probability that the butterfly is in greater than 5 patches would be 1 - 0.3006996 = 0.6993004.

If we were to sum the probabilities in a serial manner, we would arrive at what is called the *Cumulative Distribution Function*, or *CDF*. That is, we can calculate the probability of seeing 0 patches occupied, 0 or 1 patches occupied, 0, 1, or 2 patches occupied, and so on.

R has a built in function, `cumsum`, that will calculate the cumulative sum of a vector.

```
cumsum(the.probs)
```

```
plot(0:349,cum.probs,xlim=c(0,19), type="h", lwd=10, las=1,
  main="Cumulative Distribution Function p=0.02",
  xlab="No. of patches out of 349 occupied", ylab="Probability", lend=2, col="blue")
```



The r function `pbinom` will also do this for us.

The CDF is expressed as

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

Where, as before, $n$ is the number of trials and $p$ is the probability of “success”.

`dbinom` and `pbinom` are essentially telling us the same thing in a different way. Where the PMF is telling us the probability of a particular number of observations, the CDF is telling us how these probabilities accumulate.

- For example if we wanted to know the probability that the butterfly is in 5 or fewer patches, we could sum up the first 6 values provided by the PMF `sum(dbinom(0:5,349,0.02))`, or we could simply use the CDF `pbinom(5,349,0.02)`.

## How do we estimate $p$?

Let's say we conduct an experiment to estimate survival of *L. melissa* larvae on alfalfa. Say we have 100 caterpillars. We let them feed on alfalfa and observe 60 live to pupation and 40 die. Let's also say that we will consider a successful caterpillar a 'success'.

- Thus, $X = 60$. Given our data, we would like to calculate the most likely value for $p$.

The maximum likelihood estimate for the parameter $p$ is

$$\frac{number\ of\ successes}{total\ number\ of\ trials}$$

. Or, given our survial data

$$\frac{60}{100} = 0.6.$$

There we have it. The MLE is 0.6 for *melissa* survival on alfalfa for our samples. *But why?*. It's not very satisfying to say "*Because that's what Zach said*".

Let's go back and think about that PMF for a binomial distribution. Because we've chosen to model our data as being random variables drawn from a binomial distribution, we can use the PMF to determine the probability that we would get 60 sucesses and 40 deaths, given that $p = 0.6$.

```r
dbinom(60,100,0.6)
```

```
[1] 0.08121914
```

Similarly, we can also determine the probability of 60 living and 40 dead if, say, $p = 0.5$.

```r
dbinom(60,100,0.5)
```

```
[1] 0.01084387
```

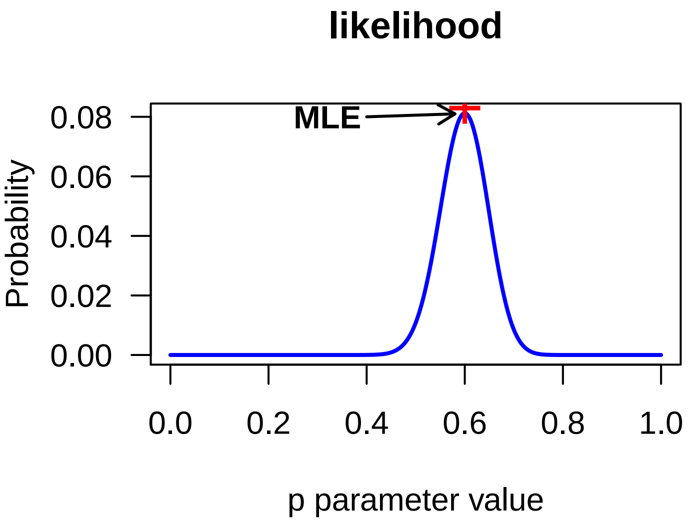
It's not as probable, but it does have a probability.

We can explore the probability of our data across a full range of possible $p$ parameter values and construct a "likelihood surface" for the $p$ parameter given our data. The peak of the curve would be where the slope = 0.

```r
pz<-seq(from=0.00001,to=0.99999,length=1000)

plot(pz, dbinom(60, 100, pz), type="l",main="likelihood", las=1,
  ylab="Probability", xlab="p parameter value", col="blue", lwd=2)

points(0.6,max(dbinom(60,100,pz)),pch="+", cex=2, col="red")
arrows(0.4,0.08,0.58,0.081, length = 0.1, lwd=1.5)
text(0.32,0.08,"MLE", font = 2)
```



So, why is

$$\frac{number\ of\ successes}{total\ number\ of\ trials}$$

the MLE for the parameter $p$?

Let's go back and think about that PMF for a binomial distribution.

$$L(data|parameter) = p^X \times (1-p)^{N-X}$$

Or, we can simply say

$$L = p^X \times (1-p)^{N-X}$$

.

We can take the log to make things easier,

$$\log L = \log(p^X) + \log((1-p)^{N-X})$$

and then take the derivative

$$\log L = X\log(p) + (n-X)\log(1-p).$$

Simple enough right?

Recall the chain rule - first you do the inside, then you do the log. So, the Log$(1-p)$ becomes $(-1)(\frac{1}{(1-p)})$, and Log$(p)$ becomes $\frac{1}{p}$

$$\frac{d\ \log L}{dp} = X\frac{1}{p} + (N-X)\ (-1)(\frac{1}{(1-p)}).$$

A little rearranging and

$$\frac{d\ \log L}{dp} = \frac{X}{p} - \frac{N-X}{1-p}.$$

Now we solve for where the derivative = 0.

$$0 = \frac{X}{p} - (\frac{N - X}{1 - p})$$

$$(\frac{N - X}{1 - p}) = \frac{X}{p}$$

$$pN - pX = X - pX$$

$$pN = X$$

$$p = \frac{X}{N}$$

And there you have it. The MLE for $p$ is the number of successes divided by the number of trials.