140.615.HW.1.Delahanty.Jeremy

Jeremy Delahanty

2024-02-11

1. It has been determined that the chance of becoming infected by a disease after contact with a particular parasite is 15%. If 50 individuals were in contact with the parasite:

For all future uses of the word Probability in my document's mathematical descriptions, I will use `Pr()`. For this question, I will reference all relevant events as `Infected` or `Not Infected` for clarity. As defined in the Diez, Cetinkaya-Rundel, and Barr textbook:

"The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times."

For this series of questions, we must next state some probability rules defined in lecture:

1. For any event (in this case, being infected by the parasite), the probability of the event occurring is between 0 and 1.

- $0 \leq \ Probability(Infected) \ \leq 1$

Next, we must define the sample space.

2. The set of possible outcomes is either `Infected` or `Not Infected`.

- $S = \ \{Infected, \ Not \ Infected\} \ = \ 1$

In this case, the outcomes of being infected or not are "disjoint" or mutually exclusive since it is impossible for someone to be both infected and not infected at the same time.

3. Thus, we can state that the probability of being infected *or* not infected is equal to the sum of their probabilities:

- $Pr(Infected \ or \ Not \ Infected) = Pr(Infected) \ + \ Pr(Not \ Infected)$

Next, we can use the *complement rule* to demonstrate that the probability of *not* being infected is equal to 1 minus the probability of being infected:

- $Pr(Not \ Infected) \ = \ 1 \ - \ Pr(Infected)$

The question states that the chance of becoming infected with the parasite is 15%, or 0.15. We can evaluate the simple function to:

$$\begin{aligned} Pr(Not\ Infected) &= 1 - Pr(Infected) \\ &= 1 - 0.15 \\ &= 0.85 \end{aligned}$$

Finally, although the question does not explicitly state this, I believe it must be assumed that each incidence of exposure is independent of another individual's exposure. Given this information, we know that the multiplication rule is applied as the following:

- $Pr(Infected\ and\ Not\ Infected) = \ Pr(Infected)\ \times\ Pr(Not\ Infected)$

With these definitions, we can now move forward answering the question.

1a) What is the chance that no-one got infected?

The multiplication rule states further, as defined in Diez et al, that "if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is:

- $P(A_1) \times P(A_2) \times \cdots \times P(A_k)$

Using R, we can calculate this as:

```r
# Probability someone is NOT infected after exposure:
p_not_infected <- 0.85

# Define number of individuals, n:
n <- 50

# Use multiplication rule:
p_nobody_infected <- p_not_infected^n

cat("Pr(Zero Infections) =", p_nobody_infected)
```

```
## Pr(Zero Infections) = 0.0002957647
```

We can see that the chance nobody got infected is incredibly low with a value of 0.00029576.

1b) What is the chance that at most 10 people got infected?

In the case that we have a fixed number of trials (here, 50 individuals that may or may not get infected after exposure to the parasite) and want to determine the probability that at most 10 are infected, we might want to use the *binomial distribution*.

The binomial distribution, as defined in Diez et al, "describes the probability of having exactly k successes in n independent trials. Our earlier definitions for this problem demonstrate we satisfy each of the four conditions for using the binomial distribution:

1. The trials are independent.

This was assumed for the problem despite not being explicitly stated.

2. The number of trials, n, is fixed.

The problem defines we only have 50 individuals.

3. Each trial outcome can be classified as a success or failure.

In our case, an individual can either be “infected” (success) or “not infected” (failure).

4. The probability of a success, p, is the same for every trial.

It is stated that the probability of being infected is 0.15.

Having satisfied these conditions, we can define the relevant values n (trials), k (number of successes), p (probability of success), and q (probability of failure): - n = 50 - k = # successes (Infections) - p = 0.15 - q = 0.85

Since we want to calculate the total probability of each of these successes occurring (0, 1, 2, 3, . . . , 10), we effectively want to calculate:

$Pr(Infected\ \le 10)$

To do that, we need to sum the probabilities of each of those occurrences:

- $Pr(0) + Pr(1) + Pr(2), ..., Pr(10)$

We can use addition because the probabilities of each of these events is mutually exclusive.

The general form for calculating the probability of exactly k number of successes (infections) is described as:

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

R has a built in tool for calculating the probability of k successes called `dbinom`. We can iterate through all values $k = 1, ..., k = 10$ using a for loop or the specialized function `sapply` which is allegedly faster:

```r
# Use n defined in first part of question, n  = 50
# Define probability of infection
p_infected <- 0.15

# Use vectorized apply on a vector with values 0 through 10 and:
# define function taking each number iterated from the vector (k)
# use function dbinom()
# describe value of k, or number of infections with 50 trials and a probability of p_infected
probabilities <- sapply(0:10, function(k) dbinom(k, size=n, prob=p_infected))

# Sum the calculated probabilities
p_at_most_10 <- sum(probabilities)

# Display the values
cat("Probability of at most 10 infections is", p_at_most_10)
```

```
## Probability of at most 10 infections is 0.8800827
```

Thus, we can see that the probability of at most 10 infections is 0.88008268.

1c) What is the chance more than 5 people got infected?

The same justifications for using the binomial distribution described in `1b` can be used here.

We are attempting to calculate:

$Pr(Infected > 5)$

We thus need to sum the probabilities of each of these occurrences:

$Pr(6) + Pr(7) + \cdots + Pr(50)$

This can be accomplished using the same `R` code as above:

```r
# Use n defined in first part of question, n  = 50
# Use p_infected defined in second part of question, p_infected = 0.15

# Use vectorized apply on a vector with values 0 through 10 and:
# define function taking each number iterated from the vector (k)
# use function dbinom()
# describe value of k, or number of infections with 50 trials
# and a probability of p_infected
probabilities <- sapply(6:50, function(k) dbinom(k, size=n, prob=p_infected))

# Sum the calculated probabilities
p_more_than_5 <- sum(probabilities)

# Display the values
cat("Probability of more than 5 infections is", p_more_than_5)
```

```
## Probability of more than 5 infections is 0.7806467
```

Thus, we can see that the probability of more than 5 infections is 0.7806467.

1d) How many people got infected if the calculated probability of this event was 3.3% (rounded?)

This is asking us to go backwards in the binomial distribution calculation. It is effectively asking us what number of infections would be required to achieve a probability of 3.3%. In other words:

$0.033 = Pr(x)$

Where x represents the number of infections needed to achieve this probability.

We should be able to use the `inverse cumulative density function` or `inverse CDF` which calculates the value of a random variable for a specific probability.

Using `R`, we can do the following:

```r
# Use n defined in first part of question, n  = 50
# Use p_infected defined in second part of question, p_infected = 0.15

# Define the probability we are trying to find the value of k for:
calculated_probability = 0.033

# Use qbinom for calculating the value of k
```

```r
number_infected = qbinom(calculated_probability, size=n, prob=p_infected)

# Display the value
cat("Number of infections to achieve Pr=0.033:", number_infected)
```

```
## Number of infections to achieve Pr=0.033: 3
```

```r
# We can check the accuracy of this computation by plugging in
# this value into the dbinom function:
p = dbinom(number_infected, size=n, prob=p_infected)

# Display the values
cat("Calculated Probability from qbinom value: ", p)
```

```
## Calculated Probability from qbinom value:  0.03185806
```

#TODO: In office hours, check out why this doesn't line up exactly, must have dones something wrong.

2. A rare genetic disease with late onset is discovered. Although only one in a million people are carriers for the variant, you consider getting screened. You are told that the genetic test is extremely accurate: it is 100% sensitive (i.e., it is always correct if you have the disease) and 99.99% specific (i.e., it gives a false positive result only 0.01% of the time). Having recently learned Bayes' theorem, you decide not to take the test after all. Why?

To answer this question, it would first be helpful to define what the terms sensitivity and specificity mean. Then, a discussion of positive and negative predictive value would be helpful.

Sensitivity

Sensitivity can be thought of as the true positive rate of the test. In other words, it is the probability of a positive test result given that the individual truly is positive for a given disease. The test being always correct given you have the disease is indeed impressive sounding! Represented as a probability, it is:

$Pr(Positive\ Test \mid Disease)$

And can be calculated by:

$Sensitivity = True\ Positive/(True\ Positive\ +\ False\ Negative)$

It must be noted that sensitivity is independent of the prevalence of the disease in the population.

Specificity

Specificity can be thought of the true negative rate of the test. In other words, it is the probability of a negative test result given that the individual truly is negative for the disease. Represented as a probability, it is:

$Pr(Negative\ Test \mid Disease)$

And can be calculated by:

$Specificity = True\ Negative/(True\ Negative\ +\ False\ Positive)$

It must be noted that specificity is independent of the prevalence of the disease in the population.

The Diagnostic Value: Positive and Negative Predictive Values

To determine if the tests are truly relevant for us, we should consider the positive and negative predictive values of the test. These are directly related to how prevalent a disease actually is.

The positive predictive value describes a probability of having the disease given a positive test. As a probability, it is written:

$Pr(Disease \mid Positive\ Test)$

The negative predictive value describes a probability of *not* having the disease given a negative test. As a probability, it is written:

$Pr(No\ Disease \mid Negative\ Test)$

Why Not Take the Test?

To investigate why it would not be helpful to take the test, we should describe the positive and negative predictive values specifically in the context of an assumed prevalence of 0.1% in the population.

From lecture, we can define a table like so:

	+	-
+	TP	FP
-	FN	TN

This represents the confusion matrix of true positives/negatives as well as false positives/negatives. Represented as probabilities of Positive (P) Test conditioned on Infected (I), we can recreate the table as:

	+	-
+	Pr(P|I)	Pr(P|!I)
-	Pr(!P|I)	Pr(!P|!I)

We can now substitute in each probability given by the specificity and sensitivity.

- $Pr(P \mid I) = 1$
- $Pr(P \mid !I) = 0.01$
- $Pr(!P \mid I) = 0$
- $Pr(!P \mid !I) = 0.99$

If we calculate each of these values using a sample of 100 thousand individuals with a prevalence of 0.1%, we expect 100 individuals to truly have the disease.

Using this expected value, we can obtain the following table:

	+	-
+	100	100
-	0	99900

Now, we can calculate the positive and negative predictive values:

```r
# Calculate positive predictive value and multiply by 100 to get
# percentage
positive_predictive_value <- (100 / (100 + 100)) * 100

# Calculate negative predictive value and multiply by 100 to get
#percentage
negative_predictive_value <- (0 / (0 + 99900)) * 100

cat("Positive Predictive Value: ", positive_predictive_value)
```

```
## Positive Predictive Value:  50
```

```r
cat("Negative Predictive Value: ", negative_predictive_value)
```

```
## Negative Predictive Value:  0
```

We can see here that the test only has 50% positive predictive value and a dismal 0% (!) negative predictive value! In other words, it is completely useless at determining at predicting that you *do not* have the disease. Even if the test result is negative, the probability of the condition being absent is essentially 0!

#TODO: This can't be right, need to check in at office hours + how to demonstrate this with Bayes' rule?

3. Suppose a new drug (B) improves upon an older one (A) by increasing the probability of self-reported relief from gastric distress from 40% to 65%. Suppose drug A is given to 10 randomly-sampled patients and drug B is given to a second 10. Let YA and YB be the number out of 10 patients that will experience relief in each group. Plot the probability distributions for YA and YB.

The binomial distribution is an applicable distribution for creating and then plotting the distributions of these drugs' effects.

1. Are the trials independent?

Each patient is independent of one another for gastric relief (presumably). They take the drugs for themselves and one patient's gastric relief or taking of the drug does not depend on other patients in the study.

2. The number of trials, n, is fixed?

The number of trials (in this case, patients receiving the drug) is fixed at 10 per drug.

3. Each trial outcome can be classified as a success or failure?

Each trial can indeed be classified as a success (relief) or failure (no relief) and have the probabilities of 40% and 65% for successful relief from drug A and B, respectively.

4. The probability of success is the same for each trial?

From the question, it appears that there are no known additonal variables that change the probabilities of patient relief.

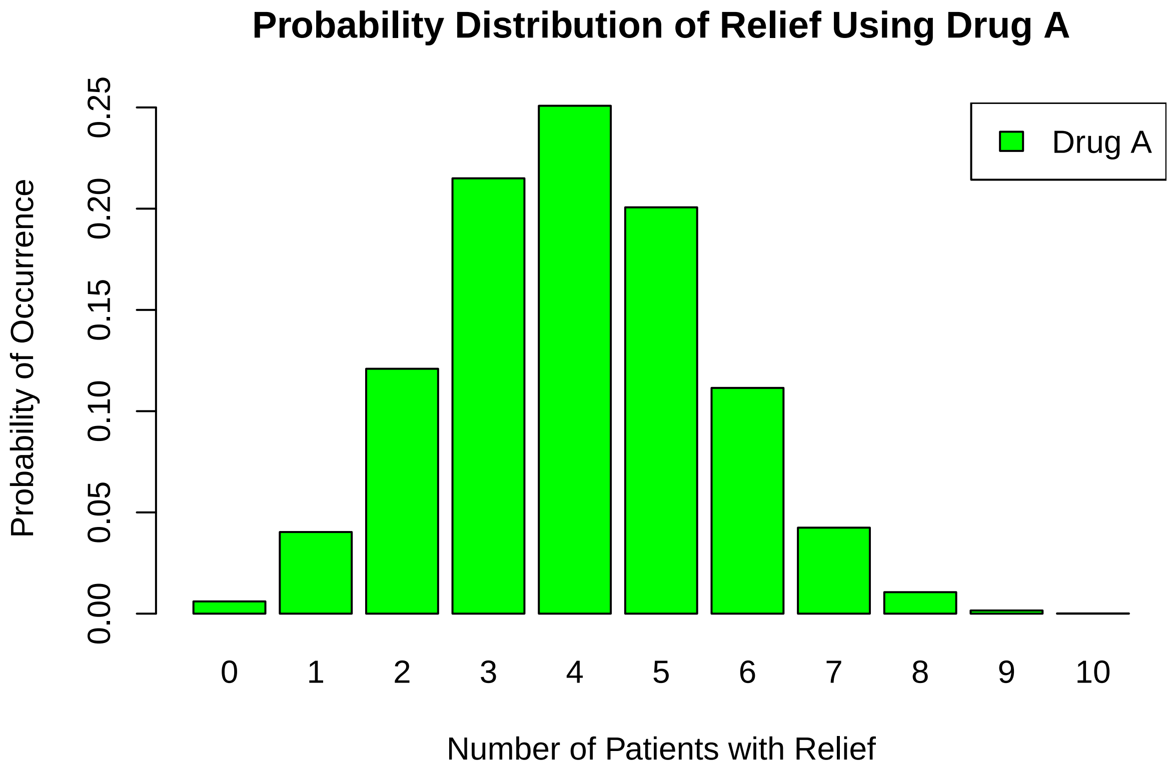
Determining the Distributions: R

We can use R and specifically call upon the dbinom function for this exercise:

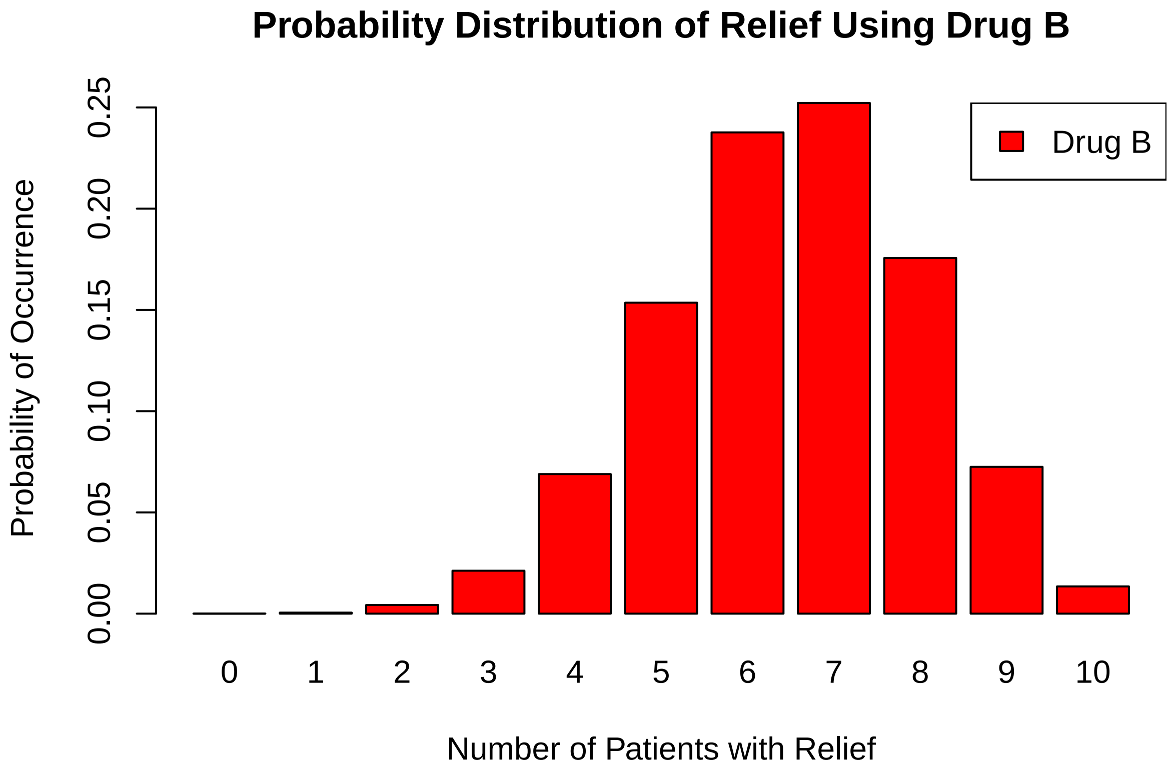
```r
# Define number of patients (n) and probabilities of relief
num_patients <- 10
p_relief_a <- 0.40
p_relief_b <- 0.65

# Make a probability distribution for drug A
prob_distribution_a <- dbinom(0:num_patients, size=num_patients,
                              prob=p_relief_a)

# Make a probability distribution for drug B
prob_distribution_b <- dbinom(0:num_patients, size=num_patients,
                              prob=p_relief_b)
```

Now that we've computed the distributions of these binomial variables, we can try plotting them individually and against one another.

```r
# We want to make a barplot for distribution a
# It has the number of patients expected to have relief on the x axis
# and a y axis describing the probability of that number of patients
# successfully experiencing relief
# So the two graphs are comparable, we want to use the max probability
# between both drug distributions.
# We then define a column color, the title of graph in "main",
# the x and y axis labels, and to have the graphs beside each other.
barplot(
  prob_distribution_a,
  names.arg=0:num_patients,
  ylim=c(0,max(prob_distribution_a, prob_distribution_b)),
  col='green',
  main="Probability Distribution of Relief Using Drug A",
  xlab="Number of Patients with Relief",
  ylab="Probability of Occurrence",
  beside=FALSE
)
# Add a legend
legend("topright", legend=c("Drug A"), fill=c("green"))
```

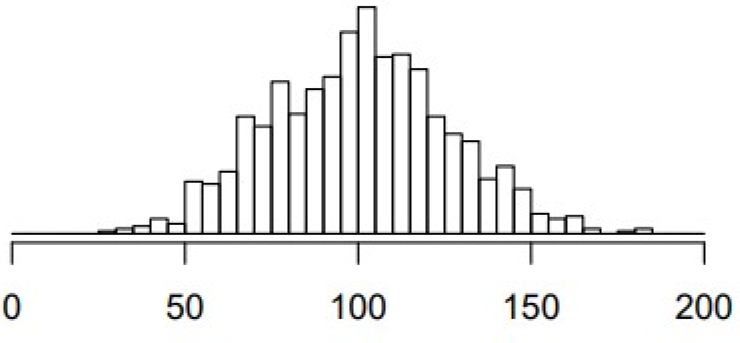


```r
# Now, do the same thing for Drug B
barplot(
  prob_distribution_b,
  names.arg=0:num_patients,
  ylim=c(0,max(prob_distribution_a, prob_distribution_b)),
  col="red",
  main="Probability Distribution of Relief Using Drug B",
  xlab="Number of Patients with Relief",
  ylab="Probability of Occurrence",
)
legend("topright", legend=c("Drug B"), fill=c("red"))
```



4. Consider the following histogram:

```
knitr::include_graphics("C:/Users/jerem/OneDrive/Desktop/hw1_question4_histogram.jpg")
```



4a. Which of the following is true? Choose one and explain briefly.

i. The mean is larger than the median
ii. The median is larger than the mean

iii. The mean and median are about the same

It is hard to definitively state which of these statements are true without the underlying data, but the shape of the distribution gives us a couple hints.

First, it appears unlikely that the mean is larger than the median. The mean is less robust to outliers and skew in the data. We see that the distribution is taller on the left side for the smaller values relative to the larger side of values in this distribution. Thus, it is likely that the *median* is a larger value in this distribution since it is less sensitive to skews in the data.

4b. Is the standard deviation about 10, about 25, or about 50? Explain briefly.

This distribution looks approximately normal. Typically, approximately 99% of the data in a normal distribution is contained within 3 standard deviations, or approximately 3σ. A summary of this typical description is: - $Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68\%$ - $Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$ - $Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99\%$

Through visual inspection, it appears that nearly all the data is present within the range of 25 - 175. With a mean of 100, it the only appropriate value listed in the question is 25. Thus, $\sigma \approx 25$