Homework 1

Kai Aragaki

Problem 1

a. **iii** - the mean and median are about the same. The distribution is roughly symmetric, and there do not appear to be any large outliers (which the mean is sensitive to).

b. Were this a normal distribution, we could assume that a standard deviation would encapsulate roughly 2/3s of the data present. By eye, a range that surrounds the putative mean (~100) by +/- **25** appears to capture roughly 2/3s of these data, while +/- 10 appears to capture too little and +/- 50 too much.

Problem 2

We will assume that when our doctor talked about infection or repair failure risk, he was talking about the risk of getting ONLY an infection or ONLY a fail to repair, not both at the same time. If that is the case:

$Pr(infection) = 3\%$

$Pr(failure) = 14\%$

$Pr(infection\ and\ failure) = 1\%$

Taking these two variables into account, there is only one other outcome that can happen: No infection and no failure. Since the probability of all outcomes must sum to 1,

```
x <- 100 - 3 - 14 - 1
x
```

```
## [1] 82
```

82% of these operations are successful and infection-free.

Problem 3

Specificity and sensitivity do not take prevalence into account. One issue that tests for low prevalence diseases have is a very low positive predictive value. This value informs how we should interpret a positive result if we get one. Let's find out what this value is.

$sensitivity = Pr(test\ pos|is\ pos) = 100\%$

$specificity = Pr(test\ neg|is\ neg) = 99.99\%$

and we want to know

$Pr(is\ pos|test\ pos)$

We know from Bayes' theorem that

$Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$

So we know that

$Pr(is\ pos|test\ pos) = \frac{Pr(test\ pos|is\ pos) \times Pr(is\ pos)}{Pr(test\ pos)}$

$Pr(test\ pos)$ can be divided into two components: tests that are positive and the testee is positive (true positive) and tests that are positive and the testee is negative (false positive):

$Pr(test\ pos) = Pr(test\ pos\ and\ is\ pos) + Pr(test\ pos\ and\ is\ neg)$

We also know that

$Pr(A\ and\ B) = Pr(B) \times Pr(A|B)$

So

$Pr(test\ pos) = Pr(is\ pos) \times Pr(test\ pos|is\ pos) + Pr(is\ neg) \times Pr(test\ pos|is\ neg)$

To simplify, we can say

$Pr(test\ pos) = prevalence \times sensitivity + (1 - prevalence) \times (1 - specificity)$

And plugging this back into Bayes' theorum:

$Pr(ispos|testpos) = \frac{Pr(testpos|ispos) \times Pr(ispos)}{prevalence \times sensitivity + (1 - prevalence) \times (1 - specificity)}$

$Pr(ispos|testpos) = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - prevalence) \times (1 - specificity)}$

```
ppv <- (1e-6) / (1e-6 + ((1 - 1e-6) * 0.0001))
ppv
```

```
## [1] 0.009901
```

The positive predictive value is 0.009901, or 0.9901%. This means that, given a positive test result, only ~1% of them are truly positive cases. In a general population, a positive test tells you very little.

Problem 4

I'm assuming that the probability of the infection of one person is independent of the probability of all others being infected

A

We can use the binomial function for this. Breaking into its components, we ask first 'how many permutations might I expect to see this outcome?'

$\binom{n}{k} = \frac{n!}{k! \times (n-k)!} = \frac{50!}{0! \times (50-0)!} = 1$

Only 1!

Then, to find its probability, we do:

$1 \times (.85)^{50} \times (.15)^{0}$

```r
a <- 1 * (.85)^50 * (.15)^0
a
```

```
## [1] 0.0002957647
```

So ~0.03% chance that no one gets infected.

B

We could do that same exercise as above for k = 10, k = 9, etc… and add all the probabilities together. However, we can also do the following:

```r
b <- pbinom(10, 50, 0.15)
b
```

```
## [1] 0.8800827
```

There is ~88% chance that 10 or fewer people got infected

C

To see if k or *more* individuals got infected, we do

```r
c <- pbinom(5, 50, 0.15, lower.tail = F) # I know. I'm using c as a variable.
c
```

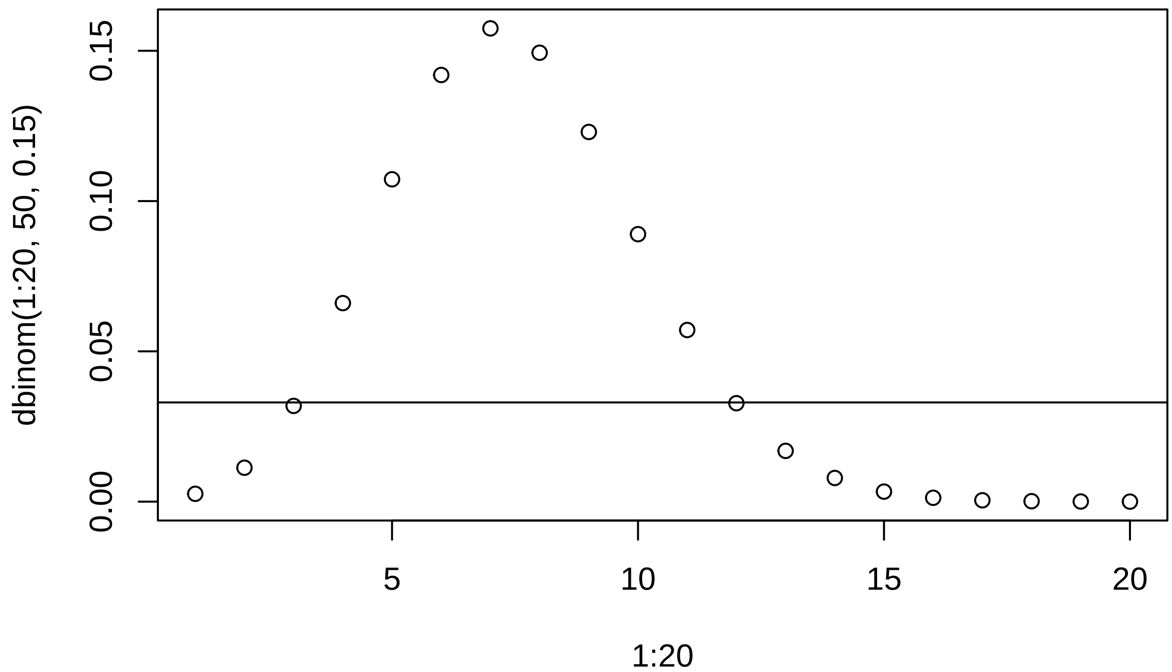
```
## [1] 0.7806467
```

There is ~88% chance that more than 5 people got infected

D

It is difficult to say for certain, since if we look at the probability density

```r
plot(1:20, dbinom(1:20, 50, 0.15)) + abline(h = 0.033)
```




```
## integer(0)
```

we note there are two points that have a probability of around 3.3%: 3 infections, and 12 infections. However, if we calculate the exact probabilities using `dbinom`...

```r
dbinom(3, 50, 0.15)
```

```
## [1] 0.03185806
```

```r
dbinom(12, 50, 0.15)
```

```
## [1] 0.03275154
```

We note that `n = 12` rounds to 0.033 (or 3.3%). This is the most likely answer.