S&DS 238 Problem Set #2

Shaun Radgowski

September 19, 2020

Problem 1

(1a)

The probability of being accepted to the club is the probability of A or B, which is:

$$P(C) = P(A \cup B) = P(A) + P(B) - P(AB)$$

$$= P(A) + P(B) - P(A)P(B) = 0.3 + 0.2 - 0.06 = 0.44$$

(1b)

The probability of being good at bowling given that the individual is good at algebra and also admitted to the Union Club is simply equal to the probability that the individual was good at bowling, because $P(AC) = P(A)$, and A and B are independent events.

$$P(B|AC) = P(B|A) = P(B) = 0.2$$

(1c)

$$P(B|C) = \frac{P(B)P(C|B)}{P(C)} = \frac{0.2}{0.44} \approx 0.455$$

(1d)

Among members of the Union Club, given that one is good at algebra, it *does* affect the conditional probability that the member is good at bowling. Knowing that the member is good at algebra *lowers* the conditional probability that the member is good at bowling from around 0.455 to 0.2.

Problem 2

(2a)

The probability of T happening on the first roll is $\frac{5}{6}$, while the chance that the first time happens on try 2 is the probability of rolling a 3 on the first roll and then rolling *not* a 3 on the second roll, or $\frac{1}{6} * \frac{5}{6}$. This can be projected out for any roll, where the probability of time t being the first instance of a number other than 3 is:

$$P\{T = t\} = \frac{5}{6} * (\frac{1}{6})^{t-1}$$

(2b)

The Law of Total Probability shows us that we can express the probability of A_k conditional on T, as such:

$$A_k = \{T = 1\}A_k \cup \{T = 2\}A_k \cup \ ... = \bigcup_{t=1}^{\infty} \{T = t\}A_k$$

$$P(A_k) = a_k = \sum_{t=1}^{\infty} P\{T = t\}P(A_k|T = t)$$

We also know by its definition that the only values of interest for T are less than 4, because it is certain that three 3s would have been rolled if the first time a non-three digit is rolled is $T \geq 4$. For $T < 4$, we are losing rolls that could have contributed to the three 3s goal, so the number of usable rolls when $T = 3$ is $k - 3$, the number of usable rolls when $T = 2$ is $k - 2$, and so on.

$$P(A_k|T \geq 4) = 1$$

$$P(A_k|T = 3) = a_{k-3}$$

$$P(A_k|T = 2) = a_{k-2}$$

$$P(A_k|T = 1) = a_{k-1}$$

Combining these equations:

$$P(A_k) = a_k = \tfrac{5}{6}a_{k-1} + \tfrac{5}{6^2}a_{k-2} + \tfrac{5}{6^3}a_{k-3} + \sum_{t=4}^{\infty} \tfrac{5}{6} * (\tfrac{1}{6})^{t-1}$$

$$= \sum_{t=1}^{3} \tfrac{5}{6} * (\tfrac{1}{6})^{t-1}a_{k-t} + \frac{\frac{5}{6}}{1-\frac{1}{6}} - \tfrac{5}{6} - \tfrac{5}{36} - \tfrac{5}{216}$$

$$= \sum_{t=1}^{3} \tfrac{5}{6} * (\tfrac{1}{6})^{t-1}a_{k-t} + \tfrac{1}{216}$$

Where $a_k = 0$ for any value of k that is less than 3.

(2c)

To determine the value of a_100, we would need to substitute $k = 100$ into the above equation. The following R code will be useful in that calculation:

```
index <- seq(1, 3)
a_ks <- rep(0, 100)
a_ks[3] <- 1 / 216

for (k in 4:100){
  term_one = sum((5/6)*((1/6)**(index - 1)) * a_ks[k - index])
  a_ks[k] = term_one + (1/216)
}

a_ks[100]
```

```
## [1] 0.318861
```

(2d)

To find the smallest number of rolls k with a probability of seeing at least one run of three consecutive 3s of at least 0.5:

$$0.5 = \sum_{t=1}^{3} \frac{5}{6^{t-1}} a_{k-t} + \sum_{t=4}^{\infty} \frac{5}{6^{t-1}}$$

A few lines of R will show that this minimum value is 180 rolls.

```
index <- seq(1, 3)
a_ks <- rep(0, 500)
a_ks[3] <- 1 / 216

for (k in 4:500){
  term_one = sum((5/6)*((1/6)**(index - 1)) * a_ks[k - index])
  a_ks[k] = term_one + (1/216)
}

min(seq(1, 500)[a_ks > 0.5])
```

```
## [1] 180
```

Problem 3

To check the approximate probability of at least one run of three consecutive 3s in 100 rolls of a die, we could use the following code:

```
# Number of simulations
n = 100000

# Number of dice rolls
k = 100

runsOf3s <- function(x){
  r <- rep(0, length(x))
  if(x[1] == 3) r[1] <- 1
  for(i in 2:length(x)){
    if(x[i] == 3){
      r[i] <- r[i-1] + 1
    }
    else{
      r[i] <- 0
    }
  }
  return(r)
}

maxRunOf3s <- function(x){
  return(max(runsOf3s(x)))
}
```

```r
successes <- 0
for(i in 1:n){
  rolls <- sample(1:6, size = k, replace = T)
  if(maxRunOf3s(rolls) > 2){
    successes = successes + 1
  }
}

# Final proportion of successes
successes/n
```

```
## [1] 0.31643
```

Problem 4

(4a)

Let's use the event conventions that X signifies a student being infected with the virus, X^C signifies a student not being infected with the virus, Y signifies a student testing positive, Y^C signifies a student testing negative, Z signifies a student being isolated, and Z^C signifies a student not being isolated. The test sensitivity (the true positive rate) is $\sigma = P(Y|X) = 0.8$, and the test specificity (the true negative rate) is $\phi = P(Y^C|X^C) = 0.998$. For the one-test plan, the probability of being isolated is simply the probability of testing positive on a single test.

i.

Given a student is infected, the probability that the student is isolated under the one-test plan would just be the probability that the test returns positive.

$$P(Z|X) = P(Y|X) = 0.8$$

ii.

The probability that a random student is isolated is the probability that a random student tests positive. Given that the true proportion of infections is 0.003, then this probability would be:

$$P(Z) = P(Y) = P(Y|X)P(X) + P(Y|X^C)P(X^C)$$

$$= P(Y|X)P(X) + (1 - P(Y^C|X^C))P(X^C)$$

$$= (0.8 * 0.003) + (1 - 0.998)(0.997) \approx 0.0044$$

iii.

According to Bayes' Theorem, the probability of a student being infected given the student being isolated would be:

$$P(X|Z) = \frac{P(Z|X)P(X)}{P(Z)} = \frac{0.8 * 0.003}{0.0044} \approx 0.545$$

(4b)

Let's use the same event conventions as above. For the three-test plan, the probability of being isolated is the probability of testing positive on a single test, times the probability of testing positive on *at least* one of the two follow-up tests.

i.

Given a student is infected, the probability that the student is isolated under the one-test plan would be the probability of testing positive on the first test, times 1 minus the probability of testing negative on both follow-up tests (according to the complement rule).

$$P(Z|X) = P(Y|X) * (1 - P(Y^C|X)^2)$$

$$= P(Y|X) * (1 - (1 - P(Y|X))^2)$$

$$= 0.8 * (1 - (1 - 0.8)^2) = 0.768$$

ii.

The probability that a random student is isolated is the probability that a random student tests positive on the first test and then tests positive on at least one follow-up test. Given that the true proportion of infections is 0.003, then this probability would be:

$$P(Z) = P(Y|X)(1 - P(Y^C|X)^2)P(X) + P(Y|X^C)(1 - P(Y^C|X^C)^2)P(X^C)$$

$$= P(Y|X)(1 - P(Y^C|X)^2)P(X) + (1 - P(Y^C|X^C))(1 - P(Y^C|X^C)^2)P(X^C)$$

$$= (0.8 * (1 - 0.2^2) * 0.003) + (1 - 0.998)(1 - 0.998^2)(0.997) \approx 0.0023$$

iii.

According to Bayes' Theorem, the probability of a student being infected given the student being isolated would be:

$$P(X|Z) = \frac{P(Z|X)P(X)}{P(Z)} = \frac{0.768 * 0.003}{0.0023} \approx 0.998$$

(4c)

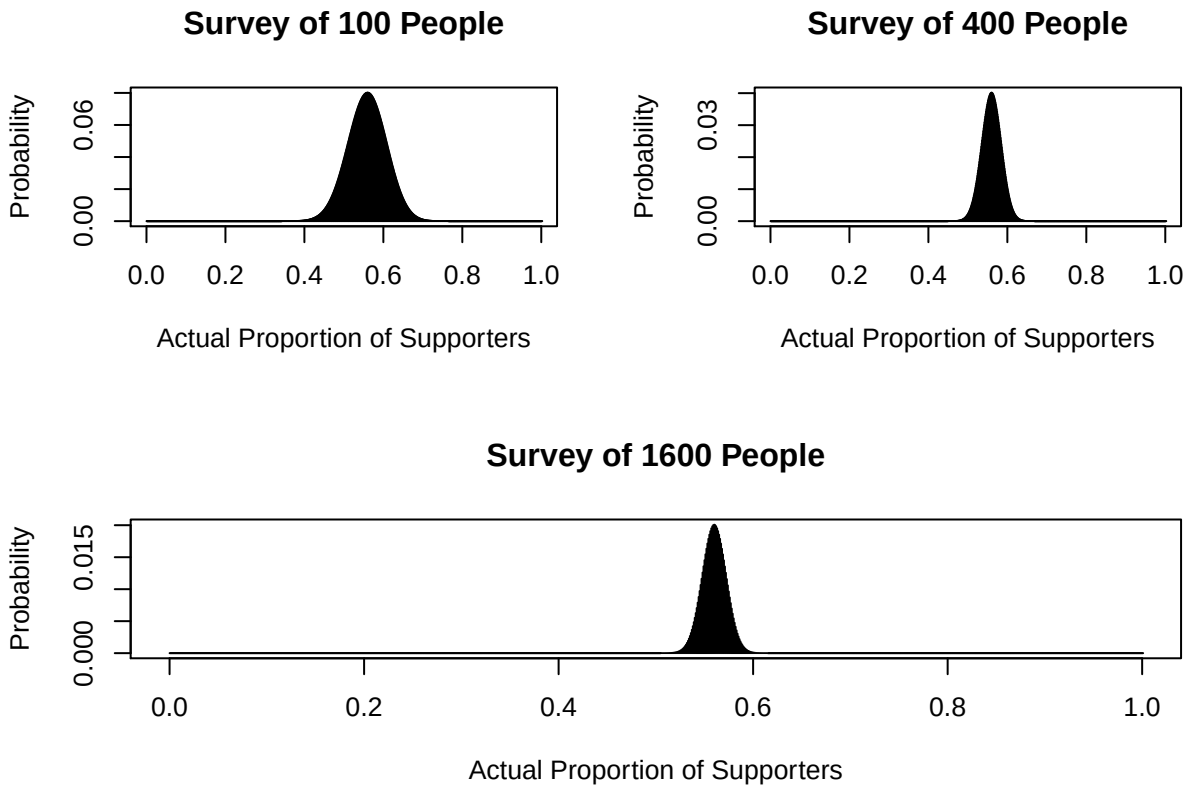
The one-test plan is slightly more effective at isolating infected students (80% of infected students are isolated as compared to 76.8% in the three-test plan). However, the three-test plan is much more effective at ensuring that the students who are in isolation are actually infected students (99.8% of isolated students are rightfully isolated as compared to 54.5% in the one-test plan). If the university's primary concern is to isolate the most infections, the one-test plan would be superior. On the other hand, if the university's primary concern is to avoid wrongfully isolating too many students, the three-test plan accomplishes that goal with only a marginal decrease in the effectiveness of isolating infected students. Of course, the three-test plan would require significantly more tests, which would also be a consideration in choosing between the two plans.

Problem 5

(5a)

i.

```
thetas <- seq(0, 1, 0.001)
k <- length(thetas)

poll <- function(n){
  liks <- dbinom(0.56 * n, n, thetas)
  plot(thetas, liks, main=paste("Survey of", n, "People"),
       xlab="Actual Proportion of Supporters", ylab="Probability", type="h")
}

layout(matrix(c(1, 2, 3, 3), 2, 2, byrow=TRUE))
poll(100)
poll(400)
poll(1600)
```



ii.

```r
posterior <- function(n){
  thetas <- seq(0, 1, 0.001)
  post <- dbinom(0.56*n, n, thetas)
  post <- post/sum(post)
  return(sum(post[thetas > 0.5]))
}

paste("Posterior Probability for n = 100:", posterior(100))
```

```
## [1] "Posterior Probability for n = 100: 0.88187479509042"
```

```r
paste("Posterior Probability for n = 400:", posterior(400))
```

```
## [1] "Posterior Probability for n = 400: 0.991332548845473"
```

```r
paste("Posterior Probability for n = 1600:", posterior(1600))
```

```
## [1] "Posterior Probability for n = 1600: 0.999999057266937"
```

iii.

Let's find a 95% confidence interval by seeing where the cumulative probability is 0.025 (for the lower limit) and where it is 0.975 (for the upper limit).

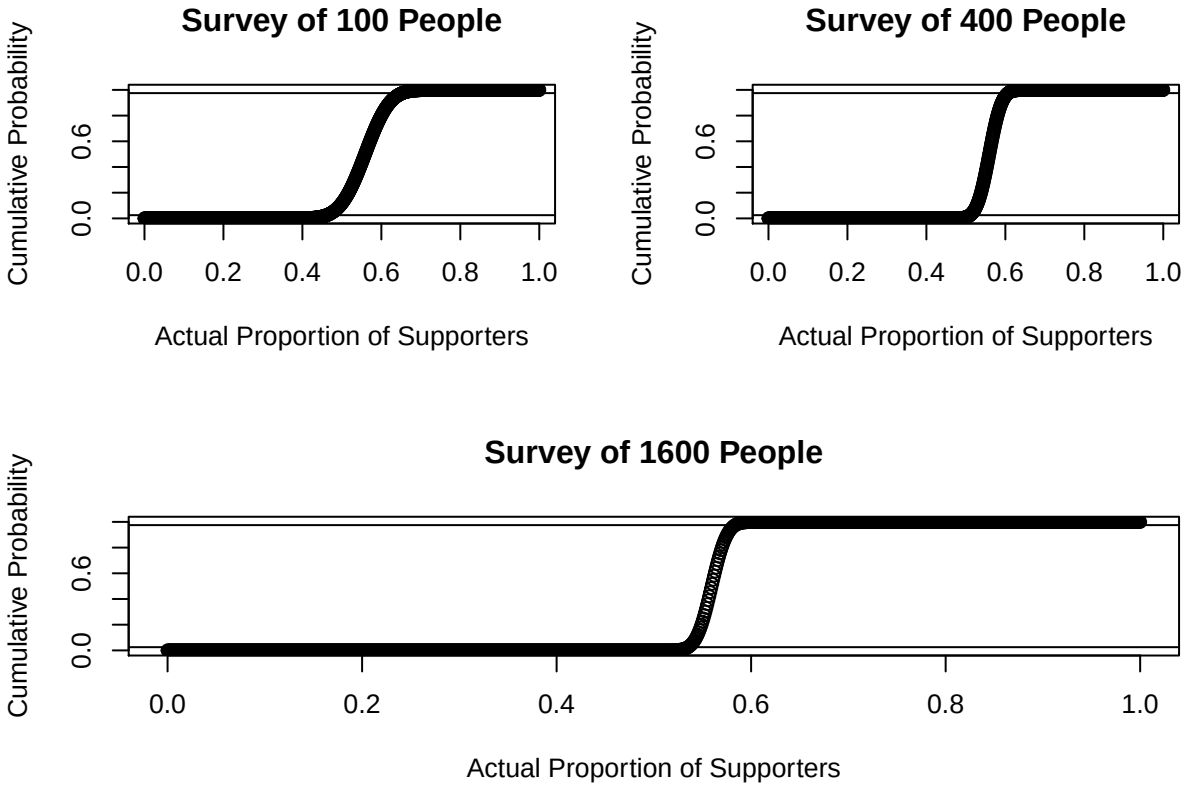
```r
interval <- function(n){
  thetas <- seq(0, 1, 0.001)
  post <- dbinom(0.56*n, n, thetas)
  post <- post/sum(post)
  cs <- cumsum(post)
  plot(thetas, cs, main=paste("Survey of", n, "People"),
       xlab="Actual Proportion of Supporters", ylab="Cumulative Probability")
  abline(h=c(0.025, 0.975))

  L <- min(thetas[cs > 0.025])
  R <- min(thetas[cs >= 0.975])
  cint <- c(L, R)
  return(cint)
}

layout(matrix(c(1, 2, 3, 3), 2, 2, byrow=TRUE))
interval(100)
```

```
## [1] 0.462 0.653
```

```r
interval(400)
```

```
## [1] 0.511 0.608
```

```
interval(1600)
```




```
## [1] 0.536 0.584
```

This code has produced the following 95% Credible Intervals: For n = 100, (0.462, 0.653) For n = 400, (0.511, 0.608) For n = 1600, (0.536, 0.584)

These can be rewritten in terms of margin of error: For n = 100, 0.5575 ± 0.0955 For n = 400, 0.5595 ± 0.0485 For n = 1600, 0.56 ± 0.024

(5b)

When the sample size *increases* by a factor of 4, the margin of error *decreases* by a factor of approximately **2**. This makes statistical sense, because a larger sample size lends to less uncertainty.