Final Report

ISYE 6644 Simulation, Team 259

Pandemic Flu Spread

Ryan Brent Keeney, Aayush Parwal, Wael Ahmad Sultan

Abstract

The Pandemic Flu Spread problem models the spread of an illness within a classroom through peer-to-peer transmission. A preliminary simulation model, in conjunction with derived formulas where applicable, has been developed using R and the outputs are presented.

In this scenario, the spread of an illness is modeled as i.i.d. Bern(p) trials, with a $P_{infection} = 0.02$. A derived solution and simulation based on $Y \sim Binomial(n = 20,\ p = 0.02)$ was utilized to analyze the expected results on day one. Furthermore, another simulation estimated the number of total sick students at the end of day two at 1.94 (including the initial infected student). The average first day with no sick students was 9.13 (5k simulations). 30.1% of the time, the pandemic ends on day 4, with no new infections.

Problem Description and Background

Pandemic Flu Spread is an application-oriented problem. In this scenario, the spread of an illness is modeled in a classroom of 21 students. At the start of the simulation (day 1), 20 students are healthy and 1 student is sick (Tommy). The probability of a sick student infecting another student is $P_{infection} = 0.02$. Each interaction is independent, so the possibility of an infection occurring can be modeled as i.i.d. Bern(p) trials. Sick students are infectious for 3 days, and always come to class. If another student becomes infected, they will then become infectious for 3 days as well, starting on the next day. Once a student has been infectious for 3 days, they are recovered and immune to further infection.

Research Questions

A. What is the distribution of the number of kids that Tommy infects on Day 1?
B. What is the expected number of kids that Tommy infects on Day 1?
C. What is the expected number of kids that are infected by Day 2?
D. Simulate the number of kids that are infected on Days 1,2,. . . . Do this many times. What are the (estimated) expected numbers of kids that are infected by Day i, i = 1, 2, . . .? Produce a histogram detailing how long the “epidemic” will last.

Methods

In parts A and B, our team used a mix of simulation and derivations. In parts C and D, our team built a simulation model to evaluate the results.

Method (Part A and B)

In part A, our team derived the expected values and discrete distributions. A simulation tool was also developed to estimate the discrete distribution. The derived formulas and results are included in the Analysis and Findings.

Method (Part C and D)

A simulation was utilized to analyze parts C and D. The simulations approach can be summarized as follows:

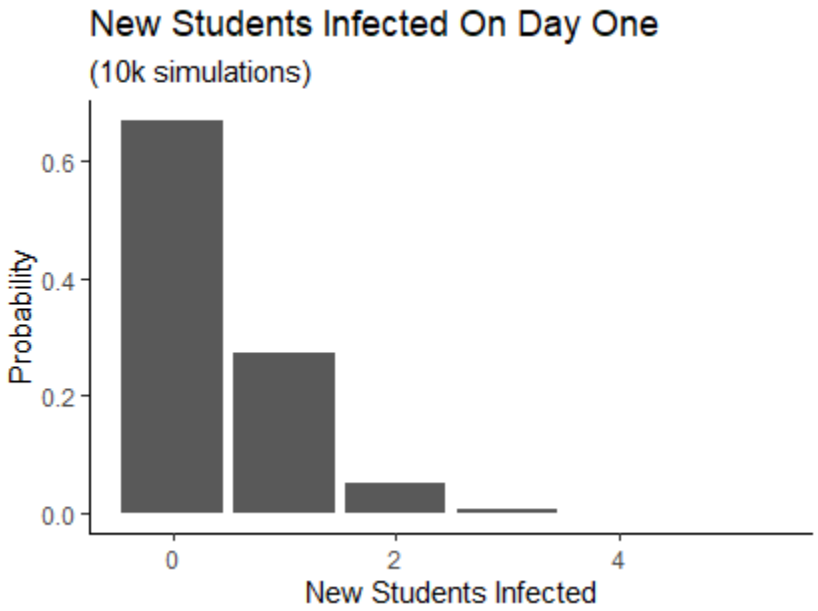
1. Set up a vector of length [n = number of students] that stores the status of the students [0 = not infected, 1 = infected, 2 = recovered]. Each index represents a student's status.
2. Set up a vector of length [n = number of students] that represents the possible number of days that a student may be infectious.
3. For each day in the simulation, track the number of:
    a. Recovered students
    b. Infected students
    c. Not infected students
4. Simulate the number of new infected students using Bernoulli(P=0.02) trials. The number of trials run is equal to the number of students who are eligible to be infected and is repeated for each infected student, or Binomial(n,p). If a "successful" transmission of the illness is passed on by multiple students to the same candidate student, only one infection is recorded (if there are two infected students, they cannot both infect a new non-infected student twice). A more detailed analysis of this methodology is documented in the Appendix, Additional Notes.
5. The number of new infection are tallied, and the following data points are recorded:
    a. Trial number, day number, infected status (recovered, infected, not infected), new infections
6. The student status vectors are updated with new infections being assigned to the left-most indexes.
7. The student possible-infected-days are reduced by 1 so that they can infect new students for a max of 3 days. If a student reaches their max infected days, their status is set to "recovered".
8. The simulation day is stepped forward and the simulation repeats until the maximum day is reached and the simulation is terminated.

This simulation is rerun 5000 times, and the results at each trial and day are saved for analysis. The simulation is designed to accept various inputs for the following attributes:

- Probability of infection
- Class size
- Days infectious
- Starting number of students infected

Analysis and Findings

In part A, the simulation estimates the discrete distribution of the number of new students who become infected during day one with a set of Bernoulli trials equal to the number of students who are eligible to be infected.



The derived distribution based on $Y \sim Binomial(n = 20,\ p = 0.02)$ distribution for the number of infected students on day 1 is given by the following formula.

$$f(y) = \binom{n}{y} p^y q^{n-y},\ y = 0, 1, \ldots, n$$

A table of the derived probabilities is provided. Based on the results, the simulation is performing as expected when the binomial(n,p) trials are staged correctly.

Day One: Infection Distribution

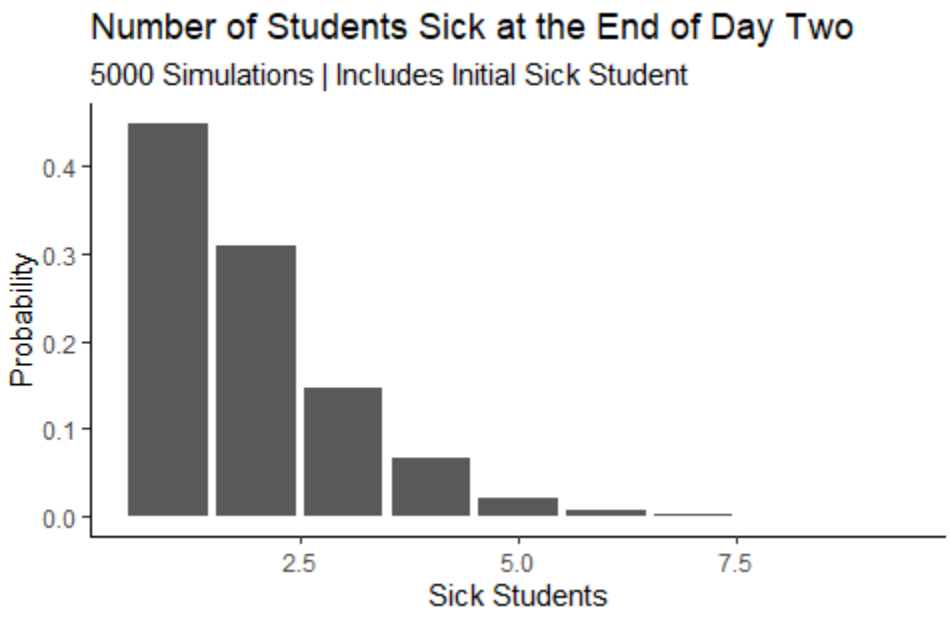
P(Y=y) given Binomial(n,p)

y	p
0	0.6676
1	0.2725
2	0.0528
3	0.0065
4	0.0006
5, … , 20	<0.0001

In part B, the simulation estimated the expected number of students who will be infected by the sick student on day 1 to be 0.41 after 10k trials. The derived expectation is 0.4, given by the following formula.

$$Y \sim Binomial(n = 20,\ p = 0.02)$$
$$E[Y] = np = 0.4$$
$$Var(Y) = npq = 0.392$$

In part C, a simulation was used to estimate the expected number of kids who are infected at the end of Day 2. With 5000 simulations, the estimated number of stick students is 1.94 (including the initial infected student).



The derived distribution is not trivial, due to the cascading effects of multiple binomial(n,p) trials in the conditions where multiple students are infected entering day two. We have chosen to leverage the simulation for this estimation rather than derive the cascading probabilities programmatically for situations where there are more than 1 sick student entering day 2. The derived probability of no additional sick students after day two is 0.45, which matches our simulations. The derived probability of zero, one, and two new infections is recorded below to demonstrate the cascading effect.

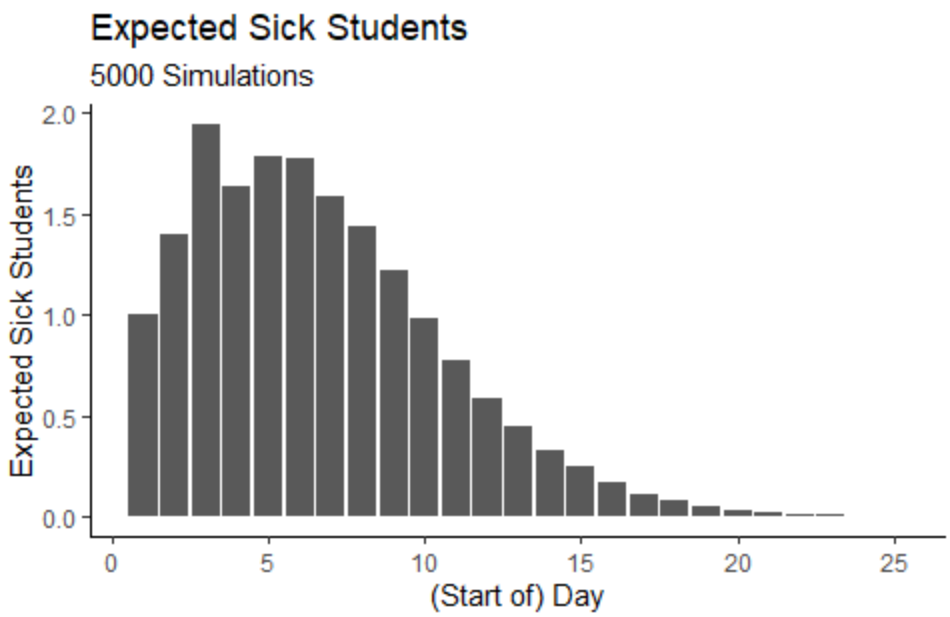
$$\begin{aligned} &Zero\ additional\ infections\text{: } P_{day\ 2\ end}(Y_{day\ 2\ end} = 0) \\ &= P_1(Y_1 = 0) \cdot P_2(Y_2 = 0 \mid Y_1 = 0) \\ &= 0.6676^2 \\ &= 0.4457 \end{aligned}$$

$$\begin{aligned} &One\ \ additional\ infections\text{: } P_{day\ 2\ end}(Y_{day\ 2\ end} = 1) \\ &= P_1(Y_1 = 0) \cdot P_2(Y_2 = 1 \mid Y_1 = 0) \\ &+ P_1(Y_1 = 1) \cdot P_2(Y_2 = 0 \mid Y_2 = 1) \end{aligned}$$

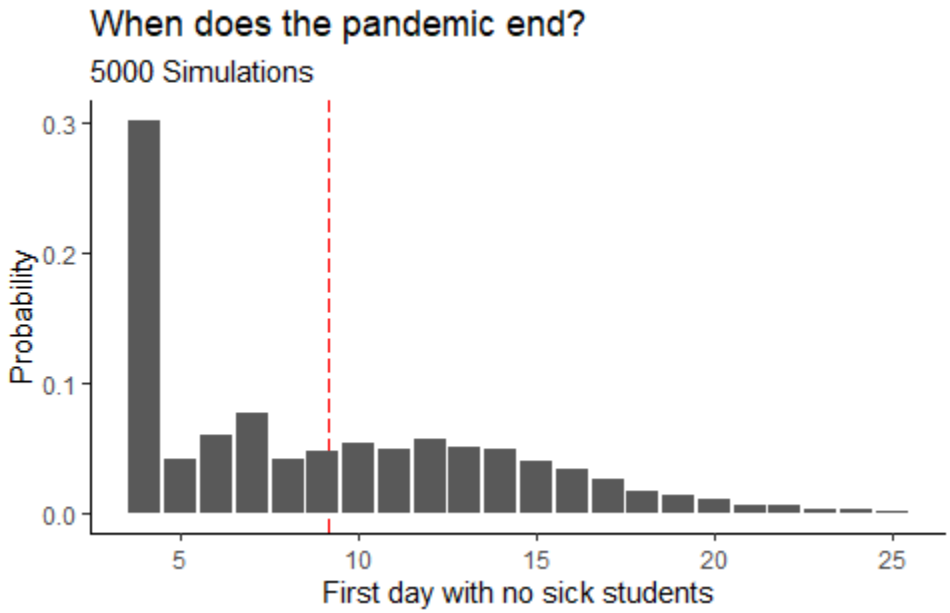
$$\begin{aligned} &Two\ \ additional\ infections\text{: } P_{day\ 2\ end}(Y_{day\ 2\ end} = 2) \\ &= P_1(Y_1 = 0) \cdot P_2(Y_2 = 2 \mid Y_1 = 0) \\ &+ P_1(Y_1 = 1) \cdot P_2(Y_2 = 1 \mid Y_1 = 1) \\ &+ P_1(Y_1 = 2) \cdot P_2(Y_2 = 0 \mid Y_1 = 2) \end{aligned}$$

Importantly, the probability of infections when there are more than 1 sick student entering the day is non-trivial, and depend on multiple binomial trials with varying sizes of n. The simulation accounts for each infection "attempt" by infectious students and does not double-count infections in the tally. For example, if there are two infectious students, and they both successfully infect a student in the same slot in the index, only one successful infection is recorded, not two.

In part D, the expected number of students who are infected at the start of each day was modeled. Day 3 had the highest expectation of sick students, followed by day 5 and 6.



The histogram of the first day with no sick students was also modeled - which is the distribution of how long the pandemic will last. The average first day with no sick students was 9.13 (5k simulations). 30.1% of the time, the pandemic ends on day 4, with no new infections. The second-most common day for the pandemic to end is day 7. The pandemic rarely lasted past 20 days (< 2% of the time). A detailed output summary is documented in the Appendix, Additional Notes.



Conclusions

For this problem, parts A and B (and sections of part C) could be solved through derivations. This step allowed us to evaluate the initial results of our simulation.

Reviewing the problem

The most likely outcome is that no additional students are infected (30.1%). If an additional student is infected, the outbreak is still likely to end with them. However, if 2 students are infected within the first few days, the pandemic is likely to be spread throughout some of the remaining students.

The pandemic rarely lasted past 20 days (< 2% of the time). A detailed output chart is documented in the Appendix, Additional Notes.

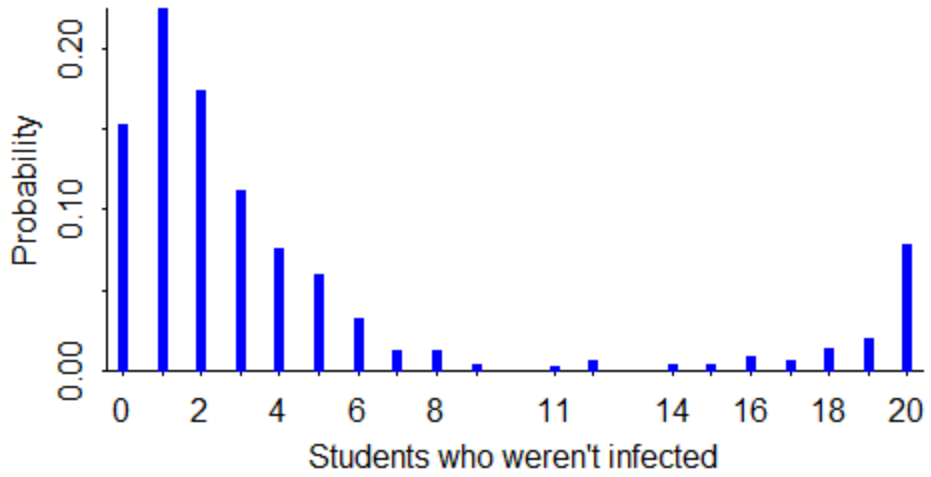
Interestingly, the average number of students who remained uninfected after the outbreak is 15.1. An infection with a $P_{infection} = 0.02$ almost never infects the entire classroom. A plot is provided in the Appendix, Additional Notes.

Simulation as a Essential Tool

While the initial probabilities can be trivially derived; however, the expanded analysis, beyond even 1 day, becomes non-trivial due to the cascading effects of multiple binomial(n,p) trials in the conditions where multiple students are infected entering day two. Importantly, the probability of infections when there are more than 1 sick student entering the day is non-trivial, and depend on multiple binomial trials with varying sizes of n.

Future Work

Our code allows for expanded analysis of the input parameters; infection probability, infection length, initial infections, and classroom size. For example, if probability of infection, $P_{infection}$, is doubled to 0.04, the rate at which the pandemic ends with the initial student is reduced by almost a factor of 4. In the initial study, the pandemic rarely impacted a majority of the students, now the pandemic has a clear bi-modal outcome and it's likely that either no new students are infected, or almost the whole class will eventually become infected.



Utilizing the simulation model, the team could model the impacts of varying starting conditions and provide recommendations for school administrators.

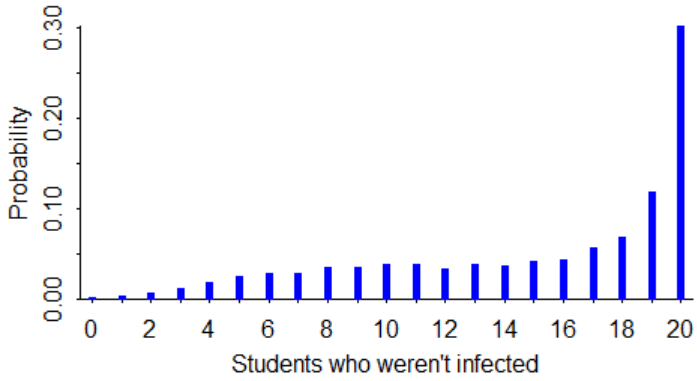
Appendix

Additional Notes

For the modeling of daily new infections, the following method was implemented.

1. Create a matrix of uniform pseudo-random numbers with the following characteristics:
    a. N of rows = students who are not infected and have not recovered (eligible infections)
    b. N of columns = infectious students
2. Set each index to TRUE or FALSE based on the probability of infection
3. Sum each row and set the resulting vector index to ONE if the value is greater than ZERO.
4. Sum the vector, this result is the new number of new infections for that day.

The following chart shows the probability of the number of students who remained uninfected after the outbreak.



For each day in the simulation, the mean outcomes are provided.

Mean Statistics Per Day

5000 Simulations

Day	Sick	Healthy	Recovered	Infections
1	1.0000	20.0000	0.0000	0.4010
2	1.4010	19.5990	0.0000	0.5416
3	1.9426	19.0574	0.0000	0.6978
4	1.6404	18.3596	1.0000	0.5420
5	1.7814	17.8176	1.4010	0.5408
6	1.7806	17.2768	1.9426	0.5012
7	1.5840	16.7756	2.6404	0.3996
8	1.4416	16.3760	3.1824	0.3222
9	1.2230	16.0538	3.7232	0.2592
10	0.9810	15.7946	4.2244	0.1938
11	0.7752	15.6008	4.6240	0.1384
12	0.5914	15.4624	4.9462	0.1140
13	0.4462	15.3484	5.2054	0.0766
14	0.3290	15.2718	5.3992	0.0544
15	0.2450	15.2174	5.5376	0.0366
16	0.1676	15.1808	5.6516	0.0232
17	0.1142	15.1576	5.7282	0.0164
18	0.0762	15.1412	5.7826	0.0112
19	0.0508	15.1300	5.8192	0.0072
20	0.0348	15.1228	5.8424	0.0048
21	0.0232	15.1180	5.8588	0.0022
22	0.0142	15.1158	5.8700	0.0014
23	0.0084	15.1144	5.8772	0.0016
24	0.0052	15.1128	5.8820	0.0004
25	0.0034	15.1124	5.8842	0.0000

Literature Review

Not Applicable.

Code

```
# Libraries ------------------------------------------------------------
library(tidyverse)
library(glue)
library(gt)
library(ggplot2)
library(arm)

# Testing --------------------------------------------------------------

# Idea, model 21 kids as vector, [0,1,2] with 1 = infected, 2 = recovered and associated days possible to be infected [3,2,1,0]

# ?rbinom
# n = students that can be infected
# size = number of trials, or number of kids who can infect others... NO (keep at 1, need to reduce size each time someone else is infected! Don't double count infections)
# prob = chance of each infection
# rbinom(n = 20, size = 4,prob = 0.02) %>% sum()



# Run trials -----------------------------------------------------------

p = 0.02
n_non_sick_students <- 20    # 20
n_sick_students <- 1        # 1
n_days_inf <- 3
n_trials <- 5000
n_days <- 25
trial_output <- tibble()
for (trial in 1:n_trials) {
   # initialize student vector
   students <- rep(0, n_non_sick_students+n_sick_students)
```

```
    # set starting sick students
    students[1:n_sick_students] <- 1

    # set student days of infection
    students_inf_days <- rep(n_days_inf,n_non_sick_students+n_sick_students)

    for (day in 1:n_days) {

        # data for start of day
        n_kids_recovered <- sum(students == 2)
        n_kids_infected <- sum(students == 1)
        n_kids_not_infected <- sum(students == 0)

        # how many more student will be infected?
        # give each infected student a chance to infect the class, if some students become infected, remove them (do
not double count).
        # possible_infected_students <- n_kids_not_infected
        # for (j in 1:n_kids_infected) {
        #     #glue('Round: {j} of {n_kids_infected}') %>% print()
        #     temp_new_infections <- rbinom(n = possible_infected_students, size = 1, prob = p) %>% sum()
        #     #glue('newly infected: {temp_new_infections}') %>% print()
        #     possible_infected_students <- possible_infected_students - temp_new_infections
        #     #glue('remaining possible: {possible_infected_students}') %>% print()
        # }
        # new_kids_infected <- n_kids_not_infected-possible_infected_students

        # idea create a vector of random numbers for each possible infection,
        # repeat for each "try"
        # sum rows, then sum column

        try_to_get_sick_matrix
<-matrix(runif(n_kids_not_infected*n_kids_infected),nrow=n_kids_not_infected,ncol=n_kids_infected)
        new_kids_infected <- sum(rowSums(try_to_get_sick_matrix<=p)>0)

        # turn on or off printing
        if (0==1) {
            glue("--- Start of Day #{day} ---") %>% print()
            glue("sick: {n_kids_infected}") %>% print()
            glue("healthy: {n_kids_not_infected}") %>% print()
```

```
    glue("recovered: {n_kids_recovered}") %>% print()
    glue("new infections: {new_kids_infected}") %>% print()
    print(students)
    print(students_inf_days)
}

# save data to trial.
trial_output_temp <- tibble(
    trial = trial,
    day = day,
    healthy = n_kids_not_infected,
    sick = n_kids_infected,
    recovered = n_kids_recovered,
    new_infections = new_kids_infected
)

trial_output <- trial_output %>% rbind(trial_output_temp)

# set up kids to infect (for countdown)
kids_to_inf <- new_kids_infected

for (n in 1:21) {
    #print(students[n])

    # if a student is sick, reduce their day by 1
    if (students[n] == 1) {
        students_inf_days[n] <- students_inf_days[n]-1
    }

    # if a student has now 0 days of inf, make them recovered
    if (students_inf_days[n] == 0) {
        students[n] <- 2
    }

    # if there is a infection left to "give", student is not sick, infect the student
    if (kids_to_inf > 0 ) {
        if (students[n] == 0) {
            students[n] <- 1
            kids_to_inf = kids_to_inf - 1
```

```
            }
          }

      }

  }
}


# Review
# ALL STATS ARE FOR START OF DAY AND WHO WILL BE INFECTED
trial_output


# Part A-B: Day 1 distribution (independent simulation) ----------------------------------------------------

# new students infected on day 1, 1000 trials
day_1 <- NULL
for (n in 1:10000) {
  #print(n)
  day_1[n] <- rbinom(n = 20, size = 1, prob = 0.02) %>% sum()
}

day_1 %>% mean()
day_1 %>% sd()
day_1 %>% discrete.histogram(xlab="New Students Infected On Day One")

day_1 %>% as_tibble() %>%
  count(value) %>%
  mutate(total = sum(n)) %>%
  mutate(p = n/total) %>%
  ggplot() +
  geom_col(aes(x=value, y = p)) +
  theme_classic() +
  labs(
    title = "New Students Infected On Day One",
    subtitle = "(10k simulations)",
    x = "New Students Infected",
    y = "Probability"
```

```
    )



# Derived day 1 distribution
options(scipen = 999)
der_1 <- tibble(
      y = 0:4,
      p = round(choose(20,y)*p^y*(1-p)^(20-y),4)
      ) %>%
      rbind(c("5, ... , 20","<0.0001"))

der_1 %>% gt() %>%
      tab_header(
            title = md("Day One: Infection Distribution"),
            subtitle = html("<em>P(Y=y) given Binomial(n,p)</em>")
      )

# Day 2 is simulated, the actual distribution depends:
#      Day 1 results
#      P(Y=y) with repeated trials for each infected student.
#      This is not trivial



# Part C: Day 2 infected ----------------------------------------------------

# select day 3, this is number of people sick (and able to infect) at the END of day 2
trial_output %>% filter(day ==3) %>% pull(sick) %>% mean()

trial_output %>% filter(day ==3) %>%
    dplyr::select(sick) %>%
    count(sick) %>%
    mutate(total = sum(n)) %>%
    mutate(p = n/total) %>%
    ggplot() +
    geom_col(aes(x=sick, y = p)) +
    theme_classic() +
    labs(
        title = "Number of Students Sick at the End of Day Two",
```

```
        subtitle = glue("{n_trials} Simulations | Includes Initial Sick Student"),
        x = "Sick Students",
        y = "Probability"
    )


# Part D: End of pandemic ----------------------------------------------------

# Expected # of infected on each day
#graphic
trial_output %>%
    group_by(day) %>%
    mutate(m_sick = mean(sick)) %>%
    filter(trial == 1) %>%
    ggplot() +
    geom_col(aes(x=day, y = m_sick)) +
    theme_classic() +
    labs(
        title = "Expected Sick Students",
        subtitle = glue("{n_trials} Simulations"),
        x = "(Start of) Day",
        y = "Expected Sick Students"
    )

# table
trial_output %>%
    group_by(day) %>%
    mutate(
        Sick = mean(sick),
        Healthy = mean(healthy),
        Recovered = mean(recovered),
        Infections = mean(new_infections),
        Day = day
        ) %>%
    ungroup() %>%
    filter(trial == 1) %>%
    dplyr::select(Day, Sick,Healthy,Recovered, Infections) %>%
    gt() %>%
    tab_header(
```

```
      title = md("Mean Statistics Per Day"),
      subtitle = glue("{n_trials} Simulations")
    )



# End of pandemic
avg_day_end = trial_output %>%
      group_by(trial) %>%
      filter(sick == 0) %>%
      filter(day == min(day)) %>%
      ungroup() %>%
      arrange(trial) %>%
      dplyr::select(day) %>%
      count(day) %>%
      mutate(total = sum(n)) %>%
      mutate(p = n/total) %>%
      mutate(ex_day = day*p) %>%
      pull(ex_day) %>% sum()

avg_day_end

trial_output %>%
    group_by(trial) %>%
    filter(sick == 0) %>%
    filter(day == min(day)) %>%
    ungroup() %>%
    arrange(trial) %>%
    dplyr::select(day) %>%
    count(day) %>%
    mutate(total = sum(n)) %>%
    mutate(p = n/total) %>%
    ggplot() +
      geom_vline(xintercept = avg_day_end,color='red', linetype = "longdash") +
      geom_col(aes(x=day, y = p)) +
      theme_classic() +
      labs(
        title = "When does the pandemic end?",
        subtitle = glue("{n_trials} Simulations"),
```

```
    x = "First day with no sick students",
    y = "Probability"
  )


# Number of students who didn't get sick
trial_output %>% group_by(trial) %>%
    summarize(spared=min(healthy)) %>% pull(spared) %>%
    discrete.histogram(xlab="Students who weren't infected")
```