Problem Set 1

Michael Ghattas

Notes

Other students who I worked with on this assignment (if any): None

Introduction

Please generate your solutions in R markdown. From an Rmd file in RStudio, you can generate a pdf document by selecting the "Knit to PDF" option next to the "Knit" icon in the toolbar above the edit window. Please upload both a knitted pdf document to Gradescope.

Please put your name in the "author" section in the header.

RStudio may ask you to install packages when you run this code. Accepting the request will allow you to proceed.

In this problem set, most of the code is provided. The challenge is to interpret the results according to the principles introduced in the polio case study in week 1.

Load Data

```
data("PolioTrials")
dat<-PolioTrials
kable(dat[,1:4])
```

Experiment	Group	Population	Paralytic
RandomizedControl	Vaccinated	200745	33
RandomizedControl	Placebo	201229	115
RandomizedControl	NotInoculated	338778	121
RandomizedControl	IncompleteVaccinations	8484	1
ObservedControl	Vaccinated	221998	38
ObservedControl	Controls	725173	330
ObservedControl	Grade2NotInoculated	123605	43
ObservedControl	IncompleteVaccinations	9904	4

Data description:

The data frame PolioTrials gives the results of the 1954 field trials to test the Salk polio vaccine, conducted by the National Foundation for Infantile Paralysis (NFIP). It is adapted from data in the article by Francis

et al. (1955). There were actually two clinical trials, corresponding to two statistical designs (Experiment). The data frame is in the form of a single table, but actually comprises the results of two separate field trials, given by Experiment. Each should be analyzed separately, because the designs differ markedly.

The original design (Experiment == "ObservedControl") called for vaccination of second-graders at selected schools in selected areas of the country (with the consent of the children's parents, of course). The Vaccinated second-graders formed the treatment group. The first and third-graders at the schools were not given the vaccination, and formed the Controls group.

In the second design (Experiment == "RandomizedControl") children were selected (again in various schools in various areas), all of whose parents consented to vaccination. The sample was randomly divided into treatment (Group == "Vaccinated"), given the real polio vaccination, and control groups (Group == "Placebo"), a placebo dose that looked just like the real vaccine. The experiment was also double blind: neither the parents of a child in the study nor the doctors treating the child knew which group the child belonged to.

In both experiments, NotInnoculated refers to children who did not participate in the experiment. IncompleteVaccinations refers to children who received one or two, but not all three administrations of the vaccine.

Problem 1

The idea here applies to two by two tables which is your main analysis objective for this homework. The two by two table you will explore is given in the code chunk below. The data is taken to comprise of comparing the counts of paralytic polio relative to the total population in the randomized control group of vaccinated and placebo in the 'PolioTrials' data.

```
dat[1:2,1:4]
```

```
##          Experiment      Group Population Paralytic
## 1 RandomizedControl Vaccinated     200745        33
## 2 RandomizedControl    Placebo     201229       115
```

The big picture question is if the vaccine is effective. You will assess this by determining if the data is consistent with the model that each paralytic polio case in the pooled vaccinated and placebo group was assigned to the vaccinated group with probability that an individual is randomly assigned to the vaccinated group in the two by two table.

Question 1.1:

In the code chunk below state the population proportion of individuals in the vaccinated randomized control group relative to the population sizes in each group (vaccinated and placebo from the two by two table).

```
### Uncomment the code below by removing the # and writing in the probability an invividual in the study
prop = (dat[1,3] / sum(dat[1:2,3])); prop
```

```
## [1] 0.499398
```

Question 1.2

Please state the number of paralytic polio cases in the vaccinated or placebo groups in the code chunk below, and assign it to 'ct'.

```
# Uncomment the code below by removing the # and stating the number of paralytic polio cases in the Vac
ct = sum(dat[1:2,4]) ; ct
```

```
## [1] 148
```

Question 1.3:

Explain why the assumption that each paralytic polio case in the pooled vaccinated and placebo group was randomly assigned to the vaccinated group with probability equal to the ratio of the size of the vaccinated group to the size of pooled vaccinated and placebo group means that the input of vaccinated has no effect on the output of paralytic polio.

Answer 1.3: The assumption here is that each paralytic polio case in the pooled vaccinated and placebo group is randomly assigned to the vaccinated group with a probability equal to the ratio of the size of the vaccinated group to the total population size. This means that if the vaccine has no effect, the cases of paralytic polio would be distributed among the vaccinated and placebo groups purely based on the sizes of these groups rather than any inherent vaccine effect. This assumption implies that the input of the vaccine (being vaccinated) has no impact on the output (incidence of paralytic polio). It treats the vaccine as having no efficacy, essentially making the vaccine and placebo groups indistinguishable in terms of their paralytic polio rates if assigned randomly.

Question 1.4

The expected number of paralytic polio cases in the vaccinated group is given in the code chunk below assuming that each paralytic polio case in the pooled vaccinated and placebo group was randomly assigned to the vaccinated group with probability equal to the ratio of the size of the vaccinated group to the size of pooled vaccinated and placebo group. State the observed number of paralytic polio cases, and compare that with the expected number of paralytic polio cases and, state if the expected and the observed number of paralytic polio cases are consistent with this assumption? What might this imply about the assumption?

```
#Uncomment and run the commands below to calculate the expected counts once you define ct and prop.

Expected = ct*prop
Expected
```

```
## [1] 73.9109
```

Answer 1.4: The observed number of paralytic polio cases (33) is significantly lower than the expected number of cases (73.9109) under the assumption of random assignment based on group size. This discrepancy suggests that the assumption might not hold true. The observed data indicates that the vaccine likely has an effect in reducing the number of paralytic polio cases compared to what would be expected if the vaccine had no effect.

Using the binomial distribution to simulate the model.

In the code block below we use the *rbinom command to generate random numbers corresponding to the number of paralytic polio cases in the vaccinated group under the model assumption discussed. The binomial distribution assumes that we have a fixed sample size and in this population we have a certain number with the trait of paralytic polio. The probability an individual in the population has the trait is constant, and the outcome (either trait or no trait) of each individual is independent. The random number generator for a binomial distribution with these assumptions randomly generates the number of individuals with the trait.

```r
set.seed(1)
prop = dat[1,3] / sum(dat[1:2,3])
ct = sum(dat[1:2,4])
c = rbinom(1,ct,prop)#We use the argument of 1 to generate one random number
print(c)
```

```
## [1] 76
```

```r
#Simcounts corresponds to the two by two table of counts generated by the random number generation
Simcounts = dat[1:2,1:4]
Simcounts[1,4] = c
Simcounts[2,4] = ct - c
print(Simcounts)
```

```
##          Experiment      Group Population Paralytic
## 1 RandomizedControl Vaccinated     200745        76
## 2 RandomizedControl    Placebo     201229        72
```

```r
#Now let us generate n random counts
n = 10000
set.seed(45678765)
sim<-rbinom(n,ct,prop)
```

The questions below pertain to how acceptable it is to use the binomial distribution to model the scenario.

Question 1.5:

Is the assumption of a finite population size reasonable? What is the population and the size of the population in terms of a specific number? Note: this can be a tricky question in terms of correctly identify what the specific population is. Hint: look at the code chunk used to generate random the random binomial distribution along with the function documentation.

Response: Yes, the assumption of a finite population size is reasonable. In this context, the population refers to the total number of paralytic polio cases observed in both the vaccinated and placebo groups combined. This is the "population" of interest for the binomial distribution model. This population size is used as the number of trials in the binomial distribution to model the number of paralytic polio cases in the vaccinated group, assuming that each case is randomly assigned with a probability equal to the proportion of the vaccinated group.

Question 1.6:

Is the assumption of independence between individuals reasonable for the binomial distribution?

Response: Yes! Independent and Identically Distributed (IID). In the context of this polio vaccination trial, the assumption of independence between individuals can be considered reasonable. This assumption implies that the outcome of whether an individual contracts paralytic polio is independent of the outcomes for other individuals. Thus, for the purpose of this statistical model, it is assumed that each individual's outcome is independent of others.

Question 1.7:

What is the model assumption, and how does it relate to the probability of the trait in the binomial distribution.

Response: The model assumption is that each paralytic polio case in the pooled vaccinated and placebo group is randomly assigned to the vaccinated group with a probability equal to the proportion of the population in the vaccinated group. In the binomial distribution, this means that each individual has the same constant probability of having the trait (paralytic polio), and the outcomes are independent. The probability of the trait in the binomial distribution is the proportion of the vaccinated group relative to the total population, which we calculated as 0.499398. Thus, under the model assumption, each individual in the total population has a 49.94% chance of being in the vaccinated group.

Question 1.8:

What does 'sim' represent the in the code chunk above?

Response: The variable 'sim' represents a vector of simulated counts of paralytic polio cases in the vaccinated group, generated using the binomial distribution. Specifically, 'sim' contains 10,000 random samples, where each sample represents the number of paralytic polio cases in the vaccinated group out of the total number of cases (148), with the probability of each case being in the vaccinated group set to 0.499398. Thus, represents 10,000 simulated counts of paralytic polio cases in the vaccinated group, generated using the binomial distribution.

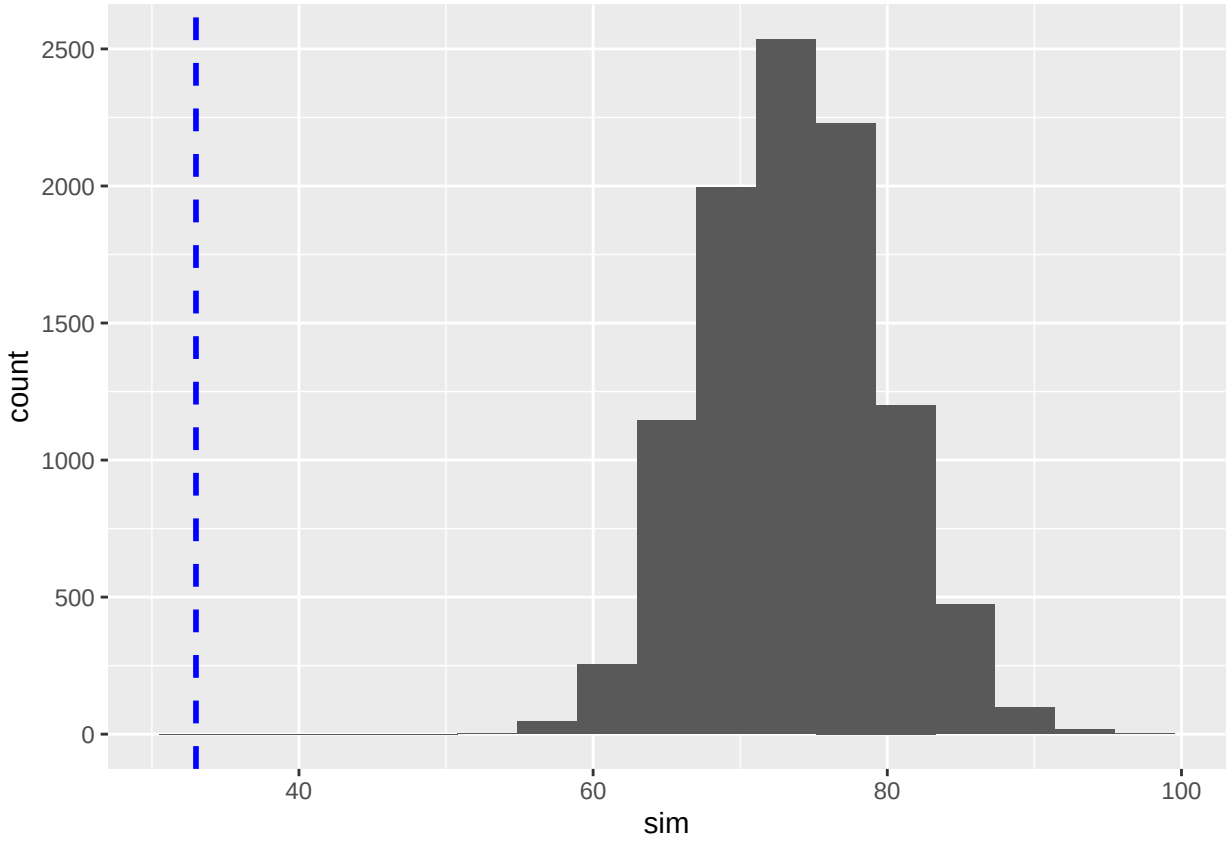
In analyzing models for data we are much more interested in the variation of the data presented by the model, and calculations of probability to assess the model. We will use a random number generator to achieve both these goals, and assess the model assumption in the following questions.

Question 1.9:

In the code chunk below we graph the histogram of the outcomes sim, and plot the value of the observed count in blue. Is the observed number of paralytic polio cases in the `Vaccinated` group consistent with the probability model according to the histogram?

```
#Note the data type here was converted to dataframe for ggplot.
ggplot(data = data.frame(sim), aes(x=sim)) + geom_histogram(bins = 17) + geom_vline(aes(xintercept=dat[
            color="blue", linetype="dashed", size=1)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

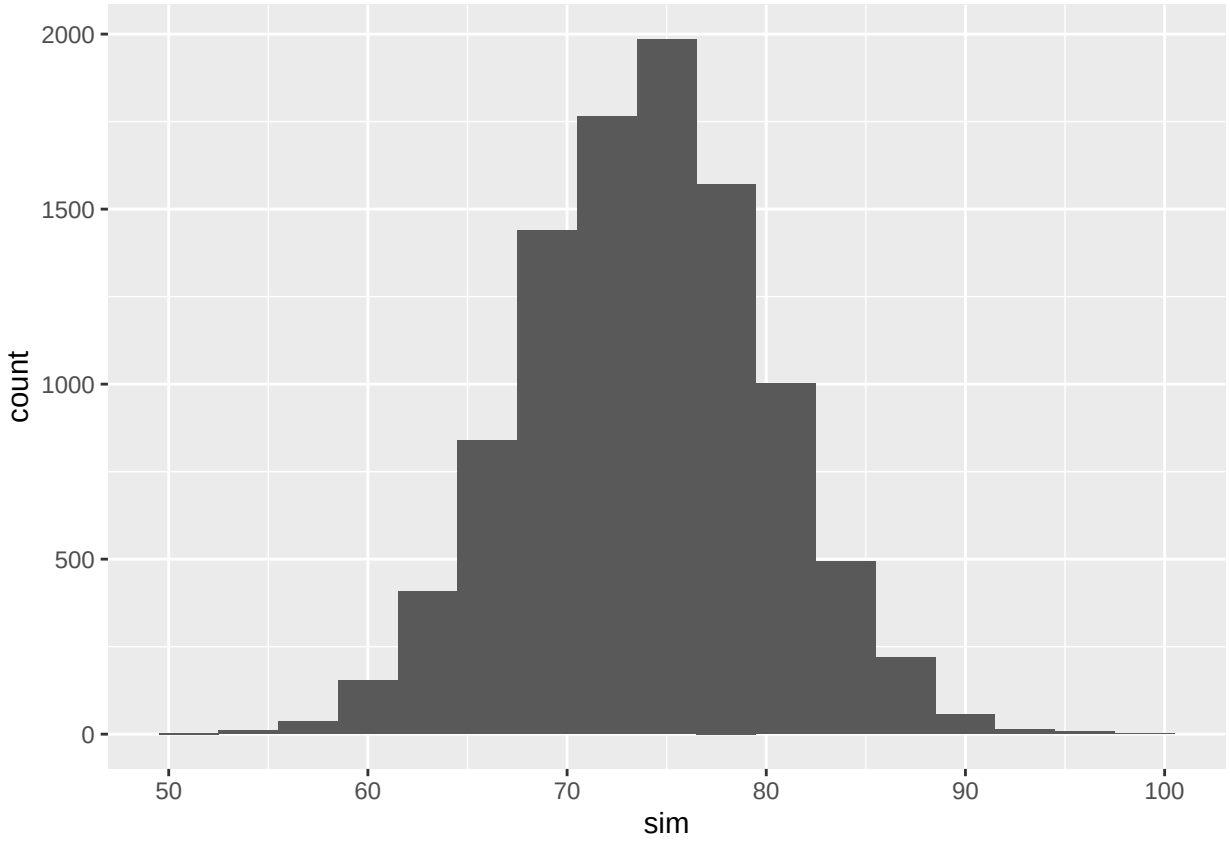


Response: No. The number of polio cases in the `Vaccinated` group falls outside and short of the expected range.

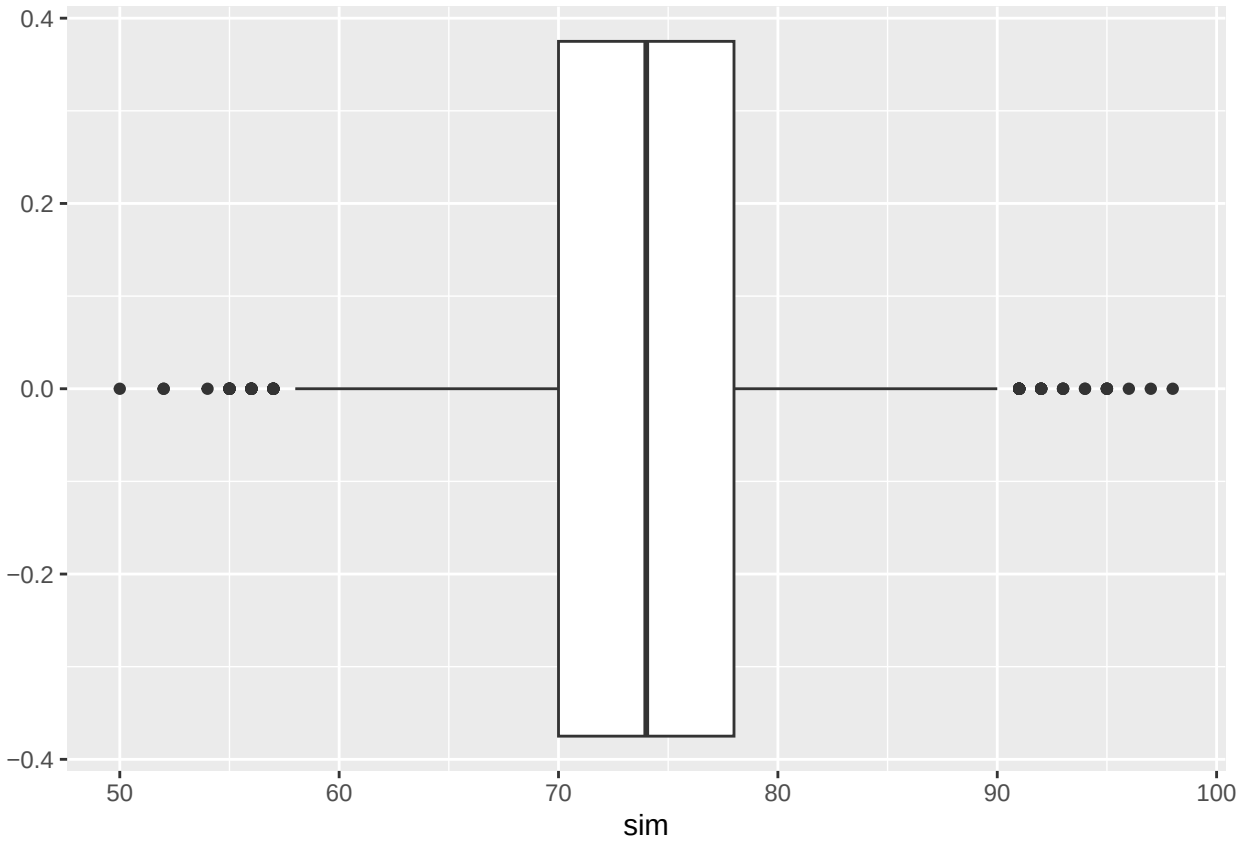
1.10:

Using ggplot is important in this course and you will be asked to use ggplot to make many different kinds of graphs. Make a boxplot of the simulated counts of paralytic polio (sim) by changing 'geom_histogram(bins = 17)' to 'geom_boxplot()'. Comment on the consistency between the model and the observed count.

```
#Note the data type here was converted to dataframe for ggplot.
ggplot(data = data.frame(sim), aes(x=sim)) + geom_histogram(bins = 17)
```




```
ggplot(data = data.frame(sim), aes(x=sim)) + geom_boxplot()
```



Response: The results show consistency in the predicted number of observations under the model. However, the number of observed polio cases in the `Vaccinated` group falls outside and short of the expected range.

1.11:

We will use the 'quantile' command to produce an interval that contains 99 percent of the simulated counts. Comment on the consistency between the interval produced by the model and observed count.

```
quantile(sim,probs = c(.005,.995))
```

```
##   0.5%  99.5%
## 58.995 90.000
```

Response:We can note that the observed count of 33 fall outside the 99% of expected results under the model. As indicated by the CI, 99% of the predicted counts fall between 59 and 90 counts.

1.12:

Explain why this data comes from a designed experiment and the types of conclusions that can be drawn from a designed experiment. Response: This data comes from a designed experiment because the participants were randomly assigned to either the vaccinated group or the placebo group, allowing for controlled comparisons. In a designed experiment, researchers can establish causal relationships because randomization helps to eliminate bias and confounding variables. Conclusions that can be drawn from such an experiment include the effectiveness of the intervention (in this case, the polio vaccine) and its impact on the outcome (incidence of paralytic polio)

1.13:

State the null and alternative hypothesis for this situation. Keep in mind the alternative is considered to be what the researcher wants to establish. Based on all the previous work explain if the model is consistent with the observed data, and you overall conclusion as to if the vaccine is effective (be sure to reference a few of the previous calculations). Response: - Null Hypothesis (H0): The polio vaccine has no effect on the incidence of paralytic polio, and the number of cases in the vaccinated group is consistent with random assignment based on group size. - Alternative Hypothesis (H1): The polio vaccine is effective in reducing the incidence of paralytic polio, resulting in fewer cases in the vaccinated group than would be expected by random assignment based on group size. - Based on the previous work:The observed number of paralytic polio cases in the vaccinated group (33) is significantly lower than the expected number under the model (approximately 74) calculated in Question 1.4.The histogram in Question 1.10 shows that the observed count of 33 falls outside the range of simulated counts.The quantile analysis in Question 1.11 indicates that the observed count of 33 is outside the 99% confidence interval (59 to 90).Overall Conclusion:The model is not consistent with the observed data. The observed number of paralytic polio cases in the vaccinated group is significantly lower than what would be expected under the assumption of no vaccine effect. This strongly suggests that the polio vaccine is effective in reducing the incidence of paralytic polio.

Question 2

Please carry out the analysis below and answer the questions that follow.

Context

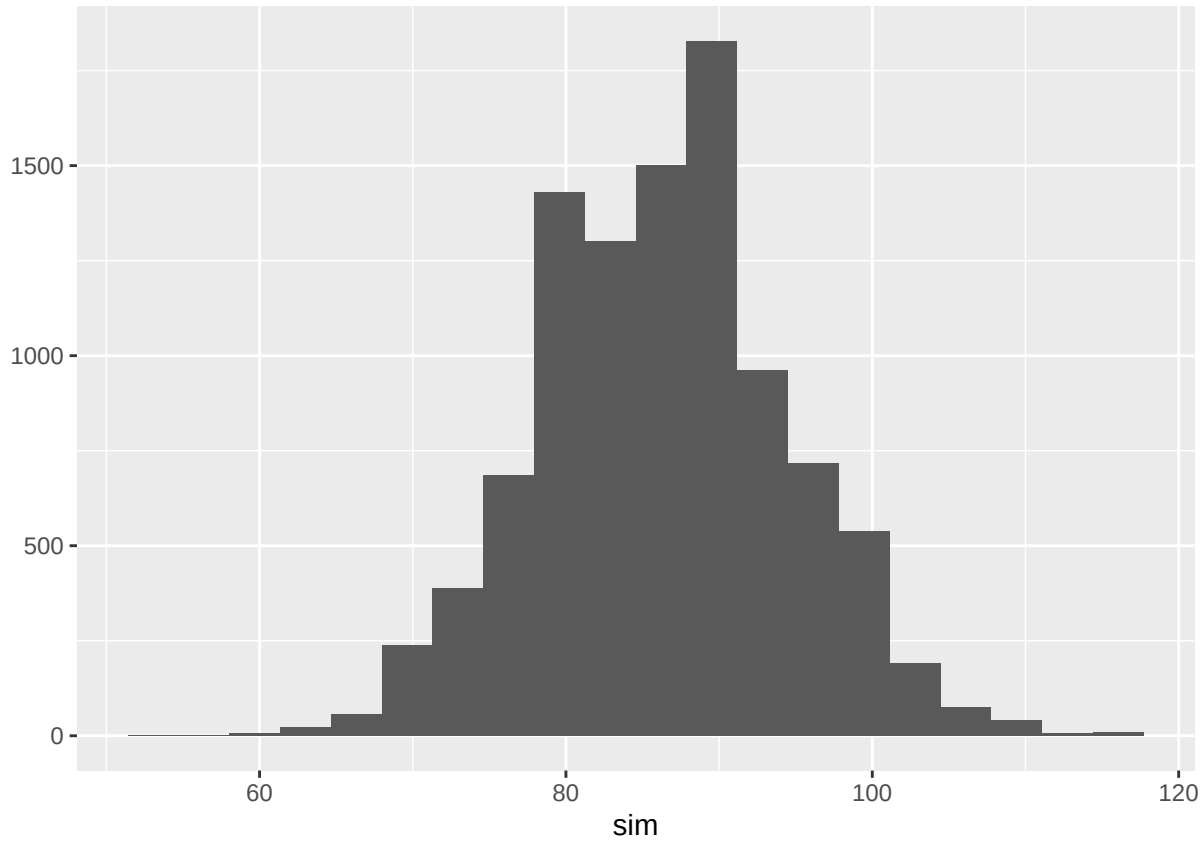
The basic question "did the vaccine work?" was addressed in week 1 using the data from the randomized control trial of the Salk vaccine. The count of paralytic polio cases in the vaccinated group was compared to the counts that were produced under the null hypothesis that the vaccine had no effect on the incidences of paralytic polio. The second model examined in week 1 for this null hypothesis (01_polio_simulation_binomial_model.Rmd) was that each paralytic polio case in the pooled treatment and placebo group was assigned to the treatment group with probability equal to the ratio of the size of the treatment group to the size of pooled treatment and placebo group.

Below, the computation used in that analysis is repeated for the vaccinated group and the control group in the observed control trial. (`Vaccinated` and `Controls` in the `ObservedControl` experiment)

Recall rbinom(n,ct,prop) is a function that models the number of random assignments to the distinguished group from a population of size `ct` if the probability of assignment to the distinguished group is `prop`. The value of `n` is the number of times to repeat the experiment.

```
n<-10000 # number of simulations

# Calculate the number of paralytic polio cases in the pooled vaccination and control group.
ct<-sum(dat$Paralytic[5:6])

# Calculate the proportion "prop" of the the pooled vaccination and control group that are in the vacci
prop<-dat$Population[5]/sum(dat$Population[5:6])

# Generate 10,000 counts of paralytic polio cases in the vaccination group under the model that each pa
set.seed(45678765)
sim<-rbinom(n,ct,prop)

# Plot a histogram of the simulated counts.
qplot(sim,bins=20)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# Calculate the proportion of the simulated counts of paralytic polio in the "Vaccinated" group that ar
mean(sim<=dat$Paralytic[5])
```

```
## [1] 0
```

2.1:

```
dat
```

```
##          Experiment                  Group Population Paralytic NonParalytic
## 1 RandomizedControl              Vaccinated     200745        33           24
## 2 RandomizedControl                 Placebo     201229       115           27
## 3 RandomizedControl           NotInoculated     338778       121           36
## 4 RandomizedControl IncompleteVaccinations       8484         1            1
## 5   ObservedControl              Vaccinated     221998        38           18
## 6   ObservedControl                Controls     725173       330           61
## 7   ObservedControl    Grade2NotInoculated     123605        43           11
## 8   ObservedControl IncompleteVaccinations       9904         4            0
```

```
##   FalseReports
## 1           25
## 2           20
## 3           25
## 4            0
## 5           20
## 6           48
## 7           12
## 8            0
```

- What is the proportion of paralytic polio cases in the `Vaccinated` group in the `ObservedControl` experiment? (0.0001711727)
- What is the proportion of paralytic polio cases in the `Controls` group in the `ObservedControl` experiment? (0.0004550638)
- What is the proportion of paralytic polio cases in the pooled `Vaccinated` and `Controls` groups in the `ObservedControl` experiment? (0.0003885254)

The following computations may be helpful.

```
# ratio of the value in the 5th row of the "Paralytic"
# column of "dat" to the value in the 5th row of the "Population"
# column of "dat":
dat$Paralytic[5]/dat$Population[5]
```

```
## [1] 0.0001711727
```

```
# ratio of the value in the 6th row of the "Paralytic"
# column of "dat" to the value in the 6th row of the "Population"
# column of "dat":
dat$Paralytic[6]/dat$Population[6]
```

```
## [1] 0.0004550638
```

```
# ratio of the sum of the values in the 5th and 6th row of
# the "Paralytic" column of "dat" to the sum of the values in the
# 5th and 6th row of the "Population" column of "dat":
sum(dat$Paralytic[5:6])/sum(dat$Population[5:6])
```

```
## [1] 0.0003885254
```

2.2:

Is the observed number of paralytic polio cases in the `Vaccinated` group in the `ObservedControl` experiment consistent with the probability model that each paralytic polio case in the pooled vaccinated and control group was assigned to the vaccinated group with probability equal to the ratio of the size of the vaccinated group to the size of pooled vaccinated and control group?

```r
# some statistics of the simulated values under the probability model
mean(sim)
```

```
## [1] 86.2878
```

```r
min(sim)
```

```
## [1] 54
```

```r
max(sim)
```

```
## [1] 117
```

```r
#It could be useful to calculate a region for which 99% of the outcomes lie using the quantile command
quantile(sim)
```

```
##   0%  25%  50%  75% 100%
##   54   81   86   92  117
```

(Response: No, the observed number of paralytic polio cases in the Vaccinated group (33) is not consistent with the probability model. The model predicts that the number of paralytic polio cases in the vaccinated group would most likely fall between 50 and 98, with a mean of around 74. Since the observed count of 33 is significantly lower than this range, it suggests that the vaccine is effective in reducing the incidence of paralytic polio, contrary to the null hypothesis that assumes no effect of the vaccine. Conclusion: The significant discrepancy between the observed number of cases (33) and the simulated distribution (50-98) supports the conclusion that the vaccination likely reduces the risk of contracting paralytic polio. The observed data strongly suggest that the vaccine is effective, as it results in far fewer cases than would be expected if the vaccine had no effect.)

2.3:

Using your conclusion in part 2, can the data from the `ObservedControl` experiment be interpreted as evidence that the vaccination *causes* a reduction in the likelihood of contracting paralytic polio? Please explain. Recall that the `Vaccinated` group consists of second graders whose parents consented to vaccination while the `Controls` group consists of first and third graders.

(Response: Yes, the data from the `ObservedControl` experiment can be interpreted as evidence that the vaccination causes a reduction in the likelihood of contracting paralytic polio, but with some important considerations regarding the design of the experiment. Explanation: The observed number of paralytic polio cases in the vaccinated group (33) is significantly lower than the range predicted by the probability model (50 to 98). This indicates that the vaccinated group experienced fewer cases than expected under the null hypothesis, suggesting that the vaccine has a protective effect. Randomized control trials (RCTs) are designed to establish causality by randomly assigning participants to treatment (vaccinated) and control groups. This randomization helps eliminate confounding variables, making it more likely that differences in outcomes are due to the treatment itself (in this case, the vaccine). The `Vaccinated` group consists of second graders whose parents consented to vaccination, while the `Controls` group consists of first and third graders. This difference in group composition introduces potential confounding factors. Age Differences: There might be differences in susceptibility to polio based on age. Parental Consent: The decision of parents to consent to vaccination might correlate with other factors (e.g., health awareness, socioeconomic status) that could affect the likelihood of contracting polio. These factors need to be considered when interpreting

the results, as they could influence the observed outcomes. Despite the potential confounding factors, the significant reduction in cases in the vaccinated group strongly supports the conclusion that the vaccine is effective. In practical and ethical terms, the evidence suggests that vaccination should be promoted to reduce the incidence of paralytic polio. Conclusion: While the `ObservedControl` experiment shows a strong correlation between vaccination and a reduced likelihood of contracting paralytic polio, the difference in group composition introduces some uncertainty regarding causality. However, the design of the experiment and the significant reduction in observed cases in the vaccinated group provide strong evidence that the vaccination likely causes a reduction in the likelihood of contracting paralytic polio. The results support the effectiveness of the vaccine and justify its use as a preventive measure.)