Lesson 5 MA206Y

Nicholas Clark

Admin

Exploration Lesson 4

1. Based on the description of the study, state the research question:

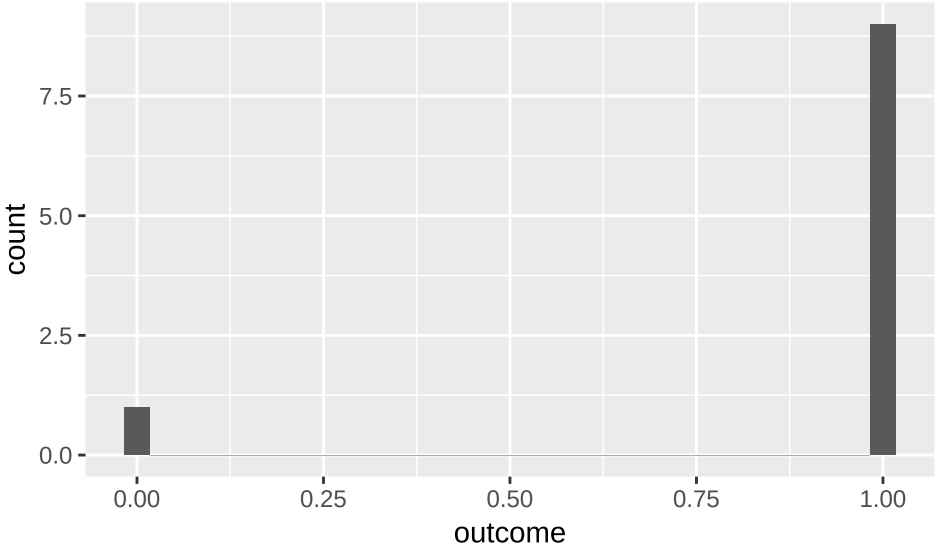
Can dogs understand both human and nonhuman gestures

2. What are the observational units?

Harley's trials

3. Correct or incorrect, this is categorical
4. 10
5. Observed statistic is 9/10

```
library(tidyverse)
trials=data.frame(outcome=c(rep(1,9),0))
ggplot(data=trials,aes(x=outcome))+geom_histogram()
```



6. Yes it is, 9/10 is greater than 50 %
7. Sure, certainly *possible*
8. Doubtful
9. π = Probability of picking currect cup, if $\pi = .5$ he is guessing at random
10. If $\pi > .5$ he is not just guessing at random
11. If he is guessing at random, that could be represeented by 'heads'. So we could flip a coin 10 times to represent one set of his 10 attempts
12. If he was guessing at random, on average we would expect 5 heads

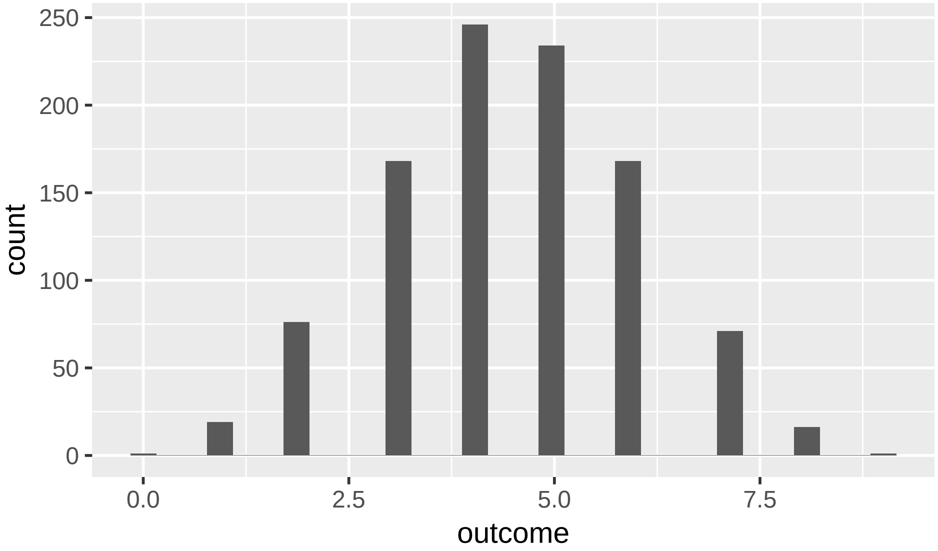
13.

```r
ledger=data.frame(students=seq(1,18),outcome=NA)
for(i in 1:18){
  ledger[i,]$outcome=rbinom(1,9,.5)
}

#ggplot(ledger,aes(x=outcome))+geom_dotplot() #Source:http://www.sthda.com/english/wiki/ggplot2-dot-plo
```

From here we see none of our students got 9 heads, so very unlikely that Harley was just guessing which cup to choose

14.

```r
ledger=data.frame(students=seq(1,1000),outcome=NA)
for(i in 1:1000){
  ledger[i,]$outcome=rbinom(1,9,.5)
}

ggplot(ledger,aes(x=outcome))+geom_histogram() #I like this better than dotplot
```



To better see how rare 9 is we can find:

```r
ledger %>%filter(outcome >= 9)%>%
  summarise(total=n())
```

```
##   total
## 1     1
```

So this is definitley an unlikely result

15. Our result is in the very right hand tail of the histogram so very unlikely to have picked the correct dish 9 times if he was just guessing
16. Yes, the results appear to be statistically significant (unlikely to happen by chance)
17. No, the results suggest that Harley isn't just guessing

Lesson 5

People spend a lot of money on bottled water. But do they really prefer bottled water to ordinary tap water? Researchers at Longwood University investigated this question by presenting people who came to a booth a local festival with four cups of water. Three cups contained different brands of bottled water, and one cup was filled with tap water. Each **subject** was asked which of the four cups of water they most preferred. Researchers kept track of how many people chose tap water in order to see whether tap water was chosen significantly less often than would be expected by random chance.

What is the research question that the researchers hoped to answer?

Identify the observational units in this study.

Identify the response variable. Is it quantitative or categorical?

As a binary variable this can be written as:

$$i = \text{Person} Y_i = 1 \text{ If tap water selected } 0 \text{ Otherwise}$$

What is the parameter of interst?

If subjects have equal preference among all four waters (meaning you are getting ripped off by buying expensive water!), what is the long-run proportion a subject would select tap water?

If subjects are less likely to prefer tap water than bottled water what values would π take on?

In statistics we sometimes have a **null hypothesis**. The null hypothesis (denoted as H_0) typically represents the “by random chance alone” explanation. In terms of π we can write this as:

The **alternative hypothesis** is the “there is an effect” explanation. In terms of our parameter, π we can write this as:

The researchers found that 3 out of 27 subjects selected tap water. This means our **estimate** of π, or $\hat{p}$ is:

What is the value of n or the sample size for this study?

Note that our H_0 and H_a are about the TRUE values of the parameter, π, but our estimate of π, $\hat{p}$ isn't specified in H_0 and H_a. Why?

If H_0 is true, meaning $\pi = 1/3$ would it be possible to observe $\hat{pi} = 3/27$?

Let's go back to our ledger:

```r
possible.outcomes=2 #This is still 2, either we observe tap water or we don't
p=1/4 #This is the probability of obtaining a 1
sample.size=27 #This is n
num.experiments=1000 #I have 1000 students doing the experiment
ledger=data.frame(trial=seq(1,num.experiments),stats=NA)
#I am making a blank object that I'm going to fill in
for(j in 1:num.experiments){
  sample=rbinom(sample.size,possible.outcomes-1,p)
  ledger[j,]$stats=sum(sample)
}
```

How rare is it, under H_0 that we observe a value of 3 or smaller?

If we observed $\hat{p} = 2/27$ would that provide more or less evidence that H_0 was not true, or in other words that $\pi < 1/4$?

A key idea is that values of the statistic that are far from the hypotheiszed parameter result in small p-values and more evidence against the null hypothesis. Which begs the question, what is a p-value?

If we conclude that people chose tap water less often than would be expected by chance alone, what's the probability we made an erroneous conclusion?

Let's go through the alternate analysis on page 44.