Lsn 22 -AY23

Clark

Admin

A 2003 study explored whether the wording on a driver's license application might impact the likelihood of applicants agreeing to be an organ doner. 161 participants were recruited for the study and were randomly assigned to two groups, one that had to opt in to be an organ donor, the second group just had a neutral question "Do you wish to be an organ donor?" The question being explored is, does the prompt impact the probability of being an organ donor.

The sources of variation diagram is:

Our response variable is:

This makes a slightly different statistical model then perhaps we're used to:

The null hypothesis then is that there is no difference in the propbabilty of an individual becoming an organ donor. In symbols we have:

A common way to depict data of this sort is through a 2×2 contingency table:

```
donor.dat<-read.table("donor.txt",header=T)
donor.dat <-donor.dat %>% filter(Default !="opt-out")%>%droplevels()
my.table<-table(donor.dat$Choice,donor.dat$Default)
my.table
```

```
##
##         neutral opt-in
##   donor      44     23
##   not        12     32
```

Our parameter we are interested in is $\pi_1 - \pi_2$, which can naturally be estimated by:

Under H_0 we would expect $\hat{p}_1 - \hat{p}_2$ to be centered at zero (why?)

Our statistic here is $44/56 - 23/55 = 0.368$

To see how rare this is under H_0 we can employ our shuffling strategy:

```
M<-5000
results<-data.frame(trial=seq(1,M),stat=NA)
for(i in 1:M){
  donor.dat.mod <- donor.dat %>% mutate(shuff.cat=sample(Default))
  my.table<-table(donor.dat.mod$Choice,donor.dat.mod$shuff.cat)
  p.tabl<-prop.table(my.table,2)
  results[i,]$stat<-p.tabl[1,1]-p.tabl[1,2]
}
results %>% ggplot(aes(x=stat))+geom_histogram()+
  geom_vline(xintercept=0.367,lwd=2,color="red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


So, pretty rare...

Because simulations can be a bit of a pain, we can also use a theory based approach if we have at least 10 successes and 10 failures **in each group**. In this canse we can say that the CLT has kicked in and we can use:

A simple way to impliment in R is using `prop.test()` though we have to make sure we input the data in correctly

```
my.table<-table(donor.dat$Default,donor.dat$Choice)
#NOTE THIS IS DIFFERENT THAN OUR VISUALIZATION
#Look at ?prop.test
prop.test(my.table,correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  my.table
## X-squared = 15.665, df = 1, p-value = 7.56e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1985851 0.5364798
## sample estimates:
##    prop 1    prop 2
## 0.7857143 0.4181818
```

Here the CI comes from:

```
p1=.7857
p2=.4182
n1=56
n2=55
se.ci=sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
multiplier=qnorm(.975)
(p1-p2)-multiplier*se.ci
```

```
## [1] 0.1985504
```

```
(p1-p2)+multiplier*se.ci
```

```
## [1] 0.5364496
```

As we see here we also get a χ^2 statistic. The statistic can be found through calculating:

$$\sum_{Cells} \frac{(Obs - Exp)^2}{Exp}$$

Our Observed are the values in the table. Our expected is calculated from what we would have expected to get in each cell assuming $\pi_1 = \pi_2$. So, for instance, if $\pi_1 = \pi_2$ then we could find a common estimate of $\pi = \pi_1 = \pi_2$ through

```
pi=(44+23)/(44+23+12+32)
pi
```

```
## [1] 0.6036036
```

So, if H_0 is true, we would have expected, out of 54 people given the neutral wording, $54 * .60$, or 32.4 to have been a donor, and 21.6 not to have donated. With opt-in wording, we had 55 people, so we would have expected $55 * .60$ or 33 to have been a donor and 22 to not have donated. So our statistic is found through:

```
cell11=(44-32.4)^2/32.4
cell21=(12-21.6)^2/21.6
cell12=(23-33)^2/33
cell22=(32-22)^2/22
cell11+cell21+cell12+cell22
```

```
## [1] 15.99551
```

We compare this to a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, or in this case, 1 degree of freedom.

Yet another statistic (that we will see is quite useful in some cases) is the odds ratio. The odds ratio is formed by comparing the odds of success from group 1 to the odds of success from group 2. Odds are:

So from our dataset, we compute the odds of success having been given Treatment A (neutral wording) as:

```
p1=44/(44+12)
Odds1=p1/(1-p1)
Odds1
```

```
## [1] 3.666667
```

The odds of success having been given Treatment B (opt-in wording) is:

```
p2=23/(23+32)
Odds2=p2/(1-p2)
Odds2
```

```
## [1] 0.71875
```

Under H_0 (there is no difference in the proportion of successes between Treatments), what should the Ratio between `Odds1/Odds2` equal?

To see how rare OUR OR is, we can again simulate:

```
Our.OR<-Odds1/Odds2
M<-1000
results<-data.frame(trial=seq(1,M),stat=NA)
for(i in 1:M){
  donor.dat.mod <- donor.dat %>% mutate(shuff.cat=sample(Default))
  my.table<-table(donor.dat.mod$Choice,donor.dat.mod$shuff.cat)
  p.tabl<-prop.table(my.table,2)
  Odds.sim1<-p.tabl[1,1]/(1-p.tabl[1,1])
  Odds.sim2<- p.tabl[1,2]/(1-p.tabl[1,2])
  results[i,]$stat<-Odds.sim1/Odds.sim2
}
results %>% ggplot(aes(x=stat))+geom_histogram()+geom_vline(xintercept = Our.OR,lwd=2,color="red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

So again, strong evidence against the null. In order to use a theory based test involving Odds, it turns out that it often times is better to use log-Odds. Though harder to interpret, let's take a look at the distribution under H_0.

```
results <- results %>% mutate(log.stat=log(stat))
results %>% ggplot(aes(x=log.stat))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

As this statistic is generally symmetric, it turns out that log-Odds converges super quickly under the CLT to a normal distribution, which makes life really nice.

Next lesson we will continue to use log-Odds as a statistic. Our statistical model will become:

Here we see we are placing structure on log-Odds similarly to how we place structure on μ in a linear regression model. This model is called a logistic regression model.

This can be fit in R using what is known as a Generalized Linear Model. This class of models is extremely flexible

```
donor.dat <-donor.dat %>% mutate(bin.outcome=ifelse(Choice=="donor",1,0))
my.glm<-glm(bin.outcome~Default,data=donor.dat,family="binomial")
coef(my.glm)
```

```
##   (Intercept) Defaultopt-in
##      1.299283     -1.629525
```

So the log-Odds of someone with Neutral wording is 1.299, or the Odds are `exp(1.299)=3.665` and the log-Odds of someone with opt-in wording is 1.299-1.629, or the Odds are `exp(1.299-1.629)=.719`