Gov. 1010 Non-Response Simulation Notes

Mark Hill

11/7/2018

We've frequently talked about non-response in the class and this is often one of the first things noticed in a methodology report, but why (and when) does non-response actually matter?

This simulation will present the results from a survey of a made up population giving their favorability toward a certain policy. First we'll create the adult population.

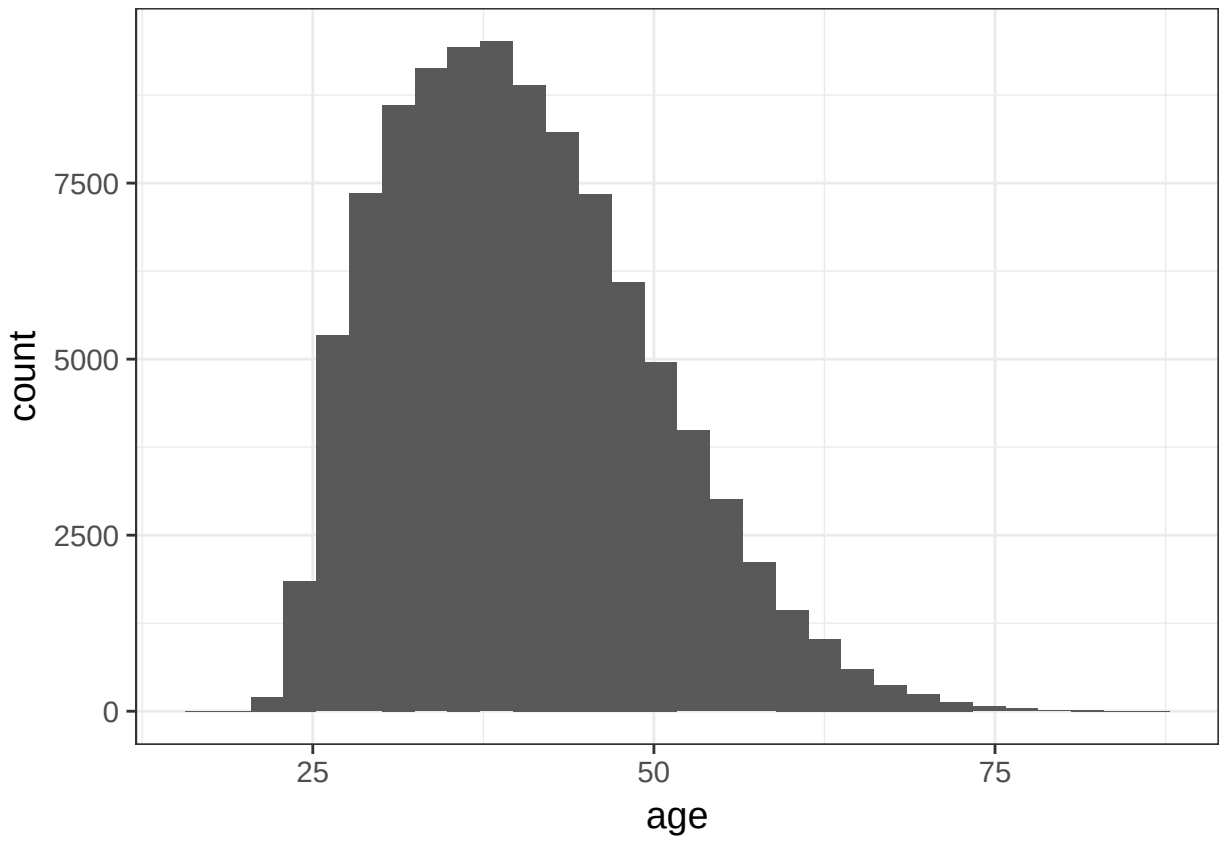
```
# draw 100,000 ages from skewed normal distribution
set.seed(02138)
age <- rsn(n=100000,xi=37, omega=12, alpha=7, tau=1)
# check min and max ages and remove people who aren't adults
min(age); max(age)
```

```
## [1] 16.77077
```

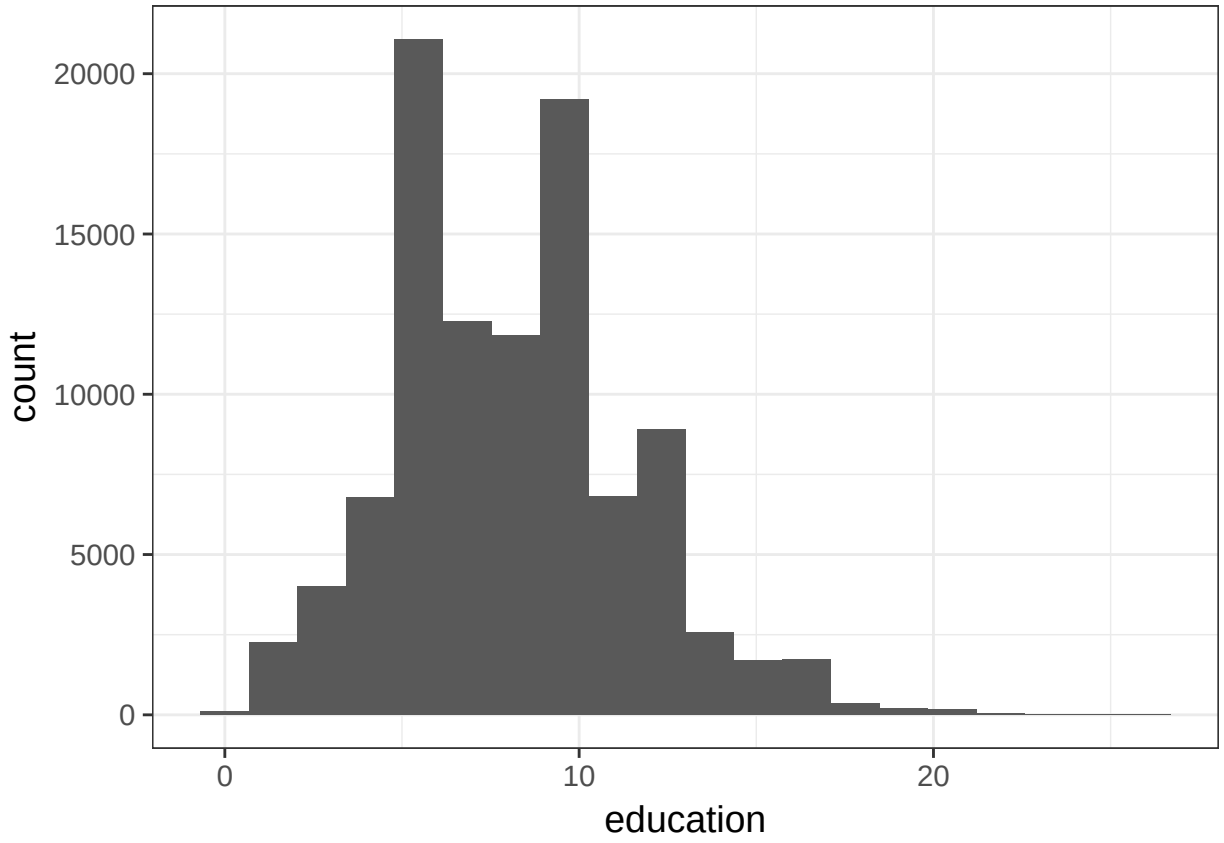
```
## [1] 87.7519
```

```
age[age < 18] <- 18
# view distribution
qplot(age) + theme_bw(base_size = 14)
```

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

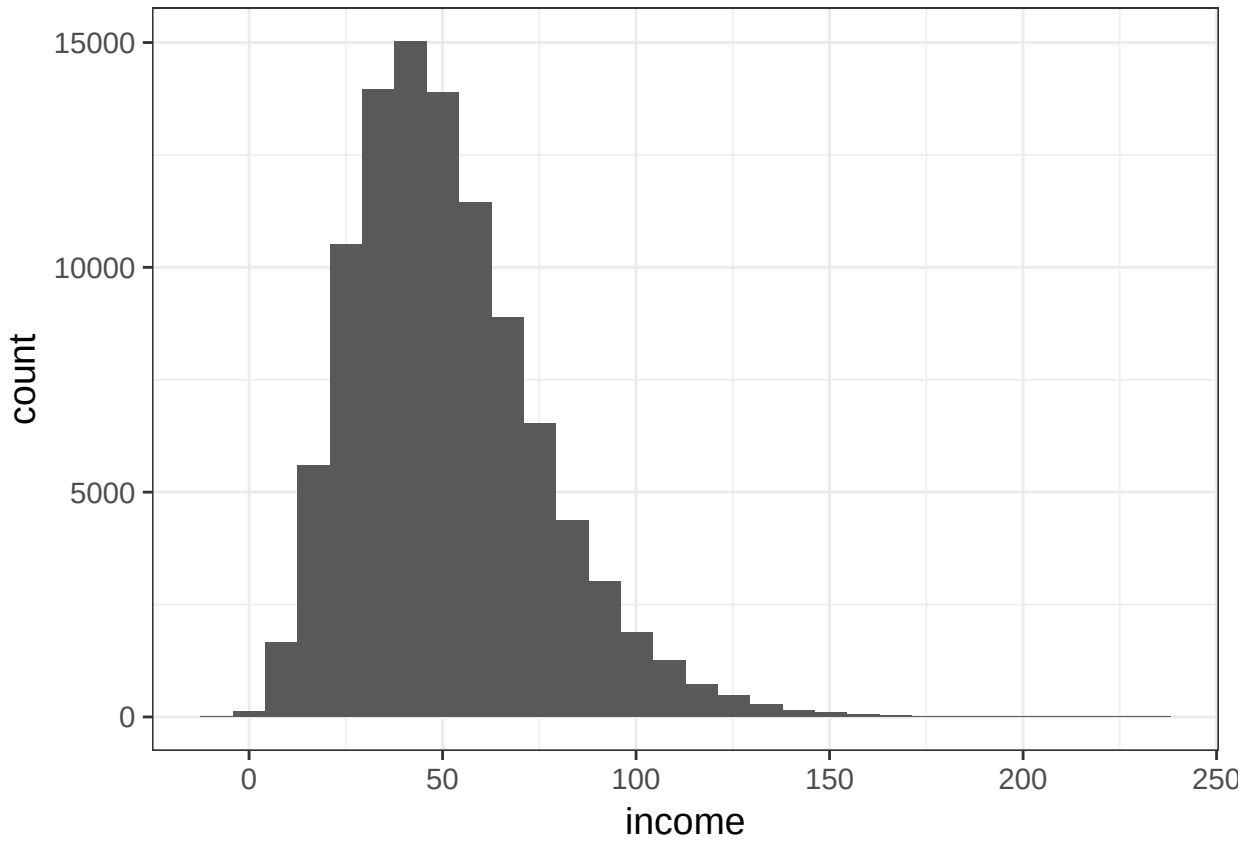


```r
# gender from bernoulli distribution probability .5
female <- rbinom(100000, 1, .5)

# years of education as a function of gender and different ages from Poisson distribution
education <- rpois(100000, lambda = female + 10*I(age > 22 & age < 35) + 6.5*I(age >=35 & age <= 65)
                   + 4*I(age > 65))

# view distribution
qplot(education, bins=20) + theme_bw(base_size = 14)
```




```r
# income from random skewed normal distribution as a function of gender, age, education, education^2
income <- rsn(n=100000,xi= (1.3*female + .33*I(age > 25 & age < 55) + 2.5*education + .2*education^2),

# view distribution
qplot(income) + theme_bw(base_size = 14)
```

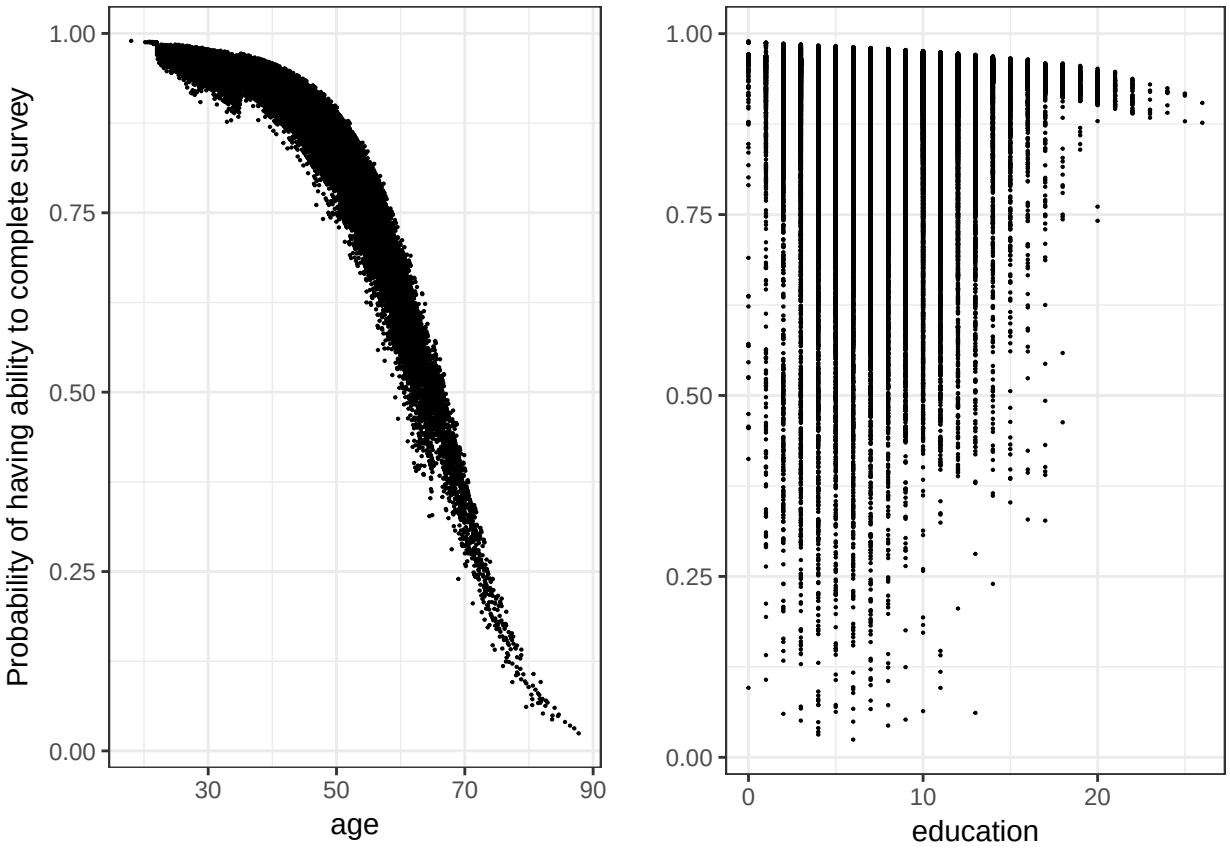
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# ability to answer survey as a function of age and education
p <- plogis(5-age/150-(age^2)/1000-education/15)
ability <- rbinom(100000, 1, p)
mean(ability)
```

```
## [1] 0.89901
```

```
p1 <- qplot(age,p, size=I(0.1)) + theme_bw() + ylab("Probability of having ability to complete survey")
p2 <- qplot(education, p, size=I(0.1)) + theme_bw() + ylab("")
grid.arrange(p1, p2, nrow = 1)
```



Now with our population's baseline demographics created we know many things are correlated. Education is related to age and gender, income is related to age, gender, and education, and one's ability to even be able to complete this survey is related to age and education. Thinking about non-response, why is the above plot important?

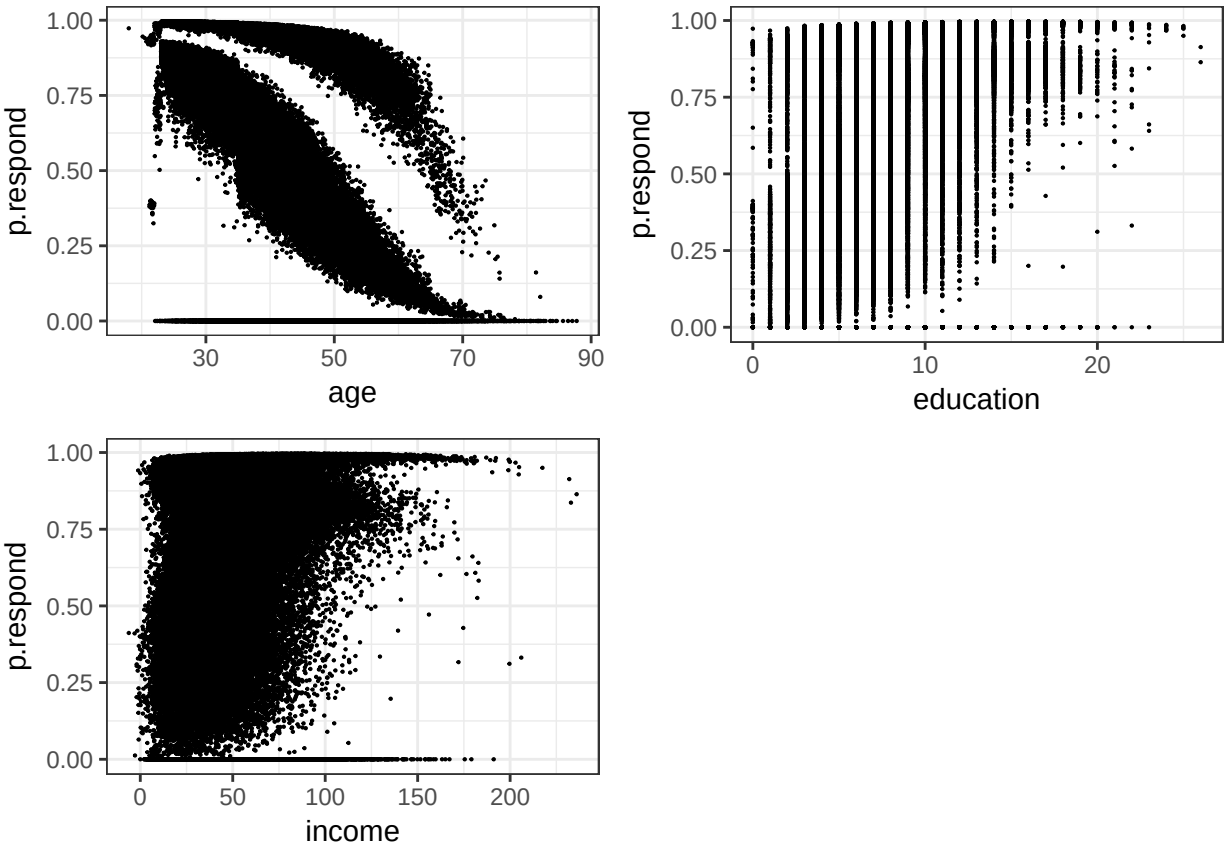
People's ability to complete the survey, and therefore non-response, is systematically related to one's age and education. If age and education play an important role in determining one's ideology and policy preferences, then this will almost certainly result in bias. 89.9% of the people in this population have the ability to complete the survey.

Next, we'll give each person a probability of responding. It will be a function of income, age, and ability to respond. If a person does not have the ability to respond, their probability of responding will be 0.

```
p.respond <- as.numeric(plogis(1.5-1*I(age>18 & age<23)-income/400-age/40+education/5-(age^2)/1250-(inc
p.respond[ability==0] <- 0

respond <- rbinom(100000, 1, p.respond)

mean(respond)
```

```
## [1] 0.68313
```

```
p1 <- qplot(age, p.respond, size=I(0.1)) + theme_bw()
p2 <- qplot(education, p.respond, size=I(0.1)) + theme_bw()
p3 <- qplot(income, p.respond, size=I(0.1)) + theme_bw()

grid.arrange(p1,p2,p3, ncol=2)
```

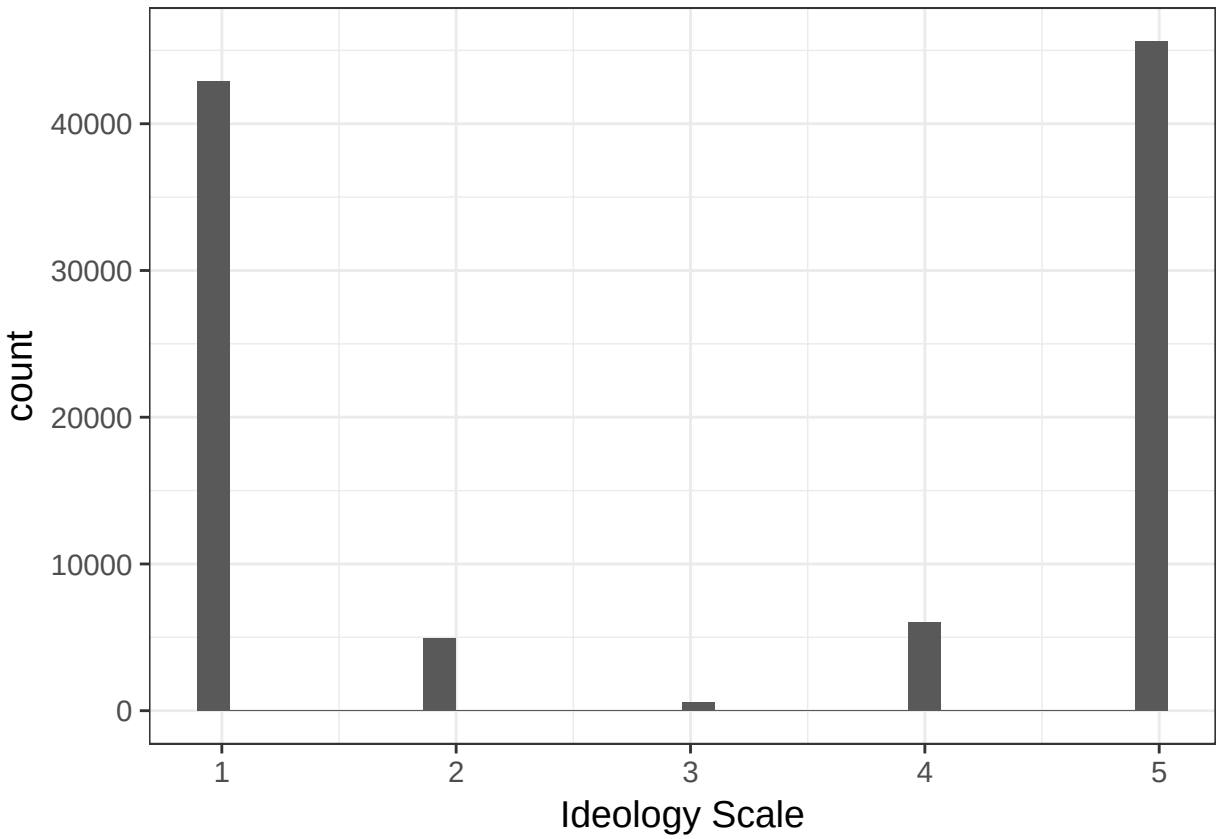


As we can see above, one's probabily of responding is correlated with variables that will likely determine ideology and supprt for policy. Arounf 70% of the population will respond if contacted.

Next the population's ideology is set up to mimic ordered logit data with values ranging from 1-5. Ideology is a function of age, gender, education, and income. The resulting distribution of ideology is below.

```
##
##   TRUE
## 500000
```

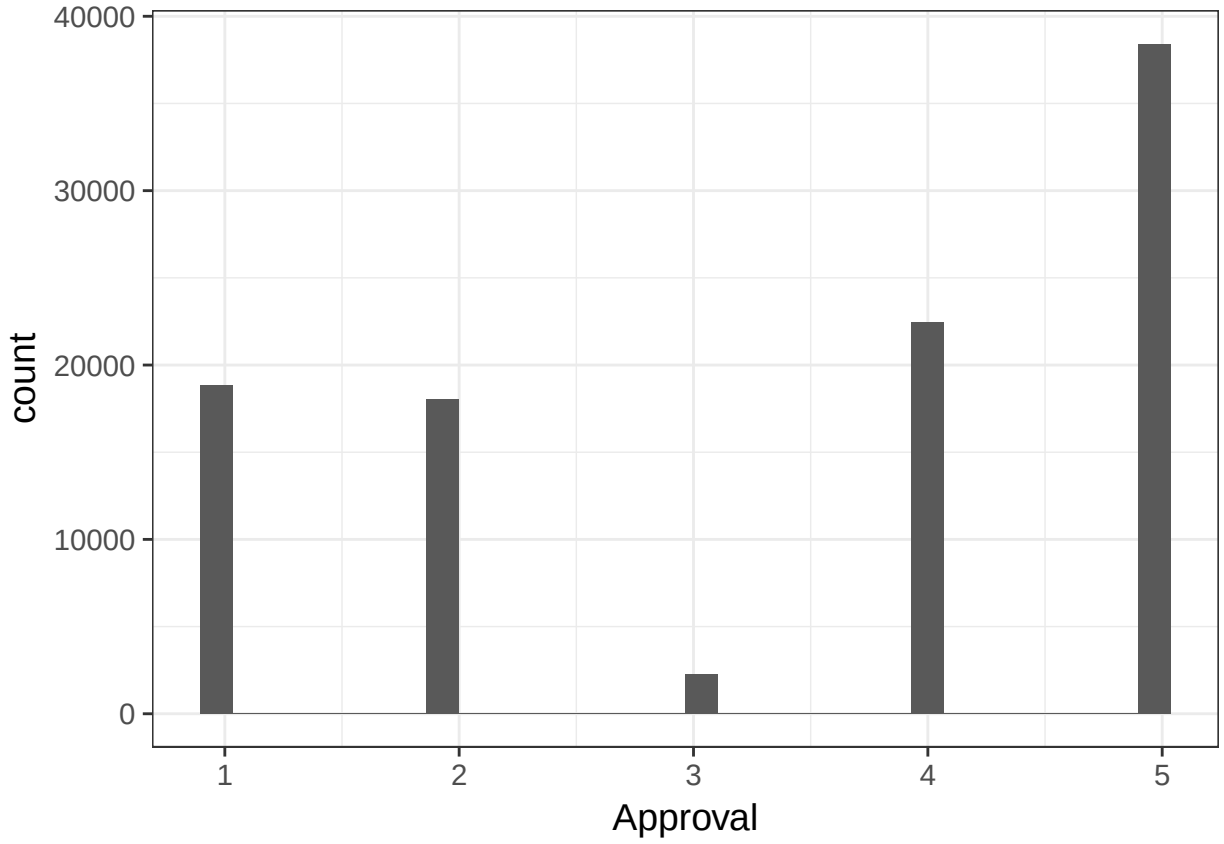
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This population is clearly divided by ideology. Now we will create support for policy as a function of ideology on a scale from 1-5 the same way the data above was created.

```
##
##   TRUE
## 500000
```

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




```
# create population
population <- data.frame(age, female, education, income, respond, support, p.respond)
truth <- mean(population$support>=4); truth
```

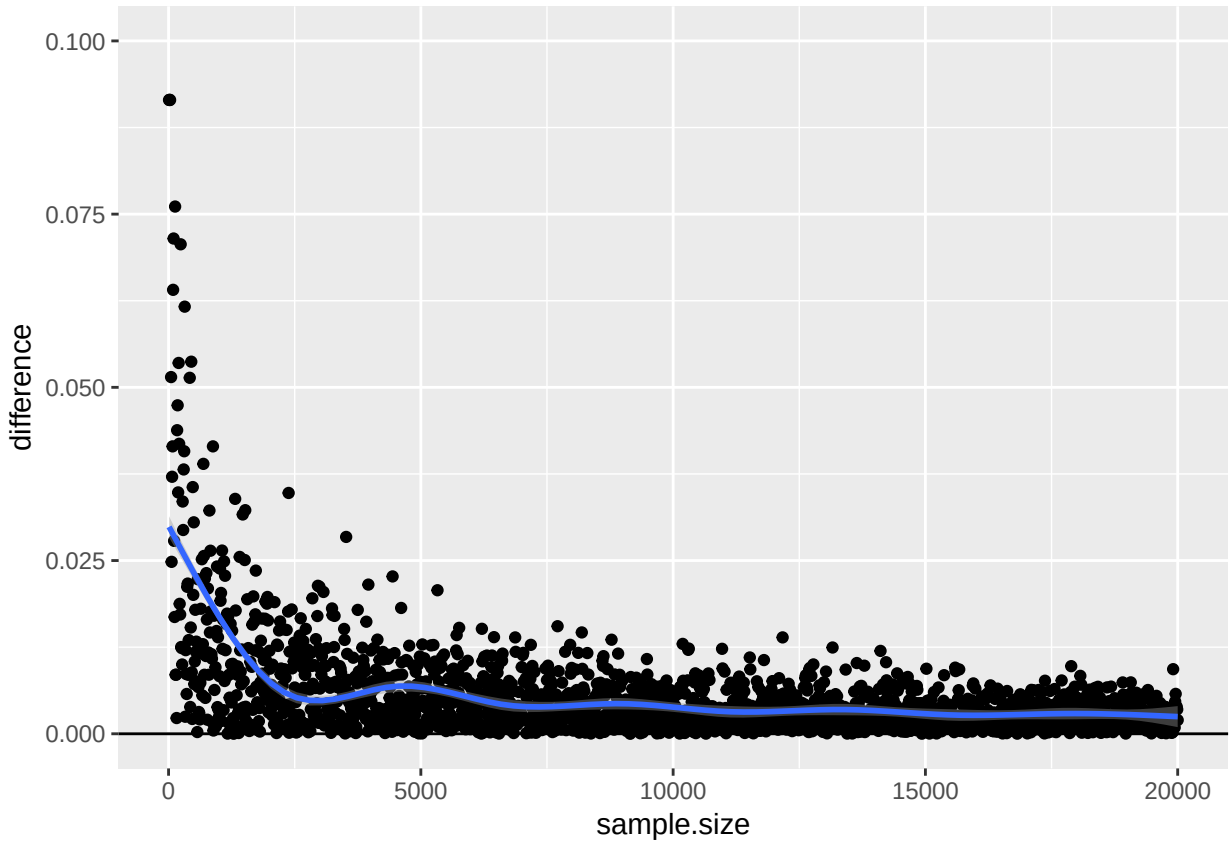
```
## [1] 0.60852
```

```
sample.size <- seq(10,20000,10)
difference <- rep(NA, length(sample.size))
i=1
for(size in sample.size){
  draw <- population[sample(1:nrow(population), size),]
  difference[i]  <- abs(truth - mean(draw$support>=4))
  i=i+1
}
qplot(sample.size,difference) + geom_hline(yintercept = 0) + geom_smooth() + ylim(0,.1)
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



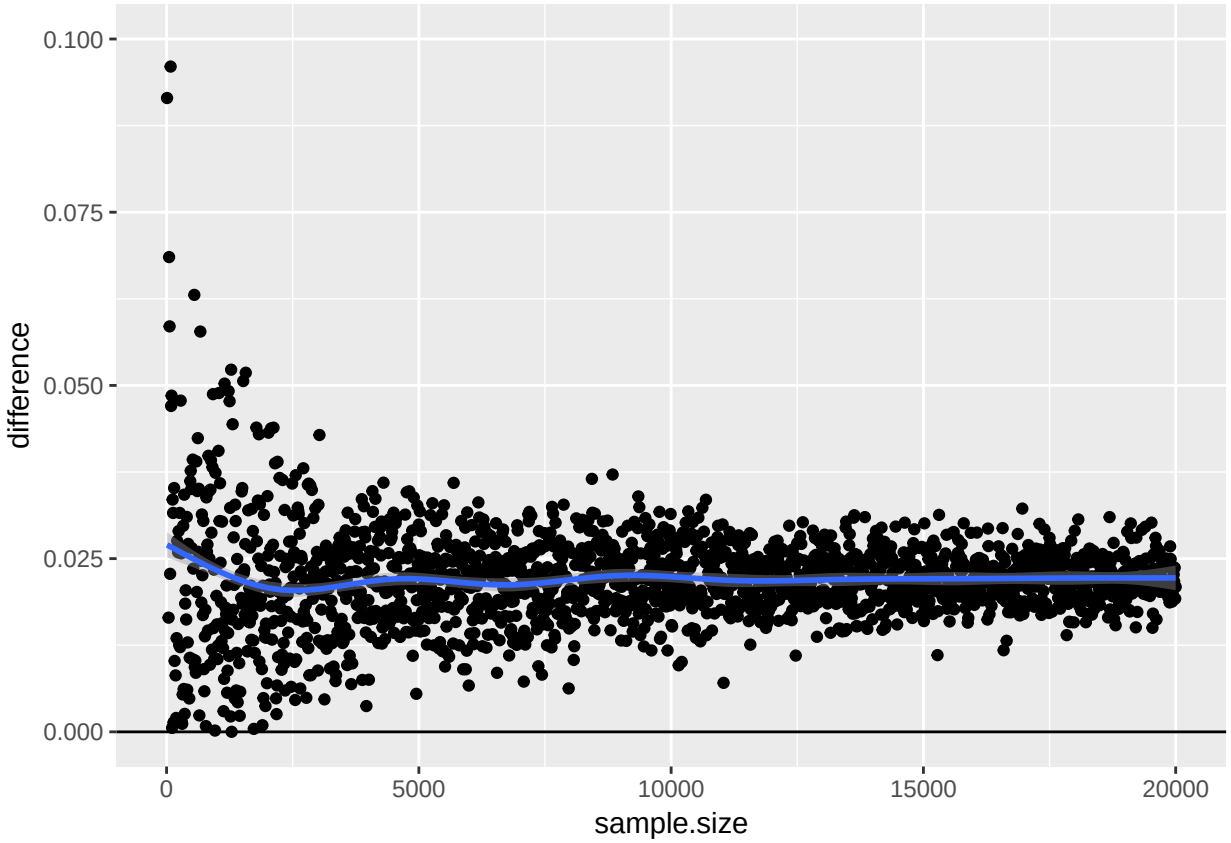
The above plot shows the absolute difference of the estimated policy approval for random samples up to N=20,000. As we would expect, as sample size increases the estimate approval converges to the true value.

```
sample.size <- seq(10,20000,10)
difference <- rep(NA, length(sample.size))
difference2 <- rep(NA, length(sample.size))
i=1
for(size in sample.size){
  draw <- population[sample(1:nrow(population), size),]
  new.pop <- population[-as.numeric(row.names(draw)),]
  draw <- draw[-which(draw$respond == 0),]
  while(nrow(draw)<size){
    x <- sample(1:nrow(new.pop),size-nrow(draw))
    new <- new.pop[x,]
    new.pop <- new.pop[-x,]
    new <- new[!new$respond == 0,]
    draw <- rbind(draw,new)
  }
  difference[i]  <- abs(truth - mean(draw$support>=4))
  difference2[i]  <- abs(truth - weighted.mean(draw$support>=4, 1/draw$p.respond))
  i=i+1
}

qplot(sample.size,difference) + geom_hline(yintercept = 0) + geom_smooth() + ylim(0,.1)
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

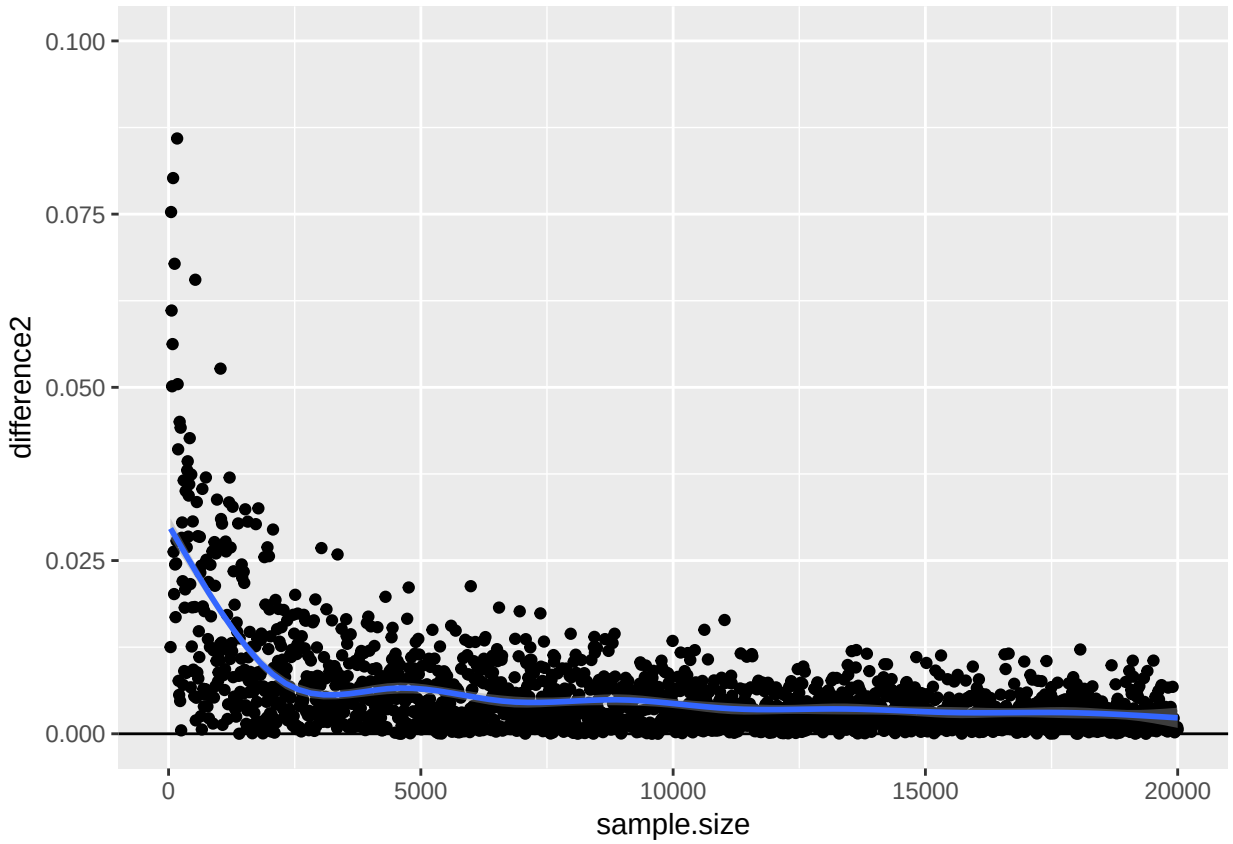
```
## Warning: Removed 3 rows containing missing values (geom_point).
```



The above plot shows the absolute difference of the estimated policy approval for samples up to N=20,000 where people are missing systematically according to their demographics. Now it appears that no matter the sample size, the estimated policy approval will be 2.5% away from the true value.

```
qplot(sample.size,difference2) + geom_hline(yintercept = 0) + geom_smooth() + ylim(0,.1)
```

```
## `geom_smooth()` using method = 'gam'
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
## Warning: Removed 3 rows containing missing values (geom_point).
```


0.100
0.075
0.050
0.025
0.000
difference2
0
5000
10000
15000
20000
sample.size