Project

Will Curkan, Alina Ila, Rachel Uc, Jocelyn Serrot, Henry Zerep

2022-11-18

Introduction

Abortion is a sensitive topic in the United States with the legendary court ruling of *Roe vs. Wade* being overturned in June 2022. While sensitive, it is still an interesting topic because of the different viewpoints: are we killing a baby, or are we killing a fetus? The biology is irrelevant, though, in statistical analyses. Because of the amount of data, it is a good topic of analyses and we shall compare statistics among the categorized demographics of recorded abortions, to see if there are any disparities*****, and double-check claims.

We are using the "Pregnancies, Births and Abortions in the United States: National and State Trends by Age" dataset called `NationalAndStatePregnancy_PublicUse` from the Guttmacher Institute website which sources its data from numerous different organizations like the World Health Origanization (WHO) and UNICEF.

We will use the columns: state, year, abortionratelt15, abortionrate1517, abortionrate1819, abortionrate2024, abortionrate2529, abortionrate3034, abortionrate3539, abortionrate40plus, which is the rate of abortion per 1000 women in the age range. For example `abortionratelt15` is the rate of abortions of the given U.S. state and year for girls less than 15 years old, and `abortionrate2024` is the rate of abortions for ladies of ages 20-24, per 1000 people.

Some questions to ask: - Can we find an estimation region for the true mean abortions for all age groups of women. - What is an estimation region for the true mean of abortions of women and girls in Alabama.

- Is there a statistically significant difference in mean abortion rates among age groups of women and girls.
- Is there a statistically significant difference in mean abortion rates among states.
- Is there a statistically significant difference in mean abortion rates among the years.

NEED TO CHECK IF DIFFERENCE WITHOUT FILLING VALUES

First, let's look at the most recent mean abortion rate per thousand by age group for all states. This is for the year 2017.

```
##    year abortionratelt15 abortionrate1517 abortionrate1819 abortionrate2024
## 45 2017        0.6923077         2.428846         7.990385         19.73269
##    abortionrate2529 abortionrate3034 abortionrate3539 abortionrate40plus
## 45         18.08654         12.07308         7.180769           2.648077
```

The mean abortion rate of all United States looks good. On average, the abortion rate for girls under 15 is .69, and the rate for women 40 and older is 2.65. Of course, this is given samples from each state in 2017 and is not necessarily representative of all abortions that occurred. We shall investigate further with statistical inference tests.

The Guttmacher institute claims that the declining abortion rates are reversing as of 2017 saying "An increase in abortion numbers is a positive development if it means people are getting the health care they want and need" [2]. But, due to the possible uncertainty in the samples, we want to see if there is statistically significant evidence that the abortion rate was dropping in the first place.

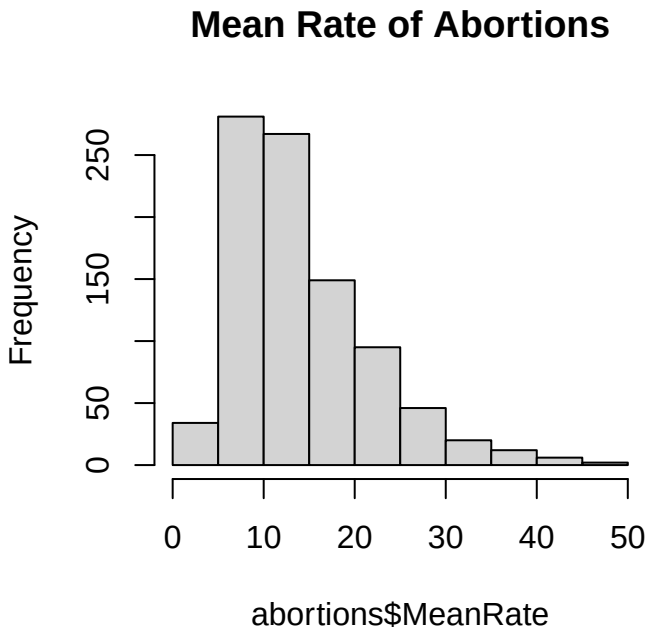
Question: - Is there a statistically significant difference in mean abortion rates from the years 1988 - 2017.

We will use the ANOVA permutation test to see if there is a difference in means among the years.

$H_0 : \mu_{1988} = ... = \mu_{2017}$

$H_A : \mu_{1988} \neq ... \neq \mu_{2017}$

```
hist(abortions$MeanRate, main = 'Mean Rate of Abortions')
```




```
## [1] 1e-04
```

1e-04

P-value is very small, so we confirm that there is a statistically significant difference in mean abortion rate among the years. We will reject the null hypothesis at a 1% level of significance that the mean abortion rate among the years is the same.

We can inspect which years are different from the others with a TukeyHSD test.

```
#TukeyHSD(aov(abortions$MeanRate ~ year1))
```

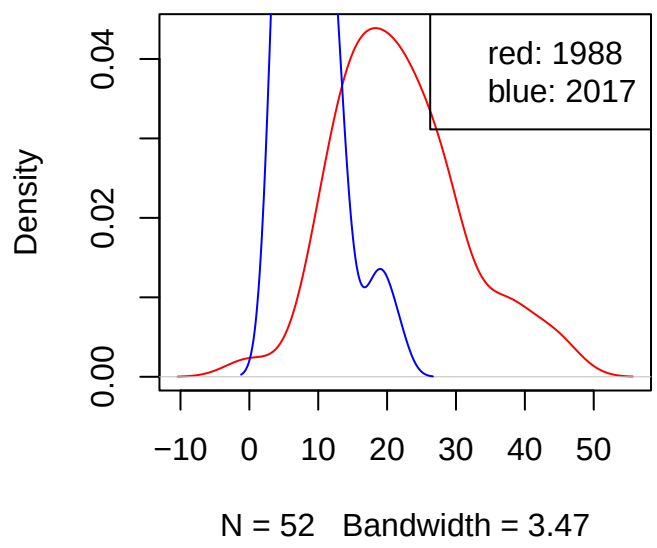
Is the variance equal

Question: Is there a difference between the mean abortion rate for 1988 and 2017 in the U.S.

$H_0 : \mu_{1988} = \mu_{2017}$ $H_A : \mu_{1988} \neq \mu_{2017}$

We subsetted the state for the specific years and plot a histogram to see. . . .

```
plot(density(year1988$MeanRate), col = 'red',
     main = 'Density of the Mean Abortion Rate: 1988 and 2017')
lines(density(year2017$MeanRate), col = 'blue')
legend('topright', c('red: 1988', 'blue: 2017'))
```



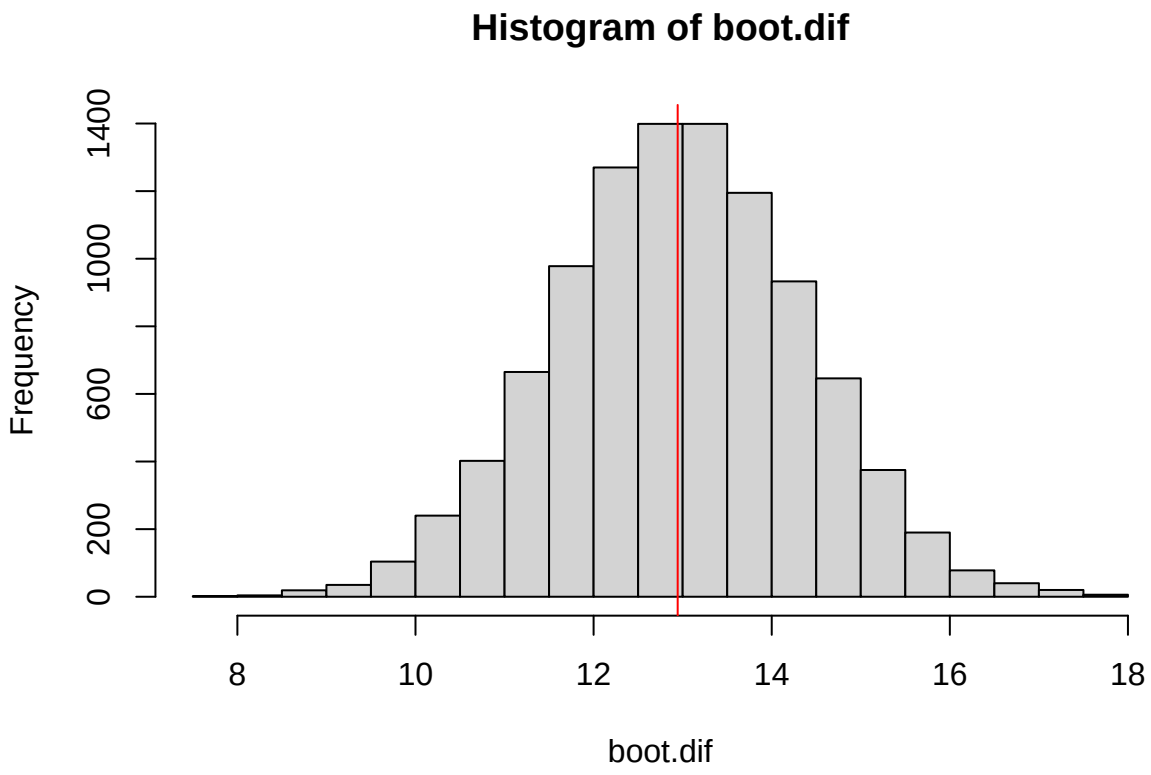
From plotting the density of the histograms, we see that in the given sample the means are close because the variances overlap. Also the data is not normal so we need to use non-normal testing methods.

```
# qqnorm(year1988$MeanRate)
# qqline(year1988$MeanRate)
# qqnorm(year2017$MeanRate)
# qqline(year2017$MeanRate)



#BOOTSTRAP 2-sample test for two population means
N <- 10^4

xbar1988 <- mean(unlist(year1988$MeanRate))
xbar2017 <- mean(unlist(year2017$MeanRate))

n1 <- length(year1988$MeanRate)
```

```
n2<- length(year2017$MeanRate)

mean.dif <- xbar1988-xbar2017

boot.dif <- numeric(N)

for (i in 1:N){
x <- sample(year1988$MeanRate, n1, replace = TRUE)
y <- sample(year2017$MeanRate, n2, replace = TRUE)
boot.dif[i] <- mean(x) - mean(y)
}

quantile(boot.dif, c(.05,.95))
```

```
##       5%      95%
## 10.65645 15.24945
```

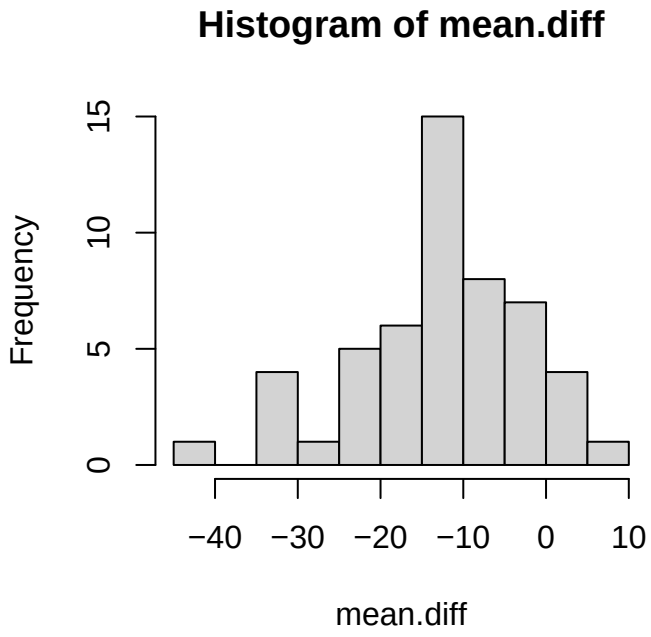
```
hist(boot.dif)
abline(v = mean.dif, col="red")
```




```
# (sum(boot.dif >= observed) + 1) / (N + 1)
```

From the result we see there is a significant difference, but we know in practice that the permutation test is shown to be more powerful. We will use a permutation test with the assumption that there is no difference in mean rates by pooling the mean rates together, then drawing samples without replacement from the pooling.

```
## [1] 52
```




```
##         5%        95%
## -32.459375   0.734375
```

Using a permutation test of the mean rates, we see that 0 is contained in the interval that contains the true difference in mean rates. We concluse that there is not enough evidence to suggest the mean rate is different in 2017 from 1988

Quick 2-sample bootstrap-t

```
n1 <- length(year1988$MeanRate)
n2 <- length(year2017$MeanRate)
Tstar <- numeric(N)
SE <- sqrt(var(year2017$MeanRate)/n2 + var(year1988$MeanRate)/n1)

for (i in 1:N)
{
  bootx <- sample(year2017$MeanRate, n2, replace = TRUE)
  booty <- sample(year1988$MeanRate, n1, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - obs.diff) /
              sqrt(var(bootx)/n2 + var(booty)/n1)
}

obs.diff - quantile(Tstar, c(.99, .01)) * SE
```

```
##      99%       1%
## 34.11553 45.73919
```

```
t.test(year2017$MeanRate, year1988$MeanRate)
```

```
##
##  Welch Two Sample t-test
##
## data:  year2017$MeanRate and year1988$MeanRate
## t = -9.2091, df = 71.178, p-value = 9.445e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.74649 -10.14149
## sample estimates:
## mean of x mean of y
##  8.854087 21.798077
```

New YORK!

```
abortions %>%
  select(state,MeanRate) %>%
  group_by(state) %>%
  mutate('state_mean' = mean(MeanRate)) %>%
  arrange(desc(state_mean))
```

```
## # A tibble: 912 x 3
## # Groups:   state [52]
##    state MeanRate state_mean
##    <chr>    <dbl>      <dbl>
##  1 NY        40.1       31.4
##  2 NY        43.8       31.4
##  3 NY        38.9       31.4
##  4 NY        37.6       31.4
##  5 NY        35.2       31.4
##  6 NY        34.2       31.4
##  7 NY        33.2       31.4
##  8 NY        33.7       31.4
##  9 NY        32.3       31.4
## 10 NY        30.9       31.4
## # ... with 902 more rows
```

NY has highest rate of abortion with mean = 31.43. We want to perform the interval test to confirm the true mean interval of NY

```
NY_rate <- filter(abortions, state == "NY") %>% select(MeanRate)
NY_rate <- as.vector(unlist(NY_rate))
NY_rate
```

```
##  [1] 40.1250 43.7750 38.8875 37.6250 35.2375 34.1500 33.1500 33.7125 32.3000
## [10] 30.9125 29.7000 27.2375 24.8125 24.6875 23.7250 22.6875 21.6500
```

```r
summary(NY_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.65   24.81   32.30   31.43   35.24   43.77
```

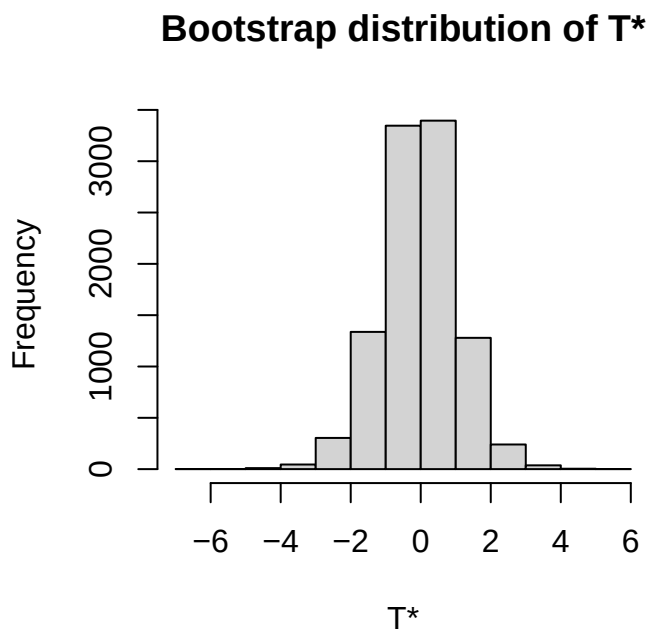
```r
xbar <- 31.43
N <- 10^4
n <- 17 #number of years

Tstar <- numeric(N)

for (i in 1:N)
{
x <-sample(NY_rate, size = n, replace = T)
Tstar[i] <- (mean(x)-xbar)/(sd(x)/sqrt(n))
}
quantile(Tstar, c(0.05, 0.95)) # the first value is negative and the second positive, so we switch
```

```
##        5%       95%
## -1.778321  1.686379
```

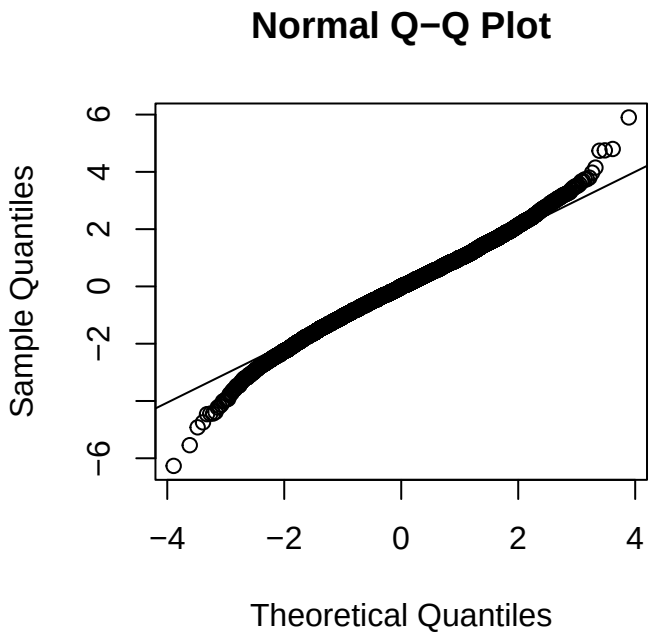
```r
xbar - quantile(Tstar, c(.95, .05))*sd(NY_rate)/sqrt(n)
```

```
##      95%       5%
## 28.73561 34.27129
```

```r
hist(Tstar, xlab = "T*", main = "Bootstrap distribution of T*")
```



```
qqnorm(Tstar)
qqline(Tstar)
```



From the bootstrap distribution, we can see the 90% CI for the true mean is (28.73, 34.305)

References

[2] https://www.guttmacher.org/article/2022/06/long-term-decline-us-abortions-reverses-showing-rising-need-abortion-supreme-court