MATH 392 Problem Set 2

EJ Arce

2 February 2018

3.4

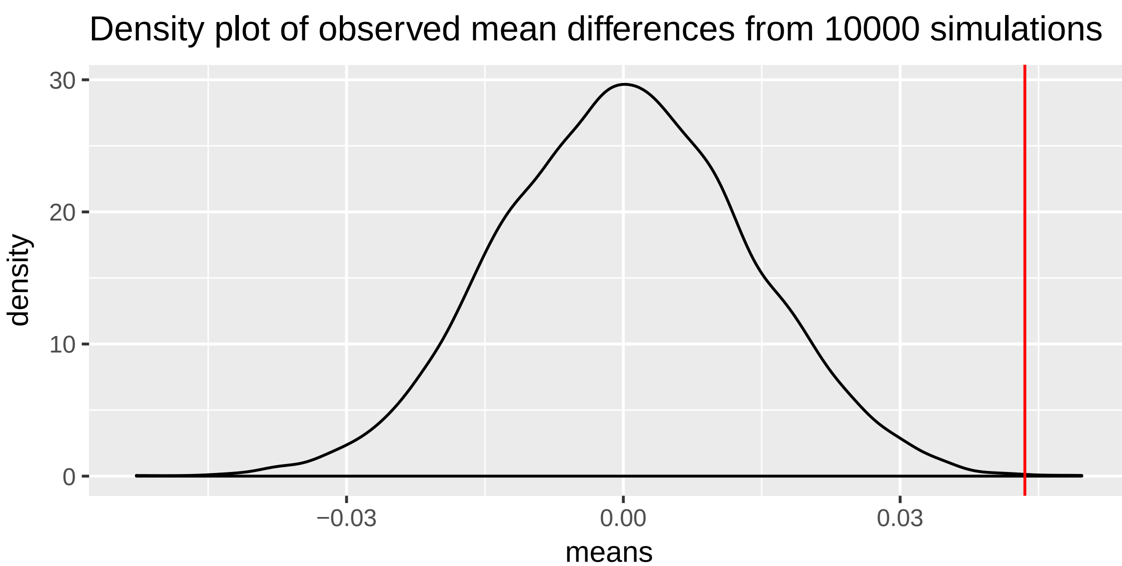
a

```
FlightDelays <- FlightDelays %>%
  mutate(Delayed20 = ifelse(Delay > 20, 1, 0))

# Calculating observed difference in the two groups' mean proportions
xobs <- mean(FlightDelays$Delayed20[FlightDelays$Carrier == "UA"] -
               mean(FlightDelays$Delayed20[FlightDelays$Carrier=="AA"]))

# Running simulation for hypothesis testing
nsim <- 10000
means <- rep(NA, nsim)
for(i in 1:nsim){
  perm <- sample(FlightDelays$Carrier, replace=F)
  means[i] <- mean(FlightDelays$Delayed20[perm=="UA"] -
                     mean(FlightDelays$Delayed20[perm=="AA"]))
}
simdf <- data.frame(means)

# Plotting the simulated null distribution
ggplot(simdf, aes(x = means)) +
  geom_density() +
  geom_vline(xintercept = xobs, col = "red") +
  ggtitle("Density plot of observed mean differences from 10000 simulations")
```



The red vertical line indicates the difference in proportions observed in the actual dataset. The p-value for a two-tailed test is calculated below.

```r
(sum(means > xobs) + 1)/(length(means) + 1) * 2
```

```
## [1] 0.0009999
```

b

We are testing to see if the variance in flight delays for United Airlines is greater than the variance for American Airlines. Thus, we are conducting a one-tailed significance test. Speicically,

$$H_0 : \rho_{UA} \leq \rho_{AA}$$

$$H_A : \rho_{UA} > \rho_{AA}$$

```r
# Variance in flight delay lengths for each carrier
varUA <- var(FlightDelays$Delay[FlightDelays$Carrier == "UA"])
varAA <- var(FlightDelays$Delay[FlightDelays$Carrier == "AA"])
varUA
```

```
## [1] 2038
```

```r
varAA
```

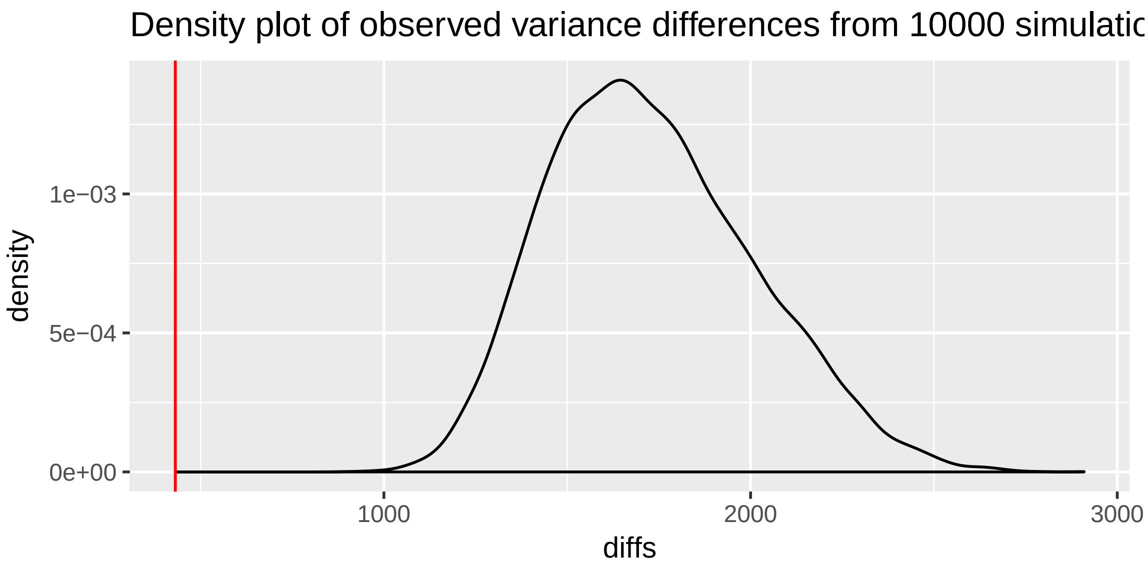
```
## [1] 1606
```

varUA and varAA indicate the variances of United Airlines' and American Airlines' delay times, respectively. A simulation just like the last problem will be used to test if the difference in these variances is statistically significant.

```r
obs.diff <- varUA-varAA

nsim <- 10000
diffs <- rep(NA, nsim)
for(i in 1:nsim){
  perm <- sample(FlightDelays$Carrier, replace=F)
  diffs[i] <- var(FlightDelays$Delay[perm=="UA"] -
                    var(FlightDelays$Delay[perm=="AA"]))
}
simdf <- data.frame(diffs)

# Plotting the simulated null distribution
ggplot(simdf, aes(x = diffs)) +
  geom_density() +
  geom_vline(xintercept = obs.diff, col = "red") +
  ggtitle("Density plot of observed variance differences from 10000 simulations")
```




```
(sum(diffs < obs.diff) + 1)/(length(diffs) + 1)
```

```
## [1] 9.999e-05
```

3.16

a

```
table(GSS2002$Gender,GSS2002$Pres00)
```

```
##
##          Bush Didnt vote Gore Nader Other
##   Female  459          5  492    26     3
##   Male    426          5  289    31    13
```

b

```
chisq.test(GSS2002$Gender,GSS2002$Pres00, simulate.p.value = T)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  GSS2002$Gender and GSS2002$Pres00
## X-squared = 33, df = NA, p-value = 5e-04
```

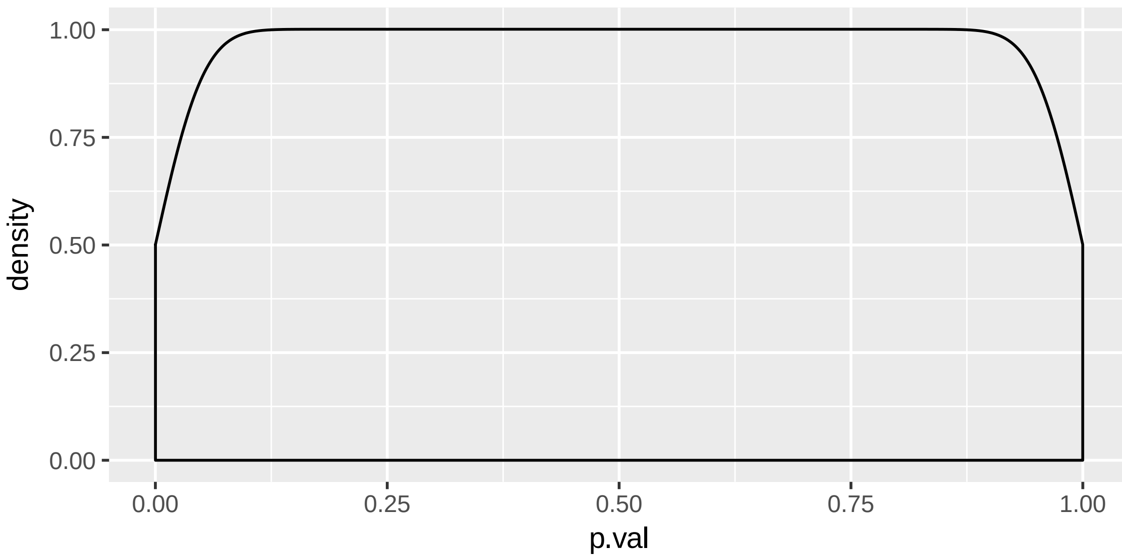
c

3.22

3.31

Empirical Solution

```
nsim <- 10000
t.obs <- rnorm(nsim,0,1)
ts <- data.frame(t.obs)
ts <- ts %>%
  arrange(t.obs) %>%
  mutate(t.obs = abs(t.obs))
p.val <- rep(NA,nsim)
for(i in 1:nsim){
  p.val[i] <- (sum(ts$t.obs>ts$t.obs[i])+1)/(length(ts$t.obs)+1)
}
ts <- cbind(ts, p.val)
ggplot(ts, aes(x=p.val)) +
  geom_density()
```



As we'd expect, the simulated density plot follows a uniform distribution.

Analytical Solution

Consider a test statistic t = $T(x_1, ..., x_n)$. Its corresponding p-value is calculated by solving

$$p = Pr(T(X) \geq t|H_0).$$

This makes p a random variable as well, since its calculated probability depends on the random variable T(X). Thus the p-value follows some probability distribution P(T). Since the p-values are drawn from the distribution of T(X), then the p-values have a one-to-one correspondence to each observed test statistic t. Thus,

$$Pr(P(T) \geq p) = Pr(T(X) \geq t) = p.$$

This shows that $P(T)$ follows a uniform distribution, where $Pr(p) = 1/n$.