MATH 392 Problem Set 2 (Corrected)

EJ Arce

9 February 2018

3.4

a

We are testing to see if the distributions of the proportion of flights delayed by more than 20 minutes, θ, differs by airline:

$$H_0 : \theta_{UA} - \theta_{AA} = 0$$

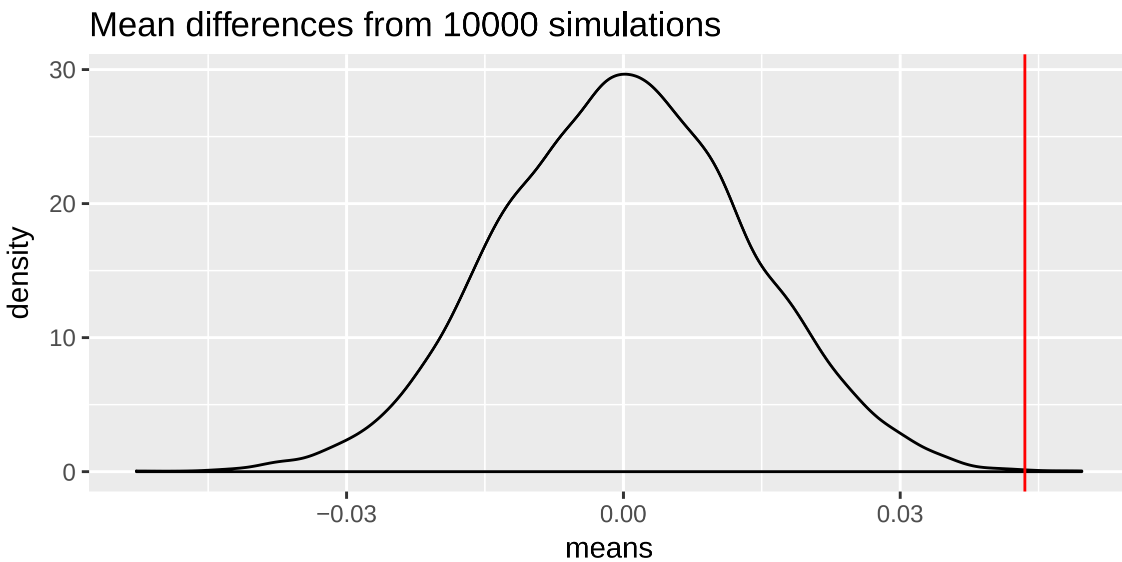
$$H_A : \theta_{UA} - \theta_{AA} \neq 0$$

```
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("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.

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

```
## [1] 0.00139986
```

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,

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

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

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

```
## [1] 2037.525
```

```
varAA
```

```
## [1] 1606.457
```

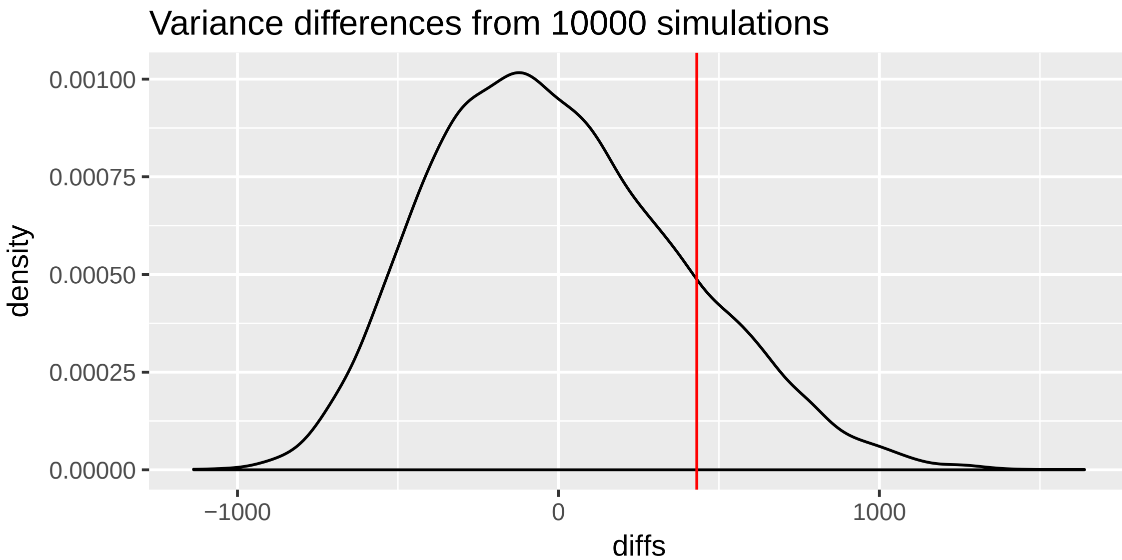
varUA and varAA indicate the variances of United Airlines' and American Airlines' delay times, respectively. A simulation just like the last problem will be used to test if the difference in these variances is statistically significant.

```
# Calculate test statistic
obs.diff <- varUA-varAA

# Run simulation
nsim <- 10000
diffs <- rep(NA, nsim)
```

```r
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("Variance differences from 10000 simulations")
```




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

```
## [1] 0.1450855
```

The density plot and calculated p-value show that the observed variance for United Airlines delays is not significantly greater than observed variances for American Airlines.

3.16

a

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

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

b

```r
gender <- GSS2002$Gender
pres <- GSS2002$Pres00
chisq.test(gender,pres)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gender and pres
## X-squared = 33.29, df = 4, p-value = 1.042e-06
```

c

```r
# Remove NAs
GSS2002 <- GSS2002 %>%
  filter(!is.na(Gender),
         !is.na(Pres00))
# Chi-squared test using permutations
x2.obs <- chisq.test(gender,pres)$statistic
nsim <- 10000
x2.stats <- rep(NA,nsim)
for(i in 1:nsim){
  perm <- xtabs(~sample(Gender, replace=F) + Pres00, data = GSS2002)
  x2.stats[i] <- chisq.test(perm)$statistic
}
(sum(x2.stats > x2.obs) + 1)/(nsim+1)
```

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

None of the simulated χ^2 values were greater than our observed χ^2_{obs} value of 33.29, resulting in our very low p-value.

3.22

```r
q <- c(.2,.4,.6,.8)
d <- data.frame("quantile" = q)
obs.stats <- c(12.57,16.87,20.73,24.66)
d<-cbind(d,obs.stats)
exp.stats <- rep(NA,4)
for(i in 1:4){
  exp.stats[i] <- qnorm(q[i],22,7)
}
d <- cbind(d,exp.stats)
d
```

```
##   quantile obs.stats exp.stats
## 1      0.2     12.57  16.10865
## 2      0.4     16.87  20.22657
## 3      0.6     20.73  23.77343
## 4      0.8     24.66  27.89135
```

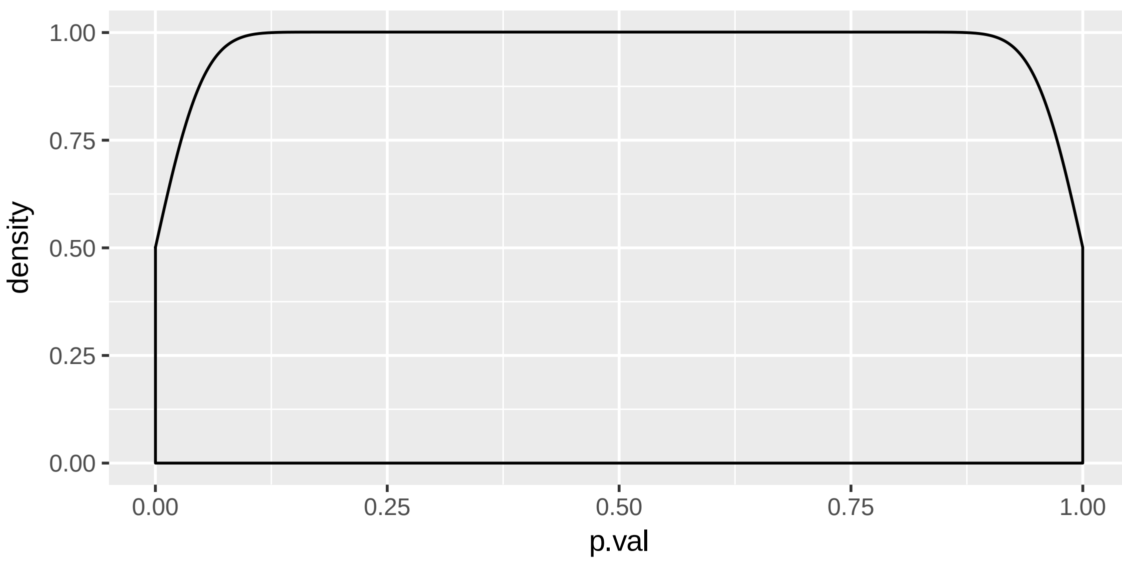
```r
# Calculate observed chi-squared value
((3.54^2)/16.11) + ((3.36^2)/20.23) + ((3.04^2)/23.77) + ((3.23^2)/27.89)
```

```
## [1] 2.098805
```

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

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

This makes p a random variable as well, since its calculated probability depends on the random variable T(X). Thus the p-value p follows some probability distribution $P = F_T(T)$. 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,

$$F_P(p) = P(P \leq p) = Pr(T(X) \leq t) = Pr(F_T(T) \leq p).$$

$$F_P(p) = P(F_T^{-1}F_T(T) \leq F_T^{-1}(p))$$

$$F_P(p) = P(T \leq F_T^{-1}(p))$$

$$F_P(p) = F_T(F_T^{-1}(p)) = p$$

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

3.32

Let Z denote the standard normal random variable. Then Z ~ N(0,1). Suppose $X = Z^2$. Show that X ~ $\chi^2_{df=1}$.

The pdf of Z is already known to be

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

Using the cdf method,

$$F_X(x) = P(X \leq x) = P(Z^2 \leq x) = P(-\sqrt{x} \leq Z \leq \sqrt{x}) = F_Z(\sqrt{x}) - F_Z(-\sqrt{x})$$

$$P(-\sqrt{x} \leq Z \leq \sqrt{x}) = F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$$

$$F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) = 2\int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$$

$$F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) = 2\frac{1}{2}erf(\frac{\sqrt{x}}{\sqrt{2}})$$

$$F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) = erf(\frac{\sqrt{x}}{\sqrt{2}})$$

Since the cdf we are differentiating over is $P(-\sqrt{x} \leq Z \leq \sqrt{x})$, we can rexpress this cdf as $2F_Z(\sqrt{x}) - 1$, so when taking the partial derivative, the coefficient 2 remains in front of the pdf $f_Z(\sqrt{x})$. Thus,

$$f_X(x) = \frac{\partial}{\partial x}(2F_Z(\sqrt{x}) - 1) = 2f_Z(\sqrt{x})\frac{1}{2}x^{-\frac{1}{2}}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x}{2}}x^{-\frac{1}{2}}$$

$$f_X(x) = \frac{x^{-\frac{1}{2}}e^{-\frac{x}{2}}}{\sqrt{2\pi}}$$

Notice that

$$\Gamma(1/2) = \int_0^\infty t^{-1/2}e^{-t}dt = \sqrt{\pi},$$

so we get

$$f_X(x) = \frac{x^{-\frac{1}{2}} e^{-\frac{x}{2}}}{\sqrt{2}\Gamma(1/2)}$$

Thus,

$f_X(x) \sim \chi^2_{df=1}$