non_parametric2

Joshua Burkhart

November 28, 2015

Class 17: More nonparametric methods for two sample problems

Overview

Today, you'll learn more about non-parametric analyses for two sample problems. You will report answers to the your turn at the end in a knit PDF emailed to upload.Class_1.ae5vul0n5o@u.box.com by end of day Monday November 30- this will be part deux of homework 4.

Your turn

Take the data you did the WMW test for (either MASS::birthwt or your final replication project dataset), and do some kind of two-sample permutation test. The possibilities are endless! You could do a simple difference in means, medians, calculate a permuted t-statistic…This is homework 4, part II. Also, feel free to watch the rest of the video- we'll return to our sneetches when we discuss paired sample tests next week!

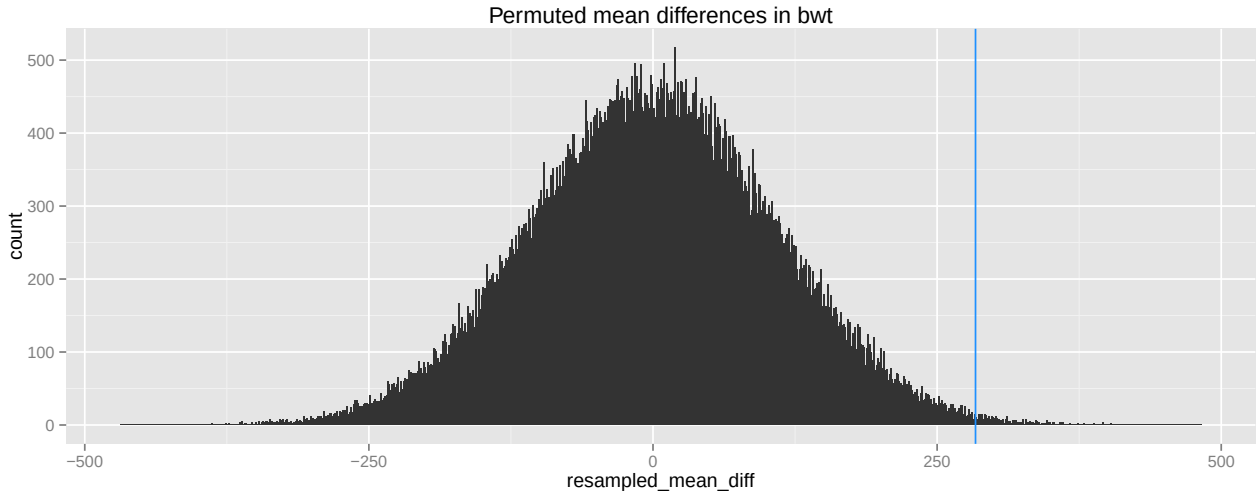
We'll do a simple difference in means.

```
smoke0 <- birthwt %>% filter(smoke == 0) %>% select(bwt)
smoke1 <- birthwt %>% filter(smoke == 1) %>% select(bwt)

smoke1_na <- t(t(append(smoke1[,1],rep(NA,41))))
bwts <- data.frame(smoke0,smoke1_na)

m_0 <- mean(smoke0[,1])
m_1 <- mean(smoke1_na[,1], na.rm = TRUE)

obs_mean_diff <- with(bwts, m_0 - m_1)
obs_mean_diff
```

```
[1] 283.7767
```

```
set.seed(0)
bwt_ary <- birthwt %>% select(bwt)
B <- 10^5-1  #set number of times to repeat this process
resampled_mean_diff <- numeric(B) # space to save the random differences
for(i in 1:B){
  resample <- sample(nrow(birthwt), size = nrow(smoke0), replace = FALSE) # sample of numbers from 1:18
  resampled_mean_diff[i] <- mean(bwt_ary[resample, ]) - mean(bwt_ary[-resample, ])
}
```

```
ggplot(data = NULL, aes(x = resampled_mean_diff)) +
  geom_histogram(binwidth = 1.25) +
  geom_vline(aes(xintercept = obs_mean_diff), colour = "dodgerblue") +
  ggtitle("Permuted mean differences in bwt")
```




```
sum(resampled_mean_diff >= obs_mean_diff) # greater than or equal to
```

```
[1] 395
```

```
sum(resampled_mean_diff <= obs_mean_diff) # less than or equal to
```

```
[1] 99604
```

```
min_sum <- min(sum(resampled_mean_diff >= obs_mean_diff), sum(resampled_mean_diff <= obs_mean_diff))
min_sum
```

```
[1] 395
```

```
#Compute P-value
min_p <- sum(min_sum + 1)/(B + 1)
c(min_p, 2*min_p)
```

```
[1] 0.00396 0.00792
```

So a difference of 283.7767 is significant at p = 0.05