Maps Search A/B Experiment

Ankit Mathur

3/18/2019

We perform Welch's t-test on the overall dataset as well as on specific cohorts within the dataset to check for significance.

```r
# Read the data
apple_data = read.csv("AB-test.csv")
```

```r
# Create a function to compute and return CTRs and p-values
cal_ab <- function(abtest){
  n_A <- nrow(subset(abtest, abtest$Branch=="A"))
  n_B <- nrow(subset(abtest, abtest$Branch=="B"))

  clk_A <- nrow(subset(abtest, abtest$Branch=="A" & abtest$Click.Through==1))
  clk_B <- nrow(subset(abtest, abtest$Branch=="B" & abtest$Click.Through==1))

  ctr_A <- clk_A / n_A
  ctr_B <- clk_B / n_B

  # Formulae based on Bernoulli trial
  std_err <- sqrt( ctr_A*(1-ctr_A) / n_A + ctr_B*(1-ctr_B) / n_B )
  t_val <- abs(ctr_B - ctr_A) / std_err
  p_val <- 1 - pt( t_val, min(n_A-1, n_B-1) )

  return(c(ctr_A, ctr_B, p_val))
}
```

```r
# Subset the dataset by Country and iOS version
ios10_data <- subset(apple_data, apple_data$iOS==10)
ios11_data <- subset(apple_data, apple_data$iOS==11)
us_data <- subset(apple_data, apple_data$Country == "US")
uk_data <- subset(apple_data, apple_data$Country == "UK")
jp_data <- subset(apple_data, apple_data$Country == "JP")

# Compute CTRs and p-values of the original and the subsetted datasets
overall_res <- cal_ab(apple_data)
us_res = cal_ab(us_data)
uk_res = cal_ab(uk_data)
jp_res = cal_ab(jp_data)
ios10_res = cal_ab(ios10_data)
ios11_res = cal_ab(ios11_data)
```

```
## [1] "Overall:: CTR-A: 0.297405 CTR-B: 0.338677 p-value: 0.080688"

## [1] "US::      CTR-A: 0.299465 CTR-B: 0.309973 p-value: 0.377774"

## [1] "UK::      CTR-A: 0.294737 CTR-B: 0.371134 p-value: 0.131279"

## [1] "Japan::   CTR-A: 0.281250 CTR-B: 0.580645 p-value: 0.008748"

## [1] "iOS10::   CTR-A: 0.355030 CTR-B: 0.305221 p-value: 0.145256"
## [1] "iOS11::   CTR-A: 0.268072 CTR-B: 0.372000 p-value: 0.004150"
```

At an overall level, branch B does not perform better than branch A as we fail to reject the null hypothesis with a p-value of 0.08. However, when we slice the data based on Country, we observe that the p-value drops to 0.0087 for Japan. Hence, we can say with 99% confidence that branch B performs significantly better than branch A in Japan.

Additionally, upon slicing the original dataset by iOS version, the p-value corresponding to iOS version 11 turns out to be 0.004. Therefore, we are 99% confident that branch B performs significantly better than branch A in the iOS version 11 cohort.

```r
# Subset the dataset further based upon the combinations of
# Country and iOS version for unsignificant results
us_ios10_data <- subset(us_data, us_data$iOS==10)
us_ios11_data <- subset(us_data, us_data$iOS==11)
uk_ios10_data <- subset(uk_data, uk_data$iOS==10)
uk_ios11_data <- subset(uk_data, uk_data$iOS==11)

us_ios10_res <- cal_ab(us_ios10_data)
us_ios11_res <- cal_ab(us_ios11_data)
uk_ios10_res <- cal_ab(uk_ios10_data)
uk_ios11_res <- cal_ab(uk_ios11_data)

us_ios10_res
```

```
## [1] 0.36000000 0.28494624 0.08434282
```

```r
us_ios11_res
```

```
## [1] 0.26907631 0.33513514 0.07039379
```

```r
uk_ios10_res
```

```
## [1] 0.3235294 0.3333333 0.4626957
```

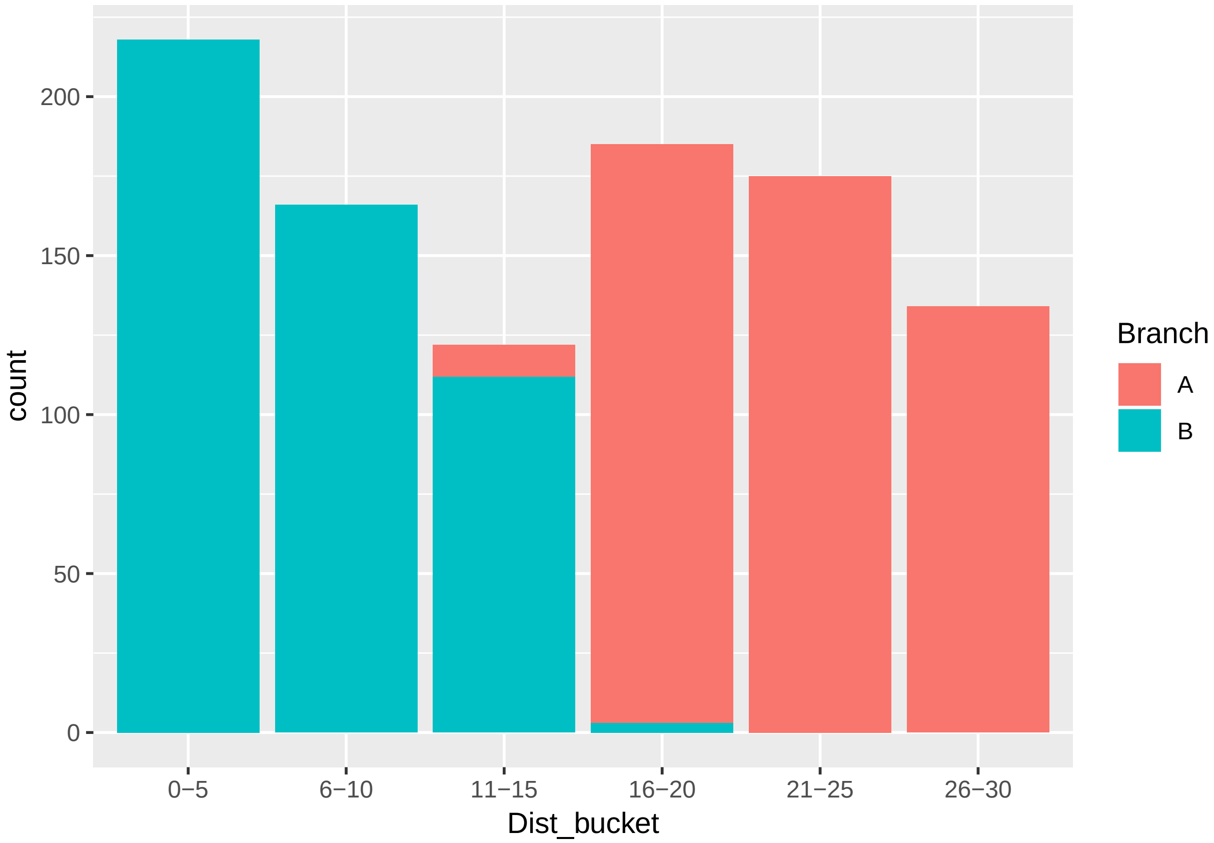
```r
uk_ios11_res
```

```
## [1] 0.27868852 0.41304348 0.07676455
```

Nothing interesting is observed when we try different combinations of countries and iOS versions that had previously failed the significance test.

```r
apple_data$Dist_bucket <- cut_number(apple_data$Distance, n=6, label=c("0-5", "6-10", "11-15", "16-20",
dist_vec <- c("0-5", "6-10", "11-15", "16-20", "21-25", "26-30")
```

```r
ggplot(apple_data, aes(x=Dist_bucket)) + stat_count(aes(fill=Branch))
```



As we can see in the plot above, the branches in the A/B experiment are split at ~15 mile mark with a very small overlap. Thus, we can deduce that users were assigned the A/B bucket based upon the distance of their search result from their location.