Statistical Experiments and Significance Testing

Fabian Peri

October 7, 2018

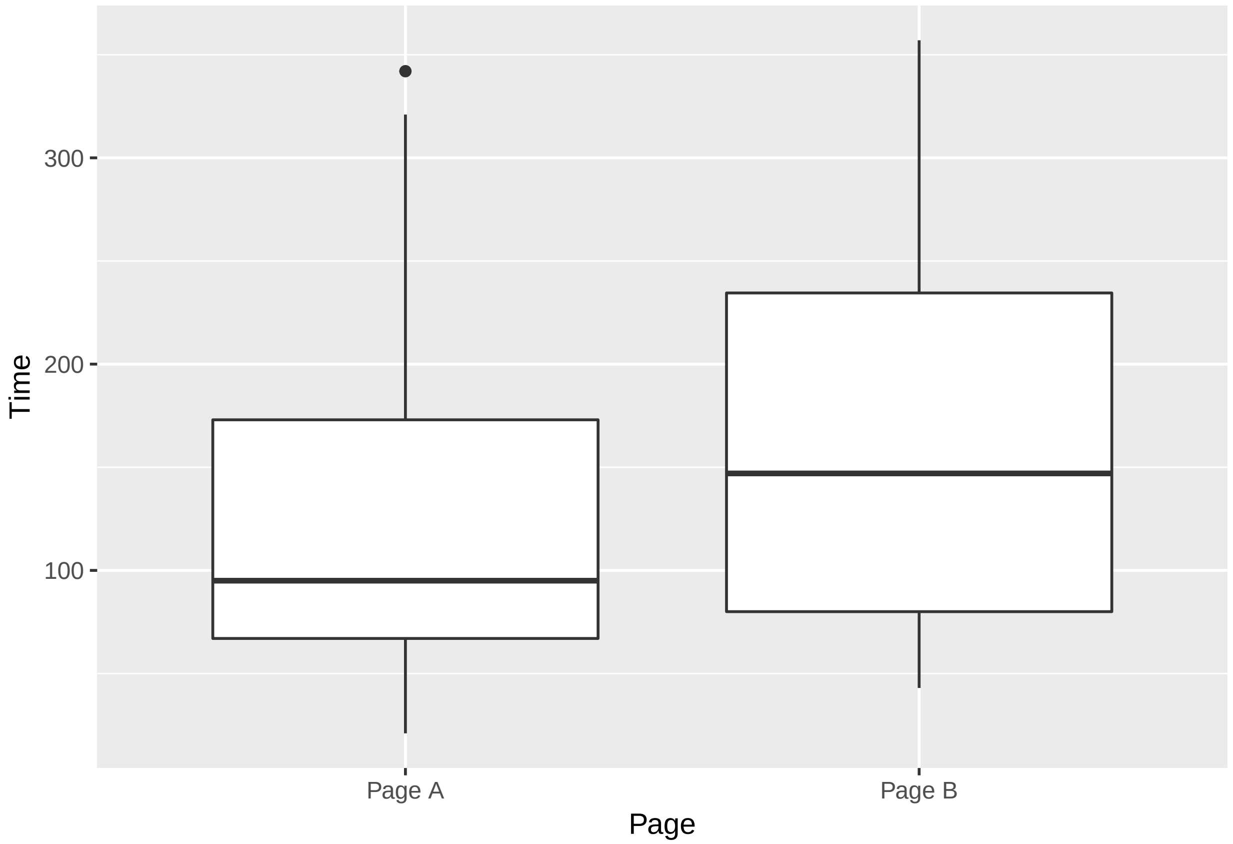
```
# packages needed for chapter 3
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lmPerm)
```

```
# Import the datasets needed for chapter 3
PSDS_PATH <- file.path('C:/Users/fabia/Desktop', 'psds_data')

session_times <- read.csv(file.path(PSDS_PATH, 'data', 'web_page_data.csv'))
session_times[,2] <- session_times[,2] * 100
four_sessions  <- read.csv(file.path(PSDS_PATH, 'data', 'four_sessions.csv'))
click_rate <-  read.csv(file.path(PSDS_PATH, 'data', 'click_rates.csv'))
imanishi <-  read.csv(file.path(PSDS_PATH, 'data', 'imanishi_data.csv'))
```

```
## Code snippet 3.1
ggplot(session_times, aes(x=Page, y=Time)) +
  geom_boxplot()
```



```
## Code for Figure 3
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0303.png'),  width = 4, height=4, units='in', res=30
ggplot(session_times, aes(x=Page, y=Time)) +
  geom_boxplot() +
  labs(y='Time (in seconds)') +
  theme_bw()
dev.off()
```

```
## pdf
##   2
```

```
mean_a <- mean(session_times[session_times['Page']=='Page A', 'Time'])
mean_b <- mean(session_times[session_times['Page']=='Page B', 'Time'])
mean_b - mean_a
```

```
## [1] 35.66667
```

```
## Permutation test example with stickiness
perm_fun <- function(x, n1, n2)
{
  n <- n1 + n2
  idx_b <- sample(1:n, n1)
  idx_a <- setdiff(1:n, idx_b)
  mean_diff <- mean(x[idx_b]) - mean(x[idx_a])
  return(mean_diff)
}
```

```
## Code for Figure 4
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0304.png'),  width = 4, height=4, units='in', res=30
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = perm_fun(session_times[,'Time'], 21, 15)
par(mar=c(4,4,1,0)+.1)
hist(perm_diffs, xlab='Session time differences (in seconds)', main='')
abline(v = mean_b - mean_a)
dev.off()
```

```
## pdf
##   2
```

```
mean(perm_diffs > (mean_b - mean_a))
```

```
## [1] 0.142
```

```
## Code for Figure 5
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0305.png'),  width = 4, height=4, units='in', res=30

obs_pct_diff <- 100*(200/23739 - 182/22588)
conversion <- c(rep(0, 45945), rep(1, 382))
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = 100*perm_fun(conversion, 23739, 22588 )
hist(perm_diffs, xlab='Conversion rate (percent)', main='')
abline(v = obs_pct_diff, lty=2, lwd=1.5)
text("  Observed\n  difference", x=obs_pct_diff,  y=par()$usr[4]-20, adj=0)

dev.off()
```

```
## pdf
##   2
```

```
mean(perm_diffs > obs_pct_diff)
```

```
## [1] 0.318
```

```
prop.test(x=c(200,182), n=c(23739,22588), alternative="greater")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(200, 182) out of c(23739, 22588)
## X-squared = 0.14893, df = 1, p-value = 0.3498
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.001057439  1.000000000
## sample estimates:
##      prop 1      prop 2
## 0.008424955 0.008057376
```

```
## Histogram of resample
## t-test
t.test(Time ~ Page, data=session_times, alternative='less' )
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  Time by Page
## t = -1.0983, df = 27.693, p-value = 0.1408
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 19.59674
## sample estimates:
## mean in group Page A mean in group Page B
##             126.3333             162.0000
```

```
## session times

## four groups ANOVA

## Code for Figure 6
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0306.png'),  width = 4, height=4, units='in', res=30

ggplot(four_sessions, aes(x=Page, y=Time)) +
  geom_boxplot() +
  labs(y='Time (in seconds)') +
  theme_bw()

dev.off()
```

```
## pdf
##   2
```

```
summary(aovp(Time ~ Page, data=four_sessions))
```

```
## [1] "Settings:  unique SS "
```

```
## Component 1 :
##             Df R Sum Sq R Mean Sq Iter Pr(Prob)
## Page         3    831.4    277.13 4054  0.08905 .
## Residuals   16   1618.4    101.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Time ~ Page, data=four_sessions))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Page         3  831.4   277.1    2.74 0.0776 .
## Residuals   16 1618.4   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Chi square test
clicks <- matrix(click_rate$Rate, nrow=3, ncol=2, byrow=TRUE)
dimnames(clicks) <- list(unique(click_rate$Headline), unique(click_rate$Click))

chisq.test(clicks, simulate.p.value=TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  clicks
## X-squared = 1.6659, df = NA, p-value = 0.4783
```

```
chisq.test(clicks, simulate.p.value=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  clicks
## X-squared = 1.6659, df = 2, p-value = 0.4348
```

```
## Code for Figure 7
x <- seq(1, 30, length=100)
chi <- data.frame(df = factor(rep(c(1, 2, 5, 10), rep(100, 4))),
                  x = rep(x, 4),
                  p = c(dchisq(x, 1), dchisq(x, 2), dchisq(x, 5), dchisq(x, 20)))

png(filename=file.path(PSDS_PATH, 'figures', 'psds_0307.png'),  width = 5, height=3, units='in', res=30

ggplot(chi, aes(x=x, y=p)) +
  geom_line(aes(linetype=df)) +
  theme_bw() +
  labs(x='', y='')

dev.off()
```

```
## pdf
##   2
```

```
## Fishers exact test
fisher.test(clicks)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  clicks
## p-value = 0.4824
## alternative hypothesis: two.sided
```

```
## Tufts example


## Code for Figure 8
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0308.png'),  width = 4, height=4, units='in', res=30
imanishi$Digit <- factor(imanishi$Digit)

ggplot(imanishi, aes(x=Digit, y=Frequency)) +
  geom_bar(stat="identity") +
  theme_bw()

dev.off()
```

```
## pdf
##   2
```