Homework 5

Anishka Chauhan

2022-08-01

```
library(dplyr)
```

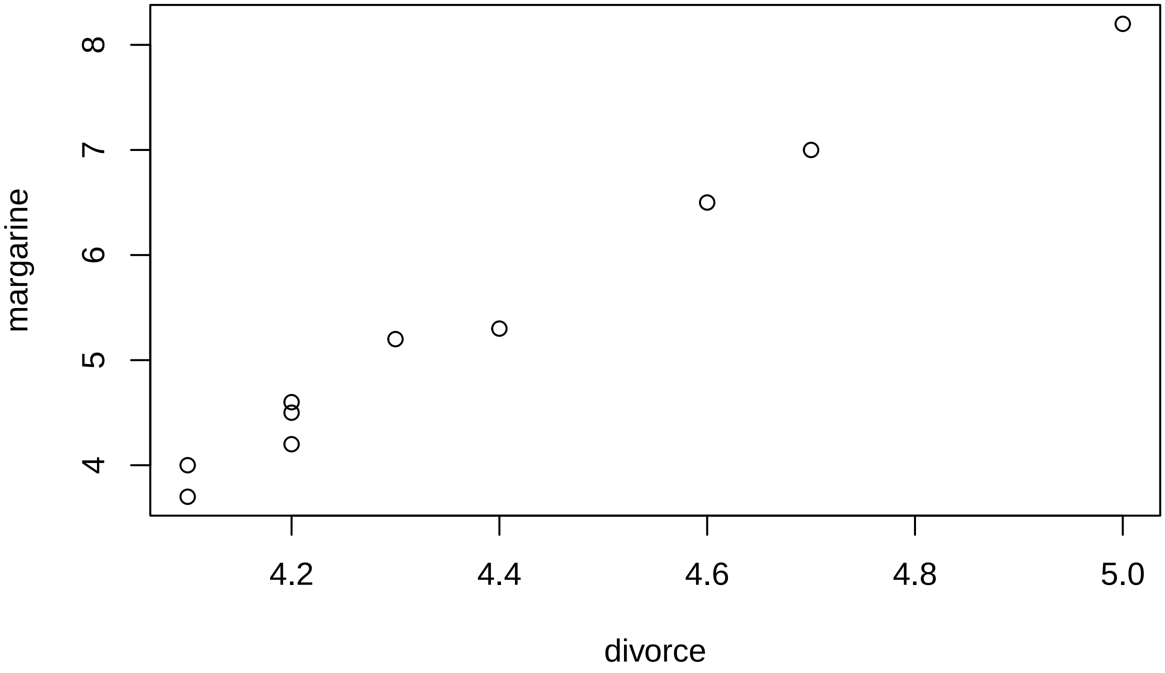
```
##
## Attaching package: 'dplyr'

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

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

```
library(ggplot2)
```

```
# from http://tylervigen.com/view_correlation?id=1703
divorce = c(5, 4.7, 4.6, 4.4, 4.3, 4.1, 4.2, 4.2, 4.2, 4.1)
margarine = c(8.2, 7, 6.5, 5.3, 5.2, 4, 4.6, 4.5, 4.2, 3.7)
plot(divorce, margarine)
```



The variables appear to have a linear correlation between each other.

2)

```
cor(divorce, margarine)
```

```
## [1] 0.9925585
```

```
boot_cor <- function(x, y, B=10000){

  n <- length(x)
  boot_stats <- matrix(nrow=B)

  for(i in 1:B){
    indices <- sample(n, replace=TRUE)
    boot_stats[i] <- cor(x[indices], y[indices])
  }

  return(boot_stats)
}
cor_sd = sd(boot_cor(divorce, margarine))
cor_sd
```

```
## [1] 0.01010831
```

```
c(cor(divorce, margarine) - 1.96*cor_sd, cor(divorce, margarine) + 1.96*cor_sd)
```

```
## [1] 0.9727462 1.0123707
```

0 is not in this confidence interval, therefore the null hypothesis should be rejected, meaning divorce and margarine consumption are not related.

3) Time trends could be a reason why these two variables show an association. It could be that over time the rate of divorce and the rate of margarine consumption could increase.

4) It assumes that the sample the bootstrap is drawing from is unbiased. That assumption can't hold as the null hypothesis is rejected.

5)

```
nbc_data = sample(c("Clinton", "Trump", "Other"), 40816, replace = TRUE, prob = c(0.51, 0.44,0.05))

boot_poll = function (x, B = 1000) {

   n <- length(x)
  boot_stats <- matrix(nrow=B)

  for(i in 1:B){
    indices <- sample(n, replace=TRUE)
    s = prop.table(table(x[indices]))
    pC = s[1]
    pT = s[3]
    boot_stats[i] <- pC - pT
  }

  return(boot_stats)
}
nbc_samples = boot_poll(nbc_data)
nbc_sd = sd(nbc_samples)
paste0("NBC/Survey Monkey Interval: ", (0.51 - 0.44) - 1.96*nbc_sd, ", ",  (0.51 - 0.44) + 1.96*nbc_sd)
```

```
## [1] "NBC/Survey Monkey Interval: 0.0605580850336431, 0.0794419149663569"
```

```
abc_data = sample(c("Clinton", "Trump", "Other"), 1128, replace = TRUE, prob = c(0.48, 0.47,0.05))
abc_samples = boot_poll(abc_data)

abc_sd = sd(abc_samples)
paste0("ABC/Washington Post Interval ", (0.48 - 0.47) - 1.96*abc_sd, ", ",  (0.48 - 0.47) + 1.96*abc_sd
```

```
## [1] "ABC/Washington Post Interval -0.0462581083185233, 0.0662581083185233"
```

6)

```
nbc_null= sample(c("Clinton", "Trump", "Other"), 40816, replace = TRUE, prob = c(0.475, 0.475,0.05))
abc_null = sample(c("Clinton", "Trump", "Other"), 1128, replace = TRUE, prob = c(0.51, 0.44,0.05))

nbc_nullsd = sd(boot_poll(nbc_null))
abc_nullsd = sd(boot_poll(abc_null))

paste0("NBC/Survey Monkey Interval Under Null: ",  - 1.96*nbc_nullsd, ", ", 1.96*nbc_nullsd)
```

```
## [1] "NBC/Survey Monkey Interval Under Null: -0.00946959270488642, 0.00946959270488642"
```

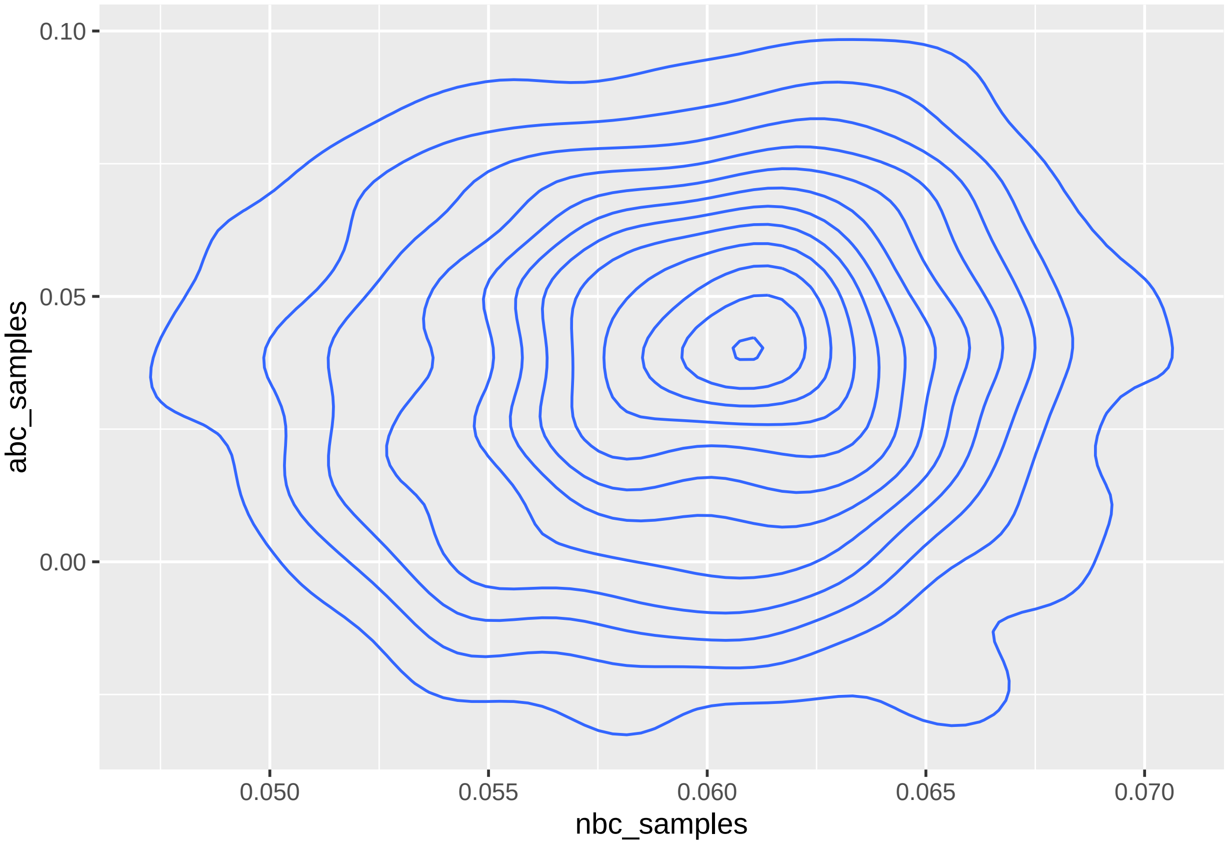
```
paste0("ABC/Washington Post Interval Under Null: ",  - 1.96*abc_nullsd, ", ", 1.96*abc_nullsd)
```

```
## [1] "ABC/Washington Post Interval Under Null: -0.0560839039286535, 0.0560839039286535"
```

Best estimate is 0.475

7)

```
x = as.data.frame(cbind(nbc_samples, abc_samples))
colnames(x) = c("nbc_samples", "abc_samples")
ggplot(x, aes(nbc_samples, abc_samples)) + geom_density_2d()
```



The density plot is centered at the point where the nbc difference is 0.075 and the abc difference is 0.04.