Biostatistics Homework 1

2023-02-21

By Lumi (12112618)

```
library(ggplot2)
library(ggrepel)
library(ggpubr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("/Users/lumizhang/Documents/sustech/biology/classes/Biostatistics/Homework/hw1")
```

1 Discrete or Continuous

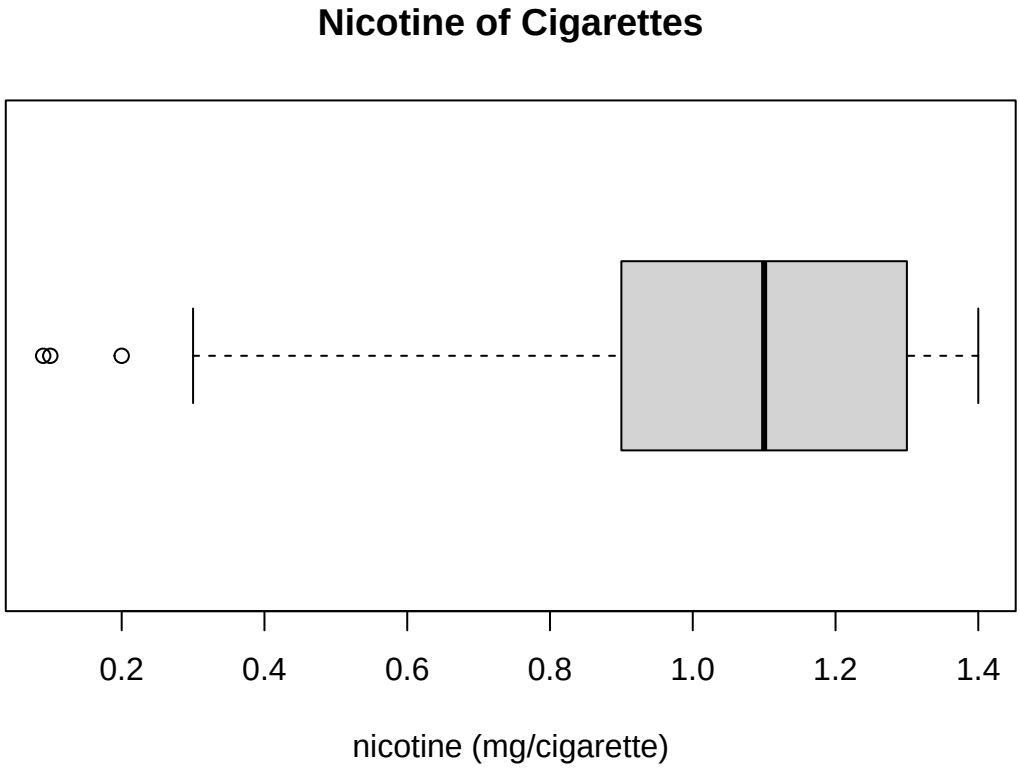
1.1 - 1.4 Discrete, Continuous, Continuous, Discrete.

2 Canadian Cigarettes

2.1

As we can see from the data, the majority of the concentration of nicotine per cigarette lay within the range of 0.9 to 1.3, with a median of 1.1. There are three outliers with extreme low concentration of nicotine, which are Brand 35, 7 and 23.

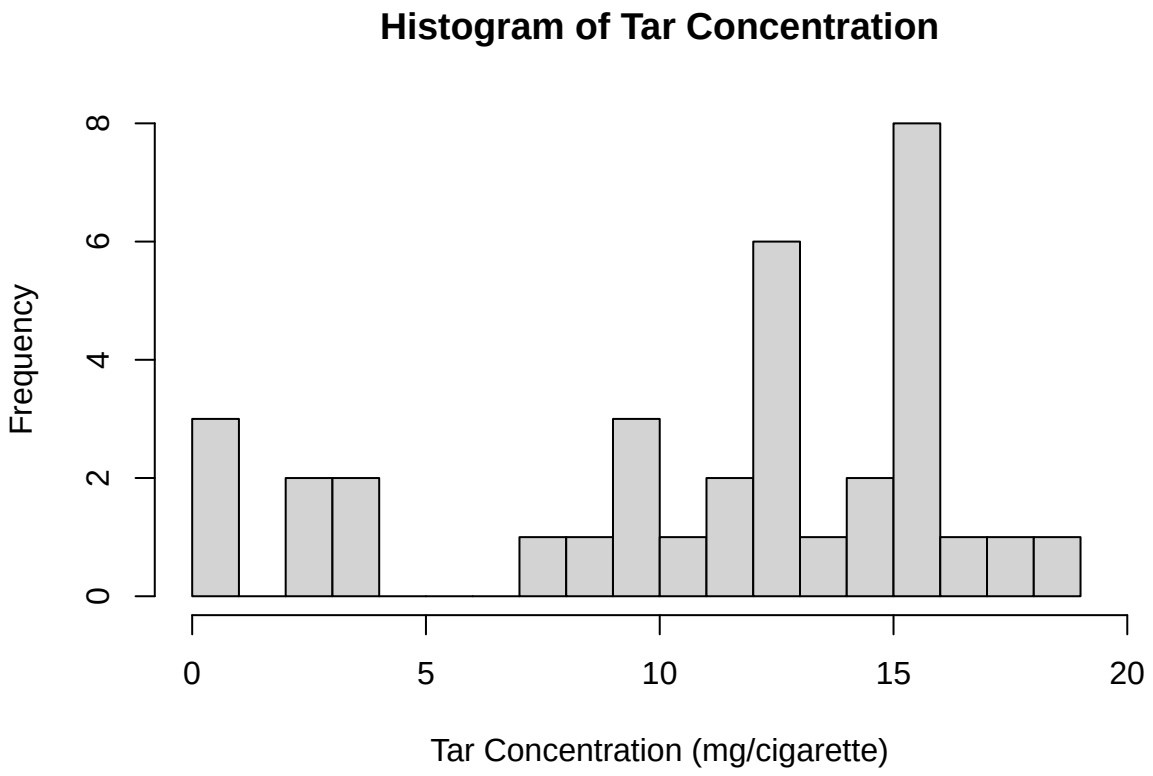
```
input_data <- read.csv("cigarettes.csv")
boxplot(input_data[,"nicotine"], main = "Nicotine of Cigarettes", xlab = "nicotine (mg/cigarette)", hor
```



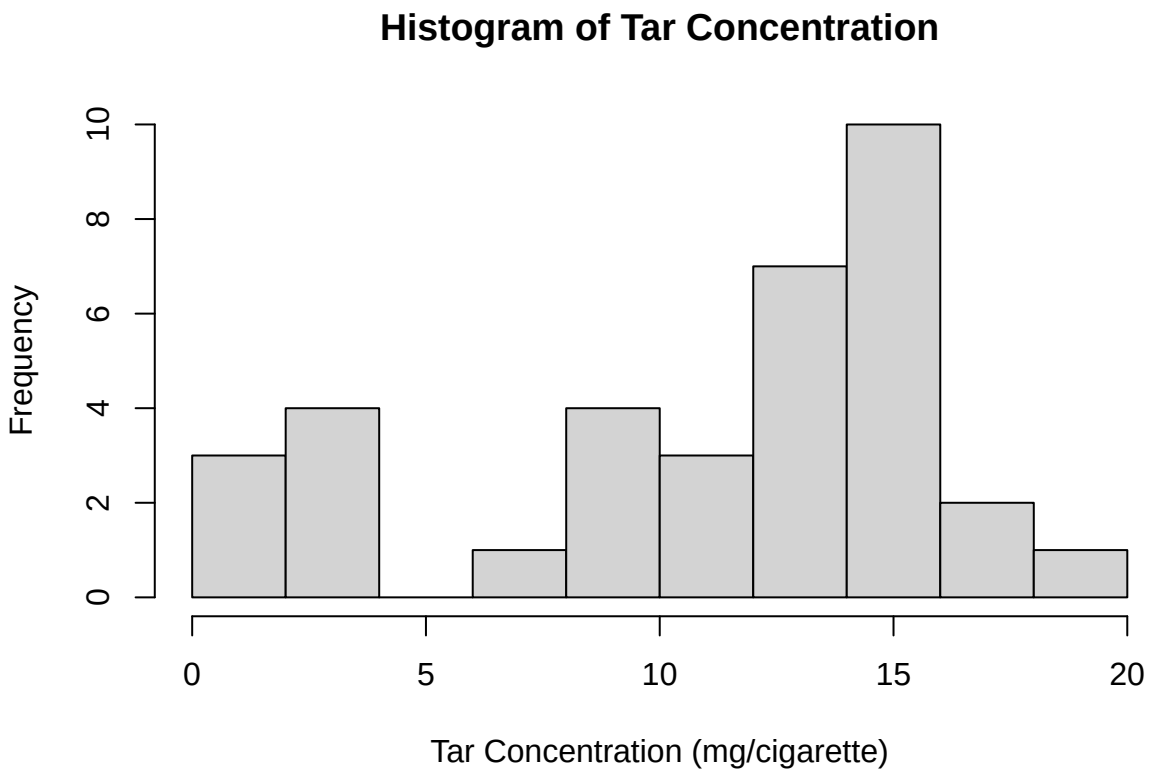
2.2

Most of the brands have a tar concentration over 10mg. With 16mg/cigarette being the most frequent value, and 13mg being the second most frequent. There are also some brands of cigarettes with very small amount of tar, 7 brands have tar concentration under 5mg.

```
hist(input_data[, "tar"],
     main = "Histogram of Tar Concentration",
     xlab = "Tar Concentration (mg/cigarette)",
     breaks = 20, xlim = c(0,20))
```



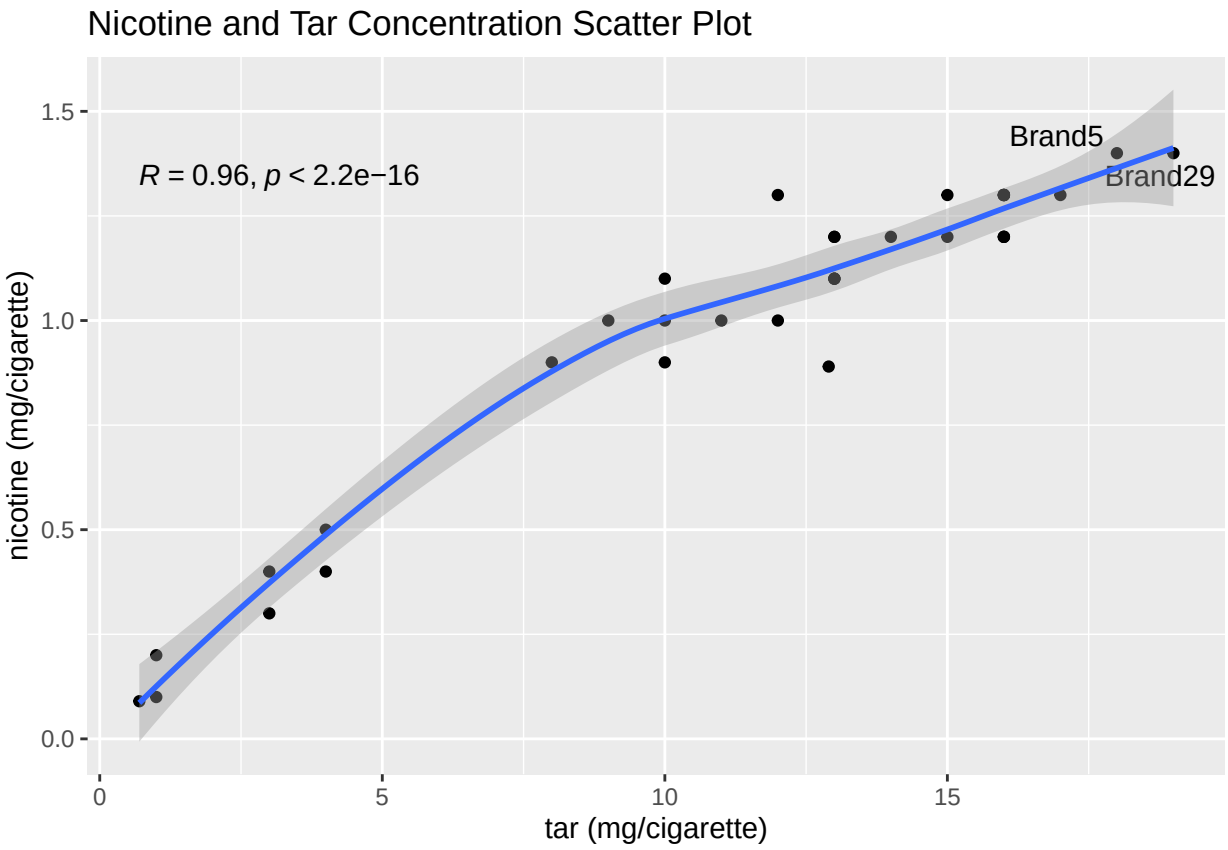

```
hist(input_data[, "tar"],
     main = "Histogram of Tar Concentration",
     xlab = "Tar Concentration (mg/cigarette)",
     breaks = 10)
```



2.3

```
ggplot(input_data, mapping = aes(x = tar, y = nicotine)) +
  geom_point() +
  geom_text_repel(data = input_data[input_data$tar > 15 & input_data$nicotine > 1.3,],
                  mapping = aes(label = Brand)) +
  labs(
    x = "tar (mg/cigarette)",
    y = "nicotine (mg/cigarette)",
    title = "Nicotine and Tar Concentration Scatter Plot"
  ) +
  geom_smooth() +
  stat_cor()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



2.4

It appears that the concentration of tar and nicotine in a cigarette is highly positively correlated (with correlation coefficient R=0.96).

3 The Unusual Die

This is not a very normal problem, because of the constraints.

3.1

There are a total of sixteen possible outcomes. The sample space is shown below. (Here we are assuming that all the dice have numbers 1-4.

```
dice_data <- read.csv("Dice.csv")
t(dice_data[,1:2])
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## Red     1    1    1    1    2    2    2    2    3     3     3     3     4     4
## Blue    1    2    3    4    1    2    3    4    1     2     3     4     1     2
##      [,15] [,16]
## Red      4     4
## Blue     3     4
```

3.2

In order for the probability of a particular outcome by proportional to the sum, we have to get the sum of all possible probabilities.

```
total <- sum(unique(dice_data$Sum))
total
```

```
## [1] 35
```

Which is 35. Since all outcomes that result in a particular sum are equally likely

```
tmp <- as.data.frame(table(dice_data$Sum))
tmp$Sum <- as.numeric(tmp$Var1) + 1
tmp$Var1 <- NULL
tmp$P <- tmp$Sum / (tmp$Freq * total)
dice_data <- merge(dice_data, tmp, by = "Sum")
dice_data[,c(1,6)]
```

```
##    Sum          P
## 1    2 0.05714286
## 2    3 0.04285714
## 3    3 0.04285714
## 4    4 0.03809524
## 5    4 0.03809524
## 6    4 0.03809524
## 7    5 0.03571429
## 8    5 0.03571429
## 9    5 0.03571429
## 10   5 0.03571429
## 11   6 0.05714286
## 12   6 0.05714286
## 13   6 0.05714286
## 14   7 0.10000000
## 15   7 0.10000000
## 16   8 0.22857143
```

So we can see from this table that:

$$P(A) = \frac{8}{35} = 0.22857143$$

3.3

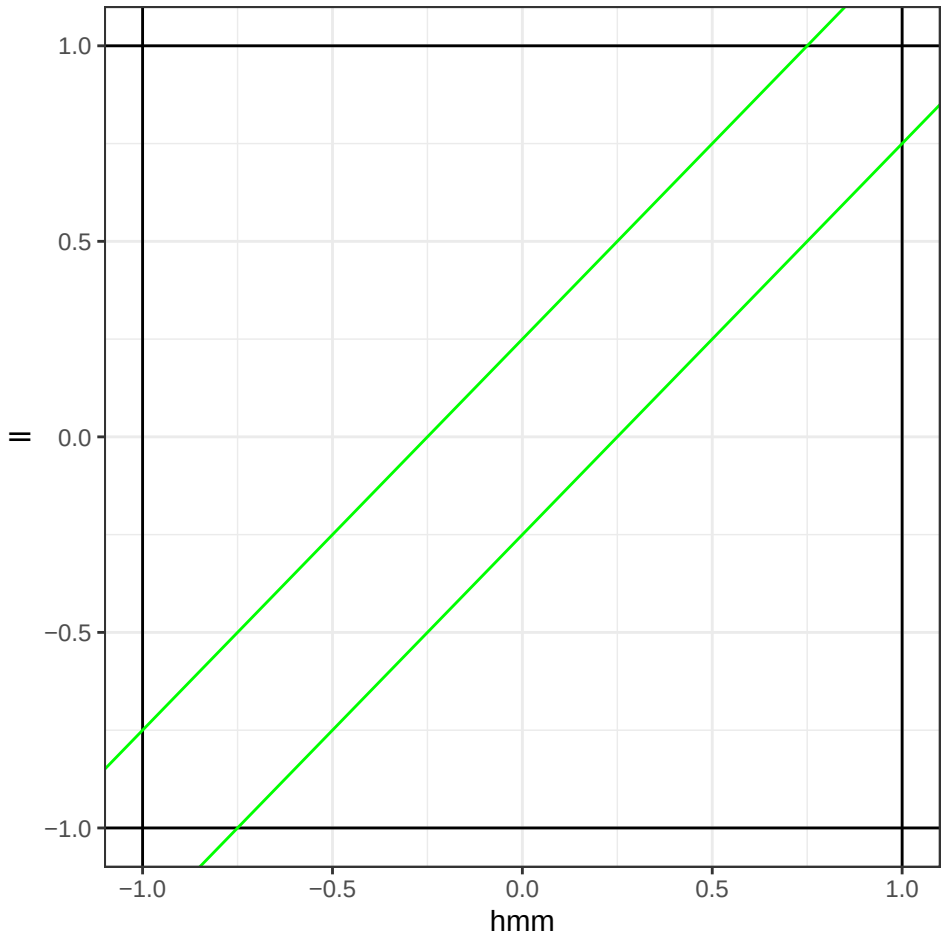
```
odd_sum <- subset(dice_data,dice_data$Sum %% 2 == 1)
sum(odd_sum$P)
```

```
## [1] 0.4285714
```

$$P(B) = \frac{3}{7} = 0.4285714$$

4 Dating

```
hmm <- c(-1,0,1)
ll <- c(-1,0,1)
ggplot(mapping = aes(hmm, ll), height = 1, width = 1)+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  geom_vline(xintercept = 1)+
  geom_vline(xintercept = -1) +
  geom_abline(intercept = 0.25, colour = "green") +
  geom_abline(intercept = -0.25, colour = "green") +
  scale_x_continuous(limits = c(-1, 1)) +
  scale_y_continuous(limits = c(-1, 1)) +
  theme_bw()
```



From the figure, we can figure out that if the two people want to meat each other, their arriving time must fall in the space between the two green lines. So we use the area's fraction to compute the possibility of them having a date.

$$P(\text{Dating}) = \frac{\text{area within green and blue line}}{\text{total area}} = \frac{15}{64} = 0.234375$$

5 Messing Round with Letters

5.1

The sample space is shown in the table.

```
sustech <- c("S", "U", "T", "E", "C", "H")
science <- c("S", "C", "I", "E", "N", "C", "E")
# colnames(sustech) <- "letters"
# colnames(science) <- "letters"
data_5 <- expand.grid(sustech = sustech, science = science)
data_5$sample_space <- paste("(",data_5$sustech,",",data_5$science, ")")
as.data.frame(t(data_5$sample_space))
```

```
##          V1        V2        V3        V4        V5        V6        V7
## 1 ( S , S ) ( U , S ) ( T , S ) ( E , S ) ( C , S ) ( H , S ) ( S , C )
##          V8        V9       V10       V11       V12       V13       V14
## 1 ( U , C ) ( T , C ) ( E , C ) ( C , C ) ( H , C ) ( S , I ) ( U , I )
##         V15       V16       V17       V18       V19       V20       V21
## 1 ( T , I ) ( E , I ) ( C , I ) ( H , I ) ( S , E ) ( U , E ) ( T , E )
##         V22       V23       V24       V25       V26       V27       V28
## 1 ( E , E ) ( C , E ) ( H , E ) ( S , N ) ( U , N ) ( T , N ) ( E , N )
##         V29       V30       V31       V32       V33       V34       V35
## 1 ( C , N ) ( H , N ) ( S , C ) ( U , C ) ( T , C ) ( E , C ) ( C , C )
##         V36       V37       V38       V39       V40       V41       V42
## 1 ( H , C ) ( S , E ) ( U , E ) ( T , E ) ( E , E ) ( C , E ) ( H , E )
```

5.2

$$P(\text{outcome of } (\mathrm{H}, \mathrm{E})) = P(\text{H from SUSTECH}) \cdot P(\text{E from SCIENCE}) = \frac{1}{7} \cdot \frac{2}{7} = \frac{2}{49}$$

5.3

Suppose Event A is getting an S from SUSTECH, and Event B is getting an S from science. Then Event getting at least one S is $A \cup B$.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{7} + \frac{1}{7} - \frac{2}{7} \cdot \frac{1}{7} = \frac{19}{49}$$

6 Very Serious Fevers

6.1

```
temp <- c("L", "M", "H", "U")
med <- c(0,1)
ss <- expand.grid(temp, med)
colnames(ss) <- c('temperature', 'medical_ins')
ss$sample <- paste("(",ss$temperature, ",", ss$medical_ins, ")")
ss$sample
```

```
## [1] "( L , 0 )" "( M , 0 )" "( H , 0 )" "( U , 0 )" "( L , 1 )" "( M , 1 )"
## [7] "( H , 1 )" "( U , 1 )"
```

Sample space are as shown above.

6.2

```
ss[ss$temperature == "H" | ss$temperature == "U",]
```

```
##   temperature medical_ins    sample
## 3           H           0 ( H , 0 )
## 4           U           0 ( U , 0 )
## 7           H           1 ( H , 1 )
## 8           U           1 ( U , 1 )
```

All the possible outcomes of A are listed above in the sample column.

6.3

```
ss[ss$medical_ins == 0,]
```

```
##   temperature medical_ins    sample
## 1           L           0 ( L , 0 )
## 2           M           0 ( M , 0 )
## 3           H           0 ( H , 0 )
## 4           U           0 ( U , 0 )
```

All possible outcomes of B are listed above in the sample column.

6.4

```
ss[ss$medical_ins == 1 | ss$temperature == "H" | ss$temperature == "U",]
```

```
##   temperature medical_ins    sample
## 3           H           0 ( H , 0 )
## 4           U           0 ( U , 0 )
## 5           L           1 ( L , 1 )
## 6           M           1 ( M , 1 )
## 7           H           1 ( H , 1 )
## 8           U           1 ( U , 1 )
```

All the possible outcomes of $B^c \cup A$ are listed above in the sample column.

6.5

No you can't because the probability of all the outcomes are not equally likely. For example, you are more likely to have a mild fever than a very serious one.

7 A Weird Spinner Game

7.1

I think it have non transitive and transitive properties.

When you choose different spinner, one spinner is not guaranteed to win over the other spinner, that might infer the non transitive properties of this game (If being non transitive means there are no absolute hierarchy between the groups. However, if it mean that you have to form a circle, then it is not a non transitive event.).

But when you look at the final results, that is to say when you look at the numbers, they are having a transitive property. Because certain numbers are larger that the others, and therefor will win.

7.2

```
a <- c(1,5,9)
b <- c(3,4,8)
c <- c(2,6,7)
```

If Han.M.M choose spinner b:

```
b_results <- expand.grid(a,b)
b_results$win <- b_results$Var2 > b_results$Var1
b_results
```

```
##   Var1 Var2   win
## 1    1    3  TRUE
## 2    5    3 FALSE
## 3    9    3 FALSE
## 4    1    4  TRUE
## 5    5    4 FALSE
## 6    9    4 FALSE
## 7    1    8  TRUE
## 8    5    8  TRUE
## 9    9    8 FALSE
```

We can see from the table that H.M.M can win in four situations, with a winning probability:

$$P(\text{win}_b) = \frac{4}{9}$$

If Han.M.M choose spinner c:

```
c_results <- expand.grid(a,c)
c_results$win <- c_results$Var2 > c_results$Var1
c_results
```

```
##   Var1 Var2   win
## 1    1    2  TRUE
## 2    5    2 FALSE
## 3    9    2 FALSE
## 4    1    6  TRUE
## 5    5    6  TRUE
## 6    9    6 FALSE
## 7    1    7  TRUE
## 8    5    7  TRUE
## 9    9    7 FALSE
```

We can see that H.M.M can win in five situations, with winning probability:

$$P(\text{win}_c) = \frac{5}{9}$$

Since $P(\text{win}_c) > P(\text{win}_b)$, choosing the spinner c will give H.M.M a better chance of winning.

8 Messing Around with Arrows

8.1

```
rings <- data.frame(c(1:10)/10)
colnames(rings) <- "diameter"
rings$cum_size <- rings$diameter^2 * pi
rings$size <- rings$cum_size
for(i in 2: 10){
  rings$size[i] = rings$size[i] - rings$cum_size[i-1]
}
rings$ring <- c(10:1)
rings[, c("size", "ring")]
```

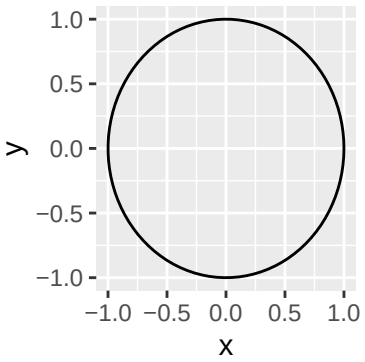
```
##          size ring
## 1  0.03141593   10
## 2  0.09424778    9
## 3  0.15707963    8
## 4  0.21991149    7
## 5  0.28274334    6
## 6  0.34557519    5
## 7  0.40840704    4
## 8  0.47123890    3
## 9  0.53407075    2
## 10 0.59690260    1
```

From the table, we can see the sizes of different rings. So if event A is getting a score of 10 when you randomly shoot an arrow, then:

$$P(\text{A}) = \frac{\text{size of ring 10}}{\text{size of target}} = \frac{0.03141593}{\pi} = \frac{1}{100}$$

8.2

```
circleFun <- function(center = c(0,0),diameter = 1, npoints = 100){
    r = diameter / 2
    tt <- seq(0,2*pi,length.out = npoints)
    xx <- center[1] + r * cos(tt)
    yy <- center[2] + r * sin(tt)
    return(data.frame(x = xx, y = yy))
}
circle <- circleFun(diameter = 2)
ggplot(circle, mapping = aes(x, y))+
  geom_path()
```



The sample space is:

$$\Omega = \{(x, y)|x^2 + y^2 \leq 1\}$$

This is uncountable.

8.3

```
rings$ring
```

```
##  [1] 10  9  8  7  6  5  4  3  2  1
```

Sample space of score you can get is shown above. Which is obviously discrete and countable.

8.4

We compute the weighted sum to get the average score.

```
rings$probability <- rings$size / pi
sum(rings$ring * rings$probability)
```

```
## [1] 3.85
```

$$\bar{x} = \sum_{i=1}^{10} p(x_i)x_i$$

And the final average score is 3.85.

8.5

No difference. Getting a score of 10's probability is still 0.01, and the sample space is still $\Omega = \{(x, y)|x^2+y^2 \leq 1\}$.