Dr Isabella Gollini

Probability and Statistics

2.2 Example: Diamonds cut and color

We use again the dataset `diamonds` included in the `ggplot2` package, but this time we focus on the variables `cut` and `color`:

- The variable `cut` represents the quality of the cut: Fair, Good, Very Good, Premium, Ideal.
- The variable `color` represents the quality of the diamond color, from J (worst) to D (best)

```
library(ggplot2)

## Warning:  package 'ggplot2' was built under R version 3.2.4

data(diamonds)
attach(diamonds)
```

We want to create the probability table that summarises the information about the diamond quality given by the variables `cut` and `color`:

```
quality <- table(cut, color) / nrow(diamonds)
quality

##            color
## cut                   D           E           F           G           H
##   Fair      0.003021876 0.004152762 0.005784205 0.005821283 0.005617353
##   Good      0.012272896 0.017296997 0.016852058 0.016147571 0.013014461
##   Very Good 0.028049685 0.044493882 0.040118650 0.042621431 0.033815350
##   Premium   0.029718205 0.043325918 0.043214683 0.054208380 0.043752317
##   Ideal     0.052539859 0.072358176 0.070930664 0.090545050 0.057749351
##            color
## cut                   I           J
##   Fair      0.003244346 0.002206155
##   Good      0.009677419 0.005691509
##   Very Good 0.022321098 0.012569522
##   Premium   0.026473860 0.014979607
##   Ideal     0.038802373 0.016611049
```

If we want to visualise the table rounded at the third digit after the decimal point:

```
round(quality, 3)
```

We want to find the probability that a randomly selected diamond is of top quality.

To be of top quality the diamond must be simultaneously of Ideal quality of cut and have the best quality of color (color `D`).

```
quality["Ideal", "D"]

## [1] 0.05253986
```

What is the probability that a randomly selected diamond is of the lowest quality?

Write here your answer

```
sum(quality["Fair","J"])

## [1] 0.002206155
```

What is the probability that a randomly selected diamond has quality color `F`?

Write here your answer

```
sum(quality[,"F"])

## [1] 0.1769003
```

What is the probability that a randomly selected diamond has a “Good” quality of the cut?

Write here your answer

```
sum(quality["Good",])

## [1] 0.09095291
```

What is the probability that a randomly selected diamond has a “Premium” quality of the cut, or quality `G` of the color, or both?

Write here your answer

```r
sum(quality["Premium",]) +
  sum(quality[,"G"]) -
  quality["Premium", "G"]

## [1] 0.4108083
```

3 Cards

We go through the example of Section 2.3 of the lecture notes; consider drawing a card at random from a standard pack of 52 cards.
What is the probability that a spade or a king is drawn?

We create a vector `values` that contains all the values a card can take:

```r
values <- c("A", 2:10, "J", "Q", "K")
values

##  [1] "A"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "J"  "Q"  "K"
```

We create the vector `suits` that contains all the card suits:

```r
suits <- c("diamonds", "clubs", "spades", "hearts")
suits

## [1] "diamonds" "clubs"    "spades"   "hearts"
```

The entire sample space can be created by using the function `expand.grid` that allows us to find all the possible combinations of values and suits.

```r
cards <- expand.grid(values = values, suits = suits)
head(cards)

##   values    suits
## 1      A diamonds
## 2      2 diamonds
## 3      3 diamonds
## 4      4 diamonds
## 5      5 diamonds
## 6      6 diamonds
```

The number of rows in `cards` corresponds to $|S|$:

```
nS <- nrow(cards)
nS

## [1] 52
```

Let A be the event that a card of spades is drawn.

```
isA <- cards[,"suits"] == "spades"
A <- subset(cards, isA)
A

##    values  suits
## 27      A spades
## 28      2 spades
## 29      3 spades
## 30      4 spades
## 31      5 spades
## 32      6 spades
## 33      7 spades
## 34      8 spades
## 35      9 spades
## 36     10 spades
## 37      J spades
## 38      Q spades
## 39      K spades
```

Let `nA` be the number of outcomes in A ($|A|$).

```
nA <- nrow(A)
nA

## [1] 13
```

$\Pr(A)$ is:

```
PrA <- nA / nS
PrA

## [1] 0.25
```

Let B be the event that a king is drawn.

```
isB <- cards[,"values"] == "K"
B <- subset(cards, isB)
B
```

```
##    values    suits
## 13      K diamonds
## 26      K    clubs
## 39      K   spades
## 52      K   hearts
```

```
nB <- nrow(B)
nB
```

```
## [1] 4
```

```
PrB <- nB / nS
PrB
```

```
## [1] 0.07692308
```

Now we have to find $A \cap B$, and $\Pr(A \cap B)$.

```
isAandB <- isA & isB
AiB <- subset(cards, isAandB)
AiB
```

```
##    values  suits
## 39      K spades
```

```
nAiB <- nrow(AiB)
nAiB
```

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

```
PrAiB <- nAiB / nS
PrAiB
```

```
## [1] 0.01923077
```

The event that a spades or a king is drawn is $A \cup B$. Using the result of Theorem 4,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

```
PrAuB <- PrA + PrB - PrAiB
PrAuB
```

```
## [1] 0.3076923
```

What is the probability that a face card is drawn? [A face card is a "J", "Q", or "K"]

```r
isD <- cards[,"values"] %in% c("J", "Q", "K")
D <- subset(cards, isD)
D

##    values    suits
## 11      J diamonds
## 12      Q diamonds
## 13      K diamonds
## 24      J    clubs
## 25      Q    clubs
## 26      K    clubs
## 37      J   spades
## 38      Q   spades
## 39      K   spades
## 50      J   hearts
## 51      Q   hearts
## 52      K   hearts

nD <- nrow(D)
PrD <- nD / nS
PrD

## [1] 0.2307692
```

What is the probability that a red face card is drawn? [Hearts and Diamonds are red suits]

```r
isE <- D[,"suits"] %in% c("diamonds", "hearts")
E <- subset(D, isE)
E

##    values    suits
## 11      J diamonds
## 12      Q diamonds
## 13      K diamonds
## 50      J   hearts
## 51      Q   hearts
## 52      K   hearts

nE <- nrow(E)
PrE <- nE / nS
PrE

## [1] 0.1153846
```

4 Combinations

4.1 Football Kits

A group of friends decided to form a new football team. They did not agree on the colors for their jerseys, so they decides to go on a website which sells football kits, and randomly selects one shirt, one pair of shorts, and one pair of socks.

The possible options on the website are:

- Shirts of 5 different colors: red, green, blue, white, yellow.
- Shorts of 3 different colors: black, white, blue.
- Socks of 2 different colors: white, red.

We want to find out:

a. How many different outfits can be created?

b. What's the probability of wearing a red shirt?

c. What's the probability of a single colored outfit?

d. What's the probability of wearing the Arsenal F.C. traditional colors (red and white only)? [The order does not count]

```r
shirts <- c("red", "green", "blue", "white", "yellow")
shorts <- c("black", "white", "blue")
socks <- c("white", "red")

outfits <- expand.grid(shirts = shirts,
                       shorts = shorts,
                       socks = socks)
head(outfits)

##   shirts shorts socks
## 1    red  black white
## 2  green  black white
## 3   blue  black white
## 4  white  black white
## 5 yellow  black white
## 6    red  white white

dim(outfits)

## [1] 30  3
```

Hint to answer question (d.): The symbol `&` means that the conditions hold simultaneously.

```
isC <- (outfits[,"shirts"] %in% c("red", "white")) &
  (outfits[,"shorts"] %in% c("red", "white")) &
  (outfits[,"socks"] %in% c("red", "white"))
```

a. How many different outfits can we create?
   The number of outfits we can create corresponds to the dimension of the sample space: that is $5 * 3 * 2$, or

```
nS <- nrow(outfits)
nS

## [1] 30
```

b. We call A the event of wearing a red shirt. This is a case of the equally likely outcomes. Only one shirt is red, out of 5 possible tops: $\Pr(A) = 1/5$.

$\Pr(A)$ can also be obtained by dividing the number of outcomes in A ($|A|$), by the number of outcomes in the sample space:

```
isA <- outfits[,"shirts"] == "red"
nA <- sum(isA)
nA

## [1] 6

PrA <- nA / nS
PrA

## [1] 0.2
```

If you want to see the outcomes contained in the event A:

```
A <- outfits[isA,]
A

##    shirts shorts socks
## 1     red  black white
## 6     red  white white
## 11    red   blue white
## 16    red  black   red
## 21    red  white   red
## 26    red   blue   red
```

c. We call B the event of obtaining a single color outfit. We can found which colors are present in shirts, shorts, and socks by using the the associative law.

```
single_col <- intersect(intersect(shirts, shorts), socks)
single_col

## [1] "white"
```

There is only one case of white shirt, white shorts, and white socks simultaneously. Since it is an equal probability case:

```
nB <- length(single_col)
PrB <- nB / nS
PrB

## [1] 0.03333333
```

d. We call C the event of all red and white kits.

The symbol `&` means that the conditions hold simultaneously.

```
isC <- (outfits[,"shirts"] %in% c("red", "white")) &
  (outfits[,"shorts"] %in% c("red", "white")) &
  (outfits[,"socks"] %in% c("red", "white"))
```

```
C <- subset(outfits, isC)
C

##    shirts shorts socks
## 6     red  white white
## 9   white  white white
## 21    red  white   red
## 24  white  white   red
```

```
nC <- nrow(C)
PrC <- nC / nS
PrC

## [1] 0.1333333
```