394D PS1

Scott Cohn

2020-07-17

1. Used Car Dealer

A used car dealer has 30 cars and 10 of them are lemons. The total number of cars is 30, $N = 30$. The marginal probability of getting a lemon is $P(L) = \frac{\text{\# of Lemons}}{N} = \frac{10}{30} = \frac{1}{3}$. Similarly, the marginal probability of getting a non-Lemon $P(\neg L) = \frac{20}{30} = \frac{2}{3}$. We are interested in the probability of getting *at least* one lemon. Let X denote the number of lemons bought so we want $P(X \geq 1)$.

We want to purchase all three cars simultaneously. We take the total number of chances of 1, 2, or 3 lemons divided by the total number of buying combinations.

$$\frac{\binom{20}{2}\binom{10}{1} + \binom{20}{1}\binom{10}{2} + \binom{20}{0}\binom{10}{3}}{\binom{30}{3}} = \frac{2800}{4060} = 0.69 = 69\%$$

So there is a 69% chance of buying at least one lemon.

2. Two Dice

We throw two dice.

What is probability that the sum of the two numbers is odd?

A single die has values 1-6, so the sum of any two die is in the closed interval $[2, 12]$. However, since each die has 6 sides, there are 36 possible combinations to achieve one of those sums. Thus, $N = 36$. Of those 36, half are odd — $n_{\text{odd}} = 18$.

It follows that the probability that the sum of the two numbers is odd is:

$$P(\text{sum is odd}) = \frac{n_{\text{odd}}}{N} = \frac{18}{36} = \frac{1}{2}.$$

What is the probability that the sum of the two numbers is less than 7?

There are 15 possible combinations of rolls, denoted set Ω, where the sum is strictly less than 7:

$$\Omega = \{(1,1),(1,2),(2,1),(1,3),(3,1),(1,4),(4,1),(1,5),(5,1),(2,2),(2,3),(3,2),(3,3),(2,4),(4,2)\}$$

Therefore,

$$P(\text{sum} < 7) = \frac{\text{length}(\Omega)}{N} = \frac{15}{36} = \frac{5}{12} \approx 42\%$$

What is the probability that the sum of the two numbers is less than 7 given that it is odd?

The above is shown mathematically as $P(\text{sum} < 7|\text{sum is odd})$. Let A denote the statement "sum < 7" and B denote the statement "sum is odd". Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

There are 18 cases where the sum is odd:

$$\Omega = \{(1,2),(2,1),(1,4),(4,1),(1,6),(6,1),(2,3),(3,2),\\ (2,5),(5,2),(3,4),(4,3),(3,6),(6,3),(4,5),(5,4),(5,6),(6,5)\}$$

We know from above that $P(B) = \frac{18}{36} = \frac{1}{2}$. It remains to find $P(A \cap B)$.

We want the intersection of statement A and statement B. Thus, $P(A \cap B)$ is equal to all of the sums in set Ω where the sum is less than 7 divided by 36 total combinations.

$$A \cap B = \{\omega \in \Omega : \omega \text{ is odd sum and } \omega \text{ sum is} < 7\}$$

There are 6 such combinations of odds with a sum less than 7:

$$\Omega_{\omega<7} = \{(1,2),(2,1),(1,4),(4,1),(2,3),(3,2)\}$$

So, $P(A \cap B) = \frac{6}{36} = \frac{1}{6}$. Putting it all together, we see that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3} \approx 33\%$$

Are these two events independent?

If the events are independent, then $P(A \cup B) = P(A) \times P(B)$. We also know that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. If these evaluate to be the same value, then events A and B are independent.

We know the following:

- $P(A) = \frac{5}{12}$
- $P(B) = \frac{1}{2}$
- $P(A \cap B) = \frac{1}{6}$

Thus,

$$
\begin{aligned}
P(A \cup B) &= P(A) \times P(B) \\
&= \frac{5}{12} \times \frac{1}{2} \\
&= \frac{5}{24}
\end{aligned}
$$

$$
\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\
&= \frac{5}{12} + \frac{1}{2} - \frac{1}{6} \\
&= \frac{3}{4}
\end{aligned}
$$

It clearly follows that $\frac{5}{24} \neq \frac{3}{4}$. Hence, the two events are *not* independent. Before doing any work, we can also observe that because their intersection is non-zero, the events are independent.

3. Random Clicker and Truthful Clicker

There are two types of users that access the website: RC and TC. RC's make up 30% of users. Then, TC = 1 - RC = 70%.

The marginal probabilities for user-type are:

- $P(TC) = 0.7$
- $P(RC) = 0.3$

It is given that RC-types will click Yes/No equally. Then the conditional probabilities for RC are:

- $P(Y|RC) = 0.5$
- $P(N|RC) = 0.5$

Using the marginal and conditional probabilities for RC-types, we compute the joint probabilities.

- $P(Y \cap RC) = P(RC)P(Y|RC) = 0.15$
- $P(N \cap RC) = P(RC)P(N|RC) = 0.15$

These probabilities are visualized in figure 1. It remains to find the conditional and joint probabilities for the TC-type. After a trial period, we know 65% said Yes and 35% said No. It follows that:

$$
\begin{aligned}
P(Y) &= P(Y \cap TC) + P(Y \cap RC) && && (1)\\
0.65 &= P(Y \cap TC) + 0.15 && \text{Plugging in.} && (2)\\
\therefore P(Y \cap TC) &= 0.5 && && (3)
\end{aligned}
$$

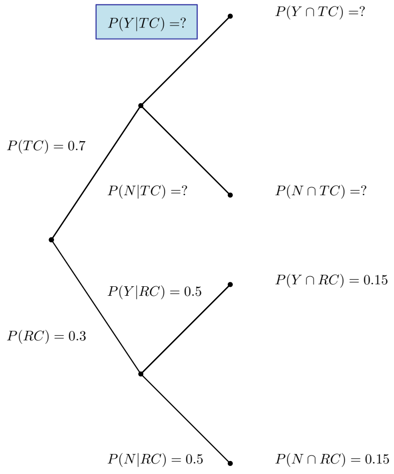

Figure 1: Tree for TC and RC

Similarly,

$$\begin{aligned} P(N) &= P(N \cap TC) + P(N \cap RC) && (4)\\ 0.35 &= P(N \cap TC) + 0.15 \qquad \text{Plugging in.} && (5)\\ \therefore P(N \cap TC) &= 0.2 && (6) \end{aligned}$$

Using the joint probability $P(Y \cap TC)$, we can compute $P(Y|TC)$ using the following formula:

$$P(Y|TC) = \frac{P(Y \cap TC)}{P(TC)} = \frac{0.5}{0.7} \approx 0.714$$

We interpret $P(Y|TC)$ as the probability that a person clicked yes *given* that they are a truthful clicker. Equivalently, this is the fraction of people who are truthful clickers that answered yes. About 71% of truthful clickers (TC) clicked yes.

4. Medical Test

Well this is a topical question ...

The marginal probabilities for having the disease are:[1]

- $P(D) = 0.000025$
- $P(\neg D) = 1 - P(D) = 0.999975$

From the sensitivity, we know the conditional probability of testing positive given that an individual has the disease is $P(+|D) = 0.993$. We can then calculate the conditional probability of testing negative given that an individual has the disease: $P(-|D) = 1 - P(+|D) = 0.007$. Thus, we have:

- $P(+|D) = 0.993$
- $P(-|D) = 0.007$

Similarly, we use the specificity to get $P(-|\neg D) = 0.9999$ and $P(+|\neg D) = 1 - P(-|\neg D) = 0.0001$.

- $P(-|\neg D) = 0.9999$
- $P(+|\neg D) = 0.0001$

Multiplying the marginal probabilities by the conditional probabilities yields the following joint probabilities:

We can check these are correct by making sure they sum to 1

- $P(+ \cap D) = 0.00002482$
- $P(- \cap D) = 0.00000018$
- $P(+ \cap \neg D) = 0.0001$
- $P(- \cap \neg D) = 0.999875$

[1] taken from bullet 3

Now supposing someone tests positive, we want the probability that they have the disease: $P(D|+)$. To find this probability, we use Bayes:

$$P(D|+) = \frac{P(D)P(+|D)}{P(+)}$$

We don't have the marginal probability of being positive, so we need to expand the denominator as follows:

$$P(D|+) = \frac{P(D)P(+|D)}{P(D)P(+|D) + P(\neg D)P(+|\neg D)}$$

Plugging in the values from above yields the following:

$$P(D|+) = \frac{(0.000025)(0.993)}{(0.000025)(0.993) + (0.999975)(0.0001)} = 0.19888 \approx 20\%$$

This result indicates that if an individual were to test positive then it is a true positive only 20 % of the time. This suggests a false positive rate that is quite large. If this is the case, an epidemiologist may ask "What if we're immune and don't know it?". To limit the rate of false positives, researchers would likely have to sample in regions where they know the disease is more prevalent, thus forcing a sample bias. Or, through very large sampling, these false positives could be factored out.

The primary concern is poor information. With high false positives, researchers are unable to accurately gauge progress in the fight against disease spread. Furthermore, patients may be exposed to treatments that they may not need. Depending on the treatment, this could be harmful. Alternatively, admittance to a hospital under the auspice of a being positive may in fact expose them to the disease when they were in fact negative in the first place.

5. One Match to Go[2]

Spiegelhalter and Ng's goal was to accurately predict soccer match results based on deviations from the norm for attacking and defense statistics, while controlling for home and away effects. They define "attack strength" as the ratio of goals scored by a team over the average number of goals scored by a team. Similarly, "defense weakness" is the ratio of goals that the team has let (or been scored against) over the average across the league. They also find the average number of goals scored by a home team and an away team to create a baseline.

Then, they ask "how many goals do we expect a team to score?" The expected number of goals comes from weighting the average home or away benchmark by the respective team's "attack strength" and the opposition team's "defense weakness". The expectation allows the authors to use a Poisson distribution to find the probability that a

[2] Data from https://www.soccerstats.com/latest.asp?league=england

team will reach a certain score in a given match-up. Furthermore, this allows the analyst to determine the probability of a given score by multiplying the each team's score probability together. Spiegelhalter and Ng use this method to compute the first, second, and third most likely result of match-ups based on scoring history. Teasing these values further, Spiegelhalter and Ng are able to construct a table of the percent probability of a home win, away win, or draw. Ex-post, they check the accuracy of their predictions with a Brier penalty, which is typically used in weather forecasting.

Table 1: State of the Premier League as of **July 13, 2020**

Team	Points	Goals for	Attack strength	Goals against	Defense weakness
Liverpool	93	76	1.60	27	0.57
Manchester C.	72	91	1.91	34	0.72
Chelsea	60	63	1.32	49	1.03
Leicester City	59	65	1.37	36	0.76
Manchester Utd	59	61	1.28	35	0.74
Wolverhampton	55	48	1.01	37	0.78
Sheffield Utd	54	38	0.80	33	0.69
Tottenham	52	54	1.14	45	0.95
Arsenal	50	51	1.07	44	0.93
Burnley	50	39	0.82	47	0.99
Everton	45	41	0.86	52	1.09
Southampton	45	45	0.95	58	1.22
Newcastle Utd	43	36	0.76	52	1.09
Crystal Palace	42	30	0.63	45	0.95
Brighton	36	36	0.76	52	1.09
West Ham Utd	34	44	0.93	59	1.24
Watford	34	33	0.69	54	1.14
Bournemouth	31	36	0.76	60	1.26
Aston Villa	30	38	0.80	65	1.37
Norwich City	21	26	0.55	67	1.41

Home teams scored on average 1.52 goals per match, while away teams scored 1.2 goals per match.[3]

First, we wish the predict the result of Liverpool (home) *versus* Tottenham (away). If Liverpool were average, we would expect them to score 1.52 goals. Liverpool is not average and scores 160% of the average number of goals.[4] Multiplying $1.52 \times 1.60 = 2.43$. We expect them to score 2.43 goals against an average team. However, we must account for the defense of Tottenham, the away team. Tottennam's "defense weakness" is 0.95. That is, they concede 95% of the average

[3] Given in online table, not calculated

[4] Liverpool's "attack strength" is 1.60

goals scored. Amending the previous calculation by factoring in the defense, 1.52 × 1.60 × 0.95 = 2.31 expected goals by Liverpool.

We apply a similar approach to the Tottenham squad: Amending the baseline of 1.2 yields 1.2 × 1.14 × 0.57 = 0.78 expected goals in a match against Liverpool.

Table 2: Expected number of goals, and percentage chance of getting a particular score for the two teams, assuming a Poisson distribution

Team	Expected goals	0	1	2	3	4	5
Liverpool	2.43	9	21	26	21	13	6
Tottenham	0.78	46	36	14	4	1	0

Next, we want to replicate the above analysis for Manchester United (home) *versus* Manchester City (away). Again, the home team — if average — will score 1.52 goals a game. Manchester United has an "attack strength" of 1.28 and Manchester City has a "defense weakness" of 0.72. Thus, we expect Manchester United to score 1.52 × 1.28 × 0.72 = 1.4 goals per game. Conversely, the average away team will score 1.2 goals per game, so adjusting by Manchester United's "defense weakness" (0.74) and Manchester City's "attack strength" (1.91) yield an expected 1.2 × 1.91 × 0.74 = 1.7 goals per game.

Table 3: Expected number of goals, and percentage chance of getting a particular score for the two teams, assuming a Poisson distribution

Team	Expected goals	0	1	2	3	4	5
Manchester Utd	1.4	25	35	24	11	4	1
Manchester C.	1.7	18	31	26	15	6	2

See Table 4 for assessed probabilities for the 2 matches.

```
liverpool <- c(0.09, 0.21, 0.26, 0.21, 0.13, 0.06)
tottenham <- c(0.46, 0.36, 0.14, 0.04, 0.01, 0.00)

manu <- c(0.25, 0.35, 0.24, 0.11, 0.04, 0.01)
manc <- c(0.18, 0.31, 0.26, 0.15, 0.06, 0.02)

# win_loss_draw matrix
# cols are liverpool rows are tott
wld_lt <- matrix(rep(NA,36),ncol=6)
# cols are manu rows are manc
wld_mm <- matrix(rep(NA,36),ncol=6)

for(i in 1:6)
  for(j in 1:i)
    wld_lt[i,j] <- liverpool[i]*tottenham[j]

for(i in 1:6)
```

```
  for(j in 1:i)
    wld_mm[i,j] <- manu[i]*manc[j]

# !! The values in the rows/cols are indexes, not scores !!
# e.g. (1, 1) is actually the P() for score (0, 0)
print(wld_lt, na.print = "")

##        [,1]   [,2]   [,3]   [,4]   [,5] [,6]
## [1,] 0.0414
## [2,] 0.0966 0.0756
## [3,] 0.1196 0.0936 0.0364
## [4,] 0.0966 0.0756 0.0294 0.0084
## [5,] 0.0598 0.0468 0.0182 0.0052 0.0013
## [6,] 0.0276 0.0216 0.0084 0.0024 0.0006    0

print(wld_mm, na.print = "")

##        [,1]   [,2]   [,3]   [,4]   [,5]  [,6]
## [1,] 0.0450
## [2,] 0.0630 0.1085
## [3,] 0.0432 0.0744 0.0624
## [4,] 0.0198 0.0341 0.0286 0.0165
## [5,] 0.0072 0.0124 0.0104 0.0060 0.0024
## [6,] 0.0018 0.0031 0.0026 0.0015 0.0006 2e-04

# replace na's with 0's to row-sum
# !! note values aren't actually 0 !!
wld_lt[is.na(wld_lt)] = 0
wld_mm[is.na(wld_mm)] = 0

# trace for probability of draws
prob_of_draw_lt <- sum(diag(wld_lt))
prob_of_draw_mm <- sum(diag(wld_mm))

# find sum of lower triangle, not including diagonal
prob_of_home_win_lt <- sum(rowSums(wld_lt * lower.tri(wld_lt, diag=FALSE)))
prob_of_home_win_mm <- sum(rowSums(wld_mm * lower.tri(wld_mm, diag=FALSE)))

# prob of home_loss = 1 - prob_win - prob_draw
prob_of_home_loss_lt <- 1 - prob_of_home_win_lt - prob_of_draw_lt
prob_of_home_loss_mm <- 1 - prob_of_home_win_mm - prob_of_draw_mm
```

Home	Away	Home win	Draw	Away win
Liverpool	Tottenham	0.70	0.16	0.13
Manchester Utd	Manchester C.	0.31	0.23	0.46

Table 4: The assessed probabilities for the 2 matches.