Lab 1: Probability Theory

W203: Statistics for Data Science

1. Meanwhile, at the Unfair Coin Factory...

You are given a bucket that contains 100 coins. 99 of these are fair coins, but one of them is a trick coin that always comes up heads. You select one coin from this bucket at random. Let T be the event that you select the trick coin. This means that $P(T) = 0.01$.

a. Suppose you flip the coin once and it comes up heads. Call this event H_1. If this event occurs, what is the conditional probability that you have the trick coin? In other words, what is $P(T|H_1)$?

b. Suppose instead that you flip the coin k times. Let H_k be the event that the coin comes up heads all k times. If you see this occur, what is the conditional probability that you have the trick coin? In other words, what is $P(T|H_k)$.

c. How many heads in a row would you need to observe in order for the conditional probability that you have the trick coin to be higher than 99%?

a)

Applying The Law of Total Probability

$P(H_1) = P(T) \cdot P(H_1|T) + P(!T) \cdot P(H_1|!T)$

$P(H_1) = 0.01 \cdot 1 + (1 - 0.01) \cdot 0.50$

$P(H_1) = .505$

$P(T) = P(T|H_1) \cdot P(H_1) + P(T|!H_1) \cdot P(!H_1)$

$0.01 = P(T|H_1) \cdot 0.505 + 0 \cdot (1 - 0.505)$

$P(T|H_1) = 0.0198 \approx 0.02$

Given 1 coin flip landing on heads, there is a 2% conditional probability that you have the trick coin.

b)

Redoing the above analysis to include # of flips (k) as a more explicitly stated variable.

Applying The Law of Total Probability

$$P(H_k) = P(T) \cdot P(H_k|T) + P(!T) \cdot P(H_k|!T)$$

$$P(H_k) = 0.01 \cdot 1 + (1 - 0.01) \cdot \frac{1}{2^k}$$

$$P(H_k) = 0.01 + \frac{0.99}{2^k}$$

$$P(T) = P(T|H_k) \cdot P(H_k) + P(T|!H_k) \cdot P(!H_k)$$

$$0.01 = P(T|H_k) \cdot (0.01 + \frac{0.99}{2^k}) + (1 - (0.01 + \frac{0.99}{2^k})) \cdot 0$$

$$P(T|H_k) = \frac{0.01}{(0.01+\frac{0.99}{2^k})}$$

If k flips of a single coin all land on heads, conditional probability is $\frac{0.01}{(0.01+\frac{0.99}{2^k})}$ **that you have the trick coin.**

c)

As $P(T|H_k)$ is an increasing function as k increases, to determine the lowest number of consecutive heads results to demonstrate a conditional probability of 99%, solve $P(T|H_k) = 0.99$ for roundup(k).

$$P(T|H_k) = \frac{0.01}{(0.01+\frac{0.99}{2^k})}$$

$$0.99 = \frac{0.01}{(0.01+\frac{0.99}{2^k})}$$

$$0.99 \cdot (0.01 + \frac{0.99}{2^k}) = 0.01$$

$$0.01 + \frac{0.99}{2^k} = \frac{0.01}{0.99}$$

$$\frac{0.99}{2^k} = \frac{0.01}{0.99} - 0.01$$

$$\frac{0.99}{2^k} = 0.000101010101010102$$

$$0.99 = 2^k \cdot 0.000101010101010102$$

$$2^k = \frac{0.99}{0.000101010101010102}$$

$$k \cdot ln(2) = ln(\frac{0.99}{0.000101010101010102})$$

$$k = 13.3$$

Roundup(k = 13.3) = 14

14 heads in a row are needed to observe that the conditional probability you have the trick coin is higher than 99%

2. Wise Investments

You invest in two startup companies focused on data science. Thanks to your growing expertise in this area, each company will reach unicorn status (valued at $1 billion) with probability 3/4, independent of the other company. Let random variable X be the total number of companies that reach unicorn status. X can take on the values 0, 1, and 2. Note: X is what we call a binomial random variable with parameters $n = 2$ and $p = 3/4$.

a. Give a complete expression for the probability mass function of X.

b. Give a complete expression for the cumulative probability function of X.

c. Compute $E(X)$.

d. Compute $var(X)$.

a)

Seeking b(x:n,p) for:

- x = 0, 1, 2
- n = 2
- p = 3/4

$$b(x:n,p) = f(x) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} & \text{x = 0,1,2} \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = \begin{cases} \binom{2}{0} \cdot (3/4)^0 \cdot (1-3/4)^{2-0} & \text{x} = 0 \\ \binom{2}{1} \cdot (3/4)^1 \cdot (1-3/4)^{2-1} & \text{x} = 1 \\ \binom{2}{2} \cdot (3/4)^2 \cdot (1-3/4)^{2-2} & \text{x} = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = \begin{cases} 1 \cdot (3/4)^0 \cdot (1-3/4)^{2-0} & \text{x} = 0 \\ 2 \cdot (3/4)^1 \cdot (1-3/4)^{2-1} & \text{x} = 1 \\ 1 \cdot (3/4)^2 \cdot (1-3/4)^{2-2} & \text{x} = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = \begin{cases} 1/16 & \text{x} = 0 \\ 6/16 = 3/8 & \text{x} = 1 \\ 9/16 & \text{x} = 2 \\ 0 & \text{otherwise} \end{cases}$$

b)

Seeking B(x:n,p) for:

- x = 0, 1, 2
- n = 2
- p = 3/4

$B(x:n,p) = P(X \leq x) = \sum_{y=0}^{x} b(x:n,p) x = 0,1,2$

$$P(X \leq x) = \begin{cases} b(0:2,3/4) & \text{x} = 0 \\ b(0:2,3/4) + b(1:2,3/4) & \text{x} = 0 \\ b(0:2,3/4) + b(1:2,3/4) + b(2:2,3/4) & \text{x} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X \leq x) = \begin{cases} 1/16 & x = 0 \\ 1/16 + 6/16 & 0 < x \leq 1 \\ 1/16 + 6/16 + 9/16 & 1 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X \leq x) = \begin{cases} 1/16 & x = 0 \\ 7/16 & 0 < x \leq 1 \\ 1 & 1 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

c)

Seeking E(X)

E(X) = $n \cdot p$ for binomial distribution

E(X) = $2 \cdot 3/4$

E(X) = 6/4

E(X) = 1 1/2 company w/ unicorn status

d)

Seeking Var(X)

Var(X) = $n \cdot p \cdot q$ for binomial distribution where q = 1-p

Var(X) = $2 \cdot 3/4 \cdot (1 - 3/4)$

Var(X) = 6/16 = 3/8

3. A Really Bad Darts Player

Let X and Y be independent uniform random variables on the interval $[-1, 1]$. Let D be a random variable that indicates if (X, Y) falls within the unit circle centered at the origin. We can define D as follows:

$$D = \begin{cases} 1, & X^2 + Y^2 < 1 \\ 0, & otherwise \end{cases}$$

Note that D is a Bernoulli variable.

a. Compute the expectation $E(D)$. Hint: it might help to remember why we use area diagrams to represent probabilites.

b. Compute the standard deviation of D.

c. Write an R function to compute the value of D, given a value for X and a value for Y. Use R to simulate a draw for X and a draw for Y, then compute the value of D.

d. Use R to simulate the previous experiment 1000 times, resulting in 1000 samples for D. Compute the sample mean and sample standard deviation of your result, and compare them to the true values in parts a. and b.

a)

E(D) = np where p represents the probability of X & Y being within the circle defined by $X^2 + Y^2 < 1. n = 1$

E(D) = $\frac{\text{Area of circle } X^2, Y^2 \text{per formula } \pi \cdot r^2}{\text{Area of square X by Y per formula X} \cdot Y}$

E(D) = $\frac{\pi \cdot 1^2}{2 \cdot 2}$

E(D) = $\frac{\pi}{4} \approx 0.785$

b)

$\sigma_D = \sqrt{n \cdot p \cdot q}$ where $n = 1, p = \frac{\pi}{4}, q = 1 - p$

$\sigma_D = \sqrt{1 \cdot \frac{\pi}{4} \cdot \frac{4-\pi}{4}}$

$\sigma_D = \frac{1}{4} \cdot \sqrt{4 \cdot \pi - \pi^2} \approx 0.411$

c)

In [1]:
```
D_calc <- function(X, Y) {
  if((X^2 + Y^2) < 1){return (1)}
  else {return (0)}
}

X <- runif(1,-1,1)
Y <- runif(1,-1,1)

D_calc(X,Y)
```

1

d)

In [1]:
```
D_calc <- function(X, Y) {
  if((X^2 + Y^2) < 1){return (1)}
  else {return (0)}
}

D_vector <- vector(mode="numeric", length=0)

for (i in 1:1000){
X <- runif(1,-1,1)
Y <- runif(1,-1,1)
D_vector <- append(D_vector, D_calc(X,Y))
}

cat("Sample mean of D: ", mean(D_vector), "  This is similar to true mean value (0.785)
 as calculated in part a), \n although there is variability each time the script is run!
\n\n")
cat("Sample standard deviation of D: ", sd(D_vector), "  This is similar to true standar
d deviation value \n (0.411) as calculated in part b), although there is variability eac
h time the script is run!\n")
cat("")
```

```
Sample mean of D:  0.796   This is similar to true mean value (0.785) as calculated in
part a),
 although there is variability each time the script is run!

Sample standard deviation of D:  0.4031706   This is similar to true standard deviation
value
 (0.411) as calculated in part b), although there is variability each time the script i
s run!
```

4. Relating Min and Max

Continuous random variables X and Y have a joint distribution with probability density function,

$$f(x,y) = \begin{cases} 2, & 0 < y < x < 1 \\ 0, & otherwise. \end{cases}$$

You may wonder where you would find such a distribution. In fact, if A_1 and A_2 are independent random variables uniformly distributed on $[0,1]$, and you define $X = max(A_1, A_2)$, $Y = min(A_1, A_2)$, then X and Y will have exactly the joint distribution defined above.

a. Draw a graph of the region for which X and Y have positive probability density.

b. Derive the marginal probability density function of X, $f_X(x)$. Make sure you write down a complete expression.

c. Derive the unconditional expectation of X.

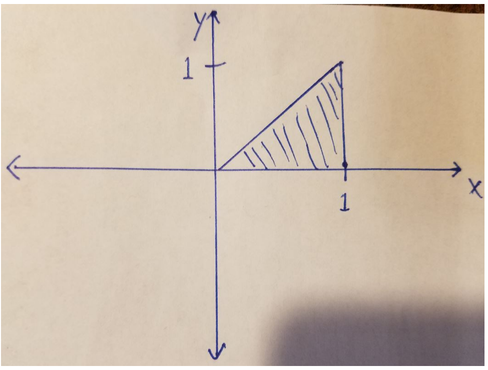
d. Derive the conditional probability density function of Y, conditional on X, $f_{Y|X}(y|x)$

e. Derive the conditional expectation of Y, conditional on X, $E(Y|X)$.

f. Derive $E(XY)$. Hint 1: Use the law of iterated expectations. Hint 2: If you take an expectation conditional on X, X is just a constant inside the expectation. This means that $E(XY|X) = XE(Y|X)$.

g. Using the previous parts, derive $cov(X, Y)$

a)



b)

$f_X(x) = \int_{-\infty}^{\infty} f(x,y) \cdot dy$ for $-\infty < x < \infty$

$f_X(x) = \int_{y=-\infty}^{0} f(x,y) \cdot dy + \int_{y=0}^{1} f(x,y) \cdot dy + \int_{y=1}^{\infty} f(x,y) \cdot dy$ all for $-\infty < x < \infty$

Note the $\int_{y=-\infty}^{0}$ & $\int_{y=1}^{\infty}$ terms reduce to 0 as f(x,y) is 0 at these y values leaving an integral of 0. Similarly, the range of x can reduce to 0 < x < 1 as other values result in f(x,y) = 0. This results in a reduced expression of:

$f_X(x) = \int_{0}^{x} f(x,y) \cdot dy$ for $0 < x < 1$

$f_X(x) = 2x$ for $0 < x < 1$

c)

$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$

Note that as stated in b), regions $-\infty < x < 0$ and $1 < x < \infty$ both will resolve to 0 for expected value as $f_X(x)$ in those regions is 0. Therefore, they will be dropped from the range .

$E(X) = \int_0^1 x \cdot 2xdx \text{ for } 0 < x < 1$

$E(X) = \frac{2x^3}{3}|_0^1 \text{ for } 0 < x < 1$

$E(X) = 2/3$

d)

$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$

$f_{Y|X}(y|x) = \frac{2}{2x} \text{ for } 0 < x < 1$

$f_{Y|X}(y|x) = \frac{1}{x} \text{ for } 0 < x < 1$

e)

$E(Y|X) = \int_y y \cdot f_{Y|X}(y|x)dy$

$E(Y|X) = \int_y y \cdot \frac{1}{x}dy \text{ for } 0 < x < 1$

$E(Y|X) = \frac{1}{x} \cdot \frac{y^2}{2}|_0^x \text{ for } 0 < x < 1$

$E(Y|X) = \frac{1}{x} \cdot \frac{x^2}{2} \text{ for } 0 < x < 1$

$E(Y|X) = \frac{1}{2}$

f)

$E(XY) = E(E(XY)|X)$

$E(XY) = E(X \cdot E(Y|X))$

$E(XY) = E(X \cdot E(Y|X))$

$E(XY) = E(X \cdot \frac{X}{2})$ for $0 < x < 1$

$E(XY) = \frac{1}{2} \cdot E(X^2)$ for $0 < x < 1$

Using relationship that $E(g(x)) = \int_x g(x) \cdot f(x)dx$ where $g(x) = X^2$

$E(XY) = \frac{1}{2} \int_x X^2 \cdot 2Xdx$ for $0 < x < 1$

$E(XY) = \frac{1}{2} \cdot [\frac{2X^4}{4}]_0^1$

$E(XY) = \frac{1}{4}$

g)

$Cov(X,Y) = E(XY) - E(X) \cdot E(Y)$

$Cov(X,Y) = E(XY) - E(X) \cdot E(Y|X) \cdot E(X)$

$Cov(X,Y) = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{2}{3}$

$Cov(X,Y) = \frac{1}{4} - \frac{2}{9}$

$Cov(X,Y) = \frac{1}{36}$