Lab 2: 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. To see if the coin you have is the trick coin, you flip it 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 $P(T|H_k)$.

b. 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%?

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 $var(X)$.

3. 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, $f_X(x)$.
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: 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 $cov(X, Y)$

4. Circles, Random Samples, and the Central Limit Theorem

Let $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ be independent random samples from a uniform distribution on $[-1, 1]$. Let D_i be a random variable that indicates if (X_i, Y_i) falls within the unit circle centered at the origin. We can define D_i as follows:

$$D_i = \begin{cases} 1, & X_i^2 + Y_i^2 < 1 \\ 0, & otherwise \end{cases}$$

Each D_i is a Bernoulli variable. Furthermore, all D_i are independent and identically distributed.

a. Compute the expectation of each indicator variable, $E(D_i)$. Hint: your answer should involve a Greek letter.

b. Compute the standard deviation of each D_i.

c. Let $\bar{D}$ be the sample average of the D_i. Compute the standard error of $\bar{D}$.

d. Now let n=100. Using the Central Limit Theorem, compute the probability that $\bar{D}$ is larger than 3/4. Make sure you explain how the Central Limit Theorem helps you get your answer.

e. Now let $n = 100$. Use R to simulate a draw for $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$. Calculate the resulting values for $D_1, D_2, ...D_n$. What is the resulting value for the statistic $\bar{D}$? How does it compare to your answer for part a?

f. Now use R to replicate the previous experiment 10,000 times, generating a sample average of the D_i each time. Plot a histogram of the sample averages.

g. Compute the standard deviation of your sample averages to see if it's close to the value you expect from part c.

h. Compute the fraction of your sample averages that are larger that 3/4 to see if it's close to the value you expect from part d.

W203 Lab 2: Probability Theory

Shan He

1\. a. $P(T|H_k) = P(T \cap H_k) / P(H_k) = P(T) \cdot P(H_k|T) / P(H_k)$

Since we know that $P(T) = 0.01$ and $P(H_k|T) = 1$

we can rewrite $P(T|H_k) = \dfrac{0.01 \cdot 1}{P(H_k)}$.

Moreover, using Law of Total Probability:

$$P(H_k) = P(H_k|T) \cdot P(T) + P(H_k|!T) \cdot P(!T)$$
$$= 1 \cdot 0.01 + (0.5)^k \cdot 0.99$$
$$= 0.01 + (0.5)^k \cdot 0.99$$

So $P(T|H_k) = \boxed{\dfrac{0.01}{0.01 + (0.5)^k \cdot 0.99}}$

b. $P(T|H_k) > 0.99 \rightarrow \dfrac{0.01}{0.01+(0.5)^k \cdot 0.99} > 0.99$

$$0.01 > 0.99(0.01 + (0.5)^k 0.99)$$

$$0.01 + (0.5)^k \cdot 0.99 < \frac{1}{99}$$
$$(0.5)^k < \left(\frac{1}{99} - \frac{1}{100}\right) / 0.99$$
$$(0.5)^k < \frac{1}{9801}$$
$$k > \log_{0.5}\left(\frac{1}{9801}\right) = 13.26$$

So, you'll need to observe 14 heads in a row.

2. a

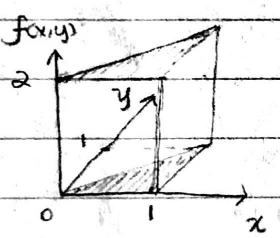
$$b(x;2,\tfrac{3}{4}) = \begin{cases} \binom{2}{x}\left(\frac{3}{4}\right)^x\left(\frac{1}{4}\right)^{2-x} & x=0,1,2 \\ 0 & \text{otherwise} \end{cases}$$

b. $B(x;2,\frac{3}{4}) = \sum_{y=0}^{x} b(y;2,\frac{3}{4}) \quad x = 0,1,2$

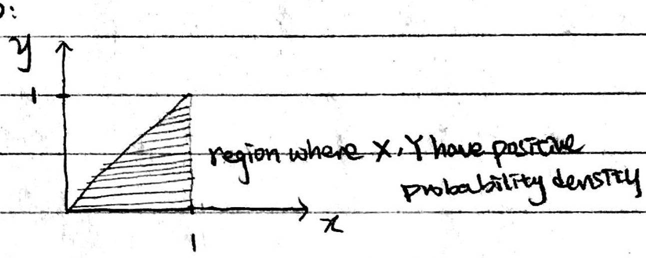
c. $E(X) = n \cdot p = 2 \cdot \frac{3}{4} = \boxed{1.5}$

d. $Var(X) = np(1-p) = 2 \cdot \frac{3}{4} \cdot \frac{1}{4} = \boxed{0.375}$

3. a) 3D:



2D:



b)

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^x 2\,dy = 2y\Big|_0^x = 2x$$

c) $$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx = \int_0^1 x \cdot 2x\,dx = \frac{2}{3}x^3\Big|_0^1 = \boxed{\frac{2}{3}}$$

d) $$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x} \quad \text{for} \quad y \in (0,x)$$

e) $$E(Y|X) = \int_0^x y \cdot f_{Y|X}(y|x)\,dy = \int_0^x y \cdot \frac{1}{x}\,dy = \frac{1}{x} \cdot y^2/2\Big|_0^x = \frac{x}{2}$$

f) $$E(XY) = E(E(XY|X)) = E(X\,E(Y|X)) = E\left(\frac{x^2}{2}\right) = \int_{-\infty}^{\infty} \frac{x^2}{2} \cdot f_X(x)\,dx$$

$$= \int_0^1 \frac{x^2}{2} \cdot 2x\,dx$$

$$= \frac{x^4}{4}\Big|_0^1 = \frac{1}{4}$$

g) $$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_y^1 2\,dx = 2-2y$$

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy = \int_0^1 2y - 2y^2\,dy = y^2 - \frac{2}{3}y^3\Big|_0^1 = \frac{1}{3}$$

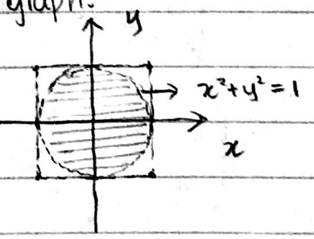
$$Cov(X,Y) = E(XY) - E(X)\cdot E(Y)$$
$$= \frac{1}{4} - \frac{2}{3}\cdot\frac{1}{3} = \frac{1}{36}$$

4. a. $E(D_i) = 1 \cdot P[X_i^2 + Y_i^2 < 1]$

$P[X_i^2+Y_i^2<1]$ for $X_i, Y_i \in [-1,1]$. can be caluclated as following:
Since X_i and Y_i are uniformly distributed between $[-1,1]$, the joint pdf $f(X,Y)$ has a evenly distributed density over the 2x2 area where $X \in [-1,1]$ and $Y \in [-1,1]$.
And writing $x^2+Y^2=1$ gives us a circle centered a $(0,0)$ with a radius of 1.
Since we know that $f(X,Y)$ has the same density over the 2x2 area, $P[(X_i^2+Y_i^2)<1]$ can be computed as:

$$P = \frac{\text{Area}(\text{circle}_{x^2+y^2=1})}{\text{Area}_{x\in[-1,1],\, Y\in[-1,1]}} = \frac{\pi}{2\times 2} = \boxed{\frac{\pi}{4}}$$

as shown in the following graph:



b. Now we know $D \sim Ber(\frac{\pi}{4})$, which follows a binomial distribution with $n=1$.

$$\sigma^2 = \frac{\pi}{4}(1-\frac{\pi}{4}) = \frac{\pi}{4} - \frac{\pi^2}{16}$$
$$\hookrightarrow \sigma = \sqrt{\frac{\pi}{4} - \frac{\pi^2}{16}}$$

c. $\sigma_{\bar{D}} = \sqrt{V(\sigma_{\bar{D}})} = \sigma/\sqrt{n} = \sqrt{\frac{\pi}{4} - \frac{\pi^2}{16}}/\sqrt{n}$

(since $n > 30$)

d. Using Central Limit Theorem, we know that $\bar{D}$ follows a normal distribution with $\mu_{\bar{D}} = E(D_i)$, $\sigma_{\bar{D}}^2 = \sigma^2/n$

$$P(\bar{D} > 3/4) \approx P\left(Z > \frac{3/4 - \pi/4}{\sqrt{\pi/4 - \pi^2/16} / \sqrt{100}}\right) = P\left(Z > \frac{-0.035}{0.041}\right)$$

$$= 1 - \Phi(-0.862)$$

$$= 0.806$$

w203_lab2_q4_SH

Shan He

10/19/2017

4e

1. Create a function that draws n of X_i, Y_i, and D_i

```
set.seed(15) #set seed for reproducible results

f <- function(n) {

  X <- runif(n,-1,1)
  Y <- runif(n,-1,1)
  D = 0

  for (i in c(1:n)){
  D[i] = ifelse( (X[i])^2 + (Y[i])^2 < 1, 1, 0)
  }

  return(D)
}
```

2. Draw D_i's for 100 X_i and Y_i

```
D_100 = f(100)
```

3. Compute $\overline{D}$

```
#sample mean
mean(D_100)
```

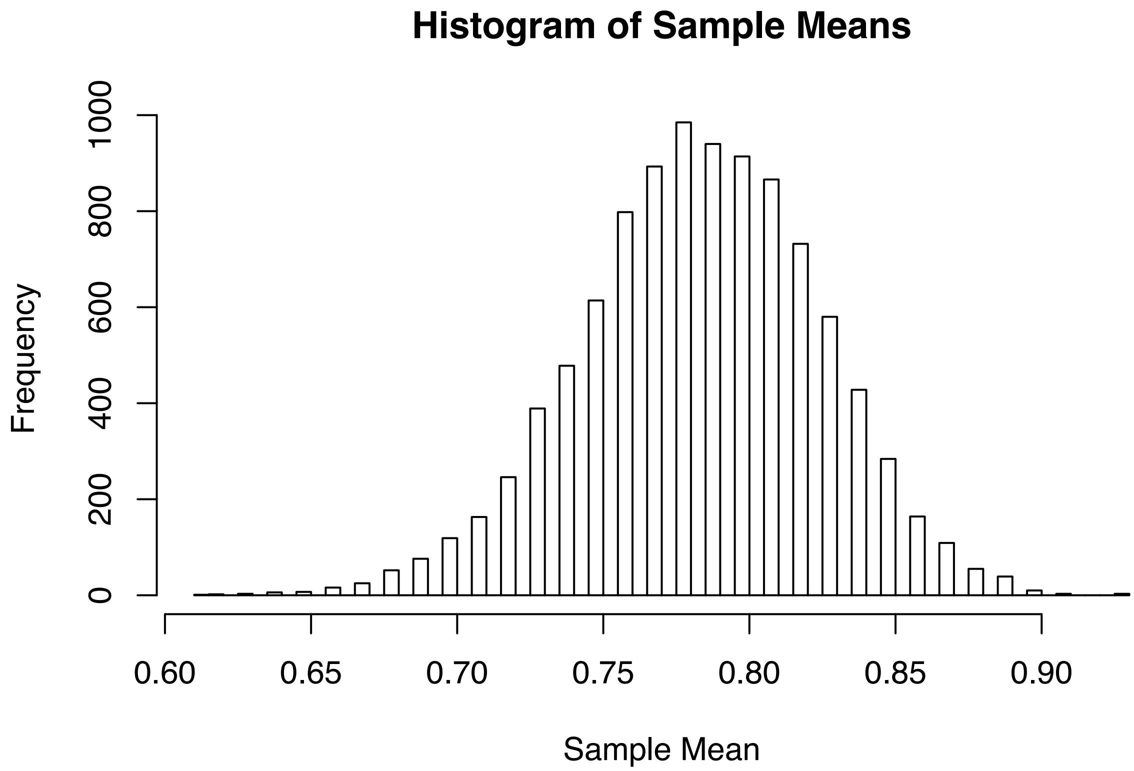
```
## [1] 0.78
```

The mean of Di's from a sample of 100 X_i's and Y_i's is 0.78, which is close to the $E(D_i)$, $\frac{\pi}{4}$ or 0.79 as calculated in part a.

4f

1. Replicate Experiments and Plot Sample Means

```
draws <- replicate(10000, mean(f(100)))
hist(draws, breaks = 50, xlab = "Sample Mean", main = "Histogram of Sample Means")
```



4g

Standard Deviation of Sample Means, or Standard Error of $\overline{D}$

```
sd(draws)
```

```
## [1] 0.04111142
```

With $n = 100$, from part c, we'd expect the standard error to be 0.041 which is very close to what we have here.

4h

Compute Fraction of $\overline{D}$ that are larger than $\frac{3}{4}$

```
sum(draws > 3/4)/10000
```

```
## [1] 0.7803
```

The value calculated from part d is 0.806, which is close to the simulated result