Lab 2: Probability Theory

Joanna Yu (Fall 2018 Tuesday 4pm)

10/15/2018

1a) What is $P(T|H_k)$.

$$\begin{aligned} P(T|H_k) &= \frac{P(H_k|T)P(T)}{P(H_k|T)P(T) + P(H_k|\bar{T})P(\bar{T})} \\ &= \frac{1 * 0.01}{1 * 0.01 + (\frac{1}{2})^k(0.99)} \\ &= \frac{1}{1 + \frac{99}{2^k}} \end{aligned}$$

1b) 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%?

$$\begin{aligned} P(T|H_k) = \frac{1}{1 + \frac{99}{2^k}} &> 0.99 \\ \frac{1}{0.99} &> 1 + \frac{99}{2^k} \\ \frac{1}{0.99} - 1 &> \frac{99}{2^k} \\ \frac{1}{99} &> \frac{99}{2^k} \\ 2^k &> 99^2 \implies k \geq 14 \end{aligned}$$

2a) Give a complete expression for the probability mass function of X.

$$P(X = 0) = \binom{2}{0}(\frac{3}{4})^0(\frac{1}{4})^2 = \frac{2!}{0!2!}(\frac{3}{4})^0(\frac{1}{4})^2 = \frac{1}{16}$$

$$P(X = 1) = \binom{2}{1}(\frac{3}{4})^1(\frac{1}{4})^1 = \frac{2!}{1!1!}(\frac{3}{4})^1(\frac{1}{4})^1 = \frac{3}{8}$$

$$P(X = 2) = \binom{2}{2}(\frac{3}{4})^2(\frac{1}{4})^0 = \frac{2!}{2!0!}(\frac{3}{4})^2(\frac{1}{4})^0 = \frac{9}{16}$$

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

2b) Give a complete expression for the cumulative probability function of X.

$$F(X) = \begin{cases} X < 0 & 0 \\ 0 \leq X < 1 & 1/16 \\ 1 \leq X < 2 & 7/16 \\ X \geq 2 & 1 \end{cases}$$

2c) Compute $E(X)$.

$$E(X) = np = 2 * \frac{3}{4} = \frac{3}{2}$$

2d) Computer $var(X)$.

$$Var(X) = np(1-p) = 2 * \frac{3}{4} * \frac{1}{4} = \frac{3}{8}$$

3a) Draw a graph of the region for which X and Y have positive probability density. $\ \ \ \ \ \ \ \ $

3b) Derive the marginal probability density function of X, $f_X(x)$

$$f_X(x) = \int_0^x f(xy)dy = \int_0^x 2dy = 2y\Big|_0^x = 2x$$

3c) Derive the unconditional expectation of X.

$$E(X) = \int_0^1 xf(x) = \int_0^1 x2xdx = \frac{2}{3}x^3\Big|_0^1 = \frac{2}{3}$$

3d) Derive the conditional probability density function of Y, conditional on X, $f_{Y|X}(y|x)$

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x,y)}{f(x)} \\ &= \frac{2}{2x} \\ &= \frac{1}{x} \end{aligned}$$

3e) Derive the conditional expectation of Y, conditional on X, E(Y|X).

$$\begin{aligned} E(Y|X) &= \int_0^x yf(y|x)dy \\ &= \int_0^x y\frac{1}{x}dy \\ &= \frac{1}{x} * \frac{1}{2}y^2\Big|_0^x \\ &= \frac{x}{2} \end{aligned}$$

3f) Derive E(XY)

$$\begin{aligned}
E(XY) &= E(E(XY|X)) \\
&= E(XE(Y|X)) \\
&= E(X(X/2)) \\
&= \frac{1}{2}E(X^2) \\
&= \frac{1}{2}\int_0^1 x^2 f(x)dx \\
&= \frac{1}{2}\int_0^1 x^2(2x)dx \\
&= \frac{1}{2}(\frac{2}{4}x^4)\Big|_0^1 \\
&= \frac{1}{4}
\end{aligned}$$

3g) Using the previous parts, derive $cov(X,Y)$

$$f(y) = \int_y^1 f(xy)dx = \int_y^1 2dx = 2x\Big|_y^1 = 2 - 2y$$

$$\begin{aligned}
cov(X,Y) &= E(XY) - E(X)E(Y) \\
&= \frac{1}{4} - \frac{2}{3}\int_0^1 yf(y)dy \\
&= \frac{1}{4} - \frac{2}{3}\int_0^1 y(2-2y)dy \\
&= \frac{1}{4} - \frac{2}{3}\int_0^1 (2y-2y^2)dy \\
&= \frac{1}{4} - \frac{2}{3}(y^2 - \frac{2}{3}y^3)\Big|_0^1 \\
&= \frac{1}{4} - \frac{2}{3} * \frac{1}{3} \\
&= \frac{1}{4} - \frac{2}{9} \\
&= \frac{1}{36}
\end{aligned}$$

4a) Compute the expectation of each indicator variable, E(Di).

$$\begin{aligned}
E(Di) &= 1 * P(Di = 1) + 0 * P(Di = 0) \\
&= P(Di = 1) \\
&= P(x^2 + y^2 < 1) \\
&= \frac{\pi(1)^2}{(1-(-1))^2} \\
&= \frac{\pi}{4}
\end{aligned}$$

4b) Compute the standard deviation of each Di.

$$
\begin{aligned}
SD(Di) = \sqrt{Var(Di)} &= \sqrt{E(Di^2) - E(Di)^2} \\
&= \sqrt{1^2 * P(Di = 1) + 0^2 * P(Di = 0) - (\frac{\pi}{4})^2} \\
&= \sqrt{P(Di = 1) - (\frac{\pi}{4})^2} \\
&= \sqrt{\frac{\pi}{4} - (\frac{\pi}{4})^2} \\
&= \frac{\sqrt{4\pi - \pi^2}}{4}
\end{aligned}
$$

4c) Let $\bar{D}$ be the sample average of the Di. Compute the standard error of $\bar{D}$. This should be a function of sample size n.

$$
\begin{aligned}
SE(\bar{D}) = \sqrt{Var(\bar{D})} &= \sqrt{Var(\frac{D_1 + D_2 + ... + D_n}{n})} \\
&= \sqrt{\frac{1}{n^2} Var(D_1 + D_2 + ...D_n)} \\
&= \sqrt{\frac{1}{n^2} Var(D_i) * n} \\
&= \sqrt{\frac{Var(D_i)}{n}} \\
&= \frac{SD(D_i)}{\sqrt{n}} \\
&= \frac{\sqrt{4\pi - \pi^2}}{4\sqrt{n}}
\end{aligned}
$$

4d) 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.

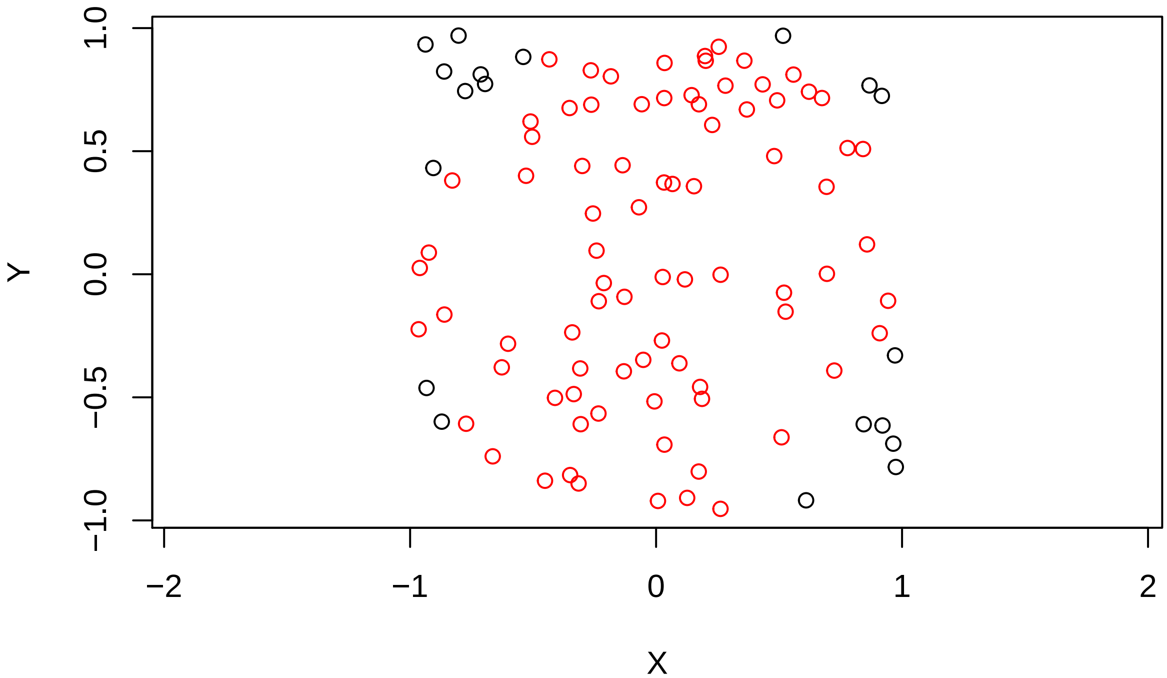
Since n is sufficiently large and the sample is not very skewed, the Central Limit Theorem tells us that the sampling distribution of the mean will approach normal. Therefore we can calculate the probability that $\bar{D}$ is larger than 3/4 by standardizing $\bar{D}$ and using the normal table.

$$
\begin{aligned}
E(\bar{D}) &= \pi/4 \approx 0.7854 \\
SE(\bar{D}) &= \frac{\sqrt{4\pi - \pi^2}}{4\sqrt{100}} \approx 0.0411 \\
P(\bar{D} > 0.75) \implies P(Z > \frac{0.75 - 0.7854}{0.0411}) &= 1 - (Z > -0.86) = 1 - 0.1949 = 0.8051
\end{aligned}
$$

4e) Now let n=100. Using R to stimulate a draw for $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$. Calculate the resulting values for $D_1, D_2, \ldots, D_n$

```r
set.seed(2529)
n = 100
X = runif(n, min=-1, max=1)
Y = runif(n, min=-1, max=1)
```

```r
D=ifelse((X^2+Y^2)<1, 1, 0)
plot(X, Y, col=D+1, asp=1)
```



4f) What value do you get for the sample average, $\bar{D}$? How does it compare to your answer for part a?

```r
mean(D)
```

```
## [1] 0.81
```

```r
pi/4
```

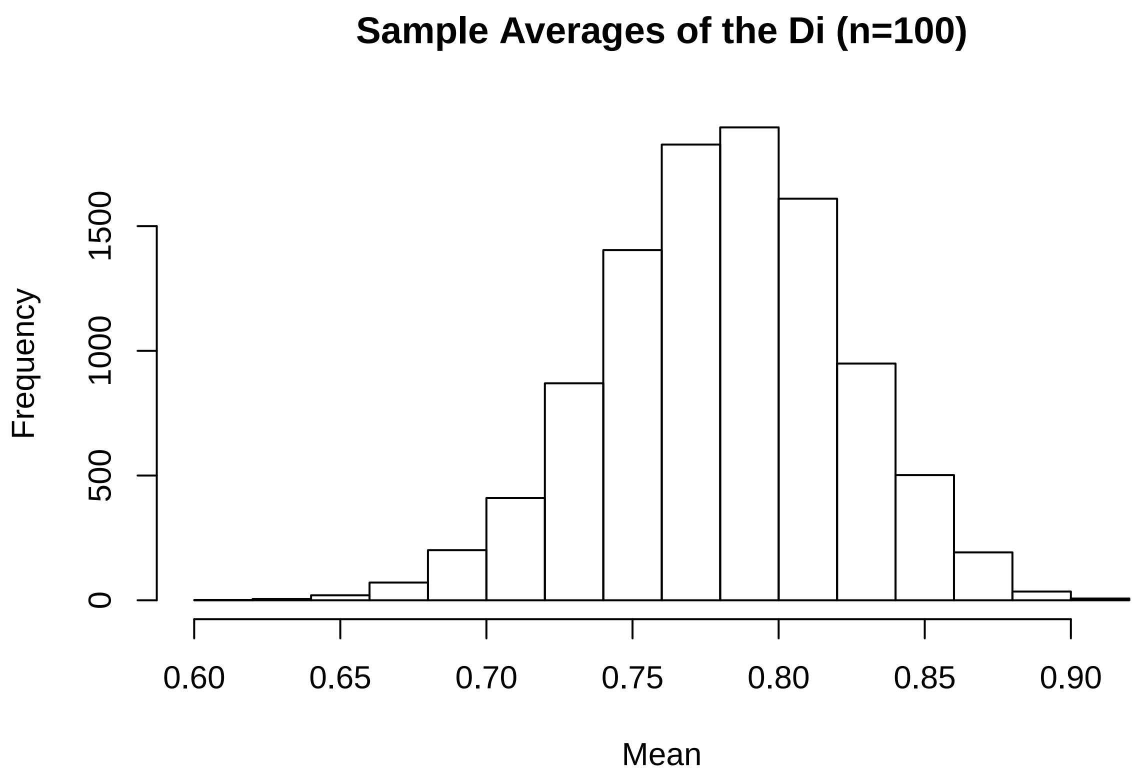
```
## [1] 0.7853982
```

The mean of $\bar{D}$ is 0.81, while my answer in part a is $\pi/4$, which is about 0.7854.

4g) Now use R to replicate the previous experiment 10,000 times, generating a sample average of the Di each time. Plot a histogram of the sample averages.

```r
execute_study = function(n) {
  X = runif(n, min=-1, max=1)
  Y = runif(n, min=-1, max=1)
  D=ifelse((X^2+Y^2)<1, 1, 0)
  mean(D)
}

result = replicate(10000,execute_study(100))
hist(result, main = "Sample Averages of the Di (n=100)", xlab = "Mean")
```



4h) Compute the standard deviation of your sample averages to see if it's close to the value you expect from part c.

```
sd(result)
```

```
## [1] 0.04113743
```

```
sqrt(4*pi-pi^2)/(4*sqrt(100))
```

```
## [1] 0.04105458
```

The standard deviation of my sample averages is 0.0411. My answer in part c is about 0.0411.

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

```
sum(ifelse(result>0.75, 1, 0))/length(result)
```

```
## [1] 0.7769
```

77.69% of my sample averages are larger than 3/4. It is close to the value I have for part d, 80.51%.