Introduction to Data Science group assignment 1

Max Callenmark Kasper Lindkvist Eskil Worm Forss
Rami Abou Zahra Oskar Ådahl

Division Of Work

Task 1: Eskil Worm Forss
Task 2: Kasper Lindkvist
Task 3: Max Callenmark
Task 4: Oskar Ådahl
Task 5: Rami Abou Zahra

1)

$$A^c \cap B^c = (A \cup B)^c \quad \text{By De Morgans law}$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \quad \text{By our assumption}$$

$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B) = 1 - \mathbb{P}(A) - \mathbb{P}(B) + P(A \cap B) =$$

$$= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A) \cdot \mathbb{P}(B) = (1 - \mathbb{P}(A)) \cdot (1 - \mathbb{P}(B)) = \mathbb{P}(A^c) \cdot \mathbb{P}(B^c) \quad \square$$

2)

a)

We can let $B =$ "the number of children with brown hair" be a random variable following a binomial distribution. $B \sim Bin(3, 1/4)$, with $n = 3$ being the number of trials or in this case children, and $p = 1/4$ being the probability of success or in this case that a child has brown hair.

Now we know that $B \geq 1$ and then what we want is

$$P(B \geq 2 | B \geq 1) = \frac{P(B \geq 2 \cap B \geq 1)}{P(B \geq 1)} = \frac{P(B \geq 2)}{P(B \geq 1)} = \frac{P(B = 3) + P(B = 2)}{P(B = 3) + P(B = 2) + P(B = 1)} \tag{1}$$

Now given $B \sim Bin(3, 1/4)$, $P(B = k) = \binom{3}{k}(\frac{1}{4})^k(\frac{3}{4})^{3-k}$, so we get:

1. $P(B = 1) = \binom{3}{1}(\frac{1}{4})^1(\frac{3}{4})^2 = \frac{27}{64}$
2. $P(B = 2) = \binom{3}{2}(\frac{1}{4})^2(\frac{3}{4})^1 = \frac{9}{64}$
3. $P(B = 3) = \binom{3}{3}(\frac{1}{4})^3(\frac{3}{4})^0 = \frac{1}{64}$

Thus, $P(B \geq 2 | B \geq 1) = \frac{\frac{1}{64} + \frac{9}{64}}{\frac{1}{64} + \frac{9}{64} + \frac{27}{64}} = \frac{\frac{10}{64}}{\frac{37}{64}} = \frac{10}{37}$

Answer:

$$P(B \geq 2 | B \geq 1) = \frac{10}{37}$$

b)

Since there is independence between the children, with the knowledge that the oldest child has brown hair the uncertainty is now in that the younger children still each has 1/4 probability of having brown hair.

So we let B_2 = "the number of younger children with brown hair" be a new random variable with binomial distribution $B_2 \sim Bin(2, 1/4)$ instead, now with only two children.

So we want $P(B \geq 2 | \text{oldest child has brown hair})$, but because of the independence this is equal to at least one of the younger children having brown hair, so: $P(B_2 \geq 1) = 1 - P(B_2 = 0) = 1 - \binom{2}{0}(\frac{1}{4})^0(\frac{3}{4})^2 = 1 - \frac{9}{16} = \frac{7}{16}$

Answer:

$$P(B_2 \geq 1) = \frac{7}{16}$$

3)

Since the area of the unit disc is π and (X, Y) is uniformly distributed on the unit disc we have that the PDF of (X, Y) is given by:

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{for} \quad x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We want to determine the PDF and CDF of $R = \sqrt{X^2 + Y^2}$. We introduce the auxiliary random variable $V = \arctan\left(\dfrac{Y}{X}\right)$. By the transformation theorem, we have that:

$$f_{R,V}(r,v) = \begin{cases} f_{X,Y}(h_1(x), h_2(y)) \cdot |\mathbf{J}| \\ 0 \quad \text{otherwise} \end{cases}$$

Where h are the unique inverses and $\mathbf{J}$ is the Jacobian. We have that the range of $\arctan(\cdot)$ is $(-\pi/2, \pi/2)$. This means that we have a 2-to-1 mapping, since the points (X, Y) and $(-X, -Y)$ correspond to the same (R, V). By symmetry and since the $|\mathbf{J}| = r$ we get:

$$f_{R,V}(r,v) = \begin{cases} 2 \cdot \dfrac{1}{\pi} \cdot r \quad \text{for} \quad 0 < r < 1 \quad \text{and} \quad -\pi/2 < v < \pi/2 \\ 0 \quad \text{otherwise} \end{cases}$$

So to get f_R, we need to integrate out f_V:

$$f_R(r) = \int f_{R,V}(r,v)dv = \int_{-\pi/2}^{\pi/2} \frac{2r}{\pi} dv = 2r$$

Then the CDF is gotten by integrating our PDF for R:

$$F_R(r) = \int_0^r 2r dr = r^2$$

Answer:

$$f_R(r) = \begin{cases} 2r & 0 \leq r \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad F_R(r) = \begin{cases} 0 & r < 0 \\ r^2 & 0 \leq r \leq 1 \\ 1 & r > 1 \end{cases}$$

4)

The probability of getting the first head on the kth toss is the same as the probability of getting $k-1$ tails in a row and then a head. Since the tosses are independent, we find that $P(X=k) = \left(\frac{1}{2}\right)^{k-1} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^{k}$. Hence, the expected value of X is

$$E(X) = \sum_{k=1}^{\infty} kP(X=k) = \sum_{k=1}^{\infty} k\left(\frac{1}{2}\right)^{k}.$$

For any $|r|<1$, the value of the geometric sum $\sum_{k=0}^{\infty} r^k$ is $\frac{1}{1-r}$. Differentiating term-wise we get that, for any $|r|<1$,

$$\sum_{k=1}^{\infty} kr^{k-1} = \left[\frac{1}{1-r}\right]' = \frac{1}{(1-r)^2}.$$

In particular,

$$E(X) = \frac{1}{2}\frac{1}{\left(1-\frac{1}{2}\right)^2} = 2.$$

Answer:

$\mathrm{E}(X) = 2$

5. (a) Since all $X_i \sim \text{Be}(p)$ and independent, then $\mathbb{E}[\widehat{p}] = \frac{1}{n} \cdot np = p$.

All variables share distribution, and are all real-valued. We may rewrite the problem as showing $\mathbb{P}(p \in I_n) \geq 1 - \alpha$ as $1 - \mathbb{P}(p \in I_n) \leq \alpha$.
Now it becomes apparent how to use Hoeffdings inequality.
Since they are Bernoulli distributed, $a = b = 0$, this follows from

$$\mathbb{P}(X_i - \mathbb{E}[X_i] \in [a, b]) = 1, \quad X_i \in \{0, 1\}, \quad X_i - p = \begin{cases} 1 - p \\ -p \end{cases} \Rightarrow \quad a = 1 - p, b = -p$$

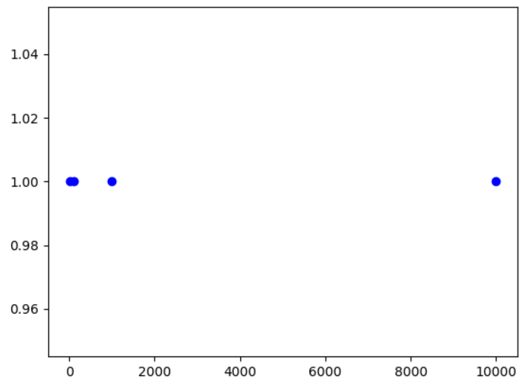
Then

$$\mathbb{P}(|\widehat{p} - p| \geq \varepsilon) \leq \exp\left\{\frac{-2n\varepsilon^2}{(b-a)^2}\right\} = \exp\left\{\frac{-2n\varepsilon^2}{(-p-1+b)^2}\right\} = \exp\left\{-2n\varepsilon^2\right\}$$

Using the provided ε_n, we get that $\mathbb{P}(|\widehat{p} - p| \geq \varepsilon_n) \leq 2\exp\left\{-2n\varepsilon^2\right\} = \alpha$, and by the remark in the beginning, this is the same as

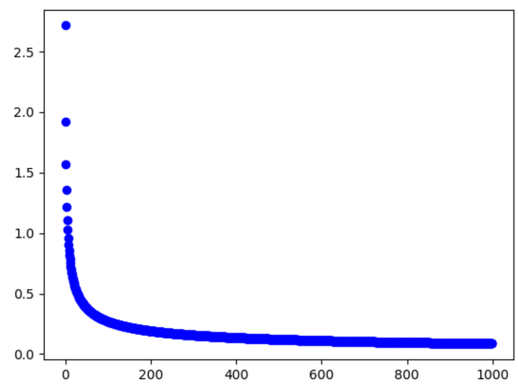
$$1 - \mathbb{P}(|\widehat{p} - p| \geq \varepsilon_n) \geq 1 - \alpha$$

Which is the same as $\mathbb{P}(p \notin I_n) \leq \alpha \Leftrightarrow \mathbb{P}(p \in I_n) \geq 1 - \alpha$



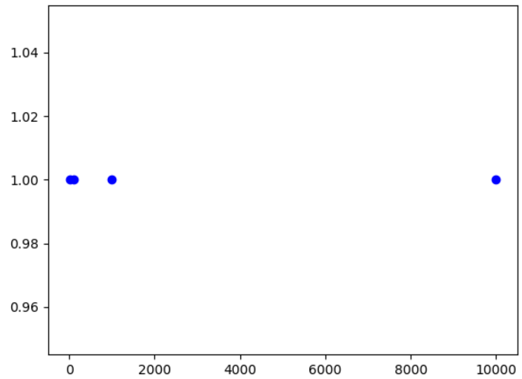
(b)

For all simulations, the confidence interval contained p



(c)

The length of the interval is $2 \cdot \varepsilon \to 0$



(d)

Since p was in all the intervals, the probability of our decision being correct is 1 for all n