Introduction to Data Science 1MS041 - Assignment 1

Holger Swartling & Nir Teyar

September 19, 2024

1 Contribution statement

The assignment as a whole was done by both group members. The group members did the assignment first individually and then checked answers with each other upon completion, thus resulting in approximately 50% contribution from each member.

Contents

1 Contribution statement **2**

2 Question 1 **4**
2.1 Question: 4
2.2 Answer: 4

3 Question 2 **4**
3.1 Part 1 4
3.1.1 Question 4
3.1.2 Answer: 5
3.2 Part 2 6
3.2.1 Question 6
3.2.2 Answer 6

4 Question 3 **6**
4.1 Question: 6
4.2 Answer: 6

5 Question 4: **8**
5.1 Question: 8
5.2 Answer: 8

6 Question 5: **9**
6.1 Part 1: 9
6.1.1 Question 9
6.1.2 Answer 9
6.2 Part 2 10
6.2.1 Question 10
6.2.2 Answer 10
6.3 Part 3 11
6.3.1 Question 11
6.3.2 Answer 11
6.4 Part 4 12
6.4.1 Question 12
6.4.2 Answer 12

2 Question 1

2.1 Question:

Suppose that A and B are independent events, show that A^c and B^c are independent.

2.2 Answer:

A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

This means that we must show that A^c and B^c are independant by showing that

$$P(A^c \cap B^c) = P(A^c)P(B^c)$$

We use the complement occurrence in order to describe the wanted outcome in a new light.

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

We also know that we can rewrite $P(A \cup B)$ as such:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Which can be used to further elaborate equation (1). It is also stated that A and B are independent which allows us to assume that $P(A \cap B) = P(A)P(B)$. With this knowledge input it all into one formula in order to complete the exercise:

$$P(A^c \cap B^c) = 1 - [P(A) + P(B) - P(A)P(B)]$$

By removing the brackets and simplifying we get the following expression:

$$P(A^c \cap B^c) = (1 - P(A))(1 - P(B)) = P(A^c)P(B^c) \tag{2}$$

Thus proving that A^c and B^c are independent events.

3 Question 2

The probability that a child has brown hair is $\frac{1}{4}$. Assume independence between children and assume there are three children.

3.1 Part 1

3.1.1 Question

If it is known that at least one child has brown hair, what is the probability that at least two children have brown hair?

3.1.2 Answer:

The probability of a child having brown hair is described as P(B) = $\frac{1}{4}$

X Describes the number of children with brown hair, hence the binomial distribution that follows is $n = 3$ and as previously mentioned $p = \frac{1}{4}$.

We start by calculating the probability of no child having brown hair:

$$P(\text{No child having brown hair}) = \left(\frac{3}{4}\right)^4 = \frac{27}{64}$$

Next we calculate one child having brown hair which is the complement occurrence:

$$P(X = 1) = 1 - \frac{27}{64} = \frac{37}{64}$$

Next we calculate for two children having brown hair using combinatorics and binomial coefficients:

$$P(X = 2) = \binom{3}{2} * \left(\frac{1}{4}\right)^2 * \left(\frac{3}{4}\right)$$

Finally we need to calculate the probability of all 3 having brown hair:

$$\left(\frac{1}{4}\right)^3 = \frac{1}{64}$$

In order to asses the probability of at least two children we will use addition as show below:

$$P(X \geq 2) = P(X = 2) + P(X = 3) \tag{3}$$

$$P(X \geq 2) = \frac{9}{64} + \frac{1}{64} = \frac{10}{64}$$

Because of the way the questions is posed we must use conditional probabilty in order to fully answer it. Conditional probability is given by the following formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

$$P(X \geq 2 \mid X \geq 1) = \frac{P(X \geq 2 \cap X \geq 1)}{P(X \geq 1)} = \frac{(P(X \geq 2)}{P(X \geq 1)} = \frac{\frac{10}{64}}{\frac{37}{64}} = \frac{10}{37}$$

Summary:

$$P(\text{At least two children have brown hair} \mid \text{At least one child has brown hair}) = \frac{10}{37}$$

3.2 Part 2

3.2.1 Question

If it is known that the oldest child has brown hair, what is the probability that at least two children have brown hair?

3.2.2 Answer

In this case we only have 2 remaining children as a specific child, also know as the oldest one, already has a determined hair colour. We will start by calculating the probability of none of the remaining children having brown hair:

$$\left(\frac{3}{4}\right)^2 = \frac{9}{16}$$

We will now use the complement occurrence in order to gain the probability of one of the two children having brown hair

$$1 - \frac{9}{16} = \frac{7}{16}$$

Summary: The probability is $\frac{7}{16}$

4 Question 3

4.1 Question:

Let (X, Y) be uniformly distributed on the unit disc, $\{(x, y) \in R^2 \mid x^2+y^2 \leq 1\}$. Set $R = \sqrt{X^2 + Y^2}$. What is the CDF and PDF of R?

4.2 Answer:

The total probability over the unit disc is 1. The density of the uniform distribution is to be described as such for $x^2 + y^2 \leq 1$, $r \leq 1$:

$$f_{XY}(x, y) = \frac{1}{\pi}$$

The CDF for R can then be described as $F_R(r)$. The probability that the radial distance R from origin is less than or equal to r can be described as such

$$F_R(r) = P(R \leq r)$$

The probability corresponds to the ratio of the area circle of the circle with radius r, and unit the unit circle. Hence we will make the following division:

$$\frac{\text{Area of circle with radius r}}{\text{Area of unit circle}} \tag{5}$$

$$\frac{\text{Area of circle with radius r}}{\text{Area of unit circle}} = \frac{\pi r^2}{\pi} = r^2$$

This applies for $0 \leq r \leq 1$.

We then acquire the PDF through differentiation of $F_R(r)$

$$f_R(r) = \frac{d}{dr} F_R(r) \tag{6}$$

$$f_R(r) = \frac{d}{dr} F_R(r) = \frac{d}{dr}(r^2) = 2r$$

This means that we have the following results:

CDF:

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

Good!

PDF

$$f_R(r) = \begin{cases} 2r, & \text{if } 0 \leq r \leq 1 \\ 0, & \text{otherwise } 0 \end{cases}$$

Summary: The CDF gives the answer r^2 within $0 \leq r \leq 1$, and PDF is $2r$ within $0 \leq r \leq 1$

5 Question 4:

5.1 Question:

A fair coin is tossed until a head appears. Let X be the number of tosses required. What is the expected value of X?

5.2 Answer:

In order to answer this question, one must understand that we are dealing with a geometric random variable.

X corresponds to the amount of tosses before getting a head. This means that we are looking for the expected value for this case $E[X]$. Since we are dealing with a fair coin, and the coin consists of two cases, heads (H) or tails (T) then it's safe to assume that $p = \frac{1}{2}$.

For a geometric variable one can calculate the expected value using the following simple formula.

$$E[X] = \frac{1}{p} \tag{7}$$

We use the formula as follows:

$$E[X] = \frac{1}{p} = \frac{1}{\frac{1}{2}} = \frac{2}{1} = 2$$

Summary: The expected value is 2 which means that you are expected to toss the coin twice before heads appear.

6 Question 5:

Let $X_1, \ldots, X_n$ be IID from Bernoulli(p).

6.1 Part 1:

6.1.1 Question

Let $\alpha > 0$ be fixed and define

$$\varepsilon_n = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}.$$

Let

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and define the confidence interval

$$I_n = \left[\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n\right].$$

Use Hoeffding's inequality to show that

$$P(p \in I_n) \geq 1 - \alpha.$$

6.1.2 Answer

Hoeffding's inequality tells us that for a sum of independent, bounded random variables $X_1, X_2, \ldots, X_n$ where $X_i \in [0, 1]$, the following holds for any $\epsilon > 0$:

$$P\left(|\hat{p}_n - p| \geq \epsilon\right) \leq 2 \exp\left(-2n\epsilon^2\right)$$

This inequality bounds the probability that the sample mean $\hat{p}_n$ deviates from the true mean p by more than ϵ.

Our first step will be to substitute our formula for ϵ inside the formula for Hoeffding's inequality.

$$P\left(|\hat{p}_n - p| \geq \varepsilon_n\right) \leq 2 \exp\left(-2n \cdot \left(\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)\right)\right)$$

The exponential term simplifies as follows:

$$P\left(|\hat{p}_n - p| \geq \varepsilon_n\right) \leq 2 \exp\left(-\log\left(\frac{2}{\alpha}\right)\right)$$

$$P\left(|\hat{p}_n - p| \geq \varepsilon_n\right) \leq 2 \cdot \frac{\alpha}{2} = \alpha$$

Thus, we have:

$$P(|\hat{p}_n - p| \geq \varepsilon_n) \leq \alpha$$

The complement of the event $|\hat{p}_n - p| \geq \varepsilon_n$ is the event $|\hat{p}_n - p| < \varepsilon_n$, which corresponds to

$$p \in [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n].$$

Therefore, we have:

$$P(p \in [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]) = P(|\hat{p}_n - p| < \varepsilon_n) = P(p \in I_n) \geq 1 - \alpha.$$

6.2 Part 2

6.2.1 Question

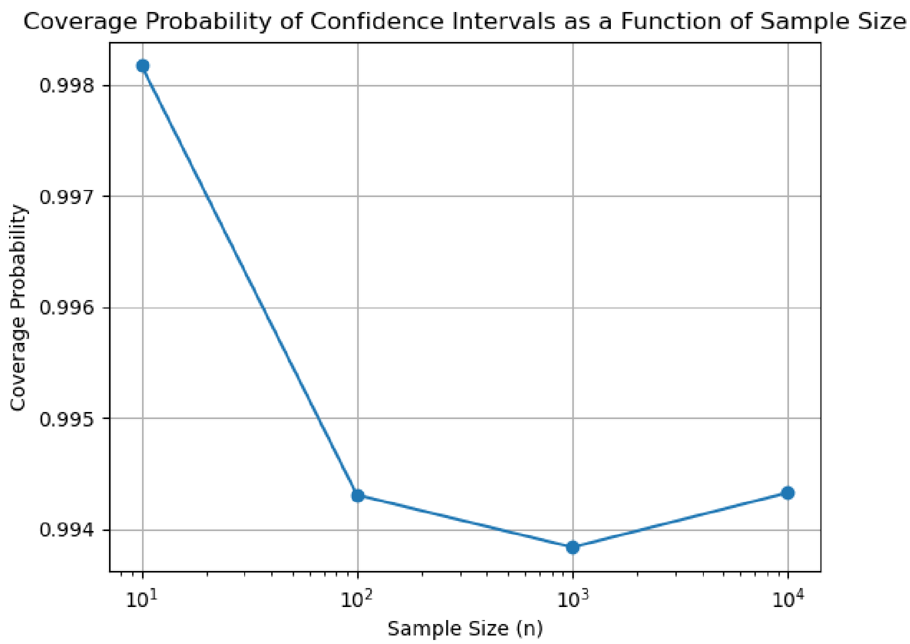
Let $\alpha = 0.05$ and $p = 0.4$. Conduct a simulation study to see how often the confidence interval I_n contains p (called *coverage*). Do this for $n = 10, 100, 1000, 10000$. Plot the coverage as a function of n.

6.2.2 Answer

We set up the simulation using the python package "numpy" and plot it using "matplotlib". We perform 100,000 simulations for each value n and average the result. In addition to the plot we also include a table of the coverage probability.

n	coverage_probability
10	0.99817
100	0.99431
1000	0.99384
10000	0.99433

Table 1: Coverage probability for different sample sizes n.



6.3 Part 3

6.3.1 Question

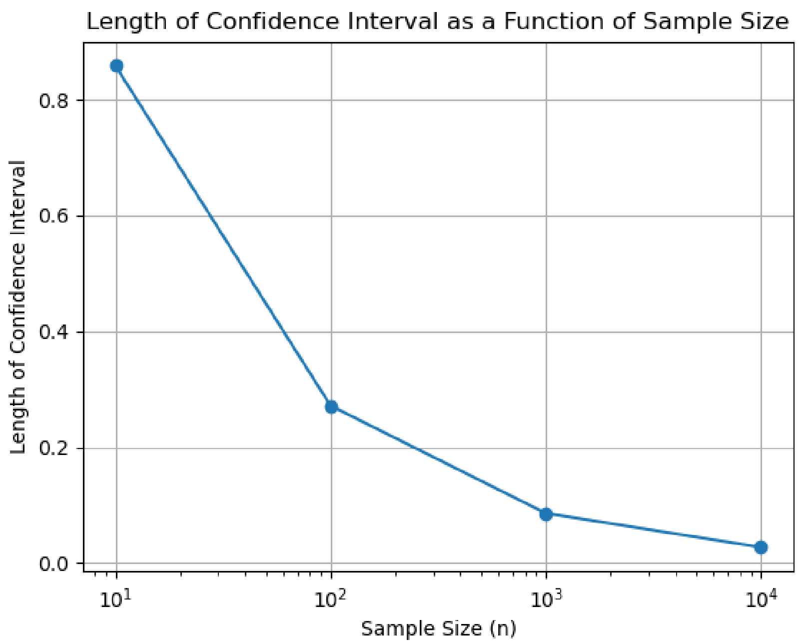
Plot the length of the confidence interval as a function of n.

6.3.2 Answer

We can clearly see that the confidence interval is markedly reduced in relation to the larger sample sizes.

n	length
10	0.85894
100	0.27162
1000	0.085894
10000	0.027162

Table 2: Lengths for different sample sizes n rounded to 5 significant digits.



6.4 Part 4

6.4.1 Question

Say that $X_1, \ldots, X_n$ represents whether a person has a disease or not.

Let us assume that, unbeknownst to us, the true proportion of people with the disease has changed from $p = 0.4$ to $p = 0.5$. We use the confidence interval to make a decision. Specifically, when presented with evidence (samples), we calculate I_n and our decision is that the true proportion of people with the disease is in I_n.

Conduct a simulation study to answer the following question: Given that the true proportion has changed, what is the probability that our decision is correct? Perform this study for $n = 10, 100, 1000, 10000$.

6.4.2 Answer

We perform the simulation from part 2 twice, once using $p_{old} = 0.4$ and once using $p_{new} = 0.5$. For every prediction we then compare the probability that p_{new} is enclosed in the span I_n calculated using p_{old} and compare it to the probability that p_{new} is enclosed in the span I_n calculated using p_{new}.

$$P(correctdecision) = \frac{P\left(p_{new} \in I_{n_{old}}\right)}{P\left(p_{new} \in I_{n_{new}}\right)}$$

As we could see in part 3 the confidence interval I_n is very large when n is small but shrinks when n increases. When $n = 1000$ the length of the confidence interval is less than the difference between p_{new} and p_{old}

$$I_{n \geq 1000} < p_{new} - p_{old} = 0.5 - 0.4 = 0.1$$

This entails that the actual value of p is no longer included in the span making all our decisions incorrect for a large enough n.

n	**correct_count_old**	**correct_count_new**	**proportion_old_vs_new**
10	99388	99830	0.99557
100	76060	99308	0.76590
1000	10	99285	0.00010
10000	0	99324	0.00000

Table 3: Comparison of correct counts and proportions for different sample sizes n.