Introduction to Data Science 1MS041

Group Assignment 1

Group 8

Course Co-ordinator

Benny Avelin

Contributors

Feruz Redi

Moutushi Sen

Samiha Akhter

Shaheryar Shaheryar

Tebogo Mitane

UPPSALA
UNIVERSITET

Problem 1

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

Solution

To show that A and B are two independent events, we can write them as follows:

$$\boldsymbol{P(A \cap B) = P(A) \,.\, P(B)}$$

So, to show that A^c and B^c are also two independent events we must show that:

$$\boldsymbol{P(A^c \cap B^c) = P(A^c) \,.\, P(B^c)}$$

As we know the complements of A and B can be written as:

$$\boldsymbol{P(A^c) = 1 - P(A)}$$

$$\boldsymbol{P(B^c) = 1 - P(B)}$$

Now using the property of complements we can relate $P(A^c \cap B^c)$ with $P(A \cup B)$ as:

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

Now we use the inclusion-exclusion principle on $P(A \cup B)$ as follows:

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

We can substitute $P(A) \,.\, P(B)$ in place of $P(A \cap B)$,

$$\boldsymbol{P(A \cup B) = P(A) + P(B) - [P(A).P(B)]}$$

Now we substitute it back into the original equation:

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

Rearranging,

$$\boldsymbol{P(A^c \cap B^c) = [1 - P(A)][1 - P(B)] = P(A^c).P(B^c)}$$

So, we have proven that:

$$\boldsymbol{P(A^c \cap B^c) = P(A^c).P(B^c)}$$

Problem 2

The probability that a child has brown hair is 1/4. Assume independence between children and assume there are three children.

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

Solution

1) Probability that at least two children have brown hair given that at least one child has brown hair

The probability that any given child has brown hair is $P(B) = \frac{1}{4}$. The probability that a child does not have brown hair is $P(B^c) = \frac{3}{4}$.

Probability of exactly 0 children with brown hair:

$$\boldsymbol{P(N = 0) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}}$$

Probability of exactly 1 child with brown hair:

$$\boldsymbol{P(N = 1) = \binom{3}{1}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2 = \frac{27}{64}}$$

Probability of exactly 2 children with brown hair:

$$\boldsymbol{P(N = 2) = \binom{3}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right) = \frac{9}{64}}$$

Probability of exactly 3 children with brown hair:

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

Now the probability that at least one child is brown haired:

$$\boldsymbol{P(N \geq 1) = 1 - P(N = 0) = 1 - \frac{27}{64} = \frac{37}{64}}$$

We want the conditional probability that at least two children have brown hair given that at least one child has brown hair. This is given by:

$$\boldsymbol{P(N \geq 2 \mid N \geq 1) = \frac{P(N \geq 2)}{P(N \geq 1)}}$$

Now we find $P(N \geq 2)$, the probability of at least two children with brown hair:

$$\boldsymbol{P(N \geq 2) = P(N = 2) + P(N = 3) = \frac{9}{64} + \frac{1}{64} = \frac{5}{32}}$$

So finally,

$$\boldsymbol{P(N \geq 2 \mid N \geq 1) = \frac{\frac{5}{32}}{\frac{37}{64}} = \frac{5}{32} * \frac{54}{37} = \frac{10}{37}}$$

2) Probability that at least two children have brown hair given that the oldest child has brown hair

Provided that the oldest child has brown hair, we only need to consider the remaining two children. For these two children, the probability of each child having brown hair remains 1/4, and they are independent of each other.

- **A:** At least 2 children have brown hair (including oldest)
- **B:** The oldest has brown hair

Basically, we must find probability of A given B, lets assume that X is the number of remaining two children (beside the oldest one).

$$\boldsymbol{P(X \geq 1) = 1 - P(X = 0)}$$

P(X=0) means the probability that both the other children have brown hair is 0:

$$\boldsymbol{P(X = 0) = \left(\frac{3}{4}\right)^2 = \frac{9}{16}}$$

Lastly,

$$\boldsymbol{P(X \geq 1) = 1 - \frac{9}{16} = \frac{7}{16}}$$

So, if the oldest child has brown hair, the probability of at least two children have brown hair is 7/16.

Problem 3

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

Solution

Our aim is to find the CDF of $\boldsymbol{F_R(r)}$ and the PDF $\boldsymbol{f_R(r)}$ of the random variable $\boldsymbol{R = \sqrt{X + Y^2}}$.

We know that CDF of R is

$$\boldsymbol{F_R(r) = P(R \leq r)}$$

$$\boldsymbol{F_R(r) = P(\sqrt{X^2 + Y^2} \leq r)}$$

It is given that the pair (X, Y) is uniformly distributed over the unit disc, the probability that $R \leq r$ corresponds to the area of a circle with radius r, scaled by the total area of the unit disc π.

We know the area of a circle with radius r is πr^2, so the CDF is:

$$\boldsymbol{F_R(r) = \frac{Area\ of\ circle\ with\ radius\ r}{Area\ of\ unit\ disc} = \frac{\pi r^2}{\pi} = r^2}$$

So, the CDF is:

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

Good!

And we know that the PDF is the derivative of CDF:

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

We also know that $F_R(r) = r^2$, so the PDF is:

Good!

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

So, the PDF is:

$$f_R(r) = \begin{cases} 0, & r < 0, \\ 2r, & 0 \le r \le 1, \\ 0, & r > 1. \end{cases}$$

Problem 4

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

Solution

The probability of getting heads in a single coin toss is $p = \frac{1}{2}$

The random variable X represents the number of tosses needed to get the first head, which means X can take values 1,2,3, …

For a geometric distribution with success probability $p = \frac{1}{2}$, the probability mass function PMF is:

$$P(X = k) = (1-p)^{k-1}.p, \text{ for } k = 1,2,3, \ldots$$

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

This means the probability of requiring exactly k tosses to get the first head is $(\frac{1}{2})^k$

Now we use the Expected Value formula to evaluate the geometric random variable with success probability p as:

$$\boldsymbol{E(X) = \frac{1}{p} = \frac{1}{\frac{1}{2}} = 2}$$

So, the expected number of coin tosses required to get the first head is 2.

Problem 5

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

1. Let $\alpha > 0$ be fixed and define

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

Let $\hat{p}_n = \frac{1}{n}\Sigma_{i=1}^{n} X_i$ and define the confidence interval $I_n = [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]$. Use Hoeffding's inequality to show that $\mathbb{P}(p \in I_n) \geq 1 - \alpha$.

2. 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.
3. Plot the length of the confidence interval as a function of n.
4. Say that say that $X_1, \dots, X_n$ represents if a person has a disease or not. Let us assume that unbeknownst to us the true proportion of people with disease has changed from $p = 0.4$ to $p = 0.5$. We use the confidence interval to decide, that is 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? Again, using $n = 10,100,1000,10000$.

Solution

1) Confidence Interval and Hoeffding's Inequality

We are given that $X_1, \dots, X_n$ i.i.d. from Bernoulli (p), and we need to derive a confidence interval for p using Hoeffding's inequality.

We know that empirical mean is:

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

And the confidence interval is given as:

$$\boldsymbol{I_n} = [\hat{\boldsymbol{p}}_{\boldsymbol{n}} - \boldsymbol{\varepsilon_n}, \hat{\boldsymbol{p}}_{\boldsymbol{n}} + \boldsymbol{\varepsilon_n}]$$

Where,

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

Hoeffding's inequality for bounded RV:

$$\mathbb{P}(|\hat{\boldsymbol{p}}_{\boldsymbol{n}} - \boldsymbol{p}| \geq \boldsymbol{\varepsilon_n}) \leq \boldsymbol{2e^{-2n\varepsilon_n^2}}$$

We can substitute ε_n with $\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$, we get:

$$\mathbb{P}\left(|\hat{\boldsymbol{p}}_{\boldsymbol{n}} - \boldsymbol{p}| \geq \sqrt{\frac{\boldsymbol{1}}{\boldsymbol{2n}} \boldsymbol{\log} \frac{\boldsymbol{2}}{\boldsymbol{\alpha}}}\right) \leq \boldsymbol{2e^{-2n.\frac{1}{2n}\log\frac{2}{\alpha}}}$$

After simplifying we get:

$$\mathbb{P}(|\hat{\boldsymbol{p}}_{\boldsymbol{n}} - \boldsymbol{p}| \geq \boldsymbol{\varepsilon_n}) \leq \boldsymbol{2e^{-\log\frac{2}{\alpha}} = 2.\frac{\alpha}{2} = \alpha}$$

Lastly,

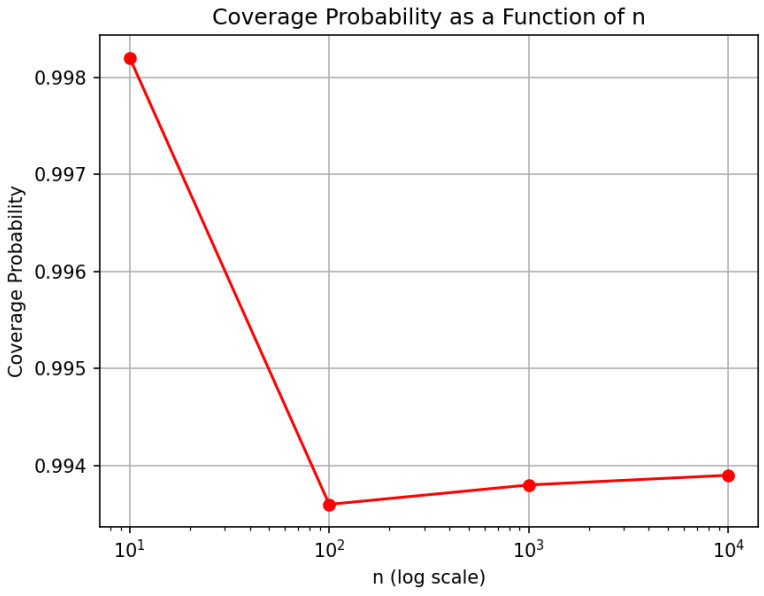
$$\mathbb{P}(|\hat{\boldsymbol{p}}_{\boldsymbol{n}} - \boldsymbol{p}| < \boldsymbol{\varepsilon_n}) \geq \boldsymbol{1 - \alpha}$$

This shows that the probability that p lies in the interval $[\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]$ is at least $1 - \alpha$ as required:

$$\mathbb{P}(\boldsymbol{p} \in \boldsymbol{I_n}) \geq \boldsymbol{1 - \alpha}$$

2) Coverage as a function of n

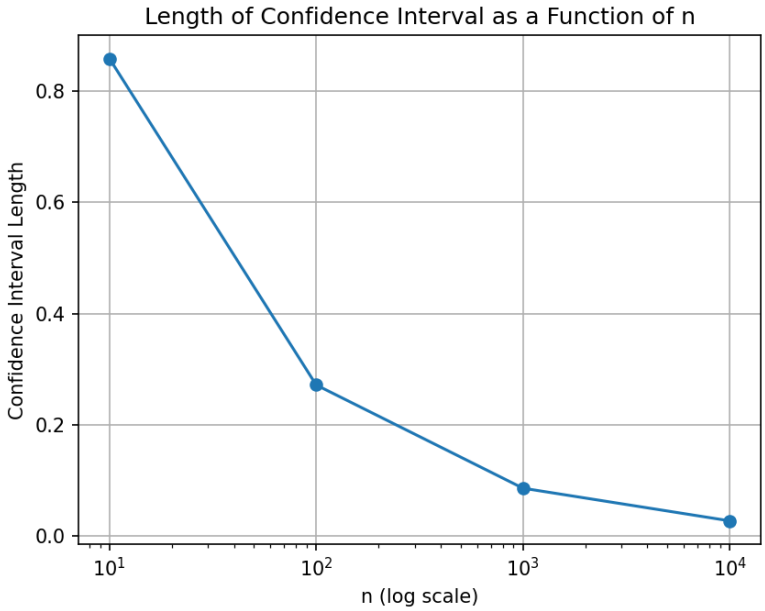
```
3) def simulate_coverage(n, p, alpha, num_simulations=1000):
4)     epsilon_n = np.sqrt(1 / (2 * n) * np.log(2 / alpha))
5)     coverage_count = 0
6)
7)     for _ in range(num_simulations):
8)
9)         X = np.random.binomial(1, p, size=n)
10)        p_hat_n = np.mean(X)
11)
12)
13)        lower_bound = p_hat_n - epsilon_n
14)        upper_bound = p_hat_n + epsilon_n
15)
16)
17)        if lower_bound <= p <= upper_bound:
18)            coverage_count += 1
19)
20)
21)    coverage_probability = coverage_count / num_simulations
22)    return coverage_probability
23)
24)alpha = 0.05
25)p = 0.4
26)n_values = [10, 100, 1000, 10000]
27)num_simulations = 10000
28)coverage_values = []
```



3) Length of confidence

```
def confidence_interval_length(n, alpha):
    epsilon_n = np.sqrt(1 / (2 * n) * np.log(2 / alpha))
    return 2 * epsilon_n

n_values = [10, 100, 1000, 10000]
interval_lengths = [confidence_interval_length(n, alpha) for n in
n_values]
```



4) Simulation for Correct Decision Probability

```
def simulate_decision_correctness(n, p_true, p_assumed, alpha,
num_simulations=1000):
    epsilon_n = np.sqrt(1 / (2 * n) * np.log(2 / alpha))
    correct_decision_count = 0

    for _ in range(num_simulations):

        X = np.random.binomial(1, p_assumed, size=n)
        p_hat_n = np.mean(X)


        lower_bound = p_hat_n - epsilon_n
        upper_bound = p_hat_n + epsilon_n


        if lower_bound <= p_true <= upper_bound:
            correct_decision_count += 1
```

```
21)
22)
23)     decision_probability = correct_decision_count / num_simulations
24)     return decision_probability
25)
26)p_true = 0.5
27)p_assumed = 0.4
28)alpha = 0.05
29)n_values = [10, 100, 1000, 10000]
30)num_simulations = 1000
31)decision_probabilities = []
```