Group Assignment 1

Kirti Barman, Sourav Sharan, Vyoma Shah,
Aditya Khadkikar, Kokila Wickramasinghe

Group 15 - Introduction to Data Science

October 24, 2024

Exercise 12.1.1. Suppose that A and B are independent events, show that A^c and B^c are independent.

Solution.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{[Addition Rule]}$$

Since A and B are independent, $P(A \cap B) = P(A)P(B)$. So,

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

$$P(A \cup B) = P(A) + P(B)(1 - P(A))$$
$$P(A \cup B) = P(A) + P(B)P(A^c)$$
$$P(A \cup B) = 1 - P(A^c) + P(B)P(A^c)$$
$$P(A \cup B) = 1 - P(A^c)(1 - P(B))$$
$$P(A \cup B) = 1 - P(A^c)P(B^c)$$
$$1 - P(A \cup B) = P(A^c)P(B^c)$$
$$P(A \cup B)^c = P(A^c)P(B^c)$$

Since $P(A \cup B)^c = P(A^c \cap B^c)$ by De Morgan's Law, we have:

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

Therefore, $P(A^c)$ and $P(B^c)$ are independent. □

Exercise 12.1.2. 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. Let B be the event where a child has brown hair, and N be the event where a child does not have brown hair.

Given, P(B) = $\frac{1}{4}$, so P(N) = 1-$\frac{1}{4}$ = $\frac{3}{4}$

There are 3 children, so there can be total 8 possible combinations : {BBB, BBN, BNN, NNN, NBB, NNB, NBN, BNB}

1. Let Y be the event that at least one child has brown hair, so it can have the following cases : {BBB, BBN, BNN, NBB, NNB, NBN, BNB}
   so, P(Y) = $\frac{1}{4}\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4}\frac{3}{4} + \frac{1}{4}\frac{3}{4}\frac{3}{4} + \frac{3}{4}\frac{1}{4}\frac{1}{4} + \frac{3}{4}\frac{3}{4}\frac{1}{4} + \frac{3}{4}\frac{1}{4}\frac{3}{4} + \frac{1}{4}\frac{3}{4}\frac{1}{4} = \frac{37}{64}$

   Let X be the event where at least 2 children have brown hair. It is considered that when X occurs, Y is also true, because having at least 2 children with brown hair, will include at least 1 child having brown hair (Y).

   X can be also given as X ∩ Y : {BBB, BBN, NBB, BNB}
   so P(X ∩ Y) = $\frac{1}{4}\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4}\frac{3}{4} + \frac{3}{4}\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{3}{4}\frac{1}{4} = \frac{10}{64}$

   So, the probability that at least two children have brown hair, given that at least one child has brown hair will be,
   P(X | Y) = $P(X \cap Y)$/P(Y) = $\frac{10}{37}$
2. Let Y be the event that the oldest child has brown hair, so it can have the following cases: {BBB, BBN, BNN, BNB}
   P(Y) = $\frac{1}{4}\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4}\frac{3}{4} + \frac{1}{4}\frac{3}{4}\frac{3}{4} + \frac{1}{4}\frac{3}{4}\frac{1}{4} = \frac{16}{64}$

   Let X be the event where at least 2 children have brown hair,
   X ∩ Y : {BBB, BBN, BNB}
   So P(X ∩ Y) = $\frac{1}{4}\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4}\frac{3}{4} + \frac{1}{4}\frac{3}{4}\frac{1}{4} = \frac{7}{64}$

   So, the probability that at least two children have brown hair, given that oldest child has brown hair will be,
   P(X | Y) = $P(X \cap Y)$/P(Y) = $\frac{7}{16}$

□

Exercise 12.1.3. Let (X,Y) be uniformly distributed on the unit disc

$$\{(x, y) \in \Re^2 | x^2 + y^2 \leq 1\}$$

Set $R = \sqrt{X^2 + Y^2}$. What is the CDF and PDF of R?

Solution. $R = \sqrt{X^2 + Y^2}$ would be the radius of the disc. The value of R can vary depending on X and Y, but it will always be between 0 and 1. Since X and Y are uniformly distributed, the probability of every point in the circle will be same, and therefore probability of R being of a particular length will uniformly increase as the area of the circle it makes increases. So the CDF of R at a perticular length, can be written down as a ratio of area of the circle over area of the whole unit circle.

The CDF of R can be found by:

$$\frac{\pi r^2}{\pi (1)^2} = r^2 \ , 0 \leq r \leq 1$$

PDF of R is found by derivating the above:

$$\frac{d}{dr} r^2 = 2r, 0 \leq r \leq 1$$

Therefore, CDF of R is

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

And, PDF of R is

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

□

Exercise 12.1.4. 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 number of tosses X required to get the first head follows a geometric distribution with probability $p = \frac{1}{2}$ (since the coin is fair). The expected value $E[X]$ of such a geometric distribution is given by:

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

$$E(X) = \sum_{n=1}^{\infty} n \cdot P(X = x)$$

$$E(X) = \sum_{n=1}^{\infty} n \cdot (\frac{1}{2})^n$$

For a fair coin, $p = \frac{1}{2}$, so the sum of an infinite geometric series can be approximated using:

$$a_1[n = 1] = 1(1/2)^1 = 1/2$$

$$a_2[n = 2] = 2(1/2)^2 = 2/4 = 1/2$$

$$a_3[n = 3] = 3(1/2)^3 = 3/8$$

$$a_4[n = 4] = 4(1/2)^4 = 4/16$$

For $n \to \infty$, the above sum approaches:

$$1 + 3/8 + 4/16 + 5/32 + 6/64 + ... \approx 2$$

Thus, the expected number of tosses E[X] to get the first head is 2.

□

Exercise 12.1.5. Let X_1, ... , X_n be IID from Bernoulli(p).

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

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

   Let $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and define the confidence interval $I_n = [\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n]$. Use Hoeffding's inequality to show that

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

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 $X_1, \ldots, X_n$ represents if 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, 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. Given $X_1, X_2, \ldots, X_n$ be IID random variables from Bernoulli(p) with $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

We have $\epsilon_n = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$. By using Hoeffding's Inequality:

$$P(|\hat{p}_n - p| \geq \epsilon_n) \leq 2\exp(-2n\epsilon_n^2).$$

Substituting the value of ϵ_n into above equation:

$$P(|\hat{p}_n - p| \geq \epsilon_n) \leq 2\exp(-2n(\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}})^2)$$
$$\Rightarrow P(|\hat{p}_n - p| \geq \epsilon_n) \leq 2\exp(-2n(\frac{1}{2n}\log\frac{2}{\alpha}))$$

$$Since, \quad \exp(-\log x) = \frac{1}{x} \Rightarrow \exp(-\log\frac{2}{\alpha}) = \frac{\alpha}{2} - (i)$$

Simplifying the exponent & replacing the formula from equation (i), we get

$$\Rightarrow P(|\hat{p}_n - p| \geq \epsilon_n) \leq 2(\frac{\alpha}{2})$$

$$\Rightarrow P(|\hat{p}_n - p| \geq \epsilon_n) \leq \alpha \qquad - (ii)$$

Now, we know that confidence interval

$$I_n = [\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n] \qquad - (iii)$$

From equation (ii), when

$$\hat{p}_n - p \leq \epsilon_n \Rightarrow p > \hat{p} - \epsilon_n \quad and \quad - \hat{p}_n + p \leq \epsilon_n \Rightarrow p \leq \hat{p}_n + \epsilon_n$$

$$\Rightarrow \hat{p}_n + \epsilon_n \geq p > \hat{p}_n - \epsilon_n$$

By using equation (iii),

$$\Rightarrow p \in I_n$$

Thus,

$$P(p \in I_n) \leq \alpha$$

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

2. The coverage probabilities for different sample sizes (n) are given below:
   $n = 10$: **0.9161**, $n = 100$: **0.9442**, $n = 1000$: **0.9506**, $n = 10000$: **0.9515**
   The plot below shows the coverage probability as a function of the sample size n.

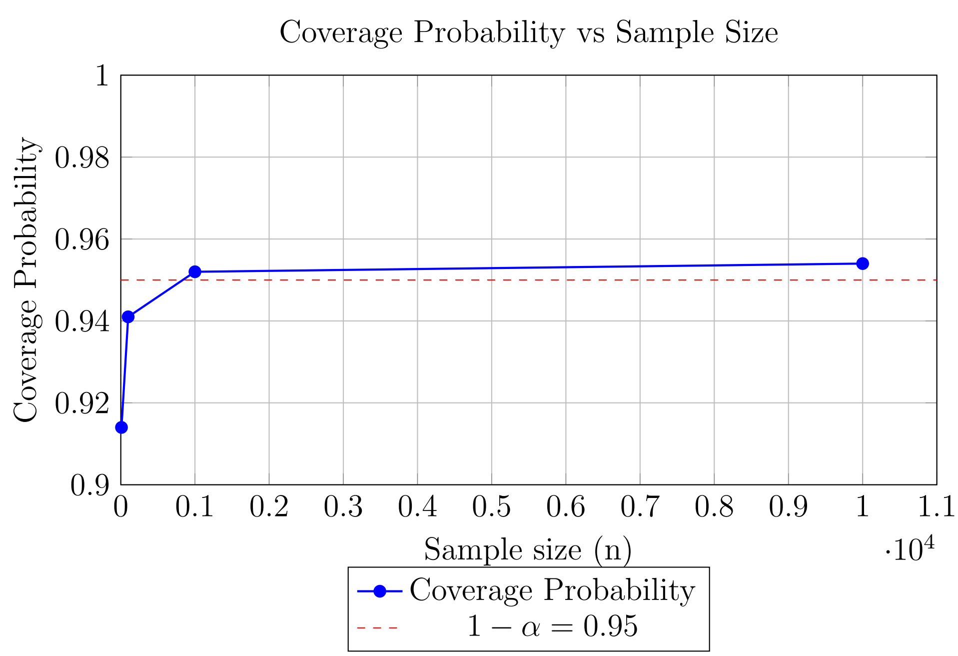

Figure 1: Coverage Probability vs. Sample Size

3. We have considered below parameters to populate the graph:

- `p_true = 0.4`: This sets the true probability of success for the Bernoulli distribution. Each trial returns 1 with a probability of 0.4 and 0 otherwise.

- `alpha = 0.05`: This is the significance level for the confidence interval, corresponding to a 95% confidence level.

- `n_values = [10, 100, 1000, 10000]`: These are the different sample sizes to be tested in the simulation.

- `num_simulations = 10000`: This defines how many times we repeat the process to estimate the coverage probability for each sample size n.

Function Definitions: `epsilon_n(n, alpha)`: This function calculates the half-width ϵ_n of the confidence interval using the formula derived from Hoeffding's inequality:

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$

It depends on both the sample size n and the significance level α.

`is_p_in_interval(sample, p_true, alpha)`: This function checks whether the true probability p_{true} is within the confidence interval for a given sample. It first computes the sample mean $\hat{p}_n$ (the proportion of successes) and then calculates the confidence interval:

$$I_n = [\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n]$$

It returns `True` if the true probability p_{true} lies within this interval and `False` otherwise.

Simulation for Coverage and Confidence Interval Length: `coverage_results = []` and `ci_lengths = []`: These lists store the coverage probability and the average length of the confidence intervals for each sample size n.

For each sample size n:

- `coverage_count = 0` and `ci_length = 0`:
  - `coverage_count`: Tracks how many times p_{true} is within the confidence interval.
  - `ci_length`: Accumulates the total length of the confidence intervals across simulations.

Simulation Loop (`for _ in range(num_simulations)`):

- **Sample Generation:** A sample of size n is drawn from a Bernoulli distribution with probability p_{true}. The sample mean $\hat{p}_n$ is computed.

- **Coverage Calculation:** The function `is_p_in_interval` checks if p_{true} lies within the computed confidence interval. If true, `coverage_count` is incremented.

- **Confidence Interval Length Calculation:** The length of the confidence interval is computed as $2 \times \epsilon_n$, and it is added to the total `ci_length`.

Storing Results: After running the simulations for a given sample size n, the coverage probability is computed as the fraction of times p_{true} is in the interval:

$$\text{coverage probability} = \frac{\text{coverage\_count}}{\text{num\_simulations}}$$

The average confidence interval length is computed by dividing the accumulated `ci_length` by the number of simulations.

Length of the Confidence Interval

We plot the average length of the confidence interval as a function of the sample size n. The plot below shows how the length of the confidence interval decreases as the sample size increases.

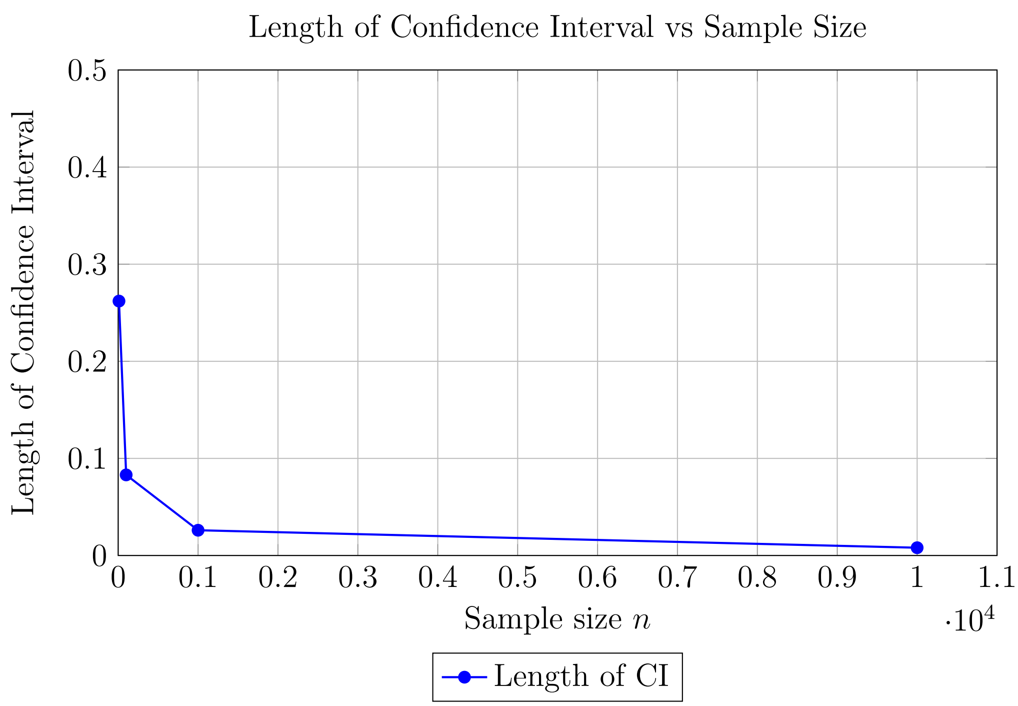

Figure 2: Length of Confidence Interval vs. Sample Size

4. **Simulation Testing**: In this study, we aim to assess the probability that our confidence interval accurately contains the true proportion of individuals with a disease after a change in the population proportion. Specifically, we analyze the scenario where the true proportion changes from $p = 0.4$ to $p = 0.5$.

To estimate the true proportion using confidence intervals, we employ the normal approximation method. The confidence interval for a binomial proportion can be calculated using the formula:

$$I_n = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{1}$$

where:

- $\hat{p}$ is the sample proportion, calculated as $\hat{p} = \frac{x}{n}$,
- x is the number of individuals with the disease in our sample,
- n is the total sample size,
- z is the z-score corresponding to the desired confidence level (e.g., $z \approx 1.96$ for a 95% confidence interval).

Simulation Procedure: We conducted a simulation study with sample sizes $n = 10, 100, 1000$, and 10000. The steps are as follows:

1. For each sample size n, we performed $10,000$ simulations.
2. In each simulation, we randomly sampled n individuals from a binomial distribution with the true proportion $p = 0.5$.
3. We calculated the sample proportion $\hat{p}$ and then constructed the confidence interval I_n using the normal approximation method.
4. We checked if the true proportion $p = 0.5$ lies within the calculated confidence interval I_n.
5. Finally, we recorded the number of times the confidence interval correctly contained $p = 0.5$.

The following code was used to implement the simulation study:

```
import numpy as np

alpha = 0.05
p_old = 0.4
p_new = 0.5

sample_sizes = [10, 100, 1000, 10000]
num_simulations = 10000
results = {}

for n in sample_sizes:
    correct_count = 0
    epsilon_n = np.sqrt(1 / (2 * n) * np.log(2 / alpha))

    for _ in range(num_simulations):
        samples = np.random.binomial(1, p_new, n)
        p_hat_n = np.mean(samples)

        lower_bound = p_hat_n - epsilon_n
        upper_bound = p_hat_n + epsilon_n

        if lower_bound <= p_old <= upper_bound:
            correct_count += 1

    results[n] = correct_count / num_simulations

# Printing Results
for n, prob in results.items():
    print(f"n = {n}: P(Correct) = {prob:.4f}")
```

Results

The probabilities observed for each sample size are as follows:

- For $n = 10$: $P(\text{correct}) \approx 0.9900$
- For $n = 100$: $P(\text{correct}) \approx 0.7558$
- For $n = 1000$: $P(\text{correct}) \approx 0.0000$
- For $n = 10000$: $P(\text{correct}) \approx 0.0000$

These results indicate that as the sample size increases, the probability of making a correct decision regarding the confidence interval more centered around p new (0.5) and the old value is less likely to show up.

□