Introduction to Data Science
Assignment 1 (Group 3)

Kartikey Sharma, Leire Unciti, Aviral Jain, Denice Wikström,
Alexander Ares Nilsson

September 26, 2024

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

Solution

Let's prove that: $P(A^c \cap B^c) = P(A^c)P(B^c)$:

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

$= 1 - P(A \cup B)$

$= 1 - (P(A) + P(B) - P(A \cap B))$

$= 1 - P(A) - P(B) + P(A \cap B)$

$= 1 - P(A) - P(B) + P(A)P(B)$

$= (1 - P(A))(1 - P(B)) = P(A^c)P(B^c)$

---

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

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

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

Solution

a. Let's define X as the number of children having brown hair.

   We know that:-
   P(child having brown hair) = 1/4
   Number of children = 3

$P(X >= 1)$
$= 1 - P(X < 1)$
$= 1 - \binom{3}{0}(\frac{3}{4})^3$
$= 1 - \frac{27}{64} = \frac{37}{64}$

$P(X >= 2)$
$= \binom{3}{3}(\frac{1}{4})^3 + \binom{3}{2}(\frac{1}{4})^2(\frac{3}{4})$
$= \frac{1}{64} + \frac{9}{64} = \frac{10}{64}$

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

P(at least 2 children have brown hair given that at least **one child** has brown hair) $= \frac{10}{37}$

b. Let A be the event that oldest child has brown hair, and we have to find $P(X >= 2|A)$

Define Y as $X >= 2|A$.
ie. if the oldest child has brown hair then at least 1 of the remaining children will have brown hair.

$= P(Y >= 1)$
$= P(Y = 2) + P(Y = 1)$
$= \binom{2}{2}(\frac{1}{4})^2 + \binom{2}{1}(\frac{1}{4})(\frac{3}{4})$
$= \frac{1}{16} + \frac{6}{16} = \frac{7}{16}$

P(at least 2 children have brown hair given that the **oldest** child has brown hair) $= \frac{7}{16}$

---

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

Solution

CDF of $R = F_R(r)$
$= P(R <= r)$
$= \frac{\text{Area of disc of radius } r}{\text{Area of disc of radius } 1} = \frac{\pi r^2}{\pi 1^2} = r^2$, where $0 \leq r \leq 1$

PDF of $R = \frac{d(F_R(r))}{dr} = \frac{d(r^2)}{dr} = \begin{cases} 2r, \ 0 \leq r \leq 1 \\ 0, \text{ otherwise} \end{cases}$

---

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?

$X \sim G(1/2)$

$\mathbb{E}[X] = \sum_{x=1}^{\infty} xf(x)$

$= \sum_{x=1}^{\infty} xp(1-p)^{x-1}$

$= p\sum_{x=1}^{\infty} x(1-p)^{x-1}$, Let $q = 1-p$

$= p\sum_{x=1}^{\infty} xq^{x-1}$

$= p\sum_{x=1}^{\infty} \frac{\partial q^x}{\partial q}$

$= p\frac{\partial}{\partial q}\sum_{x=1}^{\infty} q^x = p\frac{\partial}{\partial q}(1-q)^{-1} = p(1-q)^{-2} = p(p^{-2}) = \frac{1}{p}$

So, considering p=1/2

$\mathbb{E}[X] = 2$

---

5 Let $X_1, ..., X_n$ be IID from *Bernouli(p)*.

a. **Let $\alpha > 0$ be fixed and define**
$\epsilon_n = \sqrt{\frac{1}{2n}log(\frac{2}{\alpha})}$
Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and define the confidence interval $I_n = [\hat{p} - \epsilon_n, \hat{p} + \epsilon_n]$. Use Hoeffding's inequality to show that $P(p \in I_n) \geq 1 - \alpha$.

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

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

d. **Say $X_1, ..., X_n$ represents if a person has a disease or not. Let us assume that unbeknownst to us the true proportions 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 questions: Given that the true proportion has changed, what is the probability that our decision is correct? Again using $n = 10, 100, 1000, 10000$.**

Solution

a. We know $\alpha > 0$ and
$\epsilon_n = \sqrt{\frac{1}{2n}log(\frac{2}{\alpha})}$
$\implies 2n\epsilon_n^2 = log(\frac{2}{\alpha}) \implies e^{2n\epsilon_n^2} = \frac{2}{\alpha} \implies \alpha = 2e^{-2n\epsilon_n^2}$

By Hoeffding's Inequality,

$P(|X - \mathbb{E}[X]| \geq \epsilon_n) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$

For Bernoulli random variable, a = 0 and b = 1

$\implies P(|X - \mathbb{E}[X]| \geq \epsilon_n) \leq 2e^{-2n\epsilon^2}$

$P(p \in I_n)$

$= P(\hat{p} - \epsilon_n \leq p \leq \hat{p} + \epsilon_n)$

$= 1 - P(|\hat{p} - p| \geq \epsilon_n),$

$\geq 1 - 2e^{-2n\epsilon_n^2}$

$\geq 1 - \alpha$ [**since** $\alpha = 2e^{-2n\epsilon_n^2}$]

So we finally get $P(p \in I_n) \geq 1 - \alpha$

b. We know:

$\alpha = 0.05$

$p = 0.4$

$n = \{10, 100, 1000, 10000\}$

Plot

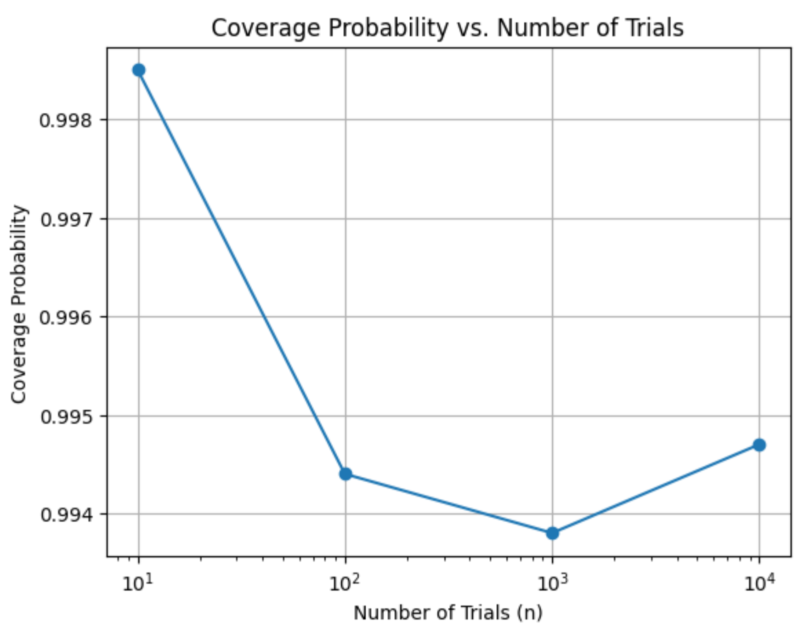

Figure 1: Coverage Probability vs. Number of Trials

Python Code

```python
import matplotlib.pyplot as plt
import numpy as np
alpha = 0.05
p = 0.4
result = []
for n in [10, 100, 1000, 10000]:
    epsilon = np.sqrt((1/(2*n))*np.log(2/alpha))
    cnt = 0
    for trials in range(10000):
        array = np.random.binomial(1, p, n)
        # randomly generating bernouli RVs
        p_hat = np.mean(array)
        if p >= (p_hat-epsilon) and p <= (p_hat+epsilon):
            cnt = cnt + 1
    result.append(float(cnt)/float(10000))
plt.plot([10,100,1000,10000], result, marker='o')
plt.xlabel('Number of Trials (n)')
plt.ylabel('Coverage Probability')
plt.title('Coverage Probability vs. Number of Trials')
plt.xscale('log'); plt.grid(True); plt.show()
```

c. *Plot*

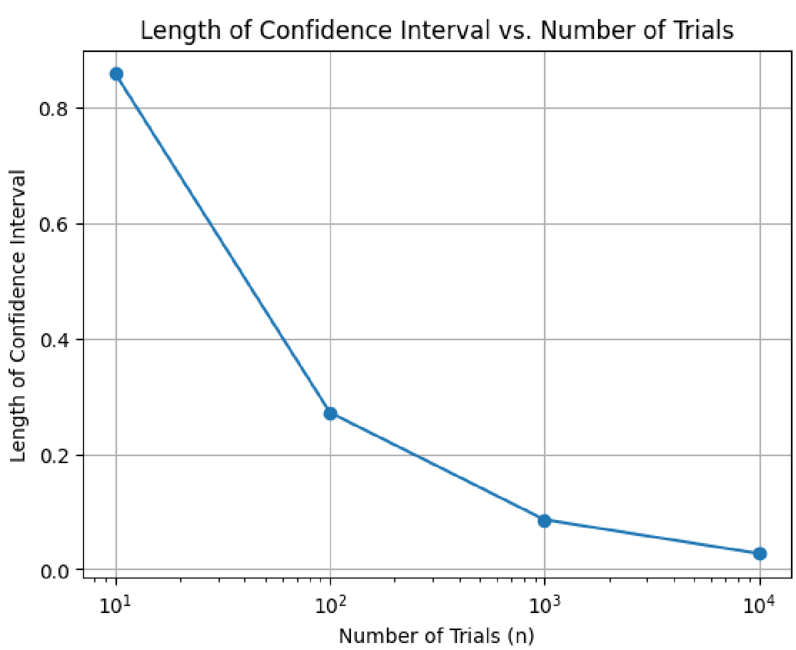

Figure 2: Length of Confidence Intervals vs. Number of Trials

Python Code

```python
import matplotlib.pyplot as plt
import numpy as np
alpha = 0.05
p = 0.4
result = []
for n in [10, 100, 1000, 10000]:
    epsilon = np.sqrt((1/(2*n))*np.log(2/alpha))
    result.append(2*epsilon)
plt.plot([10,100,1000,10000], result, marker='o')
plt.xlabel('Number of Trials (n)')
plt.ylabel('Length of Confidence Interval')
plt.title('Length Confidence Interval vs. Number of Trials')
plt.xscale('log')
plt.grid(True)
plt.show()
```

d. We DON'T know that p changed:

$p = 0.4 \to 0.5$ (Let's refer the new value to as p^*)

We compute the probability $P(p \in I_n | p \to p^*)$

Plot

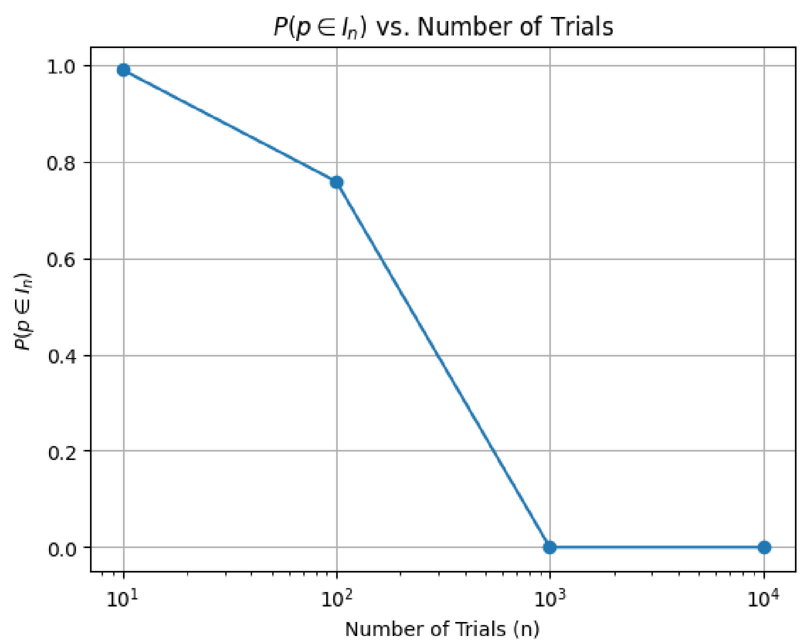

Figure 3: Probability that p lies in confidence interval vs. Number of Trials

Method

Generate samples with new proportion p^*.

Calculate interval I_n.

Calculate the probability that proportion $p \in I_n$, ie. $P(p \in I_n)$.

Python Code

```python
import numpy as np
import matplotlib.pyplot as plt


alpha = 0.05
p = 0.4 # old p
p_star = 0.5 # new p

probability_result = []

for n in [10, 100, 1000, 10000]:

    epsilon = np.sqrt((1/(2*n))*np.log(2/alpha))
    cnt = 0
    for trials in range(10000):
        array = np.random.binomial(1, p_star, n)
        p_hat = np.mean(array)

        if p_hat-epsilon <= p and p <= p_hat+epsilon:
            cnt = cnt + 1

    probability_result.append(float(cnt)/float(10000))

plt.plot([10,100,1000,10000],probability_result,marker='o')
plt.xlabel('Number of Trials (n)')
plt.ylabel(r'$P(p \in I_n)$')
plt.title(r'$P(p \in I_n)$ vs. Number of Trials')
plt.xscale('log')
plt.grid(True)
plt.show()
```