GroupAssignment1_Team10

Ismat Halabi, Anudeep Battu, Muhammad Hasnain Saeed, Muhammad Sajjad, Lavy Selvaraj

September 27, 2024

1. We proceed as follows.

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

$$\implies P(A^c) = P(A^c \cap B) + P(A^c \cap B^c) \qquad \text{lemma 1.12.3} \tag{2}$$

$$\implies P(A^c \cap B^c) = P(A^c) - P(A^c \cap B) \tag{3}$$

$$B = (A^c \cap B) \cup (A \cap B) \tag{4}$$

$$\implies P(B) = P(A^c \cap B) + P(A \cap B) \qquad \text{lemma 1.12.3} \tag{5}$$

$$\implies P(A^c \cap B) = P(B) - P(A \cap B) \tag{6}$$

Plugging equation 6 into equation 3, we get:

$$\begin{aligned}
P(A^c \cap B^c) &= P(A^c) - [P(B) - P(A \cap B)] \\
&= P(A^c) - [P(B) - P(A)P(B)] \text{ A \& B independent} \\
&= [1 - P(A)] - [P(B) - P(A)P(B)] \;\; \text{lemma 1.11.1} \\
&= [1 - P(A)] - P(B)[1 - P(A)] \\
&= [1 - P(A)][1 - P(B)] \\
&= P(A^c)P(B^c) \\
&\implies A^c \text{ and } B^c \text{ are independent by lemma 1.11.1} \blacksquare
\end{aligned}$$

+

2. Let $X \in Binomial(n = 3, p = 1/4)$ be a random variable of having k children out of 3 with Brown hair, with independent outcomes. This distribution fits the problem here. It is well known that:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{7}$$

(a) We want to compute $P(X \geq 2|X \geq 1)$.

$$\begin{aligned}
P(X \geq 2|X \geq 1) &= \frac{P(X \geq 2, X \geq 1)}{P(X \geq 1)}, \text{ def. of conditional probability} \\
&= \frac{P(X \geq 2)}{P(X \geq 1)}, \text{ intersection of } X \geq 2 \text{ and } X \geq 1 \\
&= \frac{1 - P(X \leq 1)}{1 - P(X = 0)}, \text{ Lemma 1.11} \\
&= \frac{1 - [P(X = 0) + P(X = 1)]}{1 - \binom{3}{0}\left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3}, \text{ X=0 and X=1 are disjoint events} \\
&= \frac{1 - \left[\binom{3}{0}\left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3 + \binom{3}{1}\left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^2\right]}{1 - \left(\frac{3}{4}\right)^3} \\
&= \frac{1 - \left(\frac{27}{64} + \frac{27}{64}\right)}{1 - \frac{27}{64}} = \frac{1 - \frac{54}{64}}{1 - \frac{27}{64}} = \frac{10}{37}.
\end{aligned}$$

+

(b) Let $Y \in Binomial(n = 2, p = 1/4)$. $Y = X - 1$

$$\begin{aligned}
P(X \geq 2|X = 1) &= P(Y \geq 1) \\
&= 1 - P(Y = 0) \\
&= 1 - \binom{2}{0}\left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^2 \\
&= 1 - \frac{9}{16} = \frac{7}{16}.
\end{aligned}$$

+

3. We have:

$$F_R(r) = P(R \leq r) = P(R \leq \sqrt{X^2 + Y^2}) \tag{8}$$

The random variable R is then the set of disks which have radius $0 < r < 1$ that are centered in the origin. The maximum radius is 1 since the outer boundary of X^2 and Y^2 is the unit disc, which has an area of $\pi r^2 = \pi$. Moreover, since X and Y are uniformly distributed, $F_R(r)$ is the portion of the area of the disc represented by R compared to the unit disc. The cumulative distribution function (CDF) is given by:

$$F_R(r) = P(R \leq r) = P(R \leq \sqrt{X^2 + Y^2}) = \frac{\pi r^2}{\pi} = r^2 \tag{9}$$

$$F_R(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ r^2 & \text{if } 0 \leq r \leq 1 \\ 1 & \text{if } 1 \geq r \end{cases} \tag{10}$$

The probability density function (PDF) is the derivative of its CDF. Hence:

$$f_R(r) = \begin{cases} 2r & \text{if } 0 \leq r \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

+ Good!

4. Let p be the probability of getting a head, and k be the number of times we toss a coin until we get a head. This happens when we obtain tail $(k-1)$ times and a head on the kth attempt. Hence, $P(X=k) = (1-p)^{k-1}p$. Since the coin is fair $p = 1-p = \frac{1}{2}$ and we have then $P(X=k) = \left(\frac{1}{2}\right)^k$. Hence,

$$E(X) = \sum_{k=0}^{\infty} kp^k = \sum_{k=0}^{\infty} k\left(\frac{1}{2}\right)^k \tag{12}$$

$$.\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}, \text{since it is a geometric series with } 0 < p < 1. \tag{13}$$

We derive both sides of equation 13 with respect to k obtaining:

$$\sum_{k=0}^{\infty} kp^{k-1} = \frac{1}{(1-p)^2} \tag{14}$$

Finally, we multiply each side of equation 14 with p obtaining:

$$\sum_{k=0}^{\infty} kp^k = \frac{p}{(1-p)^2} = \frac{p}{p^2} = \frac{1}{p} \tag{15}$$

Comparing equations 12 and 15, we notice that:

$$E(X) = \sum_{k=0}^{\infty} kp^k = \frac{1}{p} \tag{16}$$

Finally, we substitute $p = \frac{1}{2}$ in equation 16 obtaining: $E(X) = 2$.

+

5. (a) Definitions:

   - Let $\alpha > 0$ be fixed.
   - Let $X_1, X_2, ..., X_n \sim$ I.I.D. Bernoulli(p).
   - Interval: $I_n = [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]$, where $\varepsilon_n = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$.
   - Sample mean:

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n \tag{17}$$

$$\begin{aligned}
E[\bar{X}_n] &= E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] && \text{definition of } \bar{X}_n \\
&= \frac{1}{n}E\Big[\sum_{i=1}^{n} X_i\Big] && \text{linearity of E operator} \\
&= \frac{1}{n}\sum_{i=1}^{n} E\Big[X_i\Big] && \text{linearity of E operator} \\
&= \frac{1}{n}\sum_{i=1}^{n}[0.(1-p) + 1.p] && \text{definition of E applied to Bernoulli} \\
&= \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}np = p
\end{aligned} \tag{18}$$

Now we can apply Corollary 3.7 from the course notes related to Hoeffding's Inequality, where $b = 1$ and $a = 0$ because for a Bernoulli distributed $P(X_i) \in [a, b] = P(X_i) \in [0, 1] = 1$. We have also $\hat{p}_n = \bar{X}_n$ and $p = E[\bar{X}_n]$.

$$
\begin{aligned}
P(|\bar{X}_n - E[\bar{X}_n] \geq \varepsilon) \leq 2^{-\dfrac{2n\varepsilon^2}{(b-a)^2}} \qquad (19) \\
\implies P\left(|\hat{p}_n - p| \geq \varepsilon_n\right) \leq 2\exp\left(-2n\varepsilon_n^2\right) \qquad (20)
\end{aligned}
$$

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

$$
\begin{aligned}
2\exp\left(-2n\varepsilon_n^2\right) &= 2\exp\left(-2n\left(\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right)^2\right) \\
&= 2\exp\left(-2n \cdot \frac{1}{2n}\log\frac{2}{\alpha}\right) \\
&= 2\exp\left(-\log\frac{2}{\alpha}\right) \\
&= 2 \cdot \frac{\alpha}{2} \\
&= \alpha \qquad (21)
\end{aligned}
$$

Which makes the inequality:

$$
P\left(|\hat{p}_n - p| \geq \varepsilon_n\right) \leq \alpha \qquad (22)
$$

Thus:

$$
P\left(|\hat{p}_n - p| < \varepsilon_n\right) \geq 1 - \alpha \qquad (23)
$$

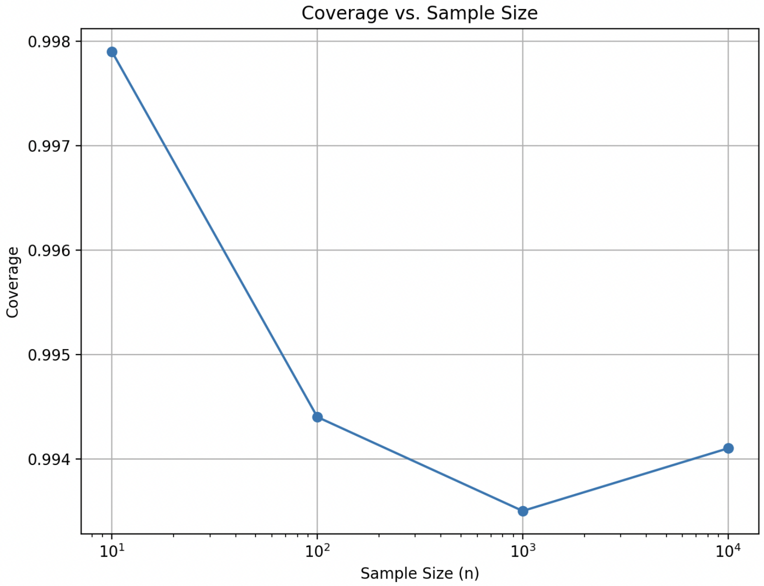
Which is equivalent to:

$$
P\left(p \in [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]\right) \geq 1 - \alpha \qquad (24)
$$

$$
P\left(p \in I_n\right) \geq 1 - \alpha \qquad (25)
$$

Therefore, the probability that p lies within the interval
$I_n = [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]$ is at least $(1 - \alpha)$.

(b) Following is a simulation study which estimates how often the confidence interval $I_n = [\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n]$ contains the true parameter p=0.4, known as the coverage probability. This was done for sample sizes n=10,100,1000,10000 with a significance level $\alpha = 0.05$



Python code:

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

p = 0.4
alpha = 0.05
sample_sizes = [10, 100, 1000, 10000]
num_simulations = 10000
ln_term = np.log(2 / alpha)

coverages = []
for n in sample_sizes:
    epsilon_n = np.sqrt(ln_term / (2 * n))
    #Binomial is a sum of independent Bernoulli
    samples = np.random.binomial(n, p, size=num_simulations)
    p_hats = samples / n
    lower_bounds = p_hats - epsilon_n
    upper_bounds = p_hats + epsilon_n
    coverage = np.mean((lower_bounds <= p) & (p <= upper_bounds))
    coverages.append(coverage)

plt.figure(figsize=(8, 6))
plt.plot(sample_sizes, coverages, marker=’o’)
plt.xscale(’log’)
plt.xlabel(’Sample Size (n)’)
plt.ylabel(’Coverage’)
plt.title(’Coverage vs. Sample Size’)
plt.grid(True)
plt.show()
```

The curve demonstrates that with increasing n, the confidence interval I_n more consistently contains p. Larger sample sizes lead to more accurate and reliable confidence intervals.

(c) The length of the confidence interval L_n is calculated by subtracting the lower limit from the upper limit:

$$L_n = (\hat{p}_n + \varepsilon_n) - (\hat{p}_n - \varepsilon_n) \tag{26}$$

$$\begin{aligned} L_n &= \hat{p}_n + \varepsilon_n - \hat{p}_n + \varepsilon_n \\ L_n &= 2\varepsilon_n \end{aligned} \tag{27}$$

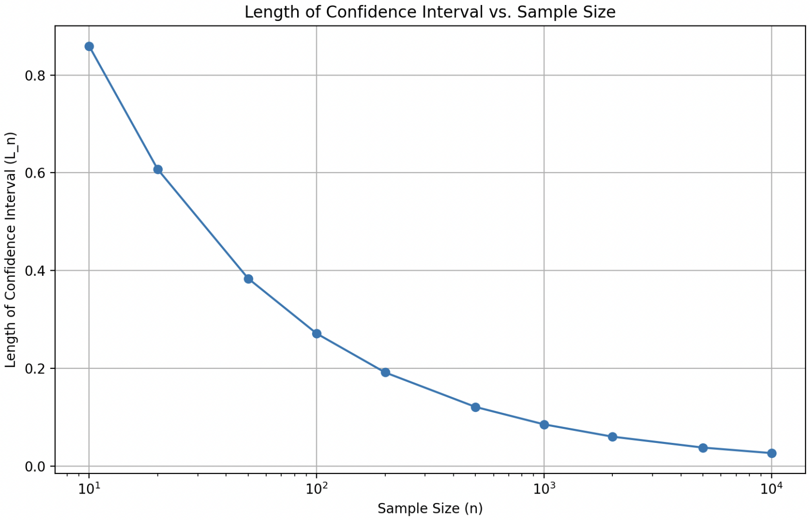
Substituting the Expression for ε_n

$$\varepsilon_n = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}$$

Substitute ε_n into L_n:

$$\begin{aligned} L_n &= 2\varepsilon_n \\ L_n &= 2\sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)} \end{aligned} \tag{28}$$

$$L_n = \sqrt{\frac{2}{n} \ln\left(\frac{2}{\alpha}\right)} \tag{29}$$



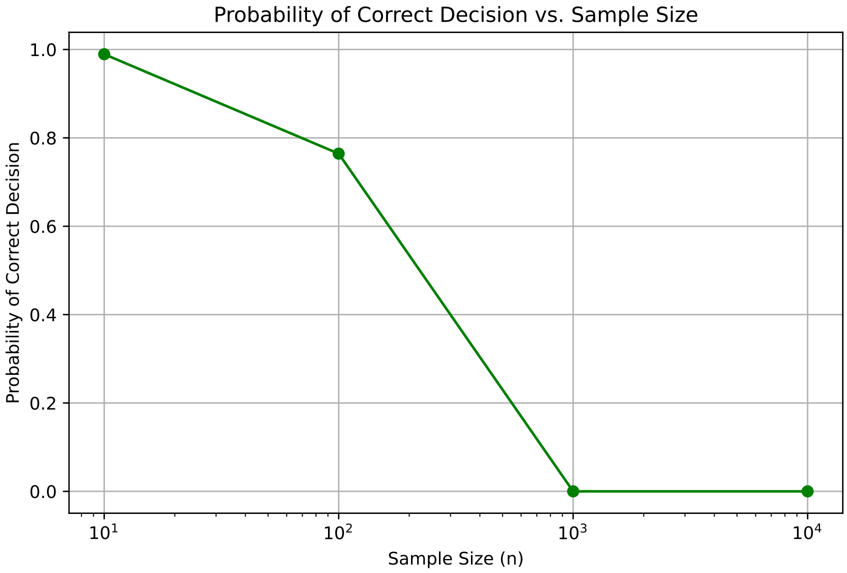
```python
import numpy as np
import matplotlib.pyplot as plt

alpha = 0.05
n_values = [10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000]
L_n = lambda n: np.sqrt((2 * np.log(2 / alpha)) / n)

L_n_values = []
for n in n_values:
    L_n_values.append(L_n(n))

plt.figure(figsize=(10, 6))
plt.plot(n_values, L_n_values, marker='o')
plt.xscale('log')
plt.xlabel('Sample Size (n)')
plt.ylabel('Length of Confidence Interval (L_n)')
plt.title('Length of Confidence Interval vs. Sample Size')
plt.grid(True)
plt.show()
```

(d) We need to plot the probability that a decision remains correct when the true probability of an event changes from 0.4 to 0.5. First we calculate the probability that the new mean falls within a confidence interval centered around the old probability by determining the interval using Hoeffding’s inequality. Then, we simulated 10,000 runs for each sample size (10, 100, 1000, and 10000) to see how sample size affects the correctness of the decision. The plot shows the decrease in probability of correct decision as the sample size increases because the interval becomes narrower and probability of new mean falling in the confidence interval of the old mean decreases.



Python code:

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

def simulate(n, old_p=0.4, new_p=0.5, alpha=0.05, num_of_simulations=10000):
    epsilon_n = np.sqrt(1 / (2 * n) * np.log(2 / alpha))
    num_of_correct_decisions = 0
    for _ in range(num_of_simulations):
        #Binomial is a sum of independent Bernoulli
        samples = np.random.binomial(1, new_p, n)
        p_hat_new = np.mean(samples)
        lower_bound = old_p - epsilon_n
        upper_bound = old_p + epsilon_n
        if lower_bound <= p_hat_new <= upper_bound:
            num_of_correct_decisions += 1
    correct_decision_probability = num_of_correct_decisions / num_of_simulations
    return correct_decision_probability

sample_sizes = [10, 100, 1000, 10000]
correct_decisions = [simulate(n) for n in sample_sizes]

plt.figure(figsize=(8, 5))
plt.plot(sample_sizes, correct_decisions, marker=’o’, linestyle=’-’, color=’g’)
plt.xscale(’log’)
plt.xlabel(’Sample Size (n)’)
plt.ylabel(’Probability of Correct Decision’)
plt.title(’Probability of Correct Decision vs. Sample Size’)
plt.grid(True)
plt.show()
```