Group Assignment 1

Introduction to Data Science 1MS041

Truls Karlsson, Nathalie Borglund, Noah Wassberg,
Andreas Larsson, Thant Zin Bo

September 2024

Problem 1

Two events E_1 and E_2 are independent if and only if

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

To show that A^c and B^c are independent we need to prove that

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

Since A and B are independent events we know that.

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

We can see that

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

and

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

which gives us

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

This can be rearranged into

$$
\begin{aligned}
&P(A^c \cap B^c) = -1 \cdot (1 - P(A)P(B) - P(A^c) - P(B^c)) \\
&P(A^c \cap B^c) = P(A^c) + P(B^c) + P(A)P(B) - 1 \\
&P(A^c \cap B^c) = (1 - P(A)) + (1 - P(B)) + (P(A)P(B) - 1) \\
&P(A^c \cap B^c) = 1 + (P(A)P(B) - P(A) - P(B)) \\
&P(A^c \cap B^c) = (1 - P(A))(1 - P(B)) \\
&P(A^c \cap B^c) = P(A^c) \cdot P(B^c)
\end{aligned}
$$

+

This shows that if A and B are independent A^c and B^c are also independent.

Problem 2

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

Problem 2a

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

The problem is of conditional probability using the formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

First, we define A as "at least two children have brown hair" and B as "at least one child has brown hair." Then, we calculate the probabilities of $P(A)$ and $P(B)$. Starting with B, we use the complement rule, which states that the probability of "at least one child has brown hair" is the same as the probability of $1-$ "no child has brown hair." To figure out the probability of "no child has brown hair," we use the equation:

$$P(k \text{ out of } n) = \binom{n}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{n-k}$$

The probability of at least one child having brown hair ($P(B)$) is:

$$P(B) = 1 - \binom{3}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{3-0} = 1 - \frac{27}{64} = \frac{37}{64}$$

Next, we calculate $P(A)$, where we redefine the problem as:

$$P(\text{at least two children have brown hair}) = \begin{aligned} &1 - P(\text{no child has brown hair}) \\ &- P(\text{exactly one child has brown hair}) \end{aligned}$$

$$P(A) = 1 - \frac{27}{64} - \binom{3}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{3-1} = 1 - \frac{27}{64} - \frac{27}{64} = \frac{10}{64}$$

Finally, we use the formula for conditional probability: +

$$P(A|B) = \frac{\frac{10}{64}}{\frac{37}{64}} = \frac{10}{37}$$

Answer: The probability of at least two children having brown hair, given that at least one child has brown hair, is $\frac{10}{37}$.

Problem 2b

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

This is a conditional probability problem where we want to calculate the probability of A given B. We define $A =$ *At least two children have brown hair* and $B =$ *The oldest child has brown hair.* Now the probability equation can be setup:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Since it is given that "The oldest child has brown hair" the $P(B) = 1$. We now focus on determining $P(A)$. The probability of at least two children having brown hair out of the two remaining children is the probability of one child having brown hair + the probability of both children having brown hair. For

this we setup an equation using combinatorics to pick the number of children $\binom{n}{k}$ and using the probabilities given in the exercise.
Note, it is also possible to take the complement approach and instead determine $P(A) = 1 -$ *neither child has brown hair.*

We define:
$C =$ *one out of two children have brown hair*
$D =$ *two out of two children have brown hair*
$E =$*k out of n children having brown hair*

$$\begin{aligned}
P(E) &= \binom{n}{k}\left(\frac{1}{4}\right)^{k}\left(\frac{3}{4}\right)^{n-k} \\
P(C) &= \binom{2}{1}\left(\frac{1}{4}\right)^{1}\left(\frac{3}{4}\right)^{2-1} = \frac{6}{16} \\
P(D) &= \binom{2}{2}\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{2-2} = \frac{1}{16} \\
P(A) &= P(C) + P(D) = \frac{6}{16} + \frac{1}{16} = \frac{7}{16}
\end{aligned}$$

Answer: The probability of at least two children having brown hair given that the oldest child has brown hair is $\frac{7}{16}$

Problem 3

Let Ω be the sample space,

$$\Omega = \{(x, y) \mid x^2 + y^2 \leq 1\}.$$

Ω represents the unit disc, where each point (x, y) is uniformly distributed and lies inside the disc with a radius of 1.

- The area of a circle with radius r is πr^2.
- The area of the unit disc with radius 1 is $\pi \cdot 1^2 = \pi$.

We know that the unit disc has a constant area of π and since the distribution is uniform we know that the joint PDF $f_{X,Y}(x, y)$ must be constant over the disc, and the total probability must sum to 1. Thus,

$$\int_{\Omega} f_{X,Y}(x, y)\, dA = 1.$$

Given that the area of the unit disc is π, we can solve $f_{X,Y}(x,y)$ as follows,

$$f_{X,Y}(x,y)\cdot\pi = 1 \quad\Rightarrow\quad f_{X,Y}(x,y) = \frac{1}{\pi}, \quad \text{for } x^2+y^2 \leq 1.$$

This means that the joint PDF is constant across the disc and has the value $\frac{1}{\pi}$.

To calculate the CDF for the radial distance $R=\sqrt{X^2+Y^2}$, we are searching for the probability that $R \leq r$, meaning that the point (X,Y) lies within the unit disc with radius 1. This can be written as,

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

Since the probability is proportional to the area of the smaller circle with radius r, we need to multiply that area by the joint PDF to ensure normalization.

$$F_R(r) = \frac{1}{\pi}\cdot\pi r^2 = r^2 \quad \text{for } 0 \leq r \leq 1.$$

The radius of a circle can never be less than 0,

$$F_R(r) = P(R \leq r) = 0, \quad r < 0.$$

The radius of the unit disc can never be more than 1,

$$F_R(r) = P(R \leq r) = 0, \quad r > 1.$$

Thus, the CDF is,

$$F_R(r) = \begin{cases} 0, & r < 0 \\ r^2, & 0 \leq r \leq 1, \\ \cancel{0}\,1, & \underline{\text{otherwise}}. \quad r > 1 \end{cases}$$

Next, we compute the PDF by taking the derivative of the CDF:

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

$$f_R(r) = \frac{d}{dr}F_R(r) = 2r, \quad 0 \leq r \leq 1$$

$$f_R(r) = \frac{d}{dr}F_R(r) = 0, \quad r < 0$$

$$f_R(r) = \frac{d}{dr}F_R(r) = 0, \quad r > 1$$

Thus, the PDF is:

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

Problem 4

Let

$$\Omega = \{w_i : w_i \in H, T\}$$

where the event H corresponds to the coin flip resulting in heads landing face up and the event T corresponds to the coin flip resulting in tails landing face up. Since the coin is fair

$$P(H) = 1/2$$

and

$$P(T) = 1/2$$

Let X be the number of tosses required for heads to appear. Since there are two possible outcomes, a constant probability of success and we are stopping after heads appears for the first time we know that X follows a geometric distribution. The formula for calculating the expected value of random variable V that follows the geometric distribution is

$$E[V] = 1/p$$

Where p is the probability of success in a trial.

Since our probability of success $P(H) = 1/2$ our p value is 1/2. Plugging this into the formula gives us

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

+

Problem 5

5.1

We have that $\alpha > 0$. Let

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

and

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

The confidence interval I_n is defined as:

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

We want to show that $P(p \in I_n) \geq 1 - \alpha$.

Using Hoeffding's inequality, which is defined as:

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

This can be simplified by noting that $X_1, \ldots, X_n$ follow a Bernoulli distribution, which means that each random variable X_i can take values from the set $\{0, 1\}$. Since 0 and 1 lie within the interval $[0, 1]$, we can describe the range of each X_i as being within this interval, or more simply:

$$[a, b] = [0, 1]$$

Plugging this and the definition of ϵ_n into Hoeffding's inequality gives us:

$$\begin{aligned} P\left(|\hat{p}_n - p| \geq \epsilon_n\right) &\leq 2\exp\left(\frac{-2n\left(\frac{1}{2n}\log\frac{2}{\alpha}\right)}{(1-0)^2}\right) \\ &\leq 2\exp\left(-\log\frac{2}{\alpha}\right) \\ &\leq 2 \cdot \frac{\alpha}{2} \\ &\leq \alpha \end{aligned}$$

Which gives us

$$P\left(|\hat{p}_n - p| \geq \epsilon_n\right) \geq 1 - \alpha$$

Which is equivalent to

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

5.2

Figure 1, the x-axis is scaled logarithmically to represent the sample sizes $n = 10, 100, 1000, 10000$. The y-axis represents the proportion of simulations in which the true parameter p is contained within the confidence interval. As the sample size increases, the coverage probability converges closer to the desired level, approaching $1 - \alpha = 0.95$.

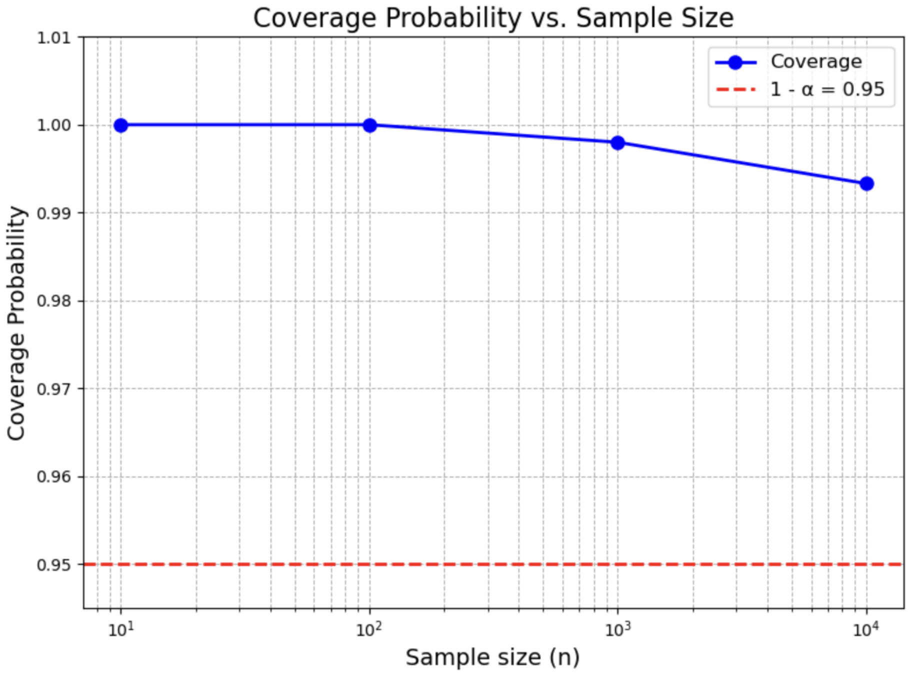

Figure 1: The plot shows the coverage probability of the confidence interval for a Bernoulli distribution with $p = 0.4$ as a function of the sample size n, using $\alpha = 0.05$.

5.3

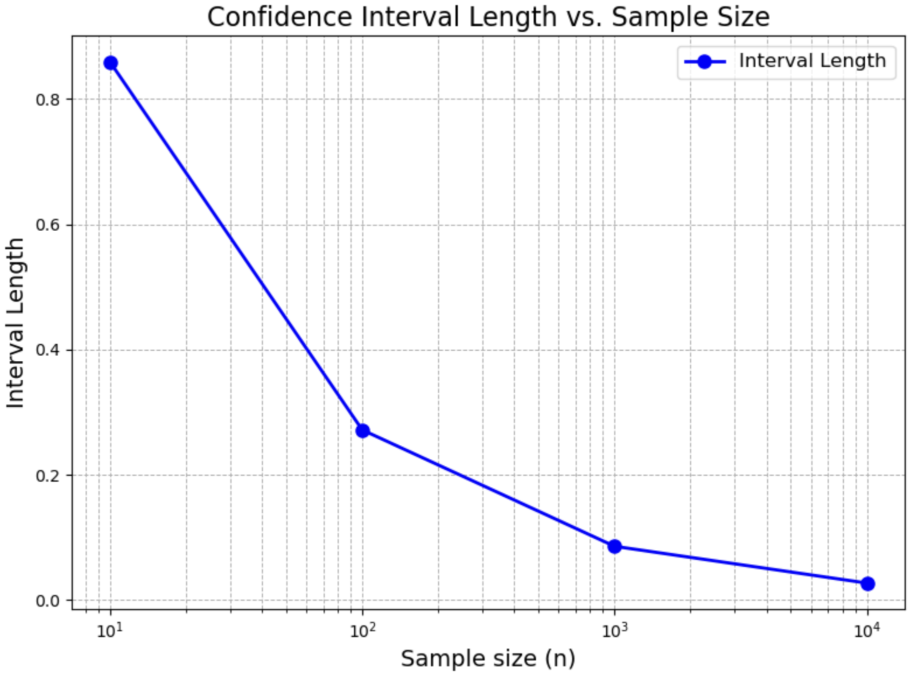

Figure 2: The plot shows the confidence interval length as a function of sample size n.

In figure 2, the interval length is computed using the formula $2\epsilon_n = 2\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$, where $\alpha = 0.05$. As the sample size n increases, the length of the confidence interval decreases, indicating that larger sample sizes yield more precise estimates. The x-axis is plotted on a logarithmic scale for better visualization.

5.4

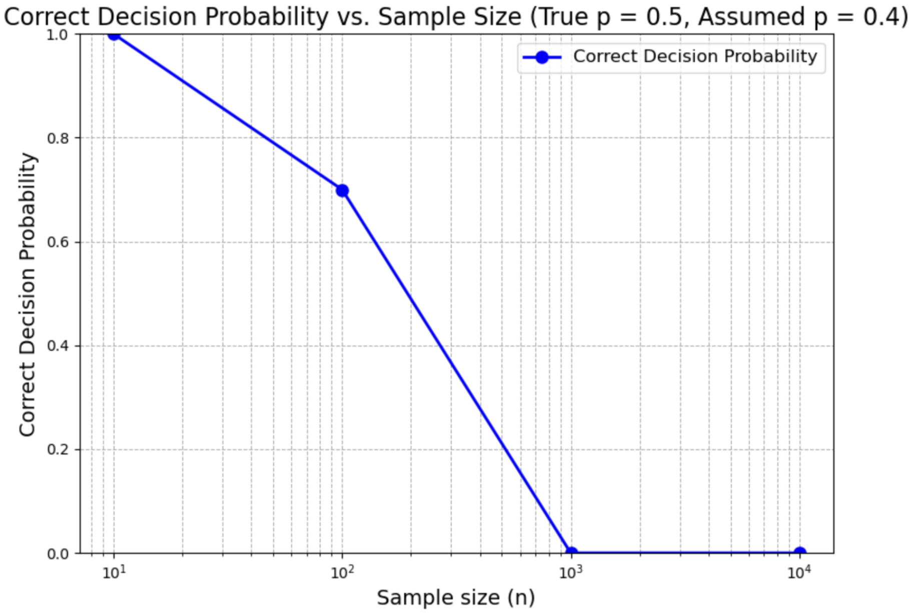

Figure 3: The plot shows the probability of making a correct decision as a function of sample size n.

In figure 3, the true proportion of people with the disease has changed to $p = 0.5$, but the confidence interval is calculated assuming $p = 0.4$. The plot shows how often the true proportion $p = 0.5$ lies within the confidence interval based on different sample sizes n. As the sample size increases, the probability of a correct decision improves. The x-axis is plotted on a logarithmic scale.