Homework 5: Statistics and probability

Casper Kristiansson

November 7, 2023

1 Probability

1.1 A

We have three boxes, one with 2 gold coins, one with 1 gold coin and one silver coin, and lastly one with 2 silver coins. We know that the first coin we pick up is a gold coin. This means the box with 2 silver coins is not selected. This means that in the two remaining boxes, there was a 50% chance of selecting the box with 2 gold coins and 50% for selecting the one with 1 gold and 1 silver coin. This means there is a 50% chance that the second coin is gold.

1.2 B

Bayes' theorem:

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

$$P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A)$$

$P(A) =$ Overall probability of breast cancer (0.3%)
$P(B) =$ Test result show positive
$P(A|B) =$ Given that the result shows positive for breast cancer and the patient has breast cancer
We can start by calculating the P(B)

$$\begin{aligned} P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A) \\ P(B) = 0.8 \times 0.003 + 0.09 \times (1 - 0.003) \\ P(B) = 0.09213 \end{aligned}$$

We then can use that to calculate the $P(A|H)$.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$
$$P(A|B) = \frac{0.8 \times 0.003}{0.09213}$$
$$P(A|B) \approx 0.02605$$

Meaning $\approx$ 2.6% of the people that get a positive result from breast cancer actually have breast cancer.

2 Confidence intervals and statistical significance

2.1 A

We can first calculate the mean value of all values which is 9.869. By using a confidence interval [5] with the most common level (95%), we can calculate that the confidence interval is 9.586 to 10.152. The calculations are derived from the standard error of the mean.

$$CI = 9.869 \pm (0.95 \times SEM)$$

Where SEM can be calculated by [3]:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

The actual calculations of the standard error of mean value were calculated using the library SciPy (scipy.stats.sem) [1].

2.2 B

To calculate the p-value we can use something called a two-sample t-test [4]. A t-test is used to compare the averages of two groups and determine if differences between them are more likely to arise from random chance. To calculate the p-value we can use a library called SciPy (scipy.stats.ttest_ind) [2] that allows the input of the two data sets. The function gives us a p-value of ≈ 0.34. Such a high value shows that there is no statistically significant difference between the two data sets' mean values. We know that values ≈ 0.05 show that the null hypothesis is rejected.

References

[1] scipy.stats.sem — scipy v1.11.3 manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.sem.html. (Accessed on 11/07/2023).

[2] scipy.stats.ttest_ind — scipy v1.11.3 manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html. (Accessed on 11/07/2023).

[3] Standard error of the mean vs. standard deviation: What's the difference? https://www.investopedia.com/ask/answers/042415/what-difference-between-standard-error-means-and-standard-deviation.asp#:~:text=SEM%20is%20calculated%20simply%20by,variability%20of%20the%20sample%20means. (Accessed on 11/07/2023).

[4] Two-sample t-test — introduction to statistics — jmp. https://www.jmp.com/en_se/statistics-knowledge-portal/t-test/two-sample-t-test.html. (Accessed on 11/07/2023).

[5] Simundic, A.-M., et al. Confidence interval. *Biochemia Medica 18*, 2 (2008), 154–161.