Using Bayesian Statistics to Model Binary Outcomes

Mario Alberto, Jatin Patel, and Dr. Sandra Hurtado Rúa

Abstract

Bayesian statistics

- Method for reallocating probability given data
- Improves the original assumptions

Our project

- Brief overview on Bayesian statistics
- We use simple examples to showcase its importance
- We show that models can be improved based on data that we gather.

Models used

- Determining the bias of a coin
- Determining the likelihood of cystic fibrosis given a positive test result.

Flipping Coins

- Assume we have 100 coins
- 99 are fair, 1 is double sided
- We flip the coin once and get heads
- What are the chances of having a fair coin?
- Intuitively: 99%
- What if we use Bayes' theorem?

Applying Bayes' theorem to Coins

T= $We\ have\ a\ fair\ coin$

$d = We\ flipped\ heads$

- $Initially : p(T) = 99\%$
- $p(d|T) = 0.5$
- $p(d) = 0.99 * 0.5 + 0.01 * 1 = .505$

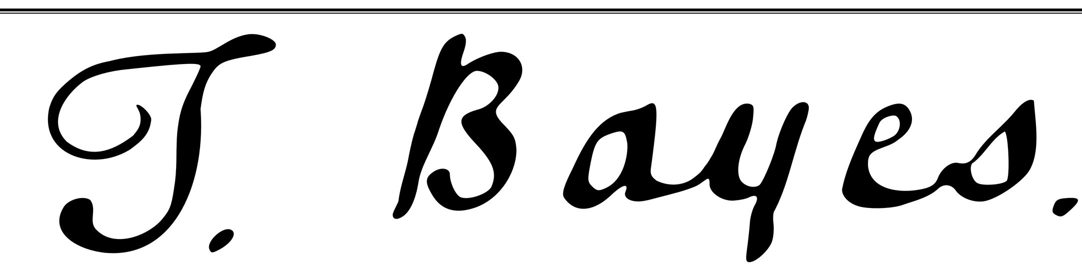
Apply Bayes' theorem:

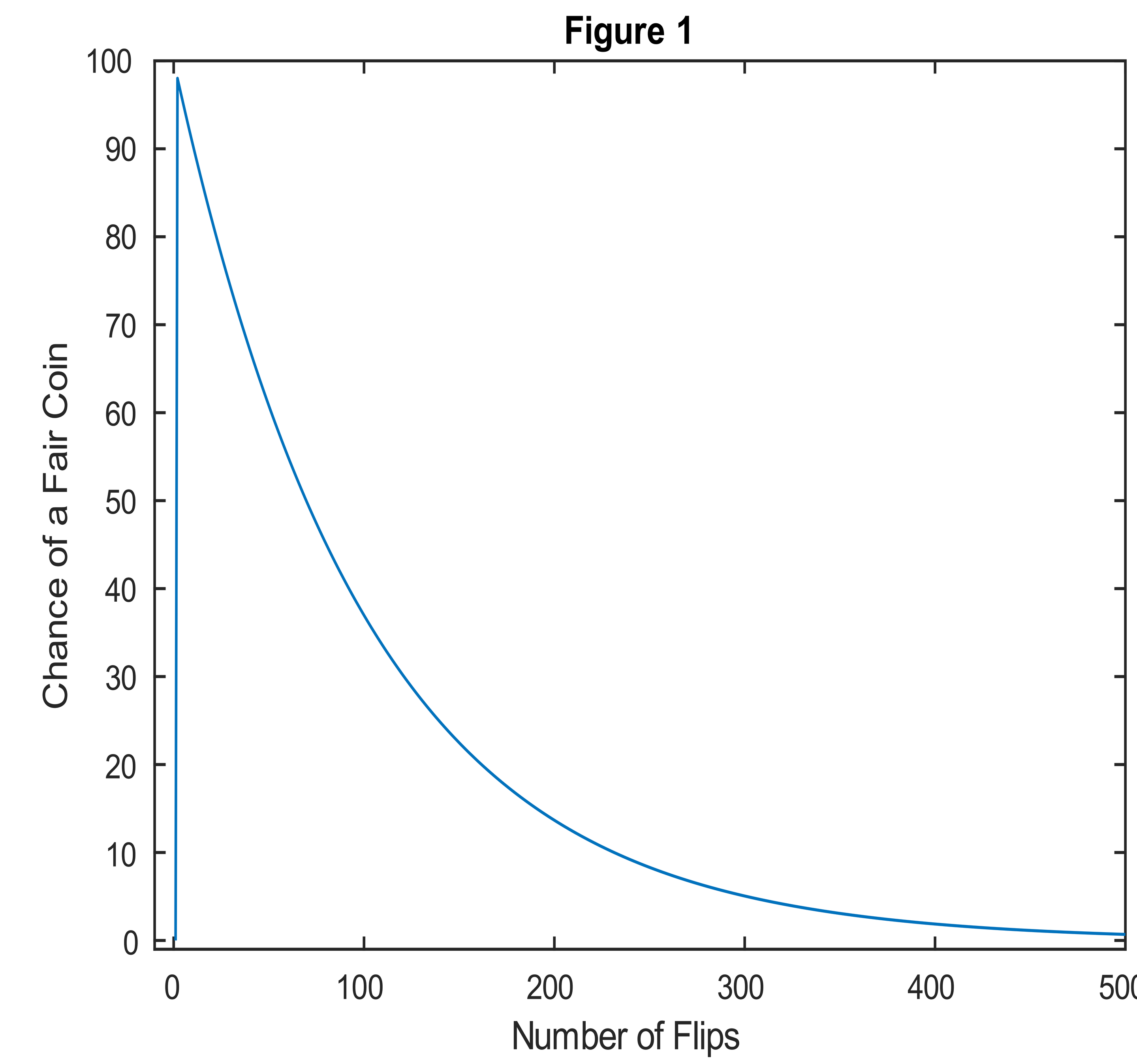
$$p(T|d) = \frac{0.5 * .99}{.505} = 0.98 = 98\%$$

- Does not match intuitive result.
- Flipping the coin again, we get heads. What is the new probability?
- Our *new* $p(T) = 0.98$

$$p(T|d) = \frac{0.5 * .98}{.505} = 0.97 = 97\%$$

The more heads that are obtained when flipping, the more likely it is that the coin is biased.





Bayes' Theorem

Likelihood

How likely is the data given that the theory is true?

Prior

How likely was the theory before observing the data?

$$p(T|d) = \frac{p(d|T) * p(T)}{p(d)}$$

Posterior

How likely is our theory given the data?

Marginal

How likely is the data to occur under *all* possible theories? How many times does the data occur?

Application to Cystic Fibrosis Screenings

T = *The patient has cystic fibrosis*

d = A positive test result

(Assay of trypsinogen concentrations)

- *Initially* $p(T) = 1\ in\ 4000$ *= 0.025%*
- $p(d|T) = 0.99$
- $p(d) = 2.5^{-4} * 0.99 + 0.99975 * 0.01 = 0.10245$

Apply Bayes' Theorem

$$p(T|d) = \frac{0.99 * 2.5 * 10^{-4}}{0.10245} = 0.0024 = 0.24\ \%$$

This illustrates why medical test with even the slightest rate of false positives are performed multiple times or in conjunction with another test

It is also important to remember that a patient exhibiting any symptoms of the disease would dramatically increase our *p(T)* value

References

Hammond, K. B., Abman, S. H., Sokol, R. J., & Accurso, F. J. (1991). Efficacy of Statewide Neonatal Screening for Cystic Fibrosis by Assay of Trypsinogen Concentrations. *New England Journal of Medicine, 325*(11), 769-774. doi:10.1056/nejm199109123251104S

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Amsterdam: Elsevier/AP.

Acknowledgments

We would like to thank Dr. Hurtado Rúa for her guidance and the Chose Ohio First program for giving us the opportunity to pursue this project.