Introduction to Bayesian learning

Let's start with an experiment. Student 1 will write down a number and then flip a coin. If flip is heads, they will honestly tell student 2 if the number is even or odd; if it's tails they will lie. Student 2 will then guess if the number is odd or even. Let $\theta$ be probability that student 2 correctly guesses whether the number is even or odd.

Before we start,

1. What's your best guess about $\theta$?

2. What's the probability that $\theta$ is greater than a half?

Now let's collect some data. Student 2 is correct in $I$ trials.

1. What's your best guess about $\theta$ now?

2. What's the probability that $\theta$ is greater than a half now?
Frequentist approach

- Definition: A procedure that quantifies uncertainty (p-value, confidence interval, etc) in terms of repeating the process that generated the data many times.

  - Definition of a sampling distribution:
    
    Let \( \hat{\theta} \) be some summary of the data used to estimate \( \theta \).
    
    If we repeatedly compute \( \hat{\theta} \) each time, the distribution of \( \hat{\theta} \) is the sampling dist.

  - Definition of a p-value:
    
    Say we have a rule to reject in null based on the data. If we repeat when the null is true, the p-value is the prob of incorrectly rejecting the null.

  - Definition of a confidence interval:
    
    A 95% CI is an interval (l,u) constructed from the data in a way so that if we repeat the 100th time \( \hat{\theta} \) will be in the interval 95% of the time.

  - Examples of repeatable data generation:

    Clinical trial

  - Sometimes it’s hard to imagine repeating the data generation:

    Climate change

- Parameters are fixed and unknown, only the data is random.

- Aims to be objective.
Follow-up questions to the guessing game:

- How would a frequentist answer these questions before collecting data?
  
  1. I don't have a guess
  2. This question is not well defined

- How would a frequentist answer these questions after collecting data?
  
  1. \( \hat{\theta} = \text{sample proportion} = \)
  2. This probability is not well defined

- Did you all have the same answer? Why or why not?
  
  All frequentists have \( \hat{\theta} = \frac{y}{n} \), so it is objective.

  In reality, humans will respond differently, so the results are subjective.
Bayesian approach

- Parameters $\theta = (\theta_1, \ldots, \theta_p)$ are fixed and unknown; analysis is conducted conditioned on data $y = (y_1, \ldots, y_n)$.

- Although the parameters are fixed (e.g., the probability of correctly guessing if the number is odd of even doesn’t change over time), we don’t know what they are.

- So Bayesians represent their uncertainty about parameters with probability distributions and treat them as random variables.

- The prior distribution, $p(\theta)$, is

  \[ \text{The prior distribution, } p(\theta) \text{, is uncertainty dist about } \theta \text{ before observing the data.} \]

  \[ \text{This person is confident } \theta \sim \frac{1}{2} \text{ This person allow for } \theta \text{ to be anything.} \]

- The likelihood, $p(y|\theta)$, is

  \[ \text{Distribution of the data given the parameters.} \]

  \[ \text{In this case } Y|\theta \sim \text{Binomial}(n, \theta) \text{ is reasonable (here } Y|x \text{ mean the dist of } Y \text{ given } \text{conditioned on } x) \]

- The posterior distribution, $p(\theta|y)$, is

  \[ \text{uncertainty dist after observing data.} \]

  \[ \text{posterior moves toward } \frac{1}{n} = 0.6. \]
Bayesian learning

- Bayesian learning combines past experience (prior) with new data (likelihood) in a mathematically coherent way (Bayes' Theorem) to form the current state of knowledge (posterior).

\[
p(\theta | y) \propto p(y | \theta) \, p(\theta)
\]

**posterior = likelihood \cdot prior**

- Bayes' Theorem:

- A Bayesian analysis is subjective (although there is a field called objective Bayes).

  Different priors lead to different results. Where to get priors?
  
  1. Literature review
  2. Expert opinion
  3. Small sample/pilot study

  4. If no information is available, pick a prior that assumes all are equally likely.
Advantages of Bayesian statistics

- Provides a formal mechanism to incorporate expert knowledge into the analysis.

  \[ \text{via the prior} \]

- Many find the subjective interpretation of a Bayesian analysis more intuitive.

  \( \text{We cannot make valid statements about uncertainty in the population!} \)
  
  - The probability that \( \theta > \frac{1}{2} \) is well defined
  
  - \( P(\theta \in (0.4)) \) is a valid interpretation of an interval
  
  - The probability that the null hypothesis is true is a valid probability

- Naturally provides a way to account for uncertainty in hard problems.

- It is an efficient way to combine data from several sources.

  \[ \text{Data source 1} \rightarrow \text{prior for the analysis} \rightarrow \text{Data source 2} \]
College football example

College football polls are an example of Bayesian learning.

- **Prior**
  
  Before the season, everybody has a personal ranking of teams

- **Likelihood**
  
  the data is the games that are played

- **Posterior**
  
  as data accrues we update our ranking. the current ranking is the posterior

- Early in the season...

  subjective: everyone has a different ranking

- By the end of the season...

  since everyone watches the same games, our rankings should converge.
BAC example

Let \( \theta \) be the true blood alcohol content (BAC) for a driver and \( Y \) be the reading from a breathalyzer. Driving with BAC > 0.08 is illegal. Experimentation shows that \( Y | \theta \sim N(\theta, \sigma^2) \) and that the measurement error standard deviation is \( \sigma = 0.005 \).

- Prior: Belief about the driver's BAC before giving the test.

- Likelihood: \( Y | \theta \sim N(\theta, 0.005^2) \)

- Posterior: Our uncertainty about the true BAC after giving the test.
BAC example

Below is a plot of the prior, likelihood, and posterior (prior*likelihood) for four examples. In all cases the observed BAC is $y = 0.09$.

1. Small measurement error, $\sigma = 0.005$, and diffuse prior, $\theta \sim N(0.8,0.1)$.
2. Large measurement error, $\sigma = 0.02$, and strong prior, $\theta \sim N(0.8,0.01)$.
3. Medium measurement error, $\sigma = 0.01$, and diffuse prior, $\theta \sim N(0.5,0.02)$.
4. Medium measurement error, $\sigma = 0.01$, and diffuse prior, $\theta \sim N(0.5,10)$.

Code is online at http://www4.stat.ncsu.edu/~reich/ST590/code/BAC.R.