GENERAL ASSEMBLY

BAYES' RULE AND INTRODUCTION TO STATISTICS

Matt Brems

Data Science Immersive, GA DC

LEARNING OBJECTIVES

- Derive and apply Bayes' Rule.
- Describe the relationship between probability and statistics.
- Understand the difference between descriptive statistics and inferential statistics.
- Identify the two fields of inferential statistics and the two most popular methods of parameter inference.

BAYES' RULE

BAYES' RULE

- Bayes' Rule relates $P(A|B)$ to $P(B|A)$.

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

BAYES' RULE

- Bayes' Rule relates $P(A|B)$ to $P(B|A)$.

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

$$P(A|B)P(B) = P(B|A)P(A)$$

BAYES' RULE

- Bayes' Rule relates $P(A|B)$ to $P(B|A)$.

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

$$P(A|B)P(B) = P(B|A)P(A)$$

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

BAYES' RULE

- Bayes' Rule relates $P(A|B)$ to $P(B|A)$.

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

$$P(A|B)P(B) = P(B|A)P(A)$$

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

- **<u>This will be very important later.</u>**

BREAKING DOWN BAYES' RULE

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

BREAKING DOWN BAYES' RULE

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

- $P(A)$ is the probability that A occurs given no supplemental information.

BREAKING DOWN BAYES' RULE

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

- $P(A)$ is the probability that A occurs given no supplemental information.
- $P(B|A)$ is the likelihood of seeing evidence (data) B assuming that A is true.

BREAKING DOWN BAYES' RULE

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

- $P(A)$ is the probability that A occurs given no supplemental information.
- $P(B|A)$ is the likelihood of seeing evidence (data) B assuming that A is true.
- $P(B)$ is what we scale $P(B|A)P(A)$ by to ensure we are only looking at A within the context of B occurring.

BREAKING DOWN BAYES' RULE

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

- $P(A)$ is the probability that A occurs given no supplemental information.
  - "Prior"
- $P(B|A)$ is the likelihood of seeing evidence (data) B assuming that A is true.
  - "Likelihood"
- $P(B)$ is what we scale $P(B|A)P(A)$ by to ensure we are only looking at A within the context of B occurring.
  - "Marginal Likelihood of B"

BAYES RULE APPLICATION

- Incorporating Context
  - iPhone, you text ‘radom.’
  - iPhone might correct to ‘random’ or ‘radon’ or leave as ‘radom,’ but which?

BAYES RULE APPLICATION

- Incorporating Context
  - iPhone, you text ‘radom.’
  - iPhone might correct to ‘random’ or ‘radon’ or leave as ‘radom,’ but which?

- In Bayesian statistics, often we let the data (or what we have observed) be y and our unknown or parameter of interest be θ.
  - Let ‘radom’ = y and we want to figure out the “truth,” or what you intended to text, labeled θ.

BAYES RULE APPLICATION

- $y =$ 'radom,' and suppose for simplicity that the three possibilities are $\theta =$ 'random,' 'radon,' or 'radom.'

- Let's find:
  - $P(\theta = random | y = radom)$
  - $P(\theta = radon | y = radom)$
  - $P(\theta = radom | y = radom)$

- Our thought process is that we'll find all three of these probabilities and then whichever probability is highest is the best θ and thus the one to which our iPhone should autocorrect.

BAYES RULE APPLICATION

- Recall:

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

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

- In order to get $P(\theta|y)$, we need $P(y|\theta)$, $P(\theta)$, and $P(y)$.

BAYES RULE APPLICATION

- Let's find:
  - $P(\theta_1 = random|y = radom)$
  - $P(\theta_2 = radon|y = radom)$
  - $P(\theta_3 = radom|y = radom)$

- We need:
  - $P(y|\theta_1)$, $P(y|\theta_2)$, and $P(y|\theta_3)$.
  - $P(\theta_1)$ $P(y|\theta_2)$, and $P(y|\theta_3)$.
  - $P(y)$.

- Brainstorm: how might we estimate these?

BAYES RULE APPLICATION

- From Google:

$\boldsymbol{\theta}$	$\boldsymbol{p(\theta)}$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

$\boldsymbol{\theta}$	$\boldsymbol{p(y = \text{“}radom\text{”}|\theta)}$
random	0.00193
radon	0.000143
radom	0.975

BAYES RULE APPLICATION

θ	$Prior: p(\theta)$	$Likelihood: p(y = \text{“radom”}\mid\theta)$	$Marginal\ Likelihood: p(y)$	$Posterior: P(\theta\mid y)$
random	7.60×10^{-5}	0.00193	$p(y)$	
radon	6.05×10^{-6}	0.000143	$p(y)$	
radom	3.12×10^{-7}	0.975	$p(y)$	

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

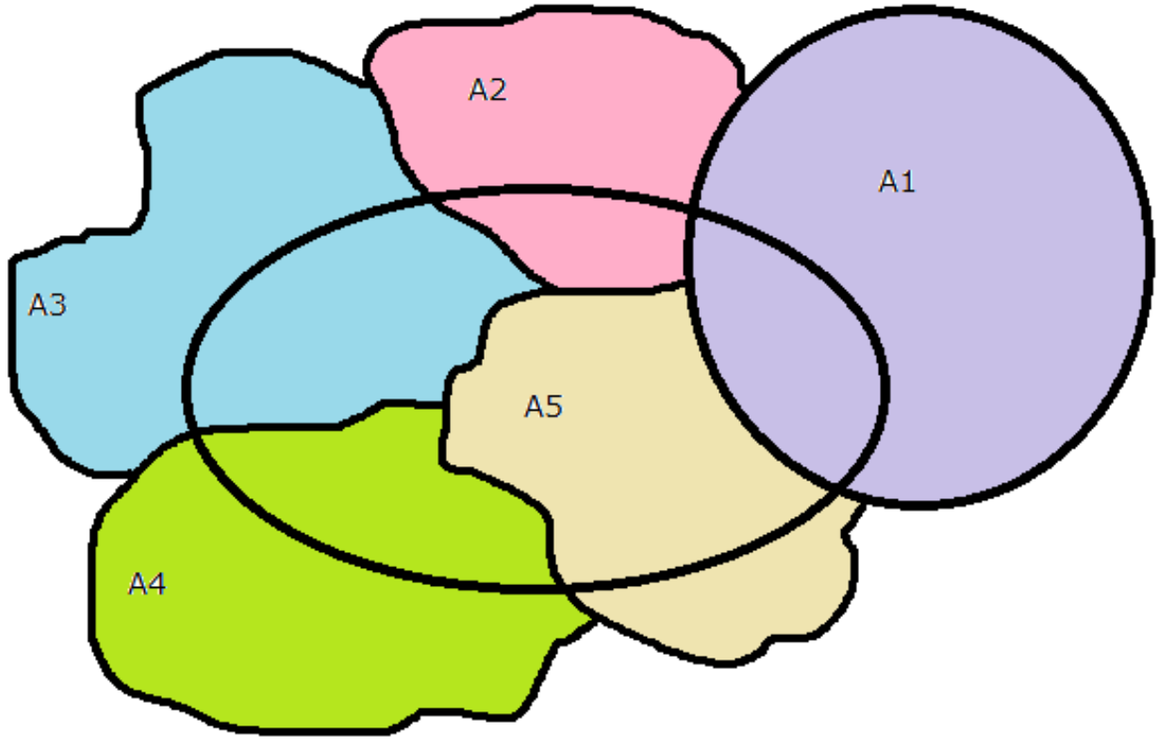
BAYES RULE APPLICATION

θ	$Prior: p(\theta)$	$Likelihood: p(y = \text{“radom”}\mid\theta)$	$Marginal\ Likelihood: p(y)$	$Posterior: P(\theta\mid y)$
random	7.60×10^{-5}	0.00193	$p(y)$	$1.47 \times 10^{-7}/p(y)$
radon	6.05×10^{-6}	0.000143	$p(y)$	$8.65 \times 10^{-10}/p(y)$
radom	3.12×10^{-7}	0.975	$p(y)$	$3.04 \times 10^{-7}/p(y)$

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

PROBABILITY RULES

- $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$
  - "Law of Total Probability"



- $P(y) = \sum_{\{i=1\}}^{n} P(y \cap \theta_i) = \sum_{\{i=1\}}^{3} P(y|\theta_i)P(\theta_i)$

BAYES RULE APPLICATION

θ	$Prior$: $p(\theta)$	$Likelihood$: $p(y = \text{“}radom\text{”}\mid\theta)$	$Marginal\ Likelihood$: $p(y)$	$Posterior$: $P(\theta\mid y)$
random	7.60×10^{-5}	0.00193	$1.47\times10^{-7} + 8.65\times10^{-10} + 3.04\times10^{-7} \approx 4.52\times10^{-7}$	$1.47 \times 10^{-7}/p(y)$
radon	6.05×10^{-6}	0.000143	$p(y) \approx 4.52\times10^{-7}$	$8.65 \times 10^{-10}/p(y)$
radom	3.12×10^{-7}	0.975	$p(y) \approx 4.52\times10^{-7}$	$3.04 \times 10^{-7}/p(y)$

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

BAYES RULE APPLICATION

θ	$Prior$: $p(\theta)$	$Likelihood$: $p(y = \text{“}radom\text{”}\mid\theta)$	$Marginal\ Likelihood$: $p(y)$	$Posterior$: $P(\theta\mid y)$
random	7.60×10^{-5}	0.00193	$1.47\times10^{-7} + 8.65\times10^{-10} + 3.04\times10^{-7} \approx 4.52\times10^{-7}$	$0.325 = 32.5\%$
radon	6.05×10^{-6}	0.000143	$p(y) \approx 4.52\times10^{-7}$	$0.002 = 0.2\%$
radom	3.12×10^{-7}	0.975	$p(y) \approx 4.52\times10^{-7}$	$0.673 = 67.3\%$

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

BAYES RULE APPLICATION

- Goal: Find posterior probability of parameter θ given our data or evidence y.
  - This is written as $P(\theta|y)$.

- Needed:
  - Prior probability of parameter θ.
  - Likelihood of data y given parameter θ.
  - Marginal likelihood of data y with no knowledge of parameter.*

BAYES RULE APPLICATION

- If your hypotheses are mutually exclusive and collectively exhaustive, the marginal likelihood is not necessary.

θ	***Prior*: $p(\theta)$**	***Likelihood*: $p(y = \text{“radom”}\mid\theta)$**	***Marginal Likelihood*: $p(y)$**	***Posterior*: $P(\theta\mid y)$**
random	7.60×10^{-5}	0.00193	$1.47\times10^{-7} + 8.65\times10^{-10} + 3.04\times10^{-7} \approx 4.52\times10^{-7}$	$1.47 \times 10^{-7}/p(y)$
radon	6.05×10^{-6}	0.000143	$p(y) \approx 4.52\times10^{-7}$	$8.65 \times 10^{-10}/p(y)$
radom	3.12×10^{-7}	0.975	$p(y) \approx 4.52\times10^{-7}$	$3.04 \times 10^{-7}/p(y)$

θ	***Prior*: $p(\theta)$**	***Likelihood*: $p(y = \text{“radom”}\mid\theta)$**	***Marginal Likelihood*: $p(y)$**	***Posterior*: $P(\theta\mid y)$**
random	7.60×10^{-5}	0.00193	$1.47\times10^{-7} + 8.65\times10^{-10} + 3.04\times10^{-7} \approx 4.52\times10^{-7}$	$0.325 = 32.5\%$
radon	6.05×10^{-6}	0.000143	$p(y) \approx 4.52\times10^{-7}$	$0.002 = 0.2\%$
radom	3.12×10^{-7}	0.975	$p(y) \approx 4.52\times10^{-7}$	$0.673 = 67.3\%$

WRAPPING UP FROM LAST TIME

AN ASIDE: PROBABILITY AND STATISTICS

- Sometimes it is less convenient to work with one particular event and more convenient to describe all possible outcomes.

AN ASIDE: PROBABILITY AND STATISTICS

- Sometimes it is less convenient to work with one particular event and more convenient to describe all possible outcomes.

- For example, rather than finding the probability that someone has an IQ of exactly 100, we might be interested in looking at all possible IQ scores and how frequently we observe each IQ value.

AN ASIDE: PROBABILITY AND STATISTICS

- Sometimes it is less convenient to work with one particular event and more convenient to describe all possible outcomes.

- For example, rather than finding the probability that someone has an IQ of exactly 100, we might be interested in looking at all possible IQ scores and how frequently we observe each IQ value.

- Recall: a distribution is the set of all possible values of a variable and how frequently the variable takes on each value.

AN ASIDE: PROBABILITY AND STATISTICS

- Every variable has a distribution.
  - The probabilities associated with each distribution must add up to 1.
  - The probability of any variable's value can never be negative.

AN ASIDE: PROBABILITY AND STATISTICS

- Every variable has a distribution.
  - The probabilities associated with each distribution must add up to 1.
  - The probability of any variable's value can never be negative.

- Let X = IQ score.
  - We might say that X follows a Normal distribution with mean 100 and standard deviation 15.

AN ASIDE: PROBABILITY AND STATISTICS

- Every variable has a distribution.
  - The probabilities associated with each distribution must add up to 1.
  - The probability of any variable's value can never be negative.

- Let X = IQ score.
  - We might say that X follows a Normal distribution with mean 100 and standard deviation 15.

- Now let Y = time it takes all American workers to get to work.
  - What do we do here?

AN ASIDE: PROBABILITY AND STATISTICS

- This gets to the relationship between probability and statistics.

AN ASIDE: PROBABILITY AND STATISTICS

- This gets to the relationship between probability and statistics.
- In probability, we know the values of these parameters (measures of a population) and can thus completely define the probability distribution.
- In statistics, we don't know the values of these parameters, so we have to estimate them.

AN ASIDE: PROBABILITY AND STATISTICS

- This gets to the relationship between probability and statistics.
- In probability, we know the values of these parameters (measures of a population) and can thus completely define the probability distribution.
- In statistics, we don't know the values of these parameters, so we have to estimate them.
- We gather a sample to learn about the population.
- We calculate statistics to learn about parameters.

RECAP

- "Tree of statistics."

- Probability is a building block to learning about statistics.
  - Samples help us to learn about populations.
  - Statistics help us to learn about parameters.

- Independence is a huge consideration that will depend on your use-case.

- Probability is complicated, but being familiar with the basics will go a long way in understanding how pieces fit together.