GENERAL ASSEMBLY

# INTRODUCTION TO PROBABILITY & BAYES' RULE

*Matt Brems, h/t Thomas Bayes*

*Data Science Immersive, GA DC*

# LEARNING OBJECTIVES

- Identify the three axioms of probability.
- Apply five basic probability rules.
- Derive and apply Bayes' Rule.
- Differentiate between frequentist and Bayesian statistics.
- Describe scenarios when Bayesian statistics is useful.

# OPENING

- What comes to mind when you hear “probability?”

# INTRODUCTION: DEFINITIONS & SETS

# DEFINITIONS

- Experiment: A procedure that can be repeated infinitely many times and has a well-defined set of outcomes.
- Event: Any collection of outcomes of an experiment.
- Sample Space: The set of all possible outcomes of an experiment, denoted $\mathcal{S}$.

# EXAMPLES

- Experiment: Flip a coin twice.
- Sample Space $\mathcal{S}$:
- Event:

- Experiment: Rolling a single die.
- Sample Space $\mathcal{S}$:
- Event:

# DEFINITIONS

- Set: A well-defined collection of distinct objects.
  - $\{Derek\ Jeter, \pi, ☺\}$
    - (Standing on the shoulders of Justin Gash for this one.)

- Element: An object that is a member of a set.
  - Derek Jeter
  - $\pi$
  - ☺

# SET OPERATIONS

- Union: $A \cup B = A \ or \ B$
- Intersection: $A \cap B = A \ and \ B$
- Example:
  - $A =$ even numbers between 1 and 10 $= \{2,4,6,8\}$
  - $B =$ prime numbers between 1 and 10$= \{2,3,5,7\}$
  - $A \cup B = ?$
  - $A \cap B = ?$

# SET OPERATIONS

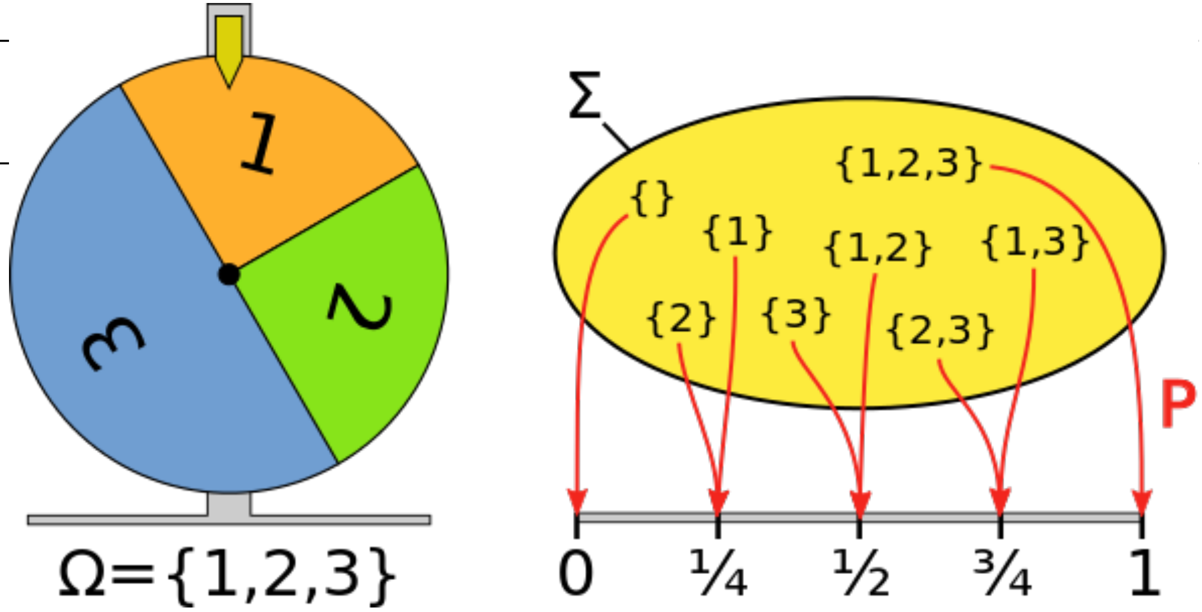
- Example:
  - $A = \{2,4,6,8\}$ & $B = \{2,3,5,7\}$
  - $A \cup B = \{2,4,6,8\} \cup \{2,3,5,7\} = \{2,3,4,5,6,7,8\}$
  - $A \cap B = \{2,4,6,8\} \cap \{2,3,5,7\} = \{2\}$

# BASICS OF PROBABILITY

# PROBABILITY BASICS

- Probability: $\mathrm{P}(\mathcal{S}, \mathcal{F}) \rightarrow [0,1]$
  - $\mathcal{S}$ is the sample space.
  - $\mathcal{F}$ is the "event space," or set of possible events.
  - $P$ is the probability function, mapping each event to the $[0,1]$ interval.

# PROBABILITY BASICS




- Probability: $P(\mathcal{S}, \mathcal{F}) \rightarrow [0,1]$
  - $\mathcal{S}$ is the sample space.
  - $\mathcal{F}$ is the "event space," or set of possible events.
  - $P$ is the probability function, mapping each event to the $[0,1]$ interval.

  - In more rigorous treatments of probability:
    - The sample space $\mathcal{S}$ is denoted by $\Omega$.
    - The "event space" is denoted either by $\mathcal{F}$ or $\Sigma$, is called a "sigma algebra" or "Borel field," and has a set of very specific properties.

# AXIOMS OF PROBABILITY (Kolmogorov Axioms)

- For any event $A$, $0 \leq P(A)$.
  - Nonnegativity.

- For the sample space $\mathcal{S}$, $P(\mathcal{S}) = 1$.
  - Unit measure.

- For mutually exclusive (or disjoint) $E_i$, $P\left(\bigcup_{\{i=1\}}^{\infty} E_i\right) = \sum_{\{i=1\}}^{\infty} P(E_i)$
  - Additivity.

- Probability must **ALWAYS** follow these three axioms.

# PROBABILITY RULES

- $P(\emptyset) = 0$
  - Note: $\emptyset$ indicates the "empty set," or the event containing zero outcomes from the experiment.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Venn diagrams can help to illustrate this – but remember that Venn diagrams are not proofs!
  - If $A$ and $B$ are disjoint, then $P(A \cap B) = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$.

# PROBABILITY RULES

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - Note: $A|B$ means “$A$ given $B$” or “$A$ conditional on the fact that $B$ happens.”
  - Example:
    - $A =$ roll a 2 $\Rightarrow P(A) = \frac{1}{6}$
    - $B =$ roll an even number $\Rightarrow P(B) = \frac{1}{2}$
    - $P(A \cap B) = P(\text{roll 2 and roll even number}) = \frac{1}{6}$
    - $P(A|B) =$ given that I roll an even, what is the probability of rolling a 2? $= \frac{1/6}{1/2} = \frac{1}{3}$

- $P(A \cap B) = P(A|B)P(B)$
  - We took the first rule on this slide, multiplied both sides of $P(B)$, and voila!

# PROBABILITY RULES

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

# PROBABILITY RULES – SUMMARY

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A|B)P(B)$
- $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$

# BAYES' RULE

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

# 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>**

# INTERPRETING PROBABILITY

# WHAT IS $P(A)$?

- We've talked a lot about probabilities of certain events, but what does this actually mean?

- There are two broad classes of probabilistic interpretations.

# TWO INTERPRETATIONS OF $P(A)$

- In the long run, how many times will $A$ occur relative to how many times we conduct our experiment?

# TWO INTERPRETATIONS OF $P(A)$

- In the long run, how many times will $A$ occur relative to how many times we conduct our experiment?

$$P(A) = \lim_{\#\ of\ exp's \to \infty} \frac{\#\ of\ times\ A\ occurs}{\#\ of\ experiments}$$

# TWO INTERPRETATIONS OF $P(A)$

- In the long run, how many times will $A$ occur relative to how many times we conduct our experiment?

$$P(A) = \lim_{\#\ of\ exp's \to \infty} \frac{\#\ of\ times\ A\ occurs}{\#\ of\ experiments}$$

$$P(heads) = \lim_{\#\ of\ coin\ tosses \to \infty} \frac{\#\ of\ heads}{\#\ of\ coin\ tosses}$$

# TWO INTERPRETATIONS OF $P(A)$

- In the long run, how many times will $A$ occur relative to how many times we conduct our experiment?

$$P(A) = \lim_{\#\, of\ exp's \to \infty} \frac{\#\, of\ times\ A\ occurs}{\#\, of\ experiments}$$

$$P(heads) = \lim_{\#\, of\ coin\ tosses \to \infty} \frac{\#\, of\ heads}{\#\, of\ coin\ tosses}$$

- This is called the **frequentist** interpretation of probability.

# TWO INTERPRETATIONS OF $P(A)$

- What is one's degree of belief in the statement $A$, possibly given evidence?

# TWO INTERPRETATIONS OF $P(A)$

- What is one's degree of belief in the statement $A$, possibly given evidence?

$$P(A) = \text{“How likely is it that } A \text{ is true?”}$$

# TWO INTERPRETATIONS OF $P(A)$

- What is one's degree of belief in the statement $A$, possibly given evidence?

$$P(A) = \text{"How likely is it that } A \text{ is true?"}$$

$$P(heads) = \text{"How likely is it that I flip a heads?"}$$

# TWO INTERPRETATIONS OF $P(A)$

- What is one's degree of belief in the statement $A$, possibly given evidence?

$$P(A) = \text{"How likely is it that } A \text{ is true?"}$$

$$P(heads) = \text{"How likely is it that I flip a heads?"}$$

- This is called the **Bayesian** interpretation of probability.

# TWO INTERPRETATIONS OF $P(A)$

- Frequentist inference and Bayesian inference have different interpretations, and these interpretations give rise to different methods of analysis.

    - Example: The average height of women at Ohio State, denoted $\mu$.
        - Frequentists treat $\mu$ as fixed: $\mu = 64 \text{ inches}$
        - Bayesians treat $\mu$ as a parameter with a distribution: $\mu \sim N(64, 2)$

    - Example: 95% confidence/credible interval
        - Frequentist interval: “I am 95% confident $\mu$ is in between 60 and 68 inches.”
        - Bayesian (credible) interval: “There is a 95% chance $\mu$ is in between 60 and 68 inches.”

# TWO INTERPRETATIONS OF $P(A)$

- Certain methods can only work relying on either frequentism or Bayesianism, but in cases where either interpretation works, your results should differ by only a negligible amount.

# THROWBACK: $p$-values

- Recall from "Stats Bomb Day" that the $p$-value for a particular experiment is "the probability that your random variable takes on a more extreme value than the one you just observed based on your data if the experiment were repeated, assuming the null hypothesis is true."

- $P(X > x|H_0 \text{ true})$
  - $X$ is the random variable.
  - $x$ is the value of your sample statistic based on the data from your experiment.
  - $H_0$ is the null hypothesis.

# THROWBACK: $p$-values

- $P(X > x|H_0 \text{ true})$
  - $X$ is the random variable.
  - $x$ is the value of your sample statistic based on the data from your experiment.
  - $H_0$ is the null hypothesis.

- This is related to, but not the same as, $P(X = x|H_0 \text{ true})$.
  - Since $x$ is the statistic based on your data, you can <u>roughly</u> think of this as $P(\text{data}|H_0 \text{ true})$.

- However, is this what we *really* want?

# THROWBACK: $p$-values

- $P(X > x | H_0 \text{ true})$
  - $X$ is the random variable.
  - $x$ is the value of your sample statistic based on the data from your experiment.
  - $H_0$ is the null hypothesis.

- This is related to, but not the same as, $P(X = x | H_0 \text{ true})$.
  - Since $x$ is the statistic based on your data, you can roughly think of this as $P(\text{data} | H_0 \text{ true})$.

- However, is this what we *really* want?
  - Wouldn't it be great if we could estimate $P(H_0 \text{ true} | \text{data})$?

# FREQUENTIST vs. BAYESIAN

- Frequentist
  - Pro: More widely understood.
  - Pro: Objective
  - Con: Relies on theoretically infinite number of experiments.
  - Con: Often does not work well with small sample sizes $n$.
  - Con: Cannot easily estimate $P(H_0 \text{ true}|\text{data})$

- Bayesian
  - Con: Less widely understood.
  - Con: "Subjective"
  - Pro: Does not rely on infinite experiments.
  - Pro: Works well even with small sample sizes $n$.
  - Pro: Can estimate $P(H_0 \text{ true}|\text{data})$

# BAYESIAN STATISTICS

# 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 $B$"

## So ***why*** do we do Bayesian statistics?

- Incorporating Context
    - iPhone, you text 'radom.'
    - iPhone might correct to 'random' or 'radon' or leave as 'radom,' but which?
        - "Bayesian Data Analysis," Gelman et al., 3rd Edition
- Sequential Updating with New Evidence
    - $P(\text{terror attack})$
    - $P(\text{terror attack} \mid 1 \text{ plane hits WTC})$
    - $P(\text{terror attack} \mid 2 \text{ planes hit WTC})$
        - "The Signal and The Noise," Nate Silver

## So ***how*** do we do Bayesian statistics?

- 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 $\theta$.
  - Let 'radom' = $y$ and we want to figure out the "truth," or what you intended to text, labeled $\theta$.

## So ***how*** do we do Bayesian statistics?

- $y =$ '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 <u>highest</u> is the best $\theta$ and thus the one to which our iPhone should autocorrect.

## So ***how*** do we do Bayesian statistics?

- 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 $P(y)$.

## So ***how*** do we do Bayesian statistics?

- 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?

## So ***how*** do we do Bayesian statistics?

- From Google:

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

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

# So ***how*** do we do Bayesian statistics?

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

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

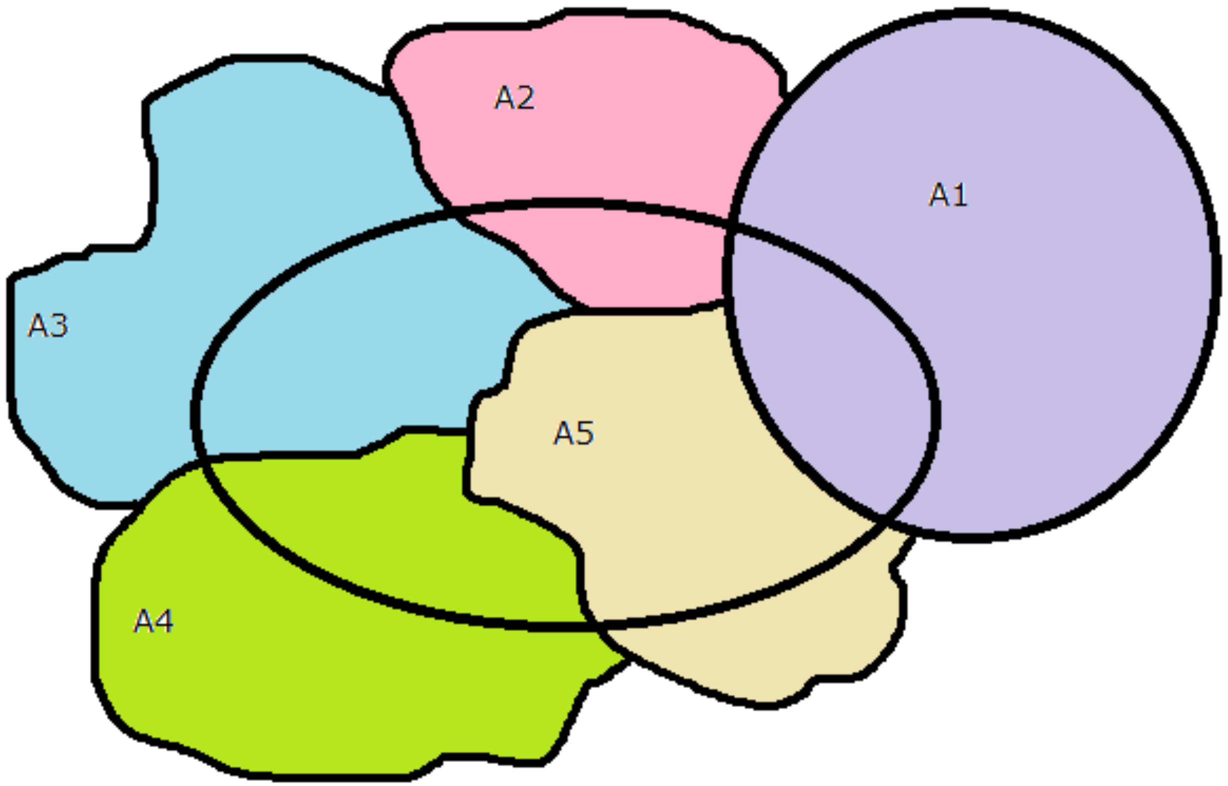
# So ***how*** do we do Bayesian statistics?

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




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

# So *how* do we do Bayesian statistics?

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

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

# So *how* do we do Bayesian statistics?

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

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

# So ***how*** do we do Bayesian statistics?

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

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

# So ***how*** do we do Bayesian statistics?

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

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

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