Bayesian Probability Estimation: The Case of Lifetime Batting Averages in Baseball

Adam Kapelner

Department of Mathematics, Queens College, CUNY

July 4, 2016
Macaulay Honors College at CUNY
Data Science Curriculum Module

Learning Objectives

- Understand Frequentist probability estimation for the probability parameter in the binomial model
- Understand Bayesian probability estimation for the probability parameter in the binomial model
- Build Bayesian credible intervals for the probability parameter in the binomial model
- Use a Bayesian model to forecast future data
- Use R to create empirical Bayes estimate of the binomial parameters

Materials Required

- Access to the Internet where readings will be linked
- The R statistical software which can be downloaded here

Readings

- Each of the concepts can be wikipediad or googled. These are the best readings...

Motivation: lifetime batting average (BA)

$$BA := \frac{\#\text{Lifetime Hits}}{\#\text{Lifetime At Bats}}$$

The ":=" means "defined as".

Motivation: lifetime batting average (BA)

$$BA := \frac{\#\text{Lifetime Hits}}{\#\text{Lifetime At Bats}}$$

The ":=" means "defined as".

Technicalities

- "At Bats" means every career plate appearance except walks, hit by pitch, bunts and others.

Motivation: lifetime batting average (BA)

$$BA := \frac{\#\text{Lifetime Hits}}{\#\text{Lifetime At Bats}}$$

The ":=" means "defined as".

Technicalities

- "At Bats" means every career plate appearance except walks, hit by pitch, bunts and others.
- "Hits" means any career instance where they reach first base via a fair ball or home run.

Motivation: lifetime batting average (BA)

$$BA := \frac{\#\text{Lifetime Hits}}{\#\text{Lifetime At Bats}}$$

The ":=" means "defined as".

Technicalities

- "At Bats" means every career plate appearance except walks, hit by pitch, bunts and others.
- "Hits" means any career instance where they reach first base via a fair ball or home run.

Note: detailed knowledge of baseball is not a prereq for this module, but many of the problems are motivated by baseball.

Lifetime Batting Average of a Single Player

We wish to answer three questions concerning the lifetime batting average:

Lifetime Batting Average of a Single Player

We wish to answer three questions concerning the lifetime batting average:

1. Given that the player has not yet retired, how do we estimate his BA?

Lifetime Batting Average of a Single Player

We wish to answer three questions concerning the lifetime batting average:

1. Given that the player has not yet retired, how do we estimate his BA?
2. Provide an interval of possible BA values and assign a probability that the BA is inside that interval

Lifetime Batting Average of a Single Player

We wish to answer three questions concerning the lifetime batting average:

1. Given that the player has not yet retired, how do we estimate his BA?
2. Provide an interval of possible BA values and assign a probability that the BA is inside that interval
3. Use historical data to improve the estimates (empirical Bayes).

Outline

- Random Variables, the Bernoulli and Binomial simple probability models

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model
- The Objective / Reference / Uninformative Prior

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model
- The Objective / Reference / Uninformative Prior
- Posterior Distribution, Bayesian Estimate, Credible Intervals

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model
- The Objective / Reference / Uninformative Prior
- Posterior Distribution, Bayesian Estimate, Credible Intervals
- Empirical Bayes

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model
- The Objective / Reference / Uninformative Prior
- Posterior Distribution, Bayesian Estimate, Credible Intervals
- Empirical Bayes
- Estimating Batting Averages with R

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number).

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number). This is important because numerical outcomes can be modeled where non-numerical outcomes cannot be.

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number). This is important because numerical outcomes can be modeled where non-numerical outcomes cannot be. (You cannot take the average of the outcomes "hit" and "walk" and "bunt").

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number). This is important because numerical outcomes can be modeled where non-numerical outcomes cannot be. (You cannot take the average of the outcomes "hit" and "walk" and "bunt"). Here is the r.v. X:

$$X : \Omega \to \mathbb{R}$$

The "support" of this random variable, $\mathbb{S}\text{upp}\left[X\right]$ is all possible values it can "spit out" i.e. the range of the function X.

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number). This is important because numerical outcomes can be modeled where non-numerical outcomes cannot be. (You cannot take the average of the outcomes "hit" and "walk" and "bunt"). Here is the r.v. X:

$$X : \Omega \to \mathbb{R}$$

The "support" of this random variable, $\mathbb{S}\text{upp}[X]$ is all possible values it can "spit out" i.e. the range of the function X.

Technically, probability is a "set function" taking in "events" (i.e. subsets of the universe of events and assigning an uncertainty value between 0 and 1:

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number). This is important because numerical outcomes can be modeled where non-numerical outcomes cannot be. (You cannot take the average of the outcomes "hit" and "walk" and "bunt"). Here is the r.v. X:

$$X : \Omega \to \mathbb{R}$$

The "support" of this random variable, $\mathbb{S}\text{upp}\,[X]$ is all possible values it can "spit out" i.e. the range of the function X.

Technically, probability is a "set function" taking in "events" (i.e. subsets of the universe of events and assigning an uncertainty value between 0 and 1:

$$\mathbb{P} : 2^{\Omega} \to [0, 1]$$

R.V. Function and the Probability Function

Technically random variables (r.v.'s) map the set of the "universe" of possible "outcomes" Ω to a numerical outcome (a real number). This is important because numerical outcomes can be modeled where non-numerical outcomes cannot be. (You cannot take the average of the outcomes "hit" and "walk" and "bunt"). Here is the r.v. X:

$$X : \Omega \to \mathbb{R}$$

The "support" of this random variable, $\mathbb{S}\text{upp}\left[X\right]$ is all possible values it can "spit out" i.e. the range of the function X.

Technically, probability is a "set function" taking in "events" (i.e. subsets of the universe of events and assigning an uncertainty value between 0 and 1:

$$\mathbb{P} : 2^{\Omega} \to [0, 1]$$

Technicalities

- The domain is only the powerset of the universe on discrete support random variables.
- Probability's mathematical definition is undisputed but its real-world definition is disputed by philosophers. This is beyond the scope of this module.
- We ignore outcomes, events and the universe Ω going forward...

Model for Lifetime Batting Average

Once again,

$$BA := \frac{\#\text{Hits}}{\#\text{At Bats}}$$

Model for Lifetime Batting Average

Once again,

$$BA := \frac{\#\text{Hits}}{\#\text{At Bats}}$$

Which can be thought of as:

$$BA := \frac{\mathbb{1}_{\text{Hit 1st at bat}} + \mathbb{1}_{\text{Hit 2nd at bat}} + \ldots + \mathbb{1}_{\text{Hit } i\text{th at bat}} + \ldots + \mathbb{1}_{\text{Hit } L\text{th at bat}}}{1 + 1 + \ldots + 1 + \ldots + 1}$$

Model for Lifetime Batting Average

Once again,

$$BA := \frac{\#\text{Hits}}{\#\text{At Bats}}$$

Which can be thought of as:

$$BA := \frac{\mathbb{1}_{\text{Hit 1st at bat}} + \mathbb{1}_{\text{Hit 2nd at bat}} + \ldots + \mathbb{1}_{\text{Hit } i\text{th at bat}} + \ldots + \mathbb{1}_{\text{Hit } L\text{th at bat}}}{1 + 1 + \ldots + 1 + \ldots + 1}$$

where the “indicator function” is defined as:

$$\mathbb{1}_{\text{A}} := \begin{cases} 1 & \text{if event } A \text{ occurred} \\ & \end{cases}$$

Model for Lifetime Batting Average

Once again,

$$BA := \frac{\#\text{Hits}}{\#\text{At Bats}}$$

Which can be thought of as:

$$BA := \frac{\mathbb{1}_{\text{Hit 1st at bat}} + \mathbb{1}_{\text{Hit 2nd at bat}} + \ldots + \mathbb{1}_{\text{Hit } i\text{th at bat}} + \ldots + \mathbb{1}_{\text{Hit } L\text{th at bat}}}{1 + 1 + \ldots + 1 + \ldots + 1}$$

where the “indicator function” is defined as:

$$\mathbb{1}_{\text{A}} := \begin{cases} 1 & \text{if event } A \text{ occurred} \\ 0 & \text{if event } A \text{ did } \textbf{not} \text{ occur.} \end{cases}$$

Model for Lifetime Batting Average

Once again,

$$BA := \frac{\#\text{Hits}}{\#\text{At Bats}}$$

Which can be thought of as:

$$BA := \frac{\mathbb{1}_{\text{Hit 1st at bat}} + \mathbb{1}_{\text{Hit 2nd at bat}} + \ldots + \mathbb{1}_{\text{Hit } i\text{th at bat}} + \ldots + \mathbb{1}_{\text{Hit } L\text{th at bat}}}{1 + 1 + \ldots + 1 + \ldots + 1}$$

where the "indicator function" is defined as:

$$\mathbb{1}_{\text{A}} := \begin{cases} 1 & \text{if event } A \text{ occurred} \\ 0 & \text{if event } A \text{ did } \textbf{not} \text{ occur.} \end{cases}$$

So if there are L lifetime at bats, then the lifetime BA can be written as:

Model for Lifetime Batting Average

Once again,

$$BA := \frac{\#\text{Hits}}{\#\text{At Bats}}$$

Which can be thought of as:

$$BA := \frac{\mathbb{1}_{\text{Hit 1st at bat}} + \mathbb{1}_{\text{Hit 2nd at bat}} + \ldots + \mathbb{1}_{\text{Hit } i\text{th at bat}} + \ldots + \mathbb{1}_{\text{Hit } L\text{th at bat}}}{1 + 1 + \ldots + 1 + \ldots + 1}$$

where the "indicator function" is defined as:

$$\mathbb{1}_{\text{A}} := \begin{cases} 1 & \text{if event } A \text{ occurred} \\ 0 & \text{if event } A \text{ did } \textbf{not} \text{ occur.} \end{cases}$$

So if there are L lifetime at bats, then the lifetime BA can be written as:

$$BA := \frac{1}{L} \sum_{i=1}^{L} \mathbb{1}_{\text{Hit } i\text{th at bat}}$$

The Bernoulli R.V.

Let's meditate on what $\mathbb{1}_{\text{Hit } i\text{th at bat}}$ really is. First of all, it can be 1 if the player got a hit on the *i*th at bat and 0 if not. And the outcome is random.

The Bernoulli R.V.

Let's meditate on what $\mathbb{1}_{\text{Hit } i\text{th at bat}}$ really is. First of all, it can be 1 if the player got a hit on the ith at bat and 0 if not. And the outcome is random. Let's call it X and call the probability of the hit θ. Thus,

$$X = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1 - \theta \end{cases}$$

The Bernoulli R.V.

Let's meditate on what $\mathbb{1}_{\text{Hit } i\text{th at bat}}$ really is. First of all, it can be 1 if the player got a hit on the ith at bat and 0 if not. And the outcome is random. Let's call it X and call the probability of the hit θ. Thus,

$$X = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1-\theta \end{cases}$$

This situation comes up a whole lot: one random "trial" or "experiment" and we wish to "count" if the event of interest happened.

The Bernoulli R.V.

Let's meditate on what $\mathbb{1}_{\text{Hit } i\text{th at bat}}$ really is. First of all, it can be 1 if the player got a hit on the ith at bat and 0 if not. And the outcome is random. Let's call it X and call the probability of the hit θ. Thus,

$$X = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1 - \theta \end{cases}$$

This situation comes up a whole lot: one random "trial" or "experiment" and we wish to "count" if the event of interest happened. This is called the "bernoulli r.v." and it is denoted via:

$$X \sim \mathrm{Bernoulli}\,(\theta)$$

where the $\sim$ means distributed as and the $\mathrm{Bernoulli}\,(\theta)$ is just shorthand for the cases "1 w.p. θ and 0 otherwise".

The Bernoulli R.V.

Let's meditate on what $\mathbb{1}_{\text{Hit } i\text{th at bat}}$ really is. First of all, it can be 1 if the player got a hit on the ith at bat and 0 if not. And the outcome is random. Let's call it X and call the probability of the hit θ. Thus,

$$X = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1 - \theta \end{cases}$$

This situation comes up a whole lot: one random "trial" or "experiment" and we wish to "count" if the event of interest happened. This is called the "bernoulli r.v." and it is denoted via:

$$X \sim \text{Bernoulli}(\theta)$$

where the $\sim$ means distributed as and the $\text{Bernoulli}(\theta)$ is just shorthand for the cases "1 w.p. θ and 0 otherwise".

What is X exactly? It's a "probability model" for a future outcome, x. Note the capital letter and lowercase letter. The future outcome is also called a "realization" because the model gets "realized" (i.e. made to be real) when the random trial is completed.

The Bernoulli R.V.

Let's meditate on what $\mathbb{1}_{\text{Hit } i\text{th at bat}}$ really is. First of all, it can be 1 if the player got a hit on the ith at bat and 0 if not. And the outcome is random. Let's call it X and call the probability of the hit θ. Thus,

$$X = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1 - \theta \end{cases}$$

This situation comes up a whole lot: one random "trial" or "experiment" and we wish to "count" if the event of interest happened. This is called the "bernoulli r.v." and it is denoted via:

$$X \sim \text{Bernoulli}\,(\theta)$$

where the $\sim$ means distributed as and the $\text{Bernoulli}\,(\theta)$ is just shorthand for the cases "1 w.p. θ and 0 otherwise".

What is X exactly? It's a "probability model" for a future outcome, x. Note the capital letter and lowercase letter. The future outcome is also called a "realization" because the model gets "realized" (i.e. made to be real) when the random trial is completed.

What is $\mathbb{S}\text{upp}\,[X]$? Since the only two legal things that could happen are the hit (1) and not a hit (0), $\mathbb{S}\text{upp}\,[X] = \{0, 1\}$.

Probability Mass Function (PMF)

The r.v. can be entered into the probability function to query the outcome of the experiment. For instance, what is:

$$\mathbb{P}(X = 1) \quad =$$

Probability Mass Function (PMF)

The r.v. can be entered into the probability function to query the outcome of the experiment. For instance, what is:

$$\begin{aligned} \mathbb{P}(X = 1) &= \theta \\ \mathbb{P}(X = 0) &= \end{aligned}$$

Probability Mass Function (PMF)

The r.v. can be entered into the probability function to query the outcome of the experiment. For instance, what is:

$$\begin{aligned} \mathbb{P}(X=1) &= \theta \\ \mathbb{P}(X=0) &= 1-\theta \end{aligned}$$

These may seem like simple questions, but ponder the following, what if we index the realization by the free variable x and ask the question.

Probability Mass Function (PMF)

The r.v. can be entered into the probability function to query the outcome of the experiment. For instance, what is:

$$\begin{aligned} \mathbb{P}(X=1) &= \theta \\ \mathbb{P}(X=0) &= 1-\theta \end{aligned}$$

These may seem like simple questions, but ponder the following, what if we index the realization by the free variable x and ask the question.

$$\mathbb{P}(X=x) = \theta^x(1-\theta)^x := \mathbb{P}(X)$$

Probability Mass Function (PMF)

The r.v. can be entered into the probability function to query the outcome of the experiment. For instance, what is:

$$\begin{aligned} \mathbb{P}(X = 1) &= \theta \\ \mathbb{P}(X = 0) &= 1 - \theta \end{aligned}$$

These may seem like simple questions, but ponder the following, what if we index the realization by the free variable x and ask the question.

$$\mathbb{P}(X = x) = \theta^x (1 - \theta)^x := \mathbb{P}(X)$$

This is an important function, it is called the "probability mass function" (PMF) and we'll denote it $\mathbb{P}(X)$. We will also call this its "density". It has the property that $\sum_{x \in \mathbb{S}\mathrm{upp}[X]} \mathbb{P}(x) = 1$.

Probability Mass Function (PMF)

The r.v. can be entered into the probability function to query the outcome of the experiment. For instance, what is:

$$\begin{aligned} \mathbb{P}(X=1) &= \theta \\ \mathbb{P}(X=0) &= 1-\theta \end{aligned}$$

These may seem like simple questions, but ponder the following, what if we index the realization by the free variable x and ask the question.

$$\mathbb{P}(X=x) = \theta^x(1-\theta)^x := \mathbb{P}(X)$$

This is an important function, it is called the "probability mass function" (PMF) and we'll denote it $\mathbb{P}(X)$. We will also call this its "density". It has the property that $\sum_{x \in \mathbb{S}\text{upp}[X]} \mathbb{P}(x) = 1$.

Technicalities

- The domain of the probability function is a subset of Ω, so technically the first query $\mathbb{P}(X=1)$ is an "abuse of notation" which is "shorthand" for $\mathbb{P}(\{\omega \in \Omega : X(\omega) = 1\}) = \mathbb{P}(\{\omega = \text{Hit}\}) = \theta$.
- x can take on only those values $\in \mathbb{S}\text{upp}[X]$. In the above, any value $\notin \mathbb{S}\text{upp}[X]$ gets probability 0. So e.g. $\mathbb{P}(X=17) = 0$ and $\mathbb{P}\left(X = \frac{2}{3}\right) = 0$, etc.
- Densities are only available for continuous r.v.'s (which we will see later in this module)

Detour: Continuous Random Variables

If a r.v. (not the bernoulli r.v.) has a support that is not countably infinite (e.g. $\{1, 2, \ldots\}$),

Detour: Continuous Random Variables

If a r.v. (not the bernoulli r.v.) has a support that is not countably infinite (e.g. $\{1, 2, \ldots\}$), then that r.v. is called "continuous".

Detour: Continuous Random Variables

If a r.v. (not the bernoulli r.v.) has a support that is not countably infinite (e.g. $\{1, 2, \ldots\}$), then that r.v. is called "continuous". A few properties:

- These r.v.'s do not have PMF's, they have "probability densitity functions" (PDF's) which we will also denote $\mathbb{P}(X)$.

Detour: Continuous Random Variables

If a r.v. (not the bernoulli r.v.) has a support that is not countably infinite (e.g. $\{1, 2, \ldots\}$), then that r.v. is called "continuous". A few properties:

- These r.v.'s do not have PMF's, they have "probability densitity functions" (PDF's) which we will also denote $\mathbb{P}(X)$.
- All probabilities of individual values is zero i.e. $\mathbb{P}(X = x) = 0$ for all values of x.

Detour: Continuous Random Variables

If a r.v. (not the bernoulli r.v.) has a support that is not countably infinite (e.g. $\{1, 2, \ldots\}$), then that r.v. is called "continuous". A few properties:

- These r.v.'s do not have PMF's, they have "probability densitity functions" (PDF's) which we will also denote $\mathbb{P}(X)$.
- All probabilities of individual values is zero i.e. $\mathbb{P}(X = x) = 0$ for all values of x.
- Probabilities are calculated only over a range via the following integration:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b \mathbb{P}(X)\, dx \quad \text{and thus} \quad \int_{x \in \mathbb{S}\text{upp}[X]} \mathbb{P}(X)\, dx = 1$$

Detour: Continuous Random Variables

If a r.v. (not the bernoulli r.v.) has a support that is not countably infinite (e.g. $\{1, 2, \ldots\}$), then that r.v. is called "continuous". A few properties:

- These r.v.'s do not have PMF's, they have "probability densitity functions" (PDF's) which we will also denote $\mathbb{P}(X)$.
- All probabilities of individual values is zero i.e. $\mathbb{P}(X = x) = 0$ for all values of x.
- Probabilities are calculated only over a range via the following integration:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b \mathbb{P}(X)\, dx \quad \text{and thus} \quad \int_{x \in \mathbb{S}\text{upp}[X]} \mathbb{P}(X)\, dx = 1$$

Technicalities

Usually, the PMF and PDF are given different notation $p(x)$ and $f(x)$. We are calling them both $\mathbb{P}(X)$ here for simplicity as they will be both used within the Bayesian paradigm interchangably — this is a slight abuse of notation as sometimes you will not know if the r.v. is discrete or continuous from its notation.

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support.

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics.

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x\, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathbb{S}\text{upp}[X]} x\, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)} \\ \mathbb{E}[X] &= \int_{x \in \mathbb{S}\text{upp}[X]} x\, \mathbb{P}(X = x)\, dx \quad \text{(for continuous r.v.'s)} \end{aligned}$$

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

$$\mathbb{E}[X] = \int_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \, dx \quad \text{(for continuous r.v.'s)}$$

For example, the expectation of the bernoulli r.v. would be:

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

$$\mathbb{E}[X] = \int_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \, dx \quad \text{(for continuous r.v.'s)}$$

For example, the expectation of the bernoulli r.v. would be:

$$\mathbb{E}[X] =$$

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

$$\mathbb{E}[X] = \int_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \, dx \quad \text{(for continuous r.v.'s)}$$

For example, the expectation of the bernoulli r.v. would be:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x \mathbb{P}(X = x) =$$

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\mathbb{S}\text{upp}[X]} x\,\mathbb{P}(X=x) \quad \text{(for non-continuous r.v.'s)}\\ \mathbb{E}[X] &= \int_{x\in\mathbb{S}\text{upp}[X]} x\,\mathbb{P}(X=x)\,dx \quad \text{(for continuous r.v.'s)}\end{aligned}$$

For example, the expectation of the bernoulli r.v. would be:

$$\mathbb{E}[X] = \sum_{x\in\mathbb{S}\text{upp}[X]} x\mathbb{P}(X=x) = \sum_{x\in\{0,1\}} x\theta^x(1-\theta)^{1-x}$$

Detour: Expectation of Random Variables

Random variables have a “balancing point” of their support. It is similar to how the “center-of-mass” is calculated in physics. We call this quantity the “expectation” or “mean” and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

$$\mathbb{E}[X] = \int_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \, dx \quad \text{(for continuous r.v.'s)}$$

For example, the expectation of the bernoulli r.v. would be:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathbb{S}\text{upp}[X]} x \mathbb{P}(X = x) = \sum_{x \in \{0,1\}} x \theta^x (1-\theta)^{1-x} \\ &= (0)\theta^0(1-\theta)^{1-0} + (1)\theta^1(1-\theta)^{1-1} =\end{aligned}$$

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)} \\
\mathbb{E}[X] &= \int_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \, dx \quad \text{(for continuous r.v.'s)}
\end{aligned}$$

For example, the expectation of the bernoulli r.v. would be:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathbb{S}\text{upp}[X]} x \mathbb{P}(X = x) = \sum_{x \in \{0,1\}} x \theta^x (1-\theta)^{1-x} \\
&= (0)\theta^0(1-\theta)^{1-0} + (1)\theta^1(1-\theta)^{1-1} = \boxed{\theta}
\end{aligned}$$

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x\, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

$$\mathbb{E}[X] = \int_{x \in \mathbb{S}\text{upp}[X]} x\, \mathbb{P}(X = x)\, dx \quad \text{(for continuous r.v.'s)}$$

For example, the expectation of the bernoulli r.v. would be:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathbb{S}\text{upp}[X]} x\mathbb{P}(X = x) = \sum_{x \in \{0,1\}} x\theta^x(1-\theta)^{1-x} \\ &= (0)\theta^0(1-\theta)^{1-0} + (1)\theta^1(1-\theta)^{1-1} = \boxed{\theta}\end{aligned}$$

Technicalities

In general, the expectation is not a parameter alone but a function of the parameter(s). Bernoulli is a special case in this regard.

Detour: Expectation of Random Variables

Random variables have a "balancing point" of their support. It is similar to how the "center-of-mass" is calculated in physics. We call this quantity the "expectation" or "mean" and denote it $\mathbb{E}[X]$ and calculate it via:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \quad \text{(for non-continuous r.v.'s)}$$

$$\mathbb{E}[X] = \int_{x \in \mathbb{S}\text{upp}[X]} x \, \mathbb{P}(X = x) \, dx \quad \text{(for continuous r.v.'s)}$$

For example, the expectation of the bernoulli r.v. would be:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathbb{S}\text{upp}[X]} x\mathbb{P}(X = x) = \sum_{x \in \{0,1\}} x\theta^x(1-\theta)^{1-x} \\ &= (0)\theta^0(1-\theta)^{1-0} + (1)\theta^1(1-\theta)^{1-1} = \boxed{\theta} \end{aligned}$$

Technicalities

In general, the expectation is not a parameter alone but a function of the parameter(s). Bernoulli is a special case in this regard.

Back to our regularly scheduled program...

Parameter

But conceptually, what is θ?

Parameter

But conceptually, what is θ? This quantity is known as a "parameter".

Parameter

But conceptually, what is θ? This quantity is known as a "parameter". The exact value of a parameters is unknowable with finite data.

Parameter

But conceptually, what is θ? This quantity is known as a "parameter". The exact value of a parameters is unknowable with finite data. In classical statistics, we provide "inference" which is loosely,

Parameter

But conceptually, what is θ? This quantity is known as a "parameter". The exact value of a parameters is unknowable with finite data. In classical statistics, we provide "inference" which is loosely,

1. estimate its value

Parameter

But conceptually, what is θ? This quantity is known as a "parameter". The exact value of a parameters is unknowable with finite data. In classical statistics, we provide "inference" which is loosely,

1. estimate its value
2. form a hypothesis of its value and then test the hypothesis (experimentation)

Parameter

But conceptually, what is θ? This quantity is known as a "parameter". The exact value of a parameters is unknowable with finite data. In classical statistics, we provide "inference" which is loosely,

1. estimate its value
2. form a hypothesis of its value and then test the hypothesis (experimentation)

   Using the Bayesian outlook, one can do the above and more,

3. find the density of θ

Parameter

But conceptually, what is θ? This quantity is known as a "parameter". The exact value of a parameters is unknowable with finite data. In classical statistics, we provide "inference" which is loosely,

1. estimate its value
2. form a hypothesis of its value and then test the hypothesis (experimentation)

   Using the Bayesian outlook, one can do the above and more,

3. find the density of θ
4. use historical data to improve the estimation

We will be covering 1, 3 and 4.

Our Parameter of Interest

But what is θ in our problem?

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \text{Bernoulli}(\theta) :=$$

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \mathrm{Bernoulli}(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ & \end{cases}$$

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \mathrm{Bernoulli}\,(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ 0 & \text{if he did not get a hit on his } i\text{th at bat} \end{cases}$$

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \text{Bernoulli}(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ 0 & \text{if he did not get a hit on his } i\text{th at bat} \end{cases}$$

This means θ is the true probability of the ith at bat.

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \text{Bernoulli}(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ 0 & \text{if he did not get a hit on his } i\text{th at bat} \end{cases}$$

This means θ is the true probability of the ith at bat. Since we only observe $x_i = 1$ or $x_i = 0$, you can intuitively see how θ can never be known

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \mathrm{Bernoulli}(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ 0 & \text{if he did not get a hit on his } i\text{th at bat} \end{cases}$$

This means θ is the true probability of the ith at bat. Since we only observe $x_i = 1$ or $x_i = 0$, you can intuitively see how θ can never be known — not enough information to determine it.

What values can θ be?

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \text{Bernoulli}(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ 0 & \text{if he did not get a hit on his } i\text{th at bat} \end{cases}$$

This means θ is the true probability of the ith at bat. Since we only observe $x_i = 1$ or $x_i = 0$, you can intuitively see how θ can never be known — not enough information to determine it.

What values can θ be? We say $\theta \in \Theta$ which is its "parameter space". Since θ is a probability, all values between 0 and 1 are allowed.

Our Parameter of Interest

But what is θ in our problem? Remember, if X_i represents the player of interest's ith at bat,

$$X_i \sim \mathrm{Bernoulli}(\theta) := \begin{cases} 1 & \text{if he got a hit on his } i\text{th at bat} \\ 0 & \text{if he did not get a hit on his } i\text{th at bat} \end{cases}$$

This means θ is the true probability of the ith at bat. Since we only observe $x_i = 1$ or $x_i = 0$, you can intuitively see how θ can never be known — not enough information to determine it.

What values can θ be? We say $\theta \in \Theta$ which is its "parameter space". Since θ is a probability, all values between 0 and 1 are allowed. But to keep the problem non-trivial, we do not allow 0 or 1, and thus we have

$\theta \in (0, 1)$ which is the "parameter space" of the Bernoulli r.v.

Identically Distributed

The next concept is subtle. Let's consider the first at bat and the second at bat: X_1 and X_2.

Identically Distributed

The next concept is subtle. Let's consider the first at bat and the second at bat: X_1 and X_2. It could be there are different probabilities of getting hits. Why? One day the player is on his game and the next day he's off his game, thus:

Identically Distributed

The next concept is subtle. Let's consider the first at bat and the second at bat: X_1 and X_2. It could be there are different probabilities of getting hits. Why? One day the player is on his game and the next day he's off his game, thus:

$$X_1 \sim \mathrm{Bernoulli}\,(\theta_1)\,, \quad X_2 \sim \mathrm{Bernoulli}\,(\theta_2) \quad s.t. \;\; \theta_1 > \theta_2$$

In which case we would be estimating two different parameters!

Identically Distributed

The next concept is subtle. Let's consider the first at bat and the second at bat: X_1 and X_2. It could be there are different probabilities of getting hits. Why? One day the player is on his game and the next day he's off his game, thus:

$$X_1 \sim \text{Bernoulli}(\theta_1), \quad X_2 \sim \text{Bernoulli}(\theta_2) \quad s.t. \quad \theta_1 > \theta_2$$

In which case we would be estimating two different parameters! Taken to the extreme, when examining n hits, there would be n different parameters. So we make a simplifying assumption: we pretend the player has "one true" probability of getting a hit, which we call θ and hence X_1 and X_2 share that same θ and are thus identically distributed.

Identically Distributed

The next concept is subtle. Let's consider the first at bat and the second at bat: X_1 and X_2. It could be there are different probabilities of getting hits. Why? One day the player is on his game and the next day he's off his game, thus:

$$X_1 \sim \text{Bernoulli}(\theta_1), \quad X_2 \sim \text{Bernoulli}(\theta_2) \quad s.t. \;\; \theta_1 > \theta_2$$

In which case we would be estimating two different parameters! Taken to the extreme, when examining n hits, there would be n different parameters. So we make a simplifying assumption: we pretend the player has "one true" probability of getting a hit, which we call θ and hence X_1 and X_2 share that same θ and are thus identically distributed. **We now assume (which will be slightly modified later) that $X_1, \ldots, X_n$ are identically distributed.**

Detour: Conditional Probability

Another subtle concept. Let's consider the following question: if you know the player got a hit on the first at bat, would that change the probability of getting a hit the next at bat?

Detour: Conditional Probability

Another subtle concept. Let's consider the following question: if you know the player got a hit on the first at bat, would that change the probability of getting a hit the next at bat? If the answer is that the probability would go up, this is known as "hitting streak" in baseball.

Detour: Conditional Probability

Another subtle concept. Let's consider the following question: if you know the player got a hit on the first at bat, would that change the probability of getting a hit the next at bat? If the answer is that the probability would go up, this is known as "hitting streak" in baseball. In our notation this would be:

$$\mathbb{P}(X_2 = 1 \mid X_1 = 1) > \mathbb{P}(X_2 = 1)$$

You may remember the l.h.s.'s notation with the vertical line symbol (the "pipe") is known as "conditional probability"

Detour: Conditional Probability

Another subtle concept. Let's consider the following question: if you know the player got a hit on the first at bat, would that change the probability of getting a hit the next at bat? If the answer is that the probability would go up, this is known as "hitting streak" in baseball. In our notation this would be:

$$\mathbb{P}\left(X_2 = 1 \mid X_1 = 1\right) > \mathbb{P}\left(X_2 = 1\right)$$

You may remember the l.h.s.'s notation with the vertical line symbol (the "pipe") is known as "conditional probability" and the notation on the right — the "default" probability notation — is known as an "unconditional probability".

Detour: Conditional Probability

Another subtle concept. Let's consider the following question: if you know the player got a hit on the first at bat, would that change the probability of getting a hit the next at bat? If the answer is that the probability would go up, this is known as "hitting streak" in baseball. In our notation this would be:

$$\mathbb{P}(X_2 = 1 \mid X_1 = 1) > \mathbb{P}(X_2 = 1)$$

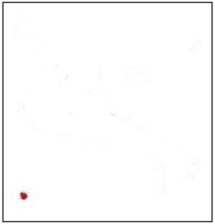
You may remember the l.h.s.'s notation with the vertical line symbol (the "pipe") is known as "conditional probability" and the notation on the right — the "default" probability notation — is known as an "unconditional probability". In our case, knowing that there was a hit just before this at bat (i.e. that $X_1 = 1$) makes the second at bat more likely to result in a hit than if you didn't have the previous information (just the r.h.s.).

Definition of Conditional Probability

How do we define conditional probability? Two images are below.

Definition of Conditional Probability

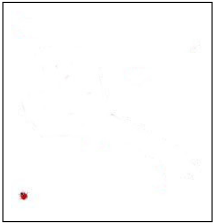
How do we define conditional probability? Two images are below.



Define the event A as you throw a pin on the left image and you hit the bug.

Definition of Conditional Probability

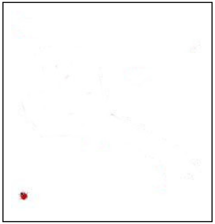
How do we define conditional probability? Two images are below.



Define the event A as you throw a pin on the left image and you hit the bug. Safe to say $\mathbb{P}(A)$ is small.

Definition of Conditional Probability

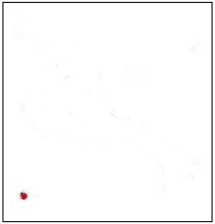
How do we define conditional probability? Two images are below.



Define the event A as you throw a pin on the left image and you hit the bug. Safe to say $\mathbb{P}(A)$ is small. Now imagine you throw a pin inside the field of the magnifying glass on the right. Safe to say $\mathbb{P}(A)$ is large there.

Definition of Conditional Probability

How do we define conditional probability? Two images are below.

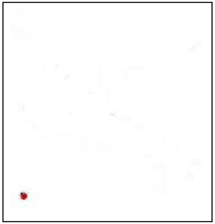


Define the event A as you throw a pin on the left image and you hit the bug. Safe to say $\mathbb{P}(A)$ is small. Now imagine you throw a pin inside the field of the magnifying glass on the right. Safe to say $\mathbb{P}(A)$ is large there.

But they're not the same events! Consider the event B where you have the magnifying glass. Pinning the bug, A, while you have B raises the probability of A, thus

Definition of Conditional Probability

How do we define conditional probability? Two images are below.



Define the event A as you throw a pin on the left image and you hit the bug. Safe to say $\mathbb{P}(A)$ is small. Now imagine you throw a pin inside the field of the magnifying glass on the right. Safe to say $\mathbb{P}(A)$ is large there.

But they're not the same events! Consider the event B where you have the magnifying glass. Pinning the bug, A, while you have B raises the probability of A, thus

$$\mathbb{P}(A \mid B) > \mathbb{P}(A).$$

Bayes Rule

By how much is $\mathbb{P}(A \mid B)$ bigger than $\mathbb{P}(A)$?

Bayes Rule

By how much is $\mathbb{P}(A \mid B)$ bigger than $\mathbb{P}(A)$? The bug on the right is in intersection of A and B at the same time. The zoom factor is whatever the original size of the field was (the box from the left image) divided by the new size. On the left, the whole box has probability 1 since it represents the universe. So the zoom factor is $\frac{1}{\mathbb{P}(B)}$.

Bayes Rule

By how much is $\mathbb{P}(A \mid B)$ bigger than $\mathbb{P}(A)$? The bug on the right is in intersection of A and B at the same time. The zoom factor is whatever the original size of the field was (the box from the left image) divided by the new size. On the left, the whole box has probability 1 since it represents the universe. So the zoom factor is $\frac{1}{\mathbb{P}(B)}$. Thus,

$$\mathbb{P}(A \mid B) = \underbrace{\mathbb{P}(AB)}_{\text{original size of bug}} \times \underbrace{\frac{1}{\mathbb{P}(B)}}_{\text{zoom factor}} = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$$

This rule is known as **Bayes Rule**

Bayes Rule

By how much is $\mathbb{P}(A \mid B)$ bigger than $\mathbb{P}(A)$? The bug on the right is in intersection of A and B at the same time. The zoom factor is whatever the original size of the field was (the box from the left image) divided by the new size. On the left, the whole box has probability 1 since it represents the universe. So the zoom factor is $\frac{1}{\mathbb{P}(B)}$. Thus,

$$\mathbb{P}(A \mid B) = \underbrace{\mathbb{P}(AB)}_{\text{original size of bug}} \times \underbrace{\frac{1}{\mathbb{P}(B)}}_{\text{zoom factor}} = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$$

This rule is known as **Bayes Rule** and it will form the basis for what we do in the next unit.

Bayes Rule

By how much is $\mathbb{P}(A \mid B)$ bigger than $\mathbb{P}(A)$? The bug on the right is in intersection of A and B at the same time. The zoom factor is whatever the original size of the field was (the box from the left image) divided by the new size. On the left, the whole box has probability 1 since it represents the universe. So the zoom factor is $\frac{1}{\mathbb{P}(B)}$. Thus,

$$\mathbb{P}(A \mid B) = \underbrace{\mathbb{P}(AB)}_{\text{original size of bug}} \times \underbrace{\frac{1}{\mathbb{P}(B)}}_{\text{zoom factor}} = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$$

This rule is known as **Bayes Rule** and it will form the basis for what we do in the next unit.

Technicalities

This example is a bit confusing. Here, $AB = A$ since A is a proper subset of B. The original bug size is A which is also AB. Thus, $\mathbb{P}(A \mid B) = \mathbb{P}(A) / \mathbb{P}(B)$ only in this case though.

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2.

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2. Imagine if knowing what happened at the first at bat X_1 did not inform anything about X_2.

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2. Imagine if knowing what happened at the first at bat X_1 did not inform anything about X_2. Then the conditional and the unconditional would be one and the same:

$$\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \mathbb{P}(X_2 = x_2) \quad \text{and shorthand,} \quad \mathbb{P}(X_2 \mid X_1) = \mathbb{P}(X_2)$$

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2. Imagine if knowing what happened at the first at bat X_1 did not inform anything about X_2. Then the conditional and the unconditional would be one and the same:

$$\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \mathbb{P}(X_2 = x_2) \quad \text{and shorthand,} \quad \mathbb{P}(X_2 \mid X_1) = \mathbb{P}(X_2)$$

This is known as "independence".

Via Bayes Rule, we have:

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2. Imagine if knowing what happened at the first at bat X_1 did not inform anything about X_2. Then the conditional and the unconditional would be one and the same:

$$\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \mathbb{P}(X_2 = x_2) \quad \text{and shorthand,} \quad \mathbb{P}(X_2 \mid X_1) = \mathbb{P}(X_2)$$

This is known as "independence".

Via Bayes Rule, we have:

$$\mathbb{P}(X_2 \mid X_1) = \frac{\mathbb{P}(X_1, X_2)}{\mathbb{P}(X_1)} = \mathbb{P}(X_2) \Rightarrow \mathbb{P}(X_1, X_2) = \mathbb{P}(X_1)\,\mathbb{P}(X_2)$$

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2. Imagine if knowing what happened at the first at bat X_1 did not inform anything about X_2. Then the conditional and the unconditional would be one and the same:

$$\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \mathbb{P}(X_2 = x_2) \quad \text{and shorthand,} \quad \mathbb{P}(X_2 \mid X_1) = \mathbb{P}(X_2)$$

This is known as "independence".

Via Bayes Rule, we have:

$$\mathbb{P}(X_2 \mid X_1) = \frac{\mathbb{P}(X_1, X_2)}{\mathbb{P}(X_1)} = \mathbb{P}(X_2) \Rightarrow \mathbb{P}(X_1, X_2) = \mathbb{P}(X_1)\,\mathbb{P}(X_2)$$

So independence implies that joint events factor and you can multiply them.

Independence

So now that we know what conditional probability is, let's return to the two at bats, X_1 and X_2. Imagine if knowing what happened at the first at bat X_1 did not inform anything about X_2. Then the conditional and the unconditional would be one and the same:

$$\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \mathbb{P}(X_2 = x_2) \quad \text{and shorthand,} \quad \mathbb{P}(X_2 \mid X_1) = \mathbb{P}(X_2)$$

This is known as "independence".

Via Bayes Rule, we have:

$$\mathbb{P}(X_2 \mid X_1) = \frac{\mathbb{P}(X_1, X_2)}{\mathbb{P}(X_1)} = \mathbb{P}(X_2) \Rightarrow \mathbb{P}(X_1, X_2) = \mathbb{P}(X_1)\,\mathbb{P}(X_2)$$

So independence implies that joint events factor and you can multiply them. **We now assume (which will be slightly modified later) that $X_1, \ldots, X_n$ are independent.**

IID and Total Hits

We say that the r.v.'s are IID and denoted

$X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Bernoulli}\,(\theta)$

IID and Total Hits

We say that the r.v.'s are IID and denoted $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ if we have the situation where the probability that any of the at bats will yield a hit is the same, θ (identical distribution) and that knowing any of the outcomes of the at bats does not affect any other hit at any other at bat (independence).

IID and Total Hits

We say that the r.v.'s are IID and denoted $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ if we have the situation where the probability that any of the at bats will yield a hit is the same, θ (identical distribution) and that knowing any of the outcomes of the at bats does not affect any other hit at any other at bat (independence).

Now consider tallying all the hits in n at bats. This would be equivalent to the sum of all the X_i's since we only care about summing the 1's, call this X (even though it's confusing),

IID and Total Hits

We say that the r.v.'s are IID and denoted $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Bernoulli}(\theta)$ if we have the situation where the probability that any of the at bats will yield a hit is the same, θ (identical distribution) and that knowing any of the outcomes of the at bats does not affect any other hit at any other at bat (independence).

Now consider tallying all the hits in n at bats. This would be equivalent to the sum of all the X_i's since we only care about summing the 1's, call this X (even though it's confusing),

$$X = \sum_{i=1}^{n} X_i$$

Can we find the PMF of X?

IID and Total Hits

We say that the r.v.'s are IID and denoted $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ if we have the situation where the probability that any of the at bats will yield a hit is the same, θ (identical distribution) and that knowing any of the outcomes of the at bats does not affect any other hit at any other at bat (independence).

Now consider tallying all the hits in n at bats. This would be equivalent to the sum of all the X_i's since we only care about summing the 1's, call this X (even though it's confusing),

$$X = \sum_{i=1}^{n} X_i$$

Can we find the PMF of X? Can we ask what is the probability of yea so many hits out of n?

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$.

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

$$
\begin{array}{cccc}
HHHOOO & HHOHOO & HOHHOO & OHHHOO \\
HHOOHO & HOOHHO & OOHHHO & HHOOOH \\
HOOOHH & OOOHHH & HOOHOH & HOHOOH \\
HOHOHO & OHOHOH & OHHOHO & OOHHOH \\
OOHOHH & OHHOHO & OHHOOH & OHOOHH
\end{array}
$$

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

$$\begin{array}{cccc} HHHOOO & HHOHOO & HOHHOO & OHHHOO \\ HHOOHO & HOOHHO & OOHHHO & HHOOOH \\ HOOOHH & OOOHHH & HOOHOH & HOHOOH \\ HOHOHO & OHOHOH & OHHOHO & OOHHOH \\ OOHOHH & OHHOHO & OHHOOH & OHOOHH \end{array}$$

Since each string of H's and O's is independent,

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

$$
\begin{array}{cccc}
HHHOOO & HHOHOO & HOHHOO & OHHHOO \\
HHOOHO & HOOHHO & OOHHHO & HHOOOH \\
HOOOHH & OOOHHH & HOOHOH & HOHOOH \\
HOHOHO & OHOHOH & OHHOHO & OOHHOH \\
OOHOHH & OHHOHO & OHHOOH & OHOOHH
\end{array}
$$

Since each string of H's and O's is independent, the probability of each will always be $0.25^3 0.75^3$

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

$$\begin{array}{cccc} HHHOOO & HHOHOO & HOHHOO & OHHHOO \\ HHOOHO & HOOHHO & OOHHHO & HHOOOH \\ HOOOHH & OOOHHH & HOOHOH & HOHOOH \\ HOHOHO & OHOHOH & OHHOHO & OOHHOH \\ OOHOHH & OHHOHO & OHHOOH & OHOOHH \end{array}$$

Since each string of H's and O's is independent, the probability of each will always be $0.25^3 0.75^3$ since each the probability of each at bat can be multiplied.

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

HHHOOO	*HHOHOO*	*HOHHOO*	*OHHHOO*
HHOOHO	*HOOHHO*	*OOHHHO*	*HHOOOH*
HOOOHH	*OOOHHH*	*HOOHOH*	*HOHOOH*
HOHOHO	*OHOHOH*	*OHHOHO*	*OOHHOH*
OOHOHH	*OHHOHO*	*OHHOOH*	*OHOOHH*

Since each string of H's and O's is independent, the probability of each will always be $0.25^3 0.75^3$ since each the probability of each at bat can be multiplied. So now we need to count the number of ways to make 3 H's in 6 positions.

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

$$\begin{array}{cccc} HHHOOO & HHOHOO & HOHHOO & OHHHOO \\ HHOOHO & HOOHHO & OOHHHO & HHOOOH \\ HOOOHH & OOOHHH & HOOHOH & HOHOOH \\ HOHOHO & OHOHOH & OHHOHO & OOHHOH \\ OOHOHH & OHHOHO & OHHOOH & OHOOHH \end{array}$$

Since each string of H's and O's is independent, the probability of each will always be $0.25^3 0.75^3$ since each the probability of each at bat can be multiplied. So now we need to count the number of ways to make 3 H's in 6 positions. Recall from basic probability that this is $\binom{6}{3} = \frac{6!}{3!(6-3)!} = 20$. So,

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

HHHOOO	*HHOHOO*	*HOHHOO*	*OHHHOO*
HHOOHO	*HOOHHO*	*OOHHHO*	*HHOOOH*
HOOOHH	*OOOHHH*	*HOOHOH*	*HOHOOH*
HOHOHO	*OHOHOH*	*OHHOHO*	*OOHHOH*
OOHOHH	*OHHOHO*	*OHHOOH*	*OHOOHH*

Since each string of H's and O's is independent, the probability of each will always be $0.25^3 0.75^3$ since each the probability of each at bat can be multiplied. So now we need to count the number of ways to make 3 H's in 6 positions. Recall from basic probability that this is $\binom{6}{3} = \frac{6!}{3!(6-3)!} = 20$. So,

$$\mathbb{P}(X = 3) = \binom{6}{3} 0.25^3 0.75^3 = \texttt{dbinom(3, 6, 0.25)} \approx 0.132$$

The PMF of the sum of Bernoullis

Imagine for a moment there are $n = 6$ at bats and the probability of a hit $\theta = 0.25$. How could we find the probability of 3 hits out of 6 at bats i.e. $\mathbb{P}(X = 3)$?

Denote hit as H and not hit as O. There are 20 ways to make 3 hits occur out of 6:

HHHOOO	*HHOHOO*	*HOHHOO*	*OHHHOO*
HHOOHO	*HOOHHO*	*OOHHHO*	*HHOOOH*
HOOOHH	*OOOHHH*	*HOOHOH*	*HOHOOH*
HOHOHO	*OHOHOH*	*OHHOHO*	*OOHHOH*
OOHOHH	*OHHOHO*	*OHHOOH*	*OHOOHH*

Since each string of H's and O's is independent, the probability of each will always be $0.25^3 0.75^3$ since each the probability of each at bat can be multiplied. So now we need to count the number of ways to make 3 H's in 6 positions. Recall from basic probability that this is $\binom{6}{3} = \frac{6!}{3!(6-3)!} = 20$. So,

$$\mathbb{P}(X = 3) = \binom{6}{3} 0.25^3 0.75^3 = \texttt{dbinom(3, 6, 0.25)} \approx 0.132$$

Where "`dbinom(3, 6, 0.25)`" is the `R` code to compute the answer.

The Binomial Model

Again consider n at bats with arbitrary probability of a hit θ (which is limited by its parameter space).

The Binomial Model

Again consider n at bats with arbitrary probability of a hit θ (which is limited by its parameter space). The r.v. we spoke about is actually called the binomial r.v.,

$$X_n = \sum_{i=1}^{n} X_i \sim \mathrm{Binomial}\,(n,\, \theta)$$

The Binomial Model

Again consider n at bats with arbitrary probability of a hit θ (which is limited by its parameter space). The r.v. we spoke about is actually called the binomial r.v.,

$$X_n = \sum_{i=1}^{n} X_i \sim \mathrm{Binomial}\,(n,\, \theta)$$

Its PMF can be found by generalizing the reasoning from the last slide:

The Binomial Model

Again consider n at bats with arbitrary probability of a hit θ (which is limited by its parameter space). The r.v. we spoke about is actually called the binomial r.v.,

$$X_n = \sum_{i=1}^{n} X_i \sim \mathrm{Binomial}\,(n,\, \theta)$$

Its PMF can be found by generalizing the reasoning from the last slide:

$$\mathbb{P}\,(X = x) = \mathrm{Binomial}\,(n,\, \theta) := \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

The Role of Data

Given that we have n at bats and we assume each at bat is an $\stackrel{iid}{\sim}$ bernoulli with parameter θ,

The Role of Data

Given that we have n at bats and we assume each at bat is an $\stackrel{iid}{\sim}$ bernoulli with parameter θ, we now focus on estimating θ and testing ideas about θ.

The Role of Data

Given that we have n at bats and we assume each at bat is an $\overset{iid}{\sim}$ bernoulli with parameter θ, we now focus on estimating θ and testing ideas about θ.

To do so, we use **data** to estimate θ and in our case, that's just $x_1, x_2, \ldots, x_n$ which looks like a string of n 0's and/or 1's.

The Role of Data

Given that we have n at bats and we assume each at bat is an $\overset{iid}{\sim}$ bernoulli with parameter θ, we now focus on estimating θ and testing ideas about θ.

To do so, we use **data** to estimate θ and in our case, that's just $x_1, x_2, \ldots, x_n$ which looks like a string of n 0's and/or 1's.

Technicalities

θ isn't exactly the lifetime batting average, it's deeper: it's the intrinsic propensity for the batter under scrutiny to create a hit out of an at-bat. If the number of lifetime at-bats is large, the lifetime BA and θ will be quite similar. We will ignore this detail.

The Role of Data

Given that we have n at bats and we assume each at bat is an $\overset{iid}{\sim}$ bernoulli with parameter θ, we now focus on estimating θ and testing ideas about θ.

To do so, we use **data** to estimate θ and in our case, that's just $x_1, x_2, \ldots, x_n$ which looks like a string of n 0's and/or 1's.

Technicalities

θ isn't exactly the lifetime batting average, it's deeper: it's the intrinsic propensity for the batter under scrutiny to create a hit out of an at-bat. If the number of lifetime at-bats is large, the lifetime BA and θ will be quite similar. We will ignore this detail.

So how do we use data?

The Role of Data

Given that we have n at bats and we assume each at bat is an $\overset{iid}{\sim}$ bernoulli with parameter θ, we now focus on estimating θ and testing ideas about θ.

To do so, we use **data** to estimate θ and in our case, that's just $x_1, x_2, \ldots, x_n$ which looks like a string of n 0's and/or 1's.

Technicalities

θ isn't exactly the lifetime batting average, it's deeper: it's the intrinsic propensity for the batter under scrutiny to create a hit out of an at-bat. If the number of lifetime at-bats is large, the lifetime BA and θ will be quite similar. We will ignore this detail.

So how do we use data? We first begin with the popular tools of frequentist estimation and then move to the Bayesian tools which is the focus of this module.

Homework Problems

1. Draw the PMF for $X \sim \text{Bernoulli}(\theta)$.
2. Imagine two Bernoulli r.v.'s X_1 and X_2 which model two fair coin flips where Heads is mapped to 1 and tails is mapped to 0. The probability of heads is $1/2$. Explain using the definition of r.v. independence why these two r.v.'s are *dependent*.
3. Using the same two sorcery-controlled coins, explain using the definition of equality in distribution why or why not X_1 and X_2 are equally distributed.
4. Are $X_1, X_2 \overset{iid}{\sim} \text{Bernoulli}(\theta)$ if they are modeled by these two sorcery-controlled coins?
5. A new situation: the probability of a bundle of coins (scotch-taped together) landing on its side is $\mathbb{P}(S) = 1/11$. Heads and tail probability are $5/11$. Let's call landing on its side the event we seek, create a Bernoulli r.v. for this event indicate its support, its parameter and the parameter space.

Homework Problems

6 I flip the coin bundle once. What is the probability of the trial of interest being found? Write a probability statement.

7 Let's say we flip 10 times. What is the probability that we get one (and only one) success? I want to see a probability model. Write "$X \sim$" something below. Then I want to see a probability statement. Then I want to see a computation. Answer then in decimal rounded to two digits.

8 Let's say we flip 10 times. What is the probability that we get 5 (and only 5) successes?

9 Let's say we flip 10 times. What is the probability that we get 8 (and only 8) successes?

10 Let's say we flip 10 times. What is the probability we get one or two successes?

Homework Problems

11 Consider three NYC buildings: One World Trade Center (104 floors), the Empire State Building (86 occupied floors) and the Bank of America Tower (55 floors). Consider walking into one of them at random and picking a random floor. What is the probability of choosing the Empire State Building's 40th floor?

12 What is the probability he on the 30the floor?

13 If you know he's on the 40th floor, what is the probability he is in the empire state building?

14 Look up the Monte Hall game. Explain why switching is a good strategy based on conditional probability.

15 Is the $\overset{iid}{\sim}$ bernoulli model for at bats a good model in baseball? This is complicated.

Likelihood function

Recall from before the PMF for our data (which is a function of the free variable X),

$$\mathbb{P}(X;\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}.$$

Likelihood function

Recall from before the PMF for our data (which is a function of the free variable X),

$$\mathbb{P}(X;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

I've wrote it differently. Here I use θ as information you need to calculate the function.

Likelihood function

Recall from before the PMF for our data (which is a function of the free variable X),

$$\mathbb{P}(X;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

I've wrote it differently. Here I use θ as information you need to calculate the function.

Technicalities

This is similar to the following polynomial: $f(x;a) = ax^2$.

Here, a is not to be considered a second free variable in the function, but a constant tuning parameter which provides customization for the shape of the parabola (thin to fat and everything in between). This is exactly what θ does above.

Likelihood function

Recall from before the PMF for our data (which is a function of the free variable X),

$$\mathbb{P}\left(X;\theta\right) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}.$$

I've wrote it differently. Here I use θ as information you need to calculate the function.

Technicalities

This is similar to the following polynomial: $f(x;a) = ax^2$.

Here, a is not to be considered a second free variable in the function, but a constant tuning parameter which provides customization for the shape of the parabola (thin to fat and everything in between). This is exactly what θ does above.

What if we "flip" or invert the role of the data and the parameter? We still get the same thing, but I'll use different notation:

$$L\left(\theta;X\right) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}.$$

Likelihood function

Recall from before the PMF for our data (which is a function of the free variable X),

$$\mathbb{P}(X;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

I've wrote it differently. Here I use θ as information you need to calculate the function.

Technicalities

This is similar to the following polynomial: $f(x;a) = ax^2$.

Here, a is not to be considered a second free variable in the function, but a constant tuning parameter which provides customization for the shape of the parabola (thin to fat and everything in between). This is exactly what θ does above.

What if we "flip" or invert the role of the data and the parameter? We still get the same thing, but I'll use different notation:

$$L(\theta;X) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

The L stands now for likelihood and the θ is now the free variable and the data X is the parameter. (but otherwise it is the same as the PMF!)

Maximum Likelihood and the MLE

Why did we do this?

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data.

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative,

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101.

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\text{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function.

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$,

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; \ell(\theta; X)$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\; \ell(\theta; X)$$

and that will be our definition going forward.

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

0

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=}$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=} \frac{\partial}{\partial \theta}[\ell(\theta; X)] =$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \stackrel{set}{=} \frac{\partial}{\partial\theta}\left[\ell(\theta; X)\right] = \frac{\partial}{\partial\theta}\left[\ln\left(\binom{n}{X}\right) + X\ln(\theta) + (n - X)\ln(1-\theta)\right] =$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=} \frac{\partial}{\partial \theta}\left[\ell(\theta; X)\right] = \frac{\partial}{\partial \theta}\left[\ln\left(\binom{n}{X}\right) + X\ln(\theta) + (n - X)\ln(1-\theta)\right] = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max} \; \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=} \frac{\partial}{\partial \theta}\left[\ell(\theta; X)\right] = \frac{\partial}{\partial \theta}\left[\ln\left(\binom{n}{X}\right) + X\ln(\theta) + (n-X)\ln(1-\theta)\right] = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

$$\Rightarrow (1-\theta)X =$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=} \frac{\partial}{\partial \theta}[\ell(\theta; X)] = \frac{\partial}{\partial \theta}\left[\ln\left(\binom{n}{X}\right) + X\ln(\theta) + (n - X)\ln(1-\theta)\right] = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

$$\Rightarrow (1-\theta)X = \theta(n-X) \Rightarrow X - \cancel{X\theta} =$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \arg\max_{\theta \in \Theta} \; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \arg\max_{\theta \in \Theta} \; \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=} \frac{\partial}{\partial\theta}[\ell(\theta; X)] = \frac{\partial}{\partial\theta}\left[\ln\left(\binom{n}{X}\right) + X\ln(\theta) + (n - X)\ln(1-\theta)\right] = \frac{X}{\theta} - \frac{n - X}{1 - \theta}$$

$$\Rightarrow (1-\theta)X = \theta(n - X) \Rightarrow X - \cancel{X\theta} = n\theta - \cancel{X\theta} \Rightarrow \hat{\theta}_{\mathrm{MLE}}$$

Maximum Likelihood and the MLE

Why did we do this? Because we now can ask the question: which value of θ is the **most likely** for this data. We call this the maximum likelihood estimator (MLE):

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta\in\Theta}{\arg\max}\; L(\theta; X)$$

This comes down to taking the derivative, setting it equal to zero and solving for θ just like in Calculus 101. Note that the maximum of the function is immune to monotonic transformations of the function. One such convenient monotonic transformation is the natural log, $\ell(\theta; X) := \ln(L(\theta; X))$, so we can now calculate the maximum likelihood estimator via

$$\hat{\theta}_{\mathrm{MLE}} := \underset{\theta\in\Theta}{\arg\max}\; \ell(\theta; X)$$

and that will be our definition going forward. So let's begin solving for the MLE of θ given our data from n at bats:

$$0 \overset{set}{=} \frac{\partial}{\partial\theta}\left[\ell(\theta; X)\right] = \frac{\partial}{\partial\theta}\left[\ln\left(\binom{n}{X}\right) + X\ln(\theta) + (n - X)\ln(1-\theta)\right] = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

$$\Rightarrow (1-\theta)X = \theta(n-X) \Rightarrow X - \cancel{X\theta} = n\theta - \cancel{X\theta} \Rightarrow \hat{\theta}_{\mathrm{MLE}} = \frac{X}{n}$$

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s,

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn).

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Note how everything is lowercase now.

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\text{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\text{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Note how everything is lowercase now. But through an abuse of notation, I didn't change the notation $\hat{\theta}_{\text{MLE}}$.

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Note how everything is lowercase now. But through an abuse of notation, I didn't change the notation $\hat{\theta}_{\mathrm{MLE}}$. But more importantly — you've seen this before!

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Note how everything is lowercase now. But through an abuse of notation, I didn't change the notation $\hat{\theta}_{\mathrm{MLE}}$. But more importantly — you've seen this before!

Remember adding up all the numbers and then dividing by the number of numbers from high school? This is called the "sample average".

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\text{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\text{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Note how everything is lowercase now. But through an abuse of notation, I didn't change the notation $\hat{\theta}_{\text{MLE}}$. But more importantly — you've seen this before!

Remember adding up all the numbers and then dividing by the number of numbers from high school? This is called the "sample average". (In the our special case of the binomial θ estimation, you may see it written as $\hat{p}$ instead).

The Sample Average

If you recall, X was defined as the sum of Bernoulli r.v.'s, hence we can rewrite the MLE as:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

This is the **estimator** (the r.v. that represents how the estimate is drawn). To find the maximum likelihood **estimate** you can plug in the data into the estimator:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Note how everything is lowercase now. But through an abuse of notation, I didn't change the notation $\hat{\theta}_{\mathrm{MLE}}$. But more importantly — you've seen this before!

Remember adding up all the numbers and then dividing by the number of numbers from high school? This is called the "sample average". (In the our special case of the binomial θ estimation, you may see it written as $\hat{p}$ instead).

This is our best guess as to the value of θ.

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit.

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000,$$

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000, \quad \bar{x} = \frac{2}{2} = 1.000$$

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000, \quad \bar{x} = \frac{2}{2} = 1.000$$

All these batting averages are absurd! Why?

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000, \quad \bar{x} = \frac{2}{2} = 1.000$$

All these batting averages are absurd! Why?

1. The first says he will never make a hit. Well why is he paid millions of dollars then? This estimate is clearly off.

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000, \quad \bar{x} = \frac{2}{2} = 1.000$$

All these batting averages are absurd! Why?

1. The first says he will never make a hit. Well why is he paid millions of dollars then? This estimate is clearly off.
2. The second says he will always make a hit. That is clearly wrong.

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000, \quad \bar{x} = \frac{2}{2} = 1.000$$

All these batting averages are absurd! Why?

1. The first says he will never make a hit. Well why is he paid millions of dollars then? This estimate is clearly off.
2. The second says he will always make a hit. That is clearly wrong.

The "Frequentist" view forces us to be purely "objective" and let the data (and only the data) speak on behalf of the underlying unknown parameter.

Trouble with the Sample Average

So if there were 100 at bats and 29 hits, the batting average is 0.290 and that is our best guess as to the true propensity of the batter to get a hit. So computing the BA is a good idea!!

But what if there were two at bats and there were no hits? Or two hits?

$$\bar{x} = \frac{0}{2} = 0.000, \quad \bar{x} = \frac{2}{2} = 1.000$$

All these batting averages are absurd! Why?

1. The first says he will never make a hit. Well why is he paid millions of dollars then? This estimate is clearly off.
2. The second says he will always make a hit. That is clearly wrong.

The "Frequentist" view forces us to be purely "objective" and let the data (and only the data) speak on behalf of the underlying unknown parameter. But this is clearly silly sometimes!

Homework Problems

1. Rederive the MLE (estimator and estimate) for the Binomial likelihood model.
2. Try to derive it without using the monotonic transformation of the log of the likelihood.
3. Instead of the maximum likelihood, write an expression for the 90%ile of the likelihood.
4. Given 345 at bats and 132 hits, what is the maximum likelihood of the hit probability?
5. Is θ the same as lifetime batting average? Discuss.
6. Why is an MLE of 0.000 or 1.000 a bad thing? Discuss.
7. Devise a means to fix these two pathological situations without looking ahead in the module.

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates?

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

The Bayesian Paradigm Shift

Imagine that θ is still a single value still unknown,

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

The Bayesian Paradigm Shift

Imagine that θ is still a single value still unknown, but now we are able to quantify our uncertainty about where θ is using a distribution, $\mathbb{P}(\theta)$.

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

The Bayesian Paradigm Shift

Imagine that θ is still a single value still unknown, but now we are able to quantify our uncertainty about where θ is using a distribution, $\mathbb{P}(\theta)$. Now, θ becomes its own r.v!

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

The Bayesian Paradigm Shift

Imagine that θ is still a single value still unknown, but now we are able to quantify our uncertainty about where θ is using a distribution, $\mathbb{P}(\theta)$. Now, θ becomes its own r.v! We are interested in what happens to this uncertainty with the introduction of data (X),

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

The Bayesian Paradigm Shift

Imagine that θ is still a single value still unknown, but now we are able to quantify our uncertainty about where θ is using a distribution, $\mathbb{P}(\theta)$. Now, θ becomes its own r.v! We are interested in what happens to this uncertainty with the introduction of data (X),

$$\theta \quad \xrightarrow{data} \quad \theta \mid X$$

This is called "Bayesian Conditionalism".

Introduction

How do we solve this problem of 0.000 and 1.000 batting average estimates? Bayesian statistics is one approach we will explore here.

The Bayesian Paradigm Shift

Imagine that θ is still a single value still unknown, but now we are able to quantify our uncertainty about where θ is using a distribution, $\mathbb{P}(\theta)$. Now, θ becomes its own r.v! We are interested in what happens to this uncertainty with the introduction of data (X),

$$\theta \quad \xrightarrow{data} \quad \theta \mid X$$

This is called "Bayesian Conditionalism". We begin with what is called the "prior" and we update it to find the "posterior" — data is able to refine our ideas about where θ is likely to be.

Bayes Rule Again

How does this update work?

Bayes Rule Again

How does this update work? Through Bayes Rule!

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}}$$

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \mathbb{P}(X, \theta)$$

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} =$$

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}}$$

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data.

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities).

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X.

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1.

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}\left(\theta \mid X\right)}_{\text{posterior}} = \frac{\mathbb{P}\left(X, \theta\right)}{\mathbb{P}\left(X\right)} = \underbrace{\frac{\mathbb{P}\left(X \mid \theta\right)}{\mathbb{P}\left(X\right)}}_{\text{update factor}} \underbrace{\mathbb{P}\left(\theta\right)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}\left(X \mid \theta\right)$ divided by $\mathbb{P}\left(X\right)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}(X \mid \theta)$ divided by $\mathbb{P}(X)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Technicalities

Previously we defined likelihood as $\mathbb{P}(X; \theta)$ but now it's $\mathbb{P}(X \mid \theta)$.

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}(X \mid \theta)$ divided by $\mathbb{P}(X)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Technicalities

Previously we defined likelihood as $\mathbb{P}(X; \theta)$ but now it's $\mathbb{P}(X \mid \theta)$. Why did the semicolon become a pipe character?

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}(X \mid \theta)$ divided by $\mathbb{P}(X)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Technicalities

Previously we defined likelihood as $\mathbb{P}(X; \theta)$ but now it's $\mathbb{P}(X \mid \theta)$. Why did the semicolon become a pipe character? Previously θ was considered a fixed constant, a parameter in the function. But now it is its own random variable with full-citizen rights!

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}(X \mid \theta)$ divided by $\mathbb{P}(X)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Technicalities

Previously we defined likelihood as $\mathbb{P}(X; \theta)$ but now it's $\mathbb{P}(X \mid \theta)$. Why did the semicolon become a pipe character? Previously θ was considered a fixed constant, a parameter in the function. But now it is its own random variable with full-citizen rights! We thereby give it a true pipe character.

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}(X \mid \theta)$ divided by $\mathbb{P}(X)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Technicalities

Previously we defined likelihood as $\mathbb{P}(X; \theta)$ but now it's $\mathbb{P}(X \mid \theta)$. Why did the semicolon become a pipe character? Previously θ was considered a fixed constant, a parameter in the function. But now it is its own random variable with full-citizen rights! We thereby give it a true pipe character.

Note that the posterior is "given X".

Bayes Rule Again

How does this update work? Through Bayes Rule!

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior}} = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \underbrace{\frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}}_{\text{update factor}} \underbrace{\mathbb{P}(\theta)}_{\text{prior}}$$

The l.h.s. is a density for θ under your data. You can query it with ideas about what you think θ is and it returns probabilities (or probability densities). The update factor is how likely the data is under an idea of θ out of all possibilities for X. If the data supports a given θ, then the update factor is greater than 1, otherwise less than 1. Note that the update factor is the likelihood $\mathbb{P}(X \mid \theta)$ divided by $\mathbb{P}(X)$, a term called the "prior predictive distribution" which we will not focus on too much (you'll see why later).

Technicalities

Previously we defined likelihood as $\mathbb{P}(X; \theta)$ but now it's $\mathbb{P}(X \mid \theta)$. Why did the semicolon become a pipe character? Previously θ was considered a fixed constant, a parameter in the function. But now it is its own random variable with full-citizen rights! We thereby give it a true pipe character.

Note that the posterior is "given X". We will return to this shortly after we speak about kernels.

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1.

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1. Thus if they're multiplied by a constant $1/c$, they'll sum (or integrate) to $1/c$.

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1. Thus if they're multiplied by a constant $1/c$, they'll sum (or integrate) to $1/c$. For instance the binomial PMFs sums to 1,

$$\sum_{x=0}^{n} \mathbb{P}(X) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = 1$$

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1. Thus if they're multiplied by a constant $1/c$, they'll sum (or integrate) to $1/c$. For instance the binomial PMFs sums to 1,

$$\sum_{x=0}^{n} \mathbb{P}(X) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} = 1$$

but the following quantity is equal to $1/n!$,

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1. Thus if they're multiplied by a constant $1/c$, they'll sum (or integrate) to $1/c$. For instance the binomial PMFs sums to 1,

$$\sum_{x=0}^{n} \mathbb{P}(X) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = 1$$

but the following quantity is equal to $1/n!$,

$$\sum_{x=0}^{n} \frac{1}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = 1/n!$$

Since $1/n!$ is not a function of the free variable x, we say that

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1. Thus if they're multiplied by a constant $1/c$, they'll sum (or integrate) to $1/c$. For instance the binomial PMFs sums to 1,

$$\sum_{x=0}^{n} \mathbb{P}(X) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = 1$$

but the following quantity is equal to $1/n!$,

$$\sum_{x=0}^{n} \frac{1}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = 1/n!$$

Since $1/n!$ is not a function of the free variable x, we say that

$$\frac{1}{x!(n-x)!}\theta^x(1-\theta)^{n-x} \propto \mathbb{P}(X)$$

Review: What is a kernel?

Recall from your probability class that a PMF (or PDF) sums (or integrates) to 1. Thus if they're multiplied by a constant $1/c$, they'll sum (or integrate) to $1/c$. For instance the binomial PMFs sums to 1,

$$\sum_{x=0}^{n} \mathbb{P}(X) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!}\theta^{x}(1-\theta)^{n-x} = 1$$

but the following quantity is equal to $1/n!$,

$$\sum_{x=0}^{n} \frac{1}{x!(n-x)!}\theta^{x}(1-\theta)^{n-x} = 1/n!$$

Since $1/n!$ is not a function of the free variable x, we say that

$$\frac{1}{x!(n-x)!}\theta^{x}(1-\theta)^{n-x} \propto \mathbb{P}(X)$$

The "proportionality" sign indicates it differs from the true PMF by only a constant which can be recovered by doing the sum.

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^{x} \propto \mathbb{P}(X)$$

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF.

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$!

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X = a)}{\mathbb{P}(X = b)}$$

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X=a)}{\mathbb{P}(X=b)}$$

since the constant will cancel.

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^{x} \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X=a)}{\mathbb{P}(X=b)}$$

since the constant will cancel.

Now let's return to Bayes Rule, note that:

$$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}\mathbb{P}(\theta) \propto$$

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^{x} \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X=a)}{\mathbb{P}(X=b)}$$

since the constant will cancel.

Now let's return to Bayes Rule, note that:

$$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X,\theta)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}\mathbb{P}(\theta) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) =$$

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X=a)}{\mathbb{P}(X=b)}$$

since the constant will cancel.

Now let's return to Bayes Rule, note that:

$$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X, \theta)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}\mathbb{P}(\theta) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta) = \text{likelihood } \times$$

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X=a)}{\mathbb{P}(X=b)}$$

since the constant will cancel.

Now let's return to Bayes Rule, note that:

$$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X,\theta)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}\mathbb{P}(\theta) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta) = \text{likelihood} \times \text{prior}$$

Review: What is a kernel?

We can go further and remove all constant multiples that do not depend on x to obtain the "kernel" which we'll denote $k(x)$,

$$k(x) := \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x \propto \mathbb{P}(X)$$

Kernels are interesting... if we find them, we have quantity proportion to $\mathbb{P}(X)$ and therefore we know it is the same as the true PMF. So if you see $k(x)$ above you know $X \sim \mathrm{Binomial}(n, \theta)$! They also can be used to calculate proportions of probabilities:

$$\frac{k(a)}{k(b)} = \frac{\mathbb{P}(X=a)}{\mathbb{P}(X=b)}$$

since the constant will cancel.

Now let's return to Bayes Rule, note that:

$$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X,\theta)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X \mid \theta)}{\mathbb{P}(X)}\mathbb{P}(\theta) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) = \text{likelihood } \times \text{ prior}$$

(since $1/\mathbb{P}(X)$ is clearly a constant and not a function of the free variable θ).

The prior

We know what the likelihood is from before, but what is the prior?

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data.

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly".

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly". First of all, what could θ be in our batting average problem? It's the probability of getting a hit, the parameter in the underlying Bernoulli.

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly". First of all, what could θ be in our batting average problem? It's the probability of getting a hit, the parameter in the underlying Bernoulli. It should be fair to say that every possibly parameter value should be represented in a prior, thus

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly". First of all, what could θ be in our batting average problem? It's the probability of getting a hit, the parameter in the underlying Bernoulli. It should be fair to say that every possibly parameter value should be represented in a prior, thus

$$\mathbb{S}\text{upp}[\theta] = \text{the parameter space of the Bernoulli}$$

which is a link between parameter spaces and supports.

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly". First of all, what could θ be in our batting average problem? It's the probability of getting a hit, the parameter in the underlying Bernoulli. It should be fair to say that every possibly parameter value should be represented in a prior, thus

$$\mathbb{S}\text{upp}\left[\theta\right] = \text{the parameter space of the Bernoulli}$$

which is a link between parameter spaces and supports.

Now if we want to be as "objective" as possible or to be as "uninformative" or as "indifferent" as possible, we can use a "uniform" prior (also called a "reference" prior).

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly". First of all, what could θ be in our batting average problem? It's the probability of getting a hit, the parameter in the underlying Bernoulli. It should be fair to say that every possibly parameter value should be represented in a prior, thus

$$\mathbb{S}\text{upp}\left[\theta\right] = \text{the parameter space of the Bernoulli}$$

which is a link between parameter spaces and supports.

Now if we want to be as "objective" as possible or to be as "uninformative" or as "indifferent" as possible, we can use a "uniform" prior (also called a "reference" prior). This is five names for the same thing:

$$\mathbb{P}(\theta) = \mathrm{U}(0,\,1) = \begin{cases} 1 & \text{if} \quad \theta \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

The prior

We know what the likelihood is from before, but what is the prior? $\mathbb{P}(\theta)$ is your belief over all θ of what θ is before you see data. This is a weird concept! So let's "break it in slowly". First of all, what could θ be in our batting average problem? It's the probability of getting a hit, the parameter in the underlying Bernoulli. It should be fair to say that every possibly parameter value should be represented in a prior, thus

$$\mathbb{S}\text{upp}[\theta] = \text{the parameter space of the Bernoulli}$$

which is a link between parameter spaces and supports.

Now if we want to be as "objective" as possible or to be as "uninformative" or as "indifferent" as possible, we can use a "uniform" prior (also called a "reference" prior). This is five names for the same thing:

$$\mathbb{P}(\theta) = \mathrm{U}(0,\,1) = \begin{cases} 1 & \text{if} \quad \theta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

This means no matter what the value of θ, we don't think any of them are more likely than any other!

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)$.

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$.

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)(1)$ and for our binomial likelihood model,

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$$

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$$

Since we are given the data X and we know n as well by assumption,

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Since we are given the data X and we know n as well by assumption, we can make the kernel a bit tighter:

$$\mathbb{P}(\theta \mid X) \propto \theta^x(1-\theta)^{n-x} := k(\theta \mid X)$$

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Since we are given the data X and we know n as well by assumption, we can make the kernel a bit tighter:

$$\mathbb{P}(\theta \mid X) \propto \theta^x(1-\theta)^{n-x} := k(\theta \mid X)$$

We know that $k(\theta \mid X)$ must correspond to a density of θ which is a constant c times the kernel. To find its constant, we integrate over the parameter space,

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Since we are given the data X and we know n as well by assumption, we can make the kernel a bit tighter:

$$\mathbb{P}(\theta \mid X) \propto \theta^x(1-\theta)^{n-x} := k(\theta \mid X)$$

We know that $k(\theta \mid X)$ must correspond to a density of θ which is a constant c times the kernel. To find its constant, we integrate over the parameter space,

$$\frac{1}{c} := \int_{\theta\in[0,1]} \theta^x(1-\theta)^{n-x}dx$$

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Since we are given the data X and we know n as well by assumption, we can make the kernel a bit tighter:

$$\mathbb{P}(\theta \mid X) \propto \theta^x(1-\theta)^{n-x} := k(\theta \mid X)$$

We know that $k(\theta \mid X)$ must correspond to a density of θ which is a constant c times the kernel. To find its constant, we integrate over the parameter space,

$$\frac{1}{c} := \int_{\theta\in[0,1]} \theta^x(1-\theta)^{n-x}dx$$

This is a very famous integral!

Posterior Kernel

Previously we showed that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)$. With the uniform prior we just discussed, $\mathbb{P}(\theta) = 1$. This means that $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)(1)$ and for our binomial likelihood model,

$$\mathbb{P}(\theta \mid X) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Since we are given the data X and we know n as well by assumption, we can make the kernel a bit tighter:

$$\mathbb{P}(\theta \mid X) \propto \theta^x(1-\theta)^{n-x} := k(\theta \mid X)$$

We know that $k(\theta \mid X)$ must correspond to a density of θ which is a constant c times the kernel. To find its constant, we integrate over the parameter space,

$$\frac{1}{c} := \int_{\theta\in[0,1]} \theta^x(1-\theta)^{n-x}dx$$

This is a very famous integral! It is known as the beta integral and its solution is the non-computable (but approximable) beta function.

The Beta Function

The beta function is:

$$B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$$

The Beta Function

The beta function is:

$$B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$$

So the integral from the previous page becomes:

$$B(x+1, n-x+1) = \int_{\theta \in [0,1]} \theta^x (1-\theta)^{n-x} dx$$

The Beta Function

The beta function is:

$$B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$$

So the integral from the previous page becomes:

$$B(x+1, n-x+1) = \int_{\theta \in [0,1]} \theta^x (1-\theta)^{n-x} dx$$

and thus the density becomes (since the l.h.s is $1/c$),

$$\mathbb{P}(\theta \mid X) = \frac{1}{B(x+1, n-x+1)} \theta^x (1-\theta)^{n-x}$$

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\,\beta) := \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\, \beta) := \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

where $\mathbb{S}\text{upp}\,[Y] = (0, 1)$

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\, \beta) := \frac{1}{B(\alpha, \beta)} y^{\alpha - 1}(1 - y)^{\beta - 1}$$

where $\mathbb{S}\mathrm{upp}\,[Y] = (0, 1)$ and the parameter space is $\alpha > 0$

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\, \beta) := \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

where $\mathbb{S}\mathrm{upp}\,[Y] = (0, 1)$ and the parameter space is $\alpha > 0$ and $\beta > 0$.

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\,\beta) := \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

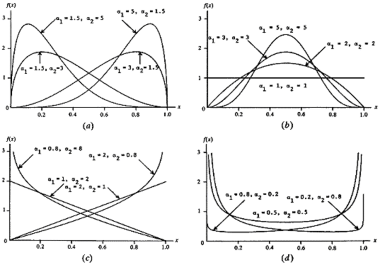
where $\mathbb{S}\mathrm{upp}\,[Y] = (0,1)$ and the parameter space is $\alpha > 0$ and $\beta > 0$. Since there are two parameters, there are many different shapes this can look like:

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\, \beta) := \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

where $\mathbb{S}\mathrm{upp}\,[Y] = (0, 1)$ and the parameter space is $\alpha > 0$ and $\beta > 0$. Since there are two parameters, there are many different shapes this can look like:

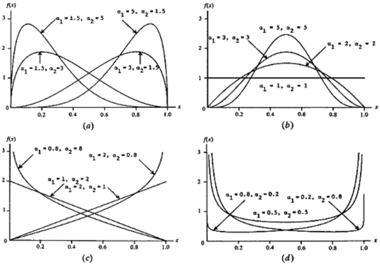


The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\,(\alpha,\,\beta) := \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

where $\mathbb{S}\mathrm{upp}\,[Y] = (0,1)$ and the parameter space is $\alpha > 0$ and $\beta > 0$. Since there are two parameters, there are many different shapes this can look like:



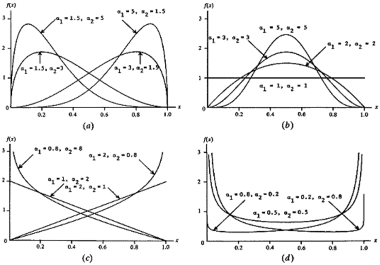
So, now we have moved from $\theta \sim U(0,1) \rightarrow$

The Beta Distribution

Just like the beta function, this is the density of a famous continuous r.v. called the "beta",

$$Y \sim \mathrm{Beta}\left(\alpha,\, \beta\right) := \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

where $\mathbb{S}\mathrm{upp}\left[Y\right] = (0, 1)$ and the parameter space is $\alpha > 0$ and $\beta > 0$. Since there are two parameters, there are many different shapes this can look like:



So, now we have moved from $\theta \sim U(0, 1) \to \theta \mid X \sim \mathrm{Beta}\left(x+1,\, n-x+1\right)$ using Bayesian conditionalism.

Posterior — what is it?

So we have the posterior. What can we do with it?

Posterior — what is it?

So we have the posterior. What can we do with it? Anything you want!

Posterior — what is it?

So we have the posterior. What can we do with it? Anything you want! What if you want to know the probability the batting average is under 200?

Posterior — what is it?

So we have the posterior. What can we do with it? Anything you want! What if you want to know the probability the batting average is under 200? Then,

$$\mathbb{P}\left(\theta \leq 0.2 \mid X\right) = \int_{0}^{0.2} \frac{1}{B(x+1, n-x+1)} \theta^{x} (1-\theta)^{n-x} d\theta$$

Posterior — what is it?

So we have the posterior. What can we do with it? Anything you want! What if you want to know the probability the batting average is under 200? Then,

$$\mathbb{P}\left(\theta \leq 0.2 \mid X\right) = \int_0^{0.2} \frac{1}{B(x+1, n-x+1)} \theta^x (1-\theta)^{n-x} d\theta$$

where the integral can be figured out by numerical integration using R via the command `pbeta(0.2, x+1, n-x+1)`.

Posterior — what is it?

So we have the posterior. What can we do with it? Anything you want! What if you want to know the probability the batting average is under 200? Then,

$$\mathbb{P}(\theta \leq 0.2 \mid X) = \int_0^{0.2} \frac{1}{B(x+1, n-x+1)} \theta^x (1-\theta)^{n-x} d\theta$$

where the integral can be figured out by numerical integration using R via the command `pbeta(0.2, x+1, n-x+1)`. Any probability can now be queried about θ!

Posterior — what is it?

So we have the posterior. What can we do with it? Anything you want! What if you want to know the probability the batting average is under 200? Then,

$$\mathbb{P}\left(\theta \leq 0.2 \mid X\right) = \int_{0}^{0.2} \frac{1}{B(x+1, n-x+1)} \theta^{x}(1-\theta)^{n-x} d\theta$$

where the integral can be figured out by numerical integration using R via the command `pbeta(0.2, x+1, n-x+1)`. Any probability can now be queried about θ! This was not possible with the frequentist / classical approach.

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ.

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution?

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess?

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

$$\mathbb{E}[Y] =$$

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^1 y \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy \\ &= \end{aligned}$$

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^1 y \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 y^{\alpha}(1-y)^{\beta-1} dy \\
&=
\end{aligned}$$

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^1 y \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} dy \\
&= \frac{1}{B(\alpha, \beta)} \int_0^1 y^{\alpha}(1-y)^{\beta-1} dy \\
&= \frac{B(\alpha-1, \beta)}{B(\alpha, \beta)} =
\end{aligned}$$

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^1 y \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 y^{\alpha}(1-y)^{\beta-1} dy \\ &= \frac{B(\alpha-1,\beta)}{B(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta} \end{aligned}$$

where the last equality is due to the beta function being expressible as the gamma function

Bayesian Estimate

We have the whole distribution $\mathbb{P}(\theta \mid X)$ but what if you were forced to give one answer as to your best guess of θ. Recall the MLE was

$$\hat{\theta}_{\mathrm{MLE}} = \bar{x} = \frac{x}{n}$$

But what do we do with a whole distribution? What is the best guess? Why not what the average guess is? That's what the expectation is, so let's compute the expectation for the beta density,

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^1 y \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 y^{\alpha}(1-y)^{\beta-1} dy \\
&= \frac{B(\alpha-1,\beta)}{B(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

where the last equality is due to the beta function being expressible as the gamma function and properties of the gamma function.

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

which is called the "Bayes-Laplace rule inverse probability".

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

which is called the "Bayes-Laplace rule inverse probability". Note that this is very similar to the MLE, $\frac{x}{n}$.

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

which is called the "Bayes-Laplace rule inverse probability". Note that this is very similar to the MLE, $\frac{x}{n}$.

Technicalities

There are other Bayesian estimates regularly in use which we won't discuss further.

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

which is called the "Bayes-Laplace rule inverse probability". Note that this is very similar to the MLE, $\frac{x}{n}$.

Technicalities

There are other Bayesian estimates regularly in use which we won't discuss further.

- The maximum a posteriori i.e. the mode of the posterior is $\arg\max_{\theta} \{\mathbb{P}(\theta \mid X)\}$

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

which is called the "Bayes-Laplace rule inverse probability". Note that this is very similar to the MLE, $\frac{x}{n}$.

Technicalities

There are other Bayesian estimates regularly in use which we won't discuss further.

- The maximum a posteriori i.e. the mode of the posterior is $\arg\max_\theta \{\mathbb{P}(\theta \mid X)\}$
- The minimum absolute error i.e. the median of the posterior

The Posterior Expectation

Thus in our setup, with $\theta \sim \mathrm{U}(0, 1)$, our Bayesian estimate, using the posterior expectation we just derived, is

$$\mathbb{E}[\theta \mid X] = \frac{(x+1)}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

which is called the "Bayes-Laplace rule inverse probability". Note that this is very similar to the MLE, $\frac{x}{n}$.

Technicalities

There are other Bayesian estimates regularly in use which we won't discuss further.

- The maximum a posteriori i.e. the mode of the posterior is $\arg\max_{\theta} \{\mathbb{P}(\theta \mid X)\}$
- The minimum absolute error i.e. the median of the posterior

We use the posterior expectation because it minimizes the squared error loss from the truth θ and this is usually an appropriate loss function unless there are other considerations.

Credible Intervals

What if we don't need a single point estimate but instead want the equivalent of the 95% confidence interval?

Credible Intervals

What if we don't need a single point estimate but instead want the equivalent of the 95% confidence interval? We want to provide an interval of possible θ values where there is 95% probability the true θ resides inside.

Credible Intervals

What if we don't need a single point estimate but instead want the equivalent of the 95% confidence interval? We want to provide an interval of possible θ values where there is 95% probability the true θ resides inside. The easiest way to do this is to use the posterior and return the 2.5% quantile and the 97.5% quantile (i.e. the "middle" 95% of the distribution) via a numerical solver algorithm.

Credible Intervals

What if we don't need a single point estimate but instead want the equivalent of the 95% confidence interval? We want to provide an interval of possible θ values where there is 95% probability the true θ resides inside. The easiest way to do this is to use the posterior and return the 2.5% quantile and the 97.5% quantile (i.e. the "middle" 95% of the distribution) via a numerical solver algorithm. This can be done in R via the following commands:

```
qbeta(0.025, x + 1, n - x + 1)
qbeta(0.975, x + 1, n - x + 1)
```

The Beta Prior

Note that our prior, $\theta \sim U(0, 1)$ can be thought of as

$$\theta \sim \text{Beta}\,(1,\, 1)$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1}$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1)\;=\frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1}=\frac{1}{(1)}(1)(1)=1$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1)\; = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) \;=\frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more “options” than just a $U(0,1)$?

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) \;=\; \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

$$\mathbb{P}\,(\theta \mid X) \quad \propto \quad \mathbb{P}\,(X \mid \theta)\,\mathbb{P}\,(\theta)$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) \;=\; \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

$$\begin{aligned} \mathbb{P}\,(\theta \mid X) \quad &\propto \quad \mathbb{P}\,(X \mid \theta)\,\mathbb{P}\,(\theta) \\ &\propto \quad \left(\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right) \end{aligned}$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) \;=\frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1}=\frac{1}{(1)}(1)(1)=1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

$$\begin{aligned}
\mathbb{P}\,(\theta \mid X) \quad &\propto \quad \mathbb{P}\,(X \mid \theta)\,\mathbb{P}\,(\theta) \\
&\propto \quad \left(\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right)\left(\frac{1}{B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}\right)
\end{aligned}$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

$$\begin{aligned} \mathbb{P}\,(\theta \mid X) &\propto \mathbb{P}\,(X \mid \theta)\,\mathbb{P}\,(\theta) \\ &\propto \left(\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right)\left(\frac{1}{B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}\right) \\ &\propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

$$\begin{aligned}
\mathbb{P}\,(\theta \mid X) &\propto \mathbb{P}\,(X \mid \theta)\,\mathbb{P}\,(\theta) \\
&\propto \left(\binom{n}{x}\theta^x(1-\theta)^{n-x}\right)\left(\frac{1}{B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}\right) \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}
\end{aligned}$$

Where this is now the kernel of the beta!

The Beta Prior

Note that our prior, $\theta \sim U(0,1)$ can be thought of as

$$\theta \sim \mathrm{Beta}\,(1,\,1) = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{(1)}(1)(1) = 1 \overset{d}{=} U(0,1)$$

Could it be the we can generalize our prior to a beta and thereby have more "options" than just a $U(0,1)$? Let's see, let's let $\theta \sim \mathrm{Beta}\,(\alpha,\,\beta)$,

$$\begin{aligned}
\mathbb{P}\left(\theta \mid X\right) &\propto \mathbb{P}\left(X \mid \theta\right)\mathbb{P}\left(\theta\right) \\
&\propto \left(\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right)\left(\frac{1}{B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}\right) \\
&\propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}
\end{aligned}$$

Where this is now the kernel of the beta! So,

$$\theta \mid X \sim \mathrm{Beta}\,(x+\alpha,\; n-x+\beta)$$

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta,

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta!

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model".

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior?

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this).

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}[\theta \mid X] = \frac{x + \alpha}{n + \alpha + \beta}$$

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}[\theta \mid X] = \frac{x+\alpha}{n+\alpha+\beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$.

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}[\theta \mid X] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$. If we don't have any "prior information" we have $\alpha = 0$ and $\beta = 0$ and we get back the MLE.

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$. If we don't have any "prior information" we have $\alpha = 0$ and $\beta = 0$ and we get back the MLE. This is known as the "Haldane Prior"

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}[\theta \mid X] = \frac{x+\alpha}{n+\alpha+\beta} \neq \frac{x}{n} = \hat{\theta}_{\text{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$. If we don't have any "prior information" we have $\alpha = 0$ and $\beta = 0$ and we get back the MLE. This is known as the "Haldane Prior" and it is "improper"

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$. If we don't have any "prior information" we have $\alpha = 0$ and $\beta = 0$ and we get back the MLE. This is known as the "Haldane Prior" and it is "improper" since $\alpha = 0$ and $\beta = 0$ are not technically in the parameter space of the beta density.

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$. If we don't have any "prior information" we have $\alpha = 0$ and $\beta = 0$ and we get back the MLE. This is known as the "Haldane Prior" and it is "improper" since $\alpha = 0$ and $\beta = 0$ are not technically in the parameter space of the beta density. Note that whatever prior we pick, as long as $n \to \infty$, $\lim_{n \to \infty} \mathbb{E}\left[\theta \mid X\right] = \hat{\theta}_{\mathrm{MLE}}$

Beta is the "Conjugate Prior"

If we have the likelihood be a binomial and the prior, a beta, then the posterior is also a beta! This is known as "conjugacy" and we say "the beta is the conjugate prior for the binomial model". Although priors are still arbitrary, this gives us a feeling of "naturalness" about having the beta prior.

Then we have two choices to make, what is α and what is β in the prior? (Under the uniform, $\alpha = 1$ and $\beta = 1$ but now we are generalizing this). Let's take a look at the Bayesian point estimate, the posterior expectation,

$$\mathbb{E}[\theta \mid X] = \frac{x + \alpha}{n + \alpha + \beta} \neq \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$$

Here, α can be interpreted as the "pseudocount" number of previous hits in a pseudocount number of at-bats $\alpha + \beta$. If we don't have any "prior information" we have $\alpha = 0$ and $\beta = 0$ and we get back the MLE. This is known as the "Haldane Prior" and it is "improper" since $\alpha = 0$ and $\beta = 0$ are not technically in the parameter space of the beta density. Note that whatever prior we pick, as long as $n \to \infty$, $\lim_{n \to \infty} \mathbb{E}[\theta \mid X] = \hat{\theta}_{\mathrm{MLE}}$ so that is a nice property to know that our arbitrary prior matters less and less as we get more data.

Homework Problems

1. What is the Bayesian paradigm shift? Compare and contrast the frequentist view to the Bayesian view.
2. Under Bayesian Conditionalism, what changes from before to after?
3. What is the update factor?
4. Derive the kernel of the Binomial r.v.
5. Derive the kernel of the Gaussian r.v.
6. Derive the kernel of the Beta r.v.
7. What does it mean that the kernel is "proportional" to the density?

Homework Problems

8. Prove that the posterior is proportional to the likelihood times the prior.
9. Why is the support of the prior the same as the parameter space of the likelihood?
10. What is an objective prior? Why does it make sense?
11. What is the objective prior for θ in the binomial model?
12. What is the kernel of the posterior of the binomial-uniform model?
13. Derive the posterior and show it is a beta.
14. Create an integral that will compute the probability theta is greater than 0.4.

Homework Problems

15 If $x = 4$ and $n = 10$, compute the probability from (14) in `R`.

16 Compute a credible Interval / Region for the data in (15) using `R`.

17 Show that a standard uniform is a Beta(1,1).

18 Show that under binomial likelihood and a general beta prior, the posterior is beta and find its posterior parameters.

19 Derive the posterior expectation under (18).

20 Explain how α and β can be interpreted as "pseudocounts".

21 Explain how the posterior expectation limits to the MLE.

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior.

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an “objective” or “uninformative” prior. The logic was that if we choose this, we're “indifferent” to any θ.

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior. The logic was that if we choose this, we're "indifferent" to any θ. We can now solve the problem from before: two at bats and 0 hits or two hits, the estimates become

$$\mathbb{E}[\theta \mid X] = \frac{0+1}{2+2} = \frac{1}{4} = 0.25 \neq 0,$$

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior. The logic was that if we choose this, we're "indifferent" to any θ. We can now solve the problem from before: two at bats and 0 hits or two hits, the estimates become

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0+1}{2+2} = \frac{1}{4} = 0.25 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2+1}{2+2} = \frac{3}{4} = 0.75 \neq 1$$

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior. The logic was that if we choose this, we're "indifferent" to any θ. We can now solve the problem from before: two at bats and 0 hits or two hits, the estimates become

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0+1}{2+2} = \frac{1}{4} = 0.25 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2+1}{2+2} = \frac{3}{4} = 0.75 \neq 1$$

The prior $\theta \sim U(0,1)$ makes sense, but we know it's silly.

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior. The logic was that if we choose this, we're "indifferent" to any θ. We can now solve the problem from before: two at bats and 0 hits or two hits, the estimates become

$$\mathbb{E}[\theta \mid X] = \frac{0+1}{2+2} = \frac{1}{4} = 0.25 \neq 0, \quad \mathbb{E}[\theta \mid X] = \frac{2+1}{2+2} = \frac{3}{4} = 0.75 \neq 1$$

The prior $\theta \sim U(0,1)$ makes sense, but we know it's silly. If $\theta = 0.05$, that's a really bad hitter and he probably wouldn't make it to the major leagues!

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior. The logic was that if we choose this, we're "indifferent" to any θ. We can now solve the problem from before: two at bats and 0 hits or two hits, the estimates become

$$\mathbb{E}[\theta \mid X] = \frac{0+1}{2+2} = \frac{1}{4} = 0.25 \neq 0, \quad \mathbb{E}[\theta \mid X] = \frac{2+1}{2+2} = \frac{3}{4} = 0.75 \neq 1$$

The prior $\theta \sim U(0,1)$ makes sense, but we know it's silly. If $\theta = 0.05$, that's a really bad hitter and he probably wouldn't make it to the major leagues! If $\theta = 0.5$ that's crazy since the best career batting average was Ty Cobb's at 0.366!

How to Choose a Smarter Prior

We explored $\theta \sim U(0,1)$ as an "objective" or "uninformative" prior. The logic was that if we choose this, we're "indifferent" to any θ. We can now solve the problem from before: two at bats and 0 hits or two hits, the estimates become

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0+1}{2+2} = \frac{1}{4} = 0.25 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2+1}{2+2} = \frac{3}{4} = 0.75 \neq 1$$

The prior $\theta \sim U(0,1)$ makes sense, but we know it's silly. If $\theta = 0.05$, that's a really bad hitter and he probably wouldn't make it to the major leagues! If $\theta = 0.5$ that's crazy since the best career batting average was Ty Cobb's at 0.366! So we shouldn't be "indifferent" across all values of θ because some are clearly improbable!

Empirical Bayes

Why not let our "previous knowledge" about what career batting averages to give us an "informed" prior?

Empirical Bayes

Why not let our "previous knowledge" about what career batting averages to give us an "informed" prior?

We now develop this idea using **this** blog post by David Robinson on 9/30/15.

Empirical Bayes

Why not let our "previous knowledge" about what career batting averages to give us an "informed" prior?

We now develop this idea using **this** blog post by David Robinson on 9/30/15. We first install necessary packages in R,

```
options(repos = structure(c(CRAN =
  "http://cran.revolutionanalytics.com/")))
tryCatch(library(dplyr),
  error = function(e){install.packages("dplyr")},
  finally = library(dplyr))
tryCatch(library(tidyr),
  error = function(e){install.packages("tidyr")},
  finally = library(tidyr))
tryCatch(library(Lahman),
  error = function(e){install.packages("Lahman")},
  finally = library(Lahman))
?Batting
```

We see this has Lahman's Baseball Database, 1871-2014, 99,846 player stints measuring 22 baseball statistics (we will only make use of "hits" and "at bats").

Load Data

Now we load up the data:

Load Data

Now we load up the data:

```
career <- Batting %>%
  filter(AB > 0) %>%
  anti_join(Pitching, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB)) %>%
  mutate(average = H / AB)
career <- Master %>%
  tbl_df() %>%
  select(playerID, nameFirst, nameLast) %>%
  unite(name, nameFirst, nameLast, sep = " ") %>%
  inner_join(career, by = "playerID") %>%
  select(-playerID)
```

and we can see a sample of what our manicured data frame looks like:

Load Data

Now we load up the data:

```
career <- Batting %>%
  filter(AB > 0) %>%
  anti_join(Pitching, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB)) %>%
  mutate(average = H / AB)
career <- Master %>%
  tbl_df() %>%
  select(playerID, nameFirst, nameLast) %>%
  unite(name, nameFirst, nameLast, sep = " ") %>%
  inner_join(career, by = "playerID") %>%
  select(-playerID)
```

and we can see a sample of what our manicured data frame looks like:

```
head(career)
```

The first colum is the baseball player's full name, followed by hits, at bats and the MLE.

Fit an Empirical Bayes Beta Prior

We can now use this historical data to form a prior about the batting average θ for estimating future batting averages when we're short on data for a new player.

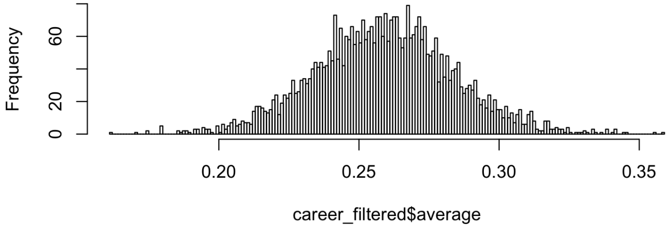
Fit an Empirical Bayes Beta Prior

We can now use this historical data to form a prior about the batting average θ for estimating future batting averages when we're short on data for a new player. So let's look at all historical batters with 500 or more at-bats.

Fit an Empirical Bayes Beta Prior

We can now use this historical data to form a prior about the batting average θ for estimating future batting averages when we're short on data for a new player. So let's look at all historical batters with 500 or more at-bats. That should give us a pretty good idea as to their career batting averages:

```
career_filtered <- career %>% filter(AB >= 500)
hist(career_filtered$average, br = 200)
```



Maximum Likelihood Estimates of α and β

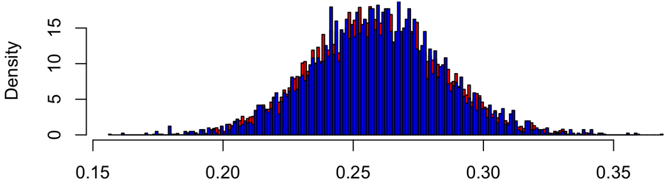
We now fit a beta to the historical θ estimates, plot atop and obtain the MLE's of α and β:

```
m <- MASS::fitdistr(career_filtered$average, dbeta,
                    start = list(shape1 = 1, shape2 = 10))
alpha0 <- m$estimate[1]
beta0 <- m$estimate[2]
round(alpha0, 1) # 79.0
round(beta0, 1) #225.9
hist(rbeta(10000, alpha0, beta0), br = 200, col = "red", prob = TRUE)
hist(career_filtered$average, br = 200, col = "blue", add = TRUE, prob = TRUE)
```

Maximum Likelihood Estimates of α and β

We now fit a beta to the historical θ estimates, plot atop and obtain the MLE's of α and β:

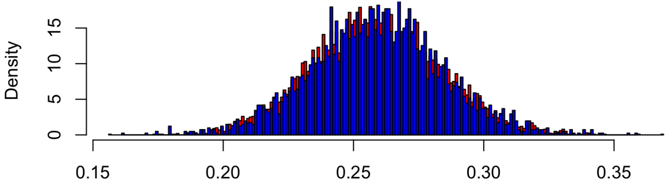
```
m <- MASS::fitdistr(career_filtered$average, dbeta,
                    start = list(shape1 = 1, shape2 = 10))
alpha0 <- m$estimate[1]
beta0 <- m$estimate[2]
round(alpha0, 1) # 79.0
round(beta0, 1) #225.9
hist(rbeta(10000, alpha0, beta0), br = 200, col = "red", prob = TRUE)
hist(career_filtered$average, br = 200, col = "blue", add = TRUE, prob = TRUE)
```



Maximum Likelihood Estimates of α and β

We now fit a beta to the historical θ estimates, plot atop and obtain the MLE's of α and β:

```
m <- MASS::fitdistr(career_filtered$average, dbeta,
                    start = list(shape1 = 1, shape2 = 10))
alpha0 <- m$estimate[1]
beta0 <- m$estimate[2]
round(alpha0, 1) # 79.0
round(beta0, 1) #225.9
hist(rbeta(10000, alpha0, beta0), br = 200, col = "red", prob = TRUE)
hist(career_filtered$average, br = 200, col = "blue", add = TRUE, prob = TRUE)
```



As we can see from the agreement of the red (the fit beta) and the blue (the true data), the $\theta \sim \mathrm{Beta}(79.0, 225.9)$ is a very nice fit for the data!

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0,$$

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0,\quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

We found that our 0/2 got shrunk up to 0.257

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

We found that our 0/2 got shrunk up to 0.257 and our 2/2 got shrunk down to 0.264.

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

We found that our 0/2 got shrunk up to 0.257 and our 2/2 got shrunk down to 0.264. "Shrinking" means moving towards the prior mean $\alpha/(\alpha + \beta) = 79.0/304.9 = 0.259$.

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

We found that our 0/2 got shrunk up to 0.257 and our 2/2 got shrunk down to 0.264. "Shrinking" means moving towards the prior mean $\alpha/(\alpha+\beta) = 79.0/304.9 = 0.259$. This is a **very** good strategy in practice —

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

We found that our 0/2 got shrunk up to 0.257 and our 2/2 got shrunk down to 0.264. "Shrinking" means moving towards the prior mean $\alpha/(\alpha + \beta) = 79.0/304.9 = 0.259$. This is a **very** good strategy in practice —if you don't have any data,

Empirical Bayes Estimates

So now let's use the prior and we get a posterior of:

$$\theta \mid X \sim \mathrm{Beta}\left(x + 79.0,\ n - x + 225.9\right)$$

where the expectation (best Bayesian estimate) is:

$$\mathbb{E}\left[\theta \mid X\right] = \frac{x + 79.0}{n + 304.9}$$

Now we can return to our problem, with two at bats, what are the best estimates?

$$\mathbb{E}\left[\theta \mid X\right] = \frac{0 + 79.0}{2 + 304.9} = 0.257 \neq 0, \quad \mathbb{E}\left[\theta \mid X\right] = \frac{2 + 79.0}{2 + 304.9} = 0.264 \neq 1$$

We found that our 0/2 got shrunk up to 0.257 and our 2/2 got shrunk down to 0.264. "Shrinking" means moving towards the prior mean $\alpha/(\alpha + \beta) = 79.0/304.9 = 0.259$. This is a **very** good strategy in practice —if you don't have any data, pretend this player is like the average player all throughout history!

Homework Problems

1. What is empirical Bayes estimation?
2. Repeat the exercise of building an empirical Bayes estimator with all players who had more than 600 at bats. What are the α and β MLE estimates?
3. Verify the beta density with the estimates from (2) is a good fit for the historical data.
4. What is the posterior mean?
5. Use your estimates from (2) to estimate the batting average of a hitter with 5 at bats and 3 hits.
6. What is the probability this hitter from (5) has a BA greater than 300?
7. Provide a credible region for the BA from the hitter of (5).
8. What kind of evidence would be required under (2) for a batter to have a BA estimate of 500?

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model
- The Objective / Reference / Uninformative Prior
- Posterior Distribution, Bayesian Estimate, Credible Intervals
- Empirical Bayes
- Estimating Batting Averages with `R`

Outline

- Random Variables, the Bernoulli and Binomial simple probability models
- Parameters, Estimators and Estimates, Maximum Likelihood
- The main problem with the Frequentist Estimator
- Bayesian Machinery for the binomial likelihood model
- The Objective / Reference / Uninformative Prior
- Posterior Distribution, Bayesian Estimate, Credible Intervals
- Empirical Bayes
- Estimating Batting Averages with `R`

You now know how to do all of this!

Outline for Next Module!

Outline for Next Module!

- Posterior Predictive Distributions, interevals and estimates — given a Bayesian setup, what will happen in the future?

Outline for Next Module!

- Posterior Predictive Distributions, interevals and estimates — given a Bayesian setup, what will happen in the future?
- Bayesian Hypothesis Testing — how to do tests on θ.

Outline for Next Module!

- Posterior Predictive Distributions, interevals and estimates — given a Bayesian setup, what will happen in the future?
- Bayesian Hypothesis Testing — how to do tests on θ.
- Hierarchical Models — we know the empirical Bayes models is wrong since BA changes over time and differs by field position of the player (e.g. pitches have lower BA's than outfielders).

Outline for Next Module!

- Posterior Predictive Distributions, interevals and estimates — given a Bayesian setup, what will happen in the future?
- Bayesian Hypothesis Testing — how to do tests on θ.
- Hierarchical Models — we know the empirical Bayes models is wrong since BA changes over time and differs by field position of the player (e.g. pitches have lower BA's than outfielders).
- Other models besides the beta-binomial model.

Outline for Next Module!

- Posterior Predictive Distributions, interevals and estimates — given a Bayesian setup, what will happen in the future?
- Bayesian Hypothesis Testing — how to do tests on θ.
- Hierarchical Models — we know the empirical Bayes models is wrong since BA changes over time and differs by field position of the player (e.g. pitches have lower BA's than outfielders).
- Other models besides the beta-binomial model.
- Cross-validation for assessing how good the Empirical-Bayes estimates are in practice for real data.