HEART: Statistics and Data Science With Networks

Joshua Agterberg

Johns Hopkins University

Fall 2021

Outline

Outline

1. Probability
2. Statistics
3. Probability and Statistics for Random Graphs Primer

Basic Probability

- Probability Theory formalizes the notion of chance in various problems
- We will not be discussing formal probability theory, which is typically covered in an AMS course.

Basic Probability

- Probability Theory formalizes the notion of chance in various problems
- We will not be discussing formal probability theory, which is typically covered in an AMS course.
- Instead, just know that probability assigns numbers between 0 and 1 to *events*

Basic Probability

- Probability Theory formalizes the notion of chance in various problems
- We will not be discussing formal probability theory, which is typically covered in an AMS course.
- Instead, just know that probability assigns numbers between 0 and 1 to *events*
- Most of this course will discuss simple probability

Basic Probability

- Probability Theory formalizes the notion of chance in various problems
- We will not be discussing formal probability theory, which is typically covered in an AMS course.
- Instead, just know that probability assigns numbers between 0 and 1 to *events*
- Most of this course will discuss simple probability
- In what follows, a *random variable* is just something whose outcome is random

Basic Probability

- Probability Theory formalizes the notion of chance in various problems
- We will not be discussing formal probability theory, which is typically covered in an AMS course.
- Instead, just know that probability assigns numbers between 0 and 1 to *events*
- Most of this course will discuss simple probability
- In what follows, a *random variable* is just something whose outcome is random
- Examples:
  - The outcome of a coin flip
  - The sum of rolling two dice
  - A graph with random edges

Bernoulli Random Variables

- A *Bernoulli* random variable X is a random variable satisfying

$$\mathbb{P}(X = 1) = p.$$

Bernoulli Random Variables

- A *Bernoulli* random variable X is a random variable satisfying

$$\mathbb{P}(X = 1) = p.$$

- Examples of Bernoulli random variables:
    - $X = 1$ if you flip a coin and it lands heads (what is p?)

Bernoulli Random Variables

- A *Bernoulli* random variable X is a random variable satisfying

$$\mathbb{P}(X = 1) = p.$$

- Examples of Bernoulli random variables:
  - $X = 1$ if you flip a coin and it lands heads (what is p?)
  - $X = 1$ if you roll a three on a single die (what is p?)

Bernoulli Random Variables

- A *Bernoulli* random variable X is a random variable satisfying

$$\mathbb{P}(X = 1) = p.$$

- Examples of Bernoulli random variables:
  - $X = 1$ if you flip a coin and it lands heads (what is p?)
  - $X = 1$ if you roll a three on a single die (what is p?)
  - $X = 1$ if you roll two dice and their sum is 7 (what is p?)

Binomial Random Variables

- A Binomial random variable is just the sum of n independent Bernoulli random variables with probability p

Binomial Random Variables

- A Binomial random variable is just the sum of n independent Bernoulli random variables with probability p
- Examples:
  - The number of times it rains in six days

Binomial Random Variables

- A Binomial random variable is just the sum of n independent Bernoulli random variables with probability p
- Examples:
  - The number of times it rains in six days
  - The number of times your coin flip lands heads in 12 flips

Binomial Random Variables

- A Binomial random variable is just the sum of n independent Bernoulli random variables with probability p
- Examples:
  - The number of times it rains in six days
  - The number of times your coin flip lands heads in 12 flips
  - Nonexample: the number of coin flips needed to get 12 heads (why?)

Normal (Gaussian) Random Variables

- A normal random variable is a random variable that formalizes the notion of *bell curve*

Normal (Gaussian) Random Variables

- A normal random variable is a random variable that formalizes the notion of *bell curve*
- It has two parameters: the mean μ and standard deviation σ

Normal (Gaussian) Random Variables

- A normal random variable is a random variable that formalizes the notion of *bell curve*
- It has two parameters: the mean μ and standard deviation σ
- Plays a central role in probability theory, statistics, and even partial differential equations

Normal (Gaussian) Random Variables

- A normal random variable is a random variable that formalizes the notion of *bell curve*
- It has two parameters: the mean μ and standard deviation σ
- Plays a central role in probability theory, statistics, and even partial differential equations
- A Normal random variable is a *continuous* random variable, meaning $\mathbb{P}(X = c) = 0$ for any real number c

Normal (Gaussian) Random Variables

- A normal random variable is a random variable that formalizes the notion of *bell curve*
- It has two parameters: the mean μ and standard deviation σ
- Plays a central role in probability theory, statistics, and even partial differential equations
- A Normal random variable is a *continuous* random variable, meaning $\mathbb{P}(X = c) = 0$ for any real number c
- Instead, the Gaussian distribution satisfies

$$\mathbb{P}(X \leq \mu) = \frac{1}{2}.$$

Gaussian Distribution

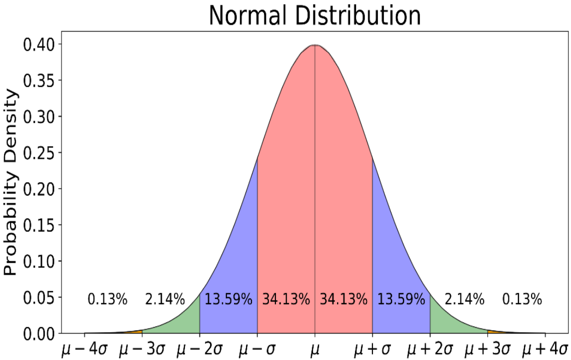

Figure: source

Other Random Variables/Distributions

- Poisson (counts)
- Exponential (times)
- Geometric (first time something happens)
- Gamma, Weibull, uniform
- ...

Expected Value

- The expected value formalizes the notion of a mean.

Expected Value

- The expected value formalizes the notion of a mean.
- The formula for a discrete random variable is

$$\mathbb{E}X = \sum_{\text{all values of } X} k\mathbb{P}(X = k).$$

Expected Value

- The expected value formalizes the notion of a mean.
- The formula for a discrete random variable is

$$\mathbb{E}X = \sum_{\text{all values of } X} k\mathbb{P}(X = k).$$

- Example:
  - Bernoulli distribution:

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

Variance

- Variance is a measure of *spread* of a random variable

Variance

- Variance is a measure of *spread* of a random variable
- Define the *second moment*:

$$\mathbb{E}X^2 = \sum_{\text{all values of } X} k^2 \mathbb{P}(X = k).$$

Variance

- Variance is a measure of *spread* of a random variable
- Define the *second moment*:

$$\mathbb{E}X^2 = \sum_{\text{all values of } X} k^2 \mathbb{P}(X = k).$$

- The variance is defined as:

$$\mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Variance

- Variance is a measure of *spread* of a random variable
- Define the *second moment*:

$$\mathbb{E}X^2 = \sum_{\text{all values of } X} k^2 \mathbb{P}(X = k).$$

- The variance is defined as:

$$\mathbb{E}X^2 - (\mathbb{E}X)^2.$$

- Bernoulli Distribution:

$$\begin{aligned} \mathbb{E}X^2 &= 1^2\mathbb{P}(X = 1) + 0^2\mathbb{P}(x = 0) \\ &= p \\ \implies \mathrm{Var}(X) &= p - p^2 \\ &= p(1 - p). \end{aligned}$$

Higher Moments

- Higher moments include the *skew* (third moment) and *kurtosis* (fourth moment)

Higher Moments

- Higher moments include the *skew* (third moment) and *kurtosis* (fourth moment)
- General formula:

$$\mathbb{E}X^p = \sum_{\text{all values of } X} k^p \mathbb{P}(X = k)$$

Independence

- Two events A and B are independent if $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B)$.

Independence

- Two events A and B are independent if $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Two random variables are independent if the above holds for all possible events.

Independence

- Two events A and B are independent if $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Two random variables are independent if the above holds for all possible events.
- Covariance:

$$\mathrm{Cov}(X, Y) = \sum_{\text{all values } k \text{ of } X \text{ and } j \text{ of } Y} \Big\{ (k - \mathbb{E}X)(j - \mathbb{E}Y) \times \mathbb{P}(X = k \text{ and } Y = j) \Big\}$$

Independence

- Two events A and B are independent if $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Two random variables are independent if the above holds for all possible events.
- Covariance:

$$\mathrm{Cov}(X, Y) = \sum_{\text{all values } k \text{ of } X \text{ and } j \text{ of } Y} \Big\{ (k - \mathbb{E}X)(j - \mathbb{E}Y) \times \mathbb{P}(X = k \text{ and } Y = j) \Big\}$$

- Correlation:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov} X, Y}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Independence

- Two events A and B are independent if $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Two random variables are independent if the above holds for all possible events.
- Covariance:

$$\mathrm{Cov}(X, Y) = \sum_{\text{all values } k \text{ of } X \text{ and } j \text{ of } Y} \Big\{ (k - \mathbb{E}X)(j - \mathbb{E}Y) \times \mathbb{P}(X = k \text{ and } Y = j) \Big\}$$

- Correlation:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov} X, Y}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- Uncorrelated does not imply independent!

Outline

Statistics vs. Probability

- Probability studies properties of distributions assuming one *knows the distribution*

Statistics vs. Probability

- Probability studies properties of distributions assuming one *knows the distribution*
- Statistics studies how to *learn the distribution*

Statistics vs. Probability

- Probability studies properties of distributions assuming one *knows the distribution*
- Statistics studies how to *learn the distribution*
- Statistics needs tools from probability and vice versa (though slightly less so)

Statistics vs. Probability

- Probability studies properties of distributions assuming one *knows the distribution*
- Statistics studies how to *learn the distribution*
- Statistics needs tools from probability and vice versa (though slightly less so)
- In practice, we do not know the Bernoulli parameter p!

Statistics vs. Probability

- Probability studies properties of distributions assuming one *knows the distribution*
- Statistics studies how to *learn the distribution*
- Statistics needs tools from probability and vice versa (though slightly less so)
- In practice, we do not know the Bernoulli parameter p!
- How do we estimate it?

Estimators

- Suppose one has observations from a distribution with some parameter θ (example: Binomial distribution with parameter $\theta = p$)

Estimators

- Suppose one has observations from a distribution with some parameter θ (example: Binomial distribution with parameter $\theta = p$)
- An *estimator*, formally, is any function of the data, but really you want to be somewhat intelligent about it

Estimators

- Suppose one has observations from a distribution with some parameter θ (example: Binomial distribution with parameter $\theta = p$)
- An *estimator*, formally, is any function of the data, but really you want to be somewhat intelligent about it
- Sample mean for observations $\{X_i\}_{i=1}^n$:

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

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be iid (independent, identically distributed).

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be iid (independent, identically distributed).
- Let $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}(X)$.

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be iid (independent, identically distributed).
- Let $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}(X)$.
- Then as $n \to \infty$

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \to N(0, 1).$$

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be iid (independent, identically distributed).
- Let $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}(X)$.
- Then as $n \to \infty$

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \to N(0, 1).$$

- I.e. $\bar{X} \approx N(\mu, \sigma^2)$ (bell curve centered at μ with standard deviation σ)

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be iid (independent, identically distributed).
- Let $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}(X)$.
- Then as $n \to \infty$

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \to N(0, 1).$$

- I.e. $\bar{X} \approx N(\mu, \sigma^2)$ (bell curve centered at μ with standard deviation σ)
- R Example

Outline

Erdos-Renyi Random Graphs

- An undirected Erdos-Renyi random graph satisfies

$$\mathbb{P}(A_{ij} = 1) = p$$

independently for all $i < j$ with $A_{ji} = A_{ij}$ for $j < i$.

Erdos-Renyi Random Graphs

- An undirected Erdos-Renyi random graph satisfies

$$\mathbb{P}(A_{ij} = 1) = p$$

  independently for all $i < j$ with $A_{ji} = A_{ij}$ for $j < i$.
- Each edge is generated with probability p.

Erdos-Renyi Random Graphs

- An undirected Erdos-Renyi random graph satisfies

$$\mathbb{P}(A_{ij} = 1) = p$$

  independently for all $i < j$ with $A_{ji} = A_{ij}$ for $j < i$.
- Each edge is generated with probability p.
- What is the distribution of the degree of the i'th vertex?

Erdos-Renyi Random Graphs

- An undirected Erdos-Renyi random graph satisfies

$$\mathbb{P}(A_{ij} = 1) = p$$

  independently for all $i < j$ with $A_{ji} = A_{ij}$ for $j < i$.
- Each edge is generated with probability p.
- What is the distribution of the degree of the i'th vertex?
- How might we estimate p?

Erdos-Renyi Random Graphs

- An undirected Erdos-Renyi random graph satisfies

$$\mathbb{P}(A_{ij} = 1) = p$$

  independently for all $i < j$ with $A_{ji} = A_{ij}$ for $j < i$.
- Each edge is generated with probability p.
- What is the distribution of the degree of the i'th vertex?
- How might we estimate p?
- R Example

Erdos-Renyi Random Graphs

- What is $\mathbb{E}A$?

Erdos-Renyi Random Graphs

- What is $\mathbb{E}A$?

$$(\mathbb{E}A)_{ij} = p\ (i \neq j) \Longrightarrow \mathbb{E}A = \begin{pmatrix} 0 & p & p & \cdots & p \\ p & 0 & p & \cdots & p \\ p & p & 0 & \cdots & \vdots \\ p & \cdots & \cdots & \ddots & p \\ p & \cdots & \cdots & p & 0 \end{pmatrix}$$

Erdos-Renyi Random Graphs

- What is $\mathbb{E}A$?

$$(\mathbb{E}A)_{ij} = p \; (i \neq j) \Longrightarrow \mathbb{E}A = \begin{pmatrix} 0 & p & p & \cdots & p \\ p & 0 & p & \cdots & p \\ p & p & 0 & \cdots & \vdots \\ p & \cdots & \cdots & \ddots & p \\ p & \cdots & \cdots & p & 0 \end{pmatrix}$$

- What if we allow self-loops?

Erdos-Renyi Random Graphs

- What is $\mathbb{E}A$?

$$(\mathbb{E}A)_{ij} = p \; (i \neq j) \Longrightarrow \mathbb{E}A = \begin{pmatrix} 0 & p & p & \cdots & p \\ p & 0 & p & \cdots & p \\ p & p & 0 & \cdots & \vdots \\ p & \cdots & \cdots & \ddots & p \\ p & \cdots & \cdots & p & 0 \end{pmatrix}$$

- What if we allow self-loops?

$$\mathbb{E}A = \begin{pmatrix} p & \cdots & p \\ \vdots & \ddots & \vdots \\ p & \cdots & p \end{pmatrix} = p11^\top \qquad 1 = \text{vector of all ones}$$

Erdos-Renyi Random Graphs

- What is the rank of $\mathbb{E}A$ with self-loops? (rank $=$ # nonzero eigenvalues)

Erdos-Renyi Random Graphs

- What is the rank of $\mathbb{E}A$ with self-loops? (rank $=$ # nonzero eigenvalues)

$$p\mathbf{1}\mathbf{1}^\top\mathbf{1} = np\mathbf{1} \implies np \text{ is an eigenvalue, all others are zero}$$

Erdos-Renyi Random Graphs

- What is the rank of $\mathbb{E}A$ with self-loops? (rank $=$ # nonzero eigenvalues)

  $p \mathbf{1}\mathbf{1}^\top \mathbf{1} = np\mathbf{1} \implies np$ is an eigenvalue, all others are zero

- So if by CLT $A \approx \mathbb{E}A$, then maybe (nonzero) eigenvalues and eigenvectors of $A \approx$ eigenvalues of eigenvectors of $\mathbb{E}A$.

Stochastic Blockmodels

- Generalization of Erdos-Renyi Random Graph

Stochastic Blockmodels

- Generalization of Erdos-Renyi Random Graph
- Suppose there are K communities.

Stochastic Blockmodels

- Generalization of Erdos-Renyi Random Graph
- Suppose there are K communities.
- Vertex i and j belong to community l and k respectively.

Stochastic Blockmodels

- Generalization of Erdos-Renyi Random Graph
- Suppose there are K communities.
- Vertex i and j belong to community l and k respectively.
- Then A is a stochastic blockmodel if

$$\mathbb{P}(A_{ij} = 1) = B_{lk}.$$

Stochastic Blockmodels

- Generalization of Erdos-Renyi Random Graph
- Suppose there are K communities.
- Vertex i and j belong to community l and k respectively.
- Then A is a stochastic blockmodel if

$$\mathbb{P}(A_{ij} = 1) = B_{lk}.$$

- Special case: one community $B_{11} = p$

Stochastic Blockmodels

- Generalization of Erdos-Renyi Random Graph
- Suppose there are K communities.
- Vertex i and j belong to community l and k respectively.
- Then A is a stochastic blockmodel if

$$\mathbb{P}(A_{ij} = 1) = B_{lk}.$$

- Special case: one community $B_{11} = p$
- One community SBM = Erdos-Renyi

Stochastic Blockmodels

With self-loops, what is $\mathbb{E}A$? (assume two communities and organized by communities)

Stochastic Blockmodels

With self-loops, what is $\mathbb{E}A$? (assume two communities and organized by communities)

$$\mathbb{E}A = \begin{pmatrix} B_{11} & \cdots & B_{11} & B_{12} & \cdots & B_{12} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ B_{11} & \cdots & B_{11} & B_{12} & \cdots & B_{12} \\ B_{21} & \cdots & B_{21} & B_{22} & \cdots & B_{22} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ B_{21} & \cdots & B_{21} & B_{22} & \cdots & B_{22} \end{pmatrix}$$

Stochastic Blockmodels

With self-loops, what is $\mathbb{E}A$? (assume two communities and organized by communities)

$$\mathbb{E}A = \begin{pmatrix} B_{11} & \cdots & B_{11} & B_{12} & \cdots & B_{12} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ B_{11} & \cdots & B_{11} & B_{12} & \cdots & B_{12} \\ B_{21} & \cdots & B_{21} & B_{22} & \cdots & B_{22} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ B_{21} & \cdots & B_{21} & B_{22} & \cdots & B_{22} \end{pmatrix}$$

R Example