COMP 680
Statistics for Computing and Data Science
Week 1: Probability and Random Variable I

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

Your Instructor

- Su Chen
- Email: Su.Chen@rice.edu
- Office: Duncan 2051
- About me

Administrative Stuff

You can find everything on the Canvas website for this course

- The syllabus
- Weekly slides, code demo and quiz
- Bi-weekly homework assignments
- Announcements and emails...

Prerequisites

- College level calculus: single & multivariate
- Basics of linear algebra: vectors and matrices
- Comfortable with mathematical reasoning...
- Basic programming skills (or not afraid to learn :)

Learning Objectives

- Identify and calculate appropriate summary and inferential statistics based on data.

Learning Objectives

- Identify and calculate appropriate summary and inferential statistics based on data.
- Gain fluency in basic programming skills in Python with a focus on simulation-based inference and statistical modeling.

Learning Objectives

- Identify and calculate appropriate summary and inferential statistics based on data.
- Gain fluency in basic programming skills in Python with a focus on simulation-based inference and statistical modeling.
- Use applied statistical knowledge to analyze real-world data, test hypotheses, build regression models, and make scientific inference.

Learning Objectives

- Identify and calculate appropriate summary and inferential statistics based on data.
- Gain fluency in basic programming skills in Python with a focus on simulation-based inference and statistical modeling.
- Use applied statistical knowledge to analyze real-world data, test hypotheses, build regression models, and make scientific inference.
- Interpret inferences and modeling results in real-world contexts and communicate the findings effectively.

Course Work

- Weekly classes schedule: MW 2-3:15pm
  - Attendance
  - Participation
  - Feedback

Course Work

- Weekly classes schedule: MW 2-3:15pm
  - Attendance
  - Participation
  - Feedback
- Major assignments:
  - Bi-weekly homework assignments
  - Weekly quizzes
  - Midterm in-class exam
  - Final takehome exam

Grades

- Assignment: 45%
  - homework: 30%
  - quizzes: 15%

Grades

- Assignment: 45%
  - homework: 30%
  - quizzes: 15%
- Exams: 50%
  - midterm: 20%
  - final: 30%

Grades

- Assignment: 45%
  - homework: 30%
  - quizzes: 15%
- Exams: 50%
  - midterm: 20%
  - final: 30%
- Other: 5%
  - class participation and attendance: 3%
  - course exit survey: 2%

Etc.

- Disability accommodations.
- Tentative schedule.
- Syllabus is subject to change with reasonable advance notice by the instructor.

What is Data Science?

- A definition: draw useful information and actionable knowledge from data using computation: so that we can better understand the world and solve problems!

What is Data Science?

- A definition: draw useful information and actionable knowledge from data using computation: so that we can better understand the world and solve problems!
- math + stats + computing + **domain knowledge**

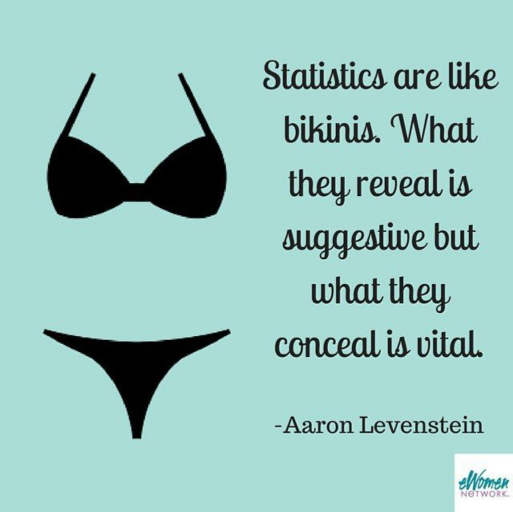
What is Data Science?

- A definition: draw useful information and actionable knowledge from data using computation: so that we can better understand the world and solve problems!
- math + stats + computing + **domain knowledge**
- A Brief History of Data Science

What is Data Science?

- A definition: draw useful information and actionable knowledge from data using computation: so that we can better understand the world and solve problems!
- math + stats + computing + **domain knowledge**
- A Brief History of Data Science
- Rooted in Statistics, powered by Computer Science, driven by application, interdisciplinary in nature
  - big data, cloud computing, data mining
  - machine learning/deep learning, artificial intelligence ...

What is Statistics?



What is Statistics??

- Origin of the word "statistics"
  - information useful to the state
  - early human statistical and census methods

What is Statistics??

- Origin of the word “statistics”
  - information useful to the state
  - early human statistical and census methods
- Wiki: history of statistics
  - Seven Pillars of Statistical Wisdom

What is Statistics??

- Origin of the word "statistics"
  - information useful to the state
  - early human statistical and census methods
- Wiki: history of statistics
  - Seven Pillars of Statistical Wisdom
- Statistics as a discipline:
  - procedures and techniques used to collect, process, and analyze the numerical data to make inferences and to reach appropriate decisions in situations of uncertainty

What is Statistics??

- Origin of the word "statistics"
  - information useful to the state
  - early human statistical and census methods
- Wiki: history of statistics
  - Seven Pillars of Statistical Wisdom
- Statistics as a discipline:
  - procedures and techniques used to collect, process, and analyze the numerical data to make inferences and to reach appropriate decisions in situations of uncertainty
- Statistics as problem solving: harnessing the power of uncertainty
  - descriptive / exploratory
  - predictive / decision-making
  - inferential / causal

Some Misconceptions

- Stats require hardcore math

Some Misconceptions

- Stats require hardcore math
  - applied stats is about reason with data

Some Misconceptions

- Stats require hardcore math
  - applied stats is about reason with data
- Stats need lot of data

Some Misconceptions

- Stats require hardcore math
  - applied stats is about reason with data
- Stats need lot of data
  - bigger data $\neq$ better data

Some Misconceptions

- Stats require hardcore math
  - applied stats is about reason with data
- Stats need lot of data
  - bigger data $\neq$ better data
- Stats provide unambiguous results & unique solution

Some Misconceptions

- Stats require hardcore math
  - applied stats is about reason with data
- Stats need lot of data
  - bigger data $\neq$ better data
- Stats provide unambiguous results & unique solution
  - "I report, you decide."

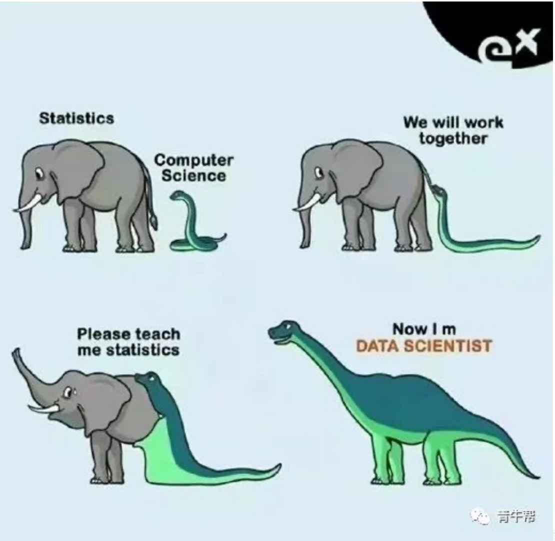
Some Misconceptions

- Stats require hardcore math
    - applied stats is about reason with data
- Stats need lot of data
    - bigger data $\neq$ better data
- Stats provide unambiguous results & unique solution
    - "I report, you decide."
- "Lies, damn lies, and statistics."

Some Misconceptions

- Stats require hardcore math
  - applied stats is about reason with data
- Stats need lot of data
  - bigger data $\neq$ better data
- Stats provide unambiguous results & unique solution
  - "I report, you decide."
- "Lies, damn lies, and statistics."
  - statistics don't lie, people do.

Statistics vs. Data Science



Statistics vs. Data Science

- “Statistics is the science of learning from data...” – ASA

Statistics vs. Data Science

- "Statistics is the science of learning from data..." – ASA
- Is data science just applied stats???

Statistics vs. Data Science

- "Statistics is the science of learning from data..." – ASA
- Is data science just applied stats???
    - origin of the field

Statistics vs. Data Science

- "Statistics is the science of learning from data..." – ASA
- Is data science just applied stats???
  - origin of the field
  - goals and the types of problems

Statistics vs. Data Science

- "Statistics is the science of learning from data..." – ASA
- Is data science just applied stats???
  - origin of the field
  - goals and the types of problems
  - data and technology involved

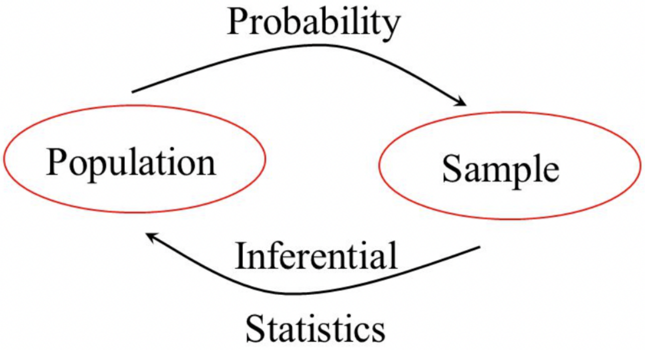
Statistics vs. Data Science

- "Statistics is the science of learning from data..." – ASA
- Is data science just applied stats???
  - origin of the field
  - goals and the types of problems
  - data and technology involved
- " fair to consider applied statistics as a subset of data science":
  - harvesting, processing, storing and cleaning are more central to data science than hard core statistics

Outline

Probability and Statistics

"Probability is the vehicle of statistics."



Probability Spaces

Definition: A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of:

- A sample space Ω, which contains all possible outcomes.
- A set of events $\mathcal{F}$, which must be a σ-algebra defined on Ω. $\mathcal{F}$ is a collection of subsets of Ω such that:

Probability Spaces

Definition: A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of:

- A sample space Ω, which contains all possible outcomes.
- A set of events $\mathcal{F}$, which must be a σ-algebra defined on Ω. $\mathcal{F}$ is a collection of subsets of Ω such that:
  - $\Omega \in \mathcal{F}$
  - if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
  - if $A_1, A_2, \cdots A_i, \cdots \in \mathcal{F}$, then $\cup_i A_i \in \mathcal{F}$
- A probability measure $\mathbb{P}$, which can assign probabilities to the events in $\mathcal{F}$.

Probability Spaces Example

- Flip a coin:
  - sample space Ω:
  - σ-algebra $\mathcal{F}$:
  - probability measure $\mathbb{P}$:

Probability Spaces Example

- Flip a coin:
    - sample space Ω:
    - σ-algebra $\mathcal{F}$:
    - probability measure $\mathbb{P}$:
- Generalize to any finite sample space:

Probability Measure

Definition: A probability measure $\mathbb{P}$ assigns each even $A \in \mathcal{F}$ to a real number, and it satisfies the following **three axioms**:

1. $0 \leq \mathbb{P}(A) \leq 1, \quad \forall A \in \mathcal{F}$

Probability Measure

Definition: A probability measure $\mathbb{P}$ assigns each even $A \in \mathcal{F}$ to a real number, and it satisfies the following **three axioms**:

1. $0 \leq \mathbb{P}(A) \leq 1, \quad \forall A \in \mathcal{F}$
2. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$

Probability Measure

Definition: A probability measure $\mathbb{P}$ assigns each even $A \in \mathcal{F}$ to a real number, and it satisfies the following **three axioms**:

1. $0 \leq \mathbb{P}(A) \leq 1, \quad \forall A \in \mathcal{F}$
2. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$
3. If $A_1, A_2, \cdots A_n \ldots$ are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Probability Measure

Definition: A probability measure $\mathbb{P}$ assigns each even $A \in \mathcal{F}$ to a real number, and it satisfies the following **three axioms**:

1. $0 \leq \mathbb{P}(A) \leq 1, \quad \forall A \in \mathcal{F}$
2. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$
3. If $A_1, A_2, \cdots A_n \ldots$ are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Based on 1 - 3, can you show $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$?

Conditional Probability

- Intuition:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \times \mathbb{P}(B) = \mathbb{P}(B|A) \times \mathbb{P}(A)$$

Conditional Probability

- Intuition:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \times \mathbb{P}(B) = \mathbb{P}(B|A) \times \mathbb{P}(A)$$

- **Bayes Theorem**:

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

Conditional Probability

- Intuition:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \times \mathbb{P}(B) = \mathbb{P}(B|A) \times \mathbb{P}(A)$$

- **Bayes Theorem**:

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

- **Law of total probability**: $A_1, A_2, \cdots$ be a partition of Ω

$$\mathbb{P}(S) = \sum_{i=1} \mathbb{P}(S \cap A_i) = \sum_{i=1} \mathbb{P}(S|A_i)\mathbb{P}(A_i)$$

Bayes Theorem

- Apply Bayes Theorem and (simple case) of Law of total probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

Bayes Theorem

- Apply Bayes Theorem and (simple case) of Law of total probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

- Example: Suppose 1% of population has certain cancer. A medical test can detect this cancer with 100% accuracy, but also has 1% false positive rate. If a person randomly selected tested positive, what is the probability this person actually has the cancer?

Independence

Definition: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are independent **if and only if**:

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

Independence

Definition: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are independent **if and only if**:

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

Equivalently, if A and B are independent then:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Conditional Independence

Definition: Two events $A, B \in \mathcal{F}$ are independent conditioned on another event C **if and only if**:

$$\mathbb{P}(A|B, C) = \mathbb{P}(A|C)$$

- It is possible that A and B are not independent, but conditioned on C they are (and vice versa!)

Conditional Independence

Definition: Two events $A, B \in \mathcal{F}$ are independent conditioned on another event C **if and only if**:

$$\mathbb{P}(A|B, C) = \mathbb{P}(A|C)$$

- It is possible that A and B are not independent, but conditioned on C they are (and vice versa!)
- A box contains two coins: a regular coin and one fake two-headed coin. I choose a coin at random and toss it twice. Define the following events:
    - $A=$ 1st coin toss is a head.
    - $B=$ 2nd coin toss is a head.
    - $C=$ the regular coin has been selected.

Monte Carlo Simulation

- How to calculate the probability without doing the math?
  - interpretation of probability: long-term (relative) frequency

Monte Carlo Simulation

- How to calculate the probability without doing the math?
  - interpretation of probability: long-term (relative) frequency
- Run a computer program to repeatedly simulate a random process!
  - Monte Carlo simulation
  - random number generator in computers

Monte Carlo Simulation

- How to calculate the probability without doing the math?
  - interpretation of probability: long-term (relative) frequency
- Run a computer program to repeatedly simulate a random process!
  - Monte Carlo simulation
  - random number generator in computers
- An approximation of the true probability
  - good enough approximation for “sufficient” repetition

Outline

Definition

What is a random variable?

- English: a variable whose value depends on some random event.

Definition

What is a random variable?

- English: a variable whose value depends on some random event.
- Mathy: given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X is a function that maps the random events in $\mathcal{F}$ to some real numbers $\mathbb{R}$.

Definition

What is a random variable?

- English: a variable whose value depends on some random event.
- Mathy: given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X is a function that maps the random events in $\mathcal{F}$ to some real numbers $\mathbb{R}$.
- Rigorous: in measure theory and advanced probability theory, for example here.

Definition

What is a random variable?

- English: a variable whose value depends on some random event.
- Mathy: given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X is a function that maps the random events in $\mathcal{F}$ to some real numbers $\mathbb{R}$.
- Rigorous: in measure theory and advanced probability theory, for example here.

Notation:

- capital letter X for the random variable
- lower case x for values the random variable can take

Discrete Random Variables

- A discrete random variable X can only take countable values in $\mathbb{R}$.

Discrete Random Variables

- A discrete random variable X can only take countable values in $\mathbb{R}$.
- A probability mass function (pmf) $P_X(x)$ is defined as:

$$P_X(x) = \mathbb{P}(X = x)$$

Discrete Random Variables

- A discrete random variable X can only take countable values in $\mathbb{R}$.
- A probability mass function (pmf) $P_X(x)$ is defined as:

$$P_X(x) = \mathbb{P}(X = x)$$

- Important discrete random variables:
  - Uniform
  - Bernoulli
  - Binomial
  - Poisson
  - See more here

Continuous Random Variables

- A continuous random variable X can take values in interval(s) in $\mathbb{R}$.

Continuous Random Variables

- A continuous random variable X can take values in interval(s) in $\mathbb{R}$.
- A cumulative distribution function (cdf) $F_X(x)$ is defined as:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Continuous Random Variables

- A continuous random variable X can take values in interval(s) in $\mathbb{R}$.
- A cumulative distribution function (cdf) $F_X(x)$ is defined as:

$$F_X(x) = \mathbb{P}(X \leq x)$$

- A probability density function (pdf) $f_X(x)$ is defined as:

$$f_X(x) = F_X'(x)$$

for all points x at which $F_X(x)$ is differentiable.

Continuous Random Variables

Some properties of cdf $F_X(x)$ and pdf $f_X(x)$:

Continuous Random Variables

Some properties of cdf $F_X(x)$ and pdf $f_X(x)$:

- $0 \leq F_X(x) \leq 1, \quad F_X(-\infty) = 0, \quad F_X(\infty) = 1,$

Continuous Random Variables

Some properties of cdf $F_X(x)$ and pdf $f_X(x)$:

- $0 \leq F_X(x) \leq 1, \quad F_X(-\infty) = 0, \quad F_X(\infty) = 1,$
- $F_X(x)$ is non-decreasing: $F_X(x_1) \leq F_X(x_2)$ for all $x_1 \leq x_2$

Continuous Random Variables

Some properties of cdf $F_X(x)$ and pdf $f_X(x)$:

- $0 \leq F_X(x) \leq 1, \quad F_X(-\infty) = 0, \quad F_X(\infty) = 1,$
- $F_X(x)$ is non-decreasing: $F_X(x_1) \leq F_X(x_2)$ for all $x_1 \leq x_2$
- $f_X(x)$ is non-negative: $f_X(x) \geq 0$ for all $x \in \mathbb{R}$

Continuous Random Variables

Some properties of cdf $F_X(x)$ and pdf $f_X(x)$:

- $0 \leq F_X(x) \leq 1, \quad F_X(-\infty) = 0, \quad F_X(\infty) = 1,$
- $F_X(x)$ is non-decreasing: $F_X(x_1) \leq F_X(x_2)$ for all $x_1 \leq x_2$
- $f_X(x)$ is non-negative: $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_X(x) = 1$

Continuous Random Variables

- Important continuous random variables:
  - Uniform
  - Normal (Gaussian)
  - Student-t
  - Cauchy
  - Exponential
  - Gamma
  - Beta
  - χ^2
- See more here

Functions of Random Variables

- Let X be a random variable and $Y = g(X)$ be another random variable defined by a function $g : \mathbb{R} \to \mathbb{R}$.

Functions of Random Variables

- Let X be a random variable and $Y = g(X)$ be another random variable defined by a function $g : \mathbb{R} \to \mathbb{R}$.
- The cdf of Y is given by:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \int_{\{x|g(x)\leq y\}} f_X(x)dx$$

Functions of Random Variables

- Let X be a random variable and $Y = g(X)$ be another random variable defined by a function $g : \mathbb{R} \to \mathbb{R}$.
- The cdf of Y is given by:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \int_{\{x|g(x)\leq y\}} f_X(x)dx$$

- If g is a strictly increasing function, then

$$F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)dx = F_X(g^{-1}(y))$$

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)$$

Generating Random Variables

- Let X be a random variable with a continuous CDF $F_X(x)$ and define the random variable $Y = F_X(x)$. Can you show $Y \sim U[0, 1]$?

Generating Random Variables

- Let X be a random variable with a continuous CDF $F_X(x)$ and define the random variable $Y = F_X(x)$. Can you show $Y \sim U[0, 1]$?
- This is how most computer programs are generating random variables of common distributions:
    - sample from Uniform [0, 1] $u_1, u_2, \cdots u_n$
    - get $x_i = F_X^{-1}(u_i)$ for a specific distribution with cdf $F_X(\cdot)$

Outline

Joint Distributions

A joint probability mass function:

	Y=0	Y=1	
X=0	1/10	2/10	3/10
X=1	3/10	4/10	7/10
	4/10	6/10	1

- Marginal distributions of X and Y?

Joint Distributions

A joint probability mass function:

	Y=0	Y=1	
X=0	1/10	2/10	3/10
X=1	3/10	4/10	7/10
	4/10	6/10	1

- Marginal distributions of X and Y?
- Conditional distribution of $X|Y=0$ and $Y|X=1$?

Joint Distributions

A joint probability density function:

$$f_{X,Y}(x,y) = \begin{cases} x + y & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Marginal distributions of X and Y?

Joint Distributions

A joint probability density function:

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Marginal distributions of X and Y?
- Conditional distribution of $X|Y=1$?

Independence and IID Samples

- Two random variables X and Y are independent **if and only if** the joint pdf $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all values x and y.

Independence and IID Samples

- Two random variables X and Y are independent **if and only if** the joint pdf $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ for all values x and y.
- If $X_1, X_2, \cdots X_n$ are independent and each has the same marginal distribution of pdf $f_X(\cdot)$, we say that $X_1, X_2, \cdots X_n$ are IID (independent and identically distributed) and has joint pdf:

$$f_{X_1,X_2,\cdots X_n}(x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} f_X(x_i)$$

Outline

Python Jupyter Notebook

- Birthday problem.
- Monty Hall problem.
- Su's favorite probability brain teaser.

Python Jupyter Notebook

- Birthday problem.
- Monty Hall problem.
- Su's favorite probability brain teaser.
- Assignment this week:
  - Install Python 3 and Jupyter Notebook through Anaconda
    - install tutorial for mac and window
  - Run the demo notebook and make sure nothing breaks.
  - Homework 1 posted, due in 2 weeks, please start early!