Introduction to Statistics

Dr. Supaporn Erjongmanee

Department of Computer Engineering
Kasetsart University
fengspe@ku.ac.th

Outline

- Introduction
- Review
  - Random variables
  - Discrete random variables
  - Continuous random variables

What is Statistics?

- Study of collection, analysis, interpretation, presentation, and organization of data [1]
- A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data [2]
- The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample. [3]

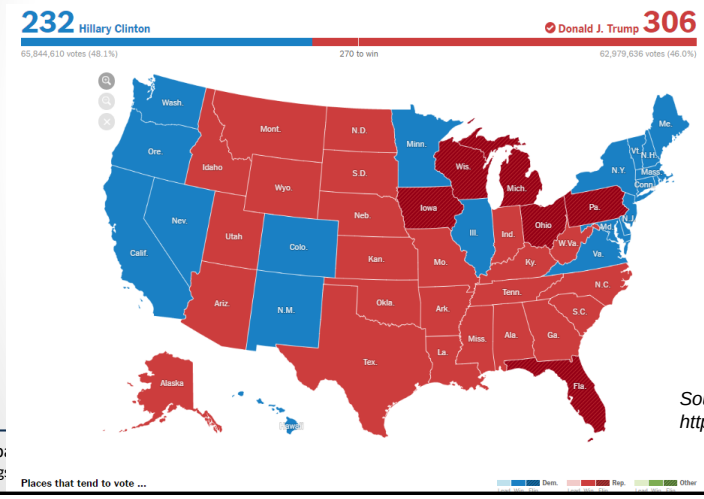
[1] http://en.wikipedia.org/wiki/Statistics
[2] http://www.merriam-webster.com/dictionary/statistics
[3] http://www.google.com

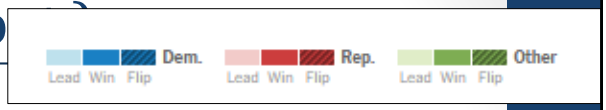
Questions about Data

1. Descriptive
   - Explain what actually happens
2. Relational
   - Focus on relationship between two or more variables
3. Causal
   - Figure out cause/effect of one variable on the others

Questions about Data (co

1. Descriptive
   - Explain what actually happens





Much more Republican

STATE	DEM.	REP.	OTH.	% RPT.
Ala.	34%	62%	2%	100%
Alaska	37%	51%	6%	99%
Ark.	34%	61%	3%	100%
Idaho	28%	59%	7%	100%
Kan.	36%	57%	5%	99%

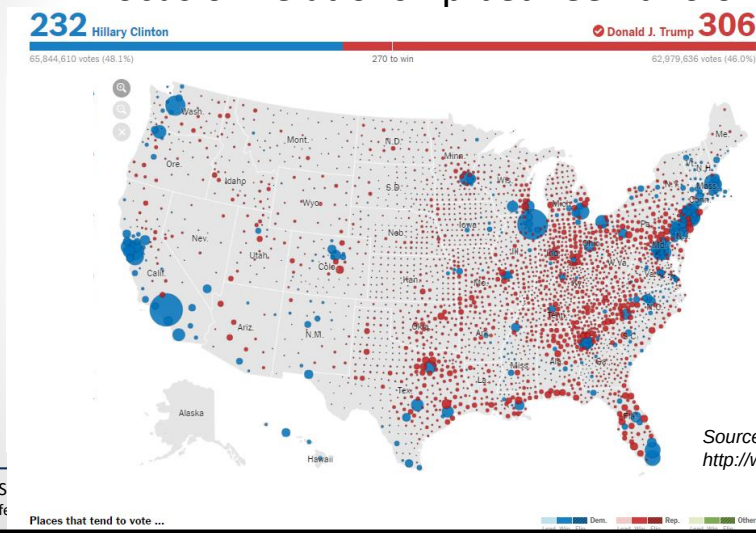
Much more Democratic

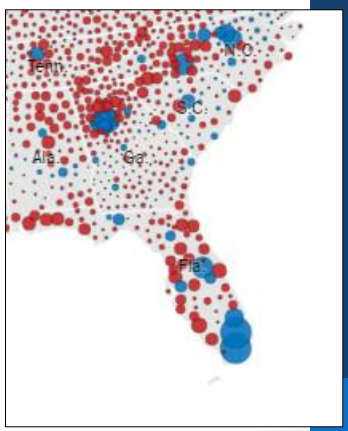
STATE	DEM.	REP.	OTH.	% RPT.
Calif.	62%	32%	3%	100%
D.C.	91%	4%	2%	100%
Hawaii	62%	30%	4%	100%
Md.	60%	34%	3%	99%
Mass.	60%	33%	4%	100%

Source: http://www.nytimes.com/elections/results/president

Questions about Data (cont.)

2. Relational
   - Focus on relationship between two or more variables



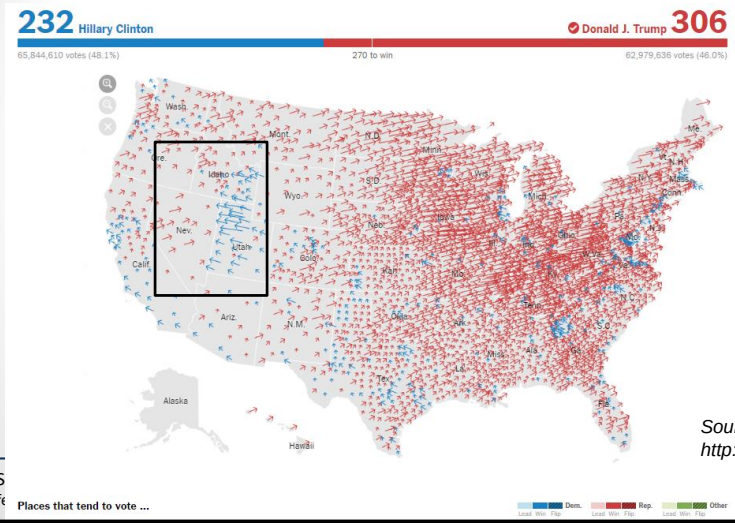


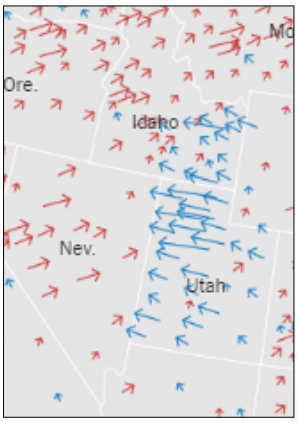
Source: http://www.nytimes.com/elections/results/president

Questions about Data (cont.)

2. Relational

- Focus on relationship between two or more variables



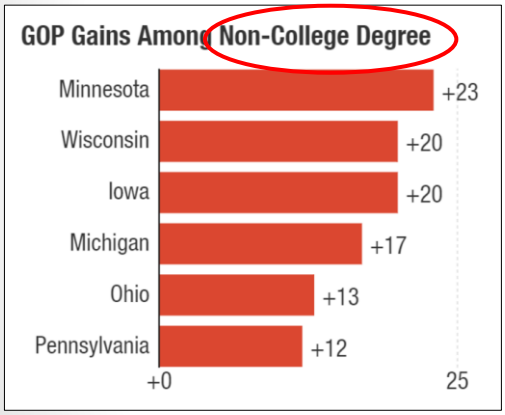


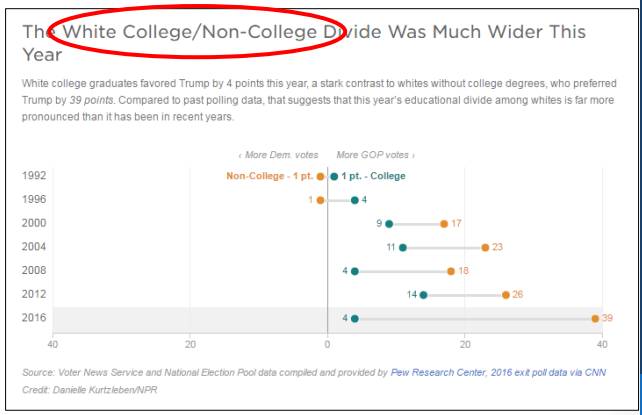
Source: http://www.nytimes.com/elections/results/president

Questions about Data

3. Causal

- Figure out cause/effect of one variable on the others





Source: http://www.npr.org/2016/11/12/501848636/7-reasons-donald-trump-won-the-presidential-election

Questions about Data (cont.)

1. Descriptive
   - Explain what actually happens
2. Relational
   - Focus on relationship between two or more variables
3. Causal
   - Figure out cause/effect of one variable on the others

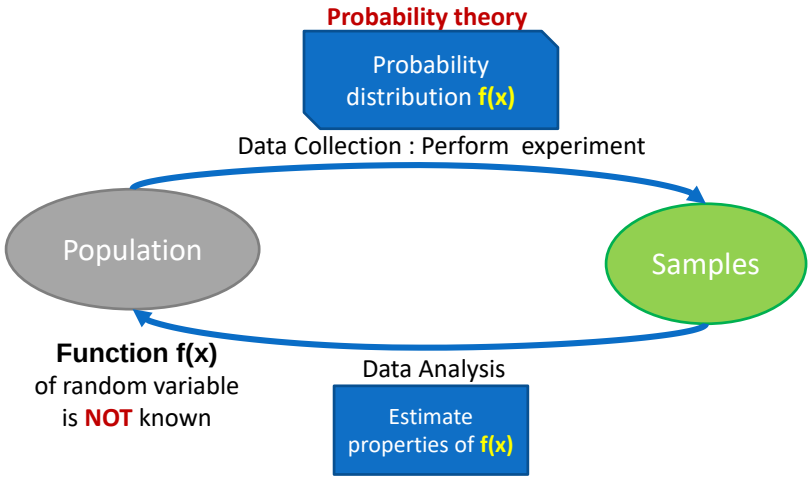
What we are familiar with: mean, variance

What we will learn in this course.

Terminologies

- Population
  - Collection of all data
- Sample
  - Subset of data
  - Also called observation or data

Inferential Statistics



Outline

- Introduction
- Review
  - Random variables
  - Discrete random variables
  - Continuous random variables

Random Variable (RV)

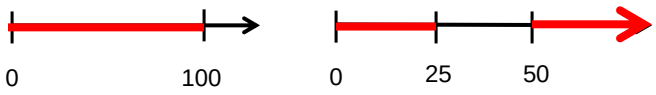
- Function that associates a numerical value with each outcome in experiment
- Random variable is represented by uppercase letter
  - Example: X, Y, Z
  - Note: lower case letter is used for outcome value.
- Example: coin toss
  - Outcome of coin-tossing experiment = {head, tail}

$$X = \begin{cases} 1 & \text{if outcome=head} \\ 0 & \text{if outcome=tail} \end{cases}$$

Bernoulli RV:
Outcome with two possible values

Types of Random Variables (RV)

- Discrete random variable
  - Outcomes come from finite set
  - Example:
    - Coin: {0, 1}
    - Dice: {1, 2, 3, 4, 5, 6}
- Continuous random variable
  - Outcomes come from interval(s) of numbers (infinite set)

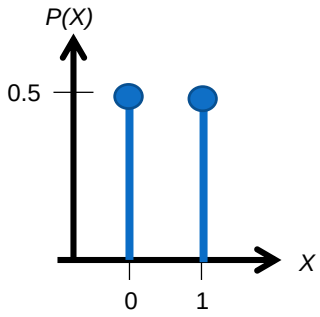


Probability Distribution

- Probability = likelihood that each outcome will occur
- Probability density function(pdf) for continuous RV
- Probability mass function (pmf) for discrete RV
  - Links each outcome with probability of its occurrence
- Example: coin toss
  - Unbiased coin

$$X = \begin{cases} 1 & \text{if outcome=head} \\ 0 & \text{if outcome=tail} \end{cases}$$

$P(X = 1) = 0.5$
$P(X = 0) = 0.5$



Probability Distribution (cont.)

- Example 2: Standard normal variable
  - Probability distribution = Gaussian distribution

$X \sim Gaussian\ RV$

$x = (-\infty, \infty)$

$$P(X = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

with mean = 0,
standard deviation = 1

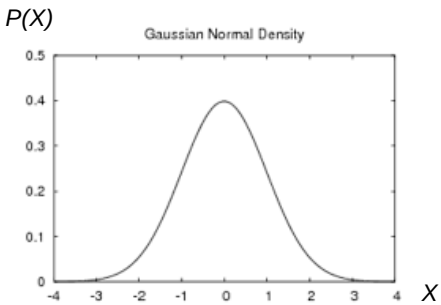


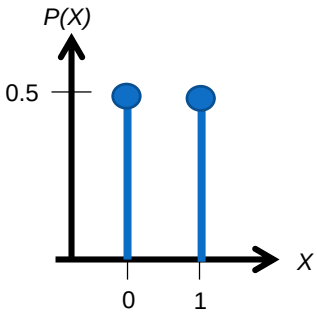
Image source: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRwf9Dtb8TAVdl4BdmOWIwuvvD3D3Dmk9gBcddQNHvZsAflJZqd

Basic Properties of PDF/PMF

- $0 \le P(x) \le 1$
- $\sum P(X) = 1$ for all possible values of X

Or 100%

Example: Coin toss



Example 2: Standard normal

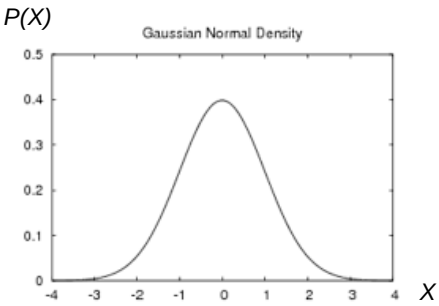


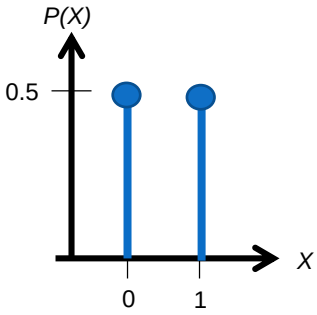
Image source: https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcQKI9cHZ-SxMZMlIZuZFqy8wRSq4j-mOX8rIU1bkmYW9hf2Pw5WeQ

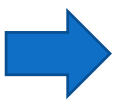
Cumulative Distribution Function (CDF)

- Sum of probabilities up to x

$$F(x) = P(X \le x) = \sum_{y:y \le x} P(y)$$

Example: Coin toss





Example: Coin toss

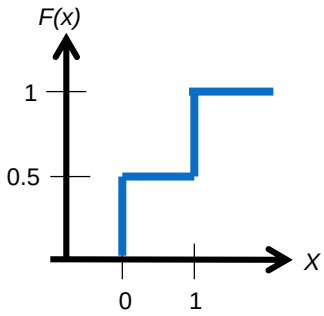


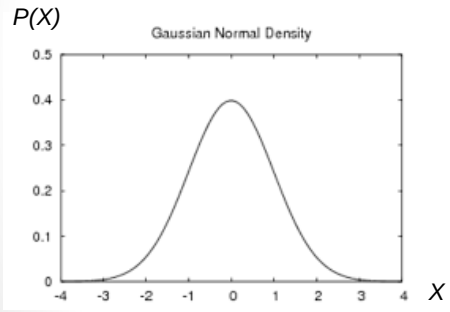
Image source: http://oneau.files.wordpress.com/2011/02/norm_cdf.png

Cumulative Distribution Function (CDF)

- Sum of probabilities up to x

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} P(y)$$

Example 2: Standard normal



Example 2: Standard normal

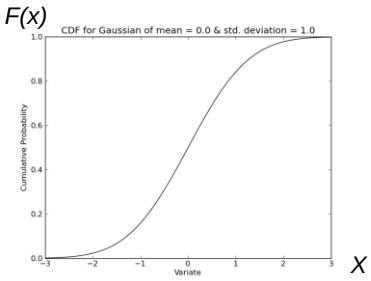


Image source: http://oneau.files.wordpress.com/2011/02/norm_cdf.png

Outline

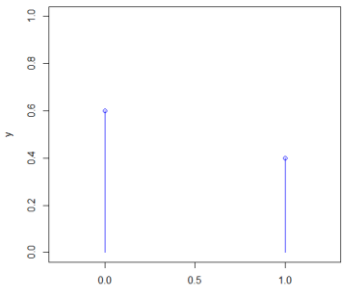
- Introduction
- Review
  - Random variable
  - Discrete random variables
  - Continuous random variables

Bernoulli Random Variable

- Random variable X with two outcomes
- Range of outcome value = {0,1}

$$P(X) = \begin{cases} p & ,x = 1 \\ 1-p & ,x = 0 \end{cases}$$

Mean	p
Variance	p(1-p)



p = 0.4

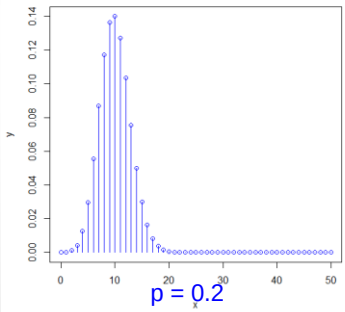
Example:

- Head vs. Tail

Binomial Random Variable

- Random experiments that independently repeat n times
- Let X = number of times that a certain event appears
- Range of outcome value = {0,1,2, …, n-1, n}

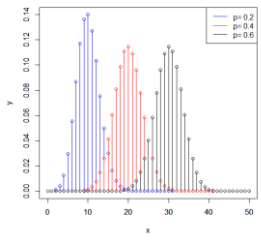
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k \in \{0,1,\dots,n-1,n\}$$



p = 0.2

Example:

- Toss coins n times and observe k heads

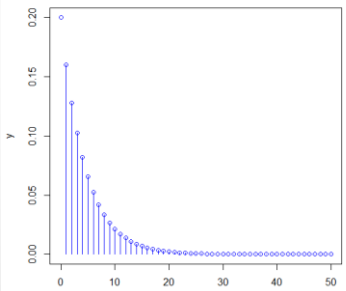


Mean	np
Variance	np(1-p)

Geometric Random Variable

- Random experiments that independently repeat n times
- Let X = number of failures until a success occurs
- Range of outcome value = {0,1,2, …}

$$P(X = k) = (1-p)^k p \ , \quad p = prob(success), k \ \in \{0,1,2, \dots\}$$



p = 0.2

Example:

- Observe k system-failures, until a system success occurs

Mean	(1-p)/p
Variance	$(1-p)/p^2$

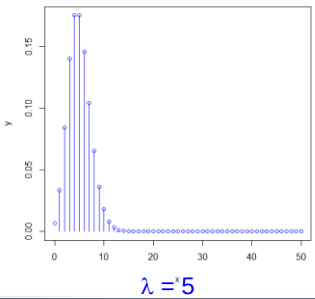
Poisson Random Variable

- Let X = number of events in a certain period
  - Event occurs randomly both in time and space
- Range of outcome value = {0,1,2, …}

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k \ \in \{0,1,2 \dots\}$$

Mean	λ
Variance	λ

λ = average number of events



λ = 5

Example:

- Particle emission of radioactive substances
- Test a system and find out possibility of k successes (or failures) to occur

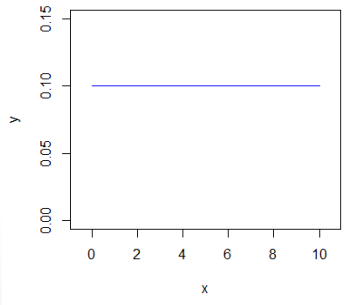
Outline

- Introduction
- Review
  - Random variable
  - Discrete random variables
  - Continuous random variables

Uniform Random Variable

- Random variable that has equal probability
- Range of outcome value = [a,b]

$$P(X = x) = \frac{1}{b - a}$$



Example:

- Real value that is likely to occur between [a,b]

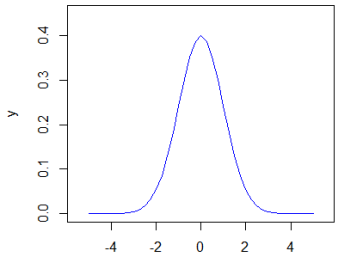
Mean	(a+b)/2
Variance	$(b-a)^2/12$

Normal Random Variable

- Random variable that has distribution Gaussian (normal) distribution
- Common and well-known distribution

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean	μ
Variance	σ^2



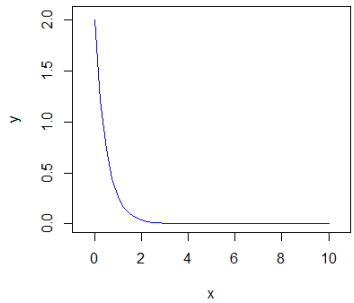
$\mu = 0, \sigma^2 = 1$

Exponential Random Variable

- Random variable that models time between occurrences of events

$$P(X = x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

λ = rate that event occurs



Mean	$1/\lambda$
Variance	$1/\lambda^2$

References

1. J.L. Devore and K.N.Berk, Modern Mathematical Statistics with Applications, Springer, 2012.
2. A.L. Leon-Garcia, Probability and Random Processes for Electrical Engineering, Addison-Wesley, 1994.