ECSE 343 Numerical Methods in Engineering

Roni Khazaka
Dept. of Electrical and Computer Engineering
McGill University

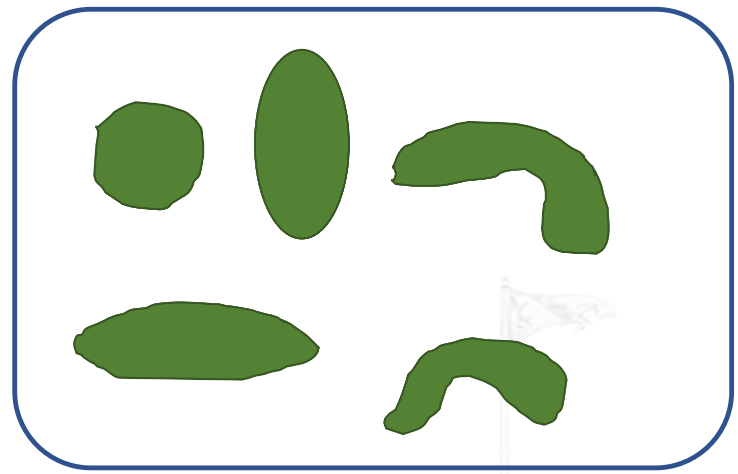
Axioms of Probability

- Consider a sample space S
- Associate each Event $A \in S$ with a number $P(A)$.
- $P(A)$ is a probability iff:

○ Axiom 1: $P(A) \geq 0$

○ Axiom 2: $P(S) = 1$

○ Axiom 3: If $\{A_1, A_2, \cdots\}$ is a sequence of mutually exclusive events (i.e. $A_i \cap A_j = \phi$, for $i \neq j$) then: $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$



S

Theorem 1: $P(\phi) = 0$

Proof: Consider the sequence $\{A_1, A_2, \cdots\}$ such that:

- $A_1 = S$ and,
- $A_i = \phi$ for all $i \geq 2$

Then

- $A_i \cap A_j = \phi$, for $i \neq j$ ➔ This is a sequence of mutually exclusive events.
- Axiom 3: $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
- $P(A_1) = P(A_1) + \sum_{i=2}^{\infty} P(A_i) = P(A_1) + \sum_{i=2}^{\infty} P(\phi)$
- $P(\phi) = 0$

Theorem 2

Consider the sequence $\{A_1, A_2, \cdots, A_n\}$ of mutually exclusive events. Then:

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

Proof: Consider the sequence $\{A_1, A_2, \cdots, A_n, \cdots\}$ such that:

$A_i = \phi$ for all $i > n$, then

- $A_i \cap A_j = \phi$, for $i \neq j$ ➔ This is a sequence of mutually exclusive events.
- $P(\bigcup_{i=1}^{n} A_i) = P(\bigcup_{i=1}^{\infty} A_i)$
- Axiom 3: $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(A_i) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\phi) = \sum_{i=1}^{n} P(A_i)$
- $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

Theorem $P(A^c) = 1 - P(A)$

Note: A^c is the complement of A

Proof:

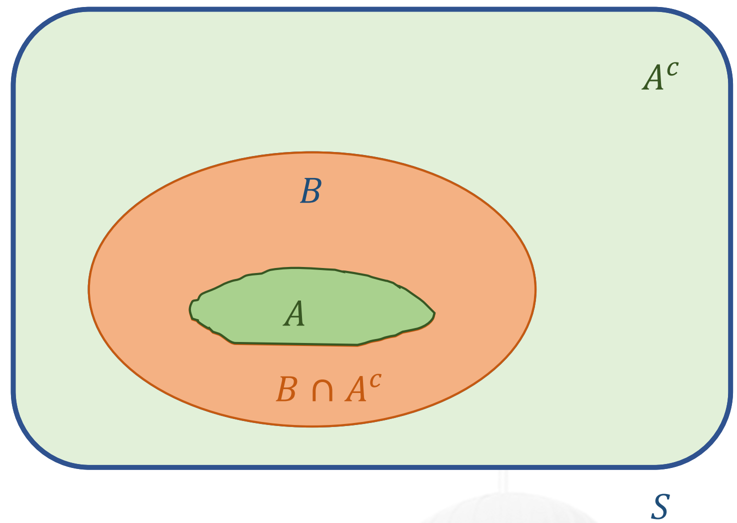
$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

Theorem If $A \subseteq B$ then $P(A) \leq P(B)$

Proof:

$$B = A \cup (B \cap A^c)$$

$$P(B) = P(A) + P(B \cap A^c) \geq P(A)$$



Theorem

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

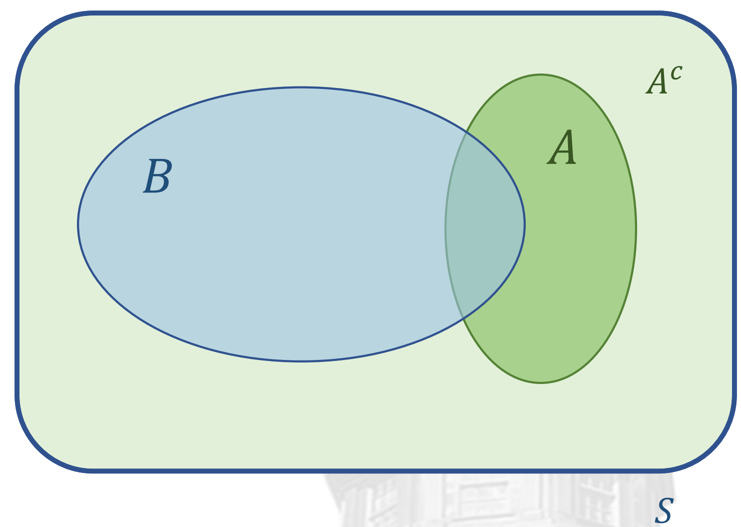
(inclusion exclusion)

Proof: $A \cup B = A \cup (B \cap A^c)$

$$P(A \cup B) = P(A) + P(B \cap A^c)$$

$$P(B \cap A) + P(B \cap A^c) = P(B)$$

$$P(A \cup B) = P(A) + P(B) - P(B \cap A)$$



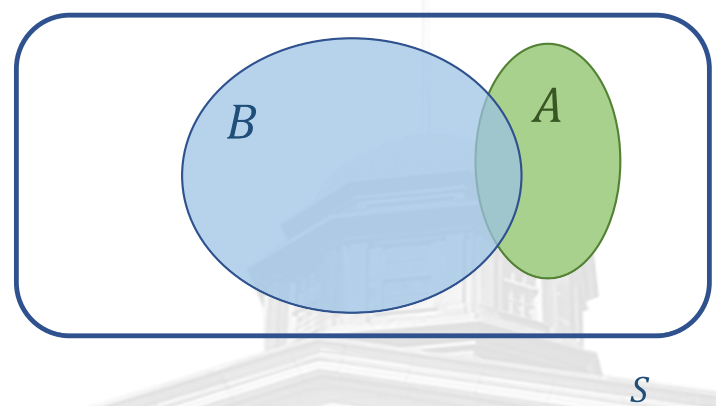
Can extend this to many events.

Independence

Definition: Events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Note: Completely different from A and B being disjoint



Conditional Probability

Update probabilities based on evidence:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) > 0$$

A and B independent

Update probabilities based on evidence:

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

Knowing that B happened, does not give us in information about A

Conditional Probability

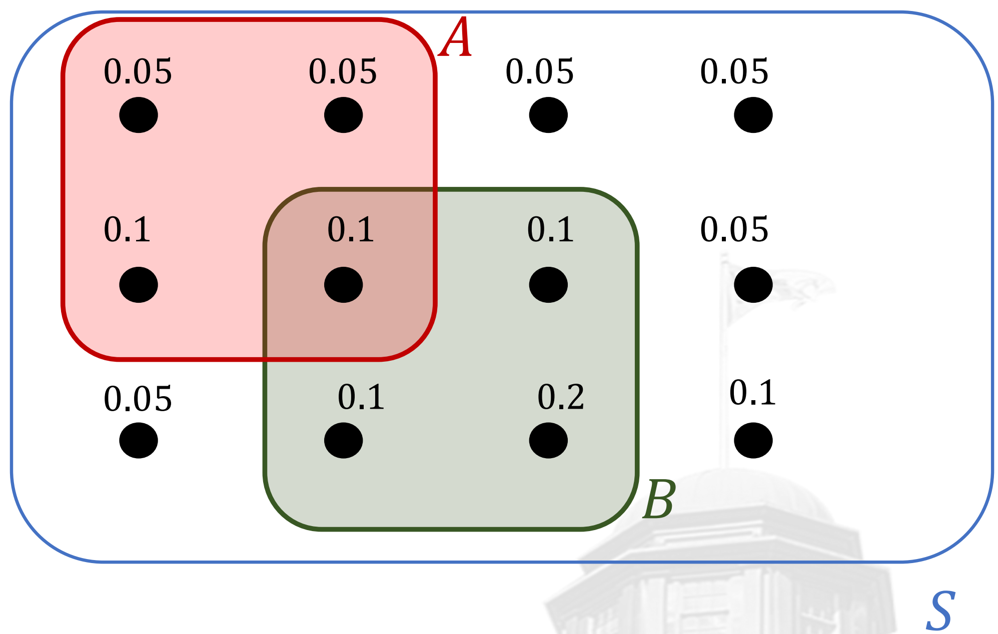
$P(S) = 1$

$P(A) = 0.3$

$P(B) = 0.5$

$P(A \cap B) = 0.1$

$$P(A \cap B) \neq P(A)P(B)$$



Conditional Probability

$P(A|B) = ?$

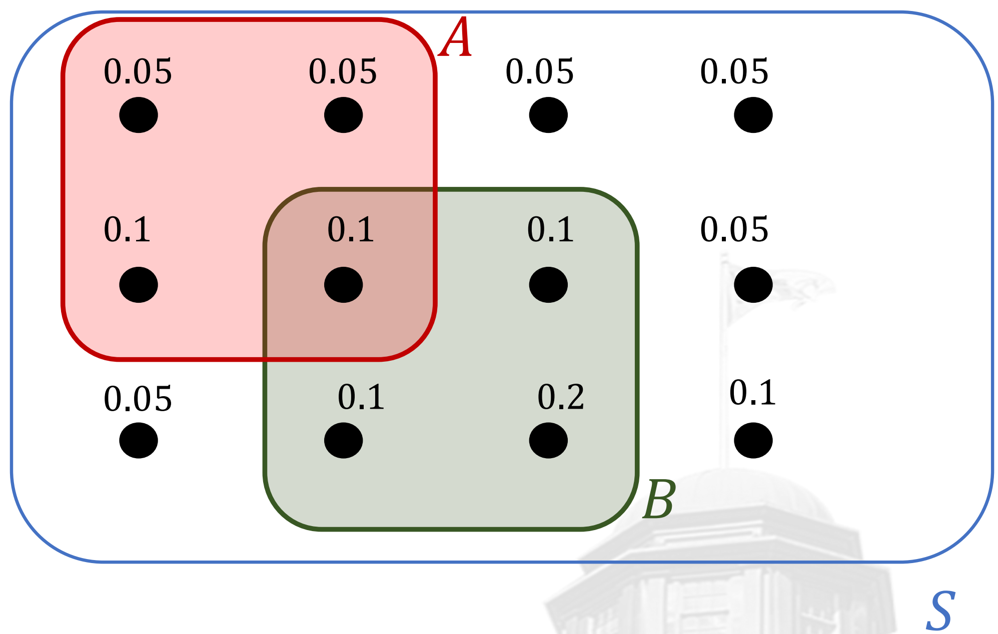
We know that B occured

At this point A can only occur if $A \cap B$ occur.

B is now our sample space

Normalize the probabilities so that $P(B) = 1$ (new sample space)

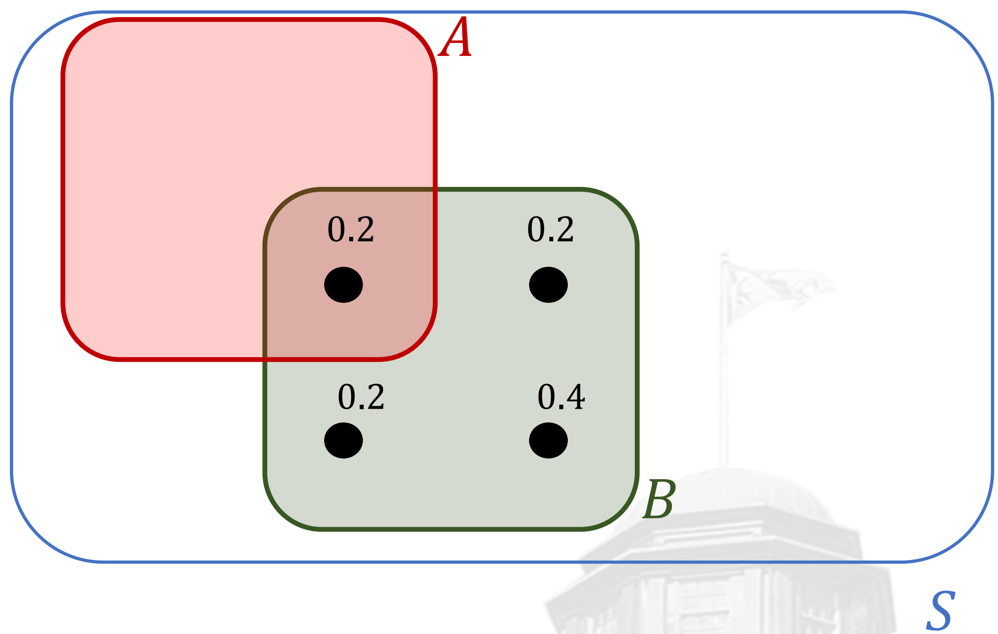
i.e. divide by $P(B) = 0.5$



Conditional Probability

Divide by $P(B)$

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



Conditional Probability

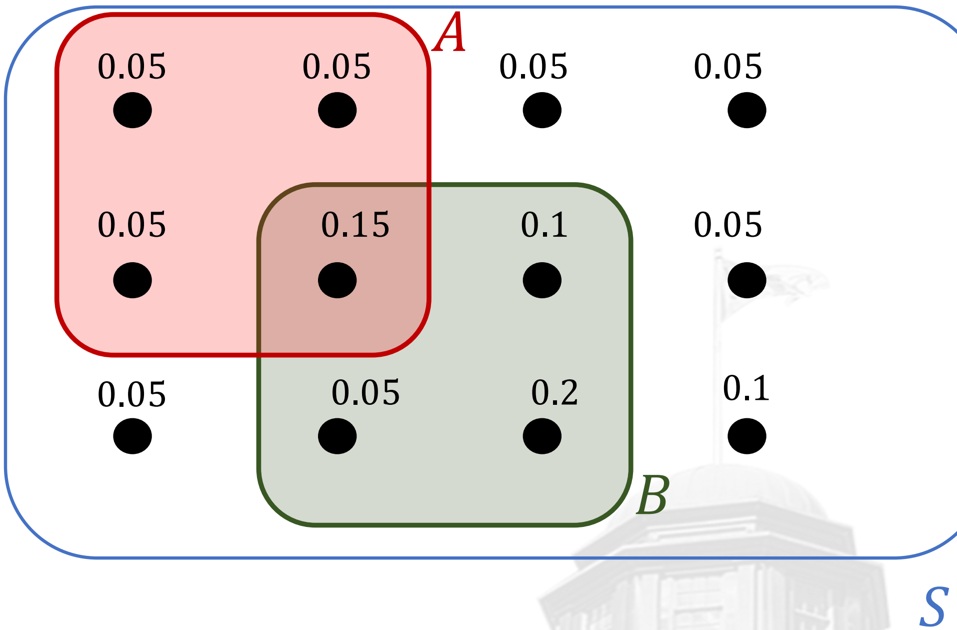
$P(S) = 1$

$P(A) = 0.3$

$P(B) = 0.5$

$P(A \cap B) = 0.15$

$$P(A \cap B) = P(A)P(B)$$



Useful Theorems

Theorem 1: $P(A \cap B) = P(A|B)P(B)$

Follows directly from conditional probability

Theorem 2: $P(A|B)P(B) = P(B|A)P(A)$

$$\left.\begin{array}{l} P(A \cap B) = P(A|B)P(B) \\ P(B \cap A) = P(B|A)P(A) \\ P(A \cap B) = P(B \cap A) \end{array}\right\} P(A|B)P(B) = P(B|A)P(A)$$

Also known as Bayes' rule: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

Law of Total Probability

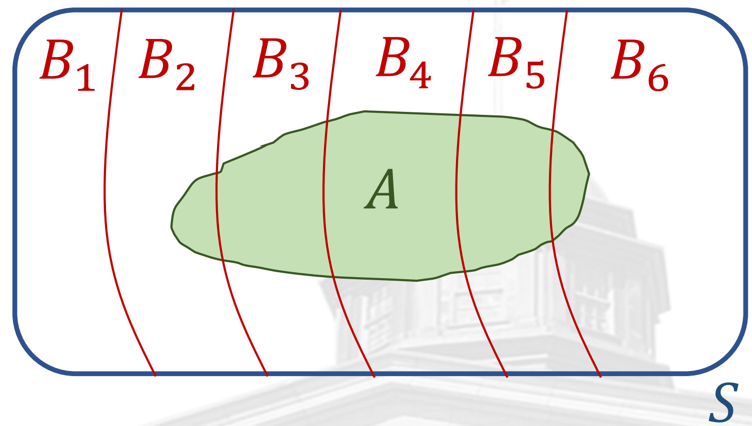
If A and B are disjoint (i.e. $A \cap B = \phi$) then: $P(A \cup B) = P(A) + P(B)$

Strategy for computing $\mathrm{P}(A)$: Divide S into disjoint events B_1, B_2,…, B_n such that:

$$B_1 \cup B_2 \cup \cdots \cup B_n = S$$

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_6)$$

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$



Random Variables

Definition: A random variable is a function from the sample space S to $\mathbb{R}$

Discrete Random Variables

A discrete random variable can take a discrete set of values:

$$x_o, x_1, x_2, \cdots$$

(could be finite, could be infinite)

Bernoulli RV: Bern(p)

A random variable X has a Bernoulli distribution with parameter p if X has only 2 possible values, 0 and 1, and

$$\underbrace{P(X = 1)}_{\substack{\text{Event } S \text{ such that} \\ X(S) = 1}} = p \qquad \underbrace{P(X = 0)}_{\substack{\text{Event } S \text{ such that} \\ X(S) = 0}} = 1 - p = q$$

Binomial RV: Bin(n,p)

A random variable X has a Binomial distribution with parameters n, and p if X can be represented as the number of successes in n independent $Bern(p)$ trials.

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

Probability Mass Function (PMF)

Probability Mass Function

The probability mass function (PMF) completely define a distribution.

$$P(X = k) \quad \text{For all } k$$

$$\sum_i P(X = x_i) = 1$$

PMFs are only defined for discrete random variables (not continuous)

Cumulative Distribution Function (CDF)

$X \leq x$ is an event.

$F(x) = P(X \leq x)$ is the CDF of X

The Cumulative Distribution Function (CDF) completely define a distribution. It has the following properties:

- Increasing.
- Right continuous.
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$

Independence of Random Variables

X, and Y are independent RVs if:

$$\underbrace{P(X \leq x, Y \leq y)}_{\text{Joint PMF} \quad F(x,y)} = P(X \leq x)\mathrm{P}(Y \leq y) \qquad \forall x, y$$

Expected Value

$$E(X) = \sum_{X} xP(X = x)$$

Example: $X \sim Bern(p)$

$$E(X) = \sum_{X} xP(X = x) = 1P(X = 1) + 0P(X = 0) = p$$

Expected Value

$$E(X) = \sum_{X} xP(X = x)$$

Example: $X \sim Bin(np)$

$$E(X) = np$$

Linearity

$$E(X+Y) = E(X) + E(Y)$$

$$E(cX) = cE(X) \qquad c \text{ is a constant}$$

Continuous Random Variables

Probability Mass Function (PMF) for a discrete Random Variable:

$$P(X = x)$$

For a continuous RV:

$$P(X = x) = 0$$

Need another equivalent concept.

Probability Density Function (PDF)

A Random Variable X has a PDF $f_X(x)$ if for all a and b:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

A valid PDF satisfies the following conditions:

$$f_X(x) \geq 0 \qquad\qquad \int_{-\infty}^{\infty} f_X(x)dx = 1$$

Cumulative Distribution Function (CDF)

If a Random Variable X has a PDF $f_X(x)$, its CDF is:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

If a Random Variable X has a CDF $F_X(x)$, its PDF is:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Cumulative Distribution Function (CDF)

If a Random Variable X has a CDF $F_X(x)$, its PDF is:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$$

Expected Value $\mathrm{E}(X)$

Discrete Random Variable X: $\mathrm{E}(X) = \sum_{X} xP(X = x)$

Continuous Random Variable X: $\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$

Expected Value

Valid for any distribution:

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Variance

For both Continuous or Discrete Random Variable X:

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X)^2]$$

Standard Deviation:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Variance

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2]$$
$$= \mathrm{E}[X^2 - 2X\mathrm{E}(X) + \mathrm{E}(X)^2]$$
$$= \mathrm{E}[X^2] - 2\mathrm{E}(X)\mathrm{E}(X) + \mathrm{E}(X)^2$$
$$= \mathrm{E}[X^2] - \mathrm{E}(X)^2$$

$$\mathrm{Var}(X + c) = \mathrm{Var}(X)$$
$$\mathrm{Var}(cX) = \mathrm{c}^2\mathrm{Var}(X)$$

Uniform Distribution

$X \sim \text{Unif}(a, b)$

- Random point in interval $[a, b]$
- All equal sized subintervals are equally likely
- Probability is proportional to "length" of interval

Uniform Distribution

$X \sim \mathrm{Unif}(a, b)$

PDF: $f_X(x) = \begin{cases} c & \text{For } a \le x \le b \\ 0 & \text{Otherwise} \end{cases}$

Note: $1 = \int_{-\infty}^{\infty} f_X(x)dx = \int_a^b c\,dx = c(b-a)$

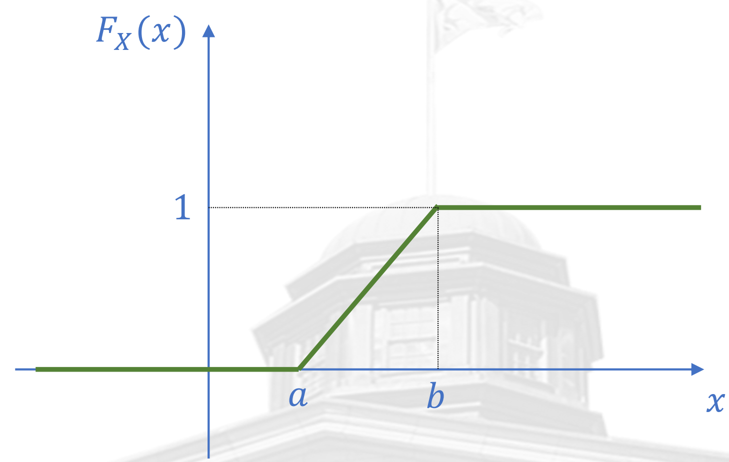
$\Rightarrow \quad c = \dfrac{1}{b-a}$

Uniform Distribution

$X \sim \text{Unif}(a, b)$

CDF: $F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{a}^{x} f_X(t)dt$

$$F_X(x) = \begin{cases} 0 & \text{For } x < a \\ \dfrac{x-a}{b-a} & \text{For } a \le x \le b \\ 1 & \text{For } x > b \end{cases}$$



Expected Value

$$f_X(x) = \begin{cases} c & \text{For } a \leq x \leq b \\ 0 & \text{Otherwise} \end{cases} \qquad c = \frac{1}{b-a}$$

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_a^b x f_X(x)dx = \int_a^b \frac{x}{b-a}dx$$

$$= \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

Variance of Unif(a,b)

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}(X)^2$$

$$\mathrm{E}(X)^2 = \frac{(b+a)^2}{4}$$

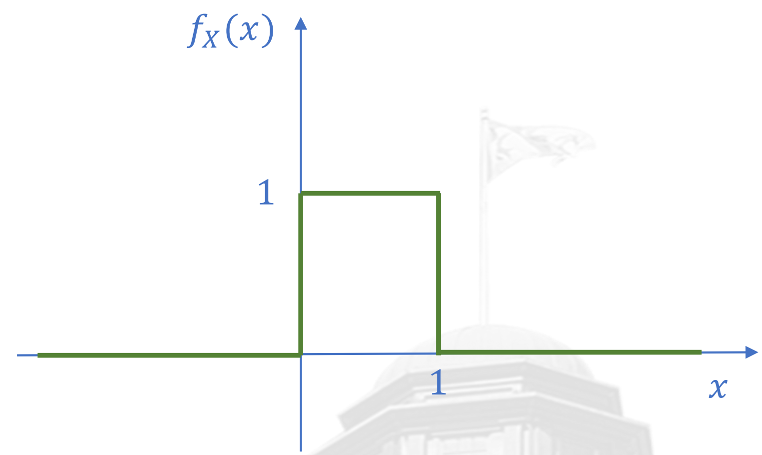
$$\mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx \quad = \int_a^b \frac{x^2}{b-a} dx \quad = \frac{1}{b-a}\left[\frac{1}{3}x^3\right]_a^b$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Unif(0,1)

$X \sim \text{Unif}(0,1)$

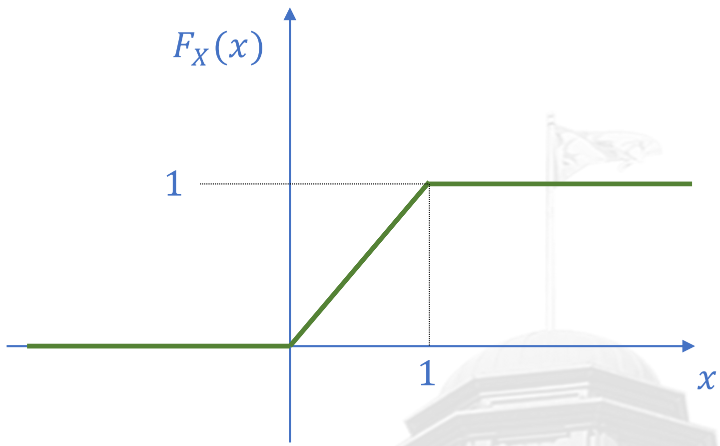
$$f_X(x) = \begin{cases} 1 & \text{For } 0 \leq x \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$



Unif(0,1)

$X \sim \text{Unif}(0,1)$

$$F_X(x) = \begin{cases} 0 & \text{For } x < 0 \\ x & \text{For } 0 \le x \le 1 \\ 1 & \text{For } x > 1 \end{cases}$$



Unif(0,1)

$X \sim \text{Unif}(0,1)$

$$\text{E}(X) = \frac{1}{2}$$

$$\text{Var}(X) = \text{E}[X^2] - \text{E}(X)^2$$

$$\text{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx \quad = \int_0^1 x^2\, dx = \frac{1}{3}$$

$$\text{Var}(X) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Universality of the Uniform Distribution

Let $U \sim \text{Unif}(0,1)$

Let $F(x)$ be a CDF. (Assume $F(x)$ is strictly increasing)

Let $X = F^{-1}(U)$

Then $X \sim F(x)$

Proof: The CDF of X is.

$$P(X \leq x) = P(F^{-1}(U) \leq x)$$
$$= P\big(U \leq F(x)\big) = F(x)$$

Note: $x = F^{-1}(u)$

$\Updownarrow$

$u = F\,(x)$

Independence of Random Variables

Definition: The random Variables $X_1, X_2, \cdots, X_n$ are independent iff:

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$$

For Discrete RVs

$$P(X_1 = x_1, \cdots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

Maximum Likelihood Estimation

<u>Regression:</u> Choose parameters to minimize the least square error between model and data.

<u>MLE:</u> Choose parameters to maximize the likelihood of observing the data. Assumes a model for the noise on observed data.

Experiment

We have a coin, and we would like to determine the distribution of Heads vs Tails if we flip the coin. We assume a Bern(p) distribution. The goal is to estimate the parameter p.

We run an experiment: Flip a coin 100 times.
We observe: 40 Heads and 60 Tails.

Choose p that maximizes the likelihood of this observation.

Experiment

$X \sim \text{Bern}(p)$

$$P(X = H) = p$$
$$P(X = T) = 1 - p = q$$

Assume p is given

$$P(40H, 60T|p) = \underbrace{\binom{100}{40} p^{40}(1-p)^{60}}$$

Choose p to maximize

➔ Hard Problem

Experiment

Assume p is given $\quad P(40H, 60T|p) = \underbrace{\binom{100}{40} p^{40}(1-p)^{60}}_{\text{Choose } p \text{ to maximize}}$

➔ Logarithm is monotonously increasing.

➔ Maximize $\log(P(40H, 60T|p))$ instead (Equivalent problem)

$$\log P(40H, 60T|p) = \log\binom{100}{40} + 40\log(p) + 60\log(1-p)$$

Experiment

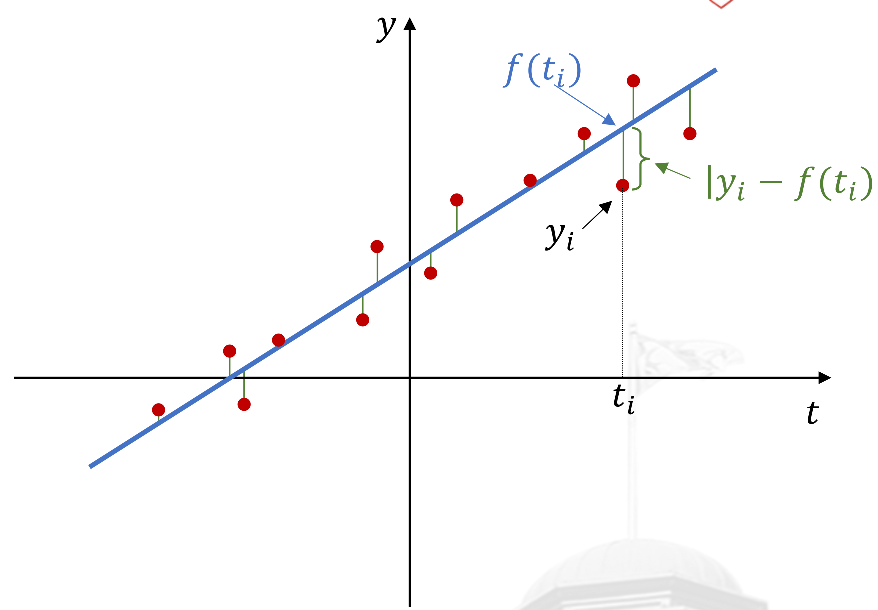
$$\log P(40H, 60T|p) = \underbrace{\log\binom{100}{40} + 40\log(p) + 60\log(1-p)}_{\text{Choose } p \text{ to maximize}}$$

$$\frac{d}{dp}\log P(40H, 60T|p) = \frac{40}{p} - \frac{60}{1-p} = 0$$

$$40(1-p) = 60p \qquad\qquad p = \frac{40}{100} = 0.4$$

Least Squares Approximation

Consider n data points (t_i, y_i)



Approximate data with a model:

$$y = f(t) = a_o + a_1 t$$

a_o and a_1 are the model parameters.

Choose the parameters to minimize:

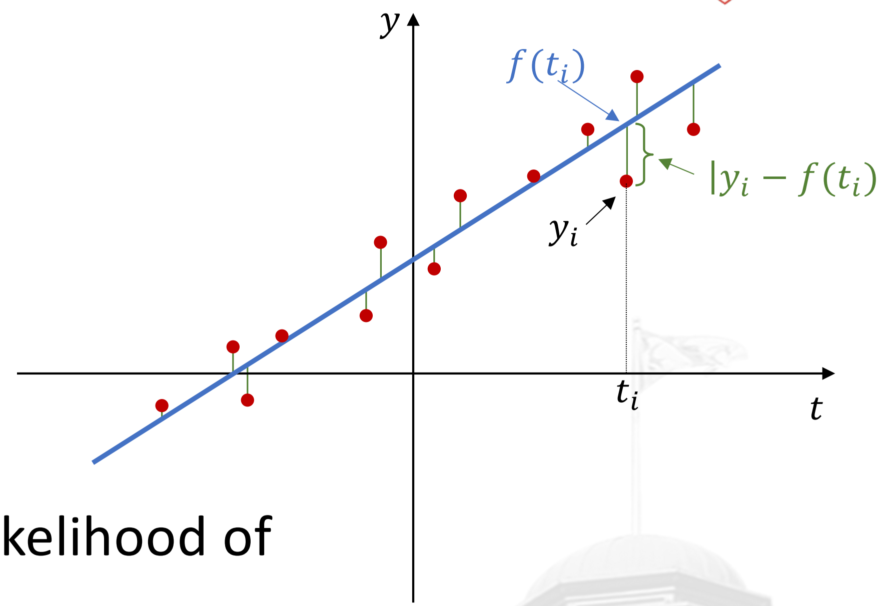
$$e = \sum_{i=1}^{n} (f(t_i) - y_i)^2$$

Maximum Likelihood Estimation MLE

Consider n data points (t_i, y_i)

Approximate data with a model:

$$y = f(t) = a_o + a_1 t + \underbrace{n_i}_{\text{noise}}$$



a_o and a_1 are the model parameters.

Choose the parameters to maximize the likelihood of observing the data (t_i, y_i)

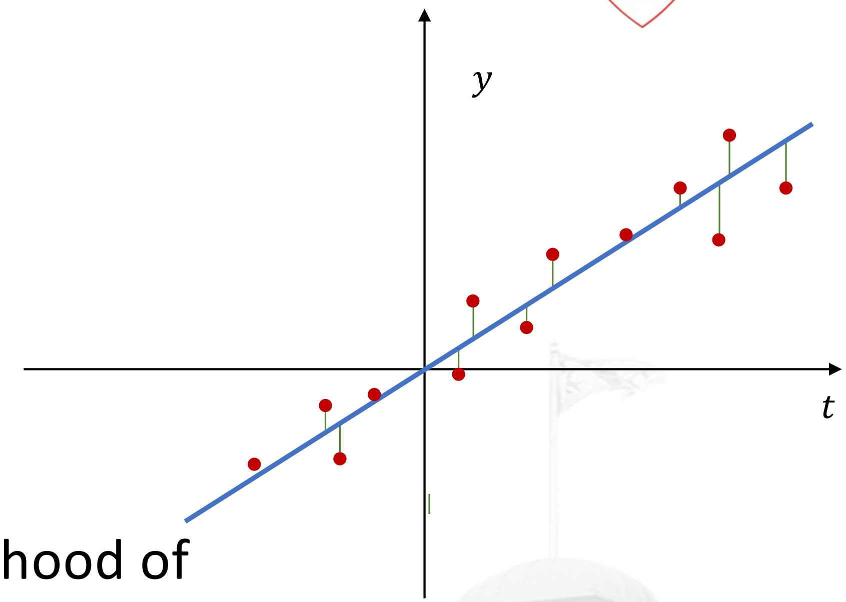
Maximum Likelihood Estimation MLE

Consider n data points (t_i, y_i)

Approximate data with a model:

$$y = f(t) = at + \underbrace{n_i}_{\text{noise}}$$

a is the model <u>parameter.</u>



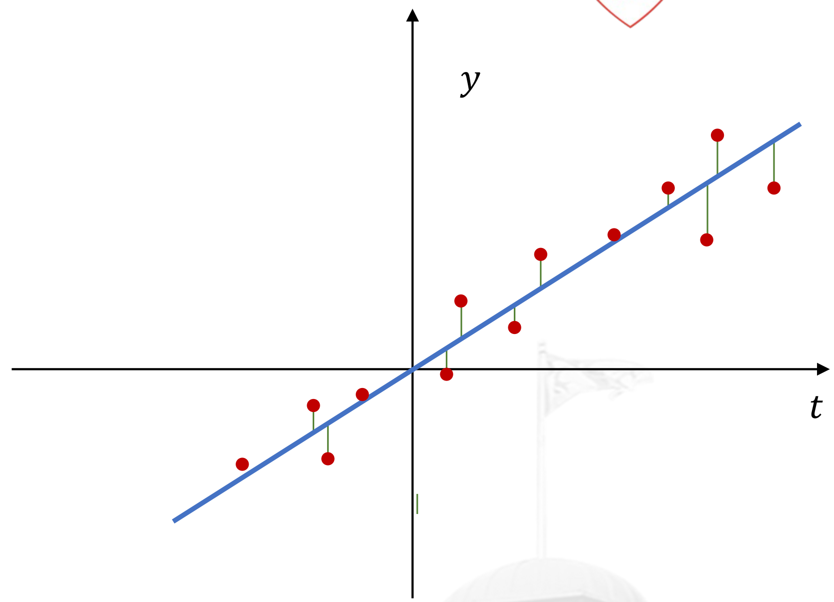
Choose the parameters to maximize the likelihood of observing the data (t_i, y_i)

Maximum Likelihood Estimation MLE

Consider n data points (t_i, y_i)

Approximate data with a model:

$$y = f(t) = at + \underbrace{n_i}_{\text{noise}}$$



$y_i = at_i + n_i$

$Y = [y_1, y_2, \cdots, y_n]$

parameter

$P(y_1, y_2, \cdots, y_n|a) = P(Y|a)$ Probability of data given parameter

data

➔ Choose p to maximize

Maximum Likelihood Estimation MLE

Assume independence:

$$P(Y|a) = P(y_1, y_2, \cdots, y_n|a) = P(y_1|a)P(y_2|a)\cdots P(y_n|a) = \underbrace{\prod_{i=1}^{n} P(y_i|a)}_{\text{Choose } a \text{ to maximize}}$$

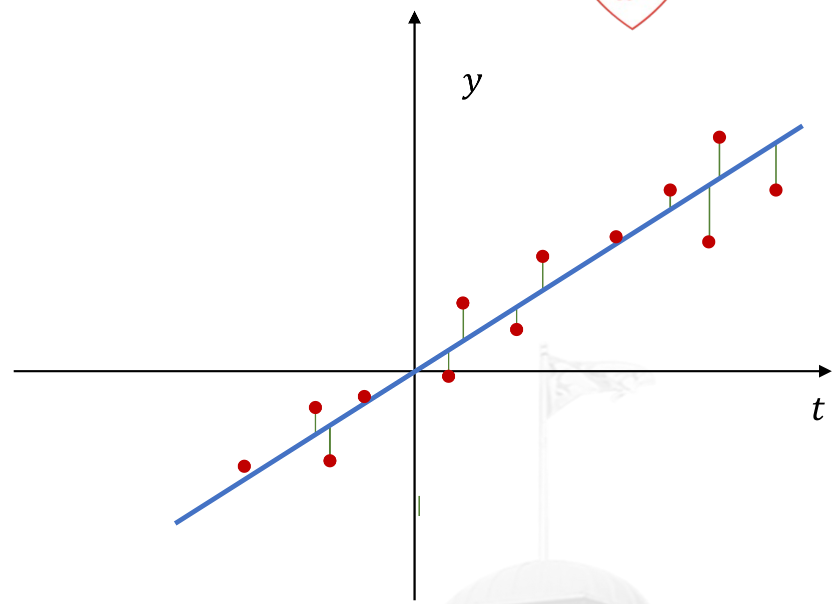
$$\log P(Y|a) = \sum_{i=1}^{n} \log P(y_i|a)$$

Maximum Likelihood Estimation MLE

Consider n data points (t_i, y_i)

Approximate data with a model:

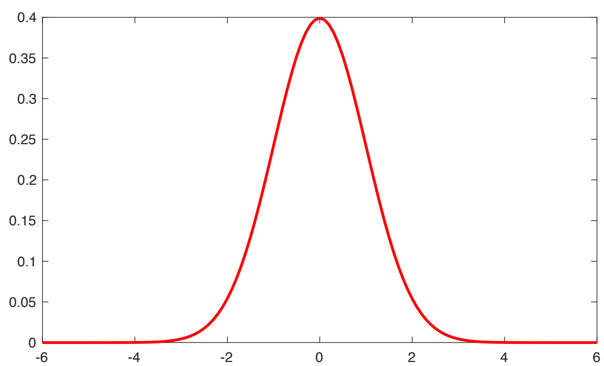
$$y = f(t) = at + \underbrace{n_i}_{\text{i.i.d. zero mean Gaussian}}$$



Maximum Likelihood Estimation MLE

n_i is i.i.d. zero mean Gaussian

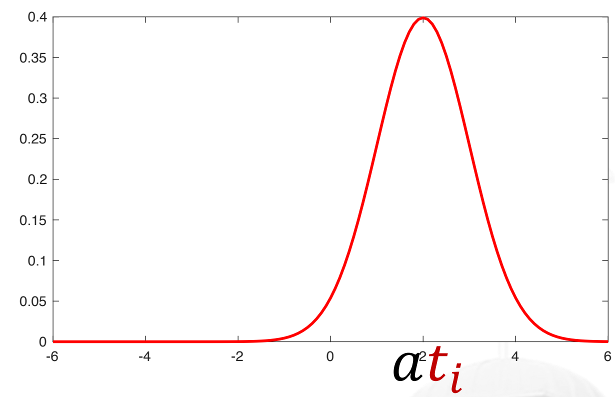
Distribution of n_i



$$f(n_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{n_i^2}{2\sigma^2}}$$

Distribution of

$y_i = at_i + n_i$



$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - at_i)^2}{2\sigma^2}}$$

Maximum Likelihood Estimation MLE

Assume independence:

$$P(Y|a) = P(y_1, y_2, \cdots, y_n|a) = P(y_1|a)P(y_2|a)\cdots P(y_n|a) = \underbrace{\prod_{i=1}^{n} P(y_i|a)}_{\text{Choose } a \text{ to maximize}}$$

$$\log P(Y|a) = \sum_{i=1}^{n} \log P(y_i|a)$$

Choose a to maximize $\displaystyle\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - a t_i)^2}{2\sigma^2}}\right)$

Maximum Likelihood Estimation MLE

Choose a to maximize

$$\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - a t_i)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - a t_i)^2}{2\sigma^2}\right]$$

$$\frac{d}{da}\sum_{i=1}^{n}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - a t_i)^2}{2\sigma^2}\right] = \sum_{i=1}^{n}\frac{2(y_i - a t_i)t_i}{2\sigma^2} = 0$$

Maximum Likelihood Estimation MLE

Choose a such that $$\sum_{i=1}^{n} \frac{2(y_i - at_i)t_i}{2\sigma^2} = 0$$

$$\sum_{i=1}^{n} y_i t_i = \sum_{i=1}^{n} a t_i^2 = a \sum_{i=1}^{n} t_i^2$$

$$a = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i^2}$$

Equivalent to Least Squares (if noise is assumed to be zero mean Gaussian)

Least Squares / Regression

$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} a = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Normal Equations: $$[t_1 \quad t_2 \quad \cdots \quad t_n] \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} a = [t_1 \quad t_2 \quad \cdots \quad t_n] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$a = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i^2}$$