香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

DDA2001: Introduction to Data Science

Lecture 9: Probability Review

Zicheng Wang

Announcement

- First assignment due on March. 10

- Midterm (not confirmed yet): **March. 16 (Saturday) 2-4 PM.**
- Zero tolerance policy when it comes to cheating so, DON’T CHEAT
- Missing the midterm or final exam without prior notification to and approval of the instructors will automatically result in the "0" grade for the exam.

- Guest Lecture (not confirmed yet): **March. 1 (Friday) 5:30-7 PM.**
- No class on this Wednesday

What is Probability?

- An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a **random experiment**.

- Probability is used to quantify the likelihood, or chance, that an outcome of a random experiment will occur.

What is Probability?

- Random Experiment:

- Consider one possible outcome： ω
- The outcome ω happens with probability P(ω)

- **It means:**
  - If we repeat such experiment **N** times
  - We observe **n** observations that the outcome is ω.
  - Then if N goes to infinity, $\boldsymbol{n/N}$ will approach P(ω).

Terminologies

- **Random Experiment**: a repeatable procedure
- **Sample space**: set of all possible outcomes Ω.
- **Event**: a subset of the sample space.
- **Probability function**, P(ω): gives the probability for each outcome $\omega \in \Omega$
  - Probability is between 0 and 1
  - Total probability of all possible outcomes is 1.
  - If A=$\{\omega_1, \omega_2, \omega_3, \ldots\}$, P(A)=$P(\omega_1)+P(\omega_2)+P(\omega_3)+\cdots$

Sample Space

- Discrete or continuous: countable (listable) or not?
- A sample space is discrete if it consists of a finite or countable infinite set of outcomes.
- A sample space is continuous if it contains an interval (or a union of multiple intervals) of real numbers.

Events

- Events are sets:
  - ✓ Can describe in words
  - ✓ Can describe in notation
- Experiment: toss a coin 2 times.
- Event -- You get 1 or more heads

  = {HH, HT, TH}

Set operations

- Events are sets, so we can use set operations

  ✓ Unions
  ✓ Intersections
  ✓ Complements

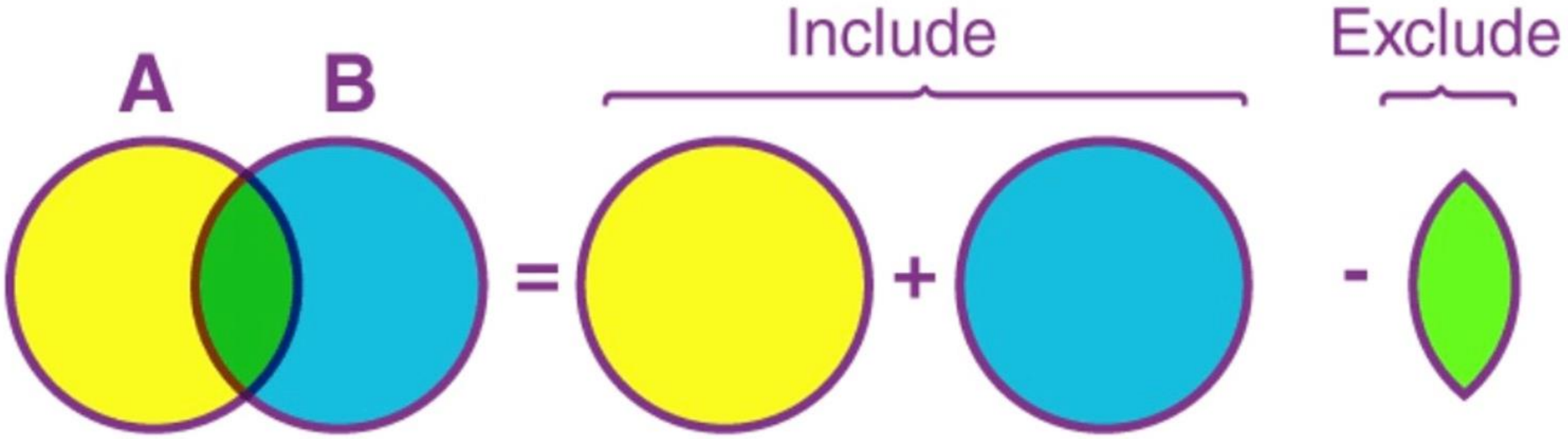
Set operations

- We denote the union as (A or B) in words, and $\mathrm{A} \cup \mathrm{B}$ in notation
- We denote the intersection as (A and B) in words, and $\mathrm{A} \cap \mathrm{B}$ in notation
- Two events A and B, such that $\mathrm{A} \cap \mathrm{B} = \emptyset$ are said to be **mutually exclusive**.
- We denote the complement as (not A) in words, and A' or A^c in notation

Set operations

- The commutative laws: $A \cup B = B \cup A$, $A \cap B = B \cap A$
- The associative laws: $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$
- The distributive laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- De Morgan's laws: $(A \cup B)' = A' \cap B'$, $(A \cap B)' = A' \cup B'$

Principle of Inclusion and Exclusion

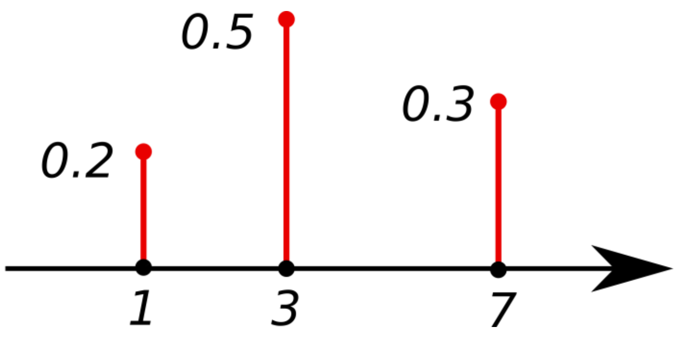


$$P(A \text{ or } B) = P(A) + P(B) - P(\ A \text{ and } B)$$

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

Probability function

- Discrete:
  - ✓ Probability mass function.
  - ✓ P(ω): gives the probability for **each** outcome ω ∈ S

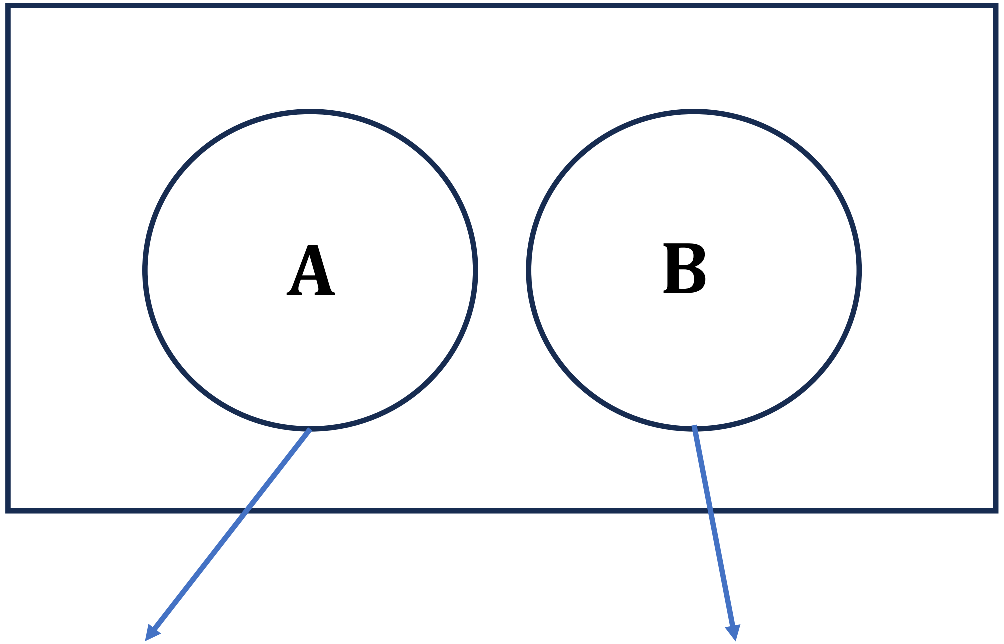


Continuous case will be defined later.

Relation between two events

- Independence: P(A and B) = P(A)P(B)
- Disjoint (mutually exclusive): P(A or B) = P(A)+P(B)

Disjoint≠Independent



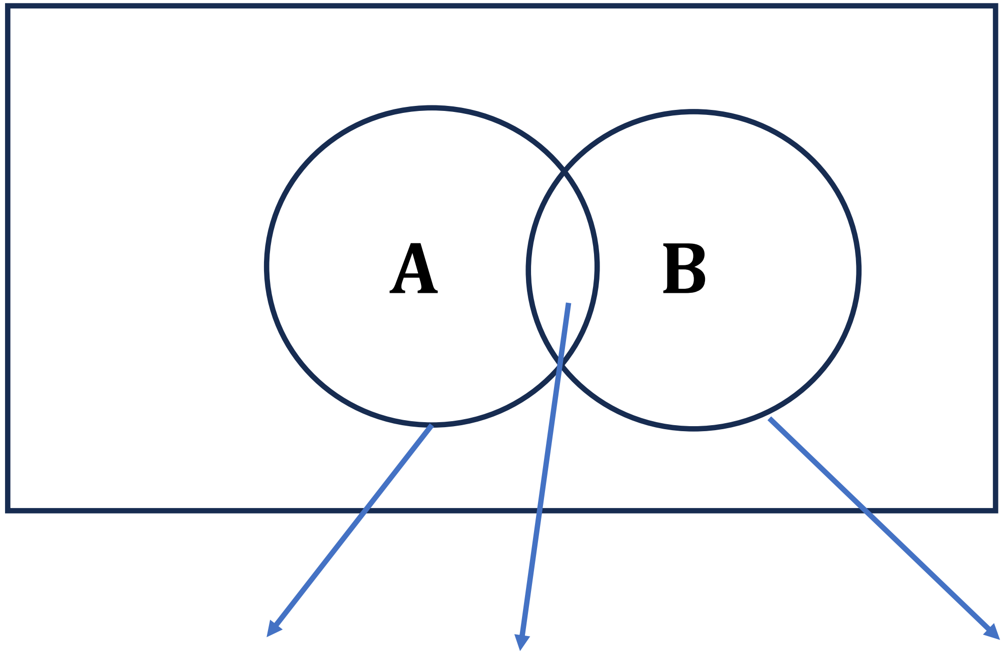
$P(A) = 0.2$

$P(B) = 0.2$

$$P(A \cap B) = 0 \neq P(A)P(B)$$

Disjoint $\nrightarrow$ Independent

Disjoint≠Independent



$$P(\,A \cup B) = 0.36 \neq P(A) + P(B)$$

$P(A) = 0.2$ $\quad$ $P(A \cap B) = 0.04$ $\quad$ $P(B) = 0.2$

Disjoint $\nleftarrow$ Independent

Random Variable (Discrete)

- X is called a random variable as it takes a numerical value that depends on the outcome of an experiment.

- Range of X: the set of possible values for X.

- Probability distribution: the probability distribution of a random variable X is a description of the probabilities associated with the possible values of X.

Probability Mass Function

- For a discrete random variable X with possible values $x_1, x_2, \dots, x_n$. A probability mass function $f(\cdot)$ is a function such that:

✓ $f(x_i) \geq 0$ for all $x_1, x_2, \dots, x_n$.

✓ $\sum_{i=1}^{n} f(x_i) = 1$

✓ $f(x_i) = \boxed{P(X = x_i)}$ for all $x_1, x_2, \dots, x_n$.

Probability that the random variable takes value x_i.

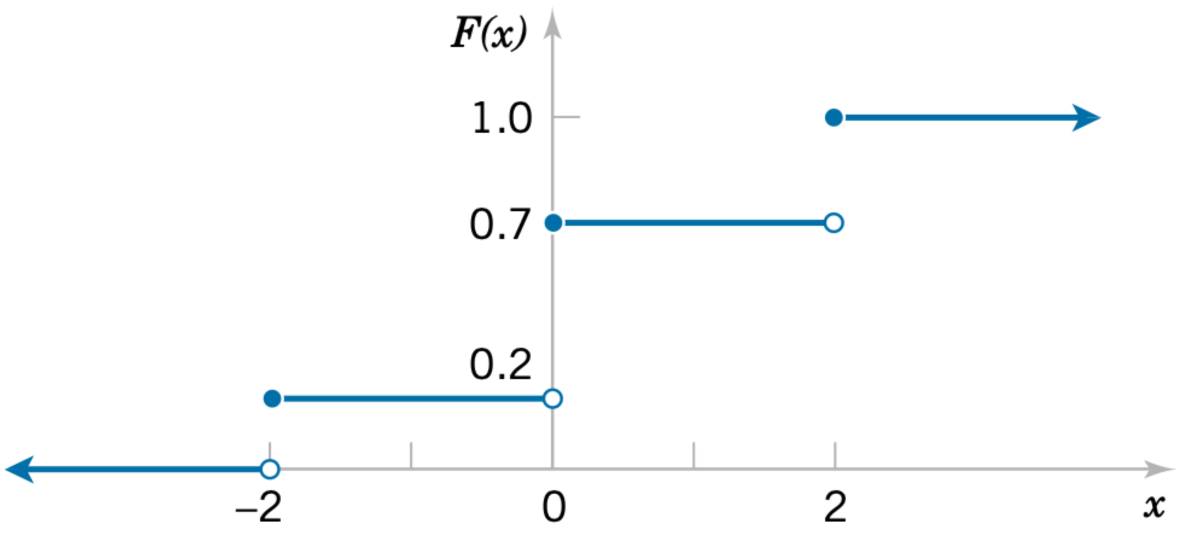
Cumulative Distribution Function

- The cumulative distribution function (cdf) gives the probability that the random variable X is less than or equal to x and is usually denoted by $F(x)$

- $F(x) = P(X \leq x)$

- $f(x) = F(x) - \lim_{y \uparrow x} F(y)$

$$F(x) = \begin{cases} 0 & x < -2 \\ 0.2 & -2 \le x < 0 \\ 0.7 & 0 \le x < 2 \\ 1 & 2 \le x \end{cases}$$



$f(-2)=0.2 \quad f(0)=0.5 \quad f(2)=0.3$

Mean and Variance

- Mean

E$[X]= \Sigma_x x\ P(X = x) = \Sigma_x x\ f(x)$

- Variance

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

Linearity: $E[\sum_i C_i X_i] = \sum_i C_i E[X_i]$

C_i is a constant

No assumption on X_i

Expectation of a function of X:

$$\mathrm{E}[g(X)] = \sum_x g(x)P(X = x) = \sum_x g(x)f(x)$$

Expectation of a constant: $E[C] = C$

Variance

$$E[\underbrace{(X - E[X])^2}_{X^2 - 2XE[X] + (E[X])^2}]$$

- $E[C] = C$ for constant C
- $E[X]$ is a constant
- $E\big[E[X]\big] = E[X]$

$$E[(X - E[X])^2] = E[X^2] \underbrace{- 2E[X]E[X] + (E[X])^2}_{-\,(E[X])^2}$$

More useful

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

1.Bernoulli distribution

- Take value 1 with probability p and value 0 with probability $1 - p$.

- $X \sim Bernoulli(p)$

- Mean=p
- Variance=p(1-p)

2.Binomial distribution

- N independent experiments
- Each experiment: success (with probability p) or failure (with probability $1 - p$).
- X: the number of success (failure).

- $X \sim Binomial(N, p)$

- Mean=Np
- Variance=Np(1-p)

3.Geometric distribution

- Continuously draw a Bernoulli R.V.
- The X-th sample is the first success.
- X follows a geometric distribution.

- $X \sim Geometric(p)$ or $X \sim Geo(p)$

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

Continuous R.V.

- A continuous random variable can take any value within its range (an interval of a union of multiple intervals of real numbers).

- We cannot list all the possible values and their probabilities as in the discrete case.

How to describe the probability?

- If P(E) = 0, then E is a zero-probability event.
- If E is empty, then E is impossible.

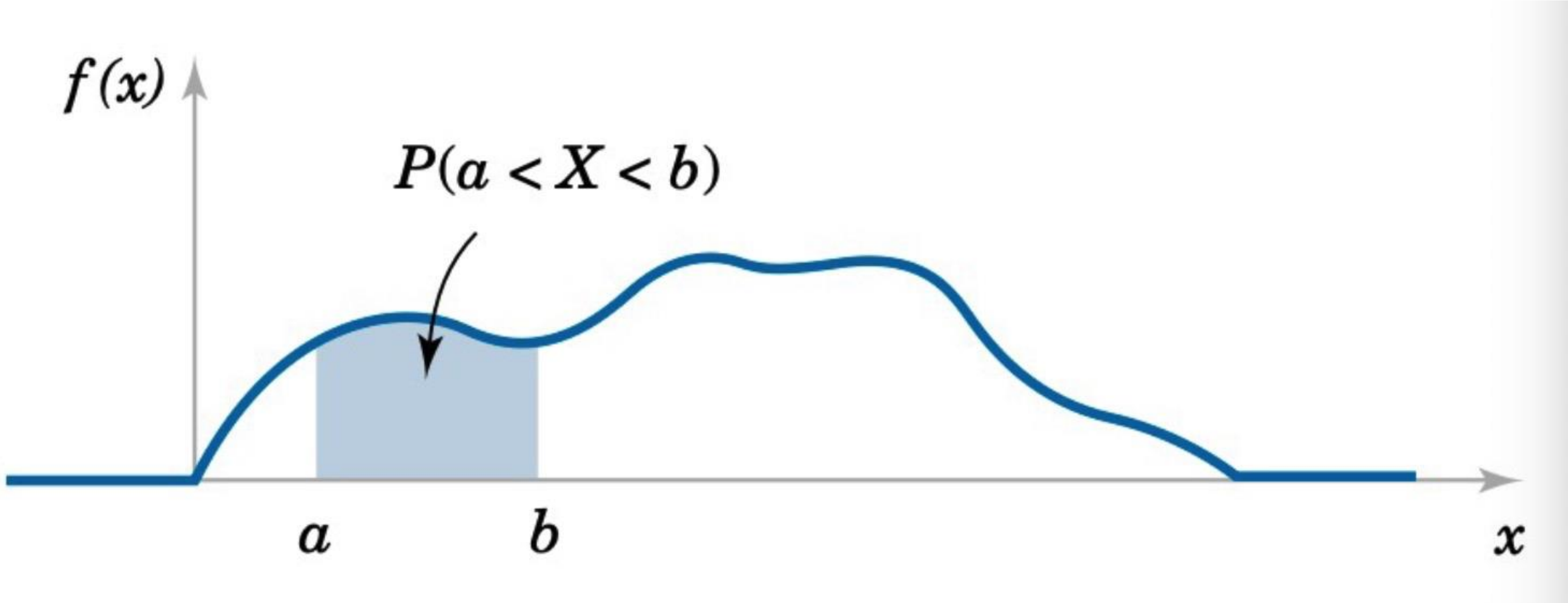
- For a continuous RV X, P(X=x) = 0 but {x} is not an impossible event.

- We will not use the probability mass function (pmf), namely P(X=x).

- Instead, we introduce a function f(ω), called the probability **density** function (pdf).
  - f(ω) > 0, if $\omega \in S$
  - f(ω) = 0, if $\omega \notin S$
  - $\int_{-\infty}^{\infty} f(x)dx = 1.$

Probability of $X \in [a, b]$

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



Properties of PDF

- For x that is not in the sample space, $f(x)=0$
- A large value of $f(x)$ means that the values around x is more likely to be observed. (remember this implication)
- As a pdf, f(x) can be larger than 1, while as a pmf, f(x) cannot be larger than 1.
  - $f(\omega) = 2$, if $\omega \in [0, 0.5]$
  - $f(\omega) = 0$, if $\omega \notin [0, 0.5]$

CDF

- Recall: the CDF of a discrete random variable X is

$$F(x) = P(X \leq x) = \sum_{\tilde{x} \leq x} f(\tilde{x})$$

- CDF for continuous random variable is defined as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

✓ $0 \leq F(x) \leq 1$

✓ If $x \leq y$, then $F(x) \leq F(y)$

For both discrete and continuous RVs

Mean and Variance

- Discrete:
  - ✓ Probability mass function.
- Continuous
  - ✓ Probability density function.

Summation ↔ Integration

- Mean

$$\mathrm{E}[X] = \sum x\, f(x)$$

- Variance

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

- Mean

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

- Variance

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} \left(x - E(X)\right)^2 f(x) dx$$

Expectation of g(X)

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

Uniform Distribution

- With the same ‘probability’, X takes a value within [a, b], where b>a.

Discrete version: toss a coin, roll a dice.

- What’s the pdf?

Uniform Distribution

- With the same probability, X takes a value within [a, b], where b>a.

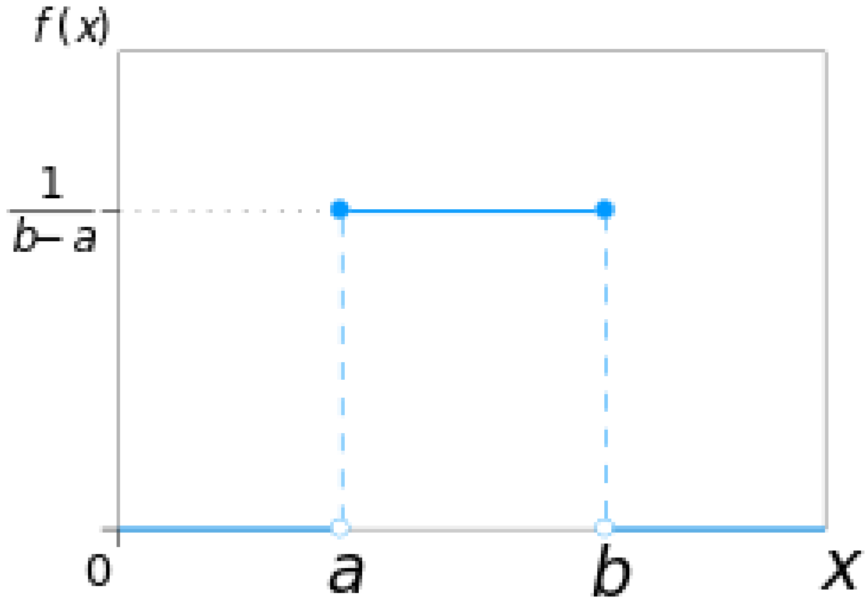
- What's the pdf?

- f(x) = c for $x \in [a, b]$ and f(x) = 0 for $x \notin [a, b]$
- As $\int_{-\infty}^{\infty} f(x)dx = c(b-a) = 1,$ we have

$$c = \frac{1}{b-a}$$

Uniform Distribution

- With the same probability, X takes a value within [a, b]
- $X \sim Uniform(a, b)$



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

Applications

- Given $X \sim Uniform(0,2)$
- What's the value of $E[2\, e^{X^2 + \cos(X)}]$?

Applications

- Given $X \sim Uniform(0,2)$
- What's the value of $E[2\ e^{X^2+\cos(X)}]$?

- f(x) = ½ for $x \in [0,2]$
- $E[2\ e^{X^2+\cos(X)}] = \int_0^2 2\ e^{x^2+\cos(x)} f(x) dx = \int_0^2 e^{x^2+\cos(x)} dx$

How to approximate $\int_0^2 e^{x^2+\cos(x)}dx$?

Given $X \sim Uniform(0,2)$, $E[2\, e^{X^2+\cos(X)}] = \int_0^2 e^{x^2+\cos(x)} dx$

- Draw N samples of $X \sim Uniform(0,2)$: $X_1, X_2, X_3, \ldots, X_N$

- Calculate $\frac{\Sigma_i 2\, e^{X_i^2+\cos(X_i)}}{N}$

Why? Expectation can be approximated by long-run average.

General Case

- How to calculate $\int_a^b h(x)dx$?

  - Draw N samples of X$\sim Uniform(a, b)$: $X_1, X_2, X_3, \ldots, X_N$
  - Calculate $\frac{\Sigma_i (\boldsymbol{b-a})\, h(X_i)}{N}$

- $E[h(x)]$ only gives you the average "height" of $h(x)$
- In order to get $\int_a^b h(x)dx$, which is the area, we need to multiply $E[h(x)]$ by $(b-a)$

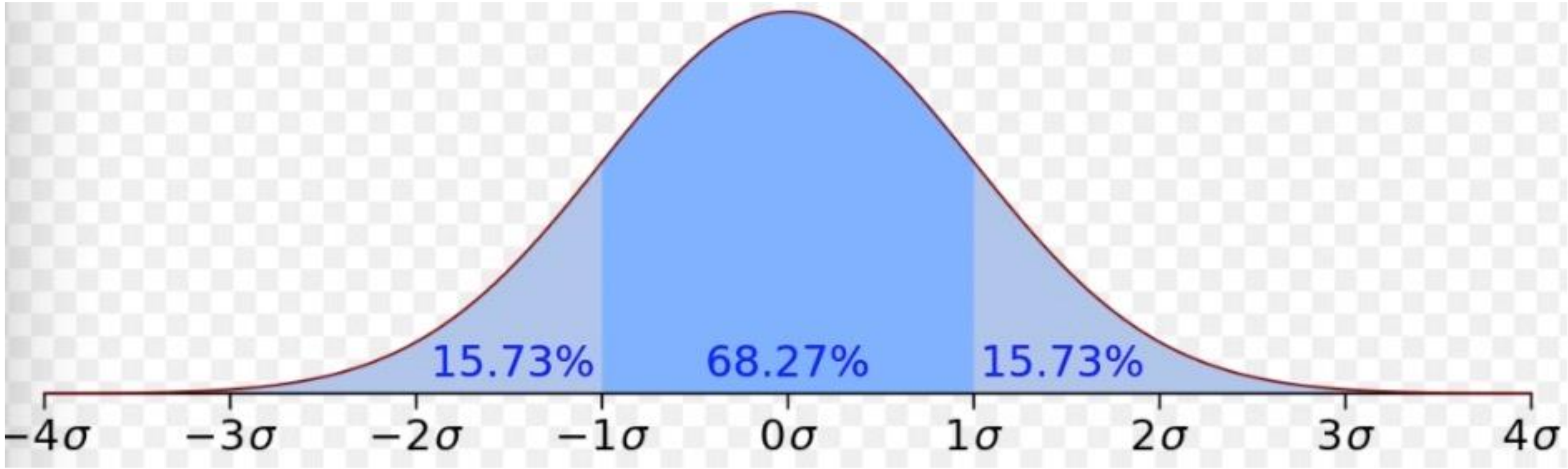
- Let $X \sim Uniform(a, b)$
- f(x) = 1/(b-a) for $x \in [a, b]$
- $E[(b-a)h(x)] = \int_a^b (b-a)h(x)f(x)dx = \int_a^b h(x)dx$

Normal Distribution

- X can be any real number
- Parameters: $\mu\ and\ \sigma$

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- $X \sim Normal(\mu, \sigma)$

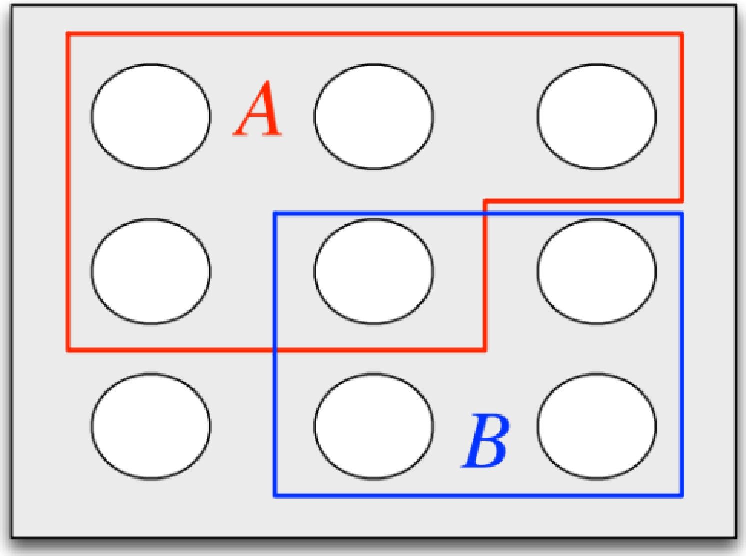


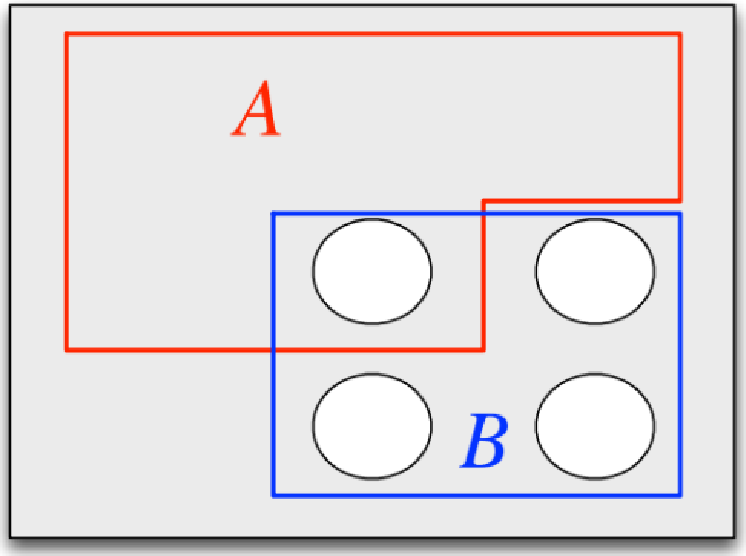
Conditional Probability

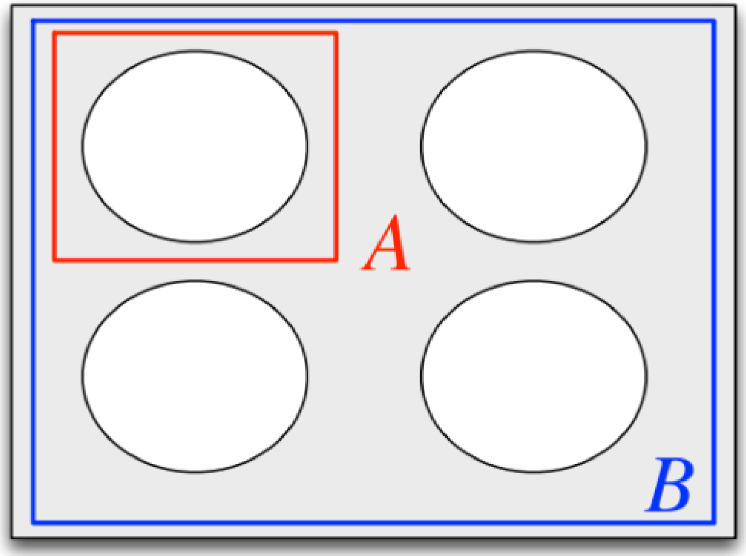
- Given the realization of event A, the probability of event B may change

- (Conditional probability) If A and B are events with P(B)>0, then the conditional probability of A given B, denoted by P(A|B), is defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$

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







$P(A|B)=?$ $P(A|B)=?$ $P(A|B)=?$

- Given the realization of event A, the probability of event B may change

- (Conditional probability) If A and B are events with P(B)>0, then the conditional probability of A given B, denoted by P(A|B), is defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- (Independence) Two events A and B are called independent if and only if $P(A \cap B) = \mathrm{P}(A)P(B)$

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

Bonus Question

Two Blind Red Envelops

Bonus Question

Independent and identically distributed

Two envelopes have i.i.d rewards R_1, R_2 that are drawn from a uniform $[0, 1]$ distribution. Because the rewards are i.i.d, selecting any of the envelopes gives you a 50% chance of selecting the one with the largest reward. Suppose that you select envelope one, but are allowed to switch to envelope two upon seeing the realization of the first envelope. A friend of yours proposes the following switching policy: pick a number $t \in (0, 1)$. If $R_1 > t$, then keep R_1 and otherwise switch to R_2.

(a) (4 points) What is the probability of selecting the envelope with higher reward under this strategy for arbitrary t?

(b) (3 points) What would be the optimal choice of t?

(c) (3 points) What would be the expected reward under the optimal t?