A CRASH COURSE ON PROBABILITY THEORY

Discrete and Continuous Probabilities

DISCRETE PROBABILITY THEORY

Events and Probabilities

- Consider a random process:
  - E.g., throw a die, pick a random card from a deck of cards
- Each possible outcome is a simple even (or sample point)
- The sample space Ω is the set of all possible simple events
- An event is a set of simple events (a subset of the sample space)
- With each simple event E we associate a real number $0 \leq \Pr(E) \leq 1$ which is the probability of event E

Probability Space – Definition

- A probability space has three components:
  1. A sample space Ω, which is the set of all possible outcomes of the random process modeled by the probability space.
  2. A family of sets F representing the allowable events, where each set in F is a subset of the sample space Ω.
     - In discrete probability space we use $F =$ "all subsets of Ω"
  3. A probability function $\Pr: F \to R$ satisfying the definition below

- A probability function is any function $\Pr: F \to R$ that satisfies the following conditions
  1. For any event $E, 0 \leq \Pr(E) \leq 1$
  2. $\Pr(\Omega) = 1$
  3. For any finite or countably infinite sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \Pr(E_i)$$

Corollary: The probability of an event is the sum of the probabilities of its simple events.

Example

- Consider the random process defined by the outcome of rolling a die:
- Each facet of the die is a simple event

$$\Omega = \{1,2,3,4,5,6\}$$

- We assume that all facets of the die are equally likely:

$$\Pr(1) = \Pr(2) = \cdots = \Pr(6) = \frac{1}{6}$$

- Event $E \;=\; \text{“odd outcome”} \;=\; \{1,3,5\}$

$$\Pr(E) = \frac{3}{6} = \frac{1}{2}$$

Example

- Rolling two dice. Sample space is the set of all ordered pairs
$$\Omega = \{(i,j): 1 \le i,j \le 6\}$$
- We assume that each simple event (i,j) has probability $\Pr(i,j) = \frac{1}{36}$

- Event $E_1 =$ "sum=2" $= \{(1,1)\}$: $\Pr(E_1) = \frac{1}{36}$
- Event $E_2 =$ "sum=3" $= \{(1,2),(2,1)\}$: $\Pr(E_2) = \frac{2}{36}$
- Event $E_3 =$ "sum at most 6" $=$
$$\{(1,1),(1,2),(1,3),(1,4),(1,5),(2,1),(2,2),(2,3),(2,4),(3,1),(3,2),(3,3),(4,1),(4,2),(5,1)\}$$
$$\Pr(E_3) = \frac{15}{36}$$
- Event $E_4 =$ "both dice have odd numbers": $\Pr(E_4) = \frac{1}{4}$
  - There are four combinations, equally likely: (odd,odd), (even, even), (odd, even), (even, odd)
- Event $E_5 = E_3 \cap E_4 = \{(1,1),(1,3),(1,5),(3,1),(3,3),(5,1)\}$: $\Pr(E_5) = \frac{6}{36}$

Conditional Probability

- In conditional probability we consider the probability that an event E_1 occurs, given that we know that an event E_2 has occurred.

- Sample space: "all the people living in Ioannina"
- Event $E_1 =$ "people living in Ioannina who were born in Ioannina"
- Event $E_2 =$ "people living in Ioannina who are students at UoI"
- Conditional probability of a person living in Ioannina to be born in Ioannina given that they are students at UoI:

$$\Pr(E_1 | E_2)$$

- Conditional probability is different from joint probability

$$\Pr(E_1 \cap E_2)$$

- This is the probability that a person living in Ioannina is born in Ioannina and is also a student at UoI

Computing Conditional Probability

The conditional probability that event E occurs given that event F occurs is

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}$$

The conditional probability is well defined only if $\Pr(F) > 0$

By conditioning on F we restrict the sample space to the set F. Thus, we are interested in $\Pr(E \cap F)$ normalized by $\Pr(F)$.

Corollary:

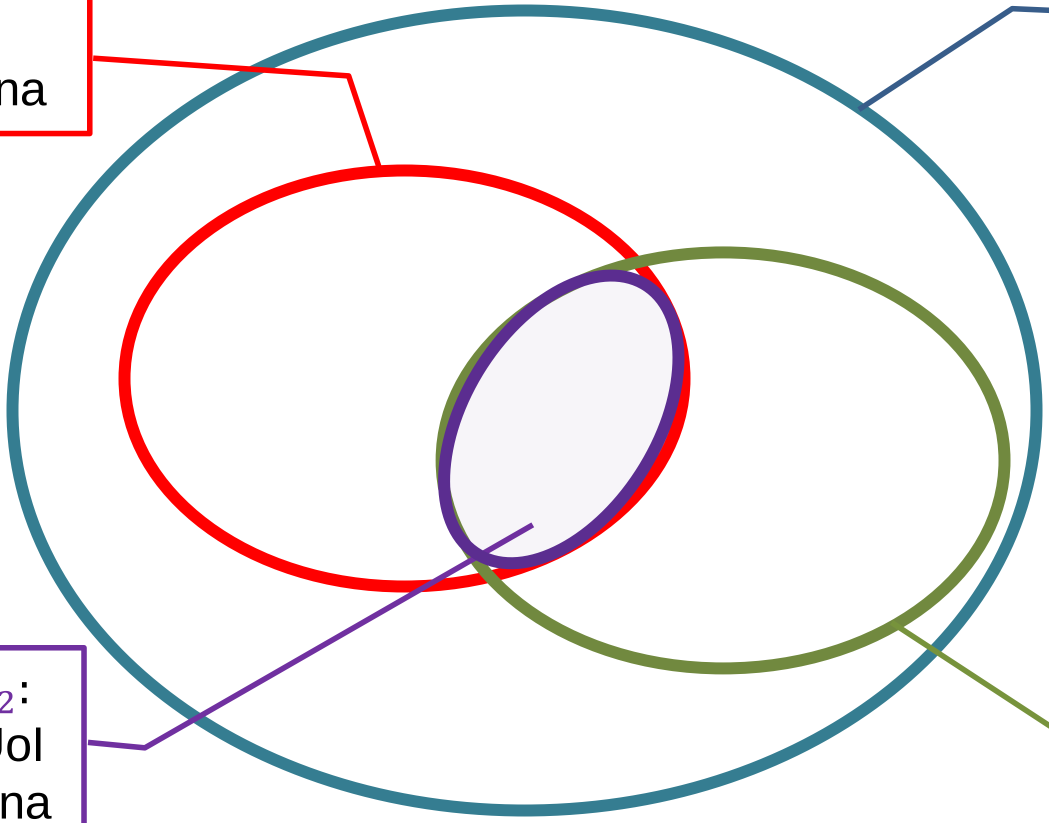
$$\Pr(E \cap F) = \Pr(E|F)\Pr(F)$$

Venn Diagrams

- We can represent events using Venn Diagrams

Event E_1:
People born in Ioannina

Sample space:
All people living in Ioannina



$$\Pr(E_1 \cap E_2) = \frac{purple}{blue}$$

The probability that a randomly selected person living in Ioannina is a student at UoI, born in Ioannina

$$\Pr(E_1 | E_2) = \frac{purple}{green}$$

The probability that a randomly selected student at UoI is born in Ioannina

Event $E_1 \cap E_2$:
Students at UoI
born in Ioannina

Event E_2:
Students at UoI

Example

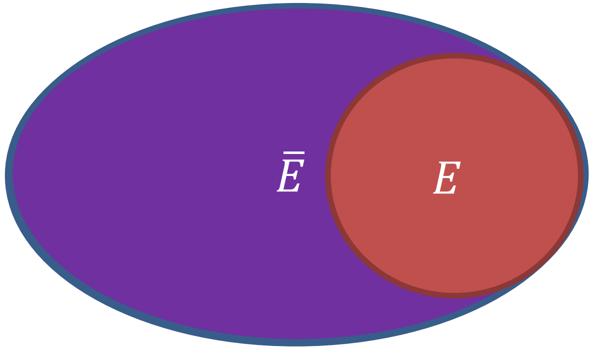
- What is the probability when rolling two dice that their sum is 8, given that their sum is even
- $E_1 =$ "sum is 8" = $\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$: $\Pr(E_1) = \frac{5}{36}$
- $E_2 =$ "sum is even": $\Pr(E_2) = \frac{1}{2}$
- $\Pr(E_1|E_2) = \frac{\Pr(\mathrm{E}_1 \cap E_2)}{\Pr(E_2)} = \frac{\frac{5}{36}}{\frac{1}{2}} = \frac{5}{18}$

$$\Pr(E_1 \cap E_2) = \Pr(E_1)$$

$E_1 \subseteq E_2$: When the sum is 8 then it is even.

Complement

- Let Ω be the sample space. If $E \subseteq \Omega$ is an event, then the complement of the event E is the event $\bar{E}$, such that
  - $E \cap \bar{E} = \emptyset$
  - $E \cup \bar{E} = \Omega$



- Example:
  - $E =$ “sum of dice is even”
  - $\bar{E} =$ “sum of dice is odd”
- Probability of the complement: $\Pr(\bar{E}) = 1 - \Pr(E)$
- Sometimes it is more convenient to work with the complement.
- Example: Compute the probability that the sum of two dice is greater than 2
  - $E =$ “sum of dice > 2”
  - $\bar{E} =$ “sum of dice = 2” $= \{(1,1)\}$
  - $\Pr(E) = 1 - \Pr(\bar{E}) = 1 - \frac{1}{36} = \frac{35}{36}$

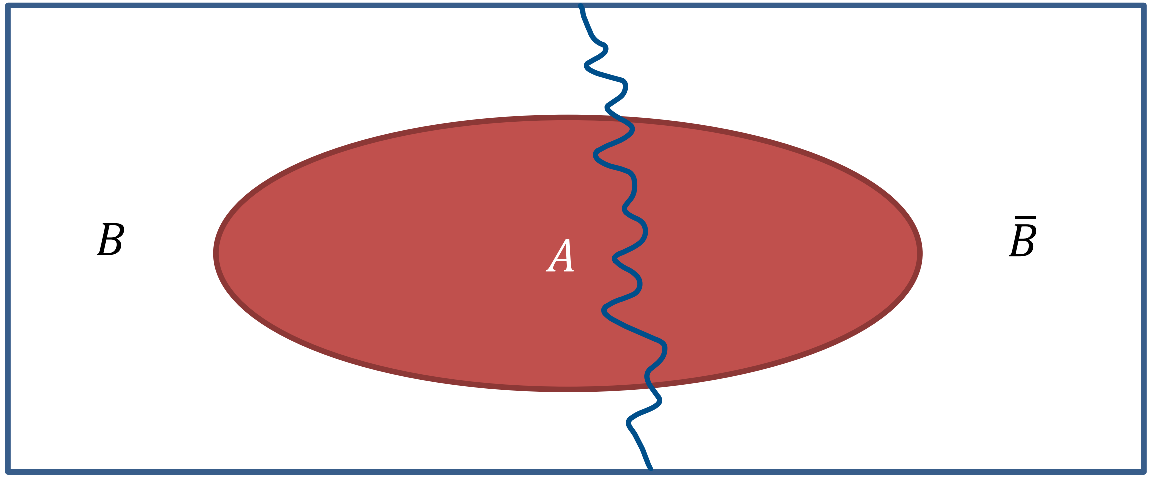
A Useful Identity

- Consider two events A, B

$A \cap B$ and $A \cap \bar{B}$ are disjoint

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \bar{B})$$
$$= \Pr(A|B)\Pr(B) + \Pr(A|\bar{B})\Pr(\bar{B})$$

Recall that $\Pr(A \cap B) = \Pr(A|B)\Pr(B)$



Application

- Compute the probability that a randomly selected person has height greater than 1.80
- Assume that we know that the probability that a man has height greater than 1.80 is 0.4, and the probability that a woman has height greater than 1.80 is 0.04
- Event A = "height greater than 1.80". We want $\Pr(A)$
- Event B = "person is a woman". $\Pr(B) = 0.51$
- We can now compute $\Pr(A)$

$$\Pr(A) = \Pr(A|B)\Pr(\mathrm{B}) + \Pr(\mathrm{A}|\bar{B})\Pr(\bar{B})$$
$$= 0.04 * 0.51 + 0.4 * 0.49 = 0.41$$

Bayes Rule

- Express the conditional probability $\Pr(E_1|E_2)$ as a function of the probability $\Pr(E_2|E_1)$

$$\Pr(E_1|E_2) = \frac{\Pr(E_2|E_1)\Pr(E_1)}{\Pr(E_2)}$$

$$= \frac{\Pr(E_2|E_1)\Pr(E_1)}{\Pr(E_2|E_1)\Pr(E_1) + \Pr(E_2|\overline{E_1})\Pr(\overline{E_1})}$$

Example: A-posteriori probability

- We are given 2 coins:
  - one is a fair coin A
  - the other coin, B, has head on both sides

- We choose a coin at random, i.e. each coin is chosen with probability $\frac{1}{2}$. We then flip the coin.

- Given that we got head, what is the probability that we chose the fair coin A???

Example: A-posteriori probability

- Event E_1 = "coin A was chosen"
- Event E_2 = "output was head"
- We want to compute $\Pr(E_1|E_2)$

- Using Bayes Rule

$$\Pr(E_1|E_2) = \frac{\Pr(E_2|E_1)\Pr(E_1)}{\Pr(E_2|E_1)\Pr(E_1) + \Pr(E_2|\overline{E_1})\Pr(\overline{E_1})}$$

$$= \frac{\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}\times\frac{1}{2} + 1\times\frac{1}{2}}$$

$$= \frac{1}{3}$$

Independent Events

- Two events E and F are independent if and only if
$$\Pr(E \cap F) = \Pr(E)\Pr(F)$$
- The probability of occurring together is equal to the product of the probabilities of occurring individually.

- Equivalently:
$$\Pr(E|F) = \Pr(E)$$
$$\Pr(F|E) = \Pr(F)$$
- The probability of one event occurring is not affected by the fact that we know the other event has occurred.

Examples

- Pick a random card from a deck:
  - $E =$ "ace was picked"
  - $F =$ "heart was picked"

Independent!
Even if we know that we have picked a heart
we still have probability $\frac{1}{13}$ to pick an ace
Two independent processes

- Roll a die:
  - $E =$ "even number" $= \{2,4,6\}$
  - $F =$ "number ≤ 4" $= \{1,2,3,4\}$

Independent!
The events are of the same process but
even if we know that we have picked a number ≤ 4
we still have probability $\frac{1}{2}$ to pick an even number

- Roll a die:
  - $E =$ "prime number" $= \{1,2,3,5\}$
  - $F =$ "number ≤ 4" $= \{1,2,3,4\}$

Not Independent!
If we know that we have picked a number ≤ 4
then we have probability $\frac{3}{4}$ to pick a prime number
while we have probability $\frac{4}{6}$ overall

Random Variables

- A random variable X on the sample space Ω is a function on Ω, that is, $X: \Omega \rightarrow R$
- A discrete random variable is a random variable that takes only a finite or countably infinite number of values.
- A random variable is a numeric quantity that we are interested in that is the by-product of the random process.
- By defining the random variable, we assign a value to every simple event in the sample space.

Examples

1. Roll a die: X_1 = "the number"
   - In this case the number we associate with each simple event is the value of the event.
2. Roll 2 dice: X_2 = "the sum of the two values"
   - For example, the simple event $(2,3)$ is assigned the value 5. Note that the same value is also assigned to the simple event $(3,2)$.
3. Flip two coins: $X_3 = \begin{cases} \$3 & \text{if two Heads} \\ \$1 & \text{otherwise} \end{cases}$
   - This random variable assigns value 3 to the event (H, H) and 1 to all other events
4. Pick a card: $X_4 = \begin{cases} 1 & \text{if card is Ace} \\ 0 & \text{otherwise} \end{cases}$
   - We assign value 1 to the event A, and zero to all other events. This models the case of "success"
5. Run QuickSort on a given matrix T: X_5 = "Running time of Quicksort"
   - The sample space is the set of all random choices made by the algorithm. Each one will result in a specific running time in $\{0, \ldots, n^2\}$

Probability Distribution

- Each value x of the random variable X, defines an event $(X = x)$ in the sample space Ω.
  - For example, for the random variable X_3 (money gained when drawing cards) the value $(X_3 = 3)$ corresponds to the event $\{(H, H)\}$, while the value $(X_3 = 1)$ corresponds to the event $\{(H, T), (T, H), (T, T)\}$
- We can thus compute the probability of a value $\Pr(X = x)$ (or $\Pr(x)$)
  - $\Pr(X_3 = 3) = \frac{1}{4}, \Pr(X_3 = 1) = \frac{3}{4}$
- The probability distribution function for random variable X gives the probability $\Pr(X = x)$ $(\Pr(x))$ for all possible values of X. The probability distribution should satisfy:
  - $0 \leq \Pr(x) \leq 1$, for all x
  - $\sum_x \Pr(x) = 1$, where the sum is over all possible values of X.

Random variables and Probability Distribution

- We sometimes define random variables simply by the set of values they take and the probability distribution, without explicit reference to the sample space
- For example, we may say that we have a random variable X that takes values $\{1,2,3,4\}$ with probability distribution $\Pr(i) = \frac{1}{4}$ (the uniform distribution)
- We often say, we have a random variable that follows the uniform distribution over the set $\{1, \ldots, n\}$
- In such cases you can think of the sample space as being the same as the field of values.

Independent Random Variables

Two random variables X and Y are independent if and only if

$$\Pr\big((X = x) \cap (Y = y)\big) = \Pr(X = x) \Pr(Y = y)$$

for all values x, y

- This definition means that all the events defined by the two variables are independent
- In simple terms, the value that one variable takes, does not depend on the value that the other variable takes.

- We also write: $\Pr(X, Y) = \Pr(X) \Pr(Y)$ or $\Pr(X|Y) = \Pr(X)$
  - $P(X, Y)$ is the joint distribution of variables X and Y
  - $P(X|Y)$ is the conditional probability distribution of variable X given Y

Example

- Rolling 5 dice:
  - The outcome of each roll is independent of the outcome of the other rolls
  - The sum of the first three rolls is independent of the sum of the last two rolls
- Drawing 3 cards:
  - The number of Aces we have is independent of the number of Hearts we get
- General rule: When repeating the same experiment multiple times, we assume that each trial is independent of the rest

Expectation

The expectation of a discrete random variable X, denoted by $E[X]$, is given by

$$E[X] = \sum_{x} x \Pr(X = x)$$

where the summation is over all values in the range of X

Think of the expectation as the mean value you would get if you took infinite values of the random variable X

Examples

- The expected value of one die roll is:

$$E[X] = \sum_{i=1}^{6} i \Pr(X = i) = \sum_{i=1}^{6} \frac{i}{6} = \frac{7}{2}$$

- The expected sum of two dice:

$$E[X] = \frac{1}{36}2 + \frac{2}{36}3 + \frac{3}{36}4 + \cdots + \frac{1}{36}12 = 7$$

- Throw two coins. If both are head you win $3, else you loose $1.1. What is the expected gain?

$$E[X] = 3\frac{1}{4} - 1.1\frac{3}{4} = -0.1\frac{3}{4}$$

Examples

- The expectation is not the most probable value. Consider random variable X that takes values $\{-2,1,2\}$ with probability $\{0.4, 0.1, 0.4\}$. The expected value is

$$E[X] = -2 \cdot 0.4 + 1 \cdot 0.1 + 2 \cdot 0.4 = 0.1$$

- The expectation may be unbounded. Consider the random variable X which takes value 2^i with probability $\frac{1}{2^i}$ for $i = 1,2,3, \dots$ (this is a distribution)

$$E[X] = \sum_{i=1}^{\infty} 2^i \frac{1}{2^i} = \sum_{i=1}^{\infty} 1 = \infty$$

Linearity of Expectation

- For any two random variables X and Y:

$$E[X+Y]=E[X]+E[Y]$$

- This holds for any random variables, X and Y do not need to be independent
- For any constant c and random variable X:

$$E[cX]=cE[X]$$

- Corollary: The expectation of a constant is the constant

$$E[c]=c$$

Examples

- Roll n dice. What is the expectation of the random variable X that is the sum of their output?
  - Define random variables $X_1, X_2, \dots, X_n$ for the output of the n dice
  - $X = \sum_{i=1}^{n} X_i$
  - $E[X] = \sum_{i=1}^{n} E[X_i] = n\frac{3}{2}$
- Roll 2 dice. What is the expectation of the random variable X that is the sum of the output of the first plus two times the output of the second?
  - Define random variables X_1, X_2 for the output of the two dice
  - $X = X_1 + 2X_2$
  - $E[X] = E[X_1] + 2E[X_2] = \frac{3}{2} + 2\frac{3}{2} = \frac{9}{2}$

Bernoulli Random Variable

- A Bernoulli Random Variable is one that takes values $\{0,1\}$. Bernoulli has a parameter p which is the probability of taking the value 1.

$$B = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

- Bernoulli variables are used as indicator variables, whether some event of interest happened or not
  - E.g., 1 if I draw an Ace, 0 otherwise

- Expectation:

$$E[B] = p \cdot 1 + (1-p) \cdot 0 = p = \Pr(B=1)$$

Binomial Random Variable

- A binomial random variable measures the number of successes in a sequence of n trials
  - E.g., toss a coin n times, random variable X is the number of times we get Head

- A binomial random variable X with parameters n, p, denoted $B(n, p)$ is defined by the following probability distribution for $k = 0,1,2, \dots, n$:

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

n:number of trials

p: probability of success

k: number of successes

$\binom{n}{k}$: number of ways to select k elements out of n elements

Expectation of a Binomial Random Variable

- We can compute the expectation using the standard formula:

$$E[X] = \sum_{k=0}^{n} k \Pr(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \cdots = np$$

- There is a simpler way. Ideas?
  - Define Bernoulli random variables $X_1, X_2, \dots, X_n$ for each trial with success probability p

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

$$E[X] = \sum_{i=1}^{n} E[X_i] = np$$

A useful formula

- Consider a discrete random variable X that takes values $1,2,3,\ldots n$. Sometimes it is easier to use the following formula to compute the expectation:

$$E[X] = \sum_{i=1}^{n} \Pr(X \geq i)$$

- Proof?

Expectation is not everything

- Consider the following two jobs:
  - One job gives salary 1000 euros per month
  - The other job gives salary 1 euro per month, plus a bonus of $1{,}000{,}000$ with probability $\frac{1}{1000}$
- Which job would you pick?

Variance

- The variance of a random variable X is
$$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

- Variance measures the expected deviation from the expected value, measured as the squared difference

- The standard deviation of a random variable X is
$$\sigma(X) = \sqrt{Var[X]}$$

Quiz

- Question: We have two events that are disjoint. Are they independent?
- Answer: No. Clearly, they are dependent. If one happens the probability of the other happening is zero.
- Question: A coin has probability p of being head. What is the probability that I throw the coin 10 times and I get all heads?
- Answer: Each coin toss is independent. Therefore, the probability is: p^{10}
- Question: A coin has probability p of being head. What is the probability that I throw the coin 10 times and I get at least one head?
- Answer: Consider the complement of this event: I get no heads. The probability of not getting a head is $1-p$. The probability of getting no heads is $(1-p)^{10}$. The probability of this not happening is $1-(1-p)^{10}$

Exercise

- Assume that N people checked coats in a restaurants. The coats are mixed up, and each person gets a random coat.
- How many people we expect to have gotten their own coats?

- Let X = "number of people that got their own coats". We want to compute $E[X] = \sum_{i=0}^{N} i \Pr(X = i)$. Not easy. Ideas?

- Define N Bernoulli random variables X_i:

$$X_i = \begin{cases} 1 & \text{person } i \text{ got their coat} \\ 0 & \text{otherwise} \end{cases}, \Pr(X_i = 1) = \frac{1}{N}$$

$$X = \sum_{i=1}^{N} X_i$$

$$E[X] = E\left[\sum_{i=0}^{N} X_i\right] = \sum_{i=0}^{N} E[X_i] = N\frac{1}{N} = 1$$

Exercise

- What is the probability that everyone gets their own coat?
- Incorrect argument: The probability that one person gets their coat is $\Pr(X_i = 1) = \frac{1}{N}$. The probability that everyone gets their coat is

$$\prod_{i=1}^{N} \Pr(X_i = 1) = \frac{1}{N^N}$$

- Where is the error in this?
- The random variables are not independent. Once one person has found their coat the probability for the rest changes.

- What is the correct probability?
- One way to compute it:

$$\Pr(X_1)\Pr(X_2|X_1)\cdots\Pr(X_N|X_{N-1},\dots,X_1) = \frac{1}{N}\frac{1}{N-1}\cdots 1 = \frac{1}{N!}$$

- It also follows from the fact that of all possible permutations of coats there is only one that is the correct one.

CONTINUOUS RANDOM VARIABLES

Continuous Random Variables

- A continuous random variable X is one that takes values on a real interval, rather than a discrete set
  - E.g., the height of a randomly selected person in Greece
  - E.g., the amount of rainfall in a specific location on a randomly selected day
- Since the range of the variable X is not countable or infinitely countable, it does not make sense to assign a probability to a specific real value
  - There are uncountably infinite of those, and also measurements are never exact.
- Instead, the probability is defined over intervals of values

Cumulative Probability Function

- Mathematically, a continuous random variable is defined through the cumulative probability function

$$F(x) = \Pr(X \leq x)$$

Which should have some nice properties (e.g. be non-decreasing and continuous)

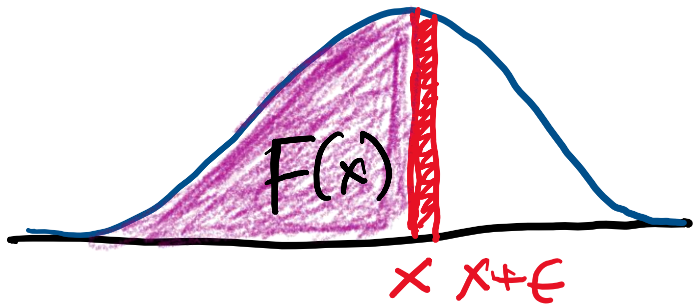
Probability Density Function

- More often, a random variable is defined by its probability density function $f(x)$.
- The function f is the derivative of the cumulative function F and it has the following two properties:
  1. $f(x) \geq 0$, for all x
  2. $\int_{-\infty}^{+\infty} f(x)dx = 1$

Probability Density Function

- The pdf is the closest analog to the probability function for the discrete case
  - It tells us how the probability mass (the samples) is distributed over the range of the random variable
- Sometimes, we may use $f(x)$ as the probability of value x
- The correct way to compute this though is to take the integral of f (the area under the curve) in the interval $(x, x+\epsilon)$

$$\Pr(x < X \leq x+\epsilon) = \int_{x}^{x+\epsilon} f(x)dx = F(x+\epsilon) - F(x)$$



$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Expectation and Indepedence

- The expectation is defined by taking the integral

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

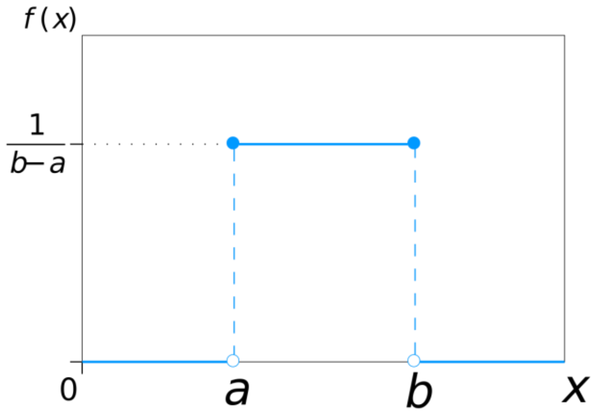
- Same properties hold for linearity of expectation

- Independence is defined using the cumulative or density function

$$F(x, y) = F(x)F(y)$$
$$f(x, y) = f(x)f(y)$$

Important continuous distributions

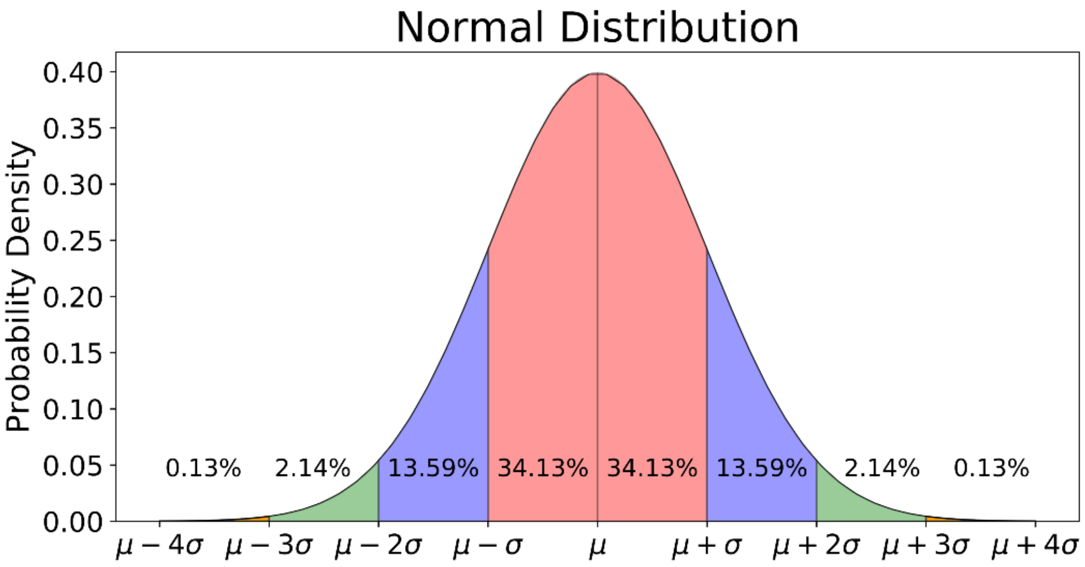
- Uniform distribution: The probability of any interval (a, b) is proportional to its length $b - a$.
  - The pdf is a flat line: equal mass everywhere.



- Gaussian/Normal distribution. Probability density function:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

It is fully characterized by the mean μ and the standard deviation σ



Central Limit Theorem

- Let $Y_1, Y_2, \ldots, Y_n$ be independent identically distributed random variables with mean μ and variance σ^2
  - For example, n height measurements from a broader population
- Let $\bar{Y} = \frac{1}{n}\sum_i Y_i$ be the mean value of the n random variables
  - Taking the mean height
- When n is large the random variable $\bar{Y}$ converges to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$
  - This means that if we repeat the height measurements multiple times, the distribution of the mean height will follow a gaussian distribution.