Introduction to Probability and Statistics

Probability is the mathematical way of quantifying uncertainty.

Put another way: probability is the study of theoretical possibilities and their likelihood of occuring.

Sample Spaces and Events

- **Sample space (*S*):** the set of all possible outcomes of our model or experiment
- **Elements:** the points in the sample space
- **Event:** a subset of the sample space

Sample Spaces and Events

- **Sample space (*S*):** the set of all possible outcomes of our model or experiment
- **Elements:** the points in the sample space
- **Event:** a subset of the sample space

Example: toss a coin twice

- **Sample space:** {HH, TT, HT, TH}
- **Elements:** {H, T}
- **Event that both tosses are the same:** {HH, TT}

Sample Spaces and Events

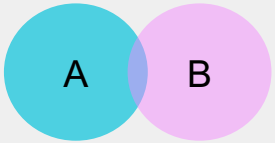
- **Complement (A^C):** Everything not in set A; for any event A, $P(A^C) = 1 - P(A)$
- **Union ($A \cup B$):** add up all events in A and B
- **Intersection ($A \cap B$):** all events that fall in both A and B

Sample Spaces and Events

- **Complement (A^C):** Everything not in set A; for any event A, $P(A^C) = 1 - P(A)$
- **Union ($A \cup B$):** add up all events in A and B
- **Intersection ($A \cap B$):** all events that fall in both A and B

- **Disjoint events**: the sets don't share any common events
  - P(A or B) = P(A) + P(B)

- **Joint events**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Probability

Every **event** A gets a **probability P(A)**, which is a real number.
Probabilities have some rules (called Axioms of Probability):

- $P(A) \geq 0$ *for every A*
- $P(Set) = 1$
- If $A_1, A_2,\ldots,A_i$ are disjoint, then $P(\cup A_1,A_2,\ldots,A_i) = \sum P(A_i)$

Two Ways of Interpreting Probabilities

- **Frequencies:** if we repeat enough trials, P(A) is the proportion of times we'll see A being true
    - E.g. If we say a fair coin has P(tossing heads) = .5, then tossing a coin lots (and lots and lots!) of times will get us 50% heads in the long term
- **Degrees of belief (Bayesian inference):** P(A) is our degree of belief that A is true (no repeated experiments necessary)
    - E.g. if we have a fair coin, we believe P(tossing heads) = .5

The difference starts to matter once we get to **inference**.

Independence

When events are independent (i.e. the occurrence of one event does not influence the occurrence of another event), the probabilities can be multiplied:

$$P(A,B) = P(A) \times P(B)$$

Independence

When events are independent (i.e. the occurrence of one event does not influence the occurrence of another event), the probabilities can be multiplied:

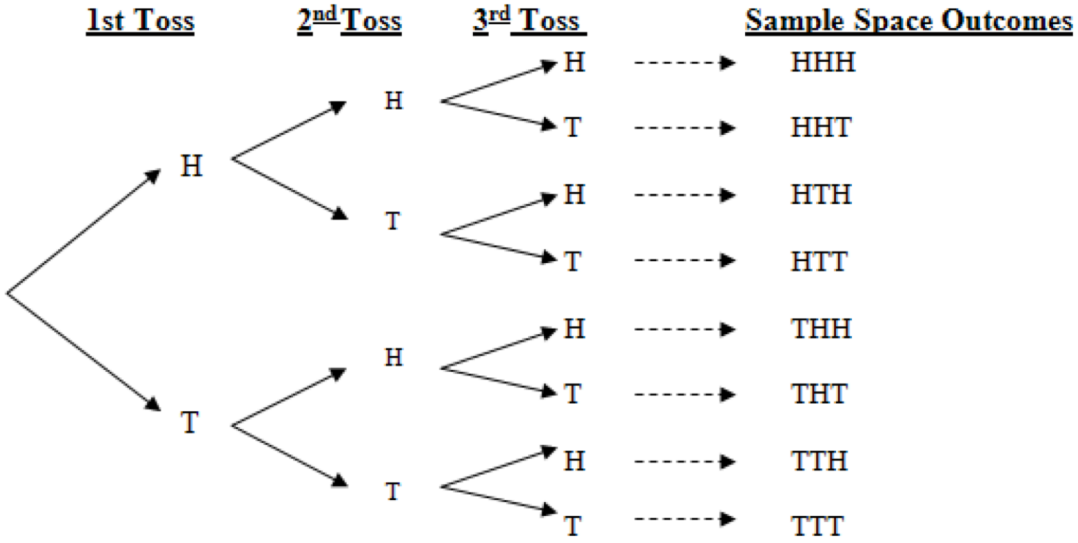
$$P(A,B) = P(A) \times P(B)$$

Independence can be:

- **Assumed:** We assume tosses of a fair coin are independent
- **Verified**: We derive then verify that $P(A,B) = P(A) \times P(B)$

Independence: Example

What's the probability of flipping three (fair) coins and getting (exactly) two heads?



$$P(H)P(H)P(T) + P(H)P(T)P(H) + P(T)P(H)P(H) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Dependence

When two events are dependent, the probability of one event occurring influences the likelihood of the other event.

Dependence

When two events are dependent, the probability of one event occurring influences the likelihood of the other event.

- What's the probability of drawing an ace from a deck of 52 cards?
- If we don't replace the drawn card, what's the probability of drawing a second ace?

$$P(A_1) = \frac{4}{52}$$

$$P(A_2|A_1) = \frac{3}{51}$$

Unions and Intersections of Events

- What's the **intersection** of two events?
  - If they're **independent**: P(A and B) = P(A) x P(B)
  - If they're **dependent**: P(A and B) = P(A) x P(B|A)

Unions and Intersections of Events

- What's the **intersection** of two events?
  - If they're **independent**: P(A and B) = P(A) x P(B)
  - If they're **dependent**: P(A and B) = P(A) x P(B|A)

- What's the **union** of two events?
  - If they're **independent**: P(A or B) = P(A) + P(B)
  - If they're **dependent**: P(A or B) = P(A) + P(B) - P(A and B)

Unions and Intersections of Events

- What's the **intersection** of two events?
  - If they're **independent**: P(A and B) = P(A) x P(B)
  - If they're **dependent**: P(A and B) = P(A) x P(B|A)

- What's the **union** of two events?
  - If they're **independent**: P(A or B) = P(A) + P(B)
  - If they're **dependent**: P(A or B) = P(A) + P(B) - P(A and B)

Draw some Venn diagrams!

Conditional Probability

Conditional probability is the probability of event A happening, given that event B has already happened. Define:

$$P(A|B) = P(A,B) / P(B)$$

Conditional Probability

Conditional probability is the probability of event A happening, given that event B has already happened. Define:

$$P(A|B) = P(A,B) / P(B)$$

Conditional probability example

- Test for a disease (D^+,$D-$) and get test results that are positive/negative (T^+, T^-)

$$P(T^+|D^+) = P(T^+ \cap D^+) / P(D)$$

Random Variables

Random variables are rules that assign a real number value to each element.

Random Variables

Random variables are rules that assign a real number value to each element.

- Toss a coin twice and let the **random variable X** be the **number of heads**.

	Probability	X
HH	1/4	2
TT	1/4	0
TH	1/4	1
HT	1/4	1

X	P(X = x)
0	1/4
1	1/2
2	1/4

Random Variables: Important Quantities

Discrete random variable: there are only finitely many values attained by the variable. (These are always thought of as functions on a fixed probability space!)

For these we can compute the following important quantities:

- Expected value: $E(X) := \mu_X := \sum_{\text{values } x \text{ of } X} P(X = x) \times x$
- Variance: $$\begin{aligned} Var(X) &:= (\sigma_X)^2 \\ &:= \sum_{\text{values } x \text{ of } X} P(X = x) \times (x - E(X))^2 \\ &= E\left((X - E(X))^2\right) \end{aligned}$$
- Standard deviation: $\sigma_X := \sqrt{Var(X)}$

Random Variables: Important Quantities

For two discrete random variables on the same probability space, we can talk about their **covariance**:

$$\text{Cov}(X, Y) = \sum P(X = x, Y = y)(x - E(X))(y - E(Y))$$

Random Variables: Inequalities for Expectation

There are some important inequalities in statistics. This one is called the **Cauchy-Schwartz inequality** and it's generally considered one of the most important inequalities in math because of its large number of applications.

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

Asymptotic Theory

As we collect a lot of data, we might be interested in what happens to our quantities of interest as we increase our data collection. This is called asymptotic analysis.

We might recall convergence from calculus (taking limits). Convergence with probabilities is a bit different. There are several different types of convergence; two important ones for us are:

1. **Convergence in distribution**
2. **Convergence in probability**

These two will give us two important theorems, discussed next (LLN and CLT).

Law of Large Numbers

Law of Large Numbers:

- Informal: The average value of a large number of independent samples of a random variable X gets arbitrarily close to its expected value E(X).
- Formal: Suppose that X_1, ..., X_n are independent random variables with the same probability densities as the random variable X, then:

$$\lim_{n\to\infty} \frac{X_1+X_2+\cdots+X_n}{n} = E(X)$$

This is a case of **converging in probability.**

Central Limit Theorem

Whereas the LLN was about samples from a distribution, the **Central Limit Theorem** is about the value of sample means from any distribution.

- Informal: Suppose X is a random variable with mean zero and finite variance. Then the sum of n trials of X divided by sqrt{n} approaches the normal distribution with mean zero and the same variance as X.
- Formal: Suppose X is a random variable with E(X) = 0 and variance less than infinity, and X_1, ..., X_n are independent random variables with the same probability distribution as X. Then the limit

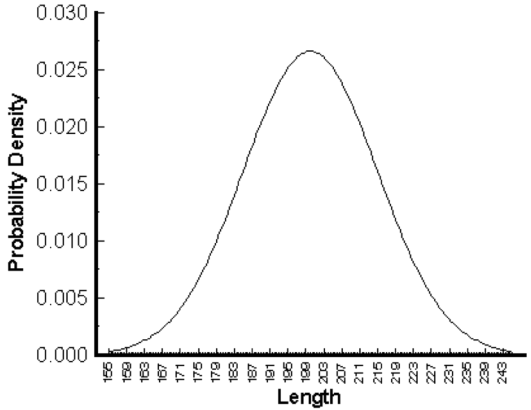
$$\lim_{n\to\infty} \frac{X_1+X_2+\cdots+X_n}{\sqrt{n}} = \mathcal{N}\left(0,\sigma^2\right)$$

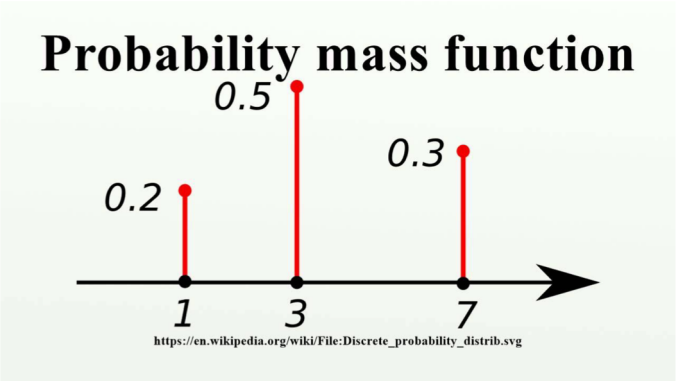
in distribution.

Probability Functions of Random Variables

Probability Density Function (PDF): For continuous random variables, a function whose value at any given sample can be interpreted as a relative likelihood that the value of the random value would equal that sample

Probability Mass Function (PMF): For discrete random variables, a function that gives the probability the the random variable X is exactly equal to some value.





Dependence and Conditional Probabilities

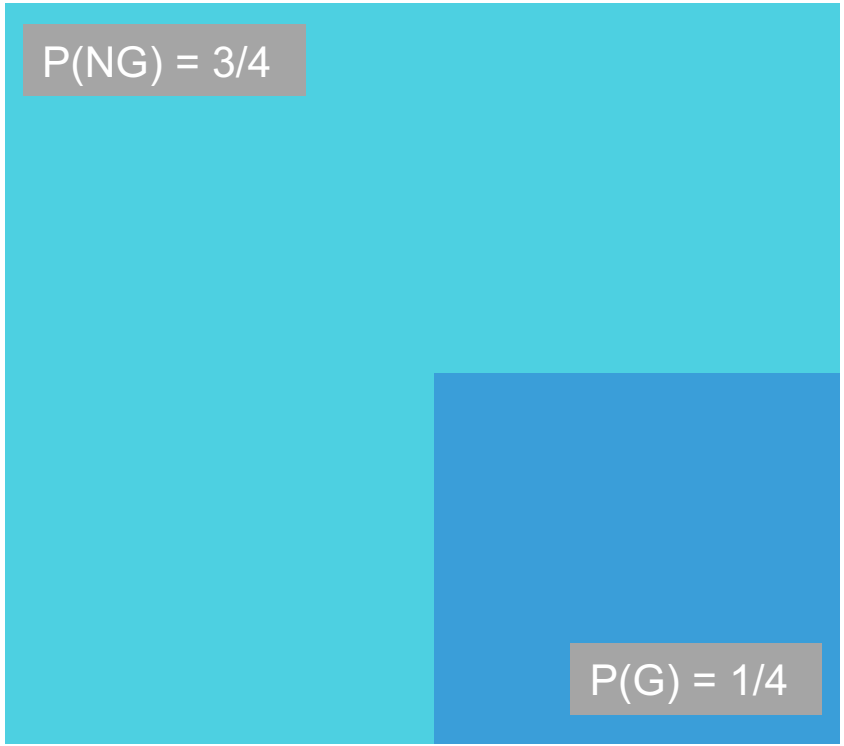
I am allergic to dogs. They make me sneeze. Sometimes dogs greet me. What is the probability that I sneeze?

P(Dog greets me) = P(G) = 1/4

P(Dog does not greet me) = P(NG) = 3/4

Visualize

P(G) = 1/4
P(NG) = 3/4



Dependence and Conditional Probabilities

Sometimes we know the **conditional probabilities** that depend on whether dogs say hello:

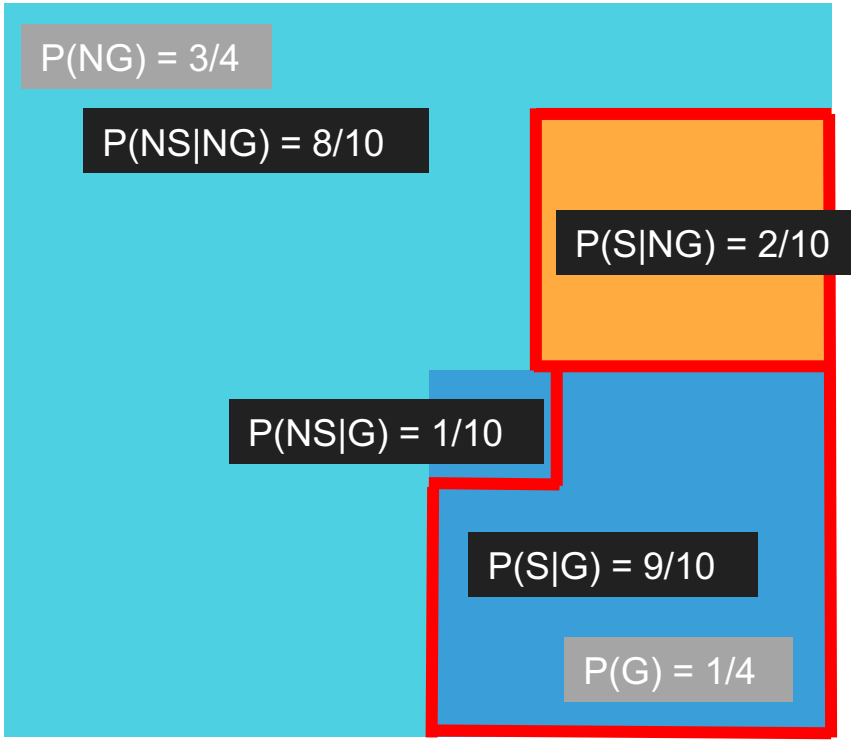
- P(Sneeze | Dog greets me) = P(S|G) = 9/10
- P(No Sneeze | Dog greets me) = P(NS|G) = 1/10
- P(Sneeze | Dog doesn't greet me) = P(S|NG) = 2/10
- P(No Sneeze | Dog doesn't greet me) = P(NS|NG) = 8/10

Visualize

P(G) = 1/4
P(NG) = 3/4

P(S|G) = 9/10
P(NS|G) = 1/10
P(S|NG) = 2/10
P(NS|NG) = 8/10

$$P(S) = P(S|G)\ P(G)\ + P(S|NG)\ P(NG)$$
$$= \frac{9}{10}\frac{1}{4} + \frac{2}{10}\frac{3}{4} = 0.375$$



Inference

Machine Learning Is Statistical Inference

Machine learning is the same thing as **statistical inference** (a case of computer science borrowing from a long history in statistics). In both cases, we're using data to learn/infer qualities of a distribution that generated our data (often termed the **DGP** or **data-generating process**).

We may care either about the whole distribution or just features, e.g. mean.

Parametric Vs. Non-parametric

If **inference** is about trying to find out our DGP, then we can say that a statistical model (of our data) is a set of possible distributions or maybe even regressions.

A **parametric model** is a particular type of statistical model: it's also a set of distributions or regressions, but they have a finite number of parameters.

Non-Parametric Statistics

In non-parametric statistics, we make fewer assumptions. In particular, we don’t assume that our data belong to any particular distribution (also called distribution-free inference).

This doesn’t mean that we know nothing, though!

Non-Parametric Inference

An example of non-parametric inference might be **creating a distribution of the data (CDF or cumulative distribution function) using a histogram**. In this case, we're not specifying parameters.

Parametric Models

A **parametric model** is a particular type of statistical model: it's also a set of distributions or regressions, but they have a finite number of parameters.

An example of a parametric model

Parametric Models: Normal Distribution

A **parametric model** is a particular type of statistical model: it's also a set of distributions or regressions, but they have a finite number of parameters.

An example of a parametric model: the Normal distribution

$$F = \{f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}, \underbrace{\mu \in \Re, \sigma > 0}_{\text{Parameters}}\}$$

Parametric Models: Maximum Likelihood

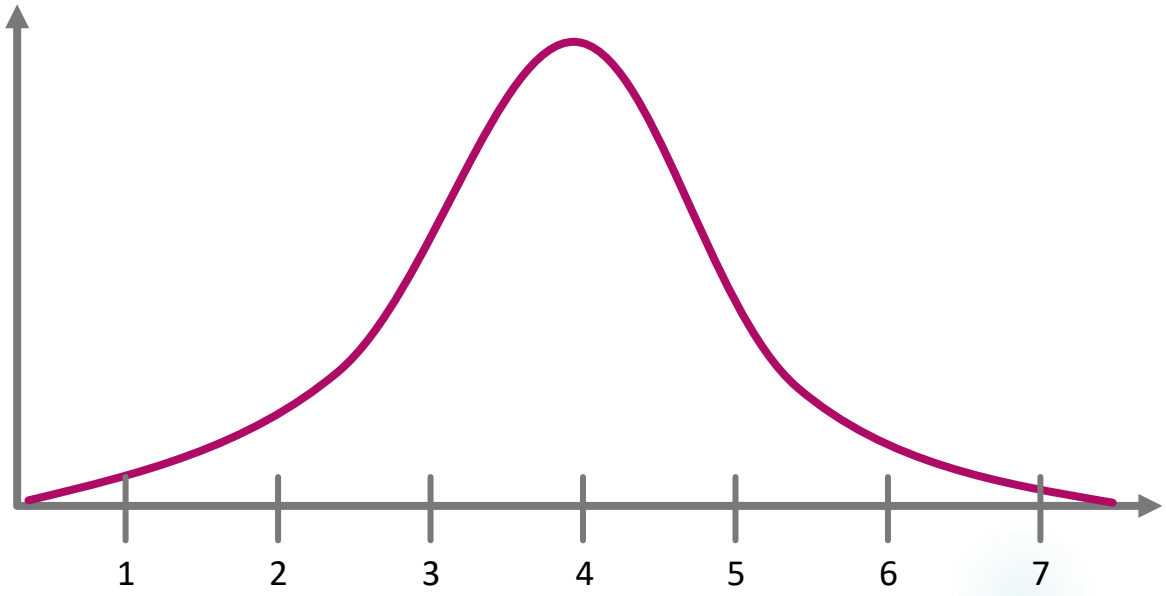
The most common way of estimating parameters in a parametric model is through **maximum likelihood estimation (MLE)**.

The **likelihood function** is related to probability and is a function of the **parameters** of the model:

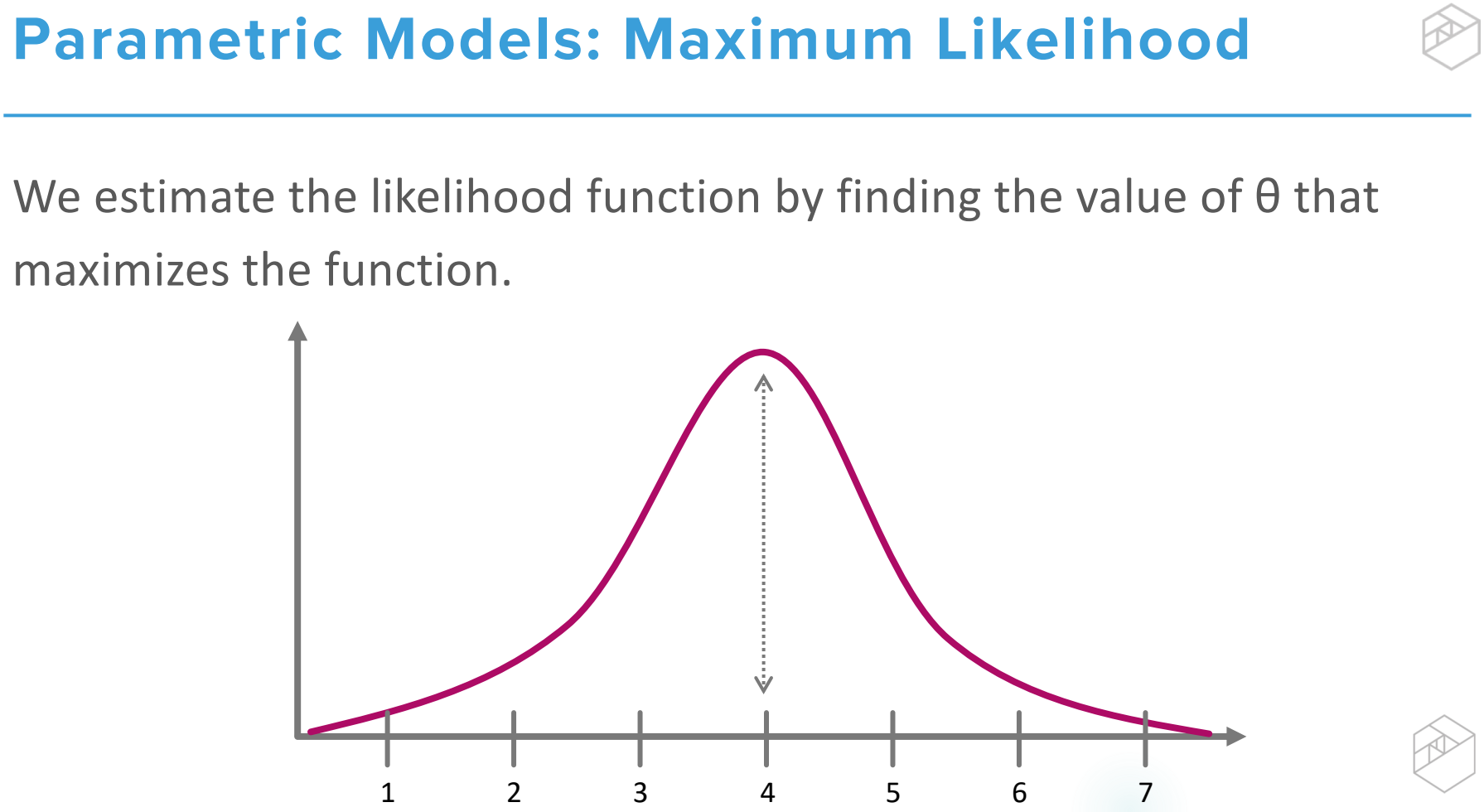
$$\underbrace{\mathcal{L}_n(\theta)}_{\text{Function of the parameters}} = \Pi_{i=1}^{n} f(X_i, \theta)$$

Parametric Models: Maximum Likelihood

We estimate the likelihood function by finding the value of θ that maximizes the function.



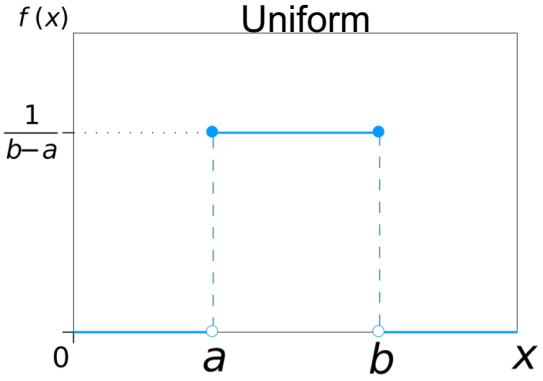
Parametric Models: Maximum Likelihood



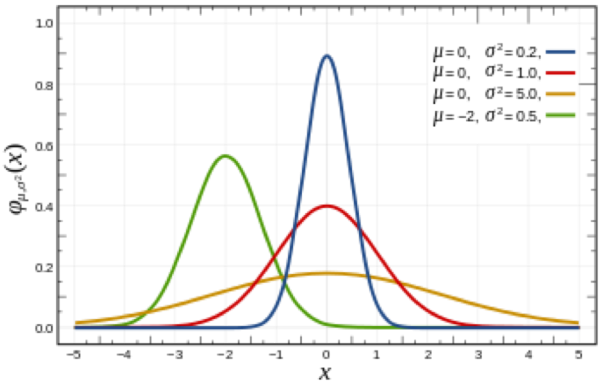
We estimate the likelihood function by finding the value of θ that maximizes the function.

Commonly Used Distributions

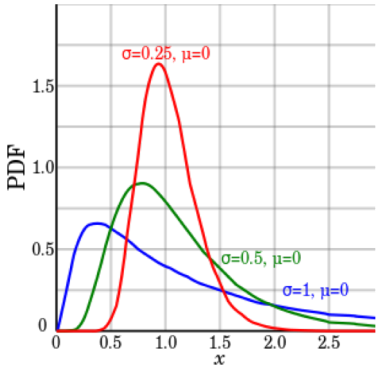
Uniform



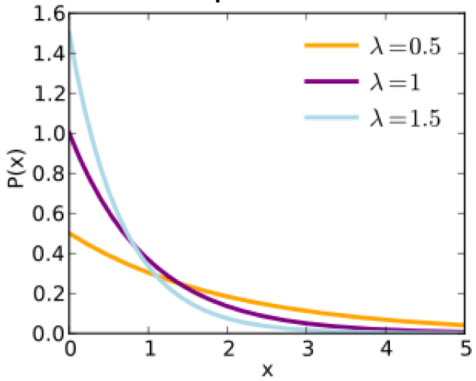
Gaussian/Normal



Log - Normal



Exponential



Poisson