Jerry Cain
CS109

May 29, 2025

Probability Reference

Contributions from Chris Piech, Lisa Yan, Tim Gianitsos, Alex Sun, Jerry Cain

Notation

This section maps between math notation used in CS109 and English. Note: "or" is not notation.

0.1 Events

E or F	Capital letters can denote events
A or B	Sometimes they denote sets
$\|E\|$ or $\|A\|$	Size of an event or set
E^C or A^C	Complement of an event or set
EF or AB	Intersection of events or sets
$E \cup F$ or $A \cup B$	Union of events or sets
$P(E)$	The probability of an event E
$P(E\|F)$	The conditional probability of an event E given F
$\binom{n}{k}$	Binomial coefficient
$\binom{n}{r_1, r_2, r_3}$	Multinomial coefficient

0.2 Random Variables

x or y or i	Lower case letters often denote regular variables
X or Y	Capital letters are used to denote random variables
$E[X]$	Expectation of X
$\text{Var}(X)$	Variance of X
$p_X(x)$	Probability mass function (PMF) of X
$p_{X,Y}(x, y)$	Joint probability mass function (PMF) of X and Y
$p_{X\|Y}(x\|y)$	Conditional probability mass function (PMF) of X given Y
$f_X(x)$	Probability density function (PDF) of X
$f_{X,Y}(x, y)$	Joint probability density function (PDF) of X and Y
$f_{X\|Y}(x\|y)$	Conditional probability density function (PDF) of X given Y
$F_X(x)$	Cumulative distribution function (CDF) of X
$F_{X,Y}(x, y)$	Joint cumulative distribution function (CDF) of X and Y
$F_{X\|Y}(x\|y)$	Conditional cumulative distribution function (CDF) of X given Y

$X \sim \text{Ber}(p)$	X is a Bernoulli random variable with parameter p
$X \sim \text{Bin}(n, p)$	X is a Binomial random variable with parameters n, p
$X \sim \text{Poi}(\lambda)$	X is a Poisson random variable with parameter λ
$X \sim \text{Geo}(p)$	X is a Geometric random variable with parameter p
$X \sim \text{NegBin}(r, p)$	X is a Negative Binomial random variable with parameters r, p
$X \sim \mathcal{N}(\mu, \sigma^2)$	X is a Gaussian random variable with mean μ and variance σ^2
$X \sim \text{Uni}(a, b)$	X is a Uniform random variable with parameters a, b
$X \sim \text{Exp}(\lambda)$	X is a Exponential random variable with parameter λ
$X \sim \text{Beta}(a, b)$	X is a Beta random variable with parameters a, b

1 Combinatorics

Inclusion-Exclusion Principle: If the outcome of an experiment can either be drawn from set A or set B, and sets A and B may potentially overlap (i.e., it is not guaranteed that $A \cap B = \emptyset$), then the number of outcomes of the experiment is $|A \cup B| = |A| + |B| - |A \cap B|$.

General Principle of Counting: If an experiment has r parts such that part i has n_i outcomes for all $i = 1, \ldots, r$, then the total number of outcomes for the experiment is $\prod_{i=1}^{r} n_i = n_1 \times n_2 \times \cdots \times n_r$

Basic Pigeonhole Principle: For positive integers m and n, if m objects are placed in n buckets, where $m > n$, then at least one bucket must contain at least two objects.

Permutations	Consider the number of ways to order n objects.
n objects are distinct (distinguishable)	$n(n-1)(n-2)\cdots 1 = n!$ ways
n_1 are indistinct (indistinguishable), n_2 are indistinct, …, and n_r are indistinct	$\dfrac{n!}{n_1!n_2!\ldots n_r!}$ ways
Combinations	Consider the number of ways to select groups of objects from a set of n distinguishable objects.
Select r objects	$\dfrac{n!}{r!(n-r)!} = \binom{n}{r}$ ways
Select r groups of objects, such that group i has size n_i, and $\sum_{i=1}^{r} n_i = n$	$\dfrac{n!}{n_1!n_2!\cdots n_r!} = \binom{n}{n_1, n_2, \ldots, n_r}$ ways
Bucketing	Consider the number of ways to place n objects into r containers.
n distinguishable objects	r^n ways
n indistinguishable objects	$\dfrac{(n+r-1)!}{n!(r-1)!} = \binom{n+r-1}{n} = \binom{n+r-1}{r-1}$ ways

2 Probability

2.1 Definitions, Axioms, and Corollaries

Frequentist definition of probability:

$$P(E) = \lim_{n\to\infty} \frac{n(E)}{n}$$

Axiom 1: $0 \le P(E) \le 1$
Axiom 2: $P(S) = 1$
Axiom 3: If E and F are mutually exclusive ($E \cap F = \emptyset$), then $P(E) + P(F) = P(E \cup F)$

Corollary 1: $P(E^C) = 1 - P(E) \quad (= P(S) - P(E))$
Corollary 2: $E \subseteq F$, then $P(E) \le P(F)$
Corollary 3: $P(E \cup F) = P(E) + P(F) - P(EF)$ (Inclusion-Exclusion Principle)

General Inclusion-Exclusion Principle:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1<\cdots<i_r} P(E_{i_1}E_{i_2}\ldots E_{i_r})$$

Define S as a sample space with equally likely outcomes. Then $P(E) = \dfrac{|E|}{|S|}$.

DeMorgan's Laws applied to probability:

$$P((E \cup F)^C) = P(E^C \cap F^C)$$
$$P((E \cap F)^C) = P(E^C \cup F^C)$$

2.2 Conditional Probability

Def. conditional probability	$P(E \mid F) = \dfrac{P(EF)}{P(F)} = \dfrac{P(E \cap F)}{P(F)}$
Chain rule	$P(EF) = P(E \mid F)P(F)$
	$P(E_1E_2\ldots E_n) = P(E_1)P(E_2 \mid E_1)\ldots P(E_n \mid E_1E_2\ldots E_{n-1})$
Law of Total Probability	$P(F) = P(F \mid E)P(E) + P(F \mid E^C)P(E^C)$
	$P(F) = \sum_{i=1}^{n} P(F \mid E_i)P(E_i)$
Bayes' Theorem	$P(E \mid F) = \dfrac{P(F \mid E)P(E)}{P(F)}$
	$= \dfrac{P(F \mid E)P(E)}{P(F \mid E)P(E) + P(F \mid E^C)P(E^C)}$
	$= \dfrac{P(F \mid E)P(E)}{\sum_i P(F \mid E_i)P(E_i)}$

Conditional paradigm: If we consistent conditionally on an event G, all of the laws of probability still hold.

2.3 Independence

Independence: Two events E and F are independent if and only if $P(EF) = P(E)P(F)$. It can be shown that independence of E and F implies:

- $P(E|F) = P(E)$ and $P(F|E) = P(F)$
- $P(E|F^C) = P(E)$ and $P(F|E^C) = P(F)$

In general, n events $E_1, E_2, \ldots, E_n$ are independent if for every subset with r elements (where $r \leq n$) it holds that:

$$P(E_{i_1}, E_{i_2}, \ldots, E_{i_r}) = P(E_{i_1})P(E_{i_2}) \ldots P(E_{i_r})$$

Two events E and F are **conditionally independent** given a third event G holds if $P(EF|G) = P(E|G)P(F|G)$.

3 Random Variables

Discrete Random Variables	
Probability Mass Function (PMF)	$p_X(x)$
PMF must sum to 1	$\sum_x p_X(x) = 1$
Probability with the PMF	$P(X = x) = p_X(x)$
Cumulative Distribution Function (CDF)	$F_X(a) = \sum_{x \leq a} p_X(x)$
Continuous Random Variables	
Probability Density Function (PDF)	$f_X(x)$
PDF must integrate to 1	$\int_{-\infty}^{\infty} f_X(x)dx = 1$
Probability with the PDF	$P(a \leq X \leq b) = \int_a^b f_X(x)dx$
Cumulative Distribution Function (CDF)	$F_X(a) = \int_{-\infty}^{a} f_X(x)dx$

We can compute the probability that the random variable X lies in an interval using the CDF, F_X: $P(a < X \leq b) = F_X(b) - F_X(a)$.

3.1 Expectation and Variance

Other names for expectation: mean, average, first moment, expected value.

Definition: $E[X] = \sum_x x p_X(x)$ — X discrete, PMF p_X
$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx$ — X continuous, PDF f_X

Linearity of Expectation: $E[aX + bY + c] = aE[X] + bE[Y] + c$

Law of the Unconscious Statistician (LOTUS):
$E[g(X)] = \sum_x g(x) p_X(x)$ — X discrete, PMF p_X
$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx$ — X continuous, PDF f_X

Linearity of expectation is often stated as: The expectation of a sum is equal to the sum of expectations.

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Definition of **variance**: $\text{Var}(X) = E[(X - E[X])^2]$.

- Most often computed as $\text{Var}(X) = E[X^2] - (E[X])^2$.
- Note: $\text{Var}(X) \geq 0$.
- Standard deviation is defined as $\text{SD}(X) = \sqrt{\text{Var}(X)}$. Note: $\text{SD}(X) \geq 0$.

3.2 Common Discrete Distributions

If a random variable follows a particular distribution we use the $\sim$ symbol to represent that the type of the random variable and pass in the appropriate parameters. For example if X follows a Normal distribution with mean 5 and variance 4 we write $X \sim \mathcal{N}(5, 4)$.

All probability mass functions (PMFs) are 0 outside the support.

Bernoulli Random Variable. $X \sim \text{Ber}(p)$
An indicator variable that takes on the value 1 ("success") or 0. Often the variable is defined to be 1 if an underlying event has occured, 0 otherwise.

PMF: $p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$ Support: $\{0, 1\}$

$E[X]$: p Var(X): $p(1 - p)$

Parameter: p: The probability that X is 1

Note: Sometimes in Machine learning algorithms, a differentiable version of the PMF is used: $p^k(1 - p)^{1-k}$. We will talk about this later.

Binomial Random Variable. $X \sim \text{Bin}(n, p)$

A variable that represents the number of successes in a fixed number of independent trials. The probability of success must be the same for each trial.

PMF:	$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$	Support:	$\{0, 1, 2, \ldots, n\}$
$E[X]$:	np	$\text{Var}(X)$:	$np(1-p)$
Parameters:	n: the number of trials		
	p: the probability of success of each trial		

Note: $\text{Bin}(1, p) = \text{Ber}(p)$.

Poisson Random Variable. $X \sim \text{Poi}(\lambda)$

The number of events occurring in a fixed interval of time or space if these events occur independently with a constant average rate.

PMF:	$p_X(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}, k \geq 0$	Support:	$\{0, 1, 2, \ldots\}$
$E[X]$:	λ	$\text{Var}(X)$:	λ
Parameter:	λ: the average number of events per fixed interval.		

Note: The Poisson RV is the number of events in an interval of time. The Exponential RV is a continuous RV that models the time until the next event occurs. They have the same parameter, λ.

Note 2: The Poisson can approximate the Binomial when λ is "moderate" (in this class, defined as $n > 20$ and $p < 0.05$ or $n > 100$ and $p < 0.1$) when the trials are mildly dependent, or even when the probability of success varies slightly between trials.

Geometric Random Variable. $X \sim \text{Geo}(p)$

The number of independent Bernoulli trials until the first success. The probability of success must be the same for each trial.

PMF:	$p_X(k) = (1-p)^{k-1} p$	Support:	$\{1, 2, \ldots\}$
$E[X]$:	$\dfrac{1}{p}$	$\text{Var}(X)$:	$\dfrac{1-p}{p^2}$
Parameter:	p: the probability of success of each trial		

Negative Binomial Random Variable. $X \sim \text{NegBin}(r, p)$

The number of independent Bernoulli trials until the r-th success. The probability of success must be the same for each trial.

PMF:	$p_X(k) = \binom{k-1}{r-1} (1-p)^{k-r} p^r$	Support:	$\{r, r+1, \ldots\}$
$E[X]$:	$\dfrac{r}{p}$	$\text{Var}(X)$:	$\dfrac{r(1-p)}{p^2}$
Parameters:	r: the total number of successes to obtain		
	p: the probability of success of each trial		

Note: $\text{NegBin}(1, p) = \text{Geo}(p)$.

3.3 Common Continuous Distributions

All probability density functions (PDFs) are 0 outside the support.

Uniform Random Variable. $X \sim \text{Uni}(a, b)$

PDF:	$f_X(x) = \dfrac{1}{b-a}$	Support:	$a \leq x \leq b$
$E[X]$:	$\dfrac{a+b}{2}$	$\text{Var}(X)$:	$\dfrac{(b-a)^2}{12}$

Exponential Random Variable. $X \sim \text{Exp}(\lambda)$

The waiting time until an event occurs when events occur independently with a constant average rate.

PDF:	$f_X(x) = \lambda e^{-\lambda x}$	Support:	$x \geq 0$
$E[X]$:	$\dfrac{1}{\lambda}$	$\text{Var}(X)$:	$\dfrac{1}{\lambda^2}$
CDF:	$F_X(x) = 1 - e^{-\lambda x}$		

Note: The Exponential RV models the time until the next event occurs. The Poisson RV is a discrete RV that models the number of events in an interval of time. They have the same parameter, λ.

Note: The Exponential RV is memoryless, in that the time you wait until the first success is distributed as an Exponential RV, independent of the amount of time you have waited so far.

Normal (Gaussian) Random Variable. $X \sim \mathcal{N}(\mu, \sigma^2)$

PDF: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Support: $-\infty < x < \infty$

$E[X]$: μ $\mathrm{Var}(X)$: σ^2

Note: When $\mu = 0$ and $\sigma^2 = 1$ ("zero mean, unit variance"), X is called a Standard Normal with CDF Φ.

Note 2: The Normal can approximate a Binomial with larger variance (in this class, defined as $np(1-p) > 10$. All trials must be independent. This approximation comes from the Central Limit Theorem.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with CDF F_X. The following properties hold:

Linearity: $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
Standard Normal: $Z = \frac{X-\mu}{\sigma}$ is the Standard Normal with CDF Φ.
Therefore $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

4 Joint Distributions

	Jointly Discrete X, Y	**Jointly Continuous** X, Y
Joint PMF	$p_{X,Y}(x,y) = P(X = x, Y = y)$	–
Joint PDF	–	$f_{X,Y}(x,y)$
Joint CDF	$F_{X,Y}(x,y) = P(X \le x, Y \le y)$	
Marginal distributions	$p_X(a) = \sum_y p_{X,Y}(a,y)$	$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)dy$
	$p_Y(b) = \sum_x p_{X,Y}(x,b)$	$f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x,b)dx$
Conditional distributions	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Independence	$p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$f_{X,Y}(x,y) = f_X(x)f_Y(y)$
	$p_{X\|Y}(x\|y) = p_X(x)$	$f_{X\|Y}(x\|y) = f_X(x)$
Bayes' Theorem	$p_{X\|Y}(x\|y) = \frac{p_{Y\|X}(y\|x)p_X(x)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{Y\|X}(y\|x)f_X(x)}{f_Y(y)}$

We can compute the probability involving two jointly distributed random variables X and Y using their joint CDF, $F_{X,Y}$: $P(a_1 < X \le a_2, b_1 < Y \le b_2) = F_{X,Y}(a_2,b_2) - F_{X,Y}(a_1,b_2) - F_{X,Y}(a_2,b_1) + F_{X,Y}(a_1,b_1)$.

In general, n random variables $X_1, X_2, \ldots, X_n$ are independent if for all $x_1, x_2, \ldots, x_n$:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \qquad \text{(jointly discrete)}$$

$$P(X_1 \le x_1, X_2 = x_2, \ldots, X_n \le_n) = \prod_{i=1}^{n} P(X_i \le x_i) \qquad \text{(jointly continuous)}$$

The n variables $X_1, X_2, \ldots, Xn$ are **independent and identically distributed** (i.i.d., iid, IID) random variables if they are independent and have the same PMF (if discrete) or PDF (if continuous).

4.1 Independent Sums of Random Variables

If X and Y are independent, then

$$P(X + Y = n) = \sum_k P(X = k)P(Y = n - k) \qquad (X, Y \text{ jointly discrete})$$

$$f_{X+Y}(\alpha) = \int_{-\infty}^{\infty} f_X(x) f_Y(\alpha - x)dx \qquad (X, Y \text{ jointly continuous})$$

Common Sums of Independent Random Variables

Independent X, Y	**Distribution of** $X + Y$
$X \sim \mathrm{Bin}(n_1, p)$, $Y \sim \mathrm{Bin}(n_2, p)$	$\mathrm{Bin}(n_1 + n_2, p)$
$X \sim \mathrm{Poi}(\lambda_1)$, $Y \sim \mathrm{Poi}(\lambda_2)$	$\mathrm{Poi}(\lambda_1 + \lambda_2)$
$X \sim \mathrm{Uni}(0,1)$, $Y \sim \mathrm{Uni}(0,1)$	$f_{X+Y}(\alpha) = \begin{cases} \alpha & 0 \le \alpha \le 1 \\ 2 - \alpha & 1 < \alpha \le 2 \\ 0 & \text{otherwise} \end{cases}$
$X \sim \mathcal{N}(\mu_1\sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$	$\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
Independent $X_1, X_2, \ldots, X_n$	**Distribution of** $\sum_{i=1}^n X_i$
$X_i \sim \mathrm{Bin}(n_i, p)$ for $i = 1, \ldots, n$	$\mathrm{Bin}(\sum_{i=1}^n n_i, p)$
$X_i \sim \mathrm{Poi}(\lambda_i)$ for $i = 1, \ldots, n$	$\mathrm{Poi}(\sum_{i=1}^n \lambda_i)$
$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$	$\mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$

4.2 Statistics of multiple RVs

Law of The Unconscious Statistician, extended to $g(X, Y)$, a function of two jointly distributed random variables:

$$E[g(X,Y)] = \sum_x \sum_y g(x,y)p_{X,Y}(x,y) \qquad (X, Y \text{ jointly discrete})$$

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dydx \qquad (X, Y \text{ jointly continuous})$$

Conditional expectation of X given $Y = y$:

$$E[X|Y=y] = \sum_x xP(X=x|Y=y) = \sum_x xp_{X|Y}(x|y) \qquad (X, Y \text{ jointly discrete})$$

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \qquad (X, Y \text{ jointly continuous})$$

Law of Total Expectation:

$$\begin{aligned} E[X] &= E[E[X|Y]] \\ &= \sum_y E[X|Y=y]P(Y=y) && (Y \text{ discrete}) \\ &= \int_{-\infty}^{\infty} E[X|Y=y]f_Y(y)dy && (Y \text{ continuous, density } f_Y(y)) \end{aligned}$$

Definition of **covariance**: $\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$.

- Most often computed as $\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$.
- Correlation of X and Y: $\rho(X,Y) = \dfrac{\text{Cov}(X,Y)}{SD(X)SD(Y)}$
- Relation to variance: $\text{Var}(X) = \text{Cov}(X,X)$
- Symmetry: $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
- Non-linear: $\text{Cov}(aX+b,Y) = a\text{Cov}(X,Y)$
- Covariance of sums: $\text{Cov}(\sum_i X_i, \sum_j Y_j) = \sum_i \sum_j \text{Cov}(X_i, Y_j)$

Variance of sums:

$$\text{Var}(X+Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X,Y) + \text{Var}(Y)$$

Independence of two random variables X and Y implies

- $E[XY] = E[X]E[Y]$ (the converse is not necessarily true), and therefore
- $\text{Cov}(X,Y) = 0$ and $\rho(X,Y) = 0$, and furthermore
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

4.3 Multinomial Distributions

Multinomial Distribution

A distribution that models the counts of outcomes $i = 1, 2, \ldots, m$, respectively, in a fixed number of independent trials, where each trial results in one of m outcomes.

Joint PMF:	$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$
Support:	$\sum_{i=1}^m c_i = n$, where c_i is a non-negative integer for $i = 1, \ldots, m$
Parameters:	n: the total number of trials
	$p_1, p_2, \ldots, p_m$: the probabilities of m outcomes, where p_i is the probability of outcome i and $\sum_{i=1}^m p_i = 1$.

4.4 The Central Limit Theorem

The Central Limit Theorem states that the sum of i.i.d. random variables is normally distributed; by extension, the sample mean of i.i.d. random variables is also normally distributed. In other words, if random variables $X_1, X_2, \ldots, X_n$ are i.i.d. such that $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$,

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \qquad \text{as } n \to \infty,$$

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \qquad \text{as } n \to \infty,$$

$$Z = \frac{\left(\sum_{i=1}^n X_i\right) - n\mu}{\sigma\sqrt{n}} \qquad \text{as } n \to \infty,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean, and $Z \sim \mathcal{N}(0,1)$ is the Standard Normal.

Most textbooks will tell you that the CLT holds if $n \geq 30$ (where n is the number of IID random variables you are summing together), but the CLT can hold for smaller n depending on the distribution of your IID random variables.

4.5 Sampling Statistics

A **sample** is defined as i.i.d. random variables $X_1, X_2, \ldots, X_n$ such that the population mean is $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. The **population mean** is μ and the **population variance** is σ^2.

Sample statistic	Notation	Expression	Notes
Sample mean	$\bar{X}$	$\frac{1}{n}\sum_{i=1}^{n} X_i$	By CLT, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Unbiased: $E[\bar{X}] = \mu$.
Sample variance	S^2	$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	Unbiased: $E[S^2] = \sigma^2$
Standard error of the mean	SE	$\sqrt{\frac{S^2}{n}}$	Estimate of standard deviation of $\bar{X}$

4.6 Bootstrapped Hypothesis Testing

We determine if two samples have significantly different statistics (e.g., sample means) if it is very unlikely that the difference arose due to chance.

The **null hypothesis** is the assumption that both samples were drawn from the same underlying distribution. The **p-value** is the probability that we would see at least the observed difference (between the original two groups) had we drawn two similar groups according to the null hypothesis. If the p-value is less than 0.05, we reject the null hypothesis.

One way to determine the p-value is to use **bootstrapping**, where we assume that the underlying distribution under the null hypothesis is the unified group of both samples. Then, we repeatedly simulate the experiment of drawing two samples of the same sizes as the original two samples (with replacement), and we report the p-value as the fraction of the samples that have a difference in statistic at least that of the observed, original difference.

5 Parameter Estimation

The task of parameter estimation centers around estimating the parameter θ of a parametric distribution from the a sample of i.i.d. random variables $X_1, X_2, \ldots, X_n$. Given a particular value of a parameter θ, the **likelihood** of a sample of i.i.d. random variables is written as

$$L(\theta) = f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta),$$

where the last equality follows if the sample is drawn i.i.d. from a distribution $f(X|\theta)$, referring to a PMF (if X discrete) or PDF (if X continuous).

The **log likelihood** is written as $LL(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$.

An **estimate** of a parameter θ is often written as $\hat{\theta}$.

5.1 Maximum Likelihood Estimation

The **maximum likelihood estimator** (MLE) of a parameter is written as

$$\theta_{MLE} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} LL(\theta).$$

Common steps to compute θ_{MLE} come from calculus, by finding $\theta = \theta_{MLE}$ for where the first derivative is zero, thereby $LL(\theta_{MLE})$ is a local maximum.

In the following table, let sample $X_1, X_2, \ldots, X_n$ be drawn i.i.d. according to the stated distribution.

Distribution	Parameter	Maximum Likelihood Estimator
Bernoulli	p	$\frac{1}{n}\sum_{i=1}^{n} X_i$
Poisson	λ	$\frac{1}{n}\sum_{i=1}^{n} X_i$
Uniform	a	$\min(X_1, \ldots, X_n)$
	b	$\max(X_1, \ldots, X_n)$
Normal	μ	$\frac{1}{n}\sum_{i=1}^{n} X_i$
	σ^2	$\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$.
Multinomial	p_i	X_i/n, where X_i is count of outcome i in sample

6 Machine Learning

Supervised machine learning tasks are where we use data to train a prediction model $\hat{Y} = g(\mathbf{X})$. $\hat{Y}$ is a prediction of an output variable Y given some input variable $\mathbf{X}$.

The **feature vector X** is a vector of m discrete features, where $\mathbf{X} = (x_1, x_2, \ldots, x_m)$. In **regression**, the label Y is continuous; in **classification**, the **class label** Y is discrete. The **training data** consists of n datapoints of input-output pairs, $(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$.

6.1 Logistic Regression

Classification task with the following model, where class Y takes on binary values 0 or 1:

$$\hat{Y} = g(\mathbf{X}) = \arg\max_{y} P(Y|\mathbf{X})$$

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma(\theta_0 + \sum_{j=1}^{m} \theta_j x_j) = \sigma(\theta^T \mathbf{x}).$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function (aka logit function). The logistic regression model has parameter $\theta = (\theta_0, \theta_1, \ldots, \theta_m)$ and the dot product notation $\theta^T \mathbf{x}$ prepends $X_0 = 1$ to the feature vector $(X_1, X_2, \ldots, X_m)$.

Training: Find the maximum likelihood estimate of θ using gradient ascent. The gradient of the log conditional likelihood function $LL(\theta) = \sum_{i=1}^{n} \log P(Y = y^{(i)} | \mathbf{X} = \mathbf{x}^{(i)})$ is

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)} \qquad \text{for } j = 0, 1, \ldots, m$$

Testing: Given a feature vector $\mathbf{X} = (x_1, \ldots, x_m)$, predict $\hat{Y} = 1$ if $P(Y = 1 | \mathbf{X} = \mathbf{x}) > P(Y = 0 | \mathbf{X} = \mathbf{x})$, and $\hat{Y} = 0$ otherwise. Equivalently, predict $\hat{Y} = 1$ if $\theta_0 + \sum_{j=1}^{m} \theta_j x_j > 0$, and $\hat{Y} = 0$ otherwise.