# MAS Math Camp Day 1: Differentiation and Integration

Samuel Baugh
Jiayi Li

Department of Statistics
University of California, Los Angeles

September 25, 2021

# Welcome to math camp!

Good morning everyone! Welcome to the first day of math camp! Sessions will include:

1. Calculus Part I: 9:00-10:45am (Sam)
2. Calculus Part II: 11:00am-12:00pm (Sam)
3. Lunch 12:00-1:30pm
4. Linear Algebra Part I 1:30pm-2:45pm (Jiayi)
5. Linear Algebra Part II 3:00pm-4:30pmpm (Jiayi)
6. Introduction to Latex 4:30pm-5:00pm (Sam)

For this morning's session I have provided a 'cheat sheet' containing the essential information on differential calculus and integral calculus. There is also a second sheet containing the practice questions we will be discussing today.
Cheat sheet note: the second row/second column of "Essential Logarithm Rules" says $e^a + e^b$ when it should say $e^a \cdot e^b$.

# Welcome to math camp!

Warm-up Question: Show

$$(x - a)^2 = x^2 - 2xa + a^2$$

# Calculus Part I: Differentiation Learning Objectives

By the end of this lesson, you will be able to:

1. Understand the intuition of derivative as the "rate of change"
2. Be able to use essential derivative rules to differentiate functions commonly seen in statistics
3. Be able to manipulate logarithms and exponential functions
4. Become comfortable with using differentiation in order to find local maxima and minima of functions with multiple variables

The most important skill you will use in statistics is (4), maximizing and minimizing functions, which uses skills 1-3. Many of the practice questions you will see today will come up in tomorrow's lesson on maximum likelihood estimation!

# Definition of the Derivative: Rate of Change

- Definition of the Derivative: Rate of Change
- Formal Definition
- Derivative of a line
- Derivative of a quadratic function

# Formal Definition

$$f'(x) = \frac{f(x+h) - f(x)}{h}$$

# Derivative of a Line: $f(x) = ax + b$

# Derivative of a Quadratic: $f(x) = x^2$

# Expanding to the Power Rule:

The three rules we have encountered so far are:

1. If $f(x) = x^2$, then $f'(x) = 2x$
2. If $f(x) = ax + b$, then $f'(x) = a$

# Sum Rule:

Rule we have encountered so far:

1. If $f(x) = x^a$, then $f(x) = ax^{a-1}$

What is the derivative of $f(x) = x^3 + x^2$?

# Fraction Trick

$$f(x) = \frac{1}{x^2} = x^{-2}$$

This allows use the power rule to find derivatives of quotients!

# A quick note about notation

The derivative of $f(x)$ can be written as $f'(x)$ or $\frac{d}{dx}f(x)$.
The second notation is preferred when when you don't necessarily want to write out our function in terms of $f(x)$, for example
$\frac{\partial}{\partial x}(x^2 + ax + b) = 2x + a$.

# Practice (sum and power rules):

The two rules we have encountered so far are:

1. The derivative of $x^a$ is $\frac{\partial}{\partial x}x^a = ax^{a-1}$ (power rule)
2. The derivative of $f(x) + g(x)$ is $(\frac{\partial}{\partial x}f(x) + \frac{\partial}{\partial x}g(x))$ (sum rule)
3. Quotients of the form $\frac{1}{x^a}$ can be written as $x^{-a}$ in order to use the power rule

# Practice (sum and power rules) answers:

What is the derivative of:

1. $f(x) = x^3 + 2x^2 + 3x + 1$?
2. $f(x) = x^3 + x^{-2}$

# Side-note: Logarithm Rules

The left and right hand sides are equivalent:

| $\log(ab)$ | $\log(a) + \log(b)$ |
|---|---|
| $e^{a+b}$ | $e^a e^b$ |
| $e^{\log(a)}$ | $a$ |
| $\log(e)$ | $1$ |
| $a\log(b)$ | $\log(b^a)$ |

These are on the cheat-sheet for future reference! Practice: show $\log(e^{\log(x)+2}) = \log(x) + 2$.

**Typo Alert:** After printing the cheat sheet I noticed that it says $e^a + e^b$ when it should say $e^a e^b$ in the second row.

# Log Rules for Exponential and Logarithmic Functions

1. $\frac{\partial}{\partial x} e^x = e^x$
2. $\frac{\partial}{\partial x} \log x = \frac{1}{x}$

# Product Rule

$$\frac{\partial}{\partial x} f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$

Example: $x^2 e^x$

# Composition Rule

$$\frac{\partial}{\partial x} f(g(x)) = f'(g(x))g'(x)$$

Example: $e^{x^2}$

# Composition Rule

$$\frac{\partial}{\partial x} f(g(x)) = f'(g(x))g'(x)$$

Another example: $(2x + 1)^3$

# Essential derivative rules: Practice!

Practice questions: differentiate the following (2, 3, 4, and 5 under "Practice: Essential Derivative Rules" on the hand-out):

- Power and composition rules:

$$f(x) = (x - a)^2$$

- Product and log rules:

$$f(x) = x^2 \log(x)$$

- Composition and log rules:

$$f(x) = \log(x^2)$$

- Composition, power, and exponent rules:

$$f(x) = e^{-(x-a)^2/b}$$

# Practice Solutions

On board

# Derivatives with respect to Multiple Variables

$$f(x, y) = x^2 y$$

# Second Derivatives

Example: $f(x) = (x-2)^2 + 3$

# Concavity: The Smile Rule :)

Example:

$$f(x) = (x-2)^2 + 3 \text{ vs } g(x) = -(x-2)^2 + 3$$
$$f'(x) = 2(x-2) \text{ vs } g'(x) = -2(x-2)$$
$$f''(x) = 2 \text{ vs } g''(x) = -2$$

# Maximization and Minimization

Where is the maximum of $f(x) = -(x-2)^2 + 3$, and what is its value? What happens to the slope at that point?

# Steps to Finding Local Maxima or Minima

Step 1: Take the derivative of your function.

$$f(x) = (x-2)^2 + 3$$

# Steps to Finding Local Maxima or Minima

Step 1: Take the derivative of your function.

$$f(x) = (x-2)^2 + 3$$

Step 2: Set your derivative to zero ($f'(x) = 0$ and solve for $x$)

$$f'(x) = 2(x-2) = 0$$

# Steps to Finding Local Maxima or Minima

Step 1: Take the derivative of your function.

$$f(x) = (x-2)^2 + 3$$

Step 2: Set your derivative to zero ($f'(x) = 0$ and solve for $x$)

$$f'(x) = (x-2)^2 + 3, f(x) = 2(x-2)$$

Step 3: Take the second derivative of your function, $f''(x)$.

$$f''(x) = 2$$

# Steps to Finding Local Maxima or Minima

Step 1: Take the derivative of your function.

$$f(x) = (x-2)^2 + 3$$

Step 2: Set your derivative to zero ($f'(x) = 0$ and solve for $x$)

$$f'(x) = (x-2)^2 + 3, f(x) = 2(x-2)$$

Step 3: Take the second derivative of your function, $f''(x)$.
Step 4: If $f''(x) > 0$, you have found a local minimum. If $f''(x) < 0$, you have found a local maximum. If $f''(x) = 0$, your point is neither a local minima or maxima.

# Steps to Finding Local Maxima or Minima

Step 1: Take the derivative of your function.

$$f(x) = (x-2)^2 + 3$$

Step 2: Set your derivative to zero ($f'(x) = 0$ and solve for $x$)

$$f'(x) = (x-2)^2 + 3, f(x) = 2(x-2)$$

Step 3: Take the second derivative of your function, $f''(x)$.
Step 4: If $f''(x) > 0$, you have found a local minimum. If $f''(x) < 0$, you have found a local maximum. If $f''(x) = 0$, your point is neither a local minima or maxima.
Step 5: Plug in the value or values you found in step 2 into your function, and check endpoints if applicable.

# Steps to Finding Local Maxima or Minima (summary)

1. Take the derivative of your function.
   *Example:* For $f(x) = (x-2)^2 + 3$ the derivative is $f'(x) = 2(x-2)$
2. Set your derivative to zero, $f'(x) = 0$, and solve for $x$.
   *Example:* Set $f'(x) = 2(x-2) = 0$ which yields the solution $x = 2$.
3. Take the second derivative of your function, $f''(x)$.
   *Example:* Calculate $f''(x) = \frac{d}{dx} f'(x) = \frac{d}{dx}(2x-4) = 2$.
4. If $f''(x) > 0$, you have found a local minimum. If $f''(x) < 0$, you have found a local maximum. If $f''(x) = 0$, your point is neither a local minima or maxima.
   *Example:* $f''(x) = 2 > 0$, so this is a local minimum.
5. Plug in the value or values you found in step 2 into your function, and check endpoints if applicable.
   *Example* The function $f(x) = (x-2)^2 + 3$ has a local minimum at $f(2) = 3$. As we do not have endpoints, this is a global minimum.

# An Example with multiple variables

What is the local optimum of

$$f(\lambda, k) = k \log(\lambda) - \lambda - \log(k!)$$

with respect to $\lambda$, over the range $0 \leq \lambda$ and $0 \leq k$? Is it a min or max?

# Maximization (with Multiple Variables): Practice!

1. Maximize $f(x, \mu, \sigma) = -\frac{(x-\mu)^2}{\sigma} - \log(\sigma)$ with respect to $\mu$
2. Maximize $f(x, \mu, \sigma) = -\frac{(x-\mu)^2}{\sigma} - \log(\sigma)$ with respect to $\sigma$, over the range $\sigma > 0$.
3. Maximize $f(p, n, k) = k \log(p) + (n - k) \log(1 - p)$ with respect to $p$, over the range $0 \leq p \leq 1$ and $0 \leq k \leq n$.

These are questions 1 and 2 under "Practice: Maximization and Minimization" on the hand-out. For these practice questions, checking the second derivative is optional!

# Maximization (with Multiple Variables) Solutions

1. Maximize $f(x, \mu, \sigma) = -\frac{(x-\mu)^2}{\sigma} - \log(\sigma)$ with respect to $\mu$ and $\sigma$, over the range $\sigma > 0$.

1.

$$\frac{\partial}{\partial \mu} f(x, \mu, \sigma) = \frac{2(x-\mu)}{\sigma} = 0$$

$$2(x-\mu) = 0$$

$$\mu^* = x$$

2.

$$\frac{\partial}{\partial \sigma} f(x, \mu, \sigma) = (x-\mu)^2 \sigma^{-2} - \sigma^{-1} = 0$$

$$\sigma^* = (x-\mu)^2$$

Maximize $f(p, n, k) = k \log(p) + (n - k) \log(1 - p)$ with respect to $p$, over the range $0 \leq p \leq 1$ and $0 \leq k \leq n$.

$$\frac{\partial}{\partial p} f(p, n, k) =$$

3.

$$\frac{k}{p} - \frac{(n - k)}{(1 - p)} = 0$$

$$\frac{k}{p} = \frac{n - k}{1 - p}$$

$$k - kp = pn - pk$$

$$p^* = \frac{k}{n}$$

# Calculus Part II: Integration Learning Objectives

By the end of this lesson, you will be able to:

1. Understand the intuition of the integral as the "reverse derivative" through the fundamental theorem of calculus
2. Understand the basic rules of indefinite integrals
3. Be able to evaluate definite integrals with finite and infinite limits
4. Be able to integrate using u-substitution

The most important use of integration statistics is using integration to find the expectation (mean) of random variables. We will probably not get to u-substitution, which will probably only come up once or twice in your stats class although I have added it to the cheat sheet as a reference.

Definition of the integral:

$$\int_a^b f(x)\,dx = \text{ area under the curve of x between a and b}$$

Example, setting $f(x) = x$ what is the value of

$$\int_0^b x\,dx$$

How about

$$\int_a^b x\,dx?$$

Definition of the integral:

$$\int_a^b f(x)\,dx = \text{ area under the curve of x between a and b}$$

Example, setting $f(x) = x$ we have

$$\int_0^b x\,dx = \frac{1}{2}b^2$$

and

$$\int_a^b x\,dx = \frac{1}{2}b^2 - \frac{1}{2}a^2$$

How would we do this in general?

$$\int_a^b f(x)\,dx = \lim_{\Delta x \to 0} \sum_x \Delta x \cdot f(x + \Delta x)$$

Recall that the definition of the derivative is

$$\lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

In a very formal we can see the similarities between the two, although in the integral we multiply by $\Delta x$ and in the derivative we divide by $\Delta x$.

Fundamental theorem of calculus:

$$\int_a^b f'(x)\,dx = F(b) - F(a)$$

Demonstration with $\int_a^b x\,dx$.

Since integration is simply the opposite of differentiation, in that if $f'(x)$ is the derivative of of $f(x)$, then $\int f'(x)dx = f(x) + C$ is the indefinite integral. Knowing this you can always check your work when integrating!

| Rule | Function | Integral |
|---|---|---|
| Constant Rule | $\int a\,dx$ | $ax + C$ |
| Linearity Rule | $\int af(x) + bf(x)\,dx$ | $a\int f(x)\,dx + b\int g(x)\,dx$ |
| Power Rule | $\int x^a\,dx$ | $\frac{1}{a+1}x^{a+1} + C$ |
| Exponential Rule | $\int e^x\,dx$ | $e^x + C$ |
| Reciprocal Rule | $\int \frac{1}{x}\,dx$ | $\log(x) + C$ |

# Integration Rules

Constant Rule $\int a\,dx = ax + C$

Example $\int 4\,dx = 4x + C$

# Integration Rules: Linearity Rule

$$\int af(x) + bf(x)\,dx = a\int f(x)\,dx + b\int g(x)\,dx$$

Example: $\int (ax + b)\,dx = \frac{a}{2}x^2 + bx + C$

# Integration Rules: Power Rule

$$\int x^a \, dx = \frac{1}{a+1} x^{a+1} + C$$

Example: $\int x^2 \, dx = \frac{1}{3} x^3 + C$

# Integration Rules: Exponential Rule

$\int e^x \, dx = e^x + C$

Example: $\int e^x + 3 \, dx = e^x + 3x + C$

# Integration Rules: Reciprocal Rule

$\int \frac{1}{x}\, dx = \log(x) + C$

Example: $\int \frac{2}{x}\, dx = 2\log(x) + C$

# Integration with Bounds

$$\int_0^2 x^2 + 1 \, dx$$

# Another Example

$$\int_1^2 (\ln(2x) + 1)\, dx$$

# Integration Rules: Practice!

1. $\int_0^1 e^{-x+2}\,dx$
2. $\int_1^2 x^{-2}\,dx$
3. $\int_0^1 \int_0^1 x^2 y^2\,dx\,dy$

# Integration Rules: Practice Solutions

1. $\int_0^1 e^{-x+2}\,dx$
2. $\int_1^2 x^{-2}\,dx$
3. $\int_0^1 \int_0^1 x^2 y^2\,dx\,dy$

# U-Substitution

By the chain rule we know that $\int f'(g(x))g'(x)\,dx = f(g(x)) + C$. In u-substitution, we let $u = g(x)$ and $du = g'(x)dx$. Then our integral becomes

$$\int f'(g(x))g'(x)\,dx = \int f'(u)\,du = f(u)$$

*Example:* In $\int 2xe^{x^2}dx$, we can let $u = x^2$ and $du = 2x\,dx$ in order to get

$$\int 2xe^{x^2}\,dx = \int e^u\,du = e^u = e^{x^2}.$$

# U-Substitution: Examples

1. $\int \lambda e^{-\lambda x}\, dx$
2. $\int (2x+2)e^{x^2+2x+x}\, dx$

# U-Substitution: Practice!

1. $\int \frac{(2x+2)}{(x^2+2x+3)^2}\, dx$
2. $\int xe^{-x^2/\sigma}\, dx$

# U-Substitution: Practice Solution

1. $\int \frac{(2x+2)}{(x^2+2x+3)^2}\,dx$

# U-Substitution: Practice Solution

2. $\int xe^{-x^2/\sigma}\, dx$

## Key terms in statistical inference:

- *Distribution:* Distributions describe frequencies through probability density or probability mass functions. Usually are characterized by one or more parameters. Example: $N(\mu, \sigma^2)$, Poisson($\lambda$), Uniform($a, b$), et cetera.
- *Parameter:* These are values that characterize a distribution, such as $\mu, \sigma^2, \lambda$, et cetera. They are often referred to using the symbol $\theta$.
- *Random variables:* Random variables are written capital letters. Random variables are given distributions, which we write as $X_i \sim N(0, 1)$. Random variables are not numbers or values, but can take on values for example $X_i = x_i$.
- *Data values:* Data values are written as lowercase letters, such as $x_i$. As above, when data values are are observations of random variables we write $X_i = x_i$.

## Frequentist vs. Bayesian Inference

| **Frequentist** | **Bayesian** |
|---|---|
| Data is viewed as a random sample from a distribution and the parameters are fixed | Data is fixed and inference is done parameters which are viewed as random |
| Point estimates are obtained through maximum likelihood estimation (MLE) | Estimates obtained through posterior distributions which contain information on mean, median, variance, et cetera |
| Uncertainty quantified through confidence intervals: if we were to re-sample our data many times, 95% of the time our sample will contain the true parameter value | Uncertainty quantified through credible intervals: there is a 95% chance that the parameter is within our interval |
| No way to incorporate prior inforation | Prior inforation can be included, however when prior information is absent it may be difficult to derive "un-informative" prior |
| Less computationally intensive | Usually more computationally intensive when the posterior distribution does not have a closed form |

## Steps to Frequentist Inference

1. Define your data, its corresponding distribution, and the unknown parameters
   *Example:* $X_1, \ldots, X_n \sim N(\mu, 1)$ with unknown parameter $\mu$
2. Write out the likelihood function
   *Example:* $L(\mu; X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2}\right)$
3. Take the log of the likelihood
   *Example:* $\ell(\mu; X) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}$
4. Find the value of the parameter that maximizes the likelihood through the differentiation or through numerical optimization
   *Example:* $\frac{\partial}{\partial \mu}\ell(\mu; X) = 0$ yields $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$
5. Find a confidence interval by deriving the distribution of your estimator
   *Example:* $\hat{\mu} \sim N\left(\overline{x}, \frac{1}{n}\right)$ so $CI_\alpha[\hat{\mu}] = \left(\overline{x} - z_{\alpha/2}\frac{1}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{1}{\sqrt{n}}\right)$

## Steps to Bayesian Inference

1. Define your data, your parameter, and the prior for your parameter
   *Example:* $X_1, \ldots, X_n \sim N(\mu, 1)$ and let the prior be $\mu \sim N(0, \sigma_0^2)$
2. Use Bayes rule to write the posterior of the parameter given the data
   *Example:* $P(\mu|X) \propto P(X|\mu)P(\mu) = \frac{1}{\sqrt{2\pi}^n} \exp\left(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma_0^2}\right)$
3. Re-arrange to see if you can get the posterior distribution to be in a familiar form. If not, you may need to use more advanced techniques such as MCMC
   *Example* Through several algebraic steps, we can see that

$$\mu|X \sim N\left(\frac{n\overline{x}}{n+\sigma_0^{-2}}, \frac{1}{n+\sigma_0^{-2}}\right)$$

   .
4. Calculate the postetrior mean and credible interval from the posterior distribution
   *Example:* $\hat{\mu} = \frac{\overline{x}}{1+\sigma_0^2}$ and $CI_\alpha[\hat{\mu}] = \left(\frac{n\overline{x}}{n+\sigma_0^{-2}} - z_{\alpha/2}\sqrt{\frac{1}{n+\sigma_0^{-2}}}, \frac{n\overline{x}}{n+\sigma_0^{-2}} + z_{\alpha/2}\sqrt{\frac{1}{n+\sigma_0^{-2}}}\right)$

## Common Probability Distributions

| Distribution | Parameters | PMF/PDF | Mean | Variance |
|---|---|---|---|---|
| Bernoulli | $p > 0$ | $P(X = 1) = p$ | $p$ | $p(1-p)$ |
| Binomial | $n > 0, 0 < p < 1$ | $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ | $np$ | $np(1-p)$ |
| Poisson | $\lambda > 0$ | $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ | $\lambda$ | $\lambda$ |
| Uniform | $a < b$ | $\frac{1}{b-a}$ for $a \leq x \leq b$ | $\frac{1}{2}(a+b)$ | $\frac{1}{12}(b-a)^2$ |
| Normal | $-\infty < \mu < \infty, \sigma^2 > 0$ | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu$ | $\sigma^2$ |
| Exponential | $\lambda > 0$ | $\lambda e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |

## Intuition for the Most Common Distributions

- *Binomial:* If $X \sim Binomial(n, p)$, then $X$ is the number of heads recorded from $n$ (biased) coin flips. Note that $Bernoulli(p) = Binomial(1, p)$, and that if $X_1, \ldots, X_n \sim Bernoulli(p)$ then $\sum_{i=1}^n X_i \sim Binomial(n, p)$.
- *Normal:* By the central limit theorem, or CLT, the average of $n$ random variables $X_1, \ldots, X_n$ has the normal distribution; formally $\frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \sigma^2)$ (as long as $n$ is large enough! Usually $n > 30$ is considered fine).
- *Poisson:* Say a call center receives an average of $\lambda$ calls per hour, and each call is randomly timed and independent. Then $X \sim Poisson(\lambda)$ is the random variable for the number of calls received within one hour.
- *Exponential:* At the call center mentioned above, the amount of time between calls is distributed $X \sim Exponential(\lambda)$.

# Linear Algebra I & II

Jiayi Li

Department of Statistics, UCLA

September 25, 2021

# Acknowledgement

These slides are modified from an original deck of slides created by Dr. Zhanhao Peng,
who recently received his Ph.D. in Statistics from UCLA.

# Operations and Properties

# Basic Notations

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with $m$ rows and $n$ columns.
- By $x \in \mathbb{R}^n$, we denote a vector with $n$ entries. An $n$-dimensional vector is often thought of as a $n \times 1$ matrix, known as a column vector. If we want to explicitly represent a row vector, we typically write $x^T$.
- The $i$-th element of vector $x$ is denoted as $x_i$, $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$
- Use notation $a_{ij}$ to denote the entry of matrix $A$ in the $i$-th row and $j$-th column.

# Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$,

$$C = AB \in \mathbb{R}^{m \times p}, \; C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

Given two vectors $x, y \in \mathbb{R}^n$, the inner product (dot product) is a real number given by

$$x^T y \in \mathbb{R} = \sum_{i=1}^{n} x_i y_i$$

Given two vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, the outer product is given by

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

# Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$,

$$AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | & & | \\ b_1 & \cdots & b_p \\ | & & | \end{bmatrix}$$

$$= \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T$$

# Matrix Multiplication

Associative:

$$(AB)C = A(BC)$$

$$((AB)C)_{ij} = \sum_{k=1}^{p}(AB)_{ik}C_{kj} = \sum_{k=1}^{p}(\sum_{l=1}^{n}A_{il}B_{lk})C_{kj}$$

$$= \sum_{k=1}^{p}(\sum_{l=1}^{n}A_{il}B_{lk}C_{kj}) = \sum_{l=1}^{n}(\sum_{k=1}^{p}A_{il}B_{lk}C_{kj}) = \sum_{l=1}^{n}A_{il}(\sum_{k=1}^{p}B_{lk}C_{kj})$$

$$= \sum_{l=1}^{n}A_{il}(BC)_{lj} = (A(BC))_{ij}$$

Distributive:

$$A(B + C) = AB + AC$$

Not commutative:

$$AB \neq BA \text{ in general}$$

# Exercise 1

1.1 Find the inner product (if it has) and the outer product of the vectors

(a) $A = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $B = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ (b) $A = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $B = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$,

(c) $A = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$, $B = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$

1.2 Find the product of the matrices

(a) $A = \begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}$, $AB =?$

(b) $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 4 & 1 \end{bmatrix}$, $AB =?$

(c) $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 4 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$, $ABC =?$

# Exercise 1

1.3 True or False

(a) For $A, B \in \mathbb{R}^{n\times n}$, $A^2 - B^2 = (A+B)(A-B)$

(b) For $A, B \in \mathbb{R}^{n\times n}$, $(AB)^2 = A^2B^2$

1.4 Find the product of the matrices

(a) $A = \begin{bmatrix} -4 & -y \\ -2x & -4 \end{bmatrix}$, $B = \begin{bmatrix} -4x & 0 \\ 2y & -5 \end{bmatrix}$, $AB = ?$

(b) $A = \begin{bmatrix} -4 & -y \\ -2x & -4 \\ z & 1 \end{bmatrix}$, $B = \begin{bmatrix} -4x & 0 & 3z \\ 2y & -5 & -z \end{bmatrix}$, $AB = ?$

# Identity and Diagonal Matrices

The identity matrix $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the digonal and zeros everywhere else.

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

For all $A \in \mathbb{R}^{m \times n}$,

$$AI = A = IA$$

(Usually the dimension of $I$ is not specified)
A diagonal matrix is a matrix where all non-digonal elements are 0, typically denoted as $D = \text{diag}(d_1, d_2, ..., d_n)$.
Clearly, $I = \text{diag}(1, 1, ..., 1)$.

# Matrix Transpose

The transpose of a matrix results from flipping the rows and columns. Given $A \in \mathbb{R}^{m \times n}$, its transpose $A^T \in \mathbb{R}^{n \times m}$, where

$$(A^T)_{ij} = A_{ji}$$

The transpose of a row (column) vector is a column (row) vector.

Tranposes have the following properties,

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

A square matrix is symmetric if $A = A^T$.

# Trace

The trace of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\text{tr}(A)$, is the sum of diagonal elements in the matrix:

$$\text{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

The trace has the following properties,

- For $A \in \mathbb{R}^{n \times n}$, $\text{tr}(A) = \text{tr}(A^T)$
- For $A, B \in \mathbb{R}^{n \times n}$, $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$, $\text{tr}(tA) = t \cdot \text{tr}(A)$
- For $A, B$ such that $AB$ is square, $\text{tr}(AB) = \text{tr}(BA)$
- For $A, B, C$ such that $ABC$ is square, $\text{tr}(ABC) = \text{tr}(BAC) = \text{tr}(CAB)$, and so on for the product of more matrices.

# Norms

Formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity)
2. $f(x) = 0$ iff $x = 0$ (definiteness)
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t|f(x)$ (homogeneity)
4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality)

A norm of a vector $x$ is informally a measure of "length" of the vector. e.g. the commonly used Euclidean or $l_2$ norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

note that $\|x\|_2^2 = x^T x$.

# Norms

By definition, $l_p$, $(p \geq 1)$ norm:

$$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$$

$l_1$ norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
$l_\infty$ norm: $\|x\|_\infty = \max_i |x_i|$
Norms can also be defined for matrices, such as the Frobenius norm,

$$\|x\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\mathrm{tr}(A^T A)}$$

# Exercise 2

2.1 For $v \in \mathbb{R}^n$, prove that $vv^T$ is a symmetric matrix.

2.2 For $v, w \in \mathbb{R}^n$, prove that $\text{tr}(vw^T) = v^T w$.

2.3 Calculate the following expressions, using the following matrices: $A = \begin{bmatrix} 2 & 3 \\ 5 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}$, $v = \begin{bmatrix} 2 \\ -4 \end{bmatrix}$

(a) $AB^T + vv^T$

(b) $Av - 2v$

(c) $v^T B$

(d) $v^T v + v^T BA^T v$

# Exercise 2

2.4 Let $a$ and $b$ be vectors in $\mathbb{R}^n$ such that

$$\|a\| = \|b\| = 1$$

and the inner product

$$a^T b = -\frac{1}{2}$$

determine $\|a - b\|$.

# Rank

A set of non-zero vectors $\{x_1, x_2, \cdots, x_n\} \subset \mathbb{R}^m$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set can be represented as a linear combination of the remaining vectors, then the vectors are said to be (linearly) dependent. That is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$, then we say that the vectors $x_1, \cdots, x_n$ are linearly dependent, otherwise they are independent. E.g.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \; x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix}, \; x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

# Rank

The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of $A$ that constitute a linearly independent set. In the same way, the row rank is the largest number of rows of $A$ that constitute a linearly independent set. For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A, and so both quantities are referred to collectively as the rank of A, denoted as rank(A). The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, A is said to be full rank.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

# Inverse

The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted $A^{-1}$, and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

Non-square matrices do not have inverses by definition. For some square matrices $A$, $A^{-1}$ may not exist. In particular, we say that $A$ is invertible or non-singular if $A^{-1}$ exists and non-invertible or singular otherwise.
In order for a square matrix $A$ to have an inverse, then $A$ must be full rank.
The following are properties of the inverse; all assume that $A, B \in \mathbb{R}^{n \times n}$ are non-singular:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1} = A^{-T}$

# Orthogonal Matrices

Two vectors $x, y \in \mathbb{R}^n$ are orthogonal if $x^T y = 0$. A vector $x \in \mathbb{R}^n$ is normalized if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being orthonormal), that is

$$U^T U = I = U U^T$$

In other words, the inverse of an orthogonal matrix is its transpose.
Operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.

$$\|Ux\| = \|x\|_2$$

for any orthogonal $U \in \mathbb{R}^{n \times n}$ , $x \in \mathbb{R}^n$

# Determinant

The determinant of a square matrix $A \in \mathbb{R}^{n\times n}$, is a function $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$, and is denoted $|A|$ or $\det A$.
Several properties of determinant:

- For $A \in \mathbb{R}^{n\times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n\times n}$ $|AB| = |A||B|$.
- For $A \in \mathbb{R}^{n\times n}$, $|A| = 0$ iff A is singular.
- For $A \in \mathbb{R}^{n\times n}$ and $A$ non-singular, $|A^{-1}| = 1/|A|$.

Define for $A \in \mathbb{R}^{n\times n}$, $A_{\setminus i,\setminus j} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the matrix that results from deleting the $i$-th row and $j$-th column from $A$, then

$$|A| = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}|A_{\setminus i,\setminus j}| = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}|A_{\setminus i,\setminus j}|$$

# Determinant

$$|a_{11}| = a_{11}$$

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}$$

$$- a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$$

# Exercise 3

3.1 Determine the rank of $A$ and $A^T A$ when

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

3.2 Let $v$ be a nonzero vector in $\mathbb{R}^n$. Then the dot product $v \cdot v = v^T v \neq 0$.
Set $a := \frac{2}{v^T v}$ and define the $n \times n$ matrix $A$ by

$$A = I - avv^T$$

where $I$ is the $n \times n$ identity matrix.
Prove that $A$ is a symmetric matrix and orthogonal.

3.3 Prove that $(A^T)^{-1} = (A^{-1})^T$.

# Exercise 3

3.4 Find the inverse matrix of the matrix

$$A = \begin{bmatrix} \frac{2}{7} & \frac{3}{7} & \frac{6}{7} \\ \frac{6}{7} & \frac{2}{7} & -\frac{3}{7} \\ -\frac{3}{7} & \frac{6}{7} & -\frac{2}{7} \end{bmatrix}$$

(hint: show that $A$ is orthogonal)

3.5 Let $A, B, C$ be the following matrices,

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{bmatrix}, C = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 5 & 6 \\ 3 & 0 & 1 \end{bmatrix}$$

Compute and simplify the following expression,

$$(A^T - B)^T + C(B^{-1}C)^{-1}$$

# Quadratic Forms and Positive Semidefinite Matrices

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a quadratic form. Written explicitly, we see that

$$x^T Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i (\sum_{j=1}^{n} A_{ij} x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

Note that

$$x^T Ax = (x^T Ax)^T = x^T A^T x = \frac{1}{2} x^T (A + A^T) x$$

The matrices appearing in a quadratic form are symmetric.

# Quadratic Forms and Positive Semidefinite Matrices

We give the following definitions:

- A symmetric matrix $A \in \mathbb{S}^n$ is positive definite (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$. This is usually denoted $A \succ 0$ (or just $A > 0$), and the set of all positive definite matrices is denoted $\mathbb{S}^n_{++}$.
- A symmetric matrix $A \in \mathbb{S}^n$ is positive semidefinite (PSD) if for all vectors $x \in \mathbb{R}^n$, $x^T Ax \leq 0$. This is usually denoted $A \succeq 0$ (or just $A \geq 0$), and the set of all positive definite matrices is denoted $\mathbb{S}^n_{+}$.
- Likewise, we define negative definite (ND) and negative semidefinite (NSD).

One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible.

# Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $A$ and $x \in \mathbb{C}^n$ is the corresponding eigenvector if

$$Ax = \lambda x, \ x \neq 0$$

note that for any eigenvector $x \in \mathbb{C}^n$, and scalar $t \in \mathbb{C}$, $A(tx) = tAx = t\lambda x = \lambda(tx)$, so $tx$ is also an eigenvector. For this reason we usually assume that the eigenvector is normalized to have length 1.
We can rewrite the equation above to state that $(\lambda, x)$ is an eigenvalue-eigenvector pair of $A$ if

$$(\lambda I - A)x = 0, \ x \neq 0$$

But $(\lambda I - A)x = 0$ has a non-zero solution to $x$ iff $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e. $|\lambda I - A| = 0$.

# Eigenvalues and Eigenvectors

We expand this expression into a polynomial in $\lambda$, where $\lambda$ will have maximum degree $n$. We then find the $n$ (possibly complex) roots of this polynomial to find the $n$ eigenvalues $\lambda_1, \cdots, \lambda_n$. To find the eigenvector corresponding to the eigenvalue $\lambda_i$, we simply solve the linear equation $(\lambda_i I - A)x = 0$ and get $x_1, \cdots, x_n$.

- The trace of $A$ is equal to the sum of its eigenvalues, $\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i$.
- The determinant of $A$ is equal to the product of its eigenvalues, $|A| = \prod_{i=1}^{n} \lambda_i$.
- The rank of $A$ is equal to the number of non-zero eigenvalues of $A$.
- If $A$ is non-singular then $1/\lambda_i$ is an eigenvalue of $A^{-1}$ with associated eigenvector $x_i$, i.e., $A^{-1}x_i = (1/\lambda_i)x_i$.

We can write all the eigenvector equations simultaneously as

$$AX = X\Lambda, \; X = (x_1, \cdots, x_n), \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$$

# Eigenvalues and Eigenvectors

Two remarkable properties come about when we look at the eigenvalues and eigenvectors of a symmetric matrix $A \in \mathbb{S}^n$. First, all the eigenvalues of $A$ are real. Secondly, the eigenvectors of $A$ are orthonormal.
We can therefore represent $A$ as $A = U\Lambda U^T$, using this, we can show that the definiteness of a matrix depends entirely on the sign of its eigenvalues. Suppose $A \in \mathbb{S}^n = U\Lambda U^T$. Then

$$x^T Ax = x^T U\Lambda U^T x = y^T \Lambda y = \sum_{i=1} \lambda_i y_i^2$$

where $y = U^T x$ (since $U$ is full rank, any vector $y \in \mathbb{R}^n$ can be represented in this form).

# Matrix Calculus

# Gradient

Suppose that $f : \mathbb{R}_{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix $A$ of size $m \times n$ and returns a real value. Then the gradient of $f$ (with respect to $A \in \mathbb{R}_{m\times n}$) is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}_{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

Note that the size of $\nabla_A f(A)$ is always the same as the size of $A$. We have

$$\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$$

$$\nabla_x(tf(x)) = t\nabla_x f(x) \ (t \in \mathbb{R})$$

# Gradient

Let $f : \mathbb{R}_m \to \mathbb{R}$ be the function defined by $f(z) = z^T z$, such that $\nabla_z f(z) = 2z$, now consider the interpretation of $\nabla f(Ax)$.

1. In the first interpretation, recall that $\nabla_z f(z) = 2z$. Here, we interpret $\nabla f(Ax)$ as evaluating the gradient at the point $Ax$, hence,

$$\nabla f(Ax) = 2(Ax) = 2Ax \in \mathbb{R}^m$$

2. In the second interpretation, we consider the quantity $f(Ax)$ as a function of the input variables $x$. More formally, let $g(x) = f(Ax)$. Then in this interpretation,

$$\nabla f(Ax) = \nabla_x g(x) \in \mathbb{R}^n$$

Keep the notation clear is extremely important!

# Hessian

Suppose that $f : \mathbb{R}_m \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R} \in n$ and returns a real number. Then the Hessian matrix with respect to $x$, written $\nabla_x^2 f(x)$ or simply as $H$ is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}_{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n x_1} & \frac{\partial^2 f(x)}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i x_j} = \frac{\partial^2 f(x)}{\partial x_j x_i}$$

# Gradient and Hessian of Quadratic and Linear Functions

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k$$

From this we can easily see that $\nabla_x b^T x = b$.

# Gradient and Hessian of Quadratic and Linear Functions

For $x \in \mathbb{R}^n$, now consider the quadratic function $f(x) = x^T Ax$ for $A \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

To take the partial derivative, we consider the terms including $x_k$ and $x_k^2$ factors separately,

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$= \frac{\partial}{\partial x_k} [\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2]$$

$$= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k = 2 \sum_{i=1}^{n} A_{ki} x_i$$

# Gradient and Hessian of Quadratic and Linear Functions

Note that the kth entry of $\nabla_x f(x)$ is just the inner product of the kth row of $A$ and $x$. Thus $\nabla_x x^T Ax = 2Ax$.
And the Hessian of the quadratic function $f(x) = x^T Ax$ is

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\partial}{\partial x_k}[\frac{\partial f(x)}{\partial x_l}] = \frac{\partial}{\partial x_k}[2\sum_{i=1}^{n} A_{ki}x_i] = 2A_{lk} = 2A_{kl}$$

Thus $\nabla_x^2 x^T Ax = 2A$ ($A$ is symmetric)

# Probability I & II

Jiayi Li

Department of Statistics, UCLA

September 26, 2021

# Acknowledgement

These slides are modified from an original deck of slides created by Dr. Zhanhao Peng,
who recently received his Ph.D. in Statistics from UCLA.

# Elements of probability

In order to define a probability on a set we need a few basic elements,

- Sample sapce $\Omega$: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.
- Set of events (or event space) $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called events) are subsets of $\Omega$.
- Probability measure: A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties (Axioms of Probability),
  1. $P(A) \geq 0$, for all $A \in \mathcal{F}$
  2. $P(\Omega) = 1$
  3. If $A_1, A_2, \cdots$ are disjoint events (i.e. $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cap_i A_i) = \sum_i P(A_i)$$

# Elements of probability

Example: Consider the event of tossing a six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. An event space $\mathcal{F}$ can be the set of all subsets of $\Omega$. One valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where $i$ is the number of elements of that set; for example

$$P(\{1, 2, 3, 4\}) = \frac{4}{6}$$

$$P(\{1, 2, 3\}) = \frac{3}{6}$$

# Elements of probability

Properties of probability:

- If $A \subseteq B \rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min(P(A), P(B))$
- (Union Bound) $P(A \cap B) \leq P(A) + P(B)$
- $P(\Omega \backslash A) = 1 - P(A)$
- (Law of Total Probability) If $A_1, \cdots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$

# Conditional probability and independence

Let $B$ be an event with non-zero probability. The conditional probability of any event $A$ given $B$ is defined as,

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

In other words, $P(A|B)$ is the probability measure of the event $A$ after observing the occurrence of event $B$. Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A|B) = P(A)$). Therefore, independence is equivalent to saying that observing $B$ does not have any effect on the probability of $A$.

By law of total probability, if $\cup_{i=1}^{k} B_i = \Omega$, then

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

# Exercise 1

1.1 Two dice are rolled, find the probability that the sum is (a) 1 (b) 4 (c) 13

1.2 A die is rolled and a coin is tossed, find the probability that the die shows an odd number and the coin shows a head.

1.3 You draw a card at random from a standard deck of 52 cards. Find each of the following conditional probabilities.
a. The card is a heart, given that it is black.
b. The card is black, given that it is a heart.
c. The card is an ace, given that it is black.
d. The card is a queen, given that it is a face card.

1.4 A tennis player $A$ has probability of $\frac{2}{3}$ of winning a set against player $B$. A match is won by the player who first wins three sets. Find the probability that $A$ wins the match.

1.5 Employment data at a large company reveal that 72% of the workers are married, that 44% are college graduates, and that half of the grads are married. What is the probability that a randomly chosen worker

a. is neither married nor a college graduate?

b. is married but not a college graduate?

c. is married or a college graduate?

1.6 Suppose that 23% of adults smoke cigarettes. It is known that 57% of smokers and 13% of nonsmokers develop a certain lung condition by age 60. a. Explain how these statistics indicate that lung condition and smoking are not independent. b. What is the probability that a randomly selected 60-year-old has this lung condition?

1.7 Two fair dice are rolled and the sum of the two numbers is observed. What is the probability that a sum of 2 appears before a sum of 6?

# Random Variables

Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space $\Omega$ are 10-length sequences of heads and tails. For example, we have $w_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.

In practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as random variables.

More formally, a random variable $X$ is a function $X : \Omega \to \mathbb{R}$. Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply $X$ (where the dependence on the random outcome $\omega$ is implied). We will denote the value that a random variable may take on using lower case letters $x$.

# Random Variables

In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses $\omega$. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a discrete random variable. Here, the probability of the set associated with a random variable $X$ taking on some specific value $k$ is

$$P(X = k) := P(\{\omega : X(\omega) = k\})$$

Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay. In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a continuous random variable. We denote the probability that $X$ takes on a value between two real constants $a$ and $b$ (where $a < b$) as

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\})$$

# Cumulative distribution functions

In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions (CDFs, PDFs, and PMFs).

A cumulative distribution function (CDF) is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as

$$F_X(x) := P(X \leq x)$$

CDF has the following properties:

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- $x \leq y \to F_X(x) \leq F_X(y)$

# Probability mass functions

When a random variable $X$ takes on a finite set of possible values (i.e., $X$ is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume. In particular, a probability mass function (PMF) is a function $p_X : \Omega \to \mathbb{R}$ such that

$$p_X(x) := P(X = x)$$

In the case of discrete random variable, we use the notation $\mathrm{Val}(X)$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $\mathrm{Val}(X) = \{0, 1, 2, \cdots, 10\}$.
PMF has the following properties:

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in \mathrm{Val}(X)} p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = P(X \in A)$

# Probability density functions

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the Probability Density Function or PDF as the derivative of the CDF, i.e.,

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

Note here, that the PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere). According to the properties of differentiation, for very small $\Delta x$,

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$$

# Probability density functions

Both CDFs and PDFs can be used for calculating the probabilities of different events.

The value of PDF at any given point $x$ is not the probability of that event, i.e., $f_X(x) \neq P(X = x)$.

PDF has the following properties:

- $f_X(x) \leq 0$
- $\int_{-\infty}^{\infty} f_X(x) = 1$
- $\int_{x \in A} f_X(x) dx = P(X \in A)$

# Expectation

Suppose that $X$ is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the expectation or expected value of $g(X)$ as

$$\mathbb{E}[g(X)] := \sum_{x \in \text{Val}(x)} g(x) p_X(x)$$

If $X$ is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as,

$$\mathbb{E}[g(X)] := \int_{\infty}^{\infty} g(x) f_X(x) dx$$

# Expectation

Properties of expectation

- $\mathbb{E}(a) = a$ for any constant $a \in \mathbb{R}$
- $\mathbb{E}(af(X)) = a\mathbb{E}(f(X))$ for any constant $a \in \mathbb{R}$
- $\mathbb{E}(f(X) + g(X)) = \mathbb{E}(f(X)) + \mathbb{E}(g(X))$

# Variation

The variance of a random variable $X$ is a measure of how concentrated the distribution of a random variable $X$ is around its mean. Formally, the variance of a random variable $X$ is defined as

$$\mathbb{V}(X) := \mathbb{E}(X - \mathbb{E}(X))^2$$

Using the properties of expectation, we have

$$\mathbb{V}(X) := \mathbb{E}(X-\mathbb{E}(X))^2 = \mathbb{E}(X^2-2\mathbb{E}(X)X+(\mathbb{E}(X))^2) = \mathbb{E}(X^2)-(\mathbb{E}$$

- $\mathbb{V}(a)$ for any constant $a \in \mathbb{R}$
- $\mathbb{V}(af(X)) = a^2\mathbb{V}(f(X))$ for any constant $a \in \mathbb{R}$

# Exercise 2

2.1 A worker can arrive to the workplace at any moment between 6 and 7 in the morning with the same likelihood.
Compute and plot the probability density function of the variable that measures the arrival time.
Compute and plot the distribution function.
Compute the probability of arriving before quarter past six and after half past six.
What is the expected arrival time?

2.2 Two balls are chosen randomly without replacement from an urn containing 8 white, 4 black, and 2 orange balls. Suppose that we win \$2 for each black ball selected and we lose \$1 for each white ball selected. We neither win nor we lose any money for selecting an orange ball. Let $X$ denote our winnings.

a. What is the expected value of $X$?

b. What is the variance of $X$?

c. Given that our winnings are negative, what is the probability that we lost exactly \$2?

# Exercise 2

2.3 Tay-Sachs disease is a rare fatal genetic disease occurring chiefly in children, especially of Jewish or Slavic extraction. Suppose that we limit ourselves to families which have (a) exactly three children, and (b) which have both parents carrying the Tay-Sachs disease. For such parents, each child has independent probability $\frac{1}{4}$ of getting the disease.
Write $X$ to be the random variable representing the number of children that will have the disease.
(a) Show that the probability distribution for $X$
(b) Find the expectation and variance of $X$
2.4 Find the probability that none of the three bulbs in a set of traffic lights will have to be replaced during the first 1200 hours of operation if the lifetime $X$ of a bulb (in thousands of hours) is a random variable with probability density function $f(x) = 6[0.25 - (x - 1.5)^2]$ when $1 \leq x \leq 2$ and $f(x) = 0$ otherwise. You should assume that the lifetimes of different bulbs are independent.

# Exercise 2

2.5 Let $X \sim U(0,1)$. Find $\mathbb{E}(X)$, $\mathbb{V}(X)$.

2.6 $f_X(x) = \lambda e^{-\lambda x}$. Find $\mathbb{E}(X)$, $\mathbb{V}(X)$.

# Joint and marginal distributions

Suppose that we have two random variables $X$ and $Y$. One way to work with these two random variables is to consider each of them separately. If we do that we will only need $F_X(x)$ and $F_Y(y)$. But if we want to know about the values that $X$ and $Y$ assume simultaneously during outcomes of a random experiment, we require a more complicated structure known as the joint cumulative distribution function of $X$ and $Y$ , defined by

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

The joint CDF $F_{XY}(x, y)$ and the joint distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)dy$$

$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y)dy$$

# Joint and marginal distributions

Here, we call $F_X(x)$ and $F_Y(y)$ the marginal cumulative distribution functions of $F_{XY}(x, y)$.

- $0 \leq F_{XY}(x, y) \leq 1$
- $\lim_{x,y\to\infty} F_{XY}(x, y) = 1$
- $\lim_{x,y\to-\infty} F_{XY}(x, y) = 0$

If $X$ and $Y$ are discrete random variables, then the joint probability mass function $p_{XY}[0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y)$$

Here, $0 \leq P_{XY}(x, y) \leq 1$ for all $x, y$, and $\sum_{x\in \mathrm{Val}(X)} \sum_{y\in \mathrm{Val}(Y)} P_{XY}(x, y) = 1$

# Joint and marginal distributions

$$p_X(x, y) = \sum_y p_{XY}(x, y)$$

and similarly for $p_Y(y)$. In this case, we refer to $p_X(x)$ as the marginal probability mass function of $X$.

Let $X$ and $Y$ be two continuous random variables with joint distribution function $F_{XY}$ . In the case that $F_{XY}(x, y)$ is everywhere differentiable in both $x$ and $y$, then we can define the joint probability density function,

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

# Joint and marginal distributions

Like in the single-dimensional case, $f_{XY}(x, y) \neq P(X = x, Y = y)$, but rather

$$\int \int_{x \in A} f_{XY}(x, y) dx dy = P((X, Y) \in A)$$

Note that the values of the probability density function $f_{XY}(x, y)$ are always non-negative, but they may be greater than 1.

Analagous to the discrete case, we define

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy$$

as the marginal probability density function (or marginal density) of $X$, and similarly for $f_Y(y)$.

# MAS Math Camp Day 2: Statistical Inference

Samuel Baugh

Department of Statistics
University of California, Los Angeles

September 26, 2021

# Lesson Plan

- Statistical inference: the basics
- Some common distributions
- Frequentist inference and maximum likelihood estimation
- Bayesian inference and posterior distributions

# Statistical Inference: The Basics

Statistical inference involves assuming that *data* are observations of *random variables*, and then using the values to infer *parameters* of a *distribution.*

Inference usually takes the form of an estimate and its uncertainty, which takes the form of a *confidence interval* or *credible interval.*

- *Distribution:* A form of a probability function. Distributions usually are characterized by one or more parameters. These are written for example as $N(\mu, \sigma^2)$, $Poisson(\lambda)$, $Uniform(a, b)$, et cetera.
- *Random variables:* Random variables are written *capital letters*. Random variables oftentimes follow distributions, which we write as $X_i \sim N(0, 1)$. Random variables are not numbers, but are characterized by their probability functions.
- *Data values:* Data values are lowercase letters, such as $x_i$. They are observations from a random variables. Data values are usually real numbers.
- *Parameter:* These are values that characterize a distribution, such as $\mu, \sigma^2, \ldots$. They are often generically referred to as $\theta$. In *frequentist inference*, they are unknown data values. In *bayesian inference*, they are random variables.

# Common Distributions

To Perform Inference you First Need a Distribution* Discrete:

- Bernoulli
- Binomial
- Poisson

Continuous:

- Uniform
- Normal
- Exponential

Your cheat-sheet for this lesson includes the parameters, PDF/PMF, mean, and variance of each of these distributions!

(*) Unless you are doing non-parametric inference

# Handy expectation and variance rules

Expectation of sum:

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

Expectation of scaled random variable:

$$E[aX] = aE[X]$$

Variance of sum of **independent** variables:

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

Variance of scaled random variable:

$$Var[aX_i] = a^2 Var[X_i]$$

# Bernoulli Distribution: The Simplest Distribution

Intuition: a single coin flip

PMF:

$$P(X = 1) = p, P(X = 0) = 1 - p$$

Mean:

$$E[X] = p$$

Variance:

$$Var[X] = p(1 - p)$$

# Binomial Distribution

Intuition: k heads out of n coin flips

Parameters:

$$n > 0, 0 < p < 1$$

PMF:

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

Mean:

$$E[X] = np$$

Variance:

$$Var[X] = np(1-p)$$

# Binomial Distribution

Note that $Bernoulli(p) = Binomial(1, p)$, and that if $X_1, \ldots, X_n \sim Bernoulli(p)$ then $\sum_{i=1}^{n} X_i \sim Binomial(n, p)$. We can use this to derive the mean and variance!

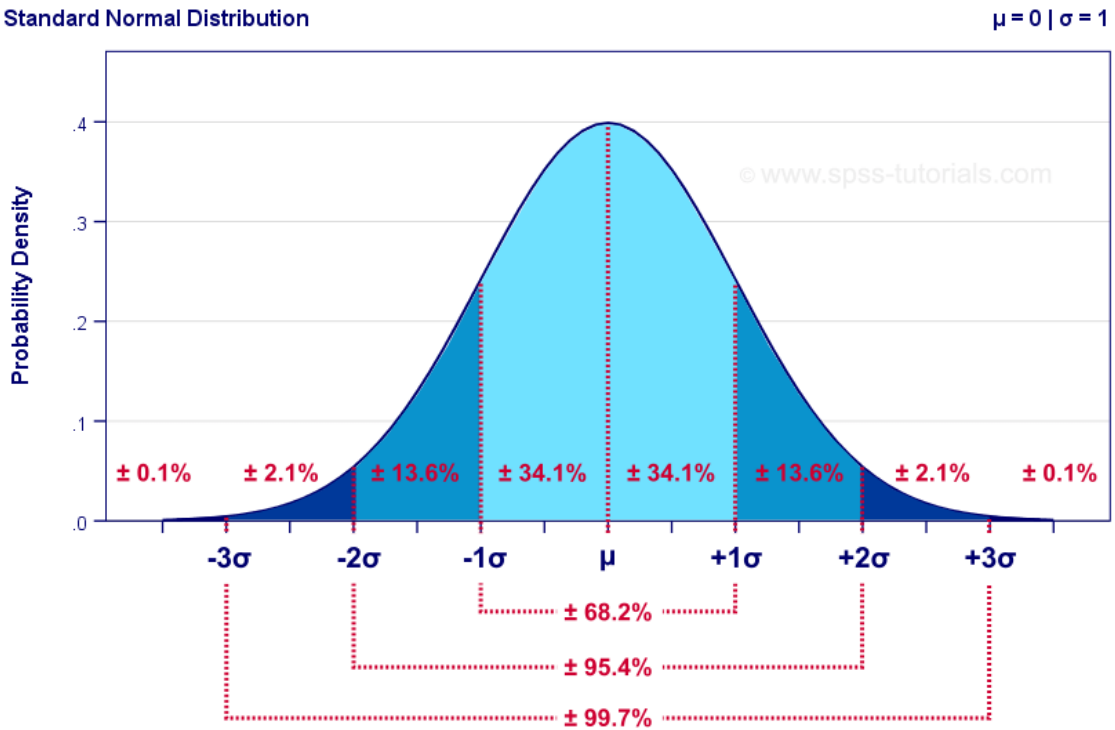
# Normal Distribution

Parameters:

$$-\infty < \mu < \infty, \sigma^2 > 0$$

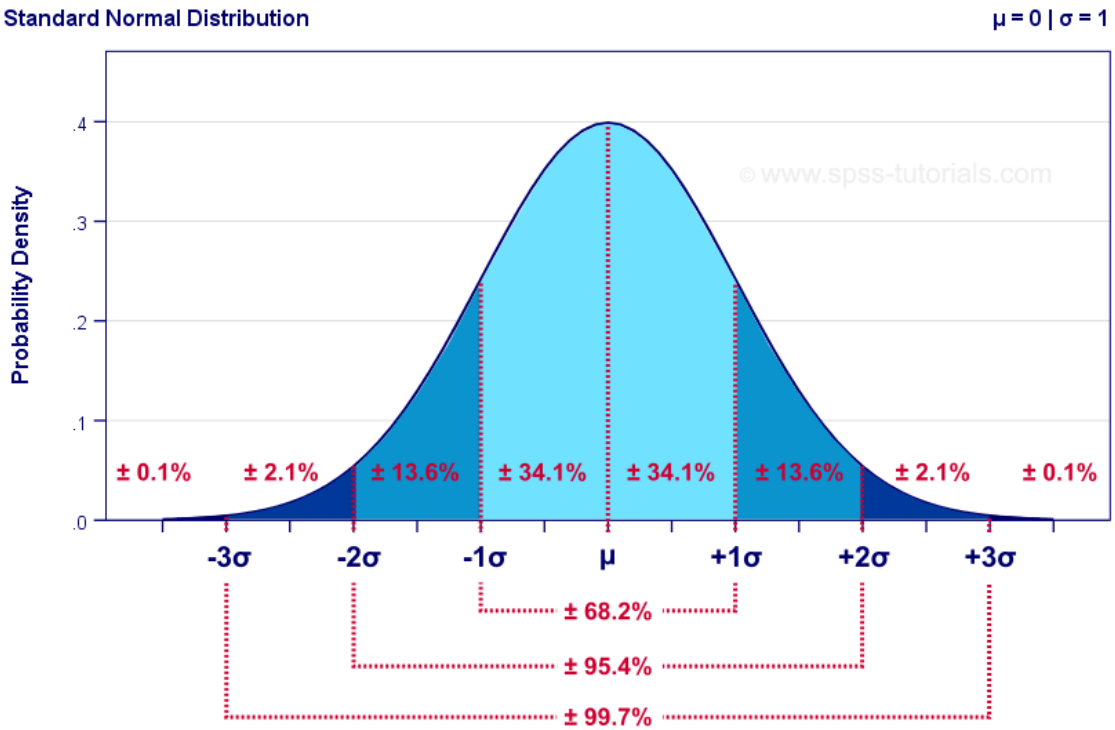
PDF:

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



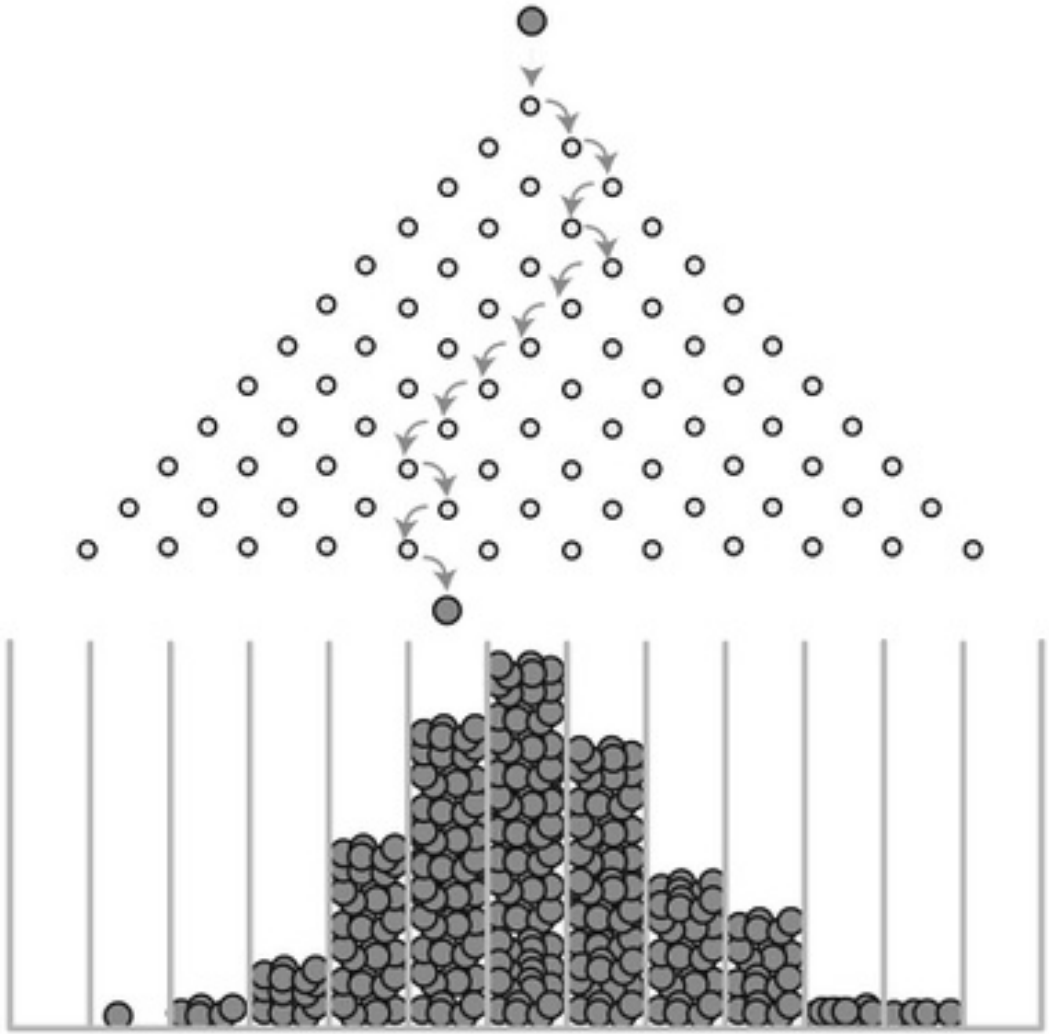
# Normal Distribution Example

Say $X \sim N(\mu, \sigma^2)$ where $\mu = 1$ and $\sigma^2 = 1$. What is the probability of observing a value less than -1? What is the probability of observing a value greater than 2?



# Normal Distribution: CLT Intuition

Central limit theorem (Galton board):

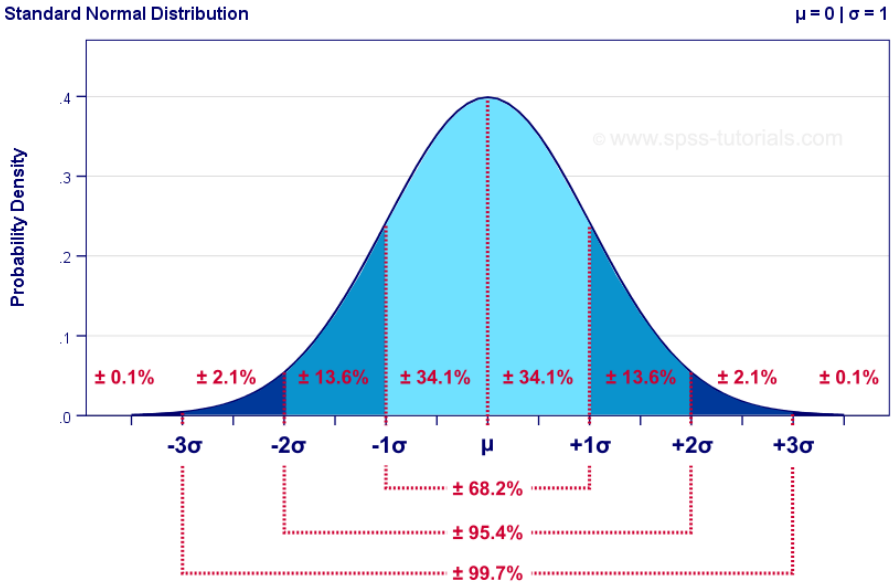


# Normal Distribution CLT Intuition

By the central limit theorem, or CLT, the average of $n$ random variables, as long as $n$ is sufficiently large (usually $n > 30$), the mean of $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, or $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$, has a normal distribution, no matter what the random variables are!

# Practice

1. We said that if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, then $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$, will have the normal distribution. What is the mean and variance of Z?
2. Say that you flip $n = 100$ fair coins. Use the central limit theorem to approximate the distribution of the sum of heads as normal. Using the diagram below, what is the approximate probability that you will get more than 55 heads? What about greater than 40 heads?



# Common Distributions: Poisson

*Intuition:* Say a call center receives an average of $\lambda$ calls per hour, and each call is randomly timed and independent. Then $X \sim Poisson(\lambda)$ is the random variable for the number of calls received within one hour.

Parameter:

$$\lambda > 0$$

PMF:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

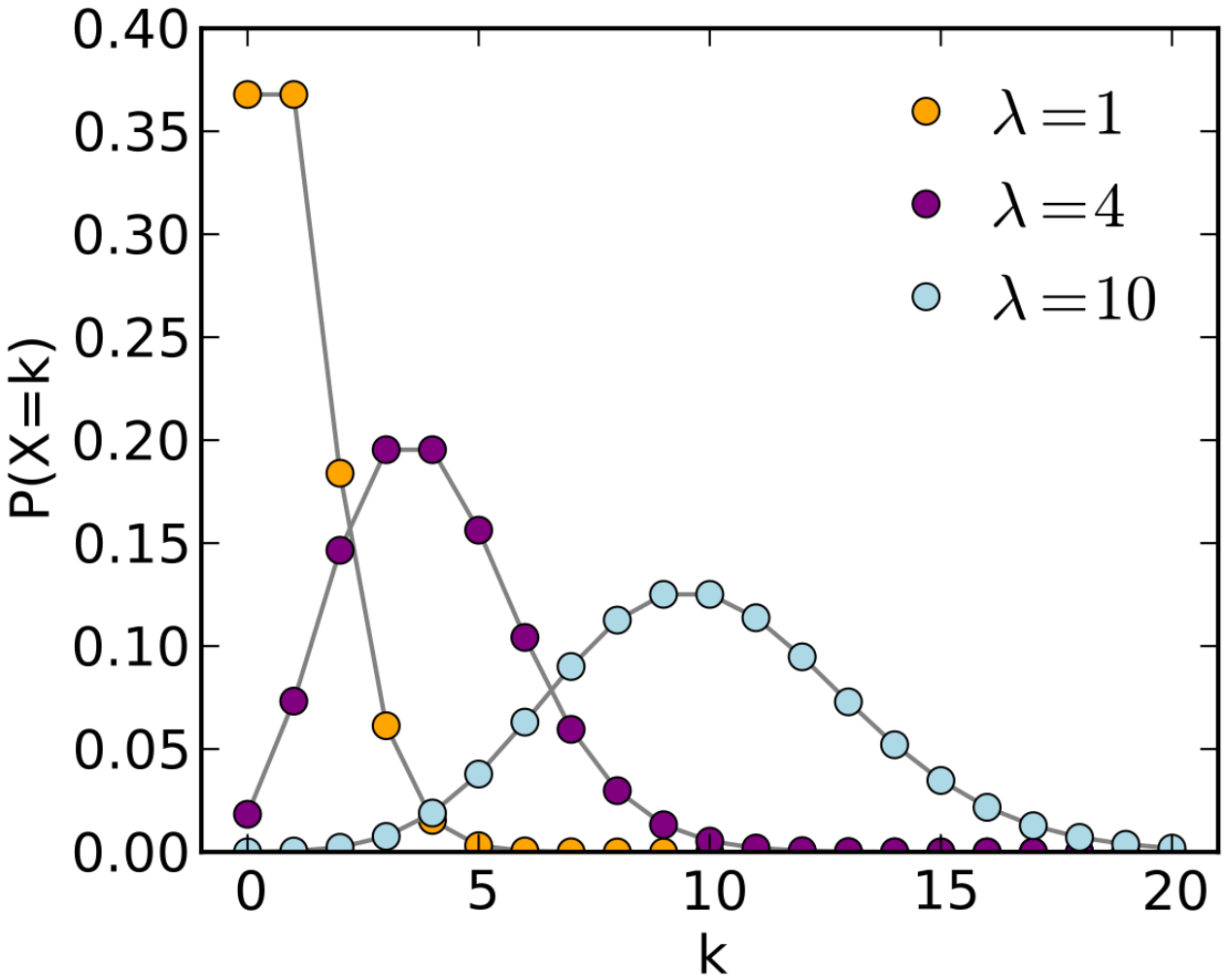
Mean:

$$E[X] = \lambda$$

Variance:

$$V[X] = \lambda$$

# Common Distributions: Poisson



# Common Distributions: Exponential

*Intuition:* Say a call center receives an average of $\lambda$ calls per hour, and each call is randomly timed and independent. Then $X \sim Exponential(\lambda)$ is the amount of time between calls.

Parameter:

$$\lambda > 0$$

PDF:

$$E[X] = \lambda e^{-\lambda x}$$

Mean:

$$E[X] = \frac{1}{\lambda}$$

Variance:

$$V[X] = \frac{1}{\lambda^2}$$

# Notation Reminder

- *Random variables:* Random variables are written in *capital letters*. Random variables follow distributions, which we write for example as $X \sim N(\mu, \sigma^2)$. Random variables are not numbers, but are characterized by their probability functions.
- *Data values:* Data values are lowercase letters. They are observations from a random variables. Data values are usually real numbers.
- *Parameter:* These are values that characterize a distribution. In *frequentist inference*, they are data values (which may be known or unknown). In *bayesian inference*, they are random variables.

A pdf written as $f_{X_i}(x_i; \mu)$ has a random variable, a data value, and a parameter. The form of $f$ here is dependent upon the distribution of $X_i$.

# Types of Inference

Inference set up: you have random variables $X_1, \ldots, X_n$, drawn from a distribution $F(\theta)$ with unknown parameters $\theta$, and observe data $x_1, \ldots, x_n$.

- **Frequentist Inference:** Parameters are data values which are unknown. They are not random, but since they are unobserved we can compute *estimators* which are random and may have distributions. Estimators are referred to as $\hat{\theta}$ (or $\hat{\mu}$, $\hat{p}$, $\hat{\sigma^2}$ $\ldots$)
- **Bayesian Inference:** Parameters are random variables which are unobserved. After observing data, we can compute the *posterior distribution*, which is written as $\theta|X$ if $X$ is the random variable corresponding to the data. Estimators may be computed from these distributions.

Similar goals, but different ways of viewing the unknown parameter and the resulting uncertainty!

| **Frequentist** | **Bayesian** |
|---|---|
| Data is viewed as a random sample from a distribution and the parameters are fixed | Data is fixed and inference is done parameters which are viewed as random |
| Point estimates are obtained through maximum likelihood estimation (MLE) | Estimates obtained through posterior distributions which contain information on mean, median, variance, et cetera |
| Uncertainty quantified through confidence intervals: if we were to resample our data many times, 95% of the time our sample will contain the true parameter value | Uncertainty quantified through credible intervals: there is a 95% chance that the parameter is within our interval |
| No way to incorporate prior inforation | Prior inforation can be included, however when prior information is absent it may be difficult to derive "uninformative" prior |
| Less computationally intensive | Usually more computationally intensive when the posterior distribution does not have a closed form |

# Likelihood Functions

Say $X_1, \ldots, X_n \sim N(\mu, 1)$ are random variables representing data points (such as heights) with values $X_1 = x_1, \ldots, X_n = x_n$. We know the random variables are independent and normally distributed, and that the variance is one, but we do not know the value of $\mu$. How should we estimate $\mu$? Demonstration on board. Recall that the normal PDF is:

$$f_{X_i}(x_i; \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2}}$$

# Deriving MLEs: Practice!

For these, you do not need to use the second derivative.

1. Say $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and $X_1 = x_1, \ldots, X_n = x_n$. Find the maximum likelihood estimator of $\sigma^2$. **Note:** For this, you should take the derivative $\frac{\partial}{\partial \sigma^2}$, **not** $\frac{\partial}{\partial \sigma}$
2. Say $X \sim \text{Poisson}(\lambda)$ and $X = k$. Find the maximum likelihood estimator of $\lambda$.
3. Say $X = k$ and $X \sim \text{Binomial}(n, p)$ for fixed integer $n$. Find the maximum likelihood estimator of $p$.

# Properties of Frequentist Estimators

Maximum likelihood estimation yields a *point-estimate* of our quantity. In Frequentist inference.

- Bias: $\mathrm{E}[\hat{\theta}] - \theta$
- Variance: $\mathrm{V}[\hat{\theta}]$

Frequentist philosophy: What are we "averaging over" in these expectations and variances? **They are taken with respect to the randomness in the data!**

Also, in frequentist inference you show your uncertainty using *confidence intervals*. This satisfies the following property:

$$P(\theta \in CI_{\alpha}[\hat{\theta}]) = \alpha$$

Demonstration: What is the bias, variance, and 95% confidence interval of $\hat{\mu}$?

# Maximum Likelihood Estimation Steps (on cheat-sheet)

1. Define your data, its corresponding distribution, and the unknown parameters
   *Example:* $X_1, \ldots, X_n \sim N(\mu, 1)$ with unknown parameter $\mu$
2. Write out the likelihood function
   *Example:* $L(\mu; X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2}\right)$
3. Take the log of the likelihood
   *Example:* $\ell(\mu; X) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}$
4. Find the value of the parameter that maximizes the likelihood through the differentiation or through numerical optimization
   *Example:* $\frac{\partial}{\partial\mu}\ell(\mu; X) = 0$ yields $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$
5. Find a confidence interval by deriving the distribution of your estimator
   *Example:* $\hat{\mu} \sim N\left(\overline{x}, \frac{1}{n}\right)$ so $CI_\alpha[\hat{\mu}] = \left(\overline{x} - z_{\alpha/2}\frac{1}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{1}{\sqrt{n}}\right)$

# Quick Note about Real Life

- For the likelihood function, even if the derivative exists, there may not be a clear solution to the MLE for some parameters. The derivative can be used to maximize the likelihood function through *gradient ascent*.
- For other likelihood functions analytic derivatives do not even exist. In this case, the derivative can be estimated using finite differencing (here, $\theta$ is )

$$\hat{L}'(\theta) = \frac{L(\theta + \Delta\theta) - L(\theta)}{\Delta\theta}$$

# Bayesian Inference

Say you have data $X_1, \ldots, X_n \sim N(\mu, 1)$ with values $x_1, \ldots, x_n$. In frequentist inference, we treated $\mu$ as a fixed yet unknown value. But what if $\mu$ was itself random, say $\mu \sim N(0, 1)$?

Then, we can use Bayes rule (using $P$ instead of $f$ for simplicity):

$$P(\mu|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|\mu)P(\mu)}{P(x_1, \ldots, x_n)}$$

# Bayesian Estimation

Since we are interested in the distribution of $\mu$ and $P(x_1, \ldots, x_n)$ does not depend on $\mu$, we usually write

$$P(\mu|x_1, \ldots, x_n) \propto P(x_1, \ldots, x_n|\mu)P(\mu)$$

**Note:** The $\propto$ operator means "proportional to". You can use this to get rid of any terms that do not involve $\mu$!

The term $P(\mu|x_1, \ldots, x_n) = L(\mu : x_1, \ldots, x_n)$ is the likelihood. So maximizing the posterior (MAP, Maximum A-Posteriori, Estimator) can be interpreted as maxizing the likelihood scaled by the prior. If the prior is uniform, say $P(\mu) = 1$, the MAP and MLE will give the same result.

# Bayesian Demonstration

Say $X_1, \ldots, X_n \sim N(\mu, 1)$ and let the prior be $\mu \sim N(0, \sigma_0^2)$. What is the posterior distribution $P(\mu|X)$?

# Bayesian Demonstration

Say $X_1, \ldots, X_n \sim N(\mu, 1)$ and let the prior be $\mu \sim N(0, \sigma_0^2)$. What is the posterior distribution $P(\mu|X)$?

Solution:

$$\mu|X \sim N\left(\frac{n\overline{x}}{n + \sigma_0^{-2}}, \frac{1}{n + \sigma_0^{-2}}\right)$$

# Practice!

Recall:

$$\mu|X \sim N\left(\frac{n\bar{x}}{n+\sigma_0^{-2}}, \frac{1}{n+\sigma_0^{-2}}\right)$$

1. Is the posterior mean of $\mu$ a biased estimator? What happens to the bias as $\sigma_0^2$ gets smaller or larger? Can you come up with an intuitive reason for the bias, and why it might be desirable?
2. What happens to the posterior variance as $\sigma_0^2$ gets large? What about when it gets very small?
3. *Challenge*: Consider a similar set-up however $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and let the prior be $\mu \sim N(0, \sigma_0^2)$ (so the variance of $X_i$ is no longer one). What is the posterior distribution $P(\mu|X)$?

# Steps to Bayesian Inference (on cheat-sheet)

1. Define your data, your parameter, and the prior for your parameter
   *Example:* $X_1, \ldots, X_n \sim N(\mu, 1)$ and let the prior be $\mu \sim N(0, \sigma_0^2)$
2. Use Bayes rule to write the posterior of the parameter given the data
   *Example:* $P(\mu|X) \propto P(X|\mu)P(\mu) =$
   $\frac{1}{\sqrt{2\pi}^n} \exp\left(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma_0^2}\right)$
3. Re-arrange to see if you can get the posterior distribution to be in a familiar form. If not, you may need to use more advanced techniques such as MCMC
   *Example* Through several algebraic steps, we can see that
   $$\mu|X \sim N\left(\frac{n\overline{x}}{n+\sigma_0^{-2}}, \frac{1}{n+\sigma_0^{-2}}\right)$$
   .
4. Calculate the posterior mean and credible interval from the posterior distribution
   *Example:* $\hat{\mu} = \frac{\overline{x}}{1+\sigma_0^2}$ and

# Note about real life

In this case the posterior distribution is normal with mean and variance terms that can be easily calculated. The reason for this is that the normal prior for $\mu$ is a *conjugate prior.* Many distributions have a conjugate prior (and there are lists online), however for most more advanced models there will not be a conjugate prior.

In this case you generally perform Markov Chain Monte-Carlo (MCMC), which is a method for approximating the posterior distribution when you do not know the distribution's form. Without going into the details, you can think of MCMC as numerical gradient descent but for Bayesian inference rather than frequentist inference. The major downside is that MCMC is often prohibitively time consuming for complex models. Bayesian inference for practical models has only been practical recently due to improvements in computing.

# Review

| **Frequentist** | **Bayesian** |
|---|---|
| Data is viewed as a random sample from a distribution and the parameters are fixed | Data is fixed and inference is done parameters which are viewed as random |
| Point estimates are obtained through maximum likelihood estimation (MLE) | Estimates obtained through posterior distributions which contain information on mean, median, variance, et cetera |
| Uncertainty quantified through confidence intervals: if we were to resample our data many times, 95% of the time our sample will contain the true parameter value | Uncertainty quantified through credible intervals: there is a 95% chance that the parameter is within our interval |
| No way to incorporate prior inforation | Prior inforation can be included, however when prior information is absent it may be difficult to derive "uninformative" prior |
| Less computationally intensive | Usually more computationally intensive |

## Key terms in statistical inference:

- *Distribution:* Distributions describe frequencies through probability density or probability mass functions. Usually are characterized by one or more parameters. Example: $N(\mu, \sigma^2)$, Poisson($\lambda$), Uniform($a, b$), et cetera.
- *Parameter:* These are values that characterize a distribution, such as $\mu, \sigma^2, \lambda$, et cetera. They are often referred to using the symbol $\theta$.
- *Random variables:* Random variables are written capital letters. Random variables are given distributions, which we write as $X_i \sim N(0, 1)$. Random variables are not numbers or values, but can take on values for example $X_i = x_i$.
- *Data values:* Data values are written as lowercase letters, such as $x_i$. As above, when data values are are observations of random variables we write $X_i = x_i$.

## Frequentist vs. Bayesian Inference

| **Frequentist** | **Bayesian** |
|---|---|
| Data is viewed as a random sample from a distribution and the parameters are fixed | Data is fixed and inference is done parameters which are viewed as random |
| Point estimates are obtained through maximum likelihood estimation (MLE) | Estimates obtained through posterior distributions which contain information on mean, median, variance, et cetera |
| Uncertainty quantified through confidence intervals: if we were to re-sample our data many times, 95% of the time our sample will contain the true parameter value | Uncertainty quantified through credible intervals: there is a 95% chance that the parameter is within our interval |
| No way to incorporate prior inforation | Prior inforation can be included, however when prior information is absent it may be difficult to derive "un-informative" prior |
| Less computationally intensive | Usually more computationally intensive when the posterior distribution does not have a closed form |

## Steps to Frequentist Inference

1. Define your data, its corresponding distribution, and the unknown parameters
   *Example:* $X_1, \ldots, X_n \sim N(\mu, 1)$ with unknown parameter $\mu$
2. Write out the likelihood function
   *Example:* $L(\mu; X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2}\right)$
3. Take the log of the likelihood
   *Example:* $\ell(\mu; X) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}$
4. Find the value of the parameter that maximizes the likelihood through the differentiation or through numerical optimization
   *Example:* $\frac{\partial}{\partial \mu}\ell(\mu; X) = 0$ yields $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$
5. Find a confidence interval by deriving the distribution of your estimator
   *Example:* $\hat{\mu} \sim N\left(\overline{x}, \frac{1}{n}\right)$ so $CI_\alpha[\hat{\mu}] = \left(\overline{x} - z_{\alpha/2}\frac{1}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{1}{\sqrt{n}}\right)$

## Steps to Bayesian Inference

1. Define your data, your parameter, and the prior for your parameter
   *Example:* $X_1, \ldots, X_n \sim N(\mu, 1)$ and let the prior be $\mu \sim N(0, \sigma_0^2)$
2. Use Bayes rule to write the posterior of the parameter given the data
   *Example:* $P(\mu|X) \propto P(X|\mu)P(\mu) = \frac{1}{\sqrt{2\pi}^n} \exp\left(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma_0^2}\right)$
3. Re-arrange to see if you can get the posterior distribution to be in a familiar form. If not, you may need to use more advanced techniques such as MCMC
   *Example* Through several algebraic steps, we can see that

$$\mu|X \sim N\left(\frac{n\overline{x}}{n+\sigma_0^{-2}}, \frac{1}{n+\sigma_0^{-2}}\right)$$

   .
4. Calculate the postetrior mean and credible interval from the posterior distribution
   *Example:* $\hat{\mu} = \frac{\overline{x}}{1+\sigma_0^2}$ and $CI_\alpha[\hat{\mu}] = \left(\frac{n\overline{x}}{n+\sigma_0^{-2}} - z_{\alpha/2}\sqrt{\frac{1}{n+\sigma_0^{-2}}}, \frac{n\overline{x}}{n+\sigma_0^{-2}} + z_{\alpha/2}\sqrt{\frac{1}{n+\sigma_0^{-2}}}\right)$

## Common Probability Distributions

| Distribution | Parameters | PMF/PDF | Mean | Variance |
|---|---|---|---|---|
| Bernoulli | $p > 0$ | $P(X = 1) = p$ | $p$ | $p(1-p)$ |
| Binomial | $n > 0, 0 < p < 1$ | $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ | $np$ | $np(1-p)$ |
| Poisson | $\lambda > 0$ | $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ | $\lambda$ | $\lambda$ |
| Uniform | $a < b$ | $\frac{1}{b-a}$ for $a \leq x \leq b$ | $\frac{1}{2}(a+b)$ | $\frac{1}{12}(b-a)^2$ |
| Normal | $-\infty < \mu < \infty, \sigma^2 > 0$ | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu$ | $\sigma^2$ |
| Exponential | $\lambda > 0$ | $\lambda e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |

## Intuition for the Most Common Distributions

- *Binomial:* If $X \sim Binomial(n, p)$, then $X$ is the number of heads recorded from $n$ (biased) coin flips. Note that $Bernoulli(p) = Binomial(1, p)$, and that if $X_1, \ldots, X_n \sim Bernoulli(p)$ then $\sum_{i=1}^n X_i \sim Binomial(n, p)$.
- *Normal:* By the central limit theorem, or CLT, the average of $n$ random variables $X_1, \ldots, X_n$ has the normal distribution; formally $\frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \sigma^2)$ (as long as $n$ is large enough! Usually $n > 30$ is considered fine).
- *Poisson:* Say a call center receives an average of $\lambda$ calls per hour, and each call is randomly timed and independent. Then $X \sim Poisson(\lambda)$ is the random variable for the number of calls received within one hour.
- *Exponential:* At the call center mentioned above, the amount of time between calls is distributed $X \sim Exponential(\lambda)$.