

# 1 Probability and statistics

## 1.1 Working with probability distributions

- Given probability distribution  $\mathbb{P}$ , sample space  $\Omega$ , and event  $A \subseteq \Omega$ :

- $\mathbb{P} \geq 0 \quad \forall A$  (probabilities are nonzero)
- $\mathbb{P}[\Omega] = 1$  (probabilities sum to 1)
- $\mathbb{P}[\emptyset] = 0$  (probability of empty set is 0)
- $\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i] = 1$

- Probabilities are independent when the joint probability is equal to the product of the marginal probabilities.

$$A \perp\!\!\!\perp B \iff \mathbb{P}[A \cap B] = \mathbb{P}[A] \mathbb{P}[B]$$

- The conditional probability of  $A$  given  $B$  is the joint probability of  $A$  and  $B$  divided by the probability of just  $B$ .

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

- The Probability Mass Function (PMF) is used to describe the behavior of *discrete* probability distributions.

$$f_X(x) = \mathbb{P}[X = x]$$

- The Probability Density Function (PDF) is the equivalent for *continuous* distributions. We use the PDF to determine the probability that random variable  $X$  is between  $A$  and  $B$ .

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x) dx$$

- The Cumulative Distribution Function (CDF) is the integral of the PDF and we use it to determine the probability that random variable  $X$  is less than or equal to  $x$ . It maps  $\mathbb{R} \rightarrow [0, 1]$  and is monotonically non-decreasing. The left and right limits are 0 and 1 ( $\lim_{x \rightarrow -\infty} = 0$  and  $\lim_{x \rightarrow \infty} = 1$ ).

$$F_X(x) = \mathbb{P}[X \leq x]$$

### 1.1.1 Notes on the normal distribution

- The normal distribution is a function of mean  $\mu$  and variance  $\sigma^2$
- The simplest case is the **standard normal distribution**,  $Z \sim \mathcal{N}(0, 1)$ , which reduces to:

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

- Interestingly, others have defined even simpler standard normals. Gauss proposed  $\sigma^2 = \frac{1}{2}$ , which reduces to:

$$\phi(x) = \frac{e^{-x^2}}{\sqrt{\pi}}$$

- Stigler proposed a formulation with  $\sigma^2 = \frac{1}{2\pi}$ , leading to:

$$\phi(x) = e^{-\pi x^2}$$

- We can convert any normally distributed variable  $X$  to a *standard normal* by subtracting the mean and dividing by the standard deviation.

$$Z = \frac{X - \mu}{\sigma}$$

- 68-95-99.7 rule:** the percentage of values that lie within 1, 2, and 3 standard deviations of the mean of a normal distribution are 68.27%, 95.45%, and 99.73% respectively. A  $\mu \pm 3\sigma$  deviation should occur at a frequency of about 1 in 370.
- The Gauss Error Function gives the probability of a RV  $Z \sim \mathcal{N}(0, 1/2)$  falling in the range  $[-x, x]$ :

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}$$

### 1.1.2 Notes on the uniform distribution

- The continuous uniform distribution is a function of the minimum and maximum values  $a$  and  $b$  with mean and median equal to  $\frac{a+b}{2}$
- The **standard uniform** is a random variable  $\sim \mathcal{U}(0, 1)$
- The PDF of a uniform distribution is a horizontal line from  $a$  to  $b$

## 1.2 Common distributions

	Type	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	<i>Discrete</i>	$\begin{cases} 0 & x < a \\ \frac{\lfloor x \rfloor - a + 1}{b - a} & a \leq x \leq b \\ 1 & x > b \end{cases}$	$\frac{I(a \leq x \leq b)}{b - a + 1}$	$\frac{a + b}{2}$	$\frac{(b - a + 1)^2 - 1}{12}$	$\frac{e^{as} - e^{-(b+1)s}}{s(b - a)}$
Bernoulli	<i>Discrete</i>	$(1 - p)^{1-x}$	$p^x (1 - p)^{1-x}$	$p$	$p(1 - p)$	$1 - p + pe^s$
Binomial	<i>Discrete</i>	$I_{1-p}(n - x, x + 1)$	$\binom{n}{x} p^x (1 - p)^{n-x}$	$np$	$np(1 - p)$	$(1 - p + pe^s)^n$
Multinomial	<i>Discrete</i>		$\frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \quad \sum_{i=1}^k x_i = n$	$\begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$	$\begin{pmatrix} np_1(1 - p_1) & -np_1p_2 \\ -np_2p_1 & \ddots \end{pmatrix}$	$\left( \sum_{i=0}^k p_i e^{s_i} \right)^n$
Poisson	<i>Discrete</i>	$e^{-\lambda} \sum_{i=0}^x \frac{\lambda^i}{i!}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\lambda$	$\lambda$	$e^{\lambda(e^s - 1)}$
Uniform	<i>Continuous</i>	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$	$\frac{I(a < x < b)}{b - a}$	$\frac{a + b}{2}$	$\frac{(b - a)^2}{12}$	$\frac{e^{sb} - e^{sa}}{s(b - a)}$
Normal	<i>Continuous</i>	$\Phi(x) = \int_{-\infty}^x \phi(t) dt$	$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$	$\mu$	$\sigma^2$	$\exp \left\{ \mu s + \frac{\sigma^2 s^2}{2} \right\}$
Log-Normal	<i>Continuous</i>	$\frac{1}{2} + \frac{1}{2} \operatorname{erf} \left[ \frac{\ln x - \mu}{\sqrt{2\sigma^2}} \right]$	$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right\}$	$e^{\mu + \sigma^2/2}$	$(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$	
Multivariate Normal	<i>Continuous</i>		$(2\pi)^{-k/2}  \Sigma ^{-1/2} e^{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)}$	$\mu$	$\Sigma$	$\exp \left\{ \mu^T s + \frac{1}{2} s^T \Sigma s \right\}$
Student's $t$	<i>Continuous</i>	$I_x \left( \frac{\nu}{2}, \frac{\nu}{2} \right)$	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left( 1 + \frac{x^2}{\nu} \right)^{-(\nu+1)/2}$	$0 \quad \nu > 1$	$\begin{cases} \frac{\nu}{\nu - 2} & \nu > 2 \\ \infty & 1 < \nu \leq 2 \end{cases}$	
Chi-square	<i>Continuous</i>	$\frac{1}{\Gamma(k/2)} \gamma \left( \frac{k}{2}, \frac{x}{2} \right)$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$	$k$	$2k$	$(1 - 2s)^{-k/2} \quad s < 1/2$
Exponential	<i>Continuous</i>	$1 - e^{-x/\beta}$	$\frac{1}{\beta} e^{-x/\beta}$	$\beta$	$\beta^2$	$\frac{1}{1 - \frac{s}{\beta}} \quad (s < \beta)$

### 1.3 Hypothesis testing

- Framework for filtering implausible scientific claims
- Basic steps:
  1. State relevant null hypothesis ( $H_0$ ) and alternative hypothesis ( $H_1$ )
    - Two-sided:  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta \neq \theta_0$
    - One-sided:  $H_0 : \theta \leq \theta_0$  vs  $H_1 : \theta > \theta_0$
  2. Determine relevant test statistic ( $T$ ) distribution, typically Student's  $t$  or normal distribution
  3. Select significance level ( $\alpha$ , often 5% or 1%)
  4. Calculate rejection region (critical region), which contains all values of  $x$  for which  $T(x)$  is greater than the critical value  $c$ :  $R = \{x : T(x) > c\}$
  5. Determine whether to accept or reject  $H_0$
- Alternatively, just calculate the  $p$ -value (probability given  $H_0$  of getting a result at least as extreme as that which was observed). Reject the null hypothesis if  $p \leq \alpha$ .
- Common ranges for  $p$ -values are:
  - $< 0.01$ : very strong evidence against  $H_0$
  - $[0.01, 0.05]$ : strong evidence against  $H_0$
  - $[0.05, 0.10]$ : weak evidence against  $H_0$
  - $> 0.1$ : yikes man
- Type I errors (false positives) occur when we incorrectly **reject** the null hypothesis. This is equivalent to  $\alpha$ .
- Type II errors (false negatives) occur when we incorrectly **fail to reject** the null hypothesis.

	Retain $H_0$	Reject $H_0$
$H_0$ true	✓	Type I Error ( $\alpha$ )
$H_1$ true	Type II Error ( $\beta$ )	✓ (power)

- *note: should probably have something on multiple comparisons here*

### 1.4 Bayesian inference

## 2 Linear algebra

### 2.1 Objects and notation

- Let scalar  $s \in \mathbb{R}$

- Let vector  $\mathbf{x} \in \mathbb{R}^n$ . We should assume that all vectors are ‘column vectors’ (ie a matrix in  $\mathbb{R}^{n \times 1}$ )
- Let 2-d matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ . We’ll identify specific elements like this:

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- We’ll denote a whole column  $i$  of a matrix as  $\mathbf{A}_{:,i}$  and a row  $j$  as  $\mathbf{A}_{j,:}$

- Tensors extend beyond 2d, eg:  $\mathbf{A}_{i,j,k}$

### 2.2 Basic matrix operations review

- The **transpose** operation mirrors the matrix across the diagonal and is denoted  $\mathbf{A}^T$ .

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

- Addition of matrices is element-wise, and therefore requires them to be the same shape.

$$C_{i,j} = A_{i,j} + B_{i,j} \quad \{A, B, C\} \in \mathbb{R}^{m \times n}$$

- The **matrix product** of  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{B} \in \mathbb{R}^{n \times p}$  is  $\mathbf{C} \in \mathbb{R}^{m \times p}$ . Note that the number of columns in the first matrix must be equal to the number of rows in the second matrix ( $n$ ). Each element in  $C_{i,j}$  can be thought of as the dot product between row  $i$  of  $\mathbf{A}$  and column  $j$  of  $\mathbf{B}$ .

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}$$

- Some matrix operation properties:
  - Distributive:  $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
  - Associative:  $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
  - **NOT** commutative:  $\mathbf{AB} \neq \mathbf{BA}$
  - Transpose product:  $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

#### 2.2.1 The identity matrix

- We’ll define the **identity matrix**  $\mathbf{I}_n$  as the matrix that does not change a vector  $\mathbf{x}$  of dimension  $n$  when they are multiplied together so that  $\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{I}_n \mathbf{x} = \mathbf{x}$ . The identity matrix is just a square matrix with 1 on the diagonal and 0 elsewhere, so for  $\mathbf{x} \in \mathbb{R}^3$ :

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

### 2.2.2 Matrix inversion

- The **matrix inverse** of  $\mathbf{A}$  is denoted  $\mathbf{A}^{-1}$  and we define it such that:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

- $\mathbf{A}$  is **invertible** if it is square ( $\in \mathbb{R}^{n \times n}$ ) and non-singular.
  - A square matrix is **singular**  $\iff$  it has a determinant of 0
  - Singular matrices have linearly dependent columns
    - \* The **determinant** of a matrix (usually denoted  $\det(\mathbf{A})$  or  $|\mathbf{A}|$ ) is a scalar factor that can be computed from the elements of a square matrix. For a  $2 \times 2$  matrix:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Rightarrow |\mathbf{A}| = ad - bc$$

### 2.3 Systems of linear equations

- We can define a **system of linear equations**,  $\mathbf{Ax} = \mathbf{b}$ .  $\mathbf{A}$  is a known matrix of coefficients,  $\mathbf{b}$  is a known vector, and we're trying to solve for vector  $\mathbf{x}$ . The matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  describes a system of  $m$  equations with  $n$  unknowns.
- This is really the same as writing:

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \dots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

- For other important properties of invertible matrices see [Wikipedia: Invertible matrix theorem](#)

## 3 Differential equations

### 3.1 Calculus refresher

- Some useful properties / rules with differentiable functions  $f(x)$  and  $g(x)$ :
  - $(cf)' = c(f')$  for any constant  $c$
  - $c' = 0$  for any constant  $c$
  - $(f + g)' = f' + g'$
  - **Power rule:**  $\frac{d}{dx}(x^n) = nx^{n-1}$
  - **Product rule:**  $(fg)' = f'g + g'f$

- **Quotient rule:**  $\left(\frac{f}{g}\right)' = \frac{f'g - g'f}{g^2}$
- **Chain rule:**  $f(g(x))' = f'(g)g'$

- Common derivatives:

- $\frac{d}{dx} x = 1$
- $\frac{d}{dx} cx = c$
- $\frac{d}{dx} e^x = e^x$
- $\frac{d}{dx} \ln x = \frac{1}{x}, \quad x > 0$
- $\frac{d}{dx} \ln |x| = \frac{1}{x}, \quad x \neq 0$
- $\frac{d}{dx} c^x = c^x \ln c$
- $\frac{d}{dx} \sin x = \cos x$
- $\frac{d}{dx} \cos x = -\sin x$
- $\frac{d}{dx} \tan x = \sec^2 x$
- $\frac{d}{dx} \log_c x = \frac{1}{x \ln c}, \quad x > 0$

- Common antiderivatives:

- $\int 0 \, dx = C$
- $\int 1 \, dx = x + C$
- $\int n \, dx = nx + C$
- $\int e^x \, dx = e^x + C$
- $\int \frac{1}{x} \, dx = \ln x + C$
- $\int x^n \, dx = \frac{x^{n+1}}{n+1} + C, \quad n \neq -1$
- $\int \sin x \, dx = -\cos x + C$
- $\int \cos x \, dx = \sin x + C$

### 3.2 1st order differential equations

### 3.3 2nd order differential equations