

COMP50008 Probability & Statistics

Imperial College London

Boxuan Tang

Spring 2023

Contents

1	Probability	2
1.1	Events	2
1.2	Random Variables	2
1.3	Discrete Random Variables	3
1.4	Continuous Random Variables	3
1.5	Joint Random Variables	3
2	Statistics	4

1 Probability

1.1 Events

Definition 1.1 (Sample Spaces) A set that describes the range of possible outcomes of a random experiment

Definition 1.2 (Events) An event E is any subset of the sample space $E \subseteq S$, a collection of some of its possible outcomes. The singleton subsets of S are the **elementary** events of S . The events are **mutually exclusive** if $\forall i, j, E_i \cap E_j = \emptyset$ and at most one of the events can occur.

Definition 1.3 (σ -algebra) The set of subsets \mathcal{F} must be

- nonempty ($S \in \mathcal{F}$)
- closed under complement ($E \in \mathcal{F} \rightarrow \bar{E} \in \mathcal{F}$)
- closed under countable union ($E_1, \dots \in \mathcal{F} \rightarrow \bigcup_i E_i \in \mathcal{F}$)

Definition 1.4 (Axioms of Probability) A **probability measure** on the pair (S, \mathcal{F}) is a mapping $P : \mathcal{F} \rightarrow [0, 1]$ satisfying the following axioms for all subsets of S

- $\forall E \in \mathcal{F}, 0 \leq P(E) \leq 1$
- $P(S) = 1$
- countably additive meaning for disjoint subsets $E_1 \dots \in \mathcal{F}, P(\bigcup_i E_i) = \sum_i P(E_i)$

Definition 1.5 (Independent Events) A set of events $\{E_1, E_2, \dots\}$ is independent iff for any finite subset $\{E_{i_1} \dots E_{i_n}\}$, $P(\bigcap_{j=1}^n E_{i_j}) = \prod_{j=1}^n P(E_{i_j})$. Two events E and F are independent iff $P(E \cap F) = P(E)P(F)$. If E and F are independent, \bar{E} and F are also independent.

Definition 1.6 (Conditional Probability) $P(E | F) = \frac{P(E \cap F)}{P(F)}$. The events E_1 and E_2 are **conditionally independent** given F iff $P(E_1 \cap E_2 | F) = P(E_1 | F)P(E_2 | F)$.

Theorem 1.7 (Bayes Theorem) $P(E | F) = \frac{P(E)P(F|E)}{P(F)}$.

Theorem 1.8 (Chain Rule) $P(A \cap B \cap C) = P(A)P(B | A)P(C | A \cap B)$

Theorem 1.9 (Partition Rule) $P(E) = \sum_i P(E | F_i)P(F_i) = P(E | F)P(F) + P(E | \bar{F})P(\bar{F})$

1.2 Random Variables

Definition 1.10 (Random Variable) A random variable X is a mapping from the sample space to real numbers, $X : S \rightarrow \mathbb{R}$. The **support** of X is the image of S under X , $\text{supp}(X) = X(S) = \{x \in \mathbb{R} | \exists s \in S, X(s) = x\}$

Definition 1.11 (Cumulative Distribution Function) The cumulative distribution function $F_x(x)$ is the probability that X takes a value less than or equal to x , $F_x(x) = P_x(X \leq x)$. For it to be valid:

- (**Monotonic**) $\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \rightarrow F_x(x_1) \leq F_x(x_2)$
- $F_x(-\infty) = 0, F_x(\infty) = 1$
- F_x is right-continuous

$$P_x(a < X \leq b) = F_x(b) - F_x(a)$$

Definition 1.12 (Expectation) **Discrete:** $\mu = E(X) = \sum_x xp(x)$.

Continuous: $\mu = E(x) = \int_{-\infty}^{\infty} xf_X(x)dx$ or generally $E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$

Linearity of Expectation: $E(aX + b) = aE(x) + b$ and $E(X + Y) = E(X) + E(Y)$.

Definition 1.13 (Variance) $\text{Var}(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2$. $\text{Var}(aX + b) = a^2\text{Var}(X)$
 $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i,j} \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$ **Standard Deviation** $\sigma = \sqrt{\text{Var}(X)}$.

Definition 1.14 (Convolution) For $Z = X + Y$, (similar for discrete)

$$F_Z(n) = \int_{k=-\infty}^{\infty} F_X(k)F_Y(n-k)dk \quad f_Z(n) = \int_{k=-\infty}^{\infty} f_X(k)f_Y(n-k)dk$$

1.3 Discrete Random Variables

Definition 1.15 (Probability Mass Function) For a discrete random variable X , the probability mass function is $p(x) = P_x(X = x)$ where $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$. Also $p(x_i) = F(x_i) - F(x_{i-1})$ and $F(x_i) = \sum_{j=1}^i p(x_j)$

Definition 1.16 (Skewness) $\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3}$

Definition 1.17 (Sum of Random Variables) Let $X_1 \dots X_n$ be random variables and $S_n = \sum_{i=1}^n X_i$. $E(S_n) = \sum_{i=1}^n E(X_i)$ and $E(\frac{S_n}{n}) = \frac{\sum_{i=1}^n E(X_i)}{n}$. $Var(S_n) = \sum_{i=1}^n Var(X_i)$ and $Var(\frac{S_n}{n}) = \frac{\sum_{i=1}^n Var(X_i)}{n^2}$. If they are independent and identically distributed where $E(X_i) = \mu_x$ and $Var(X_i) = \sigma_x^2$, $E(\frac{S_n}{n}) = \mu_x$ and $Var(\frac{S_n}{n}) = \frac{\sigma_x^2}{n}$.

Distribution	R_x	$f(x)$	μ	σ^2	γ_1	$\hat{\theta}_{MLE}$	$M_X(t)$
Bernoulli(p)	0, 1	$p^x(1-p)^{1-x}$	p	p(1-p)			$(1-p) + pe^t$
Binomial(n,p)	0 ... n	$\binom{n}{x} p^x (1-p)^{n-x}$	np	np(1-p)	$\frac{1-2p}{\sqrt{np(1-p)}}$	\bar{x}	$((1-p) + pe^t)^n$
Geometric(p)	1 ...	$p(1-p)^{x-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{2-p}{\sqrt{1-p}}$	$\frac{1}{\bar{x}}$	$\frac{pe^t}{1-(1-p)e^t}$
Poisson(λ)	0 ...	$\frac{e^{-\lambda} \lambda^x}{x!}$	λ	λ	$\frac{1}{\sqrt{\lambda}}$	\bar{x}	$e^{\lambda(e^t-1)}$
Uniform(1,n)	1 ... n	$\frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$			

When p is small and n is large, Binomial(n,p) \sim Poisson(np). Geometric is memoryless.

1.4 Continuous Random Variables

Definition 1.18 (Probability Density Function) The probability density function f_X of a **continuous** X is such that $F_X(x) = \int_{-\infty}^x f_X(u) du$ where $\forall x \in \mathbb{R}, f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(x) dx = 1$. So $f_X(x) = \frac{d}{dx} F_X(x)$. Also $P_X(a < X \leq b) = \int_a^b f_X(x) dx$.

Definition 1.19 (Quantile) α -quantile $Q_X(\alpha), 0 \leq \alpha \leq 1$ is the least number satisfying $P(X \leq Q_X(\alpha)) = \alpha$, $Q_X(\alpha) = F_X^{-1}(\alpha)$. The median is the $\frac{1}{2}$ -quantile and the k^{th} percentile is the $\frac{k}{100}$ -quantile.

Distribution	$f(x)$	$F(x)$	R_x	μ	σ^2	$\hat{\theta}_{MLE}$	$M_X(t)$
U(a,b)	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$		$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Exp(λ)	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{1}{\bar{x}}$	$\frac{\lambda}{\lambda-t}$
N(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$		all	μ	σ^2	(\bar{x}, S_{n-1}^2)	$e^{\mu t + \sigma^2 t^2 / 2}$
Lognormal(μ, σ^2)	$\frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$		+ve	$e^{\mu + \frac{\sigma^2}{2}}$	$e^{\sigma^2} - 1$		

If events in a random process \sim Poisson(λ) then the time between consecutive events \sim Exp(λ).

$X \sim N(\mu, \sigma^2) \implies \frac{X-\mu}{\sigma} \sim \Phi$ and $\Phi(1.645) \approx 0.95, \Phi(1.96) \approx 0.975, \Phi(2.58) \approx 0.995$.

Exponential is memoryless and has only support from 0 onwards.

Definition 1.20 (Moment Generating Function) $E[X^n] = \frac{d^n M_X(t)}{dt^n} \big|_{t=0}$ where

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f_X(x) dx$$

Definition 1.21 (Sum of Independent Random Variables) $M_{S_n}(t) = \prod_{j=1}^n M_{X_j}(t)$ and $E[\prod_{i=1}^n Z_i] = \prod_{i=1}^n E[Z_i]$. For 2 variables, $M_{Z_1+Z_2}(t) = M_{Z_1}(t)M_{Z_2}(t)$ and $E[Z_1 Z_2] = E[Z_1]E[Z_2]$.

Theorem 1.22 (Central Limit Theorem) Let $X_1 \dots X_n$ be independent and identically distributed random variables from any distribution with mean μ and **finite** variance σ^2 . Then $\lim_{n \rightarrow \infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \phi$

1.5 Joint Random Variables

Definition 1.23 (Joint Cumulative Distribution Function) $F(x, y) = P_Z(X \leq x, Y \leq y)$

so $F_X(x) = F(x, \infty)$ and $F_Y(y) = F(\infty, y)$

- $\forall x, y \in \mathbb{R}, 0 \leq F(x, y) \leq 1$

- **Monotonicity** $x_1 < x_2 \implies F(x_1, y_1) \leq F(x_2, y_1)$ and $y_1 < y_2 \implies F(x_1, y_1) \leq F(x_1, y_2)$
- $\forall x, y \in \mathbb{R}. F(x, -\infty) = F(-\infty, y) = 0$ and $F(\infty, \infty) = 1$

$$P_Z(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

Definition 1.24 (Joint Probability Mass Function) $p(x, y) = P_Z(X = x, Y = y)$ where $\forall x, y \in \mathbb{R}, 0 \leq p(x, y) \leq 1$ and $\sum_y \sum_x p(x, y) = 1$. So $p_X(x) = \sum_y p(x, y)$ and $p_Y(y) = \sum_x p(x, y)$.

Definition 1.25 (Joint Probability Density Function) $F(x, y) = \int_{t=-\infty}^y \int_{s=-\infty}^x f(s, t) ds dt$ where $\forall x, y \in \mathbb{R}, f(x, y) \geq 0$ and $\int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f(x, y) dx dy = 1$. So $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$. Also $f_X(x) = \int_{y=-\infty}^{\infty} f(x, y) dy$ and $f_Y(y) = \int_{x=-\infty}^{\infty} f(x, y) dx$ (**Marginal Density Functions**)

Definition 1.26 (Independence) **Discrete:** $p(x, y) = p_X(x)p_Y(y)$. **Continuous:** $f(x, y) = f_X(x)f_Y(y)$.

Definition 1.27 (Partition Rule) **Discrete:** $p_X(x) = \sum_y p_{X|Y}(x | y)p_Y(y)$.

Continuous: $f_X(x) = \int_{y=-\infty}^{\infty} f_{X|Y}(x | y)f_Y(y) dy$.

Definition 1.28 (Expectation) **Discrete:** $E(g(X, Y)) = \sum_y \sum_x g(x, y)p(x, y)$.

Continuous: $E(g(X, Y)) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x, y)f(x, y) dx dy$.

If $g(X, Y) = g_1(X) + g_2(Y)$ then $E(g(X, Y)) = E_X(g_1(X)) + E_Y(g_2(Y))$.

If $g(X, Y) = g_1(X)g_2(Y)$ and X and Y are independent, then $E(g(X, Y)) = E_X(g_1(X))E_Y(g_2(Y))$.

Definition 1.29 (Covariance) $\sigma_{XY} = E[XY] - \mu_X \mu_Y$. For independent rvs, $\sigma_{XY} = 0$.

Definition 1.30 (Correlation) $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.

Definition 1.31 (Conditional Expectation) **Discrete:** $E_{Y|X}(Y | x) = \sum_y y p_{Y|X}(y | x)$.

Continuous: $E_{Y|X}(Y | x) = \int_{y=-\infty}^{\infty} y f_{Y|X}(y | x) dy$. Also, $E_Y(Y) = E_X(E_{Y|X}(Y | X))$.

Tower Rule: $E(Y) = E_{X_n}(E_{X_{n-1}}(\dots E_{X_1}(E_Y(Y | X_1 \dots X_n) | X_2 \dots X_n) \dots | X_n))$.

Definition 1.32 (Discrete Time Markov Chain) $P(X_n = j) = (\pi_0 R^n)_j$ where $P(X_0 = i) = \pi_{0i}$ for the horizontal initial probability vector π_0 and $r_{ij} = P(X_{n+1} = j | X_n = i)$ for the transition matrix R . Since $\pi_\infty R = \pi_\infty$, R has an eigenvalue of 1 with the eigenvector π_∞ .

2 Statistics

Definition 2.1 (Bias) The bias of an estimator T for a parameter θ is $\text{bias}(T) = E[T | \theta] - \theta$. If the estimator has zero bias we say the estimator is unbiased.

Definition 2.2 (Variance) $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. **Bias-corrected Sample Variance:**

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} (\sum_i x_i^2 - n\bar{x}^2) = \frac{n}{n-1} S^2$$

Definition 2.3 (Efficiency of Estimators) For 2 unbiased estimators $\hat{\Theta}$ and $\tilde{\Theta}$, $\hat{\Theta}$ is more efficient than $\tilde{\Theta}$ if $\forall \theta. \text{Var}_{\hat{\Theta}|\theta}(\hat{\Theta} | \theta) \leq \text{Var}_{\tilde{\Theta}|\theta}(\tilde{\Theta} | \theta)$ and $\exists \theta. \text{Var}_{\hat{\Theta}|\theta}(\hat{\Theta} | \theta) < \text{Var}_{\tilde{\Theta}|\theta}(\tilde{\Theta} | \theta)$

Definition 2.4 (Consistency of Estimators) $\hat{\Theta}$ is a consistent estimator for θ if $\forall \epsilon > 0. P(|\hat{\Theta} - \theta| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$. If $\hat{\Theta}$ is unbiased then $\lim_{n \rightarrow \infty} \text{Var}(\hat{\Theta}) = 0 \implies \hat{\Theta}$ is consistent.

Definition 2.5 (Confidence Interval) For known population variance σ^2 , the $100(1-\alpha)\%$ CI for μ is

$$[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

State CLT if used. Otherwise, the $100(1-\alpha)\%$ confidence interval for μ is

$$[\bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}]$$

Definition 2.6 (Hypothesis Testing) Identify the rejection region R of T under the assumption H_0 is true, $P(T \in R | H_0) = \alpha$.

State CLT if used. State assume H_0 to be true. State at what α -level.

Testing if $\bar{X} = \mu_0$: For known population variance σ^2 , $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \Theta$. Otherwise, $T = \frac{\bar{X} - \mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$.

Testing if $\mu_X = \mu_Y$. For known population variance σ^2 , $Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Theta$. Otherwise,

$$T = \frac{\bar{X} - \bar{Y}}{S_{n_1+n_2-2} \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2} \quad S_{n_1+n_2-2}^2 = \frac{n_1-1}{n_1+n_2-2} S_{n_1-1}^2 + \frac{n_2-1}{n_1+n_2-2} S_{n_2-1}^2$$

Type I error is rejecting H_0 when it is true. **Type II** error is not rejecting H_0 when H_1 is true.

Power of Test: $P(T \in R | H_1)$, high probability of rejecting H_0 when H_1 is true.

Definition 2.7 (Chi-Squared Test)

$$X_{k-p-1}^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Define H_0 and H_1 . If $x^2 < X_{k-p-1, 1-\alpha}^2$, where k is the number of values X can take and p is the number of parameters being estimated, we do not reject the H_0 at the α significance level.

Independence Test: For $k \times l$ observed, expected value is $\hat{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$ with $(k-1)(l-1)$ degrees of freedom.

Definition 2.8 (Likelihood) $L(\theta) = P(X | \theta)$ or $l(\theta) = \log P(X | \theta)$.

Maximum Likelihood Estimate = $\text{argmax}_{\theta} L(\theta | x)$:

1. Write down $L(\theta) = \prod_{i=1}^n f(x_i | \theta)$ which is the product of the n pdf/pmf viewed as a function of θ .
2. Take the natural log of the likelihood to get $l(\theta)$.
3. Find the value of θ where $l(\theta)$ is maximised by solving $\frac{\partial}{\partial \theta} l(\theta) = 0$
4. **Check** that the estimate in step 3 is a maximum by checking that $\frac{\partial^2}{\partial \theta^2} l(\theta) < 0$

Definition 2.9 (Posterior) Posterior = Likelihood \times Prior $\times \frac{1}{\text{Evidence}}$ or $P(\theta | X) = P(X | \theta) \times P(\theta) \times \frac{1}{P(X)}$

Bayesian Estimate: $\hat{\theta}_B$ is the mean of the new distribution

Maximum A Posteriori Estimate: $\hat{\theta}_{MAP} = \text{argmax}_{\theta} [\prod_{i=1}^n P(X = x_i | \theta) \times P(\theta)]$

Definition 2.10 (Beta Prior) Used for Bernoulli, Binomial and Geometric distributions.

$$\text{Beta}(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta} \quad \max = \frac{\alpha-1}{\alpha+\beta-2} \quad \mu = \frac{\alpha}{\alpha+\beta} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Given a $\text{Beta}(\theta; \alpha, \beta)$ prior, with sample size n and mean \bar{x} , the posterior is $\text{Beta}(\theta; \alpha + n\bar{x}, \beta + n(1 - \bar{x}))$

Definition 2.11 (Gamma Prior) Used for Poisson and Exponential distributions.

$$\text{Gamma}(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1} e^{-b\theta}}{\int_0^\infty \theta^{\alpha-1} e^{-b\theta} d\theta} \quad \max = \frac{\alpha-1}{\beta} \quad \mu = \frac{\alpha}{\beta} \quad \sigma^2 = \frac{\alpha}{\beta^2}$$

Given a $\text{Gamma}(\theta; \alpha, \beta)$ prior, with sample size n and mean \bar{x} , the posterior is $B(\theta; \alpha + n\bar{x}, \beta + n)$

Definition 2.12 (Normal Prior) Used for Normal distributions. $\max = \mu, \mu = \mu$

Given a $N(\mu_0, \sigma_0^2)$ prior, with sample size n of $N(\mu, \sigma_x^2)$ where sample mean is $\bar{\mu}$ and the variance is known, the posterior is $N(\mu_1, \sigma_1^2)$ where

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma_x^2} \quad \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{n\bar{\mu}}{\sigma_x^2}$$