Introduction to Probability
CST Part IA Paper 1

Victor Zhao
xz398@cantab.ac.uk

1 Prerequisites and Introduction

1. Combinatorics:

Counting tasks on n objects			
Permutations (sort objects)		Combinations (choose r objects)	
Distinct	Indistinct	Distinct 1 group	Distinct k groups
$n!$	$\dfrac{n!}{n_1!n_2!\cdots n_r!}$	$\dbinom{n}{r} = \dfrac{n!}{r!(n-r)!}$	$\dbinom{n}{n_1, n_2, \cdots, n_k} = \dfrac{n!}{n_1!n_2!\cdots n_k!}$

   Pascal's identity: $\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r} \qquad (1 \le r \le n)$

   Binomial theorem: $(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}$

2. Probability axioms:

   Axiom 1: For any event E, $0 \le \mathbb{P}[E] \le 1$

   Axiom 2: Probability of the sample space S is $\mathbb{P}[S] = 1$

   Axiom 3: If E and F are mutually exclusive (i.e., $E \cap F = \varnothing$), then $\mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F]$
   In general, for all mutually exclusive events $E_1, E_2, \cdots$,

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} E_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[E_i]$$

3. General inclusion-exclusion principle: $\mathbb{P}\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{r=1}^{n} (-1)^{r+1} \left(\sum_{i_1 < \cdots < i_r}^{n} \mathbb{P}[E_{i_1} \cap \cdots \cap E_{i_r}] \right)$

   Case $n = 2$: $\mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F] - \mathbb{P}[E \cap F]$

4. Union bound (Boole's inequality): For any events $E_1, E_2, \cdots, E_n$,

$$\mathbb{P}\left[\bigcup_{i=1}^{n} E_i\right] \le \sum_{i=1}^{n} \mathbb{P}[E_i]$$

5. Conditional probability (original and conditioning on event G):

   Chain rule:

$$\mathbb{P}[EF] = \mathbb{P}[E|F]\mathbb{P}[F] \qquad \mathbb{P}[EF|G] = \mathbb{P}[E|FG]\mathbb{P}[F|G]$$

   Multiplication rule:

$$\mathbb{P}[E_1E_2\cdots E_n] = \mathbb{P}[E_1]\mathbb{P}[E_2|E_1]\cdots[E_n|E_1\cdots E_{n-1}]$$
$$\mathbb{P}[E_1E_2\cdots E_n|G] = \mathbb{P}[E_1|G]\mathbb{P}[E_2|E_1G]\cdots[E_n|E_1\cdots E_{n-1}G]$$

   Independence of E and F:

$$\mathbb{P}[EF] = \mathbb{P}[E]\mathbb{P}[F] \qquad \mathbb{P}[EF|G] = \mathbb{P}[E|G]\mathbb{P}[F|G]$$
$$\mathbb{P}[E|F] = \mathbb{P}[E] \qquad \mathbb{P}[E|FG] = \mathbb{P}[E|G]$$

Law of total probability:

$$\mathbb{P}[E] = \mathbb{P}[EF] + \mathbb{P}[EF^{\complement}] = \mathbb{P}[E|F]\mathbb{P}[F] + \mathbb{P}[E|F^{\complement}]\mathbb{P}[F^{\complement}]$$
$$\mathbb{P}[E|G] = \mathbb{P}[EF|G] + \mathbb{P}[EF^{\complement}|G] = \mathbb{P}[E|FG]\mathbb{P}[F|G] + \mathbb{P}[E|F^{\complement}G]\mathbb{P}[F^{\complement}|G]$$

In general, for disjoint events $F_1, F_2, \cdots, F_n$ such that $F_1 \cup \cdots \cup F_n = S$,

$$\mathbb{P}[E] = \sum_{i=1}^{n} \mathbb{P}[E|F_i]\mathbb{P}[F_i] \qquad \mathbb{P}[E|G] = \sum_{i=1}^{n} \mathbb{P}[E|F_iG]\mathbb{P}[F_i|G]$$

Bayes' theorem:

$$\mathbb{P}[F|E] = \frac{\mathbb{P}[E|F]\mathbb{P}[F]}{\mathbb{P}[E]} \qquad \mathbb{P}[F|EG] = \frac{\mathbb{P}[E|FG]\mathbb{P}[F|G]}{\mathbb{P}[E|G]}$$

6. Confusion matrix:

		Actual condition	
	Total population	Positive F	Negative $F^{\complement}$
Predicted condition	Positive E	True positive $\mathbb{P}[E\|F]$	False positive $\mathbb{P}[E\|F^{\complement}]$
	Negative $E^{\complement}$	False negative $\mathbb{P}[E^{\complement}\|F]$	True negative $\mathbb{P}[E^{\complement}\|F^{\complement}]$

2 Random Variables

1. Probability distribution functions:

Discrete random variable X:

- Probability mass function (PMF): $p(x)$
- Compute probability:

$$\mathbb{P}[X = a] = p(x)$$
$$\mathbb{P}[a \le X \le b] = \sum_{x=a}^{b} p(x)$$

- Cumulative distribution function (CDF):

$$F_X(a) = \mathbb{P}[X \le a] = \sum_{x \le a} p(x)$$

Continuous random variable X:

- Probability density function (PDF): $f(x)$
- Compute probability:

$$\mathbb{P}[X = a] = 0$$
$$\mathbb{P}[a \le X \le b] = \int_a^b f(x)dx$$

- Cumulative distribution function (CDF):

$$F_X(a) = \mathbb{P}[X \le a] = \int_{-\infty}^{a} f(x)dx$$

2. Expectation:

Discrete random variable X:

$$\mathbb{E}[X] = \sum_x xp(x)$$
$$\mathbb{E}[g(X)] = \sum_x g(x)p(x)$$

Continuous random variable X:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$$
$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Linearity of expectation: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$

Additivity of expectation: $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

3. Variance: $\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$

Scaling of variance: $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$

Standard deviation: $\mathbb{SD}[X] = \sqrt{\mathbb{V}[X]}$

Scaling of standard deviation: $\mathbb{SD}[aX + b] = |a|\mathbb{SD}[X]$

4. Discrete distributions:

   Bernoulli $\mathrm{Ber}(p)$: 1 experiment with success probability p

   $$\mathbb{P}[X=1]=p \qquad \mathbb{E}[X]=p \qquad \mathbb{V}[X]=p(1-p)$$

   Binomial $\mathrm{Bin}(n,p)$: n independent trials with success probability p

   $$\mathbb{P}[X=k]=\binom{n}{k}p^k(1-p)^{n-k} \qquad \mathbb{E}[X]=np \qquad \mathbb{V}[X]=np(1-p)$$

   Poisson $\mathrm{Pois}(\lambda)$: # successes over experiment duration, with success rate $\lambda=np$

   $$\mathbb{P}[X=k]=\frac{\lambda^k}{k!}e^{-\lambda} \qquad \mathbb{E}[X]=\lambda \qquad \mathbb{V}[X]=\lambda$$

   Geometric $\mathrm{Geo}(p)$: # independent trials until first success, with success probability p

   $$\mathbb{P}[X=n]=(1-p)^{n-1}p \qquad \mathbb{E}[X]=\frac{1}{p} \qquad \mathbb{V}[X]=\frac{1-p}{p^2}$$

   Negative binomial $\mathrm{NegBin}(r,p)$: # independent trials until r success, with success probability p

   $$\mathbb{P}[X=n]=\binom{n-1}{r-1}(1-p)^{n-r}p^r \qquad \mathbb{E}[X]=\frac{r}{p} \qquad \mathbb{V}[X]=\frac{r(1-p)}{p^2}$$

   Hypergeometric $\mathrm{Hyp}(N,n,m)$: # objects with a feature in a sample of size n (without replacement) from a population of size N that contains m items with the feature

   $$\mathbb{P}[X=n]=\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \qquad \mathbb{E}[X]=n\frac{m}{N} \qquad \mathbb{V}[X]=n\frac{m}{N}\left(1-\frac{m}{N}\right)\left(1-\frac{n-1}{N-1}\right)$$

5. Continuous distributions:

   Uniform $\mathrm{Uni}(\alpha,\beta)$: equal probability within range $[\alpha,\beta]$

   PDF: $f(x)=\begin{cases}\frac{1}{\beta-\alpha} & \text{when } \alpha\le x\le\beta\\ 0 & \text{otherwise}\end{cases}$ CDF: $F(x)=\begin{cases}0 & \text{when } x<\alpha\\ \frac{x-\alpha}{\beta-\alpha} & \text{when } \alpha\le x\le\beta\\ 1 & \text{when } x>\beta\end{cases}$

   $\mathbb{E}[X]=\dfrac{\alpha+\beta}{2}$ $\mathbb{V}[X]=\dfrac{(\beta-\alpha)^2}{12}$

   Exponential $\mathrm{Exp}(\lambda)$: time until first success occurs, with success rate λ

   PDF: $f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{when } x\ge 0\\ 0 & \text{otherwise}\end{cases}$ CDF: $F(x)=1-e^{-\lambda x}$

   $\mathbb{E}[X]=\dfrac{1}{\lambda}$ $\mathbb{V}[X]=\dfrac{1}{\lambda^2}$

   Normal (Gaussian) $\mathcal{N}(\mu,\sigma^2)$: mean μ, variance σ^2

   PDF: $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu^2)}{2\sigma^2}}$ CDF: $F(x)=\Phi\left(\dfrac{x-\mu}{\sigma}\right)$

   $\mathbb{E}[X]=\mu$ $\mathbb{V}[X]=\sigma^2$

   $X\sim\mathcal{N}(\mu,\sigma^2) \implies aX+b\sim\mathcal{N}(a\mu+b,a^2\sigma^2)$

   $X\sim\mathcal{N}(\mu_X,\sigma_X^2), Y\sim\mathcal{N}(\mu_Y,\sigma_Y^2) \implies X+Y\sim\mathcal{N}(\mu_X+\mu_Y,\sigma_X^2+\sigma_Y^2)$

6. Continuity correction:

Discrete		Continuous
$\mathbb{P}[X = a]$	$\approx$	$\mathbb{P}[a - 0.5 \le X \le a + 0.5]$
$\mathbb{P}[X > a]$	$\approx$	$\mathbb{P}[X \ge a + 0.5]$
$\mathbb{P}[X \ge a]$	$\approx$	$\mathbb{P}[X \ge a - 0.5]$
$\mathbb{P}[X < a]$	$\approx$	$\mathbb{P}[X \le a - 0.5]$
$\mathbb{P}[X \le a]$	$\approx$	$\mathbb{P}[X \le a + 0.5]$

7. Joint probability mass function (for discrete RVs): $p_{X,Y}(a, b) = \mathbb{P}[X = a, Y = b]$
   Joint distribution function (for discrete or continuous RVs): $F_{X,Y}(a, b) = \mathbb{P}[X \le a, Y \le b]$
   Joint probability density f and joint continuous distribution F (for continuous RVs):

$$F(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)dxdy \qquad f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

$$\mathbb{P}[a_1 \le X \le b_1, a_2 \le Y \le b_2] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)dxdy$$

   Marginal distribution: $F_X(a) = \mathbb{P}[X \le a] = \lim_{b\to\infty} F_{X,Y}(a, b)$

8. Covariance: $\text{Cov}[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
   Covariance of linear combinations:

$$\text{Cov}[X, a] = 0 \qquad \text{Cov}[X, X] = \mathbb{V}[X]$$
$$\text{Cov}[aX, bY] = ab\text{Cov}[X, Y] \qquad \text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$$

   Variance of sum: $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{Cov}[X, Y]$
   In general, for any random variables X_1, X_2, $\cdots$, X_n:

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i] + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{Cov}[X_i, X_j]$$

   Correlation coefficient: $\rho(X, Y) = \dfrac{\text{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} \in [-1, 1]$ ($\rho(X, Y) = 0$ if $\mathbb{V}[X] = 0$ or $\mathbb{V}[Y] = 0$)
   Scaling-invariance of correlation coefficient: $\rho(aX, bY) = \rho(X, Y)$

3 Moments and Limit Theorems

1. Markov's inequality: for any non-negative random variable X with finite $\mathbb{E}[X]$, for any $a > 0$,

$$\mathbb{P}[X \ge a] \le \frac{\mathbb{E}[X]}{a}$$

   Let $a = \delta \cdot \mathbb{E}[X]$ (where $\delta > 0$), then the inequality can be rewritten as

$$\mathbb{P}[X \ge \delta \cdot \mathbb{E}[X]] \le \frac{1}{\delta}$$

2. Chebyshev's inequality: for any random variable X with finite $\mathbb{E}[X]$ and $\mathbb{V}[X]$, for any $a > 0$,

$$\mathbb{P}[|X - \mathbb{E}[X]| \ge a] \le \frac{\mathbb{V}[X]}{a^2}$$

   Let $a = \sqrt{\delta \cdot \mathbb{V}[X]}$ (where $\delta > 0$), then the inequality can be rewritten as

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \ge \sqrt{\delta \cdot \mathbb{V}[X]}\right] \le \frac{1}{\delta}$$

3. Weak law of large numbers: let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i's are independent and identically distributed (i.i.d.) with finite expectation μ and finite variance σ^2. Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[|\overline{X}_n - \mu| > \epsilon\right] = 0$$

Strong law of large numbers:

$$\mathbb{P}\left[\lim_{n\to\infty} \overline{X}_n = \mu\right] = 1$$

4. Central limit theorem: let $X_1, X_2, \cdots, X_n$ be any sequence of i.i.d. random variables with finite expectation μ and finite variance σ^2. Let

$$Z_n = \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^{n} X_i - n\mu\right)$$

Then for any number $a \in \mathbb{R}$, it holds that

$$\lim_{n\to\infty} F_{Z_n}(a) = \Phi(a) = \frac{1}{2\pi}\int_{-\infty}^{a} e^{-\frac{x^2}{2}}\, dx$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$.

4 Applications and Statistics

1. Estimators:

   An estimator T is an unbiased estimator for the parameter θ if $\mathbb{E}[T] = \theta$ irrespective of the value θ.

   The bias of an estimator T is defined as $\mathbb{E}[T] - \theta = \mathbb{E}[T - \theta]$.

2. Unbiased estimator for the expectation and variance:

   Let $X_1, X_2, \cdots, X_n$ be identically distributed samples from a distribution with finite expectation μ and finite variance σ^2. Then

   - $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ is an unbiased estimator for μ; and
   - $S_n = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ is an unbiased estimator for σ^2.

3. Bias-variance decomposition of the mean squared error:

$$\mathrm{MSE}[T] = \mathbb{E}[(T-\theta)^2] = \underbrace{(\mathbb{E}[T]-\theta)^2}_{\text{Bias}^2} + \underbrace{\mathbb{V}[T]}_{\text{Variance}}$$

   - Estimator T_1 is better than T_2 if $\mathrm{MSE}[T_1] < \mathrm{MSE}[T_2]$;
   - If T_1 and T_2 are both unbiased, then T_1 is better than T_2 iff $\mathbb{V}[T_1] < \mathbb{V}[T_2]$.

4. Jensen's inequality: for any random variable X, and any convex function $g : \mathbb{R} \to \mathbb{R}$ (i.e., for all λ, a and b, $\lambda g(a) + (1-\lambda)g(b) \geq g(\lambda a + (1-\lambda)b)$), we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

   If g is strictly convex and X is not constant, then the inequality is strict.

5. Expected number of samples until first collision: $\sqrt{\dfrac{\pi N}{2}} - \dfrac{1}{3} + O(\dfrac{1}{\sqrt{N}})$

6. The secretary problem (maximise the probability of stopping at the best of n candidates):

   Optimal strategy: reject the first $x-1$ candidates, then accept the first candidate $i \geq x$ that is better than all candidates before

   Probability of success: $\dfrac{x-1}{n}\displaystyle\sum_{i=x}^{n}\frac{1}{i-1} \approx \frac{x}{n}\ln\left(\frac{n}{x}\right)$

   Optimal $x = \dfrac{n}{e} \implies$ maximum success probability: $\dfrac{1}{e}$