Contents

1 Review of 407

1.1 Random Variables & Some Preliminaries

(1) A random variable $X : \Omega \to \mathbb{R}$ is **continuous** if

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(x)\,\mathrm{d}x$$

for some $f_X : \mathbb{R} \to [0, \infty)$ *and for all* $a \leqslant b$. If so we say f_X is the **PDF** of X, and we define the **CDF** of X by

$$F_X(t) := \mathbb{P}(X \leqslant t).$$

(2) We say two **events** $A_1, A_2 \subset \Omega$ are **independent** if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$. *More generally, we say events* $A_1, ..., A_n \subset \Omega$ *are independent if any of the* A_i*'s satisfy that inequality; that is, for* $S \subset \{1, 2, ..., n\}$*,*

$$\mathbb{P}(\bigcup_{i \in S} A_i) = \prod_{i \in S} \mathbb{P}(A_i).$$

(3) Some examples of random variables:

(i) **Bernoulli**: let $0 < p < 1$. Define $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

(ii) **Binomial**: let $n \in \mathbb{N}$ and $0 < p < 1$. Define $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

(iii) **Geometric**: let $n \in \mathbb{N}$ and $0 < p < 1$. Define $\mathbb{P}(X = n) = (1-p)^{n-1} p$.

(iv) **Gaussian / Normal** $\mathcal{N}(\mu, \sigma^2)$: let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ (or $\sigma > 0$). Define

$$f_X(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- For a Gaussian, $\mathbb{E}X^{2n} = (2n-1)!!\sigma^{2n}$; in particular $\mathbb{E}X^4 = 3\sigma^4$. See here.
- If $X_1, ..., X_n$ are i.i.d. with $X_i \sim (N)(\mu, \sigma^2)$, then $\overline{X} := (X_1 + ... + X_n)/n \sim \mathcal{N}(\mu, \sigma^2/n)$.
- **Standard Gaussian** $Z := \mathcal{N}(0, 1)$:

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

(v) **Gamma** $\Gamma(\alpha,\beta)$: let $\alpha,\beta>0$. Define

$$f_X(x)=\frac{x^{\alpha-1}\exp(-x/\beta)}{\beta^\alpha\Gamma(\alpha)}\qquad\text{for } x\in[0,\infty),$$

where

$$\Gamma(\alpha):=\int_0^\infty e^{-x}x^{\alpha-1}\,\mathrm{d}x.$$

- Γ *interpolates the factorial since* $\Gamma(1)=1,\Gamma(n+1)=n\Gamma(n)$ *so* $\Gamma(n)=(n-1)!$ *for* $n\in\mathbb{Z}$.

(vi) **Chi-Squared** χ^2_n: a chi-squared distribution with n **degrees of freedom** is $\Gamma(n/2,2)$:

$$f_X(x)=\frac{x^{n/2-1}\exp(-x/2)}{2^{n/2}\Gamma(n/2)}\qquad\text{for } x\in[0,\infty).$$

- *If* $X_1,\dots,X_n$ *are i.i.d. standard Gaussians, then* $\sum_{i=1}^n X_i^2\sim\chi^2_n$ *(HW2 p2).*
- χ^2_n has mean n and variance $2n$.

(4) Let $A\subset\Omega$. The **indicator function** χ_A (or $\mathbb{1}_A$): $\Omega\to\{0,1\}$ is given by $\chi_A(x)=1$ if $x\in A$ and $=0$ otherwise.

1.2 Expected Values, Variance, Joint Distributions, & Related Properties

(1) (**Expected value**) If X is a nonnegative random variable, the **expected value** of X is defined by

$$\mathbb{E}X:=\int_0^\infty\mathbb{P}(X>t)\,\mathrm{d}t.$$

(*Think Lebesgue.*) *More generally, if* $\mathbb{E}|X|<\infty$*, define* $\mathbb{E}X:=\mathbb{E}(\max(X,0))-\mathbb{E}(\max(-X,0))$.

- If X is **discrete**, the above definition is equivalent to $\mathbb{E}X=\sum_x x\cdot\mathbb{P}(X=x)$.
- If X is continuous with PDF f_X, the above is equivalent to $\mathbb{E}X=\int_{-\infty}^\infty x\cdot f_X(x)\,\mathrm{d}x$.
- Expected values are **linear** and **finitely additive**:

$$\mathbb{E}(aX+b)=a\mathbb{E}X+b\qquad\text{and}\qquad\mathbb{E}(\sum_{i=1}^n X_i)=\sum_{i=1}^n\mathbb{E}X_i.$$

(2) (**Variance**) The **variance** of X is given by

$$\mathrm{var}(X):=\mathbb{E}X^2-(\mathbb{E}X)^2=\mathbb{E}(X-\mathbb{E}X)^2.$$

(*Here* $\mathbb{E}X^2$ *denotes the expected value of* X^2*, and* $\mathbb{E}(X-\mathbb{E}X)^2$ *denotes the expected value of* $(X-\mathbb{E}X)^2$.)

- Important property of variance: $\mathrm{var}(aX+b)=\mathrm{var}(aX)=a^2\,\mathrm{var}(X)$.

(3) (**Joint distribution**) Let X,Y be random variables; their **joint PDF** is the function $f_{X,Y}:\mathbb{R}\to[0,\infty)$ satisfying

$$\mathbb{P}((X,Y)\in[a,b]\times[c,d])=\iint_{[a,b]\times[c,d]}f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y\qquad\text{for all } [a,b],[c,d]\subset\mathbb{R}\cup\{\pm\infty\}.$$

- The **marginal** of X is given by

$$f_X(x):=\int_{-\infty}^\infty f_{X,Y}(x,y)\,\mathrm{d}y$$

(and likewise for that of Y). *This can be generalized to more than two variables.*

- $\mathbb{E}(XY)=\iint_{\mathbb{R}^2}xyf_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y$. More generally, if $g:\mathbb{R}^2\to\mathbb{R}$ then

$$\mathbb{E}(g(X,Y))=\iint_{\mathbb{R}^2}g(x,y)f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

(4) (**Conditional**)

(a) If $A, B \subset \Omega$, define the **conditional probability** as $\mathbb{P}(A \mid B) := \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

- (**Total Expectation Theorem**) Let $A_1, ..., A_n$ partition Ω. Then for all $B \in \Omega$,

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).$$

(b) If X, Y are random variables, define the **conditional distribution** of $X \mid Y = y$ by

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

(*Note that $X \mid Y = y$ is a random variable!*)

(c) If X, Y are random variables, let $g(y) := \mathbb{E}(X \mid Y = y)$. Define the **conditional expectation** $\mathbb{E}(X \mid Y)$ to be $g(Y)$. (*Note that $g(Y)$ is again a random variable.*)

- Conditional expectation is linear, i.e., $\mathbb{E}(X + Y \mid Z) = \mathbb{E}(X \mid Z) + E(Y \mid Z)$.
- A variant of the Total Expectation Theorem: $\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}X$. See here for proof of both •'s.

(5) (**Independence**) $X_1, ..., X_n$ are said to be **independent** if $\mathbb{P}_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} \mathbb{P}_{X_i}(x_i)$ for all $x_i \in \mathbb{R}$.

- If X_1, X_2 are independent then $\mathbb{E}(X_1X_2) = \mathbb{E}X_1\mathbb{E}X_2$. *In general, if $X_1, ..., X_n$ are independent then*

$$\mathbb{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} \mathbb{E}X_i.$$

(6) (**Covariance**) The **covariance** of X, Y is given by $\operatorname{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$.

- $\operatorname{cov}(X, X) = \operatorname{var}(X)$.
- Covariance is **bilinear** and shift-invariant.
- Another way to express covariance is $\operatorname{cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$. This shows that if X, Y are independent then $\operatorname{cov}(X, Y) = 0$. The converse is *not* true – if $\operatorname{cov}(X, Y) = 0$ we say X, Y are **uncorrelated**. Being uncorrelated does not imply being independent.

(7) (**Correlation**) The **correlation** coefficient of X, Y is $\operatorname{corr}(X, Y) := \dfrac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$.

- Correlation is invariant under *scalar* multiplication: $\operatorname{corr}(tX, Y) = \operatorname{sgn}(t)\operatorname{corr}(X, Y)$.
- $-1 \leqslant \operatorname{corr}(X, Y) \leqslant 1$ for all random variables X, Y. Proof needs Cauchy-Schwarz mentioned below.
- *If $\operatorname{corr}(X, Y)$ is close to 1 then approximately $Y = aX + b$ (i.e., affine) and their graph resembles more of a straight line. If $\operatorname{corr}(X, Y)$ is close to 0 then the scatter plot of X, Y are closer to being everywhere randomly.*

1.3 Inequalities

(1) (**Cauchy-Schwarz**) If u, v are in an inner product space then $|\langle u, v\rangle| \leqslant \|u\|\|v\|$.

- *It turns out that $\langle X, Y\rangle := \mathbb{E}(XY)$ defines an inner product on the space of random variables, so*

$$|\langle U, V\rangle| = |\mathbb{E}(UV)| \leqslant \sqrt{\langle U, U\rangle \langle V, V\rangle} = \sqrt{\mathbb{E}U^2}\sqrt{\mathbb{E}V^2}.$$

Setting $U = X - \mathbb{E}X$ and $V = Y - \mathbb{E}Y$ gives $|\operatorname{corr}(X, Y)| \leqslant 1$, as claimed.

(2) (**Jensen's**) A function $\varphi : \mathbb{R} \to \mathbb{R}$ is **convex** if for $x < y$ and $\lambda \in [0,1]$ we have

$$\varphi(\lambda x + (1-\lambda)y) \leqslant \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

If X is a random variable and $\varphi : \mathbb{R} \to \mathbb{R}$ convex, then $\varphi(\mathbb{E}X) \leqslant \mathbb{E}(\varphi(X))$.

- For example, $\varphi(x) := x^2$ is convex, so $(\mathbb{E}X)^2 \leqslant \mathbb{E}X^2$.

(3) (**Markov's**) If $\mathbb{E}|X| < \infty$ (claim trivially holds if $\mathbb{E}|X| = \infty$) then $\mathbb{P}(|X| > t) \leqslant \mathbb{E}|X|/t$ and $\mathbb{P}(|X| > t) \leqslant \mathbb{E}|X|^n/t^n$.

(4) (**Chebyshev's**) If $\mathbb{E}X < \infty$ and $\mathrm{var}(X) < \infty$, then applying Markov's inequality to $X - \mathbb{E}X$ and $n = 2$ gives

$$\mathbb{P}(|X - \mu| > t) \leqslant \frac{\mathrm{var}(X)}{t^2}.$$

1.4 Modes of Convergence

(1) $\{Y_n\}$ is said to **converge almost surely** (a.s.) to Y if $\mathbb{P}(\lim_{n\to\infty} Y_n = Y) = 1$, or in more details

$$\mathbb{P}(\{\omega \in \Omega \mid \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1.$$

(2) $\{Y_n\}$ is said to **converge in probability** to Y if

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \epsilon) = 0 \qquad \text{for all } \epsilon > 0.$$

(3) $\{Y_n\}$ is said to **converge in distribution** Y if "*the limit of the CDF is the CDF of the limit*". To put formally, for all $t \in \mathbb{R}$ such that $S \mapsto \mathbb{P}(Y \leqslant S)$ is continuous at $s = t$,

$$\lim_{n\to\infty} \mathbb{P}(Y_n \leqslant t) = \mathbb{P}(Y \leqslant t).$$

(*) (1) $\Rightarrow$ (2) $\Rightarrow$ (3) but each implication is strict. This shows (3) $\not\Rightarrow$ (2), and this shows (2) $\not\Rightarrow$ (1).

1.5 The Limit Theorems

Theorem: Week Law of Large Numbers (WLLN)

Let $\{X_i\}$ be i.i.d. with $\mu := \mathbb{E}X_1 < \infty$. Then $(X_1 + \ldots + X_n)/n$ converges *in probability* to μ as $n \to \infty$:

$$\lim_{n\to\infty} \mathbb{P}\left(|(X_1 + \ldots + X_n)/n - \mu| > \epsilon\right) = 0 \qquad \text{for all } \epsilon > 0.$$

Theorem: Strong Law of Large Numbers (SLLN)

For the same assumptions as above, $(X_1 + \ldots + X_n)/n$ converges *almost surely* to μ as $n \to \infty$:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = \mu\right) = 1.$$

Theorem: Central Limit Theorem (CLT)

Let $\{X_i\}$ be i.i.d. with $\mathbb{E}|X_1| < \infty$ and $0 < \mathrm{var}(X_1) < \infty$. Define $\mu := \mathbb{E}X_1$ and $\sigma^2 := \mathrm{var}(X_1)$. Then

$$\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

converges *in distribution* to a standard Gaussian, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \leqslant a\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-t^2/2}\, \mathrm{d}t.$$

2 Random Samples

2.1 Sampling from Normal

(1) A **random sample** of size n is a sequence $X_1, ..., X_n$ of i.i.d. random variables.

(2) A **statistic** is a function of random variable — if $X_1, ..., X_n$ is a random sample of size n and $t : \mathbb{R}^n \to \mathbb{R}^k$, then a statistic is a random varaible of form

$$Y := t(X_1, ..., X_n).$$

The distribution of Y is called the **sampling distribution**.

- The **sample mean** is a statistic defined as $\overline{X} := (X_1 + ... + X_n)/n$.
- The **sample variance** $S := \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is also a statistic.

(3) (Sample mean and variance of a Gaussian) Let $X_1, ..., X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then

(a) $\overline{X}$ and S are independent,

(b) $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, and

(c) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

(4) (***t*-distribution**) Let X be a standard Gaussian and let $Y \sim \chi^2_p$. Then $X/\sqrt{Y/p}$ has a **student's *t*-distribution** with p degrees of freedom. The PDF is given by

$$f_{X/\sqrt{Y/p}}(t) = \frac{(\Gamma(p+1)/2)}{\sqrt{p\pi}\Gamma(p/2)}\left(1 + \frac{t^2}{p}\right)^{-(p+1)/2} \qquad t \in \mathbb{R}.$$

(5) (***f*-distribution** (HW2 p3)) Let $X \sim \chi^2_p$ and $Y \sim \chi^2_q$. Then $(X/p)/(Y/q)$ is a ***f*-distribution** with p and q degrees of freedom.

2.2 The Delta Method

Theorem: The Delta Methods

Let $\theta \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$. Let $Y_1, Y_2, \ldots$ be such that

$$\sqrt{n}(Y_n - \theta)$$

converges *in distribution* to $\mathcal{N}(0, \sigma^2)$ where $\sigma^2 > 0$. Assume $f'(\theta)$ exists.

(1) (**The Delta Method**) If $f'(\theta) \neq 0$ then $\sqrt{n}(f(Y_n) - f(\theta))$ converges *in distribution* to $\mathcal{N}(0, \sigma^2(f'(\theta)^2))$. (*Note that this is $\mathcal{N}(0,1)$ times $\sigma f'(\theta)$.*)

(2) (**The Second-Order Delta Method**) If $f'(\theta) = 0$ but $f''(\theta)$ exists and is nonzero, then $n(f(Y_n) - f(\theta))$ converges *in distribution* to $\sigma^2 f''(\theta)/2$ times χ^2_1. (*Note that this is $\mathcal{N}(0,1)^2$ times $\sigma^2 f''(\theta)/2$.*)

(3) More generally, if $f'(\theta) = \ldots = f^{(m-1)}(\theta) = 0$ but $f^{(m)}(\theta)$ exists and is nonzero, then

$$n^{m/2}(f(Y_n) - f(\theta))$$

converges in distribution to $\mathcal{N}(0,1)^m$ times $\sigma^m f^{(m)}(\theta)/m!$.

2.3 Simulation of Random Variables

(1) (Inverse and pseudoinverse CDF) If X is a random variable with an invertible CDF F and if U is the uniform random variable on $(0,1)$, then $F^{-1}(U)$ is a random variable with $\mathbb{P}(F^{-1}(U) \leqslant t) = F(t)$.

- If F is not invertible, for $s \in (0,1)$ we define

$$Y(s) := \sup\{t \in \mathbb{R} : F(t) < s\} = \inf\{t \in \mathbb{R} : F(t) \geqslant s\}$$

with uniform probability law on $(0,1)$. Then we still get $\mathbb{P}(Y \leqslant t) = F(t)$. (Quiz 2 prep sheet p1)

(2) (**Box-Muller Algorithm** (Quiz 2 prep sheet p2)) Let U_1, U_2 be independent random variables uniformly distributed on $(0,1)$. Then

$$X := \sqrt{-2\log U_1}\cos(2\pi U_2) \qquad Y := \sqrt{-2\log U_1}\sin(2\pi U_2)$$

are independent standard Gaussians.

2.4 Parameter Estimation

(1) (**Estimator**) A statistic Y is called a **point estimator** if it is used to estimate a parameter θ.

- θ can be a single real number or it can be a vector. *For example if we are "guessing" the mean μ and variance σ^2 of a Gaussian, we are considering the family*

$$\{f_\theta : \theta \in \Theta\} = \left\{\frac{1}{2\pi\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) : (\mu,\sigma) \in \mathbb{R} \times [0,\infty)\right\}.$$

(2) An estimator Y for $g(\theta)$ is **unbiased** if $\mathbb{E}_\theta(Y) = g(\theta)$, i.e., *its expected value is exactly what it estimates*.

- *For example, $\overline{X}$ the population mean is unbiased as $\mathbb{E}\overline{X} = \mu$ (here $g(\mu) := \mu$).*

(3) A sequence of estimators $\{Y_n\}$ for $g(\theta)$ is **consistent** if Y_n converges *in probability* to the *constant* random variable $g(\theta)$ with respect to f_θ.

- *It may be taken for granted that $\sum_{i=1}^{n} X_i^j$ is consistent for $\mu_j := \mathbb{E}X_i^j$, assuming $\mathbb{E}|X_1|^j$ is finite.*

(*) (a) Consistent but biased estimator: for example define $X := U[0,1]$ and let

$$Y_n := \left(\sum_{i=1}^{n} X_i^2/n - (\sum_{i=1}^{n} X_i/n)^2\right)^{1/2}.$$

Since M_1, M_2 are consistent and $(a,b) \mapsto \sqrt{a-b^2}$ is continuous, Y_n is consistent for σ. However, if $n = 1$ then $\mathbb{E}X = 1/2, \mathbb{E}X^2 = 1/3, \mathrm{var}(X) = 1/12$, and $\sigma = 1/\sqrt{3}$, whereas $Y_1 = \mathbb{E}\sqrt{X^2 - X^2} = 0$.

(b) Unbiased but not consistent: let $Y_n := X$ for all n (and let X be anything but a constant random variable). Then $\mathbb{E}Y_n = \mathbb{E}X$ is unbiased for $\mathbb{E}X$, but Y_n never converges to the constant random variable $\mathbb{E}X$.

(4) (**Method of Moments**) If $g(\theta)$ can be written as a function $h(\mu_1, ..., \mu_n)$ (where $\mu_j := \mathbb{E}X_1^j$), then the **method of moments** estimator for $g(\theta)$ is given by $h(M_1, ..., M_j)$, where $M_j := \sum_{i=1}^{n} X_i^j$.

(5) (**Sufficiency**) Y is **sufficient** for θ if (*for all y and all $\theta \in \Theta$*) the conditional distribution of $X = (X_1, ..., X_n)$ given $Y = y$ does not depend on θ, i.e., $f_{X_1,...,X_n|Y}(x_1, ..., x_n \mid y)$ can be expressed without mentioning θ.

- (**Factorization Theorem**) Y is sufficient if and only if (for all $\theta \in \Theta$) the distribution $f_\theta(x)$ can be factorized as $g_\theta(Y) \cdot h(x)$ for some function g_θ of y only and h of x only. (Here $x := (x_1, ..., x_n)$.)