Contents

1 Probability Theory **1**

2 Transformations and Expectations **3**
2.1 Transformations . . . 3
2.2 Expectations . . . 3
2.3 Moments . . . 4
2.3.1 Variance . . . 4
2.3.2 Moment Generating Functions . . . 4
2.4 Other Generating Functions . . . 5

3 Common Families of Distributions **6**
3.1 Multinomial Distribution . . . 6
3.2 Exponential Families . . . 6
3.3 Location and Scale Families . . . 7
3.4 Inequalities and Identities . . . 7

4 Multiple Random Variables **8**
4.1 Facts . . . 8
4.2 Bivariate Relations . . . 8
4.3 Inequalities . . . 9
4.3.1 Numerical Inequalities . . . 9
4.3.2 Functional Inequalities . . . 9

5 Properties of a Random Sample **10**

1 Probability Theory

Theorem 1 (Laws of Set Algebra)**.** For any three sets A, B, C all subsets of S, we have that the operators $\cup$ and $\cap$ are distributive, commutative, associative and satisfy DeMorgan's Laws:

$$(A \cup B)^{\mathsf{c}} = A^{\mathsf{c}} \cap B^{\mathsf{c}}$$
$$(A \cap B)^{\mathsf{c}} = A^{\mathsf{c}} \cup B^{\mathsf{c}}$$

Definition 1 (Sigma Algebra)**.** A collections of subsets S is calles a *sigma algebra* (or *Borel Field*), denoted by $\mathcal{B}$, if it has the following three properties:

a. $\emptyset \in \mathcal{B}$

b. $A \in \mathcal{B} \implies A^{\mathsf{c}} \in \mathcal{B}$

c. If $A_i \in \mathcal{B}$ for i in $\mathcal{I}$ then $\cup_{i \in \mathcal{I}} A_i \in \mathcal{B}$, where $\mathcal{I}$ is countable. So $\mathcal{B}$ is closed under countable union.

Note that from DeMorgan's laws we have

$$\left(\bigcup_{i=1}^{\infty} A_i{}^{\mathsf{c}} \right)^{\mathsf{c}} = \bigcap_{i=1}^{\infty} A_i \tag{1}$$

which means that, using b. we get $\mathcal{B}$ is also closed under countable intersections $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$.

Definition 2 (Kolmogorov Axioms)**.** Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function $\mathbb{P}$ with domain $\mathcal{B}$ that satisfies

1. $\mathbb{P}(A) \geq 0 \quad \forall A \in \mathcal{B}$
2. $\mathbb{P}(S) = 1$
3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty}) \mathbb{P}(A_i)$

The following result makes it a bit easier to find probability functions.

Theorem 2. Let $S = \{s_1, s_2, \dots\}$ be a countable set. Let $\mathcal{B}$ be any sigma algebra of subsets of S. Let $p_1, p_2, \dots$ be nonnegative numbers that sum to 1. For any $A \in \mathcal{B}$, define $\mathbb{P}(A)$ by

$$\mathbb{P}(A) = \sum_{\{i : s_i \in A\}} p_i.$$

Then $\mathbb{P}$ is a probability function on $\mathcal{B}$.

Definition 3 (Random Variable)**.** A *random variable* is a function from a sample space S into the real numbers.

Definition 4 (cdf)**.** The *cumulative distribution function* of *cdf* of a random variable X is defined by

$$F_X(x) = \mathbb{P}_X(X \leq x) \quad \forall x$$

Theorem 3. The function $F(x)$ is a cdf if and only if:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
2. $F(x)$ is a non-decreasing function of x.
3. $F(x)$ is right-continuous, that is, for every number x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$

We say that a random variable is continuous if its cdf is continuous, and we say that it is discrete if its cdf is a step function.

Theorem 4. The following statements are equivalent:

1. X and Y are identically distributed
2. $F_X(x) = F_Y(x) \quad \forall x$

Definition 5 (probability mass function)**.** The *probability mass function* or *pmf* of a discrete random variable X is

$$f_X(x) = P(X = x) \quad \forall x$$

Definition 6 (probability density function)**.** The *probability density function* or *pdf* of a continuous random variable X is the function $f_X(x)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t \quad \forall x$$

Theorem 5. A function $f_X(x)$ is a pdf (of pmf) of a random variable X if and only if

a. $f_X(x) \geq 0 \quad \forall x$

b. $\sum_x f_X(x) = 1$ (pmf) or $\int_x f_X(x)\mathrm{d}x = 1$ (pdf).

2 Transformations and Expectations

2.1 Transformations

If X and Y are discrete random variables, with $Y = g(X)$, then

$$f_Y(y) = \sum_{\{x:g(x)=y\}} f_X(x). \tag{2}$$

For the remainder of this section we will take X and Y to be continuous random variables, with $Y = g(X)$. As such, we will define the following sets $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = g(x) \quad x \in S \subseteq \mathcal{X}\}$.

Theorem 6. Let X have cdf $F_X(s)$, let $Y = g(X)$ and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as above. Then

a. If g is increasing on $\mathcal{X}$, then $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.

b. If g is decreasing on $\mathcal{X}$, then $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.

Theorem 7. Let X have pdf $f_X(x)$ and $Y = g(X)$, where g is monotone. Define $\mathcal{X}$ and $\mathcal{Y}$ as above. Suppose $f_X(x)$ is continuous on $\mathcal{X}$ and $g^{-1}(y)$ has continuous first derivative on $\mathcal{Y}$. Then the pdf of Y is given by:

$$f_Y(y) = \mathbb{1}\{y \in \mathcal{Y}\} f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|$$

If g is not globally monotone, then we just partition $\mathcal{X}$ into subsets on which g is continuous and monotone and sum the results. If such a partition doesn't exist, then we have technical problems.

Theorem 8 (Probability integral transform)**.** Let X have continuous cdf $F_X(x)$ and define the random variable Y by $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$.

If $F_X(x) = y$ is constant on some interval then we define the inverse by

$$F_X^{-1}(y) = \inf\{x : F_X(x) = y\}.$$

2.2 Expectations

Definition 7 (Expected value)**.** The *expected value* of a random variable $g(X)$, denoted $\mathbb{E}[X]$ is defined by:

- $\mathbb{E}[X] = \int_{\mathbb{R}} g(x) f_X(x) \mathrm{d}x$ if X is continuous,
- $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} g(x) \mathbb{P}(X = x)$ if X is discrete.

Theorem 9 (Properties of expectation)**.** a. Linearity.

b. $g_1(x) \geq g_2(x)\ \forall x \quad \Longrightarrow \quad \mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$

c. $a \leq g(x) \leq b\ \forall x \quad \Longrightarrow \quad a \leq \mathbb{E}[g(X)] \leq b$

d. $\operatorname{argmin}_c \mathbb{E}[(X - c)^2] = \mathbb{E}[X]$

2.3 Moments

Definition 8 (Moment)**.** For integer n, the n^{th} *moment* of X is

$$\mu'_n = \mathbb{E}[X^n].$$

The n^{th} *central moment*, μ_n is

$$\mu_n = \mathbb{E}[(X - \mu)^n].$$

Where $\mu = \mu_1 = \mathbb{E}[X]$

2.3.1 Variance

Definition 9 (Variance)**.** The *variance* of a random variable X, written $\text{Var}[X]$ is the second central moment of X,

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The *standard deviation* of X, denoted σ_X, is given by $\sigma_X = \sqrt{\text{Var}[X]}$.

Theorem 10 (Properties of variance)**.** If X has finite variance then:

a. $\text{Var}[aX + b] = a^2\text{Var}[X]$

b. $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

2.3.2 Moment Generating Functions

Definition 10 (Moment generating function)**.** Let X be a random variable with cdf F_X. The *moment generating function (mgf)* of X, denoted $M_X(t)$, is given by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

provided that the expectation exists for t in some (open) neighbourhood of 0 (otherwise we say the mgf does not exist).

Remark 1. The mgf is the Laplace transform of the pdf.

Theorem 11. If X has mgf $M_X(t)$ then

$$\mathbb{E}[X^n] = \frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\bigg|_{t=0}$$

The mgf can be used to calculate moments, but its principal utility is in characterising a distribution. This relationship can run into some technical difficulties. If the mgf exists, it characterises an infinite set of moments. However, it is possible for two distinct random variables to give rise to the same set of moments.

The problem of uniqueness of moments does not occur if the random variables have bounded support (in this case an infinite sequence of moments uniquely determines the distribution). Further, if the mgf exists in a neighbourhood of 0 then it uniquely determines the distribution, no matter the support. Thus, existence of an infinite set of moments is not equivalent to the existence of the mgf. We have the following theorem, describing when the mgf determines the distribution.

Theorem 12 (When mgf determines distribution)**.** Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.

a. If X and Y have bounded support then $F_X(u) = F_Y(u)\ \forall u$ if and only if $\mathbb{E}[X^r] = \mathbb{E}[Y^r]\ \forall r \in \mathbb{N}$. (So the cdfs are equal if and only if all the moments agree.)

b. If the mgfs exist and are identical in some neighbourhood of 0 then the cdfs are equal.

Theorem 13 (Convergence of mgfs near 0 implies convergence of cdfs)**.** Suppose $\{X_i, i = 1, 2, \dots\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Suppose also that for all t in a neighbourhood of 0

$$lim_{i\to\infty} M_{X_i} = M_X(t)$$

where $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x at which $F_X(x)$ is continuous, we have

$$lim_{i\to\infty} F_{X_i}(x) = F_X(x).$$

Remark 2. The convergence of a sequence of moments is not enough to show the convergence of random variables. We need the moment sequence to be unique too. However, if the mgfs converge in a neighbourhood of 0 as above, then we know that the random variables converge. Convergence of mgfs is therefore a sufficient, but not necessary, condition for convergence of the random variables.

Theorem 14. For any constants a and b

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

2.4 Other Generating Functions

Definition 11 (Cumulant generating function)**.** The *cumulant generating function* is $\log(M_X(t))$. The *cumulants* of X are defined as the coefficients of the Taylor series of this function.

Definition 12 (Factorial moment generating function)**.** The *factorial moment generating function* is $\mathbb{E}[t^X]$. The name comes from

$$\frac{\mathrm{d}^r}{\mathrm{d}t^r}\mathbb{E}[t^X]\bigg|_{t=1} = \mathbb{E}[X(X-1)\cdots(X-r+1).$$

For discrete distributions this is the *probability generating function* and the coefficients of the power series give the probabilities

$$\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\mathbb{E}[t^X]\bigg|_{t=0} = \mathbb{P}(X = k).$$

Definition 13 (Characteristic function)**.** The *characteristic function* of a random variable X is

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

Remark 3. The characteristic function is the most useful of the generating functions. Every cdf has a unique characteristic function. When the moments of the cdf exist, the characteristic function can be used to calculate them.

3 Common Families of Distributions

Lots of this chapter is standard definitions of distributions, so is omitted

3.1 Multinomial Distribution

Definition 14 (Multinomial Distribution)**.** Let m and n be positive integers and let $p_1, \ldots, p_n \in [0,1]$ satisfy $\sum_{i=1}^n p_i = 1$. Then the random vector $(X_1, \ldots, X_n)$ has *multinomial distribution with m trials and cell probabilities $p_1, \ldots, p_n$* if the joint pmf of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

on the set of $(x_1, \ldots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$.

Remark 4. The marginal distributions have $X_i \sim \text{binomial}(m, p_i)$.

Theorem 15 (Multinomial Theorem)**.** Let m and n be positive integers and let $\mathcal{A}$ be the set of vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$. Then for any real numbers $p_1, \ldots, p_n$,

$$(p_1 + \cdots + p_n)^m = \sum_{\boldsymbol{x} \in \mathcal{A}} \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}.$$

3.2 Exponential Families

Definition 15 (Exponential family 1)**.** A family of pmfs/pdfs is called an *exponential family* if it can be expressed

$$f(x|\boldsymbol{\theta}) = h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^k w_i(\boldsymbol{\theta}) t_i(x) \right)$$

where $h(x) \geq 0$, the t_i are real valued functions of the observation x that do not depend on $\boldsymbol{\theta}$ and $c(\theta) \geq 0$ and the $w_i(\boldsymbol{\theta})$ are real valued functions of $\boldsymbol{\theta}$ that do not depend on x.

Theorem 16. If X is a random variable from an exponential family distribution then

$$\mathbb{E}\left[\sum_{i=1}^k \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X) \right] = -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta})$$

and

$$\text{Var}\left[\frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X) \right] = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - \mathbb{E}\left[\sum_{i=1}^k \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X) \right]$$

Definition 16 (Exponential family 2)**.** We can write another parameterisation of the exponential family

$$f(x|\boldsymbol{\eta}) = h(x) c^*(\boldsymbol{\eta}) \exp(\boldsymbol{\eta} \cdot \boldsymbol{t}(x))$$

where $\boldsymbol{\eta}$ is called the *natural parameter* and the set $\mathcal{H} = \{\boldsymbol{\eta} : \int_{\mathbb{R}} f(x|\eta) \mathrm{d}x < \infty\}$ is called the *natural parameter space* and is convex.

Remark 5. $\{\boldsymbol{\eta} : \boldsymbol{\eta} = \boldsymbol{w}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{H}$. So there may be more parameterisations here than previously.

The natural parameter provides a convenient mathematical formulation, but sometimes lacks simple interpretation.

Definition 17 (Curved exponential family)**.** A *curved exponential family* distribution is one for which the dimension of $\boldsymbol{\theta}$ is $d < k$. If $d = k$ then we have a *full exponential family.*

3.3 Location and Scale Families

Definition 18 (Location family)**.** Let $f(x)$ be any pdf. The family of pdfs $f(x-\mu)$ for $\mu \in \mathbb{R}$ is called the *location family with standard pdf* $f(x)$ and μ is the *location parameter* of the family.

Definition 19 (Scale family)**.** Let $f(x)$ be any pdf. For any $\sigma > 0$ the family of pdfs $\frac{1}{\sigma}f(x/\sigma)$ is called the *scale family with standard pdf* $f(x)$ and σ is the *scale parameter* of the family.

Definition 20 (Location-Scale family)**.** Let $f(x)$ be any pdf. For $\mu \in \mathbb{R}$ and $\sigma > 0$ the family of pdfs $\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ is called the *location-scale family with standard pdf* $f(x)$; μ is the *location parameter* and σ is the *scale parameter*.

Theorem 17 (Standardisation)**.** Let f be any pdf, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$. Then X is a random variable with pdf $\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.

Remark 6. Probabilities of location-scale families can be computed in terms of their standard variables Z

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

3.4 Inequalities and Identities

Theorem 18 (Chebychev's inequality)**.** Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

$$\mathbb{P}(g(X) \geq r) \leq \frac{\mathbb{E}[g(X)]}{r}.$$

Remark 7. This bound is conservative and almost never attained.

Remark 8 (Markov inequality)**.** The Markov inequality is the special case with $g = \mathbb{I}$.

Theorem 19. Let $X_{\alpha,\beta}$ denote a gamma(α, β) random variable with pdf $f(x|\alpha, \beta)$, where $\alpha > 1$. Then for any constants a and b:

$$\mathbb{P}(a < X_{\alpha,\beta} < b) = \beta(f(a|\alpha,\beta) - f(b|\alpha,\beta)) + \mathbb{P}(a < X_{\alpha-1,\beta} < b)$$

Lemma 1 (Stein's Lemma)**.** Let $X \sim \mathcal{N}(\theta, \sigma^2)$ and let g be a differentiable function with $\mathbb{E}[g'(x)] < \infty$. Then

$$\mathbb{E}[g(X)(X-\theta)] = \sigma^2\mathbb{E}[g'(X)]$$

The proof is just integration by parts.

Remark 9. Stein's lemma is useful for moment calculations

Theorem 20. Let χ^2_p denote a chi squared distribution with p degrees of freedom. For any function $h(x)$,

$$\mathbb{E}[h(\chi^2_p)] = p\mathbb{E}\left[\frac{h(\chi^2_{p+2})}{\chi^2_{p+2}}\right]$$

provided the expressions exist.

Theorem 21. Let $g(x)$ be a function that is bounded at -1 and has finite expectation, then

a. If $X \sim$ Poisson(λ),

$$\mathbb{E}[\lambda g(X)] = \mathbb{E}[Xg(X-1)].$$

b. If $X \sim$ negative-binomial(r, p),

$$\mathbb{E}[(1-p)g(X)] = \mathbb{E}\left[\frac{X}{r+X-1}g(X)\right].$$

4 Multiple Random Variables

4.1 Facts

+ RVs are independent if and only if their pdfs factorise
+ Functions of independent RVs are independent
+ Expectations (and hence mgfs, etc.) of independent RVs factor
+ Independent RVs have vanishing covariance/correlation, but the converse is not true in general.

4.2 Bivariate Relations

Theorem 22 (Conditional Expectation)**.**

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

provided the expectations exist.

Theorem 23 (Conditional variance)**.**

$$\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathbb{E}[X|Y]]$$

provided the expectations exist.

Definition 21 (Covariance)**.**

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

Theorem 24.

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mu_X \mu_Y$$

Theorem 25.

$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}[X, Y]$$

Definition 22 (Correlation)**.**

$$\rho_X Y = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

Remark 10. The correlation measures the strength of *linear* relation between two RVs. It is possible to have strong non-linear relationships but with $\rho = 0$.

We can use an argument similar to the standard proof of Cauchy-Schwarz to show the following

Theorem 26. Let X and Y be any RVs, then

a. $-1 \leq \rho_{XY} \leq 1$,

b. $|\rho_{XY}| = 1$ if and only if there are constants $a \neq 0, b$ such that $\mathbb{P}(Y = aX + b) = 1$. If $|\rho_{XY}| = 1$ then $\mathrm{sign}(\rho) = \mathrm{sign}(a)$.

4.3 Inequalities

4.3.1 Numerical Inequalities

Theorem 27. Let a and b be any positive numbers and let $p, q > 1$ satisfy $1/p + 1/q = 1$, then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality if and only if $a^p = b^q$.

Theorem 28 (Hölder's Inequality)**.** Let X and Y be any random variables and let $p, q > 1$ satisfy $1/p + 1/q = 1$, then

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{1/p}\mathbb{E}[|Y|^q]^{1/q}$$

Corollary 1.

- Cauchy-Schwarz is the special case $p = q = 2$
- $\mathrm{Cov}[X, Y]^2 \leq \sigma_X^2\sigma_Y^2$
- $\mathbb{E}[|X|] \leq \mathbb{E}[|X|^p]^{1/p}$
- *Liapounov's Inequality* $\mathbb{E}[|X|^r]^{1/r} \leq \mathbb{E}[|X|^s]^{1/s}$ where $1 < r < s < \infty$.

4.3.2 Functional Inequalities

Definition 23 (Convex Function)**.** A function $g(x)$ is *convex* on a set S if for all $x, y \in S$ and $0 < \lambda < 1$

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

Strictly convex is when the inequality is strict. g is *concave* if $-g$ is convex.

Lemma 2. $g(x)$ is convex on S if $g''(x) \geq 0\ \forall x \in S$.

Theorem 29 (Jensen's Inequality)**.** If $g(x)$ is convex, then for any random variable X

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

Equality holds if and only if, for every line $a + bx$ that is tangent to $g(x)$ at $x = \mathbb{E}[X]$, $\mathbb{P}\{g(X) = a + bX\} = 1$. (So if and only if g is affine with probability 1.)

Corollary 2.

- $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$
- $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$

5 Properties of a Random Sample