Contents

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 Some Distributions

Definition 14 (Normal Distribution)**.** If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{X}$ has pdf

$$f_{\boldsymbol{X}}(x_1, \ldots, x_k | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

.

3.1.1 Chi-Squared Distribution

Definition 15. The *chi-squared distribution with p degrees of freedom* has pdf

$$\chi^2_p \sim \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty.$$

Theorem 15 (Some facts)**.**

a. If $Z \sim \mathcal{N}(0,1)$ then $Z^2 \sim \chi^2_1$

b. If $X_1, \ldots, X_n$ are independent $X_i \sim \chi^2_{p_i}$ then $X_1 + \cdots + X_n \sim \chi^2_{p_1+\cdots+p_n}$.

3.1.2 Student's t-Distribution

Definition 16 (Student's t-distribution)**.** $T \sim t_p$, a *t-distribution with p degrees of freedom* if it has pdf

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{\sqrt{p\pi}} \frac{1}{(1+t^2/p)^{(p+1)/2}}, \quad t \in \mathbb{R}$$

If $p = 1$ then this is the Cauchy distribution.

Remark 4. If $X_1, \ldots, X_n$ are a random sample from $\mathcal{N}(\mu, \sigma^2)$ then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

This is often taken as the definition. Note that the denominator is independent of the numerator.

Lemma 1 (Moments and mgf of t-distribution)**.** Student's t has no mgf because it does not have moments of all orders: t_p has only $p-1$ moments. If $T_p \sim t_p$ then

$$\mathbb{E}[T_p] = 0 \quad p > 1$$
$$\mathrm{Var}[T_p] = \frac{p}{p-2} \quad p > 2$$

3.1.3 Snedcor's F-Distribution

Definition 17 (Snedcor's F-distribution)**.** A random variable $X \sim F_{p,q}$ has *F-distribution with p and q degrees of freedom* if its pdf is

$$f_X(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} (p/q)^{p/2} \frac{x^{(p/2)-1}}{(1+px/q)^{(p+q)/2}}, \quad 0 < x < \infty.$$

Remark 5. If $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m$ is an independent random sample from $\mathcal{N}(\mu_Y, \sigma_Y^2)$, then

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.$$

This is often taken as the definition.

Theorem 16 (Some facts)**.**

a. $X \sim F_{p,q} \implies 1/X \sim F_{q,p}$

b. $X \sim t_q \implies X^2 \sim F_{1,q}$

c. $X \sim F_{p,q} \implies \frac{(p/q)X}{1+(p/q)X} \sim \text{beta}(p/2, q/2)$

3.1.4 Multinomial Distribution

Definition 18 (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 6. The marginal distributions have $X_i \sim \text{binomial}(m, p_i)$.

Theorem 17 (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 19 (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 18. 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 20 (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 7. $\{\boldsymbol{\eta} : \boldsymbol{\eta} = \boldsymbol{w}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{H}$. So there may be more parameterisations here than previously.

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

Definition 21 (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 22 (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 23 (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 24 (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 19 (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 8. Probabilities of location-scale families can be computed in terms of their standard variables Z

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

3.4 Inequalities and Identities

Theorem 20 (Chebychev's inequality)**.** Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

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

Remark 9. This bound is conservative and almost never attained.

Remark 10 (Markov inequality)**.** The Markov inequality is the special case with $g = \mathbb{I}$.

Theorem 21. 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 2 (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 11. Stein's lemma is useful for moment calculations

Theorem 22. 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 23. Let $g(x)$ be a function that is bounded at -1 and has finite expectation, then

a. If $X \sim \text{Poisson}(\lambda)$,
$$\mathbb{E}[\lambda g(X)] = \mathbb{E}[Xg(X-1)].$$

b. If $X \sim \text{negative-binomial}(r, p)$,
$$\mathbb{E}[(1-p)g(X)] = \mathbb{E}\left[\frac{X}{r+X-1}g(X)\right].$$

4 Multiple Random Variables

4.1 Facts

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

4.2 Bivariate Relations

Theorem 24 (Conditional Expectation)**.**

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

provided the expectations exist.

Theorem 25 (Conditional variance)**.**

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

provided the expectations exist.

Definition 25 (Covariance)**.**

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

Theorem 26.

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

Theorem 27.

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

Definition 26 (Correlation)**.**

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

Remark 12. 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 28. 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 29. 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 30 (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 27 (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 3. $g(x)$ is convex on S if $g''(x) \geq 0\ \forall x \in S$.

Theorem 31 (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

Definition 28 (Random Sample)**.** A collection of random variables $X_1, \ldots, X_n$ is a *random sample of size n from population* $f(x)$ if they are iid with pdf/pmf f(x).

Definition 29 (Statistic)**.** $Y = T(X_1, \ldots, X_n)$ is a *statistic* if the domain of T contains the sample space of $(X_1, \ldots, X_n)$. The distribution of Y is the *sampling distribution of* Y.

Remark 13. A statistic is any function of the data. The only restriction is that the statistic is not also a function of some other parameters.

Definition 30 (Sample Mean and Sample Variance)**.** The *sample mean* $\bar{X}$ and *sample variance* S^2 of a random sample $X_1, \ldots, X_n$ are, respectively,

- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

The *sample standard deviation* is $S = \sqrt{S^2}$.

Theorem 32 (Some facts)**.** Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and finite variance σ^2, then

a. $\mathbb{E}[\bar{X}] = \mu$

b. $\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$

c. $\mathbb{E}[S^2] = \sigma^2$.

Theorem 33. Take $x_1, \ldots, x_n \in \mathbb{R}$ and let $\bar{x}$ be their mean. Then,

a. $\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

b. $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Theorem 34. Let $X_1, \ldots, X_n$ be a random sample from population $f(x|\boldsymbol{\theta})$ belonging to an exponential family

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

Define the statistics

$$T_i(X_1, \ldots, X_n) = \sum_{j=1}^n t_i(X_j), \quad i = 1, \ldots, k.$$

Then if the set $\{(w_1(\boldsymbol{\theta}), \ldots, w_n(\boldsymbol{\theta}), \theta \in \Theta\}$ contains and open subset of $\mathbb{R}^k$ then the distribution of $(X_1, \ldots, X_n)$ is an exponential family of the form

$$f(u_1, \ldots, u_n|\boldsymbol{\theta}) = H(u_1, \ldots, u_n)c(\boldsymbol{\theta})^n \exp\left(\sum_{i=1}^k w_i(\boldsymbol{\theta}) u_i\right).$$

Remark 14. The open condition eliminates curved exponential families from this result.

5.1 Sampling from the Normal Distribution

Theorem 35. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma)$ distribution. Then,

- $\bar{X}$ and S^2 are independent
- $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Lemma 4 (Covariance and independence)**.** In the case of samples from a multivariate normal

- Independence $\iff$ vanishing covariance
- Pairwise independence $\iff$ independence

5.2 Convergence Concepts

5.2.1 Convergence in Probability

Definition 31. A sequence of random variables $\{X_i : i \in \mathbb{N}\}$ *converges in probability* to a random variable X if $\forall \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

(or equivalently if $\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1$).

Theorem 36 (Weak Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be iid RVs with mean μ and $\mathbb{E}[|X_i|] < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n$. Then the sequence $\bar{X}_n$ converges in probability to μ. That is, $\forall \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mu| < \epsilon) = 1$$

Proof is using Chebychev's inequality.

Theorem 37. If $X_1, X_2, \ldots$ converges in probability to X and h is a continuous function, then $h(X_1), h(X_2), \ldots$ converges in probability to $h(X)$.

5.2.2 Almost Sure Convergence

Definition 32. A sequence of random variables $X_1, X_2, \ldots$ *converges almost surely* to a random variable X if $\forall \epsilon > 0$

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\bar{X}_n - X\right| < \epsilon\right) = 1.$$

Comments.

+ Almost sure convergence is much stronger than convergence in probability. Convergence in probability states that the sequence of measures of the sets on which the sequence has finite difference from its the limit converges to 0. Almost sure convergence states that any place where the sequence has finite difference from its limit must have measure 0. It's like a sequence of integrals converging vs. whether the integrands converge.

+ Almost sure convergence implies convergence in probability but not the other way around.

Theorem 38. If a sequence converges in probability then it is possible to find a subsequence that converges almost surely.

Theorem 39 (Strong Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be iid RVs with mean μ and $\mathbb{E}[|X_i|] < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n$. Then the sequence $\bar{X}_n$ converges almost surely to μ. That is, $\forall \epsilon > 0$

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\bar{X}_n - \mu\right| < \epsilon\right) = 1.$$

5.2.3 Convergence in Distribution

Definition 33. A sequence of random variables $X_1, X_2, \ldots$ *converges in distribution* to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points where F_X is continuous.

Remark 15. Here it is really the cdfs that converge, rather than the random variables. In this way convergence in distribution differs from the previous two concepts.

Theorem 40. Convergence in probability implies convergence in distribution

Theorem 41 (Central Limit Theorem)**.** Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then $\forall x \in \mathbb{R}$,

$$\lim_{n\to\infty} G_n(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \mathrm{d}y.$$

That is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in distribution to the standard normal.

Theorem 42 (Slutsky's Theorem)**.** If $X_n \to X$ in distribution and $Y_n \to a$ in probability with a constant, then

a. $Y_n X_n \to aX$ in distribution

b. $X_n + Y_n \to X + a$ in distribution

Remark 16. This tells us, for instance, that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \to \mathcal{N}(0, 1)$$

in distribution, since we know that $S_n \to \sigma$ in probability.

5.2.4 The Delta Method

If we are interested in the convergence of some function of a sequence of RVs, rather than the RVs themselves, then we can use the Delta Method (follows from an application of Taylor's theorem and Slutsky's theorem).

Theorem 43 (Delta Method)**.** Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is non-zero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2 g'(\theta)^2)$$

in distribution.

Remark 17. There exists a corresponding multivariate result.

If $g'(\theta) = 0$ then we take the next term in the Taylor series.

Theorem 44 (Second Order Delta Method)**.** Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is non-zero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2 g'(\theta)^2)$$

in distribution.

5.3 Generating A Random Sample

Definition 34 (Direct Method)**.** A *Direct Method* of generating a random sample uses the probability integral transform to map draws from a uniform$(0,1)$ random variable to draws from the distribution of interest.

The Probability Integral Transform states that if X has continuous cdf $F_X(x)$ then

$$F_X(X) \sim \text{uniform}(0,1).$$

Definition 35 (Accept-Reject Algorithm)**.** Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$ where f_Y and f_V have common support with

$$M = \sup_y f_Y(y)/f_V(y) < \infty.$$

To generate a random variable $Y \sim f_Y$:

a. Generate $U \sim \text{uniform}(0,1)$, $V \sim f_V$ independent.

b. If $U < \frac{1}{M} f_Y(V)/f_V(V)$, return V as a sample of Y; otherwise go back to (a.).

Remark 18.

- It is typical to call V the *candidate density* and Y the *target density.*
- One would normally try to choose a candidate density with heavier tails than the target density (e.g. Cauchy and normal) to ensure that the tails of the target are well represented. If the target has heavy tails, however, it can be hard to find a candidate that results in finite M. In this case people turn to MCMC methods.
- Note that $\mathbb{P}(\texttt{terminate}) = 1/M$. The number of trials to generate one sample of Y is therefore geometric$(1/M)$, with M the expected number of trials.
- The intuition behind this algorithm is that if we consider placing the density of a random variable Y in a box (2d for simplicity) with coordinates (v,u), we express the cdf of Y using $V, U \sim \text{uniform}(0,1)$

  $$\mathbb{P}(Y \le y) = \mathbb{P}(V \le y | U \le \frac{1}{c} f_Y(V))$$

  where $c = \sup_y f_Y(y)$. In the actual algorithm we take $U \sim \text{uniform}(0,1)$ and V to be an RV that has common support with Y.

6 Principles of Data Reduction

In this chapter, we explore how we can use functions of a sample $\boldsymbol{X}$ to make inferences about an unknown parameter (of the distribution of the sample) θ.

Definition 36 (Statistic)**.** A statistic is any function of the data.

A statistic T forms a partition of the sample space $\mathcal{X}$ according to its image, $\mathcal{T} = \{t : \exists \boldsymbol{x} \in \mathcal{X} \text{ s.t. } t = T(\boldsymbol{x})\}$. In this way a statistic provides a method of data reduction. An experimenter who observes only T will treat as equal two samples $\boldsymbol{x}, \boldsymbol{y}$ for which $T(\boldsymbol{x}) = T(\boldsymbol{y})$.

6.1 The Sufficiency Principle

Definition 37 (Sufficient Statistic)**.** A statistic $T(\boldsymbol{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\boldsymbol{X}$ given $T(\boldsymbol{X})$ does not depend on θ.

Remark 19. We ignore the fact that all points have 0 probability for continuous distributions.

Definition 38 (The Sufficiency Principle)**.** If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through $T(\boldsymbol{X})$.

Theorem 45. If $p(\boldsymbol{x}|\theta)$ is the pmf/pdf of the sample $\boldsymbol{X}$ and $q(t|\theta)$ is the pmf/pdf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if $\forall \boldsymbol{x} \in \mathcal{X}$, $p(\boldsymbol{x}|\theta)/q(t|\theta)$ is constant as a function of θ.

Remark 20 (Niceness of the exponential family)**.** It turns out that outside of the exponential family, it is rare to have a sufficient statistic that is of smaller dimension than the size of the sample.

Theorem 46 (Factorisation Theorem)**.** Let $f(\boldsymbol{x}|\theta)$ denote the pmf/pdf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exists functions $g(t|\theta)$ and $h(\boldsymbol{x})$ such that $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall \theta \in \Theta$

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x}).$$

Remark 21. This theorem shows that the identity is a sufficient statistic. It is straightforward to show from this that any bijection of a sufficient statistic is a sufficient statistic.

Theorem 47. Let $X_1, \ldots, X_n$ be iid observations from a pmf/pdf $f(x|\theta)$ from an exponential family

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

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \leq k$. Then $\boldsymbol{T}(\boldsymbol{X})$ defined by

$$T_i(\boldsymbol{X}) = \sum_{j=1}^{k} t_i(\boldsymbol{X}_j)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Definition 39 (Minimal sufficient statistic)**.** A sufficient statistic $T(\boldsymbol{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\boldsymbol{X})$, T is a function of T'.

Remark 22. By 'function of' we mean that if $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$ then $T(\boldsymbol{x}) = T(\boldsymbol{y})$ – T varies with respect to X only insofar as it varies with T'. This means that each tile of the partition of the sample space according to the image of T' is a subset of some tile in the partition according to T. This means that minimal sufficient statistics provide the coarsest possible tiling of the sample space and thus are the sufficient statistics that provide the greatest data reduction.

Theorem 48. Let $f(\boldsymbol{x}|\theta)$ be the pmf/pdf of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{X})$ such that $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ is constant as a function of θ if an only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Definition 40 (Necessary Statistic)**.** A statistic is *necessary* if it can be written as a function of every sufficient statistic.

Theorem 49. A statistic is minimal sufficient if and only if it is a necessary and sufficient statistic.

Definition 41 (Ancillary Statistic)**.** A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic*.

Definition 42 (First Order Ancillary)**.** A statistic $V(\boldsymbol{X})$ is *first order ancillary* if $\mathbb{E}[V(\boldsymbol{X})]$ is independent of θ.

Definition 43 (Complete Statistic)**.** Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\boldsymbol{X})$. The family of probability distributions is called *complete* if for every (measurable) function g

$$\mathbb{E}[g(T)] = 0 \ \forall \theta \implies \mathbb{P}(g(T) = 0) = 1 \ \forall \theta.$$

Equivalently, $T(\boldsymbol{X})$ is called a *complete statistic*.

(It is left unsaid that the function g must be independent of θ.)

Theorem 50 (Basu's Theorem)**.** If $T(\boldsymbol{X})$ is a complete and minimal sufficient statistic, then $T(\boldsymbol{X})$ is independent of every ancillary statistic.

Theorem 51 (Complete statistics in the exponential family)**.** Let $X_1, \ldots, X_n$ be iid observations from a pmf/pdf $f(x|\theta)$ from an exponential family

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

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \leq k$. Then $\boldsymbol{T}(\boldsymbol{X})$ defined by

$$T_i(\boldsymbol{X}) = \sum_{j=1}^{k} t_i(\boldsymbol{X}_j)$$

is complete if $\{(w_1(\boldsymbol{\theta}), \ldots, w_n(\boldsymbol{\theta}))\}$ contains an open set in $\mathbb{R}^k$.

Remark 23. The open set criteria excludes curved exponential families.

Theorem 52. If a minimal sufficient statistic exists, then every complete statistic is minimal sufficient.

6.2 The Likelihood Principle

Definition 44 (Likelihood)**.** Let $f(\boldsymbol{x}|\theta)$ denote the pmf/pdf of the sample $\boldsymbol{X}$. Then the *likelihood function*, given an observation $\boldsymbol{X} = \boldsymbol{x}$, is

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

as a function of θ.

Definition 45 (Likelihood Principle)**.** If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $L(\theta|\boldsymbol{x})$ is proportional to $L(\theta|\boldsymbol{y})$, that is, there exists a constant $C(\boldsymbol{x}, \boldsymbol{y})$ such that

$$L(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y}) L(\theta|\boldsymbol{y}) \quad \forall \theta,$$

then the conclusions drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical.

6.3 A Slightly More Formal Construction

Definition 46 (Experiment)**.** We define an *experiment* E to be a triple $(\boldsymbol{X}, \theta, f(\boldsymbol{x}|\theta))$, where $\boldsymbol{X}$ is a random vector with pmf/pdf f.

Knowing what experiment E was performed, an experimenter will observer $\boldsymbol{X} = \boldsymbol{x}$. The conclusions they draw about θ will be denoted $\text{Ev}(E, \boldsymbol{x})$, which stands for *the evidence about* θ *arising from* E *and* $\boldsymbol{x}$.

Definition 47 (Formal Sufficiency Principle)**.** Consider an experiment $E = (\boldsymbol{X}, \theta, f(\boldsymbol{x}|\theta))$ and suppose that $T(\boldsymbol{X})$ is a sufficient statistic for θ. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are sample points satisfying $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then $\text{Ev}(E, \boldsymbol{x}) = \text{Ev}(E, \boldsymbol{y})$.

Definition 48 (Conditionality Principle)**.** Suppose that $E_1 = \{X_1, \theta, f_1(x_1|\theta)\}$ and $E_2 = \{X_2, \theta, f_2(x_2|\theta)\}$ are two experiments, where only the unknown parameter θ need be common between the two experiments. Consider the mixed experiment in which the random variable J is observed, where $\mathbb{P}(J = 1) = \mathbb{P}(J = 2) = \frac{1}{2}$ (independent of $\boldsymbol{X}_1, \boldsymbol{X}_2, \theta$), and then the experiment E_J is performed. Formally, the experiment performed is $E^* = (\boldsymbol{X}^*, \theta, f(\boldsymbol{x}^*|\theta))$, where $\boldsymbol{X}^* = (j, \boldsymbol{X})j$ and $f^*(\boldsymbol{x}^*|\theta) = f^*((j, \boldsymbol{x}_j)|\theta) = \frac{1}{2} f_j(\boldsymbol{x}_j|\theta)$. Then

$$\text{Ev}(E^*, (j, \boldsymbol{x}_j)) = \text{Ev}(E_j, \boldsymbol{x}_j).$$

That is, information about θ depends only on the experiment run (not on the fact that the particular experiment was chosen).

Definition 49. Formal Likelihood Principle Suppose that we have two experiments, $E_1 = (\boldsymbol{X}_1, \theta, f_1(\boldsymbol{x}_1|\theta))$ and $E_2 = (\boldsymbol{X}_2, \theta, f_2(\boldsymbol{x}_2|\theta))$ where the unknown parameter θ is the same in both experiments. Suppose that $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ are sample points from E_1 and E_2 respectively, such that

$$L(\theta|\boldsymbol{x}_2^*) = CL(\theta|\boldsymbol{x}_1^*)$$

for all θ and for some constant $C(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ that is independent of θ. Then

$$\text{Ev}(E_1, \boldsymbol{x}_1^*) = \text{Ev}(E_2, \boldsymbol{x}_2^*).$$

Remark 24. Note that this is more general from the other likelihood principle since it concerns two experiments (that we can of course set to be equal).

Corollary 3 (Likelihood Principle Corollary)**.** If $E = (\boldsymbol{X}, \theta, f(\boldsymbol{x}|\theta))$ is an experiment then $\text{Ev}(E, \boldsymbol{x})$ should depend on E and $\boldsymbol{x}$ only through $L(\theta|\boldsymbol{x})$.

Theorem 53 (Birnbaum's Theorem)**.** The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.

Remark 25. Many common statistical procedures violate the Formal Likelihood Principle – the topic of the applicability of these principles is not settled. For instance, checking the residuals of a model (to grade the model) violates the Sufficiency Principle. These notions are model dependent, so may not be applicable until *after* we have decided on a model.

6.4 The Equivariance Principle

Definition 50 (Measurement Equivariance)**.** Inferences should not depend on the measurement scale used.

Definition 51 (Formal Invariance)**.** If two inference problems have the same formal structure, in terms of the mathematical model used, then the same inference procedure should be used, regardless of the physical realisation.

Definition 52 (Equivariance Principle)**.** If $\boldsymbol{Y} = g(\boldsymbol{X})$ is a change of measurement scale such that the model $\boldsymbol{Y}$ has the same formal structure as the model $\boldsymbol{X}$, then an inference procedure should be both measurement equivariant and formally invariant.

Definition 53. Let $\mathcal{F} = \{f(\boldsymbol{x}|\theta) : \theta \in \Theta\}$ be a set of pdfs or pmfs for $\boldsymbol{X}$, and let $\mathcal{G}$ be group of transformations on the sample space $\mathcal{X}$. Then $\mathcal{F}$ is *invariant under the group* $\mathcal{G}$ if for every $\theta \in \Theta$ and $g \in \mathcal{G}$ there exists a unique $\theta' \in \Theta$ such that $\boldsymbol{Y} = g(\boldsymbol{X})$ has the distribution $f(\boldsymbol{y}|\theta')$ if $\boldsymbol{X}$ has the distribution $f(\boldsymbol{x}|\theta')$.

7 Point Estimation

Definition 54 (Point Estimator)**.** A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator. An *estimate* is a realised value $w(x_1, \ldots, x_n)$.

Note that there is the implicit restriction that the estimator is not a function of the parameter you are trying to estimate.

7.1 Methods of Finding Estimators

7.1.1 Method of Moments

The method of moments is performed using a random sample $X_1, \ldots, X_n$ from a population with unknown parameters $\theta_1, \ldots, \theta_k$ by computing the first k empirical moments $m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ and matching them with the corresponding moments of the population $\mu_k' = \mathbb{E}[X^t]$.

$$\begin{aligned} m_1 &= \mu_1'(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu_k'(\theta_1, \ldots, \theta_k) \end{aligned}$$

From this you get k simultaneous equations that you can use to solve for the parameters of the population.

7.1.2 Maximum Likelihood Estimators

Definition 55 (Maximum Likelihood Estimator)**.** For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which the likelihood $L(\theta|\boldsymbol{x})$ attains its maximum as a function of θ, with $\boldsymbol{x}$ held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$.

Suppose we want to find the MLE for some function of the parameter $\tau(\theta)$.

Definition 56 (Induced Likelihood)**.** Given some function of the parameter $\tau(\theta)$, we define the *induced likelihood function* L^* by

$$L^*(\eta|\boldsymbol{x}) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|\boldsymbol{x}).$$

The value $\hat{\eta}$ that maximises $L^*(\eta|\boldsymbol{x})$ will be called the MLE of $\eta = \tau(\theta)$. It can be seen that the maxima of L^* and L coincide.

Theorem 54 (Invariance Property of MLEs)**.** IF $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Remark 26.

1) The MLE can be an unstable function of the data.

2) When verifying maxima for multi-dimensional problems try to avoid going down the hessian route, which can be tedious.

7.1.3 Bayes Estimators

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\boldsymbol{x}|\theta)$, then the posterior distribution of θ given the sample $\boldsymbol{x}$ is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})},$$

where $m(\boldsymbol{x})$ is the marginal distribution of the sample $\boldsymbol{X}$

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta.$$

Definition 57 (Conjugate Priors)**.** Let $\mathcal{F}$ denote the class of pdfs/pmfs $f(\boldsymbol{x}|\theta)$ (indexed by θ). A class Π of prior distribution is a *conjugate family* for $\mathcal{F}$ if, $\forall f \in \mathcal{F}$, $\forall$ priors $\in \mathcal{F}$ and $\forall \boldsymbol{x} \in \mathcal{X}$, the posterior distribution is in Π.

Note that this relation is not said to be symmetric.

Remark 27 (Some Examples)**.**

- $\mathcal{N}$ is self-conjugate as a family.
- Beta distribution is conjugate to binomial.

7.2 Methods of Evaluating Estimators

Definition 58 (Mean Squared Error)**.** The *mean squared error* (MSE) of an estimator W of a parameter θ is defined by $\mathbb{E}[(W-\theta)^2]$.

Lemma 5 (Bias-Variance Decomposition)**.**

$$\mathbb{E}[(W-\theta)^2] = \mathrm{Var}[W] + (\mathbb{E}[W]-\theta)^2 = \mathrm{Var}[W] + \mathrm{Bias}[W]^2$$

Definition 59 (Bias)**.** The *bias* of a point estimator W of a parameter θ is given by

$$\mathrm{Bias}[W] = \mathbb{E}[W] - \theta.$$

An estimator whose bias is 0 is called an *unbiased estimator* and has $\mathbb{E}[W] = \theta\ \forall\theta$.

Remark 28.

- Clearly, if an estimator is unbiased, then its MSE is equal to its variance.
- The MSE makes sense for location parameters but not so much for scale parameters, since it is symmetric and scale parameters have a natural floor at 0.
- The MSE may be a function of the thing you're trying to estimate. So which estimator you choose as being the 'best' may depend on the range you expect the parameter to lie within.

7.2.1 Best Unbiased Estimators

Definition 60 (Best Unbiased Estimator)**.** An estimator W^* is *best unbiased estimator* of of $\tau(\theta)$ if it satisfies $\mathbb{E}[W^*] = \tau(\theta)\ \forall\theta$, and for any other estimator with $\mathbb{E}[W] = \tau(\theta)\ \forall\theta$ we have $\mathrm{Var}[W*] \leq \mathrm{Var}[W]\ \forall\theta$. We also call W^* the *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

Remark 29. Suppose that we are trying to estimate θ and consider the class of estimators

$$\mathcal{C}_\tau = \{W : \mathbb{E}[W] = \tau(\theta)\}.$$

All estimators in this class have the same bias, so we can compare their MSEs by comparing their variances alone. (So the best estimator in this class is just the minimum variance one.) This means that the considerations of this chapter can be applied to classes like $\mathcal{C}_\tau$, even if $\tau(\theta) \neq \theta$.

The best unbiased estimator, if it exists, could be hard to find. The following lower bound at least gives a stopping criterion to our search.

Theorem 55 (Cramér-Rao Inequality)**.** Let $X_1, \ldots, X_n$ be a sample with pdf $f(\boldsymbol{x}|\theta)$ and let $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ be any estimator with finite variance satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[W(\boldsymbol{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} \left(W(\boldsymbol{x}) f(\boldsymbol{x}|\theta)\right) \mathrm{d}\boldsymbol{x}.$$

Then

$$\mathrm{Var}[W(\boldsymbol{X})] \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[W(\boldsymbol{X})]\right)^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right]}.$$

The proof of this considers the correlation between the gradient of the log likelihood and the statistic. Note that the sample in the above theorem is not necessarily iid.

Remark 30. This hold for discrete distributions too, replacing integrals with sums.

Corollary 4. If the random sample $\mathcal{X}_1, \ldots, X_n$ is iid then the result becomes

$$\mathrm{Var}[W(\boldsymbol{X})] \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[W(\boldsymbol{X})]\right)^2}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

Definition 61 (Fisher Information)**.** The quantity $\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right]$ is called the *Fisher Information* of the sample $\boldsymbol{X}$.

Lemma 6. If $f(x|\theta)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\right] \mathrm{d}x$$

(as is true for an exponential family), then the Fisher information can be written

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right].$$

Remark 31. Even if the Cramér-Rao bound is applicable, it may not be sharp – there may not be an estimator that attains this bound.

Lemma 7 (Attainment)**.** Let $X_1, \ldots, X_n$ be iid $X \sim f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér-Rao Theorem. Let $L(\theta|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\boldsymbol{X})$ is any unbiased estimator of $\tau(\theta)$ then $W(\boldsymbol{X})$ attains the Cramér-Rao lower bound if and only if $\exists a(\theta)$ such that

$$a(\theta)[W(\boldsymbol{X}) - \tau(\theta)] = \frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{x}).$$

7.2.2 Sufficiency and Unbiasedness

Theorem 56 (Rao-Blackwell)**.** Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\phi(T) = \mathbb{E}[W|T]$. Then $\mathbb{E}[\phi(T)] = \tau(\theta)$ and $\text{Var}[\phi(T)] \leq \text{Var}[W]$ $\forall\theta$. That is, $\phi(T)$ is a *uniformly better unbiased estimator* of $\tau(\theta)$.

Remark 32.

- Conditioning on any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only functions of a sufficient statistic when looking for best unbiased estimators.
- The proof doesn't require that the statistic we condition on is sufficient, but if it isn't then the resulting quantity will probably depend on the parameter we are trying to estimate.

Theorem 57. If W is a best unbiased estimator of $\tau(\theta)$ then W is unique.

The following theorem is mostly useful to show that a given estimator *isn't* best unbiased.

Theorem 58. If $\mathbb{E}[W] = \tau(\theta)$, then W is the best unbiased estimator of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.

The idea comes from considering $\phi_a = W + aU$ where $\mathbb{E}[U] = 0$, then considering the variance.

Remark 33 (Unbiased estimator of 0)**.** Note that an unbiased estimator of 0 is simply noise (one should estimate 0 with 0). If an estimator can be improved by adding noise, then it is probably defective.

We are now in a position such that, if we can characterise all of the unbiased estimators of 0 then we can check if a given estimator is best unbiased. In general this is not easy and requires conditions on the distribution. However, if a distribution is complete then it admits no unbiased estimators of 0 other than 0 itself (by definition), so we will be done.

Note that due to the Rao-Blackwell theorem, only the distribution of the sufficient statistic needs to be complete (not the underlying population distribution).

Theorem 59. Let T be a complete sufficient statistic for a parameter θ and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

Theorem 60 (Lehmann-Scheffé)**.** Unbiased estimators based on complete sufficient statistics are unique.