Auxiliary notes on Statistical Inference

Daniel Lin; Professor: George Deligiannidis

January 22, 2024

This note aims to explain the more abstract concepts in the SB2.1 Foundations of Statistical Inference course or supply some useful intuitions to enhance your understanding. Please use it in line with the lecture notes.

Contents

1 Symbols **3**

2 Distributions and their relations **3**

3 Exponential Families **4**
3.1 Affine Geometry . . . 4
3.2 Curved Exponential Family . . . 6

4 Sufficiency **7**
4.1 Minimality . . . 8

5 Quality of Estimators **9**

6 Likelihood-based Estimation **10**
6.1 Likelihood . . . 10
6.2 Intuitions behind Fisher's information . . . 11
6.3 MVUE and Cramer-Rao Lower Bound . . . 13
6.4 Complete Statistics . . . 14

7 Introduction to Bayesian Statistics **16**
7.1 Choice of Priors . . . 17
7.2 Non-informative prior . . . 18
7.2.1 Jeffrey Prior . . . 19
7.2.2 Entropy and Max Entropy Prior . . . 19
7.3 Posterior Predictive distribution . . . 20

8 Bayesian Decision Theory **21**
8.1 Risks and "good" decision rules . . . 21
8.2 Relationships between the "good" rules . . . 23
8.2.1 Bayes rule and Minimax rule . . . 23
8.2.2 Bayes rule and admissibility . . . 24
8.2.3 Minimax and Admissibility . . . 25
8.3 The posterior approach . . . 25

9 Empirical Bayes Methods **27**
9.1 Robbin's formulae . . . 27
9.2 Parametric Empirical Bayes . . . 28
9.2.1 James-Stein Estimator . . . 28
9.2.2 Connection to PEB . . . 30

10 Hypothesis Testing **31**
10.1 Quality of Test . . . 31
10.2 UMP for simple Hypothesis . . . 33
10.3 UMP for one-sided test . . . 35
10.4 Bayesian Hypothesis Testing . . . 37
10.4.1 Decision Theory and Hypothesis Testing . . . 39
10.5 Two-sided Hypothesis test . . . 41

11 Ending **44**

1 Symbols

Symbols used in this note have the following default meaning unless stated otherwise.

Symbol	Meaning
$\mathbb{R}$	the set of real numbers
$\mathscr{X}$	sample space
Θ	parameter space
X, Y, Z	random variables
x, y, z	realisations
$\mathcal{P}(X)$	the power set of set X
T, T_i or $T(x)$, $T_i(x)$	statistics (they are functions!)
η_i	canonical parameters of exponential family
$P_\theta(X)$	the probability distribution of X w.r.t. parameter θ
$E_\theta(X)$ or $E_\theta[X]$	the expectation of X w.r.t. parameter θ
$\mathrm{Var}_\theta(X)$ or $\mathrm{Var}_\theta[X]$	the variance of X w.r.t. parameter θ
$\mathrm{Cov}_\theta(X)$	the covariance of X w.r.t. parameter θ
$\hat{\theta}_{\mathrm{MLE}}$	Maximum likelihood estimator for θ
$I_X(\theta)$	Fisher information of θ w.r.t X
MVUE	minimal variance unbiased estimator
$e(T, \theta)$	efficiency of unbiased estimator T of $g(\theta)$
$\succ, \succeq, \prec, \preceq$	Loewner order (an ordering defined for matrices)
$\hat{\gamma}_T$	The Rao-Blackwell estimator of γ based on statistics T
π	distribution of the parameter θ (used in Bayesian statistics)
$l(\theta, x)$	the log-likelihood function, also written as $l(\theta; x)$ or $l(\theta \vert x)$
LP	likelihood principle
ϕ and Φ	PDF and CDF of the standard normal distribution
$L(\theta, a) = L(a, \theta)$	Loss function (used in decision theory)
$R(\theta, \Delta)$	risk function
$r(\pi, \Delta)$	Bayes risk
Δ_{Bayes}	Bayes rule(estimator)
Δ^*	minimax rule
$\Lambda_\pi(x, \Delta)$	the expected posterior loss, also called posterior risk
C	critical region
ϕ	a statistical test
$w_\phi(\theta)$	power function of test ϕ
α	size of a test/significance level
$\Lambda(x)$	the likelihood ratio (θ_1 against θ_0)
LRT	likelihood ratio test
UMP	Uniformly most powerful
UMPU	Uniformly most powerful unbiased
B	Bayes factor

2 Distributions and their relations

The statistics cookbook http://statistics.zone/ contains a comprehensive list of distributions and formulas for many areas of statistics. Distributions are not segregated species, they are closely related, as can be seen by the network of univariate distributions in the cookbook. But I will list the most commonly used ones here:

Sum of Distributions

- $\{X_i\}_{i=1,\cdots,n} \sim \mathrm{Bernoulli}(\theta)$ i.i.d. $\Leftrightarrow X = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n, \theta)$
- $\{X_i\}_{i=1,\cdots,n} \sim \mathrm{Poisson}(\lambda/n)$ i.i.d. $\Leftrightarrow X = \sum_{i=1}^n X_i \sim \mathrm{Poisson}(\lambda)$
- $\{X_i\}_{i=1,\cdots,n} \sim \mathrm{Geometric}(\theta)$ i.i.d. $\Leftrightarrow X = \sum_{i=1}^n X_i \sim \mathrm{Negative\ Binomial}(n, \theta)$
- $\{X_i\}_{i=1,\cdots,n} \sim \mathrm{Gamma}(1, \lambda)$ i.i.d. $\Leftrightarrow X = \sum_{i=1}^n X_i \sim \mathrm{Gamma}(n, \lambda)$

- $\{Z_i\}_{i=1,\cdots,n} \sim \mathrm{N}(0,1)$ i.i.d. $\Leftrightarrow X = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$
- $\{X_i\}_{i=1,\cdots,n} \sim \chi_k^2$ i.i.d. $\Leftrightarrow X = \sum_{i=1}^n X_i \sim \chi_{kn}^2$

Linearity of Normal Distribution
If $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2, \sigma_2^2), \quad X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2, \sigma_2^2)$$

If $X \sim N(\mu, \sigma^2)$, then $kX \sim N(k\mu, k^2\sigma^2)$

Linearity of Gamma distribution
$\mathrm{Gamma}(a, \lambda)$+ $\mathrm{Gamma}(b, \lambda) = \mathrm{Gamma}(a + b, \lambda)$
$\mathrm{Gamma}(a, \lambda) = \mathrm{Gamma}(a, 1)/\lambda$

Exponential and Gamma
$\mathrm{Exp}(\lambda) = \mathrm{Gamma}(1, \lambda)$. So if X_i are i.i.d $\mathrm{Exp}(\lambda)$, then $\sum_i X_i \sim \mathrm{Gamma}(n, \lambda)$.

Note $\chi_n^2 = \mathrm{Gamma}(n/2, 1/2)$. So $\chi_2^2 = \mathrm{Exp}(1/2)$ and $\mathrm{Gamma}(n, \lambda) = \chi_{2n}^2/(2\lambda)$.

Dirichlet and Gamma
If $\{U_i\}_{i=1,\cdots,k} \sim \mathrm{Gamma}(\alpha_i, \beta)$ i.i.d., then

$$\frac{1}{\sum_{i=1}^k U_i} \sim \mathrm{Dirichlet}(\alpha_1, \cdots, \alpha_k)$$

3 Exponential Families

Please find introductory notes on exponential family in the "Auxiliary Notes for Graphical Models".

3.1 Affine Geometry

There are k canonical parameters for the exponential family. Later, we will see that k decides the dimension of sufficient statistics related to the exponential family. A lower dimension is easier to deal with, so reducing k becomes important. Linear algebra is good at this. In particular, we need some affine geometry [8].

From linear combination to affine combination

Recall that the set of linear combinations of a set of vectors $\{\boldsymbol{v}_i\}_{i=1,\cdots,n}$ is defined by

$$\left\{\sum c_i \boldsymbol{v}_i \mid c_i \in \mathbb{R}\right\}$$

an affine combination requires an additional condition that coefficients c_i sum up to 1. It characterises all points on an affine hyperplane generated by $\{\boldsymbol{v}_i\}_{i=1,\cdots,n}$. (a hyperplane that does not necessarily go through the origin) Subtracting $\boldsymbol{v}_1$ from all vectors places the hyperplane at the origin. So any vector $\boldsymbol{x}$ is on the affine hyperplane iff

$$\boldsymbol{x} - \boldsymbol{v}_1 = \sum_2^n c_i(\boldsymbol{v}_i - \boldsymbol{v}_1)$$

for some scalar constants c_i. Equivalently,

$$\boldsymbol{x} = \left(1 - \sum_2^n c_i\right)\boldsymbol{v}_1 + \sum_2^n c_i \boldsymbol{v}_i$$

i.e. $\boldsymbol{x}$ is on the affine hyperplane generated by $\{\boldsymbol{v}_i\}_{i=1,\cdots,n}$ if it is a linear combination of $\boldsymbol{v}_i$ with coefficients adding up to 1.

Definition 1 (Affine Combination)**.** The affine combination of set of vectors $\{\boldsymbol{v}_i\}_{i=1,\cdots,n}$ is defined as

$$\left\{\sum c_i \boldsymbol{v}_i \mid c_i \in \mathbb{R}, \sum c_i = 1\right\}$$

Example 1. The affine combination of $\boldsymbol{v}_1, \boldsymbol{v}_2$ is $\{c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 \,|\, c_1 + c_2 = 1\}$. Let $t = c_1$, then $c_2 = 1 - t$. So affine combination is equivalent to

$$\{t\boldsymbol{v}_1 + (1-t)\boldsymbol{v}_2 = \boldsymbol{v}_2 + t(\boldsymbol{v}_1 - \boldsymbol{v}_2) \,|\, c_1 + c_2 = 1\}$$

can you recognise that the set is a line? i.e. affine hull of dimension 1 in the 2-dimensional space.

Recall the definition of linear dependence: the spanning set has a dimension less than n. Equivalently, there exists one vector which is a linear combination of the other vectors.

Definition 2 (Affine Dependence)**.** $\{\boldsymbol{v}_i\}_{i=1,\cdots,n}$ is affine dependent if one vector is an affine combination of the other vectors (i.e. on the affine hyperplane generated by the other vectors).

Theorem 3.1 (Equivalent Definitions)**.** *The following are equivalent:*

- *(i)* $\{\boldsymbol{v}_i\}_{i=1,\cdots,n} \subseteq \mathbb{R}^m$ *is affine dependent*
- *(ii) exists* c_i *not all zero s.t.* $\sum c_i = 0$ *and* $\sum c_i \boldsymbol{v}_i = 0$.
- *(iii)* $\{\boldsymbol{v}_i - \boldsymbol{v}_1\}_{i=2,\cdots,n}$ *is linearly dependent*
- *(iv) The enlarged(augmented vectors) set* $\{(1\ \boldsymbol{v}_i)\}_{i=1,\cdots,n}$ *in* $\mathbb{R}^{m+1}$ *is linearly dependent*

Proof. **(i) $\Rightarrow$ (ii)**
WLOG $\boldsymbol{v}_1$ is affine combination of the rest, i.e. exists $c_2, \cdots, c_n$ s.t. $\boldsymbol{v}_1 = \sum_2^n c_i \boldsymbol{v}_i$ and $\sum_{i=2}^n c_i = 1$,

$$\Rightarrow 1 \cdot \boldsymbol{v}_1 - \sum_2^n c_i \boldsymbol{v}_i = \boldsymbol{0}$$

Let $\alpha_1 = 1$, $\alpha_i = -c_i$ for $i = 2, \cdots, n$, then $\sum \alpha_i \boldsymbol{v}_i = 0$. Also note $\sum \alpha_i = 0$.

(ii) $\Rightarrow$ (i)
Since c_i are not all zero, WLOG assume $c_1 \neq 0$, then

$$\boldsymbol{v}_1 = \sum_{i=2}^n -\frac{c_i}{c_1}\boldsymbol{v}_i$$

Clearly, by the assumption $\sum c_i = 0$,

$$\sum_{i=2}^n -\frac{c_i}{c_1} = -\frac{-c_1}{c_1} = 1$$

So $\boldsymbol{v}_1$ is an affine combination of the rest.

(ii) $\Leftrightarrow$ (iii) and (ii) $\Leftrightarrow$ (iv) are left as exercises for linear algebra skills. □

Based on the above theorem, the affine independence is usually defined as

$$\sum c_i \boldsymbol{v}_i = 0, \text{ and } \sum c_i = 0 \Longrightarrow c_i = 0 \,\forall i$$

As you may have noticed from the example above, the dimension of the span of $n+1$ affine-independent vectors in $\mathbb{R}^m$ is n. Further, they cannot be accommodated by any hyperplane with a dimension less than n.

Statistics $T_i(x)$ and the parameters $\eta_j(\theta)$ are functions, and the concept of independence extends to functions. Another way to characterise a vector $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)$ belonging to an affine hyperplane is: $\boldsymbol{v}^T\boldsymbol{a} = c$ for some constant scalar c and constant vector $\boldsymbol{a} \in \mathbb{R}^n$. Now replace the components v_i by functions $f_i(x)$.

Definition 3 (Affine independence (functions))**.** Functions $f_1, \cdots, f_n$ are affinely independent if for any $c_0, \cdots, c_n$

$$\sum c_j f_j(x) \equiv c_0 \;\Rightarrow\; c_j = 0 \,\forall j$$

note this is equivalent to saying

$$1 \cdot c_0 - \sum c_j f_j(x) \equiv 0 \;\Rightarrow\; c_j = 0 \,\forall j$$

i.e. the set $\{1, f_1, \cdots, f_n\}$ is linearly independent.

Remark. Functions differ slightly from vector components as they have a domain and codomain. We allow a small portion (zero measure) of the domain on which the condition is not satisfied, i.e. the conditions above apply μ-almost everywhere.

Corollary 1. *If $f_1, \cdots, f_n$ are affinely independent, then they are linearly independent.*

Note: this does not hold for vectors

By throwing away functions in an affinely dependent set dependent on others one by one, one would always eventually arrive at an affinely independent set of functions. i.e. a minimal representation(definition 1.5 in lecture notes). Proposition 1.8 in lecture notes states that the dimension of this representation is unique. Because the span of canonical statistics $T_i(x)$ (plus the constant function) is equal to the span of log-likelihood functions, which is independent of the choice of $T_i(x)$. Proposition 1.9 states a useful connection: the equivalence between affine independence and being a minimal representation. (also done using the space of log-likelihood functions)

Affine independence may not be straightforward, but using proposition 1.10, checking $\mathrm{Cov}_\theta(T)$ is positive definite for one θ is enough.

3.2 Curved Exponential Family

The space spanned by $\eta_i(\theta)$ may not fill the space of all possible η_i (i.e. not fill $\mathbb{R}^k$) in the expression of the exponential family. For example, if the parameters (μ, σ^2) of normal distribution satisfies $\mu^2 = \sigma^2$, there is only one degree of freedom. The parameter space is a manifold (one-dimensional curve) in $\mathbb{R}^2$. In such cases, the exponential family is called *curved*.

Recall that

$$\exp(B(\theta)) = \int_{x \in \mathscr{X}} h(x) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) T_i(x)\right) dx$$

where $B(\theta)$ serves for normalisation of PDF. i.e. ensuring the PDF integrates into one. Therefore, the parameter space Θ for the exponential family is defined as the set of θ that ensures this integral on RHS is finite so that $B(\theta)$ can be defined properly.

Definition 4 (Parameter Space)**.**

$$\Theta := \left\{ \theta \,:\, \int_{x \in \mathscr{X}} h(x) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) T_i(x)\right) < \infty \right\}$$

Another way to control the integral is to confine the η_is, which describes the inert property of the equation of the exponential family. Such space is called a natural parameter space.

Definition 5 (Natural Parameter Space)**.**

$$\Xi := \left\{ \eta \,:\, \int_{x \in \mathscr{X}} h(x) \exp\left(\sum_{i=1}^{k} \eta_i T_i(x)\right) < \infty \right\}$$

If $\eta(\theta)$ is out of Ξ, any reparametrisation of Θ fails, as the integral is always infinity. i.e. we have relation $\eta(\Theta) \subseteq \Xi$. Though definition of Ξ does not include Θ, it has useful properties

- Ξ is convex. (Lines between any two points in the set are contained in the set) This is essential for some optimisation algorithms.
- Ξ contains a non-empty k-dimensional ball. i.e. it is properly a space of at least k-dimensions. (counter-example: curves in 2-D space do not contain 2-dimensional open set)
- In the interior $\text{int}(\Xi)$, all moments of canonical statistics T exist, and the mean and variance of T are given by derivatives of cumulant function $B(\eta)$. (theorem 1.15 in lecture notes)

A convenient way to see whether $\eta(\Theta)$ fills Ξ is to check whether $\eta(\Theta)$ contains a k-dimensional open set. Take the normal distribution mentioned in the beginning as an example, the curve $(y = x^2)$ does not contain any 2-dimensional open set. It has an empty interior. If $\eta(\Theta)$ contains a k-dimensional open set, we call the family *full rank*. Otherwise, the family is curved exponential

Definition 6 (Curved Exponential Family)**.** An exponential family with k canonical parameters is a curved exponential family if

- parameter space $\Theta \subseteq \mathbb{R}^q$ for some $q < k$ (so $\eta(\Theta)$ cannot be full-rank)
- $\{T_i\}$ are affinely independent (i.e. $\text{Cov}_\theta(T)$ is positive definite for some θ)

The second assumption ensures the number of statistics is at a minimum, with no redundant usage.

4 Sufficiency

If you have a sample from $\text{Ber}(p)$, with p unknown. The sample mean is a good approximation for p because p means the probability of sampling 1. All information that can be used to infer p is contained in the sample mean, there is no need to consult every single value of $0, 1$. We have reduced the information from n (number of samples) dimensional to 1-dimensional.

Statistics like the sample mean mentioned above are called sufficient statistics, the rigorous definition is

Definition 7 (Sufficient Statistics)**.** $T(X)$ (independent of θ) is sufficient for θ if

$$f(x \,|\, T = T(x), \theta) = f(x \,|\, T(x))$$

i.e. all information about the distribution of X is given in T.

Corollary 2. *For any function h independent of θ and sufficient statistics $T(x)$, the distribution of $h(X) \,|\, T(X)$ is independent of θ, i.e.*

$$f(h(X) \,|\, T(X), \theta) = f(h(X) \,|\, T(X))$$

The factorisation theorem says: $T(x)$ being sufficient statistics is equivalent to: with the help of T, θ and x can be separated using two functions in the PDF. i.e. $f(x; \theta) = g(T(x), \theta)h(x)$. **In practice, use this theorem to check the sufficiency of a statistic**

With the factorisation theorem, the importance of the exponential family is revealed. The canonical statistics $T_i(x)$ wraps all information, and $h(x)$ can be moved out from the exponent:

$$f(x; \theta) = \underbrace{\exp\left\{\sum_i \eta_i(\theta) T_i(x) - B(\theta)\right\}}_{=g(T(x),\theta)} h(x)$$

i.e. $T(x) := (T_i(x))$ is sufficient statistics.

4.1 Minimality

The lecture notes have fully explained the motivations behind minimal sufficiency. In summary,

- minimal sufficient statistic is the summary of data with the greatest brevity
- as a partition of $\mathscr{X}$, it is the coarsest partition.

Minimal sufficient statistics also have a convenient criterion, but note it is a logical sentence rather than a single equation
Lehmann-Schffé condition:

$$T(x) = T(y) \iff \frac{f(y;\theta)}{f(x;\theta)} \text{ is independent of } \theta \tag{1}$$

Theorem 4.1. *T is minimal sufficient statistic $\Leftrightarrow$ T satisfies (1)*

Note: with this theorem, you do not have to check the factorisation theorem first to prove minimal sufficiency.

Proof. .
Part 1 Any minimal sufficient statistic T satisfies (1)
Part 1.A show $T(x) = T(y)$ implies likelihood-ratio independent of θ:
simply use the factorisation theorem, and function g depending on θ is cancelled.
Part 1.B show likelihood-ratio independent of θ implies x, y are in the same class (for the partition given by T)
$x \sim y \Leftrightarrow T(x) = T(y)$ makes a partition Π_T,

$$x \sim y \Leftrightarrow \frac{f(y;\theta)}{f(x;\theta)} \text{ is independent of } \theta$$

also makes a partition Π_G. The task now is to show the two partitions are the same. The first partition is based on the function T taking constant value on an equivalent class $[x] \in \Pi_T$, so similarly, we aim to seek a function G that takes constant value on equivalent class $[x] \in \Pi_G$. Using the definition of minimal sufficient statistics, we could somehow write T as a function of G and prove T also takes constant value on $[x] \in \Pi_G$.
There is a trivial constant function G: take representative $\overline{x}$ of class $[x] \in \Pi_G$, define $G(y) := \overline{x}$ for any $y \in [x]$. Trivially, G is constant on equivalent classes in Π_G.
For $y \in [\overline{x}]$, by definition of Π_G, the likelihood ratio is $k(y, \overline{x})$, independent of θ. So

$$f(y;\theta) = k(y,\overline{x})f(\overline{x};\theta) = k(y, G(y))f(G(y);\theta)$$

i.e. G is a sufficient statistic (by factorisation theorem). By minimality of T, Π_T is coarser than Π_G, i.e. $T = h(G)$ for some function h. So T will be constant on $[x] \in \Pi_G$ as G does.
Therefore, the partitions Π_G and Π_T are the same.

Part 2 T satisfying (1) $\Rightarrow$ T is minimal sufficient statistic
Part 2.A Sufficiency:
Use the definition directly, because likelihood ratios are involved.

$$\begin{aligned} f(x|t,\theta) &= P_\theta(X = x|T = t) \quad \text{where } t := T(x) \\ &= \frac{P_\theta(X = x)}{P_\theta(T(x) = t)} \quad \text{definition} \\ &= \frac{f(x;\theta)}{\sum_{y:\, T(y)=t} f(y;\theta)} \quad \text{by law of total probability} \end{aligned}$$

In the last line, we used the fact statistics T is a partition of the space again. Since the likelihood ratio is independent of θ over the set $\{y : T(y) = t = T(x)\}$, $f(y;\theta)/f(x;\theta) = k(x, y)$ for some function k, so

$$f(x|t,\theta) = \frac{f(x;\theta)}{\sum_{y:\, T(y)=t} f(x;\theta)k(x,y)} = \frac{1}{\sum_{y:\, T(y)=t} k(x,y)}$$

By PDF's uniqueness, the last expression must be $f(x|t)$ (independent of θ).
Part 2.B Minimality:

If U is another sufficient statistic, aim to prove $\phi(U) = T$ for some function ϕ. By part 1.A, $U(x) = U(y) \Rightarrow$ likelihood ratio independent of $\theta \Rightarrow T(x) = T(y)$. (likelihood ratio independent of the sufficient statistic chosen)

Consider the two partitions Π_U, Π_T given by statistics U and T. by the above arguments, if $A \in \Pi_U$, then there is a set $\overline{A} \in \Pi_T$ containing A. Given u, define $A_u \in \Pi_U$ as the unique set containing u. Then $\phi(u) := T(\overline{A_u})$ satisfies our requirement. (note T is constant on $\overline{A_u} \in \Pi_T$) □

Corollary 3. *For k-parameter exponential family with canonical parameters $T(x)$, $T_{(n)}(x) := \sum_i T(x_i)$ is a minimal sufficient statistic. (x_i are samples for that family)*

Proof. By simplifying the likelihood ratio between $\boldsymbol{x}$ and $\boldsymbol{y}$, you will find the terms with θ becomes 0 iff $\sum_i T_j(x_i) = \sum_i T_j(y_i)$ for all i, i.e. $T_{(n)}(\boldsymbol{x}) = T_{(n)}(\boldsymbol{y})$. □

5 Quality of Estimators

We have seen one quality measure for estimators: sufficiency, i.e. whether the statistic summarises all information about the parameter available in the data/realisation.

In this section, suppose T is an estimator of $g(\theta)$.

Bias: $\text{bias}_\theta(T) := E_\theta(T) - g(\theta)$. Usually, we want an unbiased estimator (bias = 0). But the extent of T variation when θ changes should not be too large (stability and robustness).

MSE measures the expected distance between the estimator and the true parameter

Definition 8. If T is an estimator of $g(\theta)$,

$$\text{MSE}_\theta(T) = E_\theta[(T - g(\theta))^2]$$

The expectation is taken over the whole parameter space Θ.

Note

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) + \text{bias}_\theta(T)^2$$

which means MSE measures variation and bias of T together. Keeping both variation and bias small is desirable, but not always possible. This is known as *variance-bias trade-off* and will be studied again in Chapter 9 Empirical Bayes Methods.

Consistency: when we increase the sample size n, $T_n := T((X_1, \cdots, X_n))$ gets closer to the true parameter θ. There are three types of consistency

- **Consistency in probability/weak consistency**: probability that T_n and θ differs approaches 0,

$$\lim_{n\to\infty} P_\theta[|T_n - \theta| > \epsilon] = 0$$

- **Consistency in MSE**:

$$\lim_{n\to\infty} \text{MSE}_\theta(T_n) = 0$$

- **Strong Consistency**: T_n converges to θ with probability 1 (almost surely)

$$P_\theta(\lim_{n\to\infty} T_n = \theta) = 1$$

Since almost surely convergence implies convergence in probability, (strong consistency $\Rightarrow$ weak consistency) L2-convergence also implies convergence in probability, so (consistency in MSE $\Rightarrow$ weak consistency)

6 Likelihood-based Estimation

6.1 Likelihood

You have likely learnt likelihood before. It is a brother of probability.

- Probability $p(x;\theta)$: given parameter θ, how likely is the realisation of X to be x.
- Likelihood $L(\theta;x)$: given realisation x, how likely is the parameter being θ

The equation $p(x;\theta) = L(\theta;x)$ holds, but remember they are functions on different spaces. ($\mathscr{X}$ and Θ)

The idea of MLE is simple. Suppose the probability of drawing a white ball from a dark box with either a white or black ball is p (unknown), and in ten independent draws (with replacement), you got 4 white balls. Among all parameters $p \in [0,1]$, which value is the most likely to cause this situation? By intuition, this is $4/10 = 0.4$. Maximum likelihood estimation is the rigorous mathematical process of obtaining such value, involving maximisation/minimisation.

Definition 9 (Maximum Likelihood Estimator)**.** Given realisation x, the statistics $\hat{\theta}_{\text{MLE}} = T(x)$ is Maximum Likelihood Estimator(MLE) if

$$L(T(x);x) = \max_{\theta\in\Theta} L(\theta;x)$$

Properties of MLE:

- MLE may not be unique (multiple maximisers), and MLE may not even exist
- MLE remains invariant under transformation g. i.e. if $\hat{\theta}_{\text{MLE}}$ is MLE for θ, for any function g, $g(\hat{\theta}_{\text{MLE}})$ is the MLE for $g(\theta)$.
- Probability involves a lot of products, a trick is to use logarithms to change them to sums, which are easier for differentiation. So usually, maximisation is done on the log-likelihood $l(\theta;x) := \log(L(\theta;x))$.

Calculation of MLE becomes troublesome when the support of PDF f depends on θ. First, you cannot use logarithm as $\log 0$ is not defined. Secondly, differentiation fails.

Example 2. If $X = (X_1, \cdots, X_n)$ are iid $U[0,\theta]$ (uniform distribution), then

$$L(\theta;X) = f_\theta(X) = \prod_{i=1}^{n} \frac{1}{\theta} 1_{[0,\theta](x_i)} = \frac{1}{\theta^n} 1_{\max_i x_i \le \theta} 1_{\min_i x_i \ge \theta}$$

where the indicator function $1_{[0,\theta](x_i)} = 1$ if $x_i \in [0,\theta]$ and $1_{[0,\theta](x_i)} = 0$ otherwise.

You cannot differentiate this function. But note $1/\theta^n$ is decreasing with θ, and the smallest θ can be s.t. $L(\theta;X)$ is non-zero is $\max_i x_i$. So $\hat{\theta}_{\text{MLE}} = \max_i x_i$.

Another commonly used estimator is the Moment Estimator. What is the simplest estimator for the mean of a distribution? The sample mean $\sum_{i=1}^n x_i/n$. More generally, suppose the parameter γ can be written using only moments $m_1, \cdots, m_r$ (where $m_k := E(X^k)$), then the moment estimator

$$\gamma_{\text{MME}} := h(\hat{m}_1, \cdots, \hat{m}_r)$$

where $\hat{m}_k := \sum_{i=1}^n x_i^k/n$ is the empirical moment obtained from the data.

Properties of MME:

- Often, MME is either unbiased or asymptotically unbiased
- has strong consistency, and it is preserved under transformation
- Easy to calculate compared to MLE
- No knowledge of the sample distribution is required.

6.2 Intuitions behind Fisher's information

We care about the derivative of log-likelihood. Because the way it varies indicates how much information of the parameter θ is contained in realisation x.

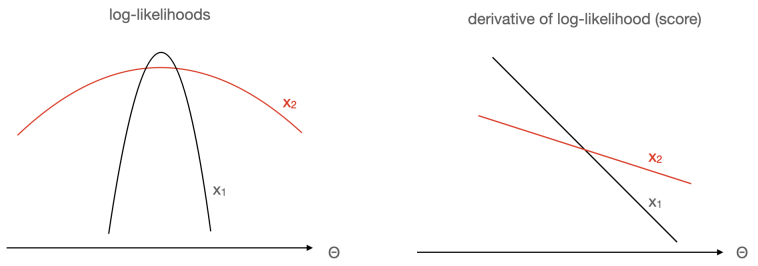

Figure 1: Comparison between log-likelihoods with different variations

If the first derivative of log-likelihood varies a lot, it means there are more sharp spikes in the log-likelihood function. For example, the black curve in figure 1 produced by observation x_1 is a sharp spike, a shift in the parameter from the peak (representing MLE estimator $\hat{\theta}_{\text{MLE}}$) causes a larger drop in log-likelihood. Therefore, we are confident that the true parameter is around the peak. So x_1 contains a lot of information about θ. Conversely, the red curve produced by observation x_2 varies less. Moving θ away from the peak, the likelihood stays high. So the peak may still be far from the true parameter, and x_2 provides less information about the true parameter.

In this section, we study the one-dimensional parameter space first.

Definition 10 (score)**.** The score function $S : \Theta \times \mathscr{X} \to \mathbb{R}$ is defined as

$$S(\theta, x) := \frac{\partial l(\theta; x)}{\partial \theta}$$

Given realisation x, the MLE is essentially the $\hat{\theta}$ that makes $S(\hat{\theta}, x) = 0$. Note the score function is also a random variable, as it depends on the value of the random variable X. Therefore, we can study it's expectation and variance.

The expectation of log-likelihood is important, based on the following proposition.

Proposition 6.1. *Under some regularity conditions that allow you to interchange integral (over $\mathscr{X}$) with differentiation (w.r.t θ), if θ_0 is the true parameter and $X \sim f_{\theta_0}$, then the function $\theta \mapsto E_{\theta_0}[l(\theta; X)]$ is maximised at θ_0.*

Note: the regularity conditions for continuous variable X are

- **Reg1 The support of PDF f does not depend on θ** (this is true for exponential families). We have seen in example 2 in the last section why this condition is important.
- **Reg2 Θ is open**. Differentiation on the boundary of sets is impossible, open sets have no boundary.
- **Reg3.1 Derivative $\partial_\theta f(x, \theta)$ exists and is finite** (we need this so that the score function can be defined)
- **Reg3.2 For each θ, there is a neighbourhood of θ on which the absolute value of $\partial_\theta f(x, \theta)$ is bounded by an integrable function.** Integrable function means a function with finite Lebesgue integral, this is required for defining the expectation and variance of the score function. (see the rigorous definition in the lecture notes)

for discrete variables, replace integration with summation.

By the above proposition, the true parameter can be recovered by maximising the function $\theta \mapsto E_{\theta_0}[l(\theta; X)]$. What about the expectation of the derivative (the score function)? By the above proposition, we have $E_{\theta_0}[S(\theta_0, X)] = 0$. Surprisingly, this holds for all $\theta \in \Theta$.

Proposition 6.2. *Under regularity conditions, $E_\theta[S(\theta, X)] = 0$ for all $\theta \in \Theta$*

The proofs of the two propositions are very similar, and the second proof is in the lecture notes. As an exercise, prove Proposition 6.1 by yourself.

We have explained at the beginning of this section that the variance of the score function measures how much information is contained in the observation x.

Definition 11 (Fisher's information(definition 1))**.** The Fisher information for variable X is

$$I_X(\theta) := \mathrm{Var}_\theta[S(\theta, X)] = E_\theta[S(\theta, X)^2]$$

note the second equality holds due to Proposition 6.2. And Fisher's information depends on θ and X.

There is another way to describe how the score function varies: the derivative of the score, i.e. the second derivative of log-likelihood. So, a second definition for Fisher's information is

Definition 12 (Fisher's information(definition 2))**.** The Fisher information for variable X is

$$I_X(\theta) := -E_\theta[\partial_\theta S(\theta, X)] = E_\theta[-l''(\theta; X)]$$

$-l''(\theta; X)$ is often called the observed information, and denoted $J(\theta; X)$. It is Fisher's information at an observation of X.

We need to take expectation as $\partial_\theta S(\theta, X)$ is also a random variable. The negative sign is there because Fisher's information should be higher around a spike of log-likelihood (where the score function has a negative gradient). The following theorem(will see it again in the next section) rigorously describes the relation between the variance of an unbiased estimator and Fisher's information

Theorem 6.3 (Carmer-Rao Lower bound)**.** *If $\Theta \subseteq \mathbb{R}$ (only one parameter considered), and regularity conditions are satisfied, then for the unbiased estimator $\hat{\theta}(X)$ and all $\theta \in int(\Theta)$,*

$$Var_\theta(\hat{\theta}(X)) \geq \frac{1}{I_X(\theta)}$$

This is an important application of Fisher's information. The variance of an unbiased estimator of θ has a lower bound given by sensitivity of θ to data X. If $I_X(\theta)$ is larger, θ is more sensitive to change in data X, so the estimator of θ can potentially be more certain (smaller variance).

Note that definition 2 is valid only if $l(\theta; X)$ is twice-differentiable. To prove definition 1 and definition 2 are equivalent, the integration should be exchangeable with the second derivative (w.r.t θ). This is called regularity condition 4. (or Reg4)

Theorem 6.4. *Under Reg1 - Reg4, the two definitions of Fisher's information are equivalent, i.e.*

$$E_\theta[-l''(\theta; X)] = var_\theta[l'(\theta; X)] = E_\theta[(l'(\theta; X))^2]$$

Proof. See the lecture notes. □

Proposition 6.5 (Properties of the Fisher information)**.**

- *(Information of independent samples) If you have independent samples $X_1, \cdots, X_n$, then the Fisher's information of $(X_1, \cdots, X_n)$ is the sum of Fisher's information for X_i. i.e. each sample provides some information on the parameter θ, and the pieces of information add up.*
- *(Reparametrisation) If reparametrisation ξ is given by $\theta = h(\xi)$ with h being differentiable, then the Fisher's information for ξ, $I^*_X(\xi)$, is*

$$I^*_X(\xi) = I_X(h(\xi)) \cdot (h'(\xi))^2$$

For the formulae of reparametrisation, the squared term $(h'(\xi))^2$ comes from the twice differentiation on the way of finding the Fisher information (twice use of chain rule).

All the definitions above can be extended to multi-dimensional parameter space using random vectors and random matrices. Please check section 3.2 of the lecture notes. The only thing to note is that we do not need total differentiation in multi-variate cases. Only partial derivatives of the log-likelihood are involved in the above processes.

6.3 MVUE and Cramer-Rao Lower Bound

It is tempting to measure the equality of the estimator merely using MSE. However, there is always an estimator with zero MSE: $\hat{\gamma}_0 := \theta_0$ where θ_0 is the true parameter. Therefore, no other parameter can beat this estimator.

MVUE: the estimator with the smallest variance among all unbiased estimators for all values of θ. This is a stringent requirement. MVUE may not exist. In figure 2, suppose all three unbiased estimators are T_1, T_2, T_3. On the left case, T_3 has the smallest variance at all θ. However, in the right case, T_2 and T_3 take turns to be the estimator with the smallest variance for different θ. So no MVUE exists.

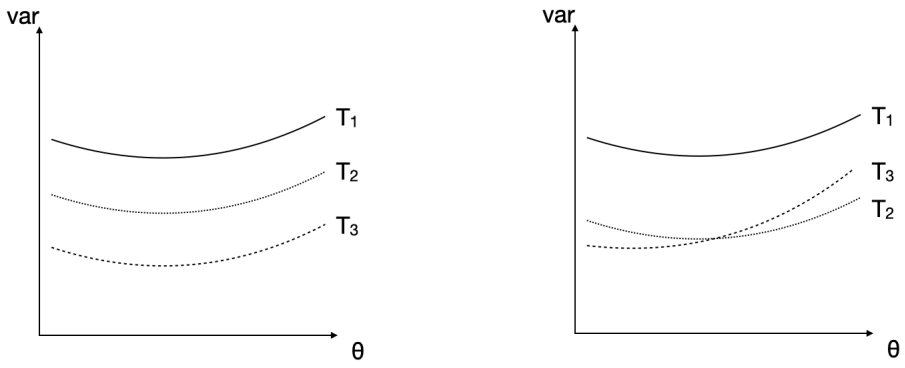

Figure 2: A demonstration of MVUE

Can the variance of an unbiased estimator become as small as you like? The answer is no. There is a lower bound (called *Cramer-Rao lower bound*) related to Fisher's information.

Theorem 6.6 (CRLB(1D))**.** *Suppose Reg2 - Reg4, and Fisher information $I_X(\theta) \in (0, \infty)$, and $\gamma = g(\theta)$ for g continuously differentiable with $g' \neq 0$.*

If T is regular (can exchange integration with integrand $T(x)$ and differentiation w.r.t. θ) unbiased estimator of γ, then

$$Var_\theta(T) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}$$

T attains the lower bound iff the score function is an affine function of T with an intercept at $g(\theta)$. More precisely,

$$T(x) - g(\theta) = g'(\theta) S(\theta; x)/I_X(\theta)$$

for all θ, x.

Proof. Aim to use the Cauchy-Schwarz inequality,

$$E_\theta(YZ)^2 \leq E_\theta(Y^2) E_\theta(Z^2)$$

let $Y = T(x) - g(\theta)$, $Z = S(\theta, X) = l'(\theta; x)$, then

$$E_\theta[(T(x) - g(\theta))S(\theta, X)]^2 \leq E_\theta[(T(x) - g(\theta))^2] E_\theta[S(\theta; x)^2]$$

Recall $E_\theta(S(\theta, X)) = 0$ and $\mathrm{Var}_\theta(S(\theta, X)) = I_X(\theta)$. So

$$E_\theta[(T(x) - g(\theta))S(\theta, X)]^2 \leq \mathrm{Var}_\theta(T) I_X(\theta)$$

remains to show $E_\theta[T(x) - g(\theta)S(\theta, X)] = \partial/\partial\theta E_\theta[T] = g'(\theta)$. This can be proved by simple manipulations of calculus and the fact that T is regular.

Cauchy-Schwarz inequality obtains equality iff $Y - cZ$ is almost surely 0. In this case, it is $T(x) - g(\theta) = cS(\theta; x)$ almost surely. Taking the inner product of this expression,

$$E_\theta[(T(x) - g(\theta))^2] = c^2 E_\theta[S(\theta; x)^2] \;\Rightarrow\; \mathrm{Var}_\theta(T) = c^2 I_X(\theta) = \frac{|g'(\theta)|^2}{I_X(\theta)}$$

so $c = g'(\theta)/I_X(\theta)$. Therefore, the equality condition is

$$T(x) - g(\theta) = g'(\theta)S(\theta; x)/I_X(\theta)$$

There is another version of Cauchy-Schwarz, which also allows you to prove CRLB:

$$\mathrm{Cov}_\theta(YZ)^2 \leq \mathrm{Var}_\theta(Y)\mathrm{Var}_\theta(Z)$$

□

Any estimator attaining CRLB is an MVUE. Efficiency measures how close the $\mathrm{Var}_\theta(T)$ is to CRLB.

Definition 13 (Efficiency)**.** Efficiency of unbiased estimator T of $g(\theta)$ is the ratio of variance and CRLB,

$$e(T, \theta) := \frac{|g'(\theta)|^2}{I_X(\theta)\mathrm{Var}_\theta(T)}$$

The highest possible efficiency is 1, which means T attains CRLB.

Proposition 6.7. *Given $X \sim f_\theta$ for $\theta \in \Theta$. Under regularity conditions, if there exists an estimator $T(x)$ attaining CRLB for $\gamma := g(\theta)$, then $\{f_\theta\}$ must be an exponential family.*

So, the expression of the exponential family is designed for attaining CRLB. For distributions attaining CRLB, any unbiased estimator is MLE (given that MLE exists). So indeed, distributions satisfying CRLB are quite useful.

The CRLB for multivariate cases uses the Loewner order, recall that $A \succeq 0$ means A is positive semi-definite, i.e. for all $u \in \mathbb{R}^m$, $u^T A u \geq 0$. The order $\succeq$ is the Loewner order.

Definition 14 (Loewner order)**.** Given two matrices $A, B \in \mathbb{R}^{m\times m}$, if for all $u \in \mathbb{R}^m$, $u^T(A - B)u \geq 0$, then $A \succeq B$

The multi-dimensional CRLB is obtained by replacing $1/I_X(\theta)$ with the inverse matrix $I_X(\theta)^{-1}$, g' with multivariate derivative(Jacobin) $D_\theta g$ and the square becomes matrix square $(D_\theta g)I_X(\theta)^{-1}(D_\theta g)^T$.

Theorem 6.8 (CRLB(Multi-dimensional))**.** *Suppose multivariate Reg1 - Reg4 holds, and Fisher information matrix $I_X(\theta)$ is not singular (non-zero determinant), then*

$$Var_\theta(T) \succeq (D_\theta g)I_X(\theta)^{-1}(D_\theta g)^T$$

Rao-Blackwell Theorem
Also by the famous statistician Rao, this theorem states that using sufficient statistics $T(X)$ instead of using whole data X directly reduces the uncertainty of any unbiased estimators. i.e. given $\gamma = g(\theta)$ unbiased, $\hat{\gamma}_T := E_\theta[\hat{\gamma} \,|\, T]$ is also an unbiased estimator with smaller variance. And $\hat{\gamma}_T$ is almost surely the unique estimator with this variance if $\mathrm{tr}(\mathrm{Cov}_\theta(\hat{\gamma})) < \infty$ (Please find the comprehensive proof in the lecture notes. Or check Rao's paper [11] where you can find proofs of both CRLB and Rao-Blackwell theorem).

6.4 Complete Statistics

Recall from linear algebra that a set of vectors $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ is called complete if they span the whole space. i.e. any $\boldsymbol{v} \in \mathbb{R}^n$ can be written as $\boldsymbol{v} = \sum_{i=1}^n a_i \boldsymbol{v}_i$, where $a_i = \langle \boldsymbol{v}, \boldsymbol{v}_i \rangle$. Therefore, if $\boldsymbol{v}$ is orthogonal to every $\boldsymbol{v}_i$, $v = 0$. Similarly, if we have a statistics $T \in \mathcal{T}$ based on a discrete random variable $X \in \mathscr{X}$, T is called complete if for any function $h : \mathcal{T} \to \mathbb{R}$

$$E_\theta(h(T)) = \sum_{t\in\mathcal{T}]} h(t)P_\theta(T = t) = 0 \;\forall\theta \in \Theta \implies h(t) = 0$$

The proper definition allows $h(t)$ to be non-zero for a set with zero probability measure.

Definition 15 (Complete Statistics)**.** Statistics T for model $\{P_\theta\,;\,\theta\in\Theta\}$ is complete if

$$E_\theta[h(T)]=0\ \forall\theta\in\Theta\ \Rightarrow\ P_\theta(h(T)=0)=1\ \forall\theta\in\Theta$$

Example 3. If $X_1,\cdots,X_n$ are i.i.d. $U(0,\theta)$, the sufficient statistics $T=\max_i X_i$ is complete.
Note $f_T(t)=nt^{n-1}/\theta^n 1_{(0,\theta)}(t)$, so if

$$E_\theta[h(T)]=\int_0^\theta h(t)f_T(t)\,dt=\frac{n}{\theta^n}\int_0^\theta h(t)t^{n-1}\,dt=0$$

for all $\theta>0$, then differentiation and applying the fundamental theorem of calculus yields the integrand $h(t)t^n=0$ for all $t>0$. But this is true only if $h(t)\equiv 0$. So $P_\theta(h(T)=0)=1$.

Properties of complete statistics

- We have seen that $T(x)=(T_i(x))_{i=1,\cdots,k}$ is sufficient for exponential family. If the exponential family is full-rank, we can further deduce that T is complete.
- complete statistics may not exist
- If a sufficient statistics T is complete, then T is minimal sufficient

Using complete statistics, we can upgrade the Rao-Blackwell theorem. If T is sufficient and complete, then the Rao estimator $\hat\gamma_T:=E_\theta[\hat\gamma\,|\,T]$ has minimum variance

Theorem 6.9 (Lehmann-Scheffe)**.** *If T is sufficient and complete, $\hat\gamma$ is an unbiased estimator for $g(\theta)$, then $\hat\gamma_T$ is MVUE for γ. In fact, it is the unique MVUE.*

Proof. Suppose $\tilde\gamma$ is unbiased another estimator and define $\tilde\gamma_T$ similarly. First by the Rao-Blackwell theorem,

$$\mathrm{Var}_\theta(\tilde\gamma_T)\le\mathrm{Var}_\theta(\tilde\gamma)\ \forall\theta\in\Theta$$

To prove $\hat\gamma_T$ is MVUE we need

$$\mathrm{Var}_\theta(\hat\gamma_T)\le\mathrm{Var}_\theta(\tilde\gamma)\ \forall\theta\in\Theta$$

so proving $P_\theta(\hat\gamma_T=\tilde\gamma_T)=1$ is enough. recall T is complete, the condition is equivalent to $E_\theta[\hat\gamma_T-\tilde\gamma_T]=0$. But both of them are unbiased estimators of γ, so

$$E_\theta[\hat\gamma_T-\tilde\gamma_T]=\gamma-\gamma=0$$

□

The uniqueness of MVUE stated in the Lehmann-Scheffe theorem can be proved by the following lemma:

Lemma 6.10. *If $g_1(T),g_2(T)$ are unbiased estimators of $g(\theta)$ where T is a complete statistics, then $g_1=g_2$ a.e. (equivalent to $P(g_1(T)-g_2(T)=0)=1$)*

Proof. Direct application of the completeness: $P(g_1(T)-g_2(T)=0)=1$ can be deduced from $E(g_1(T)-g_2(T))=0$ by completeness. (defining $h(T):=g_1(T)-g_2(T)$) But this is trivial as they are both unbiased.

$$E(g_1(T)-g_2(T))=g(\theta)-g(\theta)=0$$

□

Using Lehmann-Scheffe theorem to find MVUE

- Find unbiased estimator $\hat\gamma$ (it can be a very simple estimator), and complete, sufficient statistics T, then

  $$\hat\gamma_T:=E_\theta[\hat\gamma\,|\,T]$$

  is MVUE
- If you can find complete, sufficient statistics T and function $h=h(T)$ s.t. $E_\theta[h(T)]=g(\theta)$ (i.e. $h(T)$ is unbiased), then $h(T)$ is the MVUE. (by uniqueness)

When would there exist an MVUE? If we want $\hat\gamma$ to be MVUE, note for any other unbiased estimator $\tilde\gamma$, we can define $U:=\tilde\gamma-\hat\gamma$ (note $E_\theta(U)=0$) so that $\tilde\gamma=U+\hat\gamma$. So $\tilde\gamma$ absorbs both variations from $\hat\gamma$ and U. If $\hat\gamma$ is uncorrelated to all such U (with expectation 0), then the variance of $\hat\gamma$ is guaranteed to be smaller than $\tilde\gamma$.

Theorem 6.11 (Condition for MVUE)**.** *$\hat\gamma$ is MVUE for $\gamma:=g(\theta)$ iff $\hat\gamma$ is uncorrelated to all $U\in\mathcal{U}$ where*

$$\mathcal{U}:=\left\{h:\mathscr{X}\to\mathbb{R}\,:\,E_\theta[h(X)]=0,\ E_\theta[h^2(X)]<\infty\right\}$$

7 Introduction to Bayesian Statistics

The key difference between Bayesian and frequentist statistics is that the parameter θ is considered fixed for the frequentist's view, but Bayesian views θ as a random variable with associated distribution.

So even before collecting any data, we assign a *prior distribution* $\pi(\theta)$ based on existing knowledge or experience. For example, a good prior for probability p of a coin showing head can be a triangular distribution peaked at 0.5. Because in experience, most coins are fair, with p around 0.5, and the extreme values ($p \approx 0$, $p \approx 1$) are less likely.

Instead of looking for a good estimator for θ after collecting the data, we update the distribution of θ. The new distribution $p(\theta|X)$ is often called *posterior distribution*. Here is the fundamental formulae for Bayesian statistics

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(X|\theta)\pi(\theta)}{p(X)} \propto p(X|\theta)\pi(\theta)$$

we ignore $p(X)$ as it is unknown, and it's irrelevant to θ. After obtaining $p(X|\theta)\pi(\theta)$, use the fact that

$$\int_\Theta p(\theta|X)\ d\theta = 1$$

to normalise $p(X|\theta)\pi(\theta)$ to make it a proper distribution.

Likelihood Principle
Here we introduce a way to evaluate an inference method, *likelihood principle*(LP). As said in [5], it is a normative principle. By far, we have learnt methods like MLE, MME and Bayes inference. An inference satisfies the LP if the conclusion is the same whenever the likelihoods of two experiments are the same, regardless of the sampling method.

For example, to estimate p for Bernoulli trials. One can either fix the number of trials n and count the number of successes y, or count the number of trials required until that number of successes. The distribution is Binomial in one case but negative binomial in the other case. But likelihood functions have the same form. So any inference satisfying LP should be the same when n, y are the same.

Some methods from frequentist's statistics satisfy LP, e.g. MLE. Most Bayesian inferences satisfy LP, but Jeffrey's prior does not. Staring at the likelihood is enough when using an inference method satisfying LP.

Berstein von-Mises Theorem Is the choice of prior important? Well, there are only two terms in the Bayes formulae: the likelihood $p(X|\theta)$ and prior $\pi(\theta)$. Especially when the data size of X is small, the likelihood has a limited impact on the distribution of θ. So, the way to find a good choice of prior should be studied. On the other hand, the Berstein von-Mises Theorem told us that under certain regularity conditions when your data size n gets large enough, the posterior distribution $\pi(\theta|X)$ will look like a sampling distribution of the MLE (so we are taking a frequentist's view of the Bayes statistics), which is normal distribution centred at MLE (or other consistent estimator) with the variance being exactly CRLB. Or strictly speaking,

$$\int_\Theta \left|\pi(\theta|X) - \phi(\theta; \hat{\theta}, I(\theta_0)^{-1}/n)\right|\ d\theta \to 0$$

when $n \to \infty$. ϕ is the pdf of normal distribution. This is a convergence in probability under the true parameter θ_0.

However, the regularity conditions are strict, so it is easy to find a model where this theorem cannot apply. The conditions are as follows

- The parameter space has fixed, finite dimensions $\Theta \subset \mathbb{R}^d$, with the true parameter θ_0 lying in side.
- The log-likelihood is smooth enough. (we only need up to two derivatives, i.e. in $\mathcal{C}^2$) and the expectation of the first derivative is bounded and that of the second derivative is locally bounded.
- prior $\pi(\theta)$ is continuous and non-zero at θ_0 (otherwise, the posterior $p(\theta_0|X) = 0$ no matter what X is, we never obtain the true parameter)
- MLE or the estimator $\hat{\theta}$ you choose should be consistent

- posterior should be a consistent estimator.

The first assumption breaks for large unstructured data sets, and the last assumption is fragile. For example, the posterior mean is usually inconsistent with the true population mean.

Again, from this view, the study of choice of priors is essential.

7.1 Choice of Priors

If the estimation of Bernoulli parameter θ uses the normal distribution centred at $1/2$ (only use the part on $[0,1]$) as prior, the resulting posterior is not in some nice form

$$p(\theta|x) \propto \theta^x(1-\theta)^x \exp\left\{-\frac{1}{2\sigma^2}(\theta-1/2)^2\right\}$$

(it is already messy only with one data point x) This is not in the form of any common distributions that we know. So a bad choice of prior gives you pain in calculations and further analysis.

What is a good choice here? It is time to introduce the Beta distribution

$$\pi(\theta) = \frac{1}{\text{Beta}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where the Beta function $\text{Beta}(\alpha,\beta)$ is just a normalising constant. You can see the form of Beta blends pretty well with the Bernoulli likelihood. Indeed, the posterior distribution is another Beta distribution with parameters adjusted according to data.

$$\pi(\theta|x_1,\cdots,x_n) \sim \text{Beta}(\alpha+\sum_i x_i, \beta+n-\sum_i x_i)$$

In such case, we say Beta distribution is the *conjugate prior* for the likelihood of Bernoulli.

Definition 16 (conjugate prior)**.** Prior distribution family $(\pi_\gamma)_{\gamma\in\Gamma}$ (note: γ is the parameter, in the case of Beta, $\gamma=(\alpha,\beta)$) is said to be the *conjugate prior* for likelihood $L(\theta,x)$ if the posterior

$$\pi_\gamma(\theta|x) = L(\theta,x)\pi_\gamma(\theta) = \pi_{\gamma'}(\theta)$$

for some $\gamma'\in\Gamma$. γ' depends on the data x so it can also be denoted as $\gamma(x)$.

The beta distribution is also very flexible. For example, figure 3 depicts various shapes of the Beta distribution.

Usually, we want the first two moments of the prior(mean and variance) to match that of the sample. But even if you restrict the mean to a certain value like 0.5, i.e.

$$\frac{\alpha}{\alpha+\beta} = \frac{1}{2}$$

there are still many combinations of α and β possible.

For normal likelihood, the conjugate prior is the Gamma distribution. As long as the likelihood belongs to the exponential family, there is a conjugate prior for it. (Proposition 7.4 in lecture notes)

Table of conjugate prior

Likelihood	Conjugate prior	number of parameters
Bernoulli, Binomial, Geometric, Negative Geometric	Beta	2
Normal	Normal	2
Exponential, Poisson, Normal	gamma	2

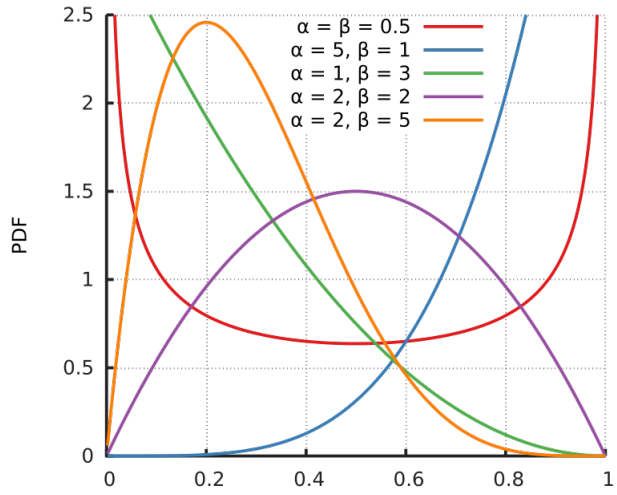

Figure 3: Shapes of Beta distribution (from Wikipedia)

Is there a good way to choose the parameter γ for conjugate prior? In the case of Beta distribution, the sum $\alpha + \beta$ increases by n after collecting n data points. (because $\alpha' = \alpha + \sum_i x_i, \beta' = \beta + n - \sum_i x_i$) So we can say $\alpha + \beta$ quantifies our confidence in the model. If you think your experience is equivalent to having seen 100 data points, then set the prior such that $\alpha+\beta = 100$. A new data point has less effect on your prior if you set higher $\alpha+\beta$.

If you calculate the posterior mean (i.e. the mean of Beta(α', β')),

$$\begin{aligned} E(\theta|x) &= \frac{\sum_i x_i + \alpha}{n + \alpha + \beta} \\ &= \frac{n}{n+\alpha+\beta}\overline{X}_n + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta} \\ &= \frac{n}{n+\alpha+\beta}\overline{X}_n + \frac{\alpha+\beta}{n+\alpha+\beta}E(\theta) \quad \text{where } E(\theta) \text{ is the prior mean} \end{aligned}$$

i.e. posterior mean is a weighted sum of the sample mean and the prior mean. Again, the term $\alpha + \beta$ appears. Higher $\alpha + \beta$ means the prior mean plays a more important role.

Conjugate priors are convenient, but sometimes flexibility in the distribution is required. e.g. a heavy-tailed t distribution instead of the normal distribution. In such cases, Monte Carlo sampling helps to avoid crazy calculations. Interested readers can refer to Chapter 4 of [6].

7.2 Non-informative prior

Alternatively, if you feel you have no prior knowledge of the parameter, a *non-informative prior* can be used. For example, a layman could not know whether a baseball player will strike the ball or not. Parameters do not have to be specified for non-informative prior. One lazy choice is to let $\pi(\theta) \propto 1$, which is the *uniform prior*.

Such priors are called *improper* as technically they are not distribution, i.e.

$$\int_\Theta \pi(\theta)\, d\theta = \infty$$

but if integral of $L(\theta; x)\pi(\theta)$ is finite, the posterior distribution still exists.

Note if we choose $\alpha = \beta = 0$ (i.e. $\alpha + \beta = 0$, no prior information) in the case of beta distribution,

$$c := \int_0^1 \frac{1}{\theta(1-\theta)} = \infty$$

so the PDF (abuse the definition a bit, it is not a PDF) $\frac{1}{c}\theta^{-1}(1-\theta)^{-1}$ is non-zero iff $\theta = 0$ or 1. i.e. with no prior knowledge, one takes whatever the first data point they see as the true value. No randomness is involved. Indeed, the posterior distribution exists and is:

$$p(\theta|x=1) \propto \frac{1}{1-\theta}, \quad p(\theta|x=0) \propto \frac{1}{\theta}$$

both posterior distributions are improper. So after normalising, $p(\theta|x=1)$ is concentrated at $\theta = 1$ and $p(\theta|x=0)$ is only non-zero when $\theta = 0$. i.e. the posterior conclusions are simply the data.

7.2.1 Jeffrey Prior

One drawback of the uniform prior is that: after reparametrisation, the uniform prior may not be uniform anymore. Take the example in the lecture notes, $\pi(\theta) = 1$ takes no preference on any θ, but after transformation $\eta := \log\theta$, the uniform prior becomes e^η, which favours higher η values. log is monotone, so the transformed prior favours a higher θ value when it shouldn't.

Recall from Proposition 6.5 that under reparametrisation $\theta = g(\phi)$, the new Fisher information is

$$I_X^*(\phi) = (g'(\phi))^2 I_X(\theta)$$

i.e. $\sqrt{I_X^*(\phi)} = \sqrt{I_X(\theta)}|g'(\phi)|$. This is similar to the transformation of variables for a probability density function. So if we let the prior distribution be

$$\pi(\theta) \propto \sqrt{I_X(\theta)}$$

then the transformed prior

$$\tilde{\pi}(\theta) \propto \pi(g(\phi))|g'(\phi)| = \sqrt{I_X(\theta)}|g'(\phi)| = \sqrt{I_X^*(\phi)}$$

so that choice is invariant. In fact, it is consistent [13]. The prior is called *Jeffrey prior*.

An example in the lecture notes shows that Jeffrey prior for λ in Poisson distribution is $\pi(\lambda) \propto \lambda^{-1/2}$. This distribution seems to be informative, as it favours λ closer to 0. But recall the meaning of Fisher's information. $I_{X_1}(\lambda) = 1/\lambda$ so when λ is small, the data gives more information about λ. So we assign more densities in the prior toward 0 so that the update by data (posterior distribution) is more accurate.

7.2.2 Entropy and Max Entropy Prior

Entropy quantifies how much information would be given on average by an event. For events with lower probabilities, like the moon has been bombarded, you will be surprised and gain a lot of information from this event. In contrast, if you see the sun rising from the east, there is no extra information for you because this happens every day, it has a high probability. In general, we wish the entropy of events with probability 1 to be 0, and the entropy of events with low probabilities to be high.

There are many decreasing functions, but we wish our information function $h : \mathcal{F} \to \mathbb{R}^+$ (negative information does not make sense) to satisfy $h(E \cap F) = h(E) + h(F)$ where E, F are two independent events. i.e. we wish the information of E, F happening together is exactly the sum of information given by E and information given by F. The probability of two independent events satisfies $P(E \cap F) = P(E)P(F)$, so clearly, h should involve probability and logarithm. Note information decreases with probability, so let's try $h(E) := \log_2(1/P(E)) = -\log_2(P(E))$ (other base also works, 2 is a conventional choice in information theory) Note $1/P(E) \in [1, \infty)$, so $h(E) \geq 0$ and $h(E) = 0$ iff $P(E) = 1$.

The entropy of a random variable is defined as the expected value of information h, i.e. for discrete variables

$$\text{entropy}(X) = E_X(h(X)) = -\sum_x P(x)\log_2(P(x))$$

if the number of possible values of X is fixed, this quantity is maximised when all probabilities are equal. For example, if $X \in \{0, 1\}$, the entropy is maximised when $P(X = 0) = P(X = 1) = 0.5$. Figure 4 is a plot of entropy against $P(X = 0), P(X = 1)$.

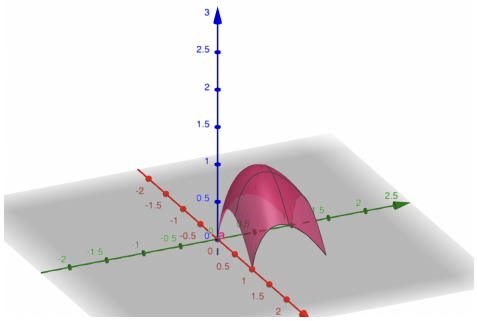

Figure 4: Demonstration of entropy on a discrete variable with two possible values

So if the entropy of distribution is maximised, on average information is low. The non-informative prior *maximum entropy prior* refers to the prior distribution that maximises the entropy under certain constraints. For example, one may wish that $E_\pi(\theta) = \mu$, $\mathrm{Var}_\pi(\theta) = \sigma^2$, then together with the restriction that π integrates to 1, the entropy is maximised when $\pi \sim N(\mu, \sigma^2)$. i.e. the normal distribution maximises entropy when the mean and variance are fixed, it is the most boring distribution (least informative). This makes sense as the central limit theorem says the distribution of the sample mean always approaches the normal distribution as sample size $n \to \infty$, no matter what distribution you are sampling from.

Theorem 8.5 in the lecture notes provide a quick way to find the maximum entropy prior

- Suppose all constraints are of the form

$$\int T_i(x)\pi(\theta)\, dx = t_i, \quad i = 1, \cdots, p$$

  for some functions T_i and constants t_i

- starting with

$$\pi(\theta) = \exp\left\{\sum_{i=1}^{p} \lambda_i T_i(\theta)\right\}$$

  plug it into the constraints and figure out the values of λ_i

- normalise the final distribution if necessary, or just report

$$\pi(\theta) \propto \exp\left\{\sum_{i=1}^{p} \lambda_i T_i(\theta)\right\}$$

  if normalisation is cumbersome

7.3 Posterior Predictive distribution

Suppose we have found the posterior distribution $p(\theta|x)$ where $x = (x_1, \cdots, x_n)$ and a new data point x_{n+1} comes in. We wish to predict its distribution, i.e. find $p(x_{n+1}|x)$. By the law of total probability,

$$p(x_{n+1}|x) = \int_\Theta p(x_{n+1} \,|\, \theta, x)\pi(\theta|x)\, d\theta$$

If we assume the new data is independent of the rest, then the formula is simplified to

$$p(x_{n+1}|x) = \int_\Theta p(x_{n+1} \,|\, \theta)\pi(\theta|x)\, d\theta$$

You can obtain a prior predictive distribution similarly. Usually, Bayes statistics involve many rounds of model update, so the posterior this round becomes the prior to the next round.

8 Bayesian Decision Theory

The decision system is a general framework that describes how we make decisions in estimations, hypothesis testing or confidence sets(general version of confidence interval) etc. In point estimation, we decide which point from the parameter space Θ is a good estimator for θ, given the data X. In hypothesis testing, we decide whether to reject a hypothesis or not (let 0 be not rejected and 1 for rejection). The ingredients we need are

- Parameter space Θ, sample space $\mathscr{X}$
- Statistical Model $\{f_\theta : \mathscr{X} \to [0,1] \,|\, \theta \in \Theta\}$
- Action space $\mathcal{A}$ (pick a decision in this space)
- Decision rule $\Delta : \mathscr{X} \to \mathcal{A}$. i.e. give data X, we make decision $\Delta(X) \in \mathcal{A}$.

For now, we restrict to the deterministic decision rule. But it is possible to define a *randomised decision rule*. For example, instead of saying "reject the null hypothesis" ($\Delta(X) = 1$), the decision can be: reject the null hypothesis with a probability of 0.7, so now the decision $\Delta(X)$ is a random variable with its probability distribution.

(WARNING: the following are from probability theory. If you don't know, it will not affect your understanding of the rest, skip this paragraph.) Recall the event space $\mathcal{F}$ is a σ-algebra on the sample space $\mathscr{X}$, then probability measure can be defined as a function $P : \mathcal{F} \to [0,1]$. We can define a probability measure on the action space $P : \mathfrak{A} \to [0,1]$ where $\mathfrak{A}$ is a σ-algebra on the action space $\mathcal{A}$. And denote the set of all probability measures by $\mathcal{P}(\mathcal{A})$.

Definition 17 (Randomised decision rule)**.** The randomised decision rule is $\mathfrak{d} : \mathscr{X} \to \mathfrak{d}_x \in \mathcal{P}(\mathcal{A})$ that is a measurable function. (i.e. for each $A \in \mathcal{A}$, $x \mapsto \mathfrak{d}_x(A)$ is measurable)
Note: $\mathfrak{d}_x$ is like a pdf defined on the action space.

With the randomised decision rule, the loss should be modified to

$$E_{\mathfrak{d}_x}[L(\theta, a)] = \int_{\mathcal{A}} L(\theta, a)\mathfrak{d}_x(a)\, da$$

Table of Statistical methods viewed as decision system

Method	Action Space $\mathcal{A}$	decision rule Δ
Point Estimation	Θ	$\Delta(X) = \hat{\theta}(X)$ i.e. the point estimator
Hypothesis Testing	$\{0,1\}$	$\Delta(X)$ is the hypothesis test
Confidence Set	$\mathcal{P}(\Theta)$, subsets of Θ	$\Delta(X) = C(X)$, the confidence set

The loss function, similar to the one in machine learning, penalises the decision for deviating from the true parameter θ. So it is a function $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$, where higher loss corresponds to decision $a \in \mathcal{A}$ not matching the true parameter θ. For example, the square loss function and absolute error loss are used for point estimation

$$L(\theta, a) = (a - \theta)^2,\ L(\theta, a) = |a - \theta|$$

For hypothesis testing, the trivial loss function is

$$L(a, \theta) = 1_{a \neq \theta}$$

i.e. wrong decision, loss is 1, correct decision, loss is 0. Loss functions for confidence set estimations are less obvious, some examples can be found in [2] and [15].

8.1 Risks and "good" decision rules

Consider the loss of a decision rule, i.e. $L(\Delta(X), \theta)$, there are two sources of randomness in this expression: X is a random variable (the frequentist's view), θ has its random distribution(the Bayesian's view). But to measure the quality of the decision rule Δ, we need to remove both randomness. Several risk functions will be defined, which discover the properties of the loss function under the randomness in θ and X. Figure 5 depicts all the risks and their relations.

Start with the frequentist's approach, eliminating the randomness of X. This is easy. Given a θ, the distribution is $X \sim f_\theta$. The average loss across all samples is called the *(frequentist's) risk*.

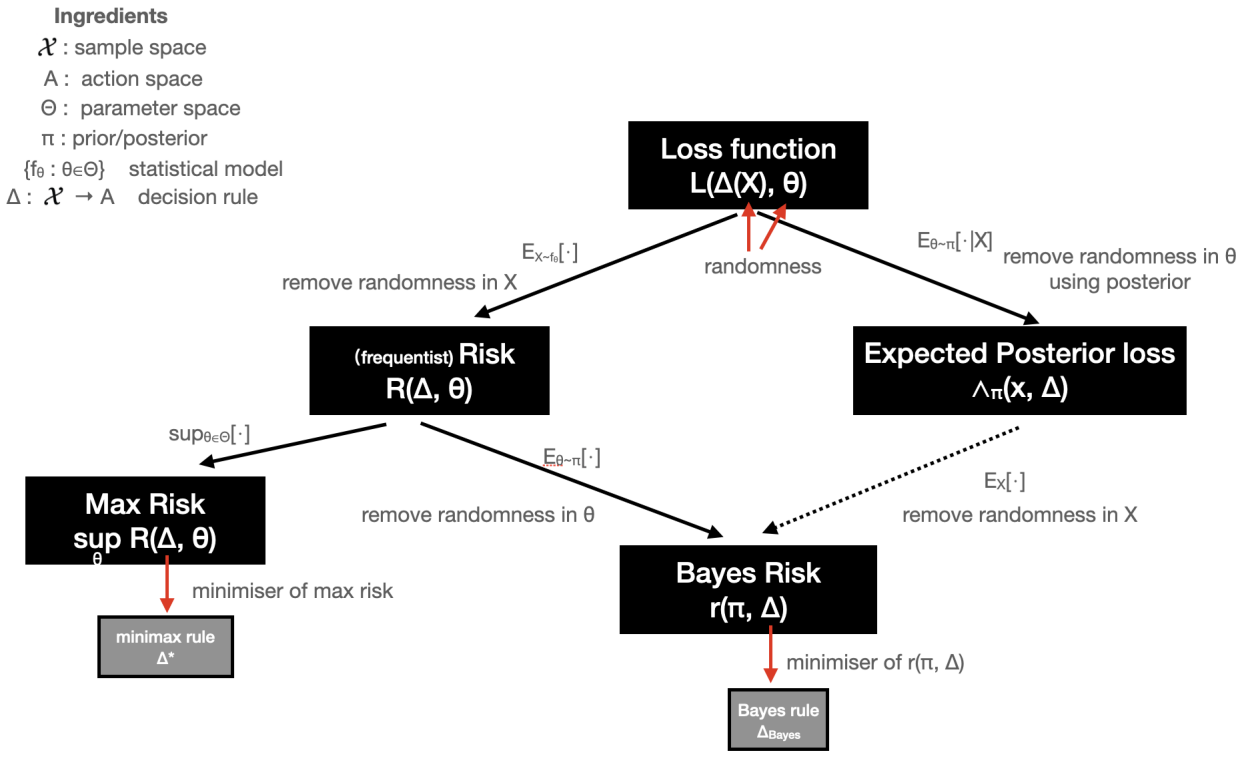

Figure 5: Relations of the loss function and all the risk functions

Definition 18 (Risk)**.** The risk $R(\theta, \Delta)$ is defined as

$$R(\theta, \Delta) = E_\theta[L(\Delta(X), \theta)] = \int_{x\in\mathcal{X}} L(\Delta(X), \theta) f_\theta(x)\, dx$$

if X is discrete, replace the integral with sum. (for the rest of this section, X is assumed to be continuous. But the discrete case is always simply replacing the integral with sum)

Example 4. For point estimation, the risk of square loss is

$$R(\theta, \Delta) = E_\theta[(\theta - \Delta(X))^2] = \mathrm{MSE}_\theta(\Delta)$$

i.e. the mean-squared error of Δ as an estimator.

For hypothesis testing with the natural loss,

$$R(\theta, \Delta) = E_\theta[1_{\Delta(X)\neq\theta}] = P_\theta(\Delta(X) \neq \theta)$$

i.e. the probability of type I and type II errors.

Risk can also be assigned to randomised loss in a similar way.

Before removing randomness in θ, note some decision rules are worse for any θ, such rules are called *inadmissible*

Definition 19. If

$$R(\theta, \Delta_1) \geq R(\theta, \Delta_2)\ \forall\theta \in \Theta, \qquad \text{and } \exists\theta \text{ s.t. } R(\theta, \Delta_1) > R(\theta, \Delta_2)$$

then Δ_1 is *strictly dominated* by Δ_2, and Δ_1 is *inadmissible*. Any rule that is not inadmissible is *admissible*

One way to further remove the randomness in θ is to find the worst-case scenario, i.e. the maximum risk among all parameters θ. Note the max risk $\sup_\theta R(\theta, \Delta)$ is only dependent on Δ, it is a quality measure for decision. The best decision should have the lowest risk, so define the *minimax rule* (or minimax estimator) to be the minimiser of max risk.

Definition 20 (Minimax Rule)**.** Δ^* is the minimiser of max risk, i.e.

$$\Delta^* = \arg\min_{\Delta \in \mathcal{A}} \sup_{\theta \in \Theta} R(\theta, \Delta)$$

Another way is to incorporate Bayesian statistics. Assume π is the prior distribution of θ, the expectation of risk under this distribution is the *Bayes integrated risk*.

Definition 21 (Bayes risk)**.** The Bayes (integrated) risk $r_\pi(\Delta) = r(\pi, \Delta)$ is defined as

$$r(\pi, \Delta) := E_\pi[R(\theta, \Delta)] = \int_\Theta R(\theta, \Delta)\pi(\theta)\, d\theta$$

Again, the Bayes risk is a measure of the quality of a decision. So we can define the Bayes rule(or Bayes estimator) w.r.t π as the best rule in the sense of having minimum Bayes risk.

Definition 22 (Bayes rule)**.** Bayes rule Δ_{Bayes} is the minimiser of Bayes risk, i.e.

$$\Delta_{\text{Bayes}} := \arg\min_{\Delta \in \mathcal{A}} r_\pi(\Delta)$$

and the risk for Bayes rule, $r_\pi(\Delta_{\text{Bayes}})$, is denoted r_π

The choice of prior matters here. If a prior gives a higher risk than all other priors, it is *least favourable*.

Definition 23 (Least Favourable Prior)**.** A prior π is least favourable if $r_\pi \geq r_{\pi'}$ for all other priors π'

When there are only two risks, and the action space is finite, the risks and estimators ("good" rules) described above can be visualised on a 2-D plot. Please find more details in [16].

8.2 Relationships between the "good" rules

So far, we have three "good" rules: Bayes, minimax, and admissible. It is worth studying how they are related to each other.

8.2.1 Bayes rule and Minimax rule

Finding a minimax rule is not easy as it involves two layers of optimisations. Bayes risk can help find the minimax estimator. Note, by definition, the max risk is always greater than the Bayes risk (which is the mean of risk under distribution π) for any decision rule. See figure 6 for an example. Therefore, if the max risk of a rule Δ_0 is a lower bound of the Bayes risk of a Bayes estimator, it must be minimax.

Theorem 8.1 (Using Bayes risk to find minimax)**.** *Fix a prior distribution π, suppose Δ_{Bayes} is a Bayes estimator with Bayes risk r_π, then any rule Δ_0 satisfying*

$$\sup_\theta R(\theta, \Delta_0) \leq r_\pi$$

is minimax. Further, if Δ_{Bayes} is unique, the Δ_0 is the unique minimax rule.

Proof. Target: prove Δ_0 is the minimiser of max risk. Suppose Δ is any rule,

$$\begin{aligned}
\sup_\theta R(\theta, \Delta) &\geq E_\pi[R(\theta, \Delta)] = r(\pi, \Delta) \\
&\geq r(\pi, \Delta_{\text{Bayes}}) \quad \text{by the definition of Bayes rule as a minimiser} \\
&= r_\pi \geq \sup_\theta R(\theta, \Delta_0)
\end{aligned}$$

If Δ_{Bayes} is the unique Bayes rule (i.e. unique minimiser of Bayes risk), then the second inequality is strict. We have $\sup_\theta R(\theta, \Delta_0) < \sup_\theta R(\theta, \Delta)$ for any decision rule Δ, and so Δ_0 is the unique minimiser of max risk. □

On the other hand, if the Bayes risk of a Bayes rule Δ_{Bayes} attains the upper bound (max risk), then Δ_{Bayes} must be a minimax rule.

Theorem 8.2 (Condition for Bayes estimator being minimax)**.** *If Δ_{Bayes} is a Bayes estimator for prior π satisfying*

$$R(\theta, \Delta_{Bayes}) \leq r_\pi \quad \forall \theta \in \Theta$$

i.e. $\sup_\theta R(\theta, \Delta_{Bayes}) = r_\pi$*, then Δ_{Bayes} is minimax rule and in that case, π is least favourable. Further, if Δ_{Bayes} is a unique Bayes estimator, then it is also a unique minimax rule.*

Proof. Apply the last theorem with $\Delta_0 = \Delta_{\text{Bayes}}$.

From the equation $\sup_\theta R(\theta, \Delta_{\text{Bayes}}) = r_\pi$, prior π is doing the worst job in the sense that it maximises the risk. So any other prior π' can do better than π. i.e. π is least favourable. □

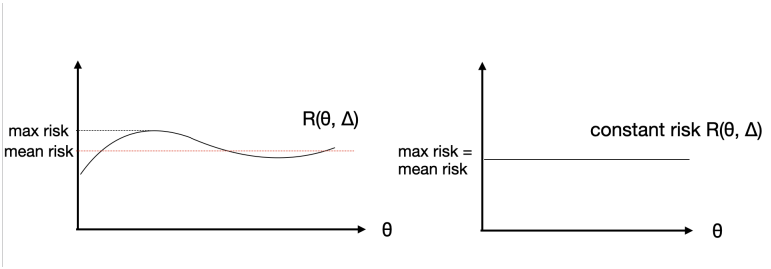

Figure 6: Visual comparison between the max risk and Bayes risk(the mean risk). If π is the uniform prior, then Bayes risk is simply the mean of risk as a function.

The condition in Theorem 8.2 can be trivially satisfied if $R(\theta, \Delta)$ is independent of θ. (illustrated at the right of figure 6) We can choose the prior π freely. So if the region on which the Bayes risk does not attain max risk is assigned probability (measure) 0, then the condition in Theorem 8.2 can also be satisfied.

Corollary 4. *If $\omega_\pi \subset \Theta$ is the set of θ at which the Bayes risk of Bayes estimator Δ_{Bayes} attains maximum risk, and $\pi(\omega_\pi) = 1$, then Δ_{Bayes} is minimax rule.*

constant risk ($R(\theta, \Delta)$ independent of θ) is a special case of the corollary, where $\omega_\pi = \Theta$ for any prior π.

8.2.2 Bayes rule and admissibility

After obtaining a Bayes rule, we need to check whether it is admissible or not.

Proposition 8.3 (Admissible Bayes rules)**.** *Bayes rule Δ_{Bayes} w.r.t. π with finite Bayes risk (i.e. $r_\pi(\Delta_{Bayes}) < \infty$) is admissible if either one of the following holds*

- *Δ_{Bayes} is the unique Bayes rule*
- *risk $\theta \mapsto R(\theta, \Delta)$ is continuous for all decision rule Δ, and prior π has positive density w.r.t. Lebesgue measure.*

Proof. **Unique Bayes rule ⇒ admissible**
Prove by contra-positive, suppose Δ_{Bayes} is not admissible, i.e. exists Δ s.t.

$$R(\theta, \Delta) \leq R(\theta, \Delta_{\text{Bayes}}) \quad \forall \theta \in \Theta \quad \text{ and } \exists \theta \text{ s.t. } R(\theta, \Delta) < R(\theta, \Delta_{\text{Bayes}})$$

note, the second inequality ensures $\Delta \neq \Delta_{\text{Bayes}}$. Expectation preserves inequality, so the Bayes risks also satisfy

$$r_\pi(\Delta) \leq r_\pi(\Delta_{\text{Bayes}})$$

Therefore, Δ is also a Bayes risk. i.e. Δ_{Bayes} is not unique.

Continuous risk, π positive density ⇒ admissible
Prove by contradiction. Assume risk is continuous w.r.t θ and π has positive density w.r.t. Lebesgue measure, but Δ_{Bayes} is not admissible. Define Δ as above, and let the difference set

$$A_\Delta := \{\theta : R(\theta, \Delta) < R(\theta, \Delta_{\text{Bayes}})\} = \{\theta : d(\theta) < 0\} \neq \emptyset$$

where $d(\theta) := R(\theta, \Delta) - R(\theta, \Delta_{\text{Bayes}})$. The function d is continuous by assumption, so A_Δ contains an open set, which has a non-zero measure under the Lebesgue measure. So $\pi(A_\Delta) > 0$. But then,

$$P_\pi(R(\theta, \Delta) < R(\theta, \Delta_{\text{Bayes}})) > 0$$

which means

$$r_\pi(\Delta) = E_\pi[R(\theta, \Delta)] < E_\pi[R(\theta, \Delta_{\text{Bayes}})] = r_\pi[\Delta_{\text{Bayes}}]$$

contradicts the definition of Bayes rule Δ_{Bayes} as the minimiser of Bayes risk. □

Remark. An admissible decision rule is defined as being not inadmissible. So when proving theorems about admissibility, try to reverse the argument first. Starting with "Δ is inadmissible" is much easier.

8.2.3 Minimax and Admissibility

One can prove in a similar way to Proposition 8.3 that a unique minimax rule is always admissible. On the other hand, if an admissible decision rule Δ has constant risk, i.e. $R(\theta, \Delta)$ is independent of θ, then Δ is a minimax rule.

Proposition 8.4. *An admissible decision rule Δ with constant risk is a minimax rule.*

Proof. Easy exercise. (hint: start by assuming Δ is NOT minimax rule, find a contradiction to the admissibility) □

A summary of the relations between three types of "good" rules can be found in figure 7.

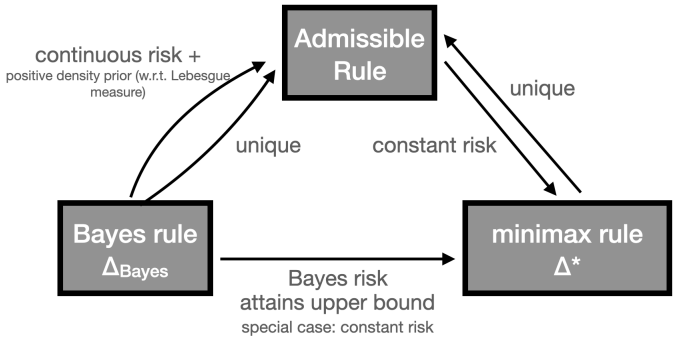

Figure 7: Relations between Bayes rule, minimax rule and admissible rule

8.3 The posterior approach

In section 7.4.1, we removed the randomness of θ and X in the loss $L(\Delta(X), \theta)$ by taking the frequentist's approach first. (i.e. taking the expectation w.r.t. $X \sim f_\theta$) But we can also take the Bayesian approach first. Pick a prior π, note X is not removed yet, so we have to take the expectation of the loss function w.r.t. the posterior distribution $\pi[\cdot|X]$

Definition 24 (Expected Posterior Loss)**.** The expected posterior loss (or posterior risk) is

$$\Lambda_\pi(x, \Delta) := E_\pi\left[L(\Delta(X), \theta) \,|\, X = x\right] = \int_\Theta L(\Delta(X), \theta)\pi(\theta \,|\, x)\, d\theta$$

But the problem is, removing randomness in X for $\Lambda_\pi(x, \Delta)$ is impossible, as the full distribution of X, $p(x) = f_X(x)$, is unknown. Let's view this from the Bayes integrated risk,

$$\begin{aligned} r_\pi(\Delta) &= \int_{\Theta\times\mathscr{X}} L(\Delta(X),\theta)p(x,\theta)\,dx\,d\theta && \text{by definition} \\ &= \int_\Theta \underbrace{\int_{\mathscr{X}} L(\Delta(X),\theta)f_\theta(x)\,dx}_{=R(\Delta,\theta)}\,\pi(\theta)\,d\theta \quad (1) && \text{as } p(x,\theta) = p(x|\theta)\pi(\theta) = f_\theta(x)\pi(\theta) \\ &= \int_{\mathscr{X}} \underbrace{\int_\Theta L(\Delta(X),\theta)\pi(\theta|x)\,dx}_{=\Lambda_\pi(x,\Delta)}\,\underbrace{p(x)}_{\text{unknown}}\,dx \quad (2) && \text{as } p(x,\theta) = \pi(\theta|x)p(x) \end{aligned}$$

(note: equations (1) and (2) correspond to the diamond of four quantities in figure 5)

Does it mean the posterior risk is useless? Fixing observation $X = x$, the minimiser of Posterior risk is easier to find than the minimiser of Bayes risk(i.e. Bayes rule) as we will see in an example later. Further, a minimiser of the posterior risk must also be a minimiser of the Bayes risk by the following theorem. Therefore, we can find the Bayes rule conveniently by minimising the posterior risk.

Theorem 8.5. *If a decision rule Δ^π minimises the posterior risk w.r.t. π for all $x \in \mathscr{X}$, then Δ^π is the Bayes rule*

Proof. Simply use equation (2) above,

$$r_\pi(\Delta) = \int_{\mathscr{X}} \Lambda_\pi(x,\Delta)p(x)\,dx$$

if $\Delta = \Delta^\pi$ minimises $\Lambda_\pi(x, \Delta)$ for all $x \in \mathscr{X}$, it also minimises RHS of the equation. Hence, it is a Bayes rule. □

Example 5. Consider the point estimation with decision rule $\Delta(X) = \hat{\theta}$ (the point estimator) and square loss function, the expected posterior loss (posterior risk) is

$$\begin{aligned} \Lambda_\pi(x,\hat{\theta}) &= E_\pi\left[(\hat{\theta}-\theta)^2\,|X=x\right] \\ &= E_\pi\left[(\hat{\theta}-\mu_x+\mu_x-\theta)^2\,|X=x\right] \quad \text{where } \mu_x := E_\pi[\theta|X=x] \text{ is posterior mean} \\ &= (\hat{\theta}-\mu_x)^2 + 2(\hat{\theta}-\mu_x)\underbrace{E_\pi[\mu_x-\theta\,|\,X=x]}_{=0 \text{ by definition of } \mu_x} + E_\pi[(\mu_x-\theta)^2\,|\,X=x] \\ &= (\hat{\theta}-\mu_x)^2 + \mathrm{Var}_\pi[\theta\,|\,X=x] \end{aligned}$$

The minimised of Λ is $\hat{\theta} = \mu_x$, the posterior mean of θ, and the associated posterior risk is the posterior variance $\mathrm{Var}_\pi[\theta\,|\,X=x]$. So μ_x is also the Bayes estimator by theorem 8.5.

What if we attempt to find the Bayes estimator by minimising the Bayes risk? The Bayes risk is

$$r_\pi(\hat{\theta}) = E_{\theta\sim\pi}\left[E_\theta[(\hat{\theta}(X)-\theta)^2]\right] = E_{\theta\sim\pi}[\mathrm{MSE}_\theta(\hat{\theta})]$$

where $E_\theta[(\hat{\theta}(X)-\theta)^2] = R(\theta,\hat{\theta})$ is the risk. But there is no good way to minimise this quantity even if you try hard. Two layers of expectation are cumbersome.

The same phenomenon occurs with absolute error loss $L(\theta,\hat{\theta}) = |\hat{\theta}-\theta|$ or the zero-one loss function

$$L(\theta,\hat{\theta}) = \begin{cases} a & \text{if } |\theta-\hat{\theta}| > b, \\ 0 & \text{otherwise} \end{cases}$$

The general reasons for the ease in minimisation of posterior risk are:

- heuristically, posterior risk takes the observed data into account. Therefore, it naturally produces a better estimate and easier form of the risk function
- The posterior distribution has a simpler parameter space as you add more data. Therefore, removing the randomness in θ is simpler compared to using prior distribution alone.

End of example.

The converse of theorem 8.5 is true provided that the Bayes risk of Δ^π is finite. (the technical proof of this result requires measure theory, so will be omitted here) Therefore, in this case, minimisations of posterior risk and Bayes risk are essentially equivalent.

Example 6. At the beginning of section 7.1, we used Beta(α, β) prior for data $X_1, \cdots, X_n \sim$ Bern(θ). By the above example, the Bayes estimator is the posterior mean

$$\overline{\theta}_{\alpha,\beta} := E[\theta|x] = \frac{\sum x_i + \alpha}{n + \alpha + \beta} = \frac{n}{n+\alpha+\beta}\overline{X}_n + \frac{\alpha+\beta}{n+\alpha+\beta}E(\theta)$$

which approaches the MLE $\overline{X}_n$ as $n \to \infty$. But when $\alpha + \beta \to \infty$, the posterior mean is the same as the prior mean $\alpha/(\alpha+\beta)$.

Previously we were struggling to decide good values of α and β, now we have another way to choose: aim to find α, β s.t. the risk $R(\theta, \overline{\theta}_{\alpha,\beta})$ is constant. This ensures the posterior mean is also a minimax rule.

9 Empirical Bayes Methods

We have learnt two ways of choosing prior distribution:

- picking a distribution that is conjugated to the likelihood function and choosing the parameters according to the decision theory framework.
- be lazy and pick a non-informative prior

In this section, a method of choosing prior based on previous data (or experience) will be introduced.

9.1 Robbin's formulae

The Bayes estimator of a point estimation problem under the square loss is the posterior mean. Robbin proposed a way to find this Bayes estimator even without having to explicitly access a prior distribution [12].

Suppose $X \sim$ Poi(θ) with $\theta \sim \pi$ for some unknown prior distribution π. The marginal distribution for x is

$$\begin{aligned} p(x) &= \int_\Theta p(x|\theta)\pi(\theta)\, d\theta \\ &= \int_\Theta \frac{e^{-\theta}\theta^x}{x!}\pi(\theta)\, d\theta \end{aligned}$$

The posterior mean is given by

$$\begin{aligned} E(\theta|x) &= \int_\Theta \theta\pi(\theta|x)\, d\theta \\ &= \int_\Theta \theta\frac{\pi(\theta)f(x|\theta)}{p(x)}\, d\theta \quad \text{by Bayes formulae} \\ &= \frac{1}{p(x)}\int_\Theta \frac{e^{-\theta}\theta^{x+1}}{x!}\pi(\theta)\, d\theta \\ &= (x+1)\frac{\int_\Theta \frac{e^{-\theta}\theta^{x+1}}{(x+1)!}\pi(\theta)\, d\theta}{\int_\Theta \frac{e^{-\theta}\theta^x}{x!}\pi(\theta)\, d\theta} \\ &= (x+1)\frac{p(x+1)}{p(x)} \end{aligned}$$

Marginal distribution is unknown, but if you have enough data, estimation is possible. Suppose n data $\{x_1, \cdots, x_n\}$ are collected, define the empirical distribution $p_n(x) :=$ (number of data points equal x) / $n = |\{i = 1, \cdots, n \mid X_i = x\}|/n$. Then the posterior mean can be estimated by

$$\hat{\theta}_n(x) = (x+1)\frac{p_n(x+1)}{p_n(x)}$$

and $\hat{\theta}_n(x) \to E(\theta|x)$ when $n \to \infty$ in this case.

In the above derivations, the prior π is hidden in the integrals. This is more like an algebraic trick, and Robbins showed similar manipulations work for binomial and geometric distributions. In general, it works for any Laplacian type distribution with the form $p(x|\theta) = e^{\theta x} f(x) h(\lambda)$.

Example 7. Suppose X is the number of traffic accidents per day at an intersection following a Poisson distribution with estimator λ. The data collected over 100 days is as follows: 70 days with 0 accident, 20 days with 1 accident, 7 days with 2 accidents and 3 days with 3 accidents. So the Bayes estimator (posterior mean) is

$$\hat{\theta}_{100}(0) = (0+1)\frac{p_{100}(1)}{p_{100}(0)} = \frac{20}{70} = 0.286$$

similarly $\hat{\theta}_{100}(1) = 0.7, \hat{\theta}_{100}(2) = 1.29$. $\hat{\theta}_{100}(x)$ cannot be estimated for $x \geq 3$, because we have not observed those data. This is one restriction for Robbin's formulae.

Notably, the data of 1-accident days is used to estimate the posterior mean for 0-accident days. Similarly, 1-accident days depend on 2-accident days etc. Whether such correlations exist is questionable.

Direct usage of Robbin's formulae in data analysis is context-related, and may not be a stable estimation. However, the core idea behind Robbin's formulae is that a large enough sample embeds a prior distribution implicitly.

9.2 Parametric Empirical Bayes

Parametric Empirical Bayes (PEB) method assigns a parameter ϕ to the prior distribution, $\theta \sim \pi(\theta, \phi)$. For example, if Beta$(\alpha, \beta)$ the prior of θ, the hyperparameter $\phi = (\alpha, \beta)$. ϕ is called hyperparameter because it is a parameter for the parameter θ. Such structure is called *hierarchical Bayes*. The hyperparameter will be estimated from the data, which can be a frequentist's estimator (MLE, MME etc.)

1. Choose a hierarchical Bayes structure, picking an appropriate prior distribution $\pi(\theta, \phi)$
2. Find an estimate $\hat{\phi}(x)$ from the data x
3. find the posterior distribution by
$$\hat{\pi}(\theta|x) \propto L(\theta, x)\pi(\theta, \hat{\phi}(x))$$

The data x is used twice in the estimation of the posterior distribution. Once in estimating the prior distribution and another in updating the prior.

In contrast to PEB, Robbin's approach was named non-parametric Empirical Bayes(NPEB) because a hyper-parameter is not assigned to the prior. Such methods do not explicitly specify the prior distribution. Another early application of NPEB is given by Good and Toulmin [4] for the missing species problem.

A summary of the early applications of PEB is given by Morris[10]. Stein's method is a remarkable application that came even before the term PEB was proposed.

9.2.1 James-Stein Estimator

Suppose there are independent variables $X_i \sim N(\mu_i, 1)$, and for simplicity assume only a single observation is made $x = (x_1, \cdots, x_n)$. (also define $\mu = (\mu_1, \cdots, \mu_n)$) Writing down the log-likelihood of joint distribution,

$$l(\mu_i) = -\frac{n}{2}\log(2\pi) - \sum_i \frac{(x_i - \mu_i)^2}{2}$$

it is not hard to see that $l(\mu|x)$ is maximised if $\mu_i = x_i$ for all i (i.e. $\hat{\mu}_{\text{MLE}} = x$). So the traditional MLE approach suggests estimating X_i's separately. For example, if X_i is the test score of student i, it makes sense to estimate the mean of his score μ_i based on only his past scores. This traditional approach had been used for a long time until James and Stein proposed a new estimator (JS estimator)[7]

$$\hat{\mu}_{\text{JSE}} = \left(1 - \frac{n-2}{\sum_{i=1}^{n} X_i^2}\right) X$$

The MLE estimator is unbiased, it has minimum variance among all unbiased estimators(MVUE). But is the MSE (risk under the square loss) optimal among all estimators (decision rules)? It turns out that the JS estimator beats MLE no matter what the true value of μ is as long as $n > 2$. (In decision theory, this indicates that the MLE estimator is inadmissible) So the JS estimator shocked the whole statistics community at that time. In our example, the JS estimator assumes that the mean test score of a student is dependent on all other students' scores.

The key to JSE beating MLE is that it gives up unbiasedness to reduce the variance. Recall that mean square error MSE = bias + var. JSE has a slightly higher bias compared to MLE, but it has a significantly lower variance, which is accomplished by the shrinkage. Note JSE takes the form $(1 - B)X$ where $B \in (0, 1)$ is a shrinkage factor. Figure 8 describes how shrinkage reduces MSE. The large red dot is the true mean μ_i and the black dots are observations of the variable X_i. If estimator $\hat{\mu}_i := X_i$, its variance is the magnitude of the black dots' variations, which is roughly measured by the area of the red circle. This estimator has zero bias (red circle centred at the true mean) but a large variance. A shrinkage estimator, on the other hand, pulls all points towards the mean 0 and reduces the area of the circle. It sacrifices some bias to achieve lower variance, and remember the shrinkage has a squared effect on the circle area (because $n = 2$) This is the *bias-variance trade-off* that later became very popular in statistical learning and machine learning.

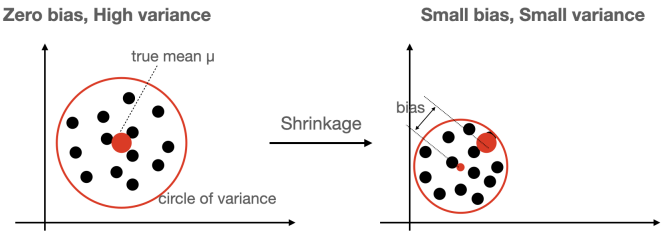

Figure 8: Illustration of how shrinkage reduces MSE

However, as you see in lower dimensions n, the trade-off may not be very beneficial. But for higher dimensions, shrinkage by $(1 - B)$ shrinks the volume of the sphere of variance by $(1 - B)^n$. This is the reason JSE is only used when $n > 2$.

Remark. The ideal case for shrinkage is when the true mean is exactly positioned at the origin, and shrinkage does not change bias at all. Since the true mean is unknown, standardisation of the data sample can be used (making the sample mean 0), which also pulls the true mean close to 0. The standardised data is $X - \overline{X}\mathbf{1}_n$ (each entry is simply $X_i - \overline{X}$), and JS estimator becomes

$$\mu_{\text{JSE}+}^{(\mu_0)} := \left(1 - \frac{a}{\|X - \overline{X}\mathbf{1}_n\|^2}\right)^{+} (X - \overline{X}\mathbf{1}_n) + \overline{X}\mathbf{1}_n$$

where a can be $n - 2$ or other values. The $(.)^+$ takes only the positive part of the expression to avoid "negative" shrinkage which makes the variance even larger. Note the mean is added back to restore the original scale. This estimator (called JS positive-part estimator) has a smaller MSE than the traditional JSE if a is picked appropriately. Therefore, many statistical learning methods work better if the data is standardised. (e.g. LASSO, ridge regressions)

Not surprisingly, JS positive-part estimator is still not the best estimator, because the shrinkage used is quite naive. Some further improvements are discussed in [14].

9.2.2 Connection to PEB

The JSE estimator can be obtained from a parametric empirical Bayes method, we assumed $X_i \mid \mu_i \sim N(\mu_i, 1)$, i.e. $X_i = \mu_i + \epsilon_i$ with $\epsilon_i \sim N(0,1)$. Now assign a prior distribution $N(0, \tau^2)$ to all μ_i. The marginal distribution on X_i is $N(0, 1+\tau^2)$ (assume μ_i are independent from ϵ_i) so in fact $\mu_i | x_i$ will also be a normal distribution.

Calculating $\pi(\mu_i|x_i) \propto L(\mu_i, x_i)\pi(\mu_i)$, the parameters of posterior distribution can be found by completing the square.

$$\mu_i | x_i \sim N\left((1 - \frac{1}{1+\tau^2})x_i, \frac{\tau^2}{1+\tau^2}\right)$$

which yields the posterior mean (Bayes estimator) $\hat{\mu}_i = (1 - 1/(1+\tau^2))x_i$. τ^2 can be estimated from the data. Note

$$\sum_i \frac{x_i^2}{1+\tau^2} \sim \chi_n^2$$

where χ_n^2 is a chi-squared distribution with a degree of freedom n.
It can be proved that $E(1/\chi_n^2) = n - 2$ (search online if you don't know how to prove), so a reasonable estimator for $1 + \tau^2$ is $\sum_i x_i^2/(n-2)$. Substituting back to the posterior mean yields exactly the JSE.

Of course, other estimators for τ can be used. The MLE estimator is

$$\hat{\tau}^2 = \frac{\sum_i X_i^2 - 1}{p}$$

In the rest of this section, a mathematical proof of why the James-Stein estimator always has lower MSE will be provided.

Theorem 9.1 (James-Stein Theorem)**.** *The James-Stein estimator strictly dominates the MLE w.r.t. quadratic loss function.*

Proof. The squared loss for any estimator $\hat{\mu}$ is $\sum_i (\hat{\mu}_i - \mu_i)^2$, use bridging to include the data x_i:

$$\begin{aligned} L(\hat{\mu}(\boldsymbol{x}), \mu) &= \sum_i (\hat{\mu}_i - x_i + x_i - \mu_i)^2 \\ &= \sum_i (x_i - \hat{\mu}_i)^2 + (x_i - \mu_i)^2 - 2(x_i - \hat{\mu}_i)(x_i - \mu_i) \end{aligned}$$

So the risk is

$$\begin{aligned} R(\hat{\mu}, \mu) &= E_x(L(\hat{\mu}(x), \mu)) \\ &= E(\|\boldsymbol{x} - \hat{\mu}\|^2) + E(\|\boldsymbol{x} - \mu\|^2) - 2\sum_i E[(x_i - \hat{\mu}_i(x))(x_i - \mu_i)] \\ &= E(\|\boldsymbol{x} - \hat{\mu}\|^2) + \mathrm{Var}(\boldsymbol{x}) - 2\sum_i E\left[\frac{\partial(x_i - \hat{\mu}_i(x))}{\partial x_i}\right] \end{aligned}$$

where the last term is obtained by Stein's identity (proved by integration by part) and it is true as long as $\hat{\mu}_i(x)$ is bounded differentiable.

Substituting $\hat{\mu}$ by MLE and JSE yields

$$R(\hat{\mu}_{\mathrm{MLE}}, \mu) = \mathrm{Var}(\boldsymbol{x}) = N$$

and

$$R(\hat{\mu}_{\mathrm{JSE}}, \mu) = \mathrm{Var}(\boldsymbol{x}) - E\left[\frac{(n-2)^2}{\|\boldsymbol{x}\|^2}\right] = N - (n-2)^2 E\left[\frac{1}{\|\boldsymbol{x}\|^2}\right]$$

Therefore, $R(\hat{\mu}_{\mathrm{JSE}}, \mu) < R(\hat{\mu}_{\mathrm{MLE}}, \mu)$ for any μ. □

.

Chapter Comment

As a closing for this chapter, both Robbin's method and Stein's method use some property of the prior distribution. (Robbin's method uses algebraic tricks on Laplacian-like PDFs, whereas Stein's method assumes normality) But in fact, the empirical Bayes method applies to any prior distribution in the exponential family, and

$$E(\mu_i|x_i) = x_i + \frac{\partial \log(f(x_i))}{\partial x_i}$$

where f is the marginal distribution of x. The term x_i corresponds to the MLE estimator and the second term is the *Bayesian correction*. This formula is called *Tweedie's formula*. So EB combines the Bayesian approach and the frequentist approach. EB is still a developing field and there are many beautiful applications.

10 Hypothesis Testing

Recall that hypothesis testing decides whether the null hypothesis $H_0(\theta \in \Theta_0)$ should be accepted or rejected against an alternative hypothesis H_1 $(\theta \in \Theta_1)$. If an event with a very low probability under the null assumption is observed, we reject H_0.

Example 8. We are testing whether a lady can taste the difference between wines. H_0: she cannot taste the difference, H_1: she can taste the difference. Out of five cups of wine (poured from 4 different bottles), she correctly identified 4 of them. Under H_0 (the lady knows nothing about wines), she only takes random guesses and the probability of getting 4 out of 5 cups right is

$$\left(\frac{1}{4}\right)^4 \frac{3}{4} = 0.29\%$$

which is pretty low. We have enough evidence that the lady can taste wines. (reject H_0)

Usually, a nice statistic $T(X)$ will be picked (for example, $T(X) = \overline{X}$ the sample mean). In the example above, $T(x)$ = number of cups the lady guessed right out of 5 cups of wine. For an observed value x, the p-value $p_x := P(\text{observing } T(x)|H_0)$ is the probability of observing this value of $T(x)$(or even a more extreme value) assuming the null hypothesis is true, and H_0 will be rejected if $p_x < \alpha$ for some small value α called *significance level*.

More generally, a hypothesis test uses a set $C \subseteq \mathscr{X}$ on which if the observed $x \in C$, the null hypothesis H_0 will be rejected. C is called the *critical region*.

Example 9. Suppose statistics $T(X)$ has a monotone increasing PDF f (lower value of $T(x)$ are less likely) and the CDF is F under the null hypothesis, then $p_x = P(T(X) < T(x)|H_0) = F(T(x))$. So the critical region is

$$\begin{aligned} C &= \{x : p_x < \alpha\} \\ &= \{x : F(T(x)) < \alpha\} \\ &= \{x : x < T^{-1}(F^{-1}(\alpha))\} \quad \text{if both } T \text{ and } F \text{ are invertible} \end{aligned}$$

In this example, we have explicitly found an inequality for x. But this is not always the case.

10.1 Quality of Test

There are two possible errors we may make for hypothesis testing, illustrated in Figure 9.

Denote the probabilities of making Type I and Type II errors as $\alpha(\theta), \beta(\theta)$ respectively. (the probabilities depend on what the true parameter θ is) Ideally, we wish $\alpha(\theta), \beta(\theta)$ to both be small, but they are usually negatively correlated.

A conceptual explanation is: suppose H_0 : the suspect is not guilty, H_1 : the suspect is guilty. Accepting H_0 means releasing the suspect and rejecting H_0 means sentencing the suspect to death. If you make strict laws and the judge is ruthless, most guilty suspects will be identified and punished(low probability of Type II errors), but many innocent people will be killed (high probability of Type I errors). The trend is reversed if you make loose laws and the judge is lenient. A mathematical explanation is given below.

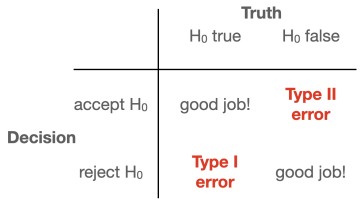

Figure 9: Type I and Type II errors in hypothesis testing

Definition 25 (Power function)**.** The power function for a true parameter θ is defined by

$$w_\phi(\theta) := P_\theta(x \in C) = P_\theta(\text{reject } H_0) = E_\theta[\phi(X)]$$

if the test ϕ in use is clear, the power function is denoted $w(\theta)$.

.

Under null hypothesis ($\theta \in \Theta_0$), the power is

$$w(\theta) = P_\theta(\text{reject } H_0 | H_0) = \alpha(\theta)$$

which equals the probability of making a Type I error. And the worst-case scenario (among all $\theta \in \Theta_0$) is called the *size* of the test, denoted by

$$\alpha := \sup_{\theta \in \Theta_0} w(\theta)$$

Under the alternative hypothesis ($\theta \in \Theta_1$),

$$w(\theta) = 1 - \beta(\theta) = \text{probability of making the correct decision to reject } H_0$$

.

Therefore, an ideal test have low $w(\theta)$ for $\theta \in \Theta_0$ and high $w(\theta)$ when $\theta \in \Theta_1$. But remember they are both controlled by the critical region C, and the true parameter is unknown. Shrinking C usually makes $w(\theta)$ lower for the whole parameter space. This is illustrated in figure 10, where the black curve (broken) illustrates an ideal test with low power on H_0 and suddenly rises on H_1, producing a low probability of Type II error.

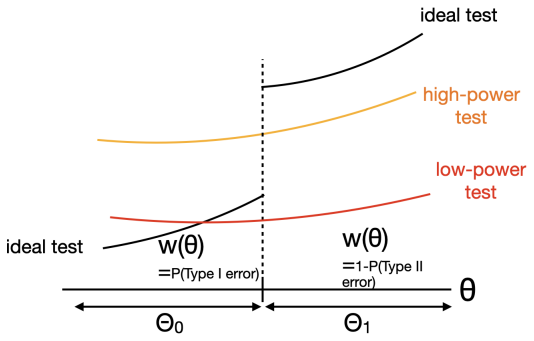

Figure 10: Explanation of negative correlation between probabilities of Type I and Type I errors, both null and alternative hypothesis are composite (not a single point)

Due to the negative correlation between two errors, a compromise is to find a test with acceptable behaviour on Θ_0 (below a threshold) and maximise the power on Θ_1.

Definition 26 (Uniformly Most Powerful Test)**.** A uniformly most powerful(UMP) test (for testing H_0 against H_1) at level α is the test ϕ s.t. (1) ϕ has size smaller than α, i.e.

$$w_\phi(\theta) \leq \alpha \quad \forall \alpha \in \Theta_0$$

(2) for any other test ϕ' with size below α, ϕ has higher power than it.

$$w_{\phi'}(\theta) \leq w_{\phi}(\theta) \quad \forall \alpha \in \Theta_1$$

An example of a UMP test is given in figure 11, with the power function being the bold curve. All other tests (red curves) have lower power on Θ_1.

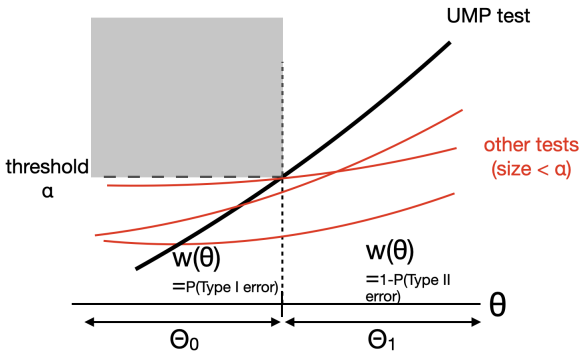

Figure 11: Illustration of UMP test. Any test with a power function crossing the grey area is banned from competing for the role of UMP at level α

For the rest of this chapter, we will find the UMP in various situations.

10.2 UMP for simple Hypothesis

Suppose H_0 and H_1 are simple (Θ_0, Θ_1 are singleton), then the Neyman Pearson theorem suggests that a likelihood ratio test(LRT) with size α is a UMP test at that level α. Conversely, a UMP test at that level α is (almost) a LRT.

Definition 27 (Likelihood ratio test)**.** Suppose $X \sim f_\theta$, for hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, the likelihood ratio is

$$\Lambda(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}$$

a larger $\Lambda(x)$ indicates that θ_1 is much more likely to generate the observation x then θ_0, providing evidence to reject H_0.
A likelihood ratio test takes the form

$$\phi(x) = \begin{cases} 1, & \Lambda(x) > k \\ 0, & \Lambda(x) \leq k \end{cases}$$

Theorem 10.1 (Neyman-Pearson)**.** *Suppose LRT ϕ_0 has size α, then it is the UMP at level α.*

Proof. Given another test ϕ with size not exceeding α, the aim is to prove $w_\phi(\theta_1) \leq w_{\phi_0}(\theta_1)$, equivalent to

$$E_{\theta_1}[\phi(X)] - E_{\theta_1}[\phi_0(X)] = \int \phi(x) f_{\theta_1}(x) - \phi_0(x) f_{\theta_1}(x)\, dx \leq 0$$

Note by assumption, size of test $\phi \leq \alpha$, so

$$\int \phi(x) f_{\theta_0}(x) - \phi_0(x) f_{\theta_0}(x)\, dx = E_{\theta_0}[\phi(X)] - E_{\theta_0}[\phi_0(X)] = E_{\theta_0}[\phi(X)] - \alpha \leq 0$$

The proof is complete if we can show

$$\int \phi(x) f_{\theta_1}(x) - \phi_0(x) f_{\theta_1}(x)\, dx \leq k \int \phi(x) f_{\theta_0}(x) - \phi_0(x) f_{\theta_0}(x)\, dx \quad (1)$$

where RHS is non-positive. Equivalently,

$$\int (\phi(x) - \phi_0(x))(f_{\theta_1}(x) - k f_{\theta_0}(x))\, dx \leq 0 \quad (2)$$

When $\phi_0(x) = 1$, $f_{\theta_1}(x) - k f_{\theta_0}(x) > 0$ (the critical region of LRT ϕ_0) and $\phi(x) - \phi_0(x) \leq 0$ because $\phi(x) \in \{0, 1\}$. Conversely, $\phi_0(x) = 0$ implies that $f_{\theta_1}(x) - k f_{\theta_0}(x) \leq 0$ and $\phi(x) - \phi_0(x) \geq 0$. So the integrand

$$U(x) := (\phi(x) - \phi_0(x))(f_{\theta_1}(x) - k f_{\theta_0}(x))$$

is always non-positive, and inequality (2) holds. □

Summarising from the proof, the LRT somehow restricts the power difference of two tests at θ_1 by that at θ_0, i.e.

$$w_\phi(\theta_1) - w_{\phi_0}(\theta_1) \leq k(w_\phi(\theta_0) - w_{\phi_0}(\theta_0))$$

and θ_0 is already controlled by our assumptions. This inequality allows us to pass that control onto θ_1.

Neyman-Pearson provides a handy way to obtain UMP. If a UMP at level α is required, find a value k s.t. the LRT has size α. But such k may not exist for discrete f_θ, where the likelihood ratio also becomes discrete. A randomised likelihood ratio test is required in such a situation.

Definition 28 (randomised Likelihood ratio test)**.** Suppose $X \sim f_\theta$, for hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, a randomised likelihood ratio test takes the form

$$\phi(x) = \begin{cases} 1, & \Lambda(x) > k \\ 0, & \Lambda(x) < k \\ \gamma, & \Lambda(x) = k \end{cases}$$

It is left to the readers to check that the randomised Likelihood ratio test with size α is still the UMP at level α.

Lemma 10.2. *For any size $\alpha \in (0, 1)$, there is always k s.t. the randomised LRT ϕ has size exactly α*

Proof. When f_θ is discrete, the CDF of the likelihood ratio $G(k) = P_{\theta_0}[\Lambda(X) \leq k]$ has some jumps ($G(k)$ measures the probability of accepting H_0 with θ_0 being true for non-randomised LRT, which means $G(k) = 1-$size). Therefore, the range of function $G(k)$ does not fill up $[0, 1]$, making it impossible to pick $G(k)$ s.t. $G(k) = 1 - \alpha$ for any $\alpha \in (0, 1)$. This is illustrated in Figure 12.

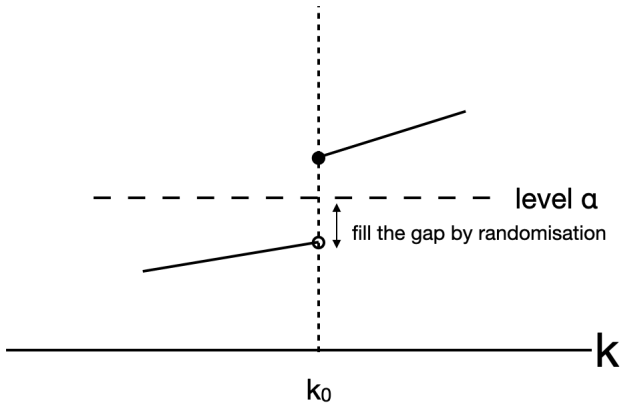

Figure 12: non-randomaised LRT may not achieve size α exactly

Now fix α (the level we require, e.g. 0.05), let k_0 be the jump that made $G(k)$ skip the value $1 - \alpha$, i.e. $G(k_0) > 1 - \alpha$, and $G(k_0 - \delta) \leq 1 - \alpha$ for any $\delta > 0$, γ in randomised LRT can help to fill the gap.

The size of the randomised LRT is

$$\begin{aligned} E_{\theta_0}[\phi] &= \gamma P(\Lambda(X) = k_0) + P(\Lambda(X) > k_0) \\ &= \gamma P(\Lambda(X) = k_0) + 1 - P(\Lambda(X) \le k_0) \end{aligned}$$

so picking

$$\gamma_0 = \frac{\alpha - 1 + P(\Lambda(X) \le k_0)}{P(\Lambda(X) = k_0))} = \frac{\alpha - 1 + G(k_0)}{G(k_0) - P(\Lambda(X) < k_0))}$$

ensures the randomised LRT has size α. □

Finally, we prove the converse of Neyman-Pearson: any UMP at level α is (almost surely) a randomised LRT.

Theorem 10.3. *Suppose the probability measures for simply hypothesis H_0 and H_1, denoted as $P_0 := P_{\theta_0}$ and $P_1 := P_{\theta_1}$ respectively, are absolutely continuous w.r.t. Lebesgue measure, then any UMP test at level α is P_0 and P_1-almost surely equals a randomised LRT.*

Remark. The fancy term "absolutely continuous w.r.t. Lebesgue measure" basically means the probability of any event defined on $\mathbb{R}$ with zero Lebesgue measure(e.g. a point event has zero measure, so does any event with finite points) is 0 for both P_0 and P_1. The term "P_0 and P_1-almost surely equal" means two things are equal except on an offset with zero probability (under P_0 and P_1).

Proof. Pick a randomised LRT ϕ_0 with size α as instructed by the lemma above (picking appropriate k_0 and γ_0). Suppose ϕ is any UMP test at level α. By definition, the power function for ϕ_0 and ϕ is the same at θ_1 ($E_{\theta_1}[\phi(X)] = E_{\theta_1}[\phi_0(X)]$), meaning that LHS of inequality (1) in the proof of theorem 10.1 is 0. The RHS term is non-negative, so it is sandwiched and guaranteed to be 0, i.e.

$$E_{\theta_0}[\phi(X)] = E_{\theta_0}[\phi_0(X)]$$

so the power functions are the same at θ_0 and θ_1, implying

$$\begin{aligned} \int U(x)\, dx &= \int (\phi(X) - \phi_0(X))(f_{\theta_1}(x) - k f_{\theta_0}(x))\, dx \\ &= \int (\phi(X) - \phi_0(X)) f_{\theta_1}(x)\, dx - k \int (\phi(X) - \phi_0(X)) f_{\theta_0}(x)\, dx \\ &= (E_{\theta_1}[\phi(X)] - E_{\theta_1}[\phi_0(X)])) - k(E_{\theta_0}[\phi(X)] - E_{\theta_0}[\phi_0(X)]) = 0 \end{aligned}$$

Recall that $U(x) \le 0$ (see the proof of theorem 10.1), so $\int U(x)\, dx = 0$ implies $U(x) = 0$ except on a set S with zero Lebesgue measure (e.g. discontinuous points).

Since P_0, P_1 are absolutely continuous, $P_0(S) = P_1(S) = 0$. For $x \in \mathcal{X} \setminus S$, $U(x) = 0$ so $\phi(x) = \phi_0(x)$ or $\Lambda(x) = k_0$. Although the γ value of the two tests may be different, the k value of test ϕ must be k_0, and $S \cup \{k_0\}$ is still a set with zero probability measure under P_0 and P_1. So $\phi = \phi_0$ P_0, P_1-almost surely. □

Notice in the last of the proof that the choice of γ in randomised LRT does not affect the power at θ_1, it is only there to adjust the size (power at θ_0). So we can pick $\gamma = 1$ and let the rejection region be $C = \{x : \Lambda(x) \ge k\}$. This test is UMP at level $\alpha := P_{\theta_0}(X \in C)$, and it is called a *Neyman-Pearson test*.

10.3 UMP for one-sided test

In this section, we focus on finding UMP for the *one-sided hypothesis testing*, formulated as below:

$$H_0 : \theta \le \theta_0 \quad H_1 : \theta > \theta_0$$

This sounds much more complicated than a simple hypothesis, especially controlling the power on $\Theta_0 = \{\theta \le \theta_0\}$ is cumbersome. But with monotonicity, the situation can be simplified to simple hypotheses testing and the Neyman-Pearson test can be applied.

Definition 29 (monotone likelihood ratio)**.** Family $\{f_\theta\}_{\theta \in \Theta \subset \mathbb{R}}$ satisfies monotone likelihood ratio (MLR) if there is a function $t(x)$ s.t. the likelihood ratio $x \mapsto f_{\theta_2}(x)/f_{\theta_1}(x)$ ($\theta_1 \le \theta_2$) is a non-decreasing function of $t(x)$. (implicitly assumes that $\Lambda(x)$ can be written in the form of $g(t(x))$ for some function g)

Example 10. Suppose $X_1, \cdots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with σ^2 known. For two means $\mu_1 < \mu_2$, the likelihood ratio of the joint distribution is

$$\begin{aligned}\frac{f_{\mu_2}(x)}{f_{\mu_1}(x)} &= \frac{(2\pi)^{-n/2}\sigma^{-n}\exp\left[-\sum_i (x_i-\mu_2)^2/(2\sigma^2)\right]}{(2\pi)^{-n/2}\sigma^{-n}\exp\left[-\sum_i (x_i-\mu_1)^2/(2\sigma^2)\right]} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_i (x_i-\mu_2)^2 - \sum_i (x_i-\mu_1)^2\right)\right\} \\ &= \exp\left\{\frac{(\mu_2-\mu_1)}{\sigma^2}\left(\sum_i x_i - \frac{n}{2}(\mu_1+\mu_2)\right)\right\}\end{aligned}$$

which is an non-decreasing function of $t(x) := \sum_i x_i$ because $\mu_2 > \mu_1$.

.
With MLR property, the Neyman-Pearson test ϕ has a rejection region

$$C = \{x : \Lambda(x) = f_{\theta_1}(x)/f_{\theta_0}(x) \geq k\} = \{x : t(x) \geq t\}$$

for some t (because $f_{\theta_1}(x)/f_{\theta_0}(x) = g(t(x))$ for some non-decreasing function g)

Karlin-Rubin theorem says with MLR property, the test with rejection region $C = \{t(x) \geq k\}$ is the UMP for the one-sided hypothesis testing. The basic idea is to show that power on Θ_0 can be controlled at θ_0 completely, and then the Neyman-Pearson test for θ_0 against $\theta \in \Theta_1$ all takes the same rejection region, and have the maximum power on Θ_1 by the Neyman Pearson theorem.

Theorem 10.4 (Karlin-Rubin)**.** *Suppose $X \sim f_\theta$ and $\{f_\theta\}_{\theta\in\Theta}$ satisfies MLR w.r.t. statistics $t(x)$. Then the UMP test at level α for one-sided hypothesis testing is*

$$\phi(x) = \begin{cases} 1, & t(x) \geq k, \\ 0, & t(x) < k \end{cases}$$

where k is a number s.t. $P_{\theta_0}(t(X) \geq k) = \alpha$. (such k must exists if $t(X)$ is continuous, but for discrete case, randomised test can be used)

Proof. **Control on Θ_0**
Let $C := \{x : t(x) \geq t\}$, then $P_\theta(x \in C)$ is non-decreasing w.r.t. θ (proved in the lemma 10.5 below). And if a test is constructed using C as the rejection region, the power $w(\theta) = P_\theta(x \in C) \leq P_{\theta_0}(x \in C) = w(\theta_0)$ for any $\theta \leq \theta_0$. So the size of the test must be $\sup_{\theta\leq\theta_0} w(\theta) = w(\theta_0)$, which indicates the power on $\Theta_0 = \{\theta \leq \theta_0\}$ is under control as long as power at θ_0 is controlled. Now aim to prove that the Neyman-Pearson test has maximum power on $\Theta_1 = \{\theta > \theta_0\}$.

Control on Θ_1
For any $\theta' > \theta_0$, consider the test with simple hypotheses

$$H_0' : \theta = \theta_0, \quad H_1' : \theta = \theta'$$

the Neyman-Pearson test $\phi_{\theta'}$ at level α has rejection region $\{f_{\theta'}(x)/f_{\theta_0}(x) \geq k'\} = \{t(x) \geq k\}$ for some k s.t. $P_{\theta_0}(t(x) \geq k) = \alpha$. ($t(x)$ is independent of θ', so k does not depend on θ', i.e. rejection region does not depend θ' and the test $\phi_{\theta'}$ can be simply denoted as ϕ for any θ') Because ϕ is UMP test (for testing H_0' against H_1'), for any other test ϕ^* with $w_{\phi^*}(\theta_0) \leq \alpha$, we have $w_{\phi^*}(\theta') \leq w_\phi(\theta')$. The assumption $w_{\phi^*}(\theta_0) \leq \alpha$ is trivially satisfied for any test (on the one-sided hypothesis testing) with size below α. Because

$$w_{\phi^*}(\theta_0) \leq \sup_{\theta\leq\theta_0} w_{\phi^*}(\theta_0) \leq \alpha$$

The choice of θ' is arbitrary, so $w_{\phi^*}(\theta) \leq w_\phi(\theta)$ for all $\theta > \theta_0$ and for any test with size α on H_0. This concludes that ϕ is UMP for testing H_0 against H_1 at level α. □

Lemma 10.5. *Under the conditions specified in Karlin-Rubin's theorem. If $C := \{x : t(x) \geq t\}$, then $P_\theta(x \in C)$ is non-decreasing w.r.t. θ*

Proof. Take any $\theta_0 < \theta_1$, the Neyman-Pearson test ϕ (for testing $\theta = \theta_0$ against $\theta = \theta_1$) has rejection region

$$C = \{x : \Lambda(x) = f_{\theta_1}(x)/f_{\theta_0}(x) \geq k\} = \{x : t(x) \geq t\}$$

for some t due to MLR. The power function of ϕ is $w_\phi(\theta) = P_\theta(x \in C)$. The aim is to prove $w_\phi(\theta_0) \leq w_\phi(\theta_1)$.

Neyman-Pearson theorem ensures ϕ is the UMP. So for any test ϕ' with $w_{\phi'}(\theta_0) \leq w_\phi(\theta_0) =: \alpha$, we have $w_{\phi'}(\theta_1) \leq w_\phi(\theta_1)$. If we pick a trivial test ϕ': flip a coin Y, if $Y = 1$, reject H_0. The coin does not need to be fair, denote $P(Y = 1) =: p$. Such ϕ' has constant power, $w_{\phi'}(\theta_1) = w_{\phi'}(\theta_0) = p$. Let $p = a$, ϕ' satisfies the condition $w_{\phi'}(\theta_0) \leq \alpha$ and so

$$\begin{aligned} w_\phi(\theta_0) = \alpha &= w_{\phi'}(\theta_1) \\ &= w_{\phi'}(\theta_1) \quad \text{because } \phi' \text{ is randomised} \\ &\leq w_\phi(\theta_1) \quad \text{because } \phi \text{ is UMP} \end{aligned}$$

□

Example 11. Again we play with the model $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known. Let the one-sided hypothesis testing be

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu \leq \mu_1$$

We have proved in the previous example that the PDF of X, f_θ, satisfies MLR for test statistics $t(X) = \sum_i X_i$ (which is continuous). By the Karlin-Rubin theorem, the UMP at level α must have the form

$$\phi(x) = \begin{cases} 1 & t(x) \geq k \\ 0 & t(x) < k \end{cases}$$

the only mission is to find k s.t.

$$P(\sum_i X_i \geq k \,|\, \mu = \mu_0) = \alpha$$

Note

$$\begin{aligned} P(\sum_i X_i \geq k \,|\, \mu = \mu_0) &= P(\overline{X}_n - \mu_0 \geq k/n - \mu_0 \,|\, \mu = \mu_0) \quad \text{Recall the sample mean } \overline{X}_n \sim N(\mu_0, \sigma^2/n) \\ &= P(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \frac{k/n - \mu_0}{\sigma/\sqrt{n}} \,|\, \mu = \mu_0) \\ &= P(Z \geq \frac{k/n - \mu_0}{\sigma/\sqrt{n}}) \quad \text{where } Z \sim N(0,1) \end{aligned}$$

so we need $(k/n - \mu_0)/(\sigma/\sqrt{n}) = z_{1-\alpha}$. By rearranging,

$$k = n\left(\mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right)$$

10.4 Bayesian Hypothesis Testing

Before finding UMP for two-sided tests, which is a hard task, it is worth mentioning the fact that Bayes statistics can be incorporated into hypothesis testing. More surprisingly, the Bayesian decision theory yields that the likelihood ratio test is the "best" test for simple hypotheses without knowing in advance that the likelihood ratio test is a good candidate. It's form $C = \{x : f_{\theta_1}(x) \geq k f_{\theta_0}(x)\}$ naturally arises when the Bayesian's view for hypothesis testing is taken.

Bayesian statistics view θ as a random variable, and assign probabilities to H_0 and H_1 (i.e. $P(\theta \in \Theta_0)$ and $P(\theta \in \Theta_1)$). These probabilities can be updated using data x, which is the core idea of Bayes statistics.
There are two ways to compare the two probabilities:

- pick the hypothesis which higher posterior probability $P(H_i : X = x)$. Such hypothesis test is *Maximum a posteriori*(MAP) test and will be discussed later.

- compare on the odd ratio scale, a higher odds (i.e. $P(H_0 \text{ is true})/P(H_1 \text{ is true})$) indicates less evidence to reject H_0.

We focus on the second one in this section, suppose the prior probabilities are π_0, π_1 respectively, by the Bayes formulae,

$$\begin{aligned} P(H_0 \,|\, X = x) &= \frac{P(X = x|H_0)P(H_0)}{P(X = x)} \\ &= \frac{\pi_0 f_0(x)}{\sum_{i=0}^{1} P(X = x|H_i)P(H_i)} \qquad \text{by law of total probability} \\ &= \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \end{aligned}$$

where f_0 and f_1 are densities of x under H_0 and H_1 respectively. Similarly,

$$P(H_1 \,|\, X = x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

and so the posterior odds

$$\frac{P(H_0 \,|\, X = x)}{P(H_1 \,|\, X = x)} = \frac{\pi_0}{\pi_1} \frac{f_0(x)}{f_1(x)}$$

The prior odd ratio π_0/π_1 is updated to the posterior odd ratio by the *Bayes factor* $f_0(x)/f_1(x)$, which is essentially the likelihood ratio. So Bayes factor has the same role as the likelihood function in Bayes statistics.

As for the general hypothesis $H_0 : \theta \in \Theta_0,\ H_1 : \theta \in \Theta_1$ (Θ_i may be singleton or composite), a probability distribution is required on Θ_0 and Θ_1. Suppose g_0, g_1 are prior density functions of $\theta \,|\, H_0$ and $\theta \,|\, H_1$ respectively. In this case, the prior probability for H_i is

$$\pi_i := \frac{\int_{\Theta_i} g_i(\theta)\, d\theta}{\int_{\Theta_0} g_0(\theta)\, d\theta + \int_{\Theta_1} g_1(\theta)\, d\theta}$$

and g_0, g_1 must be chosen so that $\pi_0 + \pi_1 = 1$. Note if $\Theta_i = \{\theta_i\}$ (single-element set), then $\int_{\Theta_i} g_i(\theta)\, d\theta = g_i(\theta_i)$. The Bayes factor can be defined in a similar fashion

Definition 30 (Bayes Factor)**.** Suppose $X \sim f_\theta$. The Bayes factor for Hypothesis testing $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$ is defined as

$$\begin{aligned} B &:= \frac{E_{\theta \sim g_0}[f_\theta(x)]}{E_{\theta \sim g_1}[f_\theta(x)]} \\ &= \frac{\int_{\Theta_0} f_\theta(x) g_0(\theta)\, d\theta}{\int_{\Theta_1} f_\theta(x) g_1(\theta)\, d\theta} \end{aligned}$$

which is the ratio of expected likelihood in Θ_0 and Θ_1.

More generally, Bayes factors for testing two probabilistic models M_0 and M_1 (e.g. one assumes X satisfies normal distribution, and the other assumes t-distribution), with parameters θ_0, θ_1 and corresponding prior densities π_0, π_1, is defined as follows

Definition 31 (Bayes Factor for two different models)**.**

$$\begin{aligned} B &:= \frac{P(x \,|\, M_0)}{p(x \,|\, M_1)} \\ &= \frac{\int f(x \,|\, \theta_0, M_0)\pi_0(\theta_0)\, d\theta_0}{\int f(x \,|\, \theta_1, M_1)\pi_1(\theta_1)\, d\theta_1} \end{aligned}$$

Recall the p-value that is used to measure the evidence to retain H_0, this method is called *Null Hypothesis Significance Testing*(NHST). Bayes factor and p-value play the same rule in hypothesis testing, but there are many differences

- They are on different scales. p-value > 0.05 usually means no evidence (to reject H_0), where the choice of 0.05 is arbitrary. On the other hand, the Bayes factor around 1 corresponds to no evidence.
- There is no "rejection region" nor "rejection boundary" for testing using the Bayes factor. Conclusions made from Bayes factors are in the form: "there is weak/strong evidence supporting H_0" or "there is weak/strong evidence supporting H_1".
- p-value is calculated merely based on H_0 with no relation with H_1. But Bayes factor uses both hypotheses
- Bayes factor depends on your choice of the pair of prior distributions π_0, π_1, so it could be different for different choices, but p-value is a unique value.
- NHST conclusion cannot be changed, but the Bayes factor can be updated further using new data points.

As we have mentioned before, the choice of prior is extremely important. If the choice of prior is careless (e.g. use a non-informative prior), the conclusions between NHST and Bayes factor may conflict as the sample size n increases, called the *Jeffrey-Lindley paradox*[9]. Some priors resolve this paradox, e.g. the Jeffreys-Zellner-Siow priors [3]. Bayesian hypothesis testing is still a deep topic under research.

The rest of this section delves into the application of Bayesian decision theory on hypothesis testing.

10.4.1 Decision Theory and Hypothesis Testing

Throughout this section, we will apply Decision theory to the hypothesis test ϕ for simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ with rejection region C, and find a "good" test".

Define a loss function

$$L_{a,b}(\theta, \phi(x)) := \begin{cases} a\phi(x) & \theta = \theta_0 \\ b(1 - \phi(x)) & \theta = \theta_1 \end{cases}$$

note if $\theta = \theta_0$ and $\phi(x) = 1$, this indicates a type I error. Conversely, for $\theta = \theta_1$, $1 - \phi(x) = 1$ indicates a type II error. The scalars a, b specify the strength of punishment on the Type I and Type II errors.

By definition, the risk function is

$$R(\theta, \phi) = \begin{cases} a\alpha_C & \theta = \theta_0 \\ b\beta_C & \theta = \theta_1 \end{cases}$$

where $\alpha_C := w(\theta_0)$ and $\beta_C := 1 - w(\theta_1)$ are probabilities of making Type I and Type II errors respectively when the critical region of ϕ is C.

Assigning probabilities $\pi(\theta_0) =: \pi_0$, $\pi(\theta_1) =: \pi_1$, the Bayes risk can be calculated directly as being

$$r_\pi(\phi) = \pi_0 a\alpha_C + \pi_1 b\beta_C$$

The *Bayes test* is the best test in the sense of minimising this Bayes risk.

Theorem 10.6. *Suppose $X \sim f_\theta$ and f_θ is (Lebesgue)-integrable. The Bayes test for simple hypotheses has the critical region*

$$C = \left\{ x : \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \geq A \right\}$$

where $A := \pi_0 a/(\pi_1 b)$.
So Bayes test under loss $L_{a,b}$ and prior π is essentially a likelihood ratio test with threshold A. And any likelihood ratio test with rejection region $C = \{\Lambda(x) = f_{\theta_1}(x)/f_{\theta_0}(x) \geq k\}$ is a Bayes test for some prior probabilities π_0, π_1.

Proof. Using the definitions $\alpha_C = P(X \in C \,|\, H_0)$ and $\beta_C = P(X \in \mathscr{X} \setminus C \,|\, H_1)$ where $\mathscr{X} \setminus C$ is the acceptance region (complement of rejection region C), the Bayes risk can be calculated as follows

$$\begin{aligned} r_\pi(\phi) &= \pi_0 a P(X \in C \,|\, H_0) + \pi_1 b P(X \in C' \,|\, H_1) \\ &= \int_C \pi_0 a f_{\theta_0}(x)\, dx + \pi_1 b \int_{\mathscr{X}\setminus C} f_{\theta_1}(x)\, dx \\ &= \int_C \pi_0 a f_{\theta_0}(x)\, dx + \pi_1 b \left[1 - \int_C f_{\theta_1}(x)\, dx\right] \quad \text{(Lebesgue) integral is additive (for union of disjoint sets)} \\ &= \pi_1 b + \int_C \pi_0 a f_{\theta_0}(x) - \pi_1 b f_{\theta_1}(x)\, dx \end{aligned}$$

the set C that minimises the last line is the set only including negative parts of the integrand, i.e.

$$C := \{x \,:\, \pi_0 a f_{\theta_0}(x) - \pi_1 b f_{\theta_1}(x) \le 0\} = \{x \,:\, \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \ge \frac{\pi_0 a}{\pi_1 b}\}$$

For any $k > 0$, it is trivial that one can find prior probabilities $\pi_0 \in [0, 1]$ s.t.

$$k = \frac{\pi_0 a}{\pi_1 b} = \frac{\pi_0 a}{(1-\pi_0) b}$$

so any likelihood ratio test is a Bayes test for some prior π. □

Note when $a = b = 1$, the loss $L_{a,b} = L_{1,1}$ is the trivial 0-1 loss $L(\theta, \phi(x)) = 1_{\theta \ne \phi(x)}$ we discussed in chapter 8 Bayesian Decision Theory. It is 1 if we made a wrong decision, and it is 0 if the decision is correct. So the risk (assuming ϕ has rejection region C) is

$$\begin{aligned} R(\theta, \phi_C) &= E(1_{\theta=\theta_0, \phi(x)=1} + 1_{\theta=\theta_1, \phi(x)=0}) \\ &= P(\theta = \theta_0, x \in C) + P(\theta = \theta_1, x \in \mathscr{X} \setminus C) \\ &= \alpha_C + \beta_C \end{aligned}$$

i.e. the sum probabilities of making Type I and Type II errors.
By theorem 10.6, given prior π, the Bayes test (for simple hypotheses H_0, H_1) for 0-1 loss has a rejection region

$$\begin{aligned} C &= \left\{\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \ge \frac{\pi_0}{\pi_1}\right\} \\ &= \left\{B \frac{\pi_0}{\pi_1} \le 1\right\} \quad B \text{ is the Bayes factor} \\ &= \left\{\frac{p(H_0 \,|\, X = x)}{p(H_1 \,|\, X = x)} \le 1\right\} \quad \text{by Bayes rule} \\ &= \{p(H_0 \,|\, X = x) \le p(H_1 \,|\, X = x)\} \end{aligned}$$

which means reject H_0 if its posterior probability is lower than that of H_1. So for 0-1 loss, the Bayes test (for simple hypotheses) is the MAP(maximum a posteriori) test.

What if any of the two hypotheses is composite (not simple)? Define the marginal likelihood of x under hypothesis H_i as

$$m_i(x) := \begin{cases} f_{\theta_i}(x) & H_i \text{ is simple}, \theta \in \{\theta_i\} \\ \int_{\Theta_i} f_\theta(x) g_i(\theta) & H_i \text{ is composite}, \theta \in \Theta_i \end{cases}$$

where $g_i(\theta)$ is the prior density of $\theta \,|\, H_i$. Then using a similar way to proving theorem 10.6, the following result is obtained

Proposition 10.7. *The Bayes test for 0-1 test has the critical region*

$$C = \{x \,:\, \frac{m_1(x)}{m_0(x)} > \frac{\pi_0}{\pi_1}\}$$

where

$$\pi_i := \frac{\int_{\Theta_i} g_i(\theta)\, d\theta}{\int_{\Theta_0} g_0(\theta)\, d\theta + \int_{\Theta_1} g_1(\theta)\, d\theta}$$

10.5 Two-sided Hypothesis test

These are the tests of the forms

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

or

$$H_0 : \theta \in [\theta_1, \theta_2] \quad H_1 : \theta > \theta_1 \text{ or } \theta < \theta_1$$

or

$$H_0 : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2 \quad H_1 : \theta \in (\theta_1, \theta_2)$$

Unfortunately, UMP may not always exist for two-sided tests.

Example 12. Use the model $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known once again. But this time the hypothesis testing is $\mathcal{H} := \{H_0, H_1\}$, where

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

we aim to find a UMP at level α.

Pick any $\mu_1 > \mu_0$, consider the simple hypotheses testing $\mathcal{H}_1 = \{H_0', H_1'\}$, where

$$H_0' : \mu = \mu_0 \quad H_1' : \mu = \mu_1$$

from the proof of the Karlin-Rubin theorem, we can see the Karlin-Rubin test ϕ_1 with rejection region $C_1 := \{x : \overline{x} \geq \mu_0 + (\sigma/\sqrt{n})z_{1-\alpha}\}$ is the UMP at level α for $\mathcal{H}_1$. So $w_{\phi_1}(\mu_1) \geq w_\phi(\mu_1)$ for any other test ϕ with $w_\phi(\mu_0) \leq \alpha$.

On the other hand, pick $\mu_2 < \mu_0$, consider the simple hypotheses testing $\mathcal{H}_2 = \{H_0'', H_1''\}$, where

$$H_0'' : \mu = \mu_0 \quad H_1'' : \mu = \mu_2$$

the hypothesis test ϕ_2 with rejection region

$$C_1 := \{x : \overline{x} \leq \mu_0 + (\sigma/\sqrt{n})z_{1-\alpha}\}$$

is the UMP at level α for $\mathcal{H}_2$(proof left as exercise) So $w_{\phi_2}(\mu_2) \geq w_\phi(\mu_2)$ for any other test ϕ with $w_\phi(\mu_0) \leq \alpha$.

Therefore, for any test ϕ with size α for hypothesis testing $\mathcal{H}$, i.e. $w_\phi(\mu_0) \leq \alpha$, the power on $\mu < \mu_0$ is dominated by the test ϕ_2 and the power on $\mu > \mu_0$ is dominated by test ϕ_1. Clearly, ϕ_1 and ϕ_2 are completely different tests. Therefore, there is no single test with maximum power on $\{\mu : \mu \neq \mu_0\}$.

This dilemma is illustrated in figure 13, where two thick black curves are powers of ϕ_1 and ϕ_2 described above. Any other test with size $< \alpha$ must not exceed the power of ϕ_1 on the right side, and not exceed the power of ϕ_2 on the left side.

Recall that in Chapter 6 Likelihood-based estimation, when the estimator with uniformly lowest variance does not exist, the scope is restricted to unbiased estimators and aims to find an unbiased estimator with minimum variance(MVUE). Similarly, for hypothesis testing, we restrict to the "unbiased" tests.

Definition 32 (UMPU)**.** A test with rejection region C is unbiased of size α if

$$P_\theta(X \in C) \leq \alpha \; \forall \theta \in \Theta_0 \quad P_\theta(X \in C) \geq \alpha \; \forall \theta \in \Theta_1$$

A test is uniformly most powerful unbiased(UMPU) at level α if it has maximum power among unbiased tests of size α.

So we are asking for control over Θ_1 as well as Θ_0, as illustrated in Figure 14.
In the case of

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

which is the one in the example above (Figure 15), requesting a test to be unbiased bans the tests ϕ_1 and ϕ_2 which cheat to achieve high power on one side by sacrificing the power on the other side. (But both sides of θ_0 are in Θ_1, where the power should be maximised) If ϕ is unbiased in this case, the power function satisfies

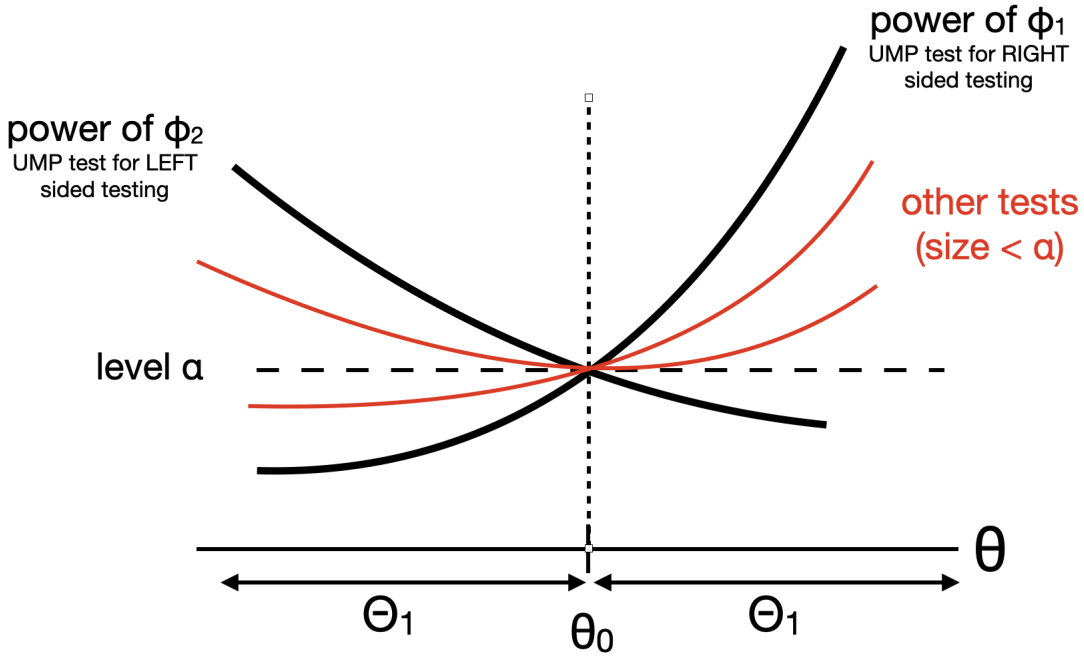

Figure 13: Explanation of why UMP does not exist for the two-sided testing

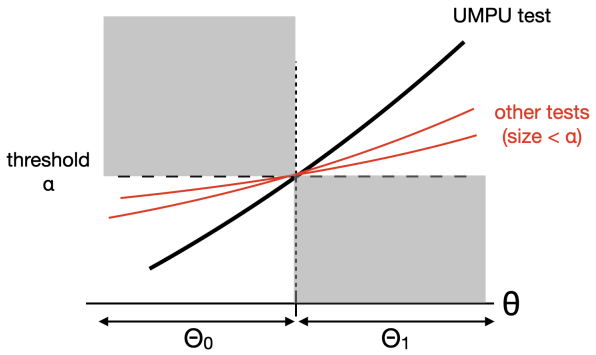

Figure 14: Illustration of UMPU test. Any test with a power function crossing any of the grey areas is banned from competing for the role of UMPU at level α

- $w_\phi(\theta_0) = \alpha$
- $w'_\phi(\theta_0 = 0)$

because the power function is usually continuous.

The following theorem gives a UMPU test in one case of two-sided Hypothesis testing

Theorem 10.8. *Suppose $X \sim f_\theta$ where $\{f_\theta\}_{\theta \in \Theta \subseteq \mathbb{R}}$ is an exponential family with single parameter*

$$f_\theta := h(x) \exp\left(\eta(\theta)t(x) - B(\theta)\right)$$

where η is strict increasing function of θ, then for hypothesis testing

$$H_0 : \theta \in [\theta_1, \theta_2] \quad H_1 : \theta > \theta_1 \text{ or } \theta < \theta_1$$

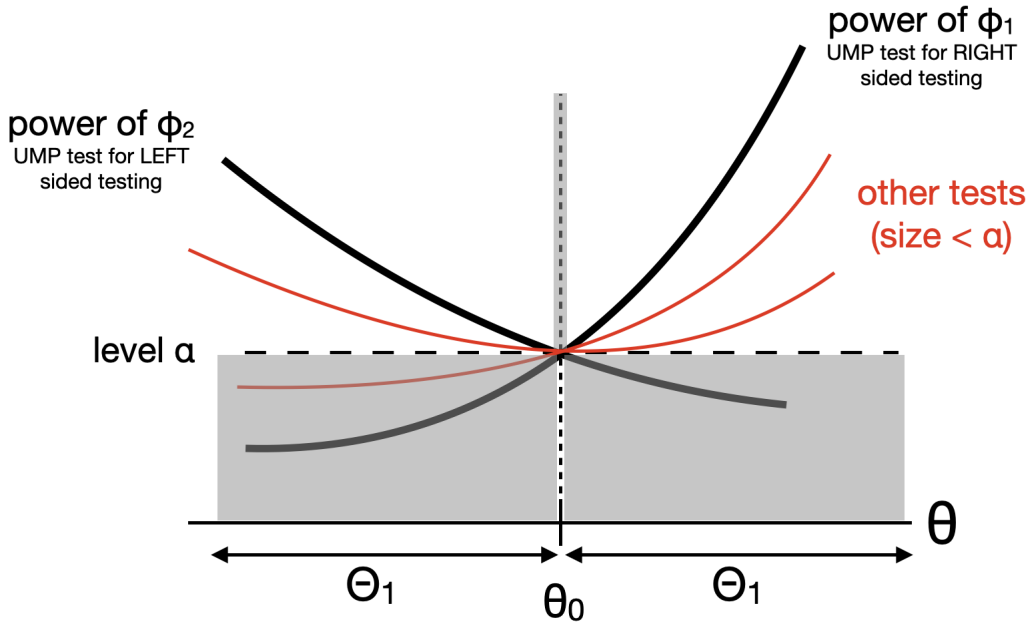

Figure 15: UMPU for the central singleton null hypothesis case

the UMPU test at level α has the following form

$$\phi(x) := \begin{cases} 1, & t(x) < c_1 \text{ or } t(x) > t_2 \\ \gamma, & t(x) = t_1 \text{ or } t(x) = t_2 \\ 0, & t(x) \in (t_1, t_2) \end{cases} \qquad (*)$$

where t_1, t_2, γ are chosen s.t. ϕ is unbiased test of size α.

Remark. The proof will be omitted (full proof to all theorems in this section can be found in [1], also on the website Uni of Washington Lecture Notes), but note that t_1, t_2, γ sometimes have no analytical solutions. Numerical solutions are required in such cases.

As for the case given in the example above(null hypothesis is a single hole and the alternative hypothesis is the rest of Θ), mild additional conditions yield the UMPU test of the same form as in (*)

Theorem 10.9. *Suppose $X \sim f_\theta$ where $\{f_\theta\}_{\theta \in \Theta \subseteq \mathbb{R}}$ is an exponential family with a single parameter defined as in the above theorem. For hypothesis testing*

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

the test in the form (*) is UMPU test at level α if the following conditions are satisfied

- *(i) $w_\phi(\theta_0) = E_{\theta_0}[\phi(X)] = \alpha$*
- *(ii) $E_{\theta_0}[t(X)\phi(X)] = E_{\theta_0}[t(X)]E_{\theta_0}[\phi(X)] = \alpha E_{\theta_0}[t(X)]$. i.e. $t(X)$ is statistically independent of $\phi(X)$. (we are NOT saying that the definition of ϕ does not involve $t(x)$)*

Example 13. Review the example at the beginning of this section again. $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known, the hypothesis testing is

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

we aim to find a UMPU at level α. The joint distribution(multi-variate normal) is an exponential family with statistics $t(X) := \sum_i X_i/n = \overline{X}$ and the only parameter involved is $\theta = \mu$. By Theorem 10.9, the UMPU test have form (*), and we need to pick t_1, t_2 (the distribution is continuous, so a randomised test is not required) s.t. the conditions hold.

- $1 - w_\phi(\mu_0) = P(t_1 \le \overline{X} \le t_2 | \mu = \mu_0) = 1 - \alpha$ by condition (i)
- Also,

$$\begin{aligned} E_{\mu_0}[t(X)(1-\phi(X))] &= E_{\mu_0}[t(X)] - E_{\mu_0}[t(X)\phi(X)] \quad \text{by linearity of expectation} \\ &= (1-\alpha)E_{\mu_0}[t(X)] \quad \text{by condition (ii)} \\ &= (1-\alpha)\mu_0 \quad \text{because } E_{\mu_0}[\overline{X}] = \mu_0 \end{aligned}$$

Now we calculate LHS using the actual distribution and note $1 - \phi(X) = 1$ iff $\overline{X} \in [t_1, t_2]$,

$$\begin{aligned} E_{\mu_0}[t(X)(1-\phi(X))] &= \int_{t_1}^{t_2} t f(t)\, dt \quad \text{where } f \text{ is PDF of } t(X) = \overline{X} \\ &= \int_{t_1}^{t_2} t \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n}{2\sigma^2}(y-\mu_0)^2\right\} dt \end{aligned}$$

It is left to the readers to check that

$$t_1 := n\mu_0 - z_{1-\alpha/2}\sqrt{n}\sigma, \quad t_2 := n\mu_0 + z_{1-\alpha/2}\sqrt{n}\sigma$$

solves the two equations

$$P(t_1 \le \overline{X} \le t_2 | \mu = \mu_0) = 1 - \alpha$$

$$\int_{t_1}^{t_2} t \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n}{2\sigma^2}(y-\mu_0)^2\right\} dt = (1-\alpha)\mu_0$$

So the test with rejection region $\{x : x < t_1 \text{ or } x > t_2\}$ is the UMPU test.
Example ends

Finally, the third form of two-sided hypothesis testing

$$H_0 : \theta \le \theta_1 \text{ or } \theta \ge \theta_2 \quad H_1 : \theta \in (\theta_1, \theta_2)$$

is the reversed version of the first one, so the test is also very similar to (*), but with the rejection region and acceptance region swapped.

Theorem 10.10. *Suppose $X \sim f_\theta$ where $\{f_\theta\}_{\theta \in \Theta \subseteq \mathbb{R}}$ is an exponential family with single parameter. For hypothesis testing*

$$H_0 : \theta \le \theta_1 \text{ or } \theta \ge \theta_2 \quad H_1 : \theta \in (\theta_1, \theta_2)$$

the UMPU test at level α has the following form

$$\phi(x) := \begin{cases} 1, & t(x) \in (t_1, t_2) \\ \gamma, & t(x) = t_1 \text{ or } t(x) = t_2 \\ 0, & t(x) < c_1 \text{ or } t(x) > t_2 \end{cases}$$

where t_1, t_2, γ are chosen s.t. ϕ is unbiased test of size α.

11 Ending

The auxiliary notes on Statistical Inference finish here. Thank you very much for reading.
Please email daniel.kansaki@outlook.com or yuhang.lin@new.ox.ac.uk if you find any typo or have suggestions for improvements.

References

[1] George Casella and Roger L. Berger. *Statistical inference*. Brooks/Cole Cengage Learning, 2001.

[2] George Casella and Jiunn Tzon Hwang. Evaluating confidence sets using loss functions - jstor, May 1982.

[3] Merlise Clyde, Mine Çetinkaya Rundel, Colin Rundel, David Banks, Christine Chai, and Lizzy Huang. An introduction to bayesian thinking, Jun 2022.

[4] I. Good and G. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43:45–63, 1956.

[5] Jason Grossman. The likelihood principle. In Prasanta S. Bandyopadhyay and Malcolm R. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 553–580. North-Holland, Amsterdam, 2011.

[6] Peter D. Hoff. *A first course in Bayesian Statistical Methods*. Springer, 2009.

[7] W. James and C. Stein. Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. Prob., Vol. I*, pages 361–379, Berkeley, 1961. University of California Press.

[8] David C. Lay, Steven R. Lay, and Judith McDonald. *Linear algebra and its applications*. Pearson Education Limited, 2022.

[9] D. V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[10] Carl N. Morris. Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

[11] C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Springer Series in Statistics*, page 235–247, 1992.

[12] Herbert Robbins. An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I*, pages 157–163. University of California Press, Berkeley and Los Angeles, 1956.

[13] Christian P. Robert, Nicolas Chopin, and Judith Rousseau. Harold jeffreys's theory of probability revisited. *Statistical Science*, 24(2), 2009.

[14] Peter Yi-Shi Shao and William E. Strawderman. Improving on the james-stein positive-part estimator. *The Annals of Statistics*, 22(3):1517–1538, 1994.

[15] Robert L. Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, Mar 1972.

[16] G. A. Young and R. L. Smith. *Decision theory*, page 4–21. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2005.