# Contents

1. Prelude 2

1.1. Frequency-Bayes Unification 3

1.2. The Likelihood Function 4

1.3. Notations 5

2. Sufficiency and Ancillarity 5

2.1. Sufficient statistics 5

2.2. Minimal Sufficient Statistics 7

2.3. Completeness and minimal sufficiency 9

2.4. Ancillarity 10

3. Unbiased point estimation 11

3.1. Unbiased point estimator 11

3.2. The Cramer-Rao Lower Bound 14

4. Method of Moments and Maximum Likelihood Estimation 18

4.1. Method of Moments Estimation 18

4.2. Maximum Likelihood Estimation 20

5. General theory for $M$-estimation 26

5.1. General framework 26

5.2. Consistency 28

5.3. Asymptotic normality 29

6. Delta method 33

6.1. Delta method 33

6.2. Variance stabilizing 34

6.3. Asymptotic normality of moment estimators 34

7. Hypothesis Testing 35

7.1. Extending MP tests for simple hypotheses to UMP tests for composite hypothesis 37

7.2. Likelihood Ratio Test 40

7.3. Asymptotic tests 41

8. Interval Estimation 45

8.1. Method-I for constructing interval estimators: inverting tests of hypotheses 46

8.2. Method-II for constructing interval estimators: using pivotal quantities 48

---

These notes benefit from scribe notes taught by Professor Tirthankar Dasgupta at Harvard.
1. Prelude

Stat 593 studies the following question: how can and should we carry out statistical inference, efficiently using observed data to learn about the unknown and make decisions in the face of uncertainty? In particular, suppose that we obtain data $y$ from a statistical model with an unknown parameter $\theta$ (imagine that $\theta$ indexes a family of possible distributions for $y$, and that the true value of $\theta$ determines the true data-generating process). We will then consider questions such as:

(1) (Point Estimation) What is a “good” estimator for $\theta$? This depends of course on the definition of “good”; we will introduce several criteria by which estimators can be judged. Just providing a point estimator $\hat{\theta}$, without any sense of its uncertainty, is usually unsatisfying; so statistical inference emphasizes accompanying point estimators with information about their uncertainties, e.g., we may be able to describe the distribution of $\hat{\theta}$ or at least say what its standard deviation is, approximately (the standard deviation of an estimator is often called its standard error).

Much of the beauty and power of statistics stems from the almost-magical fact that it is often possible not only to estimate unknown quantities, but to make valid, useful statements about how far off the estimate is likely to be from the unknown truth.

(2) (Interval Estimation) Intuitively, much more informative than just saying something like “I estimate that $\theta$ is 42” is to provide an interval, such as saying something like “I am 95% confident that $\theta$
is between 39 and 45 (inclusive)." But what does “confident” mean? If \( \theta \) is a constant, then it either is or isn’t in the interval \([39, 45]\), so what does the 95% mean? In this course we will define precisely what it means to give an interval estimate, and study ways of constructing such estimates.

3. \((Hypothesis \ Testing)\) In many applications in the physical, biological, and social sciences, a researcher is interested in testing a hypothesis which can be expressed in the form \( \theta = \theta_0 \) or, more generally, as \( \theta \in H_0 \) for some set \( H_0 \). Hypothesis testing is closely related to interval estimation (and arguably the latter is more useful), and can be approached via various perspectives.

4. \((Prediction)\) George Box famously said that “essentially, all models are wrong, but some are useful.” In some problems the parameter \( \theta \) has a direct physical interpretation, but in other problems it is just a convenient fiction which helps us obtain a distribution that fits the data well. Whether or not \( \theta \) itself is of interest, there may be tremendous interest in predicting (also known as forecasting) future data. A useful idea for doing this is first to use the observed data to learn about \( \theta \), and then to use knowledge about \( \theta \) to predict future observations.

5. \((Decision-making)\) We may want to learn about the unknown for the sake of improving our knowledge about the world (and since it’s fun to learn new things), but we may also need to make decisions based on our inferences. Decision theory, which is closely related to game theory, provides a framework for this, and also sheds light on the earlier questions since, for example, how “good” a point estimate \( \hat{\theta} \) is can be studied in terms of what the cost is of providing \( \hat{\theta} \) when the true value is \( \theta \).

**Note on terminology:** A statistic is a function of the data (it must be computable using only observed data). An estimator is a statistic designed to estimate a particular parameter; after the data are observed and a specific numerical value can be obtained, it is called an estimate (we sometimes do not distinguish between estimators and estimates since we often have to go back and forth between thinking of the data as random and thinking of the data as having “crystallized” into specific values, and it is usually clear from the context which is meant). The unknown quantity we wish to estimate is called the estimand.

\*1.1. It is crucial to distinguish clearly between parameters and data, and correspondingly to distinguish clearly between estimands and estimates.

1.1. **Frequency-Bayes Unification.** Statistical inference has produced many paradoxes and controversies in the last hundred years, most notably a sometimes-acrimonious divide between so-called Bayesians and so-called frequentists. In this course, we aim for a Frequency-Bayes unification, arguing that for
the most part, it is desirable to use procedures that are good from both perspectives.

*Conditioning is the soul of statistics.* This manifests itself in many of the debates over inference: what should be held constant, and what should be allowed to roam randomly over its possible values? In the Bayesian point of view, unknowns are treated as random variables, and inference is based on the *posterior density* $g(\theta | y)$. This is intuitively very natural since it says to condition on what we know (the data), and use probability to quantify our uncertainty about the quantity we don’t know (the parameter).

However, some statisticians reject the idea of allowing $\theta$ to have a distribution, since (in their view) it is an unknown constant but is not a random variable; they then consider possible values of $y$ which did *not* actually occur, and work with $f(y | \theta)$ rather than $g(\theta | y)$. Indeed, they may even reject the notation $f(y | \theta)$, writing instead $f_0(y)$ or $f(y; \theta)$. This is sometimes called a “frequentist” point of view, but frequentism is the idea of evaluating estimators in terms of their long-run performance in repeated sampling; thus, a procedure derived from a Bayesian point of view may in fact be excellent by frequentist standards too (and often, but now always, it is true that procedures obtained by Bayesian methods have good frequentist properties).

While discussing frequentist ideas and methods, we will use the notation $f_\theta(y)$ rather than $f(y | \theta)$. The notation $f(y | \theta)$ will be introduced again while introducing Bayesian methods, where conditioning on the parameter makes more sense.

Throughout the course, we emphasize *likelihood* as a bridge between the two schools of thought: both Bayesians and frequentists recognize the importance of the likelihood function (defined in the next section). A second bridge is provided by *decision analysis*, our study of which will reveal that in great generality, even a diehard frequentist *must* act like a Bayesian in order to be “admissible.”

### 1.2. The Likelihood Function.

In a probability problem, we typically have a random variable $Y$ whose distribution has been specified in some way, and wish to make statements about various probabilities and averages. In an inferential problem, we face an inverse problem: given that $Y$ has been observed to be $y$, we wish to make statements about the unknown parameters associated with $Y$.

Throughout, suppose that we have a model consisting of a collection of distribution functions indexed by an unknown parameter $\theta$ taking values in some space $\Theta$. We assume that $Y$ is sampled according to one of these distributions, resulting in observed data $y$.

Let $f(y | \theta)$ be the probability mass function if $Y$ is discrete, or the probability density function if $Y$ is continuous (more generally, $\{f(y | \theta) : \theta \in \Theta\}$ can be a family of densities with respect to some dominating measure, which is normally taken to be counting measure in the discrete case and Lebesgue measure in the continuous case). The model may be parametric (which we
will take to mean that \( \Theta \) is finite-dimensional) or nonparametric (e.g., the space of all CDFs, or all unimodal continuous PDFs).

The likelihood function, defined below, gives the probability of the observed data as a function of the unknowns. This gives a graphical way to see what the data say about \( \theta \), and to compare different values of \( \theta \).

**Definition 1.1.** The *likelihood function* of the model \( f(y|\theta) \) with observed data \( y \) is the function
\[
L(\theta) = f(y|\theta),
\]
viewed as a function of \( \theta \) with the observed data \( y \) fixed. Since \( L(\theta) \) involves a product, it is often more convenient to work with the logarithm of the likelihood function, so we define the *log-likelihood* function by
\[
l(\theta) = \log L(\theta).
\]

1.3. **Notations.** We will use the following generic notations, unless specified otherwise.

1. Bold mathematical symbol, e.g., \( Y \) to define a vector.
2. \( Y = (Y_1, \ldots, Y_n) \) to denote the vector of \( n \) observations \( Y_1, \ldots, Y_n \).
3. \( \theta \) to denote a parameter and \( \Theta \) to denote a vector of parameters.
4. While discussing frequentist methods, \( f_\theta(y) \) to denote a probability density/mass function and \( F_\theta(y) \) to denote a cumulative distribution function (CDF) of a random variable \( Y \), where the probability distribution is indexed by parameter \( \theta \).
5. \( P_\theta [Y \in A] \) to denote the probability that the observed data belongs to a (measurable) set \( A \), where the underlying probability distribution is indexed by parameter \( \theta \).

2. **Sufficiency and Ancillarity**

2.1. **Sufficient statistics.** A *sufficient statistic* is a statistic that contains all of the information available from the data which is pertinent for inference about \( \theta \). This can serve the purpose of data reduction, which is one major purpose of statistics (indeed, Fisher wrote that “the object of statistical methods is the reduction of data”). Also, we will see that conditioning on a sufficient statistic is useful in creating estimators.

**Definition 2.1.** Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} f_\theta(y) \). A statistic \( T(Y) \) is *sufficient* for \( \theta \) (or, more generally, for the family of distributions indexed by \( \theta \)) if the conditional distribution of \( Y | T \) does not depend on \( \theta \). That is, if we condition on a sufficient statistic \( T \), the distribution of the data no longer involves \( \theta \); the data depend on \( \theta \) only via \( T \). Mathematically, the conditional probability \( P_\theta [Y \in A|T = T(Y)] \) is free of \( \theta \) for any measurable set \( A \).

Sufficiency can also be defined via Bayesian or information theoretic perspectives.
Theorem 2.1. The following are equivalent.

(a) The statistic \( T \) is sufficient for \( \theta \), i.e., the conditional distribution of \( Y \) given \( T \) does not depend on \( \theta \).

(b) (Bayesian) The conditional distribution of \( \theta \) given \( T \) is the same as the posterior distribution of \( \theta \) given \( Y \), for any prior distribution on \( \theta \).

(c) (information theory) The chain \( \theta \rightarrow T \rightarrow Y \) is Markovian for any distribution on \( \theta \), i.e., \( \theta \) and \( Y \) are conditionally independent given \( T \).

\( 2.1 \). Prove the equivalence of the above definitions of sufficiency.

Example 2.1. Let \( Y_1, Y_2 \) be i.i.d. \( \text{Pois}(\lambda) \), and let \( T \equiv Y_1 + Y_2 \). Then \( (Y_1, Y_2) \mid T \sim \text{Multinomial}(T, (1/2, 1/2)) \), so \( T \) is sufficient. Likewise, any one-to-one function of \( T \) is sufficient. Check though that, for example, \( T_2 \equiv Y_1 + 2Y_2 \) is not sufficient.

Example 2.2. Let \( Y_1, Y_2, \ldots, Y_n \) be i.i.d. \( \text{Bern}(p) \). Then \( \sum_i Y_i \) is sufficient since the distribution of \( (Y_1, \ldots, Y_n) \) given that \( \sum_i Y_i = k \) is uniform on the \( \binom{n}{k} \) possible sequences.

Rather than using the definition of sufficiency, it is usually easier to use the following criterion. The result says that \( f_{\theta}(y) \) depends on \( \theta \) only through \( T(y) \).

Theorem 2.2 (Factorization Theorem). A statistic \( T(Y) \) is sufficient for \( \theta \) if and only if the joint probability function \( f_{\theta}(y) \) can be expressed in the form:

\[
f_{\theta}(y) = g_{\theta}(T(y)) \cdot h(y)
\]

for some functions \( g_{\theta}(\cdot) \) and \( h(\cdot) \), where \( g(\cdot) \) involves \( \theta \) and \( h(\cdot) \) is free of \( \theta \).

The Factorization Criterion holds in great generality, but the general proof requires more measure theory than we wish to get embroiled in here. For the discrete case, we can argue as follows.

Proof. Only if part: Let \( T \) be sufficient. For any \( y \) and the corresponding \( T(Y) = T(y) \),

\[
f_{\theta}(y) = P_{\theta}[Y = y] = P_{\theta} \left[ Y = y \cap T(Y) = T(y) \right] = P_{\theta} \left[ Y = y \mid T(Y) = T(y) \right] P_{\theta}[T(Y) = T(y)].
\]

By definition of sufficiency, the first term is free of \( \theta \) and the second term depends on \( \theta \) and on \( y \) through \( T(y) \). Thus, the joint pmf \( f_{\theta}(y) \) is of the desired form with \( h(y) = P_{\theta}[Y = y \mid T(Y) = T(y)] \) and \( g_{\theta}(T(y)) = P_{\theta}[T(Y) = T(y)] \).

If part: We now assume the condition

\[
f_{\theta}(y) = P_{\theta}[Y = y] = g_{\theta}(T(y)) \cdot h(y).
\]
Then, using the general fact from the proof of the only if part that \( P_\theta [Y = y] = P_\theta [Y = y \cap T(Y) = T(y)] \), it follows that the conditional probability

\[
P_\theta [Y = y | T(Y) = T(y)] = \frac{P_\theta [Y = y]}{P_\theta [T(Y) = T(y)]} = \frac{g_\theta (T(y)) h(y)}{\sum_{y: T(y) = T(y)} g_\theta (T(y)) h(y)},
\]

which is free of \( \theta \).

We now present a couple of examples to illustrate applications of the factorization theorem:

**Example 2.3.** (Exponential family) Let \( Y_1, \ldots, Y_n \) be iid observations from

\[
f_\theta(y) = \exp \left( \eta(\theta) T(y) - \psi(\eta) \right) h(y).
\]

The joint pdf is given by

\[
f_\theta(y) = \exp \left( \eta(\theta) \sum_{i=1}^n T(y_i) - n\psi(\eta) \right) \prod_{i=1}^n h(y_i).
\]

By the factorization theorem, \( T(Y) = \sum_{i=1}^n T(Y_i) \) is a sufficient statistic for \( \theta \). More generally, for \( Y_1, Y_2, \ldots, Y_n \) i.i.d. from a multi-parameter exponential family

\[
f_\theta(y) = \exp \left( \sum_{j=1}^k \eta_j(\theta) T_j(y) - \psi(\eta) \right) h(y),
\]

the statistic \( \left( \sum_{i=1}^n T_1(Y_i), \ldots, \sum_{i=1}^n T_k(Y_i) \right) \) is jointly sufficient for \( \theta \). Note that the dimension of the sufficient statistic matches the number of parameters.

**Example 2.4.** (Uniform distribution) Let \( Y_1, \ldots, Y_n \) \( \text{iid} \sim Unif[0, \theta] \). Let \( Y_{(n)} = \max(Y_1, \ldots, Y_n) \) be the largest order statistic and \( Y_{(1)} = \min(Y_1, \ldots, Y_n) \) be the smallest order statistic. The joint pdf of \( Y \) is given by

\[
f_\theta(y) = \frac{1}{\theta^n} \mathbb{1}(y_{(n)} \leq \theta) \mathbb{1}(y_{(1)} \geq 0).
\]

By the factorization theorem, \( Y_{(n)} \) is a sufficient statistic for \( \theta \).

**2.2. Minimal Sufficient Statistics.** A sufficient statistic is not unique. First of all, it is easy to see that any one-to-one transformation of a sufficient statistic gives a sufficient statistic. Moreover, note that the entire data always form a sufficient statistic, which belies the intention of keeping a simplified function of the data without losing information. We therefore define *minimality* to capture the idea of keeping just what is necessary for inference about \( \theta \).
Definition 2.2. A sufficient statistic \( T \) is **minimal** if \( T \) is a function of any other sufficient statistic. That is, if \( S \) is sufficient, then there is a function \( g \) such that \( T = g(S) \) a.s. for all \( \theta \).

Some examples follow.

1. For \( Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \) with \( \mu \) and \( \sigma^2 \) unknown, \((\bar{Y}, \sum_{i=1}^n Y_i^2)\) is minimal sufficient. Equivalently, \((\bar{Y}, \sum_{i=1}^n (Y_i - \bar{Y})^2)\) is minimal sufficient.

2. For \( Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Unif}(\theta - 1/2, \theta + 1/2), \) the 2-dimensional statistic \((\min Y_i, \max Y_i)\) is minimal sufficient. This family is not an EF since the support depends on \( \theta \).

A minimal sufficient statistic always exists (except in certain pathological measure-theoretic cases). By definition, the minimal sufficient statistic will be unique up to one-to-one transformations. The following Theorem from Casella and Berger (2002) provides us with a guideline to establish minimal sufficiency.

Theorem 2.3. Let \( Y \) constitute a vector of iid observations from a probability distribution with density \( f_\theta(y) \) and let \( T(Y) \) be a sufficient statistic for \( \theta \). If for two distinct data points \( x \neq y \), the ratio \( f_\theta(x)/f_\theta(y) \) is free of \( \theta \) only if \( T(x) = T(y) \), then \( T \) is minimal sufficient.

Proof. Let \( T' \) be any other sufficient statistic. We will show that \( T \) is a function of \( T' \), for which it suffices to show that if any two points \( x \) and \( y \) \((x \neq y)\) satisfy \( T'(x) = T'(y) \), then they also satisfy \( T(x) = T(y) \).

For two such points \( x \) and \( y \) satisfying \( T'(x) = T'(y) \), we have that

\[
\frac{f_\theta(x)}{f_\theta(y)} = \frac{g_\theta(T'(x))h(x)}{g_\theta(T'(y))h(y)} = \frac{h(x)}{h(y)},
\]

which is free of \( \theta \). Hence by the condition of the Theorem, we must have \( T(x) = T(y) \). \( \square \)

Example 2.5. Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \). For \( x \neq y \), the ratio

\[
\frac{f_\theta(x)}{f_\theta(y)} = \frac{\exp\left\{-\frac{1}{2\sigma^2} \left( \sum_{i=1}^n (x_i - \mu)^2 / (2\sigma^2) \right) \right\}}{\exp\left\{-\frac{1}{2\sigma^2} \left( \sum_{i=1}^n (y_i - \mu)^2 / (2\sigma^2) \right) \right\}} = \exp\left\{-\frac{1}{2\sigma^2} \left( \sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 \right) + \frac{\mu^2}{\sigma^2} \left( \sum_{i=1}^n x_i - \sum_{i=1}^n y_i \right) \right\},
\]

which is free of \( \mu \) and \( \sigma^2 \) (i.e., same for all values of \( \mu \) and \( \sigma^2 \)) if and only if

\[
\sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \quad \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2.
\]

Thus, \((\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)\) is a minimal sufficient statistic for \((\mu, \sigma^2)\).
2.3. Completeness and minimal sufficiency. We now introduce the notion of completeness and a complete sufficient statistic, which provides us with an alternative route to establishing minimal sufficiency.

**Definition 2.3.** A statistic $T(Y)$ is called complete for a family of distributions indexed by parameter $\theta$ if it is impossible to construct a non-trivial unbiased estimator of zero from the statistic. That is, the only way to have $E_\theta[h(T(Y))] = 0$ for all $\theta$ is $h(T)$ to be 0 almost surely, i.e., $P_\theta[h(T)] = 0$.

**Example 2.6.** (Casella and Berger, 2002) Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Ber(p)$. We shall show that the statistic $T = T(Y) = \sum_{i=1}^n Y_i$ is a complete sufficient statistic for $p$. Let $h(T)$ be a function of $T$ that satisfies $E_p[h(T)] = 0$. Since $T \sim Binom(n, p)$, this means

$$\sum_{t=0}^n h(t) \binom{n}{t} p^t(1-p)^{n-t} = 0$$

$$\Rightarrow (1-p)^n \sum_{t=0}^n h(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = 0 \quad \forall p \in (0,1).$$

The term $(1-p)^n$ is not zero for any $p \in (0,1)$. Thus we must have $\sum_{t=0}^n h(t) \binom{n}{t} r^t = 0$ where $r = p/(1-p) \in (0, \infty)$. Now $\sum_{t=0}^n h(t) \binom{n}{t} r^t$ is a polynomial in $r$ of degree $t$. To be zero for all $r$, each coefficient must be zero. Thus $P_p[h(T) = 0] = 1$ for all $p \in (0,1)$, implying that $T$ is complete.

A complete sufficient statistic will automatically be minimal (except in pathological cases), and completeness greatly simplifies the ancillarity issues as discussed below.

**Proposition 2.1.** Any complete sufficient statistic is also minimal, as long as at least one minimal sufficient statistic exists for the model.

**Proof.** Let $T$ be complete and sufficient, and let $M$ be minimal sufficient for parameter $\theta$. Consider the quantity $h(T) = E_\theta(T|M) - T$. Then the following claims can be made:

(a) $h(T)$ is a statistic, i.e., a function of $Y$ only that is free of $\theta$.

(b) $h(T)$ is a function of $T$.

(c) $h(T)$ is an unbiased estimator of zero.

Claim (a) is true because the first term $E_\theta(T|M)$ depends on the conditional distribution of $T|M$, which is free of $\theta$ because it is conditioned on the sufficient statistic $M$. Also, $E_\theta(T|M)$ is a function of $M$, which is a function of $Y$. The second term $T$ is a statistic by definition. To see why claim (b) is true, note that the first term $E_\theta(T|M)$ is a function of $M$, which is minimal sufficient, and hence a function of every other sufficient statistic, including
Thus, both terms are functions of $T$. Finally, part (c) holds because $E[E_\theta(T|M)] = E(T) - E(T) = 0$.

By (c), $E_\theta (h(T)) = 0$. By Definition 2.3, this means $h(T) = 0$ almost surely. Hence $E(T|M) = T$ almost surely, which implies that $T$ is a function of $M$ and hence $T$ is minimal sufficient.

Completeness is often difficult to prove, but it is automatic that the natural sufficient statistic in an exponential family is complete, under mild conditions. However, a complete sufficient statistic may not exist. Consider the following example: let $Y_1, \ldots, Y_n \ iid \sim \mathcal{N}(\theta, \theta^2)$ (curved normal family). It can be shown that the statistic $T = (\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$ is minimal sufficient but not complete sufficient.

2.4. Ancillarity. A statistic $A(Y)$ is ancillary if its distribution does not depend on $\theta$. Why then isn’t “ancillary” a polite word for “useless”? We have been assuming that the model $f_\theta(y)$ is correct, but if we wish to test the model, or compare different models, then ancillary information may be very valuable. Indeed, the fact that an ancillary statistic is (typically) independent of the minimal sufficient statistic is very convenient in testing the model.

**Example 2.7.** Some examples of ancillary statistic are given below.

(i) A trivial example of ancillary statistic is a constant.

(ii) Let $Y_1, \ldots, Y_n \ iid \sim \mathcal{N}(\mu, 1)$. The statistic $A(Y) = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$ is ancillary for $\mu$.

(iii) Let $Y_1, \ldots, Y_n \ iid \sim \text{expo}(\lambda)$ with pdf $f_\lambda(y) = \lambda \exp(-\lambda y), \lambda > 0, y \geq 0$. The statistic $A(Y) = Y_n/(Y_1 + \ldots + Y_n)$ is ancillary for $\lambda$. To see why this is true, consider the transformation $U_i = \lambda Y_i$ for $i = 1, \ldots, n$. Then $A(Y) = U_n/(U_1 + \ldots + U_n)$, where $U_i \ iid \sim \text{expo}(1)$. This example can be generalized to the case where $Y_1, \ldots, Y_n$ are iid from a scale family of distributions, i.e., with cdf of the form $F(y/\sigma), \sigma > 0$. Then, any statistic that depends on $Y$ only through the $n-1$ values $Y_2/Y_1, \ldots, Y_n/Y_1$ (e.g., $Y_1/\sum_{i=1}^n Y_i$) is an ancillary statistic. Refer to Section 6.2 of Casella and Berger (2002) for further details.

Basu’s Theorem is an extremely elegant and useful result relating completeness to ancillarity. The proof beautifully combines each ingredient.

**Theorem 2.4 (Basu).** If $T$ is a complete sufficient statistic for $\theta$ and $A$ is ancillary, then $A$ and $T$ are independent.

**Proof.** For any measurable set $B$, $h(T) \equiv P_\theta(A \in B|T) - P_\theta(A \in B)$ is a function of $T$ not depending on $\theta$ (by sufficiency for the first term, by ancillarity for the second). We have $E(h(T)) = 0$, so completeness gives $h(T) = 0$ almost surely. \qed
Example 2.8. Let \( Y_1, \ldots, Y_n \) be i.i.d. \( \mathcal{N}(0, \sigma^2) \). We will use Basu’s Theorem to give an easy proof that \( \bar{Y} \) and \( S = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \) are independent. To apply Basu’s Theorem, we first need to specify an inference problem for which \( \bar{Y} \) is complete and sufficient. To do this, we introduce a parameter \( \mu \) for the mean: let \( Y_1, \ldots, Y_n \) be i.i.d. \( \mathcal{N}(\mu, \sigma^2) \) with \( \mu \) unknown and \( \sigma^2 \) known. (Even though the mean is 0 in the original problem, we require an unknown parameter for which \( \bar{Y} \) is a complete sufficient statistic.) Then \( \bar{Y} \) is a complete sufficient statistic (this can be shown directly or using NEFs and the fact that the MGF (if it exists) uniquely determines the distribution). The statistic \( S \) is ancillary (indeed, so is the vector of residuals \( (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}) \)) because it is translation-invariant (add and subtract \( \mu \) in each term). By Basu’s Theorem, \( \bar{Y} \) and \( S \) are independent for all \( \mu \) and \( \sigma^2 \). In particular, they are independent for \( \mathcal{N}(0, \sigma^2) \), regardless of whether \( \sigma^2 \) is known.

Example 2.9. Let us revisit example 2.7(iii), where we argued that \( Y_n/(Y_1 + \ldots + Y_n) \) is ancillary for \( \lambda \). It is easy to see that \( \bar{Y} \) is CSS for \( \lambda \). Hence by Basu’s theorem, \( \bar{Y} \) and \( Y_n/(Y_1 + \ldots + Y_n) \) are independent.

Example 2.10. Let \( Y_1, \ldots, Y_n \) be i.i.d. \( \mathcal{N}(\mu, \sigma^2) \). Two possible estimators for \( \mu \) are the sample mean \( \bar{Y} \) and the sample median \( \text{med}(Y) \). We may wish to know how correlated these two estimators are. Using Basu’s Theorem, it is easy to find their covariance:

\[
\text{Cov}(\bar{Y}, \text{med}(Y)) = \text{Cov}(\bar{Y}, \text{med}(Y) - \bar{Y} + \bar{Y}) = \text{Cov}(\bar{Y}, \text{med}(Y) - \bar{Y}) + \text{Cov}(\bar{Y}, \bar{Y}) = \sigma^2/n,
\]

since \( \text{med}(Y) - \bar{Y} \) is ancillary (for \( \mu \), treating \( \sigma \) as known). Note that the same answer, \( \sigma^2/n \), holds if the median is replaced by any statistic \( T \) such that \( T - \bar{Y} \) is ancillary, e.g., any order statistic!

3. Unbiased point estimation

3.1. Unbiased point estimator. We now discuss how a sufficient statistic (possibly a complete or a minimal sufficient statistic) or its function can be used as an unbiased point estimator of a parameter \( \theta \) or a parametric function \( g(\theta) \). For example, in the context of iid observations \( Y_1, \ldots, Y_n \) from a \( \text{Poi}(\lambda) \) distribution, one might be interested in estimating the parameter \( \lambda \) itself, or a parametric functions like the true probability of obtaining a zero value from the distribution (i.e., \( P_{\lambda}(Y = 0) = e^{-\lambda} \)). We first introduce the concept of unbiasedness.

Definition 3.1. A statistic \( T(Y) \) is an unbiased estimator of a parametric function \( g(\theta) \) if its expected value (with respect to the distribution of the statistic over hypothetical replications of the data \( Y \)) equals \( g(\theta) \), i.e., \( E_{\theta}[T(Y)] = g(\theta) \).
An unbiased estimator is not necessarily a “good” one, as we will soon see through some examples. For example, in the Poisson example, a single observation $Y_1$ is an unbiased estimator of $\lambda$. We will however, argue, that the concepts of sufficiency and completeness will allow us to start from such “trivial” estimators, improve them, and even arrive at estimators that are “best” with respect to some criterion of goodness for point estimators. The criterion that we will consider in this Section is Mean Squared Error (MSE), defined as the expectation of the squared difference between the parametric function $g(\theta)$ and its estimator $T$, i.e., $E_\theta[(T - g(\theta))^2]$. Note that the MSE can be expressed as

$$E_\theta[(T - E_\theta(T))^2] + [E_\theta(T) - g(\theta)]^2. $$

The first term in the above expression is the variance of $T$ denoted by $\text{Var}_\theta(T)$, and the second is the square of the bias $E_\theta[T - g(\theta)]$. If $T$ is unbiased, the second term vanishes, and the MSE reduces to the variance. A common and well studied problem in Statistical inference is the “bias-variance” tradeoff. An estimator may have a small bias, but may be preferable to an unbiased estimator with a much larger variance. However, unbiasedness is an attractive (albeit nonsensical in certain situations) property on its own, and for the most part of this Section, we will keep our discussion restricted to unbiased estimators.

Based on the MSE criterion, it makes sense to find an estimator $T(Y)$ of $g(\theta)$ that has the smallest variance in the class of all unbiased estimators. The key idea that we will use to find such an estimator with minimum variance consists of the following steps:

(i) Find any (possibly trivial) unbiased estimator.

(ii) Improve this unbiased estimator by conditioning on a sufficient statistic (better unbiased estimator).

(iii) If the sufficient statistic used for conditioning is complete, then we’ve found the “best” estimator.

In this Section we will state and prove two Theorems. Theorem 3.1, due to Rao (1945) and Blackwell (1947), helps accomplish step (ii). Theorem 3.2 can be used to achieve task (iii).

**Theorem 3.1. [Rao-Blackwell Theorem]** Let $W(Y)$ be an unbiased estimator of $g(\theta)$ and $T$ be any sufficient statistic. Consider the estimator $\phi(T) = E_\theta[W|T]$. This estimator satisfies:

(a) The estimator $\phi(T)$ is an unbiased estimator of $g(\theta)$, i.e.,

$$E_\theta[\phi(T)] = g(\theta).$$

(b) The estimator $\phi(T)$ is “uniformly” better than $W$ in the sense

$$\text{Var}_\theta[\phi(T)] \leq \text{Var}_\theta[W] \quad \forall \theta \in \Theta.$$
Proof. First, note that \( \phi(T) \) is an estimator (a statistic that is free of \( \theta \) by sufficiency of \( T \)). We have that,

\[
g(\theta) = E_\theta[W] = E[E[W|T]] = E_\theta[\phi(T)].
\]

This proves part (a). To prove part (b), note that

\[
\text{Var}_\theta[W] = E[\text{Var}(W|T)] + \text{Var}[E(W|T)] = E[\text{Var}(W|T)] + \text{Var}[\phi(T)] \\
\geq \text{Var}_\theta[\phi(T)] \quad \forall \theta.
\]

□

Theorem 3.1 tells us that we can take any unbiased estimator \( W \) of \( g(\theta) \) and improve it by conditioning on any sufficient statistic. The question that naturally arises is, does this sequential improvement depend on the initial choice \( W \)? what would have happened if a different estimator, say \( \tilde{W} \) was chosen and improved upon by conditioning? Can we reach a “best” estimator by starting from either of them? Theorem 3.2 provides an answer to these questions.

**Theorem 3.2.** [Lehmann-Scheffe Theorem] An unbiased estimator of \( g(\theta) \) that is a function of a complete sufficient statistic is the unique uniformly minimum variance unbiased estimator (UMVUE) of \( g(\theta) \).

**Proof.** Let \( W \) and \( \tilde{W} \) be any two unbiased estimators of \( g(\theta) \), \( T \) a complete sufficient statistic (CSS) for \( \theta \), and \( \phi(T) = E[W|T] \) and \( \tilde{\phi}(T) = E[\tilde{W}|T] \) be the estimators obtained by conditioning \( W \) and \( \tilde{W} \) on the CSS \( T \). Then, by the Rao-Blackwell theorem, \( \phi(T) \) and \( \tilde{\phi}(T) \) are both (i) functions of the CSS \( T \), (ii) unbiased estimators of \( g(\theta) \), and (iii)

\[
\text{Var}_\theta[\phi(T)] \leq \text{Var}[W], \quad \text{Var}_\theta[\tilde{\phi}(T)] \leq \text{Var}[\tilde{W}].
\]

Let \( h(T) = \phi(T) - \tilde{\phi}(T) \). Then

\[
E_\theta[h(T)] = E_\theta[\phi(T) - \tilde{\phi}(T)] = g(\theta) - g(\theta) = 0.
\]

By completeness of \( T \), \( h(T) = 0 \) almost surely, implying that \( \phi(T) = \tilde{\phi}(T) \) almost surely. □

**Remark 3.1 (Implications of the Lehmann-Scheffe theorem).** (a) As long as we have not found a CSS, we can keep on improving an unbiased estimator by conditioning on any sufficient statistic. Once a CSS is obtained, and used for conditioning, we cannot improve further because the unique “best” estimator has been obtained by conditioning on the CSS.

(b) If we have found a CSS and a function of the CSS that is unbiased for \( g(\theta) \), that function is the UMVUE of \( g(\theta) \).
Example 3.1. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Poi}(\lambda)$. We already know that $T = \sum_{i=1}^{n} Y_i$ is CSS for $\lambda$. Start with the trivial estimator $Y_1$, which is unbiased for $\lambda$. Then,

This means $E(Y_1|T) = T/n = \bar{Y}$. By the Lehmann-Scheffe theorem, $\bar{Y}$ is the UMVUE of $\lambda$.

Example 3.2 (UMVUE can be a ridiculous estimator!). Consider a single observation $Y \sim \text{Poi}(\lambda)$. Suppose you are interested in the function $g(\lambda) = e^{-2\lambda}$. It is easy (show this) to see that $T(Y) = (-1)^Y$ is an unbiased estimator of $g(\lambda)$. Because $Y$ is a CSS, by the Lehmann-Scheffe theorem, $T(Y)$ is the UMVUE of $e^{-2\lambda}$, which means, irrespective of the observed value of $Y$, the UMVUE is $-1$ or $+1$ depending on whether $Y$ is even or odd!

Remark 3.2. Unbiasedness itself may be a ridiculous concept in many practical situations, see for example, the famous Circus Example by Basu (1971). However, it remains a useful and intuitive concept in statistical inference, and often a starting point in many inference problems.

3.2. The Cramér-Rao Lower Bound. Having discussed unbiased estimators of parametric functions, their construction and properties, we now aim at finding a lower bound of the variance of such unbiased estimators. The famous Cramér-Rao Lower Bound (Rao 1945; Cramer 1946) derives an expression for variance which is the best that an unbiased estimator can achieve, in the sense of minimizing variance (which is equivalent to minimizing MSE when working with unbiased estimators). However, in order to state and prove the theorem, we need to introduce some additional concepts related to the likelihood function defined by (1.1) and briefly discussed in Section 1.2. We thus introduce two important terms – the Score function and Fisher information – in the following two subsections and discuss their properties.

3.2.1. The Score Function. The derivative of the log-likelihood function occurs so commonly that it has its own name: the score function.

Definition 3.2. The score function for a single observation $Y$ is defined as

$$S_1(Y, \theta) = \frac{\partial \log f_{\theta}(Y)}{\partial \theta}. \hspace{1cm} (3.1)$$

Definition 3.3. The score function $S(Y, \theta)$ for the data $Y = (Y_1, \ldots, Y_n)$ is defined as

$$S_n(Y, \theta) = \frac{\partial \log f_{\theta}(Y)}{\partial \theta} = \frac{\partial l(\theta)}{\partial \theta}, \hspace{1cm} (3.2)$$

where $l(\theta)$ is the log-likelihood function.
Henceforth, for notational simplicity, we will drop the index \(n\) and write the score function for data \(Y\) as \(S(Y, \theta)\). The score function is often called the “score statistic”, but that terminology falsely suggests that it is a statistic (i.e., a quantity computable from the data); in fact, the score function depends on both \(Y\) and \(\theta\). Note that, if \(Y_1, \ldots, Y_n\) are iid observations, then from (3.2)

\[
S_n(Y, \theta) = \sum_{i=1}^{n} \frac{\partial \log f_{\theta}(Y_i)}{\partial \theta} = \sum_{i=1}^{n} S_1(Y_i, \theta).
\]

As we shall see later, an immediate application of the score function is in finding the MLE: after observing \(Y = y\), set \(S(y, \theta) = 0\) and solve for \(\theta\).

Under mild regularity conditions, the score function has zero mean.

**Proposition 3.1.** Under regularity conditions,

\[
E_{\theta}S(Y, \theta) = 0 \text{ for all } \theta \in \Theta.
\]

**Proof.** For simplicity, we will assume that the density is absolutely continuous with respect to Lebesgue measure, but the same result holds for discrete or mixed distributions (integrating with respect to the appropriate measure).

\[
E_{\theta}S(Y, \theta) = \int S(y, \theta)f_{\theta}(y)dy
= \int \frac{L'(\theta)}{L(\theta)} L(\theta)dy
= \int \frac{\partial f_{\theta}(y)}{\partial \theta} dy
= \frac{\partial}{\partial \theta} \int f_{\theta}(y)dy
= 0.
\]

The variance of the score function (equivalently, its second moment) gives an often-convenient way to compute expected Fisher information introduced in the following subsection.

**Proposition 3.2.** Under regularity conditions,

\[
\text{Var}_{\theta}(S(Y, \theta)) = E_{\theta}(S^2(Y, \theta)).
\]

This follows directly from Proposition 3.1.

3.2.2. Fisher Information. The “expected” Fisher information, or simply the Fisher information for a single observation \(Y\) is defined as:

\[
I_1(\theta) = E_{\theta} \left( \frac{\partial}{\partial \theta} \log f_{\theta}(Y) \right)^2 = E_{\theta}(S^2_1(Y, \theta)),
\]
where $S_1(Y, \theta)$ is given by (3.1).

Fisher information based on data $Y = (Y_1, \ldots, Y_n)$ is defined as

$$I_n(\theta) = E_\theta \left( \frac{\partial}{\partial \theta} \log f_\theta(Y) \right)^2 = E_\theta (S^2(Y, \theta)).$$

As in the case of score function, for notational simplicity, we will drop the index $n$ and write the Fisher information for data $Y$ as $I(\theta)$, unless otherwise specified.

**Proposition 3.3.** Under mild regulatory conditions,

$$I(\theta) = -E_\theta \left( \frac{\partial^2}{\partial \theta^2} \log f_\theta(Y) \right).$$

\$3.2.\$ Prove Proposition 3.3

We now develop some important properties of Fisher information, which will be useful in computing information and provide some intuition for the highly suggestive term “information”.

A simple but fundamental property of Fisher information is *additivity*: if $Y_1$ and $Y_2$ are independent observations, then the information about $\theta$ from $(Y_1, Y_2)$ (given by $I_2(\theta)$ defined by (3.4) with $n = 2$) is the sum of the information from $Y_1$ and the information from $Y_2$. This simple property is both intuitively reasonable (information should accumulate as independent evidence arrives) and very useful for computation. In particular, for $Y_1, \ldots, Y_n$ i.i.d., we have $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is given by (3.3)

**Example 3.3 (Poisson).** Let $Y_1, \ldots, Y_n \sim \text{Poi}(\lambda)$. It is easy to see that (check this!) $I_1(\lambda) = 1/\lambda$ and $I(\lambda) = n/\lambda$.

**Example 3.4 (Location Families).** Let $f_\theta(y) = f(y - \theta)$ with $f$ a known density and $\theta$ a location parameter. Then

$$I_1(\theta) = \int S^2(y, \theta) f(y-\theta) dy = \int \left( \frac{f'(y-\theta)}{f(y-\theta)} \right)^2 f(y-\theta) dy = \int \left( \frac{f'(u)}{f(u)} \right)^2 f(u) du$$

is a constant (not depending on $y$ or $\theta$).

**Example 3.5 (Scale Families).** Let $f_\theta(y) = \theta^{-1} f(y/\theta)$ with $f$ a known density and $\theta > 0$ a scale parameter. Then a calculation similar to that of the previous example yields $I_1(\theta) \propto 1/\theta^2$ (checking this is left as an exercise). This makes sense since we can write $Y = \theta X$ with $X \sim f$, and then $\text{Var}(\theta X) = \theta^2 \text{Var}(X)$ (so small $\theta$ corresponds to small variance, in which case an observed $y$ is very informative).

3.2.3. The Cramér-Rao Theorem.

**Theorem 3.3.** [CRLB] Let $g(\theta)$ be a parametric function of interest and $T(Y)$ be an unbiased estimator of $g(\theta)$. Assume that the following regularity conditions hold:

(i) The parameter space $\Theta$ is a nonempty open set in $\mathbb{R}$;
(ii) The densities \( f_\theta(y) \) are all with respect to a dominating measure \( m \) which is sigma-finite (i.e., the sample space is a countable union of sets of finite measure);

(iii) The support of \( f_\theta(y) \) does not depend on \( \theta \);

(iv) The density \( f_\theta(y) \) is differentiable w.r.t. \( \theta \);

(v) The function \( g(\theta) \) is differentiable w.r.t. \( \theta \);

(vi) The Fisher information \( I(\theta) \) exists and is nonzero;

(vii) The estimator \( T(Y) \) has finite variance;

(viii) The integrals \( \int f_\theta(y)dy \) and \( \int T(y)f_\theta(y)dy \) can be differentiated under the integral sign, twice for the former.

Then,

\[
\text{Var}(T) \geq \frac{[g'(\theta)]^2}{I(\theta)},
\]

where \( I(\theta) \) is the Fisher information defined by (3.4).

**Proof.** Consider the Score function \( S(Y, \theta) \) defined by (3.2), and the covariance between \( S(Y, \theta) \) and \( T(Y) \). By the Cauchy-Schwarz inequality,

\[
\{\text{Cov}(S(Y, \theta), T(Y))\}^2 \leq \text{Var}(S(Y, \theta)) \cdot \text{Var}(T(Y)).
\]

Now,

\[
\text{Cov}(S(Y, \theta), T(Y)) = E[S(Y, \theta)T(Y)] - E[S(Y, \theta)]E[T(Y)]
\]

\[
= \int \frac{\partial \log f_\theta(y)}{\partial \theta} T(y)f_\theta(y)dy
\]

\[
= \int \frac{\partial f_\theta(y)}{f_\theta(y)} T(y)f_\theta(y)dy
\]

\[
= \frac{\partial}{\partial \theta} \int T(y)f_\theta(y)dy
\]

\[
= \frac{\partial}{\partial \theta} E_\theta(T(Y)) = \frac{\partial}{\partial \theta} g(\theta) = g'(\theta).
\]

Now, by Proposition 3.2, \( \text{Var}(S(Y, \theta)) = E[S(Y, \theta)]^2 \), which equals \( I(\theta) \) by (3.4). Substituting \( \text{Var}(S(Y, \theta)) = I(\theta) \) and \( \text{Cov}(S(Y, \theta), T(Y)) = g'(\theta) \) from (3.8) into the CS-inequality (3.7) the result follows. \( \square \)

**Remark 3.3.** We will call (i)-(viii) of Theorem 3.3 regularity conditions, or obvious modifications thereof, the “usual regularity conditions” as it is obviously tedious to state them all each time. The main points are that strange things can happen if the support of the distribution depends on the parameter, e.g., for the Uniform(0, \( \theta \)) model, and that we often need to differentiate under the integral sign to obtain the desired results. Analysis books contain various criteria under which differentiation under the integral sign is permissible, but since derivatives are limits, the problem is really just the question of when can one exchange a limit and an integral, and we already have criteria such as dominated convergence for this.
Here are two examples in which CRLB is achieved with equality.

**Example 3.6.** Consider \( n \) i.i.d. \( \mathcal{N}(\mu, \sigma^2) \) observations. Then \( \text{Var}(\bar{Y}) = \frac{\sigma^2}{n} = \frac{1}{I_n(\mu)} \).

**Example 3.7.** Consider \( n \) i.i.d. \( \text{Pois}(\lambda) \) observations. Then \( \text{Var}(\bar{Y}) = \frac{\lambda}{n} = \frac{1}{I_n(\lambda)} \).

So equality holds for the Normal and the Poisson, when estimating the mean parameters; but equality would not hold in those examples if estimating a nonlinear function of the mean. You will explore the question of when in general CRLB can be achieved with equality as a home work problem.

We shall see later, that the MLE asymptotically achieves the CRLB, but it does not typically achieve it with equality for any fixed sample size.

4. **Method of Moments and Maximum Likelihood Estimation**

4.1. **Method of Moments Estimation.** Method of moments is often a straightforward way of obtaining estimators of parameters or parametric functions by equating sample moments to population moments. Recall that for a random variable \( Y \) distributed with a probability density function \( f_\theta(y) \), the \( r \)th raw moment of the distribution of \( Y \) is

\[
\mu'_r = E_\theta(Y^r), \quad r = 1, 2, \ldots.
\]

The \( r \)th central moment is

\[
\mu_r = E_\theta((Y - \mu'_1)^r) = E_\theta((Y - E_\theta(Y))^r), \quad r = 1, 2, \ldots.
\]

Thus, for example, mean of the distribution is the first raw moment and the variance the second central moment.

**Definition 4.1.** Let \( Y_1, \ldots, Y_n \) be an iid random sample from \( f_\theta(y) \), where \( \theta \in \Theta \subset \mathbb{R}^k \). A method of moments estimator of \( \theta \) is obtained by equating the first \( k \) sample moments \( \sum_{i=1}^n Y_i^r/n \) for \( r = 1, \ldots, k \) to the corresponding \( k \) population (raw) moments \( \mu'_r(\theta) = E_\theta(Y^r) \) for \( r = 1, \ldots, k \).

**Example 4.1.** Let \( Y_1, \ldots, Y_n \) \( \text{iid} \) \( \mathcal{N}(\mu, \sigma^2) \). The method of moments estimators can be obtained by equating the first two sample and population moments, i.e.,

\[
\bar{Y} = \mu, \quad \frac{1}{n} \sum_{i=1}^n Y_i^2 = E(Y_1^2) = \mu^2 + \sigma^2.
\]

Solving the above two equations, we get the following method of moments (MOM) estimators for \( \mu \) and \( \sigma^2 \):

\[
\hat{\mu}_{\text{MOM}} = \bar{Y},
\]

\[
\hat{\sigma}_{\text{MOM}}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2.
\]
Example 4.2. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Binom}(k, p)$, where both $k$ and $p$ are unknown. Equating the first two sample moments to those of the population, we have

$$
\bar{Y} = kp, \\
\frac{1}{n} \sum_{i=1}^{n} Y_i^2 = E(Y_1^2) = kp(1 - p) + k^2p^2.
$$

Solving the above two equations, the method of moments estimators of $k$ and $p$ are obtained as:

$$
\hat{k}_{\text{MOM}} = \frac{\bar{Y}^2}{\bar{Y} - \sum_{i=1}^{n} (Y_i - \bar{Y})^2}, \\
\hat{p}_{\text{MOM}} = \bar{Y}/\hat{k}_{\text{MOM}}.
$$

These MOM estimators are not good, because both of them can be negative if the sample mean is smaller than the sample variance. However, obtaining MLEs of $k$ and $p$ is a non-trivial exercise. Like many other situations where the MLE is difficult to obtain, in this example the method of moments estimators provide a reasonable starting point for obtaining good estimators.

Example 4.3 (MOM estimators for the exponential family). Consider the single parameter exponential family of distributions characterized by the pdf

$$
f_\theta(y) = \exp\{\eta(\theta)T(y) - \psi(\eta)\} h(y).
$$

To obtain the method of moments estimator for $\theta$, note that by differentiating both sides of the identity

$$
\int \exp\{\eta(\theta)T(y) - \psi(\eta)\} h(y)dy = 1
$$

within the integral with respect to $\eta$ yields

$$
\int \exp\{\eta T(y) - \psi(\eta)\} (T(y) - \psi'(\eta)) h(y)dy = 0
$$

where $\psi'(\eta) = \partial\psi(\eta)/\partial\eta$. The above implies

$$
\int T(y) \exp\{\eta T(y) - \psi(\eta)\} h(y)dy = \psi'(\eta) \int \exp\{\eta T(y) - \psi(\eta)\} h(y)dy,
$$

which implies $E_{\theta}(T(Y)) = \psi'(\eta)$.

Thus, the following equation (4.1) yields the MOM estimator for the parameter $\theta$ of an exponential family:

$$
(4.1) \quad \frac{1}{n} \sum_{i=1}^{n} T(Y_i) = \psi'(\eta).
$$
4.2. **Maximum Likelihood Estimation.** A widely used method of obtaining a point estimate for a parameter \( \theta \) is to find the maximum likelihood estimate (MLE). As the name suggests, the MLE is defined as any vector \( \hat{\theta}_n \) maximizing \( L(\theta) \).

The MLE is often a reasonable estimate, but it should not be put to use blindly without inspecting the entire likelihood function. For example, it sometimes happens that other values of \( \theta \) are far more centrally located than the MLE (the mode of a distribution need not be near the mean!). Or there may be several peaks in the likelihood function, perhaps negligibly different in height. Careful study of the shape of \( L(\theta) \) is more informative than locating a single point with slightly higher likelihood than other points.

\[\textbf{\# 4.1.} \text{Saying that the likelihood function looks like a Normal curve is a very different statement than saying that the distribution of the MLE is approximately Normal!}\]

**Example 4.4** (MLE in an exponential family). Let
\[Y_1, \ldots, Y_n \sim \text{exp}\left\{ \eta(\theta) T(y) - \psi(\eta) \right\} h(y).\]

The log-likelihood equation is
\[l(\theta, Y) = \left( \sum_{i=1}^{n} T(Y_i) \right) \eta(\theta) - n \psi(\eta) + \sum_{i=1}^{n} \log h(Y_i).\]

Differentiating both sides with respect to \( \eta \) yields
\[\frac{\partial l}{\partial \eta} = \sum_{i=1}^{n} T(Y_i) - n \psi'(\eta).\]

Equating the above to zero, we obtain, the ML equation as
\[\frac{1}{n} \sum_{i=1}^{n} T(Y_i) = \psi'(\eta),\]

which is exactly the same as equation [4.1] that yields the MOM estimator.

4.2.1. **Computing the MLE.** The following Cauchy location problem illustrates many important ideas for finding the MLE. This is not an exponential family, and can have a very complicated likelihood function with many local maxima.

**Example 4.5.** Let \( Y_1, Y_2, \ldots, Y_n \) be i.i.d. as a Cauchy location family, i.e., \( Y_i = C_i + \theta \) with \( C_i \) i.i.d. Cauchy. So
\[f_{\theta}(y) = \frac{1}{\pi(1 + (y - \theta)^2)}.\]

We wish to estimate the location parameter \( \theta \). The log-likelihood function is
\[l(\theta) = -\sum_{j=1}^{n} \ln(1 + (y_j - \theta)^2),\]
so the score function is
\[ S(y, \theta) = 2 \sum_{j=1}^{n} W_j (y_j - \theta), \]
where the \( W_j \) are “weights” defined by
\[ W_j(y, \theta) = \frac{1}{1 + (y_j - \theta)^2}. \]
The MLE \( \hat{\theta}_n \) satisfies \( S(y, \theta) = 0 \), so it satisfies
\[ \theta = \frac{\sum_{j=1}^{n} W_j y_j}{\sum_{j=1}^{n} W_j}, \]
which is a weighted average of the \( y_j \). Note that the weights depend on both \( \theta \) and on the \( y_j \), with the more extreme values of \( y_j \) given less weight.

A solution to the above equation is a fixed point of the function
\[ h(\theta) \equiv \frac{\sum_{j=1}^{n} W_j(y_j, \theta) y_j}{\sum_{j=1}^{n} W_j(\theta, y)}, \]
i.e., we are looking for \( \hat{\theta}_n \) satisfying \( h(\hat{\theta}_n) = \hat{\theta}_n \). In such a setting, iterative techniques are usually used.

We first explain how the Newton-Raphson method can be applied here. Start with a guess \( \theta_1 \) for the solution to \( S(y, \hat{\theta}_n) = 0 \) (e.g., the median of \( y_1, \ldots, y_n \)). Expand the score function as a Taylor series (up through the linear term):
\[ S(y, \theta_2) \approx S(y, \theta_1) + (\theta_2 - \theta_1) S'(y, \theta_1). \]
Setting \( S(y, \theta_2) = 0 \) (the hoped-for value) and solving for \( \theta_2 \) gives
\[ \theta_2 = \theta_1 + \frac{S(y, \theta_1)}{-S'(y, \theta_1)}. \]
(The ‘ denotes the derivative with respect to the \( \theta \) component.) To evaluate this, note that
\[ -S'(y, \theta) = 2 \sum_{j=1}^{n} W_j (2W_j - 1) = 2 \sum_{j=1}^{n} (2W_j^2 - W_j), \]
so
\[ \theta_2 = \theta_1 + \frac{\sum_{j=1}^{n} W_j(y_j - \theta)}{\sum_{j=1}^{n} (2W_j^2 - W_j)}. \]
The procedure can then be iterated to find \( \theta_3, \theta_4, \ldots \) until convergence. Note though that this may not give the global maximum (it may even give a local minimum). The same idea can be used for multi-dimensional \( \theta \), using gradients and Hessians.

A closely-related alternative is to use Fisher’s method of scoring. This method replaces \( -S'(Y, \theta_1) \) by its average \( E_{\theta_1}(-S'(Y, \theta_1)) = I(\theta_1) \). Fisher scoring has the advantages that it is sometimes easier to compute and work
with \( I(\theta_1) \) than \(-S'(Y, \theta_1)\), and that it is often less sensitive to the initial guess \( \theta_1 \).

Returning to the specifics of the Cauchy location problem, let us compute \( I(\theta) \). As shown above, in a location family \( I(\theta) \) is a constant. Thus, we can choose \( \theta \) to make the computations as simple as possible. Let us take \( \theta = 0 \).

Then
\[
W_j = \frac{1}{1 + Y_j^2} = \frac{Z_2^2}{Z_2^2 + Z_1^2}
\]
with \( Z_1, Z_2 \) i.i.d. \( \mathcal{N}(0, 1) \), using the fact that a Cauchy is the ratio of two standard Normal r.v.s. But if \( Z \sim \mathcal{N}(0, 1) \), then \( Z^2 \sim \chi_1^2 \sim 2\text{Gamma}(1/2) \), so \( W_j \sim \text{Beta}(1/2, 1/2) \).

In particular,
\[
E_0 W_j = \frac{1}{2}, \quad \text{Var}_0 W_j = \frac{1}{8}.
\]
Therefore,
\[
I(\theta) = E_\theta(-S'(Y, \theta)) = -nE_\theta(2W_1(2W_1 - 1)) = nE_\theta(4W_1^2 - 2W_1) = \frac{n}{2}.
\]
This does not depend on \( \theta \), so \( I(\hat{\theta}_n) \) also equals \( n/2 \). Thus, Fisher scoring uses
\[
\theta_2 = \theta_1 + \frac{4}{n} \sum_{j=1}^{n} W_j (y_j - \theta).
\]

4.2.2. Properties of the MLE. The following result states that the MLE of a function \( g(\theta) \) can be obtained by applying \( g \) to the MLE for \( \theta \).

**Proposition 4.1** (Equivariance of MLE). If \( \hat{\theta}_n \) is a MLE for \( \theta \) and \( \tau = g(\theta) \), with \( g \) a one-to-one function, then an MLE for \( \tau \) is \( \hat{\tau} = g(\hat{\theta}_n) \).

The above property is handy but, in a sense, too handy. For example, if \( \hat{\theta}_n \) is the MLE for \( \theta \), the equivariance property implies that \( e^{\hat{\theta}_n} \) is the MLE for \( e^{c\theta} \) for any constant \( c \). This will generally be terribly biased unless \( c \) is close to 0.

The following property establishes a connection of the MLE to a sufficient statistic:

**Proposition 4.2** (MLE and sufficiency). The MLE of a parameter \( \theta \) is a function of every statistic \( T \) that is sufficient for \( \theta \).

**Proof.** Let \( T = T(Y) \) be any sufficient statistic. Then,
\[
L(\theta) = f_\theta(Y) = g_\theta(T(Y)) h(Y),
\]
by the factorization theorem. Because \( \hat{\theta}_n \) maximizes \( L(\theta) \), it follows that \( \hat{\theta}_n \) maximizes \( g_\theta(T(Y)) \). The value of \( \theta \) that maximizes \( g_\theta(T(Y)) \) must be a function of \( T(Y) \). \( \square \)

**Consistency of MLE**

Let \( \hat{\theta}_n \) denote the MLE of a parameter \( \theta \) obtained from an iid sample \( Y_1, \ldots, Y_n \) drawn from \( f_\theta(y) \). Assume that there exists a “true value” \( \theta_0 \) of \( \theta \). Then, under mild regularity conditions, the MLE can be shown to be
a consistent estimator of \( \theta \) (i.e., it converges almost surely (strong convergence) or in probability (weak convergence) to the true value \( \theta_0 \)). We now discuss the idea behind the proof of consistency of the MLE, which can be presented at different levels of rigor.

Recall that the MLE is obtained by maximizing the log-likelihood function \( f_\theta(Y) = \sum_{i=1}^n \log f_\theta(Y_i) \), which is the same as maximizing the “average log-likelihood” \( n^{-1} \sum_{i=1}^n \log f_\theta(Y_i) \). Denote this quantity by

\[
(4.3) \quad \bar{l}_n(\theta) = n^{-1} \sum_{i=1}^n \log f_\theta(Y_i).
\]

Then, by definition,

\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \bar{l}_n(\theta).
\]

The first part of the proof involves arguing that \( \bar{l}_n(\theta) \) converges (pointwise) to

\[
\bar{l}(\theta, \theta_0) = E_{\theta_0} [\log f_\theta(Y_1)].
\]

By the strong law of large numbers (SLLN), it follows that

\[
(4.4) \quad \bar{l}_n(\theta) \xrightarrow{a.s.} \bar{l}(\theta, \theta_0),
\]

for each \( \theta \) under the regularity condition \( |\bar{l}(\theta, \theta_0)| < \infty \) for all \( \theta \) in the neighborhood of \( \theta_0 \).

The second part of the proof shows that whereas by definition, the LHS of (4.4) is maximized by the MLE \( \hat{\theta}_n \), the RHS is maximized by \( \theta_0 \), the true value of the parameter. We formally state this result in the form of the following lemma.

**Lemma 4.1.** If the density \( f_\theta(y) \) is identifiable, i.e.,

\[
\theta_1 \neq \theta_2 \implies f_{\theta_1}(y) \neq f_{\theta_2}(y)
\]

for at least one \( y \), then

\[
(4.5) \quad \bar{l}(\theta_0, \theta_0) > \bar{l}(\theta, \theta_0) \quad \forall \, \theta \neq \theta_0.
\]

**Proof.** It needs to be shown that

\[
E_{\theta_0} [\log f_{\theta_0}(Y)] > E_{\theta_0} [\log f_\theta(Y)] \quad \forall \, \theta \neq \theta_0.
\]

Consider the quantity

\[
E_{\theta_0} \left[ \log \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right].
\]

Since \( \log \) is a concave function, by Jensen’s inequality,

\[
E_{\theta_0} \left[ \log \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right] \leq \log E_{\theta_0} \left[ \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right] = \log \int \frac{f_\theta(y)}{f_{\theta_0}(y)} f_{\theta_0}(y) dy = 0.
\]

This implies

\[
E_{\theta_0} \left[ \log \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right] \leq 0
\]
and consequently,

\[ E_{\theta_0} [\log f_{\theta_0}(Y)] \geq E_{\theta_0} [\log f_{\theta}(Y)] \quad \forall \, \theta, \]

Equality holds in the above if and only if

\[ f_{\theta}(Y) = f_{\theta_0}(Y) \quad \forall \, Y. \]

Therefore, under the assumption of identifiability, the equality is strict. □

We have now established the following:

Result 1: Under Condition A:

\[ |\bar{\ell}(\theta, \theta_0)| < \infty \quad \text{for all} \quad \theta \quad \text{in the neighborhood of} \quad \theta_0, \]

\[ \bar{\ell}_n(\theta) \xrightarrow{a.s.} \bar{\ell}(\theta, \theta_0) \quad \text{pointwise, for each} \quad \theta \quad \text{(By equation (4.4)).} \]

Result 2: Under Condition B: Identifiability, i.e., \( \theta_1 \neq \theta_2 \implies f_{\theta_1}(y) \neq f_{\theta_2}(y) \) for at least one \( y \),

\( \bar{\ell}(\theta, \theta_0) \) is maximized by \( \theta_0 \). (Lemma 4.1)

Are these two results enough to argue that \( \hat{\theta}_n \), the maximizer of \( \bar{\ell}_n(\theta) \), converges to \( \theta_0 \), the maximizer of \( \bar{\ell}(\theta, \theta_0) \)? No, because what we have proven so far does not guarantee that the limit of the sequence of maximums \( \{\hat{\theta}_n\} \) is equal to the maximum of the limiting curve. We need additional conditions to establish such convergence. One such condition is the following:

Condition C: Assume that \( \log f_{\theta}(y) \) has a continuous derivative with respect to \( \theta \) in a neighborhood of \( \theta_0 \) for each \( y \).

We will choose an arbitrary interval \([\theta_0 - \delta, \theta_0 + \delta]\) around \( \theta_0 \) and show that condition C, result 1 and result 2 imply that \( \partial \bar{\ell}_n(\theta)/\partial \theta = 0 \) has a solution almost surely in this interval. Then the convergence can be established by the arbitrary choice of \( \delta \).

First, note that by result 2, for given \( \delta \), we can find \( \varepsilon > 0 \) such that

\begin{align*}
(6.6) & \quad \bar{\ell}(\theta_0, \theta_0) - \bar{\ell}(\theta_0 - \delta, \theta_0) > \varepsilon, \\
(6.7) & \quad \bar{\ell}(\theta_0 + \delta, \theta_0) - \bar{\ell}(\theta_0, \theta_0) > \varepsilon.
\end{align*}

Now, consider a subset \( \Omega_1 \) of the underlying sample space with \( P(\Omega_1) = 1 \) and such that Result 1 holds for at \( \theta = \theta_0 - \delta \) and \( \theta = \theta_0 + \delta \). For each \( \omega \in \Omega_1 \), we can choose a sample size large enough (i.e., \( \forall \, n > n(\varepsilon, \omega) \)) such that

\begin{align*}
(6.8) & \quad |\bar{\ell}_n(\theta_0 - \delta) - \bar{\ell}(\theta_0 - \delta, \theta_0)| < \varepsilon/2, \\
(6.9) & \quad |\bar{\ell}_n(\theta_0) - \bar{\ell}(\theta_0, \theta_0)| < \varepsilon/2, \\
(6.10) & \quad |\bar{\ell}_n(\theta_0 + \delta) - \bar{\ell}(\theta_0 + \delta, \theta_0)| < \varepsilon/2.
\end{align*}

Note that \( \omega \) does not explicitly appear in each inequality, but \( \{\bar{\ell}_n\}_{n=1}^{\infty} \) is actually a different sequence for each \( \omega \) in the underlying sample space.
Thus, for all $n > n(\varepsilon, \omega)$,
\[
\bar{l}_n(\theta_0) - \bar{l}_n(\theta_0 - \delta) = \bar{l}_n(\theta_0) - \bar{l}(\theta_0, \theta_0) + \bar{l}(\theta_0, \theta_0) - \bar{l}(\theta_0, \theta_0 - \delta) + \bar{l}(\theta_0, \theta_0 - \delta) - \bar{l}_n(\theta_0 - \delta) > -\varepsilon/2 + \varepsilon - \varepsilon/2 \geq 0.
\]

Similarly, we can show that $\bar{l}_n(\theta_0) > \bar{l}_n(\theta_0 + \delta)$. Assuming condition C, i.e., $l_n(\theta)$ has a continuous derivative w.r.t $\theta$ in the neighborhood of $\theta_0$ for each $y$ in the support of $\theta_0$. Substituting $\theta_0 = 0$,

\[
\bar{l}_n(\theta_0) = \bar{l}_n(\theta_0 - \delta) + \bar{l}(\theta_0, \theta_0) - \bar{l}(\theta_0, \theta_0 - \delta) + \bar{l}(\theta_0, \theta_0 - \delta) - \bar{l}_n(\theta_0 - \delta) > 0.
\]

Asymptotic normality of MLE

**Theorem 4.1.** Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f_{\theta}(y)$, where $\theta_0$ is an interior point of $\Theta$ and $f_\theta(y)$ satisfies the following regularity conditions:

(i) $\frac{\partial^2}{\partial \theta^2} \log f_\theta(y)$ exists locally at $\theta_0$, and there exists some $\delta > 0$ such that $\sup_{|\theta - \theta_0| < \delta} |\frac{\partial^2}{\partial \theta^2} \log f_\theta(y)| \leq g(y)$ holds for some $g$ where $E_{\theta_0}g(Y) < \infty$.

(ii) $-E_\theta \left( \frac{\partial^2 \log f_\theta(Y)}{\partial^2 \theta} \right) |_{\theta = \theta_0} = I_1(\theta_0)$.

Let $\hat{\theta}_n$ be the MLE of $\theta_0$ satisfying:

(4.11) $\bar{l}_n(\hat{\theta}_n) = \sum_{i=1}^{n} \frac{\partial \log f_\theta(Y_i)}{\partial \theta} = 0$.

Then if $\hat{\theta}_n \rightarrow_{p} \theta_0$,

\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \overset{d}{\rightarrow} N \left( 0, (I_1(\theta_0))^{-1} \right)
\]
as $n \rightarrow \infty$.

**Proof.** We use a Taylor series expansion of $\bar{l}_n(\hat{\theta}_n)$ around $\theta_0$:

$\bar{l}_n(\hat{\theta}_n) = \bar{l}_n(\theta_0) + \bar{l}'_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2} \bar{l}''_n(\theta^*) (\hat{\theta}_n - \theta_0)^2$,

where $\theta^*$ is between $\hat{\theta}_n$ and $\theta_0$. Substituting $\bar{l}_n(\hat{\theta}_n) = 0$ from (4.11), and rearranging terms, we have that

\[
(\hat{\theta}_n - \theta_0) = \frac{-\bar{l}_n(\theta_0)}{\bar{l}'_n(\theta_0) + \frac{1}{2} \bar{l}''_n(\theta^*)(\hat{\theta}_n - \theta_0)^2},
\]

which implies

\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) = \frac{-\bar{l}_n(\theta_0)}{\frac{1}{n} \bar{l}'_n(\theta_0) + \frac{1}{2n} \bar{l}''_n(\theta^*)(\hat{\theta}_n - \theta_0)^2}.
\]
Note that
\[ \frac{1}{n} \hat{p}_n'(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_{\theta}(Y_i)}{\partial^2 \theta} \bigg|_{\theta=\theta_0} \rightarrow P \left( \hat{\theta}_n - \theta_0 \right) = -I_1(\theta_0), \quad \text{by WLLN.} \]

Furthermore, fix \( \epsilon > 0 \) and choose \( \delta > 0 \) small enough,
\[ P_{\theta_0} \left( \left| \frac{1}{2n} \hat{p}_n''(\theta^*) \right| > \epsilon \right) \leq P_{\theta_0} \left( \sup_{\theta:|\theta-\theta_0|\leq\delta} \left| \frac{1}{2n} \hat{p}_n''(\theta) \right| > \epsilon \right) + P_{\theta_0}(\left| \hat{\theta}_n - \theta_0 \right| > \delta) \]
\[ \leq P_{\theta_0} \left( \frac{1}{2n} \sum_{i=1}^{n} g(Y_i) > \frac{\epsilon}{\delta} \right) + P_{\theta_0}(\left| \hat{\theta}_n - \theta_0 \right| > \delta) \rightarrow 0 \]
as \( n \rightarrow \infty \) followed by \( \delta \rightarrow 0 \). This shows that
\[ \frac{1}{2n} \hat{p}_n''(\theta^*)(\hat{\theta}_n - \theta_0) \rightarrow P 0. \]

Consequently, applying Slutsky's theorem, the result is established. \( \square \)

**Remark 4.1.** The asymptotic normality of MLE established through Theorem 4.1 has the following implications:

1. The MLE is asymptotically unbiased.
2. For large \( n \) the variance of the MLE can be approximated by \( (nI_1(\theta_0))^{-1} \), which is the CRLB for unbiased estimators of \( \theta \). Thus, \( \hat{\theta}_n \) has the smallest asymptotic variance among unbiased estimators, and hence is called asymptotically efficient.

## 5. General theory for M-estimation

### 5.1. General framework

The theory for the maximum likelihood estimator discussed in the previous section is only one example for a much larger class of estimation procedures. Suppose we observe \( Y_1, \ldots, Y_n \), and an M-estimator \( \hat{\theta}_n \) is defined by

\[ \hat{\theta}_n \equiv \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_{\theta}(Y_i), \]

where \( \{m_{\theta} : \theta \in \Theta\} \) is a class of criteria/loss functions indexed by the (finite-dimensional) parameter \( \theta \in \Theta \).

In many cases, solving (5.1) can be reduced to a class of estimating equations by setting the derivative(s) to be zero:

\[ \frac{1}{n} \sum_{i=1}^{n} \psi_{\hat{\theta}_n}(Y_i) = 0. \]
In particular, if $\theta \mapsto m_\theta(y)$ is differentiable for all $y$, then we may take $\psi_\theta = m'_\theta$. An estimator defined through \((5.2)\) is called an $Z$-estimator.

Let us fix some notation that will be used frequently throughout this section. Let

\[
M_n(\theta) = \mathbb{P}_n m_\theta \equiv \frac{1}{n} \sum_{i=1}^{n} m_\theta(Y_i),
\]

\[
\Psi_n(\theta) = \mathbb{P}_n \psi_\theta = \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(Y_i),
\]

and let

\[
\mathbb{G}_n(f) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(Y_i) - Pf)
\]

be the empirical process.

**Example 5.1.** Let $Y_1, \ldots, Y_n$ be i.i.d. with a common density $f_\theta$. The maximum likelihood estimator $\hat{\theta}_n$ studied in the previous section is included in the $M$-estimation framework by letting $m_\theta = \log f_\theta$:

\[
\hat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_\theta(Y_i).
\]

**Example 5.2.** Let $Y_1, \ldots, Y_n$ be i.i.d. samples from a common distribution, and we want to estimate the ‘location’ of this distribution. Here ‘location’ is a vague term, and can have different meanings:

- Solving zero for
  \[
  \sum_{i=1}^{n} (Y_i - \theta) = 0
  \]
  leads to the mean for the empirical distribution.

- Solving zero for
  \[
  \sum_{i=1}^{n} \text{sign}(Y_i - \theta) = 0
  \]
  leads to the median for the empirical distribution.

A general form for the ‘location’ estimator $\hat{\theta}_n$ is given by the zero of the estimating equation

\[
\sum_{i=1}^{n} \psi(Y_i - \theta) = 0,
\]

where $\psi$ is some monotone and odd function. Here are some further choices:

- (Huber loss) $\psi_k(y) = (y \wedge k) \vee (-k)$.
- (Quantile function) $\psi_p(y) = -(1-p)1_{y<0} + p1_{y>0}$. Note that $p = 1/2$ corresponds to the median.
5.2. **Consistency.** First we consider consistency for $M$-estimators. Since
\[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} M_n(\theta), \]
if $M_n \to M$ for some function $M$ ‘nicely’, then we may naturally expect that
\[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} M_n(\theta) \to \arg \max_{\theta \in \Theta} M(\theta) = \theta_0. \]
We make this heuristic argument rigorous below.

**Theorem 5.1.** Let $M_n$ be random functions and $M$ be a fixed function defined on a metric space $(\Theta, d)$. Suppose that for any $\varepsilon > 0$,
\[ (1) \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to_p 0, \]
\[ (2) \sup_{d(\theta, \theta_0) > \varepsilon} M(\theta) < M(\theta_0). \]
Then $\hat{\theta}_n \to_p \theta_0$.

**Proof.** Since
\[ M_n(\hat{\theta}_n) \geq M_n(\theta_0) \quad \text{by definition of } \hat{\theta}_n \]
\[ = M(\theta_0) + o_p(1) \quad \text{by (1)}, \]
it follows that
\[ M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\theta_0) + o_p(1) \]
\[ \leq \sup_{\theta \in \Theta} (M_n - M)(\theta) + o_p(1) \to_p 0. \]
The claim now follows from the assumption (2) (why?). $\square$

Next we consider $Z$-estimators. The results are broadly similar in spirit to the above theorem.

**Theorem 5.2.** Let $\psi_n$ be random functions and $\psi$ be a fixed function defined on a metric space $(\Theta, d)$. Suppose that for any $\varepsilon > 0$,
\[ (1) \sup_{\theta \in \Theta} \|\psi_n(\theta) - \psi(\theta)\| \to_p 0, \]
\[ (2) \inf_{d(\theta, \theta_0) > \varepsilon} \|\psi(\theta)\| > 0 = \|\psi(\theta_0)\|. \]
Then $\hat{\theta}_n \to_p \theta_0$.

**Proof.** Apply the previous theorem with $M_n(\theta) = -\|\psi_n(\theta)\|$ and $M(\theta) = -\|\psi(\theta)\|$. $\square$

Condition (1) in the above theorems is usually an ‘empirical process condition’: for the $M$-estimation framework in the i.i.d. case, condition (1) boils down to
\[ \sup_{\theta \in \Theta} |(\mathbb{P}_n - P)m_\theta| \to_p 0. \]
The class $\{m_\theta : \theta \in \Theta\}$ is called $P$-Glivenko-Cantelli if the above display holds true. The Glivenko-Cantelli property can typically be checked purely by the geometric properties of the class $\{m_\theta : \theta \in \Theta\}$, and has been well-known since the 1980s.
5.3. **Asymptotic normality.** For asymptotic normality results, we slightly reverse the order and start with the $Z$-estimation framework. Recall that a $Z$-estimator $\hat{\theta}_n$ is given by the zero of the estimating equation

$$
\Psi_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \psi_{\hat{\theta}_n}(Y_i) = 0.
$$

Using Taylor expansion, we have

$$
0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\Psi_n''(\tilde{\theta}_n),
$$

where $\tilde{\theta}_n$ is some point in between $\theta_0$ and $\hat{\theta}_n$. Therefore we may rewrite the above display as follows:

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}\Psi_n(\theta_0)}{\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi_n''(\tilde{\theta}_n)}.
$$

Since we expect

$$
\sqrt{n}\Psi_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(Y_i) \approx N(0, P\psi_{\theta_0}^2),
$$

$$
\Psi_n'(\theta_0) \rightarrow P\psi_{\theta_0}',
$$

$$
\frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi_n''(\tilde{\theta}_n) = o_p(1) \cdot O_p(1) = o_p(1),
$$

it is natural to expect the following result

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, P\psi_{\theta_0}^2/(P\psi_{\theta_0}'^2)).
$$

This is exactly the same development as we have seen previously in the case for the maximum likelihood estimator. One major drawback of this approach, however, is that although the asymptotic normality by itself only uses the first derivative of $\psi_{\theta}$, the proof requires at least a second derivative for $\psi_{\theta}$. Such a requirement is often too strong. Below we will use more modern (= ‘empirical process’) approach to alleviate such a strong smoothness requirement.

**Theorem 5.3.** Suppose that:

1. The map $\theta \mapsto P\psi_{\theta}$ is differentiable at $\theta_0$ with non-singular derivative matrix $V_{\theta_0}$.
2. For any $\delta_n \rightarrow 0$, $\sup_{d(\theta, \theta_0) < \delta_n} |\mathcal{G}_n(\psi_{\theta} - \psi_{\theta_0})| \rightarrow_p 0$.

If $\hat{\theta}_n \rightarrow_p \theta_0$, then

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(Y_i) + o_p(1)
$$

$$
\rightarrow_d N(0, V_{\theta_0}^{-1}P\psi_{\theta_0}\psi_{\theta_0}'V_{\theta_0}^{-1}).
$$
Proof. By consistency of \( \hat{\theta}_n \) and (2), we have
\[
o_p(1) = \mathbb{G}_n \psi_{\hat{\theta}_n} - \mathbb{G}_n \psi_{\theta_0} = \sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n}) - \mathbb{G}_n \psi_{\theta_0}
\]
(since \( \mathbb{P}_n \psi_{\hat{\theta}_n} = P\psi_{\theta_0} = 0 \))
\[
= \sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + \sqrt{n} \cdot o_p(\|\theta_0 - \hat{\theta}_n\|) - \mathbb{G}_n \psi_{\theta_0}.
\]
This implies
\[
(5.3) \quad \sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + o_p(\sqrt{n}\|\theta_0 - \hat{\theta}_n\|) = \mathbb{G}_n \psi_{\theta_0} + o_p(1).
\]
Note that if we can prove \( \sqrt{n}\|\theta_0 - \hat{\theta}_n\| = O_p(1) \), then (5.3) immediately yields that the claim of the theorem. To see this, by (5.3),
\[
\sqrt{n}\|\theta_0 - \hat{\theta}_n\| \leq \|V_{\theta_0}^{-1}\| \cdot \sqrt{n}\|\theta_0 - \hat{\theta}_n\| \leq O_p(1) + o_p(\sqrt{n}\|\theta_0 - \hat{\theta}_n\|)
\]
\[
\iff (1 - o_p(1))\sqrt{n}\|\theta_0 - \hat{\theta}_n\| = O_p(1),
\]
as desired. \( \square \)

Next consider the M-estimators.

**Theorem 5.4.** Suppose that:

(1) The map \( \theta \mapsto Pm_\theta \) admits a second-order Taylor expansion at \( \theta_0 \) with non-singular second derivative matrix (Hessian) \( V_{\hat{\theta}_0} \).

(2) For any \( M \) > 0,
\[
\sup_{\|h\| \leq M} \left| \mathbb{G}_n \sqrt{n}(m_{\theta_0 + h/\sqrt{n}} - m_{\theta_0}) - h^\top \mathbb{m}_{\theta_0} \right| \rightarrow_p 0.
\]
If \( \sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1) \), then
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, V_{\theta_0}^{-1}Pm_{\theta_0}m_{\theta_0}^\top V_{\theta_0}^{-1}).
\]

**Proof.** Condition (2) entails that for any \( h_n = O_p(1) \),
\[
o_p(1) = \mathbb{G}_n \left( \sqrt{n}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) \right) - \mathbb{G}_n (h_n^\top \mathbb{m}_{\theta_0})
\]
\[
= n\mathbb{P}_n (m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - nP(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - \mathbb{G}_n (h_n^\top \mathbb{m}_{\theta_0})
\]
\[
= n\mathbb{P}_n (m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - \frac{1}{2} h_n^\top V_{\theta_0} h_n + o(1) - \mathbb{G}_n (h_n^\top \mathbb{m}_{\theta_0}),
\]
which implies that
\[
n\mathbb{P}_n (m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2} h_n^\top V_{\theta_0} h_n + \mathbb{G}_n (h_n^\top \mathbb{m}_{\theta_0}) + o_p(1).
\]
Now let \( \hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) \) and \( \tilde{h}_n = -V_{\theta_0}^{-1}\mathbb{G}_n \mathbb{m}_{\theta_0} \), we obtain
\[
n\mathbb{P}_n (m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2} \hat{h}_n^\top V_{\theta_0} \hat{h}_n + \mathbb{G}_n (\hat{h}_n^\top \mathbb{m}_{\theta_0}) + o_p(1)
\]
\[
n\mathbb{P}_n (m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = -\frac{1}{2} (\mathbb{G}_n \mathbb{m}_{\theta_0})^\top V_{\theta_0}^{-1} (\mathbb{G}_n \mathbb{m}_{\theta_0}) + o_p(1).
\]
Using the definition of $\hat{\theta}_n$, we have
\[
\frac{1}{2} \hat{h}_n^\top V_0 \hat{h}_n + \mathbb{G}_n(\hat{h}_n^\top \hat{m}_0) + o_p(1)
\]
\[
= n \mathbb{E}_n (m_{\theta_0 + \hat{h}_n/\sqrt{n}} - m_{\theta_0})
\]
\[
\geq n \mathbb{E}_n (m_{\theta_0 + \hat{h}_n/\sqrt{n}} - m_{\theta_0})
\]
\[
= -\frac{1}{2} (\mathbb{G}_n \hat{m}_0) \mathbb{V}^{-1}_0 \hat{m}_0 \mathbb{V}_0 \mathbb{G}_n \hat{m}_0 + o_p(1),
\]
which is equivalent to
\[
\frac{1}{2} (\hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_0) \mathbb{V}_0 (\hat{h}_n + V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_0) + o_p(1) \geq 0.
\]
This shows that
\[
\hat{h}_n = \sqrt{n} (\hat{\theta}_n - \theta_0) = V_{\theta_0}^{-1} \mathbb{G}_n \hat{m}_0 + o_p(1),
\]
completing the proof.

Condition (2) in the above theorems is usually called ‘Donsker property’ for the underlying class of interest. One classical condition verifying such Donsker property is given by the Lipschitz condition:

- In the $Z$-estimation framework, we require the following: for any $\theta_1, \theta_2$ close enough to $\theta_0$, there exists some $\hat{\psi}$ with $P \hat{\psi}^2 < \infty$ such that
  \[
  \|\psi_{\theta_1}(y) - \psi_{\theta_2}(y)\| \leq \hat{\psi}(y) \|\theta_1 - \theta_2\|.
  \]
- In the $M$-estimation framework, we require the following: for any $\theta_1, \theta_2$ close enough to $\theta_0$, there exists some $\hat{m}$ with $P \hat{m}^2 < \infty$ such that
  \[
  \|m_{\theta_1}(y) - m_{\theta_2}(y)\| \leq \hat{m}(y) \|\theta_1 - \theta_2\|.
  \]

It is easy to check that this condition requires much less differentiability of the log-likelihood function in the context of maximum likelihood estimation.

**Example 5.3.** Let $Y_1, \ldots, Y_n$ be i.i.d. samples from an unknown distribution $F$, and we are interested in estimating the median $\theta_0$ of $F$. We assume that $F$ is differentiable at $\theta_0$ with positive derivative $f(\theta_0)$. Consider the sample median
\[
\hat{\theta}_n = \arg \max_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \theta|.
\]
We may apply Theorem 5.4 with $m_{\theta}(y) = -|y - \theta| + |y|$. By triangle inequality we may take $\hat{m} \equiv 1$. The map $\theta \mapsto m_{\theta}(y)$ is differentiable at
\[ \theta_0 \text{ except for } y = \theta_0 \text{ with derivative } m_{\theta_0}(y) = -\text{sign}(y - \theta_0). \] Furthermore, by a somewhat tedious calculation using integration by parts we have

\[ P_{m\theta} = -\int |x - \theta| \, dF(x) + \int |x| \, dF(x) = -2\int_0^\theta F(x) \, dx + \theta. \]

This means that

\[ \frac{d}{d\theta} P_{m\theta} = -2F(\theta) + 1 (= 0 \text{ for } \theta = \theta_0), \]

\[ \frac{d^2}{d\theta^2} P_{m\theta} = -2f(\theta). \]

Hence Theorem 5.4 yields that

\[ \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1/(4f^2(\theta_0))). \]

**Example 5.4.** Suppose we observe i.i.d. samples \((X_1, Y_1), \ldots, (X_n, Y_n)\) from the regression model

\[ Y = f_{\theta_0}(X) + \xi, \]

where \(E[\xi|X] = 0\) and \(\{f_{\theta} : \theta \in \Theta\}\) is a family of regression functions. Let \(\hat{\theta}_n\) be the least squares estimator defined by

\[ \hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^n (Y_i - f_{\theta}(X_i))^2. \]

To fit the M-framework, we may take \(m_{\theta}(x,y) = -(y - f_{\theta}(x))^2\), and we expect that under identifiability condition on \(\{f_{\theta} : \theta \in \Theta\}\),

\[ \hat{\theta}_n \to \arg\max_{\theta} P_{m\theta} = \arg\min_{\theta} P(f_{\theta} - f_{\theta_0})^2 = \theta_0. \]

On the other hand,

\[ P_{m\theta} \approx P((\theta - \theta_0)^\top \hat{f}_{\theta_0})^2 + \mathbb{E}\xi^2, \]

so we may expect \(V_{\theta_0} = 2P\hat{f}_{\theta_0}\hat{f}_{\theta_0}^\top\). Since \(m_{\theta_0}(x,y) = -2(y - f_{\theta_0}(x))\hat{f}_{\theta_0}(x)\), we expect the asymptotic variance for \(\sqrt{n}(\hat{\theta}_n - \theta_0)\) to be given by

\[ (2P\hat{f}_{\theta_0}\hat{f}_{\theta_0}^\top)^{-1}P\xi^2\hat{f}_{\theta_0}\hat{f}_{\theta_0}^\top \cdot (2P\hat{f}_{\theta_0}\hat{f}_{\theta_0}^\top)^{-1} = \frac{1}{4}P\xi^2 \cdot (P\hat{f}_{\theta_0}\hat{f}_{\theta_0}^\top)^{-1} \text{ if } X \text{ is independent of } \xi. \]

The distribution convergence follows by additional complexity condition on the class \(\{f_{\theta} : \theta \in \Theta\}\) but we shall not go into technical details for this.
6. Delta method

6.1. Delta method. Suppose $\hat{\theta}_n$ is an estimator of $\theta_0$, and we are interested in the quantity $\phi(\theta_0)$. A natural plug-in estimator is given by $\phi(\hat{\theta}_n)$, but the question is, how does $\phi(\hat{\theta}_n)$ perform as an estimator of $\theta_0$?

The following theorem, known as the delta method, quantifies the distributional performance of the estimator $\phi(\hat{\theta}_n)$ in terms of the performance of $\hat{\theta}_n$ itself.

**Theorem 6.1.** Let $\phi : \mathbb{R}^k \to \mathbb{R}^m$ be differentiable at $\theta_0$ in the sense that there exists $\phi'_{\theta_0} \in \mathbb{R}^{m \times k}$ such that

$$\phi(\theta_0 + h) = \phi(\theta_0) + \phi'_{\theta_0} \cdot h + o(\|h\|), \quad \text{as } \|h\| \to 0.$$ 

If $r_n(\hat{\theta}_n - \theta_0) \to_d Z$ for some $r_n \to \infty$ and some random vector $Z$, then

$$r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) \to_d \phi'_{\theta_0} \cdot Z.$$

**Proof.** The proof is basically Taylor expansion, but we include here some details. Note that

$$r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) = r_n\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) + r_n o(\|\hat{\theta}_n - \theta_0\|)$$

$$= \phi'_{\theta_0}(r_n(\hat{\theta}_n - \theta_0)) + o_p(1)$$

$$\to_d \phi'_{\theta_0} \cdot Z,$$

as desired. \qed

**Example 6.1.** Let $Y_1, \ldots, Y_n$ be i.i.d. samples. Let

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

be the sample variance (modulo the fact that we use the weight $n$ rather than the usual $n - 1$, but this does not matter in asymptotics). Let $\phi(x,y) = y - x^2$. Then we may write $S_n^2 = \phi(\bar{Y}, \bar{Y}^2)$. This helps us to derive limit distribution for $S_n$. To this end, let $\alpha_k = EY_1^k$. By central limit theorem, we have

$$\sqrt{n} \left( \begin{pmatrix} \bar{Y} \\ \bar{Y}^2 \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \right) \to_d N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1 \alpha_2 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \right).$$

It is easy to find the derivative of the map $\phi$, which is given by $\phi'_{(\alpha_1, \alpha_2)} = (-2\alpha_1, 1)^\top$. Then by delta method, the asymptotic variance for $\sqrt{n}(\phi(\bar{Y}, \bar{Y}^2) - \phi(\alpha_1, \alpha_2))$ is given by

$$\begin{pmatrix} -2\alpha_1 & 1 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \begin{pmatrix} -2\alpha_1 \\ 1 \end{pmatrix}$$

$$= \alpha_4 - \alpha_2^2 \text{ if } \alpha_1 = 0.$$

Hence by shift invariance, with $\mu_k = E(Y_1 - EY_1)^k$,

$$\sqrt{n}(S_n^2 - \mu_2) \to_d N(0, \mu_4 - \mu_2^2).$$
6.2. Variance stabilizing. Asymptotic results like the ones we discussed above are crucial for building confidence intervals or testing hypotheses about the parameters (to be discussed later), but typically suffer from the following problem: the asymptotic variance is a function of the unknown parameter in the sense that

\[ \sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, \sigma^2(\theta_0)). \]

Therefore, we either need to use a naive plug-in approach (i.e., substituting MLEs into the variance expressions) or use a variance stabilizing approach: transform the parameter such that the asymptotic variance is free of the unknown parameter (i.e., pivotal). The basic idea for finding such a transform is as follows: by delta method,

\[ \sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta_0)) \rightarrow_d N(0, (\phi'(\theta_0))^2 \sigma^2(\theta_0)). \]

So we may take

\[ \phi(\theta) = \int 1 \sigma(\theta) \, d\theta, \]

so that the asymptotic distribution of \( \sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta_0)) \) is pivotal.

**Example 6.2.** Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be i.i.d. bivariate normally distributed with correlation \(\rho\). Then the sample correlation

\[ r_n = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \]

satisfies

\[ \sqrt{n}(r_n - \rho) \rightarrow_d N(0, (1 - \rho^2)^2). \]

The variance of the limiting normal distribution unfortunately depends on the unknown correlation \(\rho\) so we cannot directly use \(r_n\) to conduct inference on \(\rho\). Following the variance stabilizing idea above, we may consider the following transform:

\[ \phi(\rho) = \int \frac{1}{1 - \rho^2} \, d\rho = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} = \text{arctanh} \rho. \]

6.3. Asymptotic normality of moment estimators. Consider the following general form of moment estimators:

\[ \mathbb{E}_{\hat{\theta}_n} f_j(Y) = \frac{1}{n} \sum_{i=1}^{n} f_j(Y_i), \quad 1 \leq j \leq k. \]

The classical method of moment estimator we have seen in Section 4.1 corresponds to the case \(f_j(y) = y^j\).

For notational simplicity, we write \(f = (f_1, \ldots, f_k)\).
Theorem 6.2. Suppose that the map $e(\theta) = P_0 f$ is one-to-one on an open set $\Theta \subset \mathbb{R}^k$, and is continuously differentiable at $\theta_0$ with derivative matrix $e'_0$. Further assume that $P_0 \|f\|^2 < \infty$. Then $\hat{\theta}_n$ exists with probability tending to one, and
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, (e'_0)^{-1}P_0 f f^T(e'_0)^{-\top}).
\]

Proof. By inverse function theorem, there exists an open set $U \subset \mathbb{R}^k$ containing $\theta_0$, and an open set $V \subset \mathbb{R}^k$ containing $P_0 f$ such that $e : U \rightarrow V$ is $C^1$ homeomorphism. So $\hat{\theta}_n = e^{-1}(P_n f)$ as soon as $P_n f \in V$, which happens with probability tending to 1 according to the weak law of large numbers. The claim of the theorem now follows directly from the delta method.

7. Hypothesis Testing

A hypothesis is a statement or assertion about a parameter (scalar) or vector. Typically we test a “null hypothesis” denoted as $H_0$ versus an “alternative hypothesis” $H_1$. The nature and complexity of these hypotheses depend on the type of question asked. Consider the following types of hypotheses for a scalar parameter $\theta$:

1. $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ (point or simple null versus point or simple alternative).
2. $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ (point null versus composite, one-sided alternative).
3. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (point null versus composite, two-sided alternative).
4. $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ (composite null versus composite alternative).
5. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \mathbb{R}$ (point null nested within a composite alternative).

The basic idea of hypothesis testing is to (i) partition the sample space $\Omega$ into two disjoint subsets – the rejection region ($R$) and the acceptance region ($A$), such that $R \cap A = \phi$ and $R \cup A = \Omega$, and (ii) use a decision rule that leads to rejection of the null hypothesis if the observed data $Y \in R$. Thus, there are two key aspects associated with “constructing” a test:

(i) Instead of creating a decision rule based on the entire data $Y$, it makes sense to create one based on a statistic $T(Y)$, which is possibly a sufficient statistic.

(ii) We need to create a decision rule based on $T(Y)$.

In the rest of this section, we will denote the parameter space specified by $H_0$ by $\Theta_0$ and that specified by $H_1$ by $\Theta_1$.

Example 7.1. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, 1)$. To test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, one can use the test statistic $T(Y) = \bar{Y}$ and the following decision rule: reject $H_0$ if $\sqrt{n}(\bar{Y} - \mu_0) > 2$. 
Clearly, one can have several choices for $T(Y)$ and the rejection region $R$. How does one judge the “goodness” of a test? Typically, in hypothesis testing literature, two types of errors are considered: Type-I error (rejecting $H_0$ when it is actually true) and Type-II error (accepting $H_0$ when it is actually false). The hypothesis testing procedure should ensure that the probabilities of these two types of errors are not large. Instead of the probability of Type-II error, one can also consider the power of the test, defined as the probability of rejecting the null when it is not true. The power is a function of $\theta$ unless $H_1$ is a point hypothesis, and is often denoted by $\beta(\theta) = P_\theta\{T(Y) \in R|\theta \in \Theta_1\}$.

A typical approach to hypothesis testing is to find a rejection region that guarantees that the probability of type-I error does not exceed a certain threshold $\alpha$, called level of the test. Mathematically, $P_\theta\{T(Y) \in R|\theta \in \Theta_0\} \leq \alpha$ for all $\theta \in \Theta_0$, which means $\sup_{\theta \in \Theta_0} P_\theta\{T(Y) \in R|\theta \in \Theta_0\} \leq \alpha$. If the probability of type-I error attains the exact value $\alpha$, it is called size of the test. Next, one attempts to find a test that has maximum power among all level $\alpha$ tests for all $\theta \in \Theta_1$. If such a test exists, it is called a uniformly most powerful (UMP) test of level-$\alpha$. A well-known tool used to construct MP and UMP tests is the Neyman-Pearson lemma described below.

**Theorem 7.1** (Neyman-Pearson lemma). Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with pdf $f_\theta(y)$. Consider the problem of testing the point null hypothesis $H_0 : \theta = \theta_0$ versus an alternative hypothesis $H_1 : \theta = \theta_1$. The rejection region

$$R = \{Y : \frac{f_{\theta_1}(Y)}{f_{\theta_0}(Y)} \geq c\}$$

is the most powerful level $\alpha$ test if $c$ satisfies the size-$\alpha$ condition, i.e.,

$$P_{\theta_0}(Y \in R) = \alpha.$$  

**Proof.** Consider the following diagrammatic representation of the sample space

![Diagram](image)

where $R_1 \cup R_2$ is the optimal Neyman-Pearson rejection region and $R_2 \cup R_3$ is the rejection region for any other level-$\alpha$ test. Then $R_1 \cup R_2 = \{Y : \frac{f_{\theta_1}(Y)}{f_{\theta_0}(Y)} \geq c\}$.
Since \( R_1 \cup R_2 \) satisfies the size-\( \alpha \) condition and \( R_1 \cup R_2 \) satisfies the level-\( \alpha \) condition, we must have 
\[
P_{\theta_0}(R_1 \cup R_2) = \alpha \geq P_{\theta_0}(R_2 \cup R_3),
\]
which implies that
\[
\int_{R_1 \cup R_2} f_{\theta_0}(y)dy \geq \int_{R_2 \cup R_3} f_{\theta_0}(y)dy,
\]
which, in turn, implies
\[
(7.1) \quad \int_{R_3} f_{\theta_0}(y)dy \geq \int_{R_3} f_{\theta_0}(y)dy,
\]
We need to prove that \( P_{\theta_1}(Y \in R_1 \cup R_2) \geq P_{\theta_1}(Y \in R_2 \cup R_3) \), for which it suffices to show that \( P_{\theta_1}(Y \in R_1) \geq P_{\theta_1}(Y \in R_3) \). To that effect,
\[
P_{\theta_1}(Y \in R_1) = \int_{R_1} \left[ \frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} \right] f_{\theta_0}(y)dy \geq c \int_{R_3} f_{\theta_0}(y)dy, \text{ by the condition of the N-P rejection region}
\]
\[
\geq c \int_{R_3} f_{\theta_0}(y)dy, \text{ by (7.1)}
\]
\[
= c \int_{R_3} \left[ \frac{f_{\theta_0}(y)}{f_{\theta_1}(y)} \right] f_{\theta_1}(y)dy \geq \int_{R_1} f_{\theta_1}(y)dy
\]
where the last inequality follows since \( \frac{f_{\theta_0}(y)}{f_{\theta_1}(y)} \geq 1/c \) for all \( y \in R_3 \), the region not contained in the optimal N-P rejection region. But \( \int_{R_3} f_{\theta_1}(y)dy = P_{\theta_1}(R_3) \). Thus the result is established.

**Example 7.2.** Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, 1) \). Consider testing \( H_0^{(1)} : \mu = \mu_0 \) versus \( H_1^{(1)} : \mu = \mu_1 (> \mu_0) \). Then,
\[
\frac{f_{\mu_1}(Y)}{f_{\mu_0}(Y)} = \exp \left\{ n\bar{Y}(\mu_1 - \mu_0) - (n/2)(\mu_1^2 - \mu_0^2) \right\},
\]
after some algebra. Thus \( \frac{f_{\mu_1}(Y)}{f_{\mu_0}(Y)} \geq c \) implies that \( \bar{Y} \geq c \), since \( \mu_1 > \mu_0 \). Thus the rejection region is \( R = \{ Y : \bar{Y} \geq c \} \), where \( c \) must satisfy \( P_{\mu_0} \{ \bar{Y} \geq c \} = \alpha \) and hence should be equal to \( \mu_0 + z_\alpha/\sqrt{n} \).

**7.1. Extending MP tests for simple hypotheses to UMP tests for composite hypothesis.** In Example 7.2 we saw that for the iid data from \( N(\mu, 1) \), the MP test (rejection rule) for testing \( H_0^{(1)} : \mu = \mu_0 \) versus \( H_1^{(1)} : \mu = \mu_1 (> \mu_0) \) is
\[
(7.2) \quad R_1 : \bar{Y} > \mu_0 + z_\alpha/\sqrt{n}.
\]
Note that the rejection rule \( R_1 \) in (7.2) also gives a UMP test of level \( \alpha \) for testing \( H_0^{(1)} \) versus the composite alternative \( H_1^{(2)} : \mu > \mu_0 \), because it does not depend on \( \mu_1 \). Using a similar argument, a UMP rejection rule of level \( \alpha \) for testing \( H_0^{(1)} \) versus the composite alternative \( H_1^{(3)} : \mu < \mu_0 \) will be:

\[
(7.3) \quad R_2 : \bar{Y} < \mu_0 - z_\alpha / \sqrt{n}.
\]

However, note that a UMP test for testing \( H_0^{(1)} \) versus the two-sided composite alternative \( H_1^{(4)} : \mu \neq \mu_0 \) does not exist.

\section{7.1.}
Provide an argument to justify the above statement.

Now suppose, we wish to test the composite null \( H_0^{(2)} : \mu \leq \mu_0 \) versus the composite alternative \( H_1^{(2)} : \mu > \mu_0 \). We will argue that the rejection rule \( R_1 \), given by (7.2), which is UMP for testing \( H_0^{(1)} \) versus \( H_1^{(2)} \) is also UMP for testing \( H_0^{(2)} \) versus \( H_1^{(2)} \). To show this, define \( \beta_1(\mu) \) as the power function of the test defined by \( R_1 \). Then we need to establish the following:

(a) \( \beta_1(\mu) \leq \alpha \ \forall \ \mu \leq \mu_0 \), which means \( R_1 \) is a level-\( \alpha \) test for \( H_0^{(2)} \).

(b) Let \( \tilde{R} \) be any other rejection region with power function \( \tilde{\beta}(\mu) \) that satisfies

\[
(7.4) \quad \tilde{\beta}(\mu) \leq \alpha \ \forall \ \mu \leq \mu_0,
\]

which means \( \tilde{R} \) is also a level-\( \alpha \) test for \( H_0^{(2)} \). Then,

\[
(7.5) \quad \beta_1(\mu_1) \geq \tilde{\beta}(\mu_1) \text{ for any arbitrary } \mu_1 > \mu_0.
\]

To establish (a), note that

\[
\beta_1(\mu) = P_\mu \{ \sqrt{n}(\bar{Y} - \mu) > \sqrt{n}(\mu_0 - \mu + z_\alpha / \sqrt{n}) \} \\
= P_\mu \{ Z \leq \sqrt{n}(\mu - \mu_0) - z_\alpha / \sqrt{n} \}, \quad \text{where } Z \sim N(0,1) \\
= \Phi \left( \sqrt{n}(\mu - \mu_0) - z_\alpha / \sqrt{n} \right),
\]

where \( \Phi(\cdot) \) is the CDF of a standard normal variable. Thus \( \beta_1(\mu) \) is a non-decreasing function of \( \mu \). Thus, for all \( \mu \leq \mu_0 \), \( \beta_1(\mu) \leq \beta_1(\mu_0) = \alpha \) and (a) is satisfied.

To establish (b), note that we have already established (through the NP lemma) that \( R_1 \) is the UMP test for testing \( H_0^{(1)} \) versus \( H_1^{(2)} \). This means, for any arbitrary \( \mu_1 > \mu_0 \),

\[
\beta_1(\mu_1) \geq \beta^*(\mu_1),
\]

where \( \beta^*(\mu) \) is the power function of any test whose power function also satisfies \( \beta^*(\mu_0) \leq \alpha \). But by (7.4), \( \tilde{\beta} \) satisfies this condition. Therefore for any test \( \tilde{R} \) whose power function satisfies (7.4), (7.5) must hold.

This example for testing hypotheses on the mean of the normal distribution can be generalized by (a) introducing a monotonicity condition on the likelihood ratio, and (b) applying the Karlin-Rubin Theorem (Theorem 7.2) stated later.
Definition 7.1 (Monotone Likelihood Ratio). The family of distributions $f_\theta(y)$, $\theta \in \Theta$, has **monotone likelihood ratio** (MLR) in a statistic $T(Y)$ if the ratio $f_{\theta_2}(Y)/f_{\theta_1}(Y)$ can be expressed as a function of $(T(Y), \theta_1, \theta_2)$ and for each $\theta_1 < \theta_2$, the ratio is non-decreasing in $T(Y)$ when at least one of the numerator and denominator is positive.

Example 7.3. Consider the exponential family with pdf

$$f_\theta(Y) = \exp\{T(Y)\eta(\theta) - \psi(\eta)\} h(Y).$$

For $\theta_1 < \theta_2$, the log of the likelihood ratio is

$$\lambda = T(Y)\{\eta(\theta_2) - \eta(\theta_1)\} - \{\psi(\eta)|_{\theta_2} - \psi(\eta)|_{\theta_1}\},$$

which implies $\partial \lambda / \partial T = \eta(\theta_2) - \eta(\theta_1) \geq 0$ if $\eta(\theta)$ is a non-decreasing function of $\theta$.

Theorem 7.2 (Karlin-Rubin). Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Let $T$ be a sufficient statistic for $\theta$, and suppose the pdf $f_\theta(Y)$ has MLR in $T$. Then, for any $t_0$, the test with rejection region $T > t_0$ is UMP level-$\alpha$, where $\alpha = P_{\theta_0}(T > t_0)$.

The proof of the Karlin-Rubin Theorem can be found in Chapter 8 of Casella and Berger (2002), and is similar to our argument for the normal example. However, a key element of the proof involves showing that the MLR condition leads to a non-decreasing power function.

Example 7.4. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(p)$. Consider testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$. First consider testing $H_0 : p = p_0$ versus $H_1 : p = p_1$ where $p_1 > p_0$. By the NP lemma, the MP test (rejection rule) is:

$$\frac{f_{p_1}(Y)}{f_{p_0}(Y)} > \text{constant},$$

where the constant is to be determined from the size-$\alpha$ condition. This implies

$$\left(\frac{p_1}{p_0}\right)^T \left(\frac{1-p_0}{1-p_1}\right)^T > \text{constant},$$

where $T = \sum_{i=1}^n Y_i$. The function on the LHS is increasing in $T$ for any $p_1 > p_0$, and hence the family has an MLR in $T$. Thus, the MP rejection region is $T > c$, where $c$ is to be determined from the level-$\alpha$ condition, i.e.,

$$P_{p_0}(T > c) = P_{p_0}(T \geq c + 1) = \alpha,$$

which implies that

$$\sum_{j=c+1}^n \binom{n}{j} p_0^j (1-p_0)^{n-j} = \alpha.$$  

For fixed $\alpha$, there may not exist a $c$ that satisfies the above condition. However, it is possible to find a $c$ that satisfies

$$\sum_{j=c+1}^n \binom{n}{j} p_0^j (1-p_0)^{n-j} < \alpha < \sum_{j=c}^n \binom{n}{j} p_0^j (1-p_0)^{n-j}.$$
Then, by the Karlin-Rubin theorem, the test will be UMP of level

\[ \alpha^* = \sum_{j=c+1}^{n} \binom{n}{j} p_j^2 (1 - p_0)^j. \]

### 7.2. Likelihood Ratio Test

Consider testing \( H_0 : \theta \in \Theta_0 \) versus \( H_1 : \theta \in \Theta_1 \), where \( \Theta_0 \cup \Theta_1 = \Theta \). Define the likelihood ratio (LR) test statistic as:

\[ (7.6) \quad \Lambda = \sup_{\theta \in \Theta_0} \frac{L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}. \]

If the likelihood is maximized by some \( \theta^* \in \Theta_0 \), then \( \Lambda \) equals one; otherwise it is smaller than 1. The smaller the ratio, the more likely it is that the likelihood is maximized by some \( \theta \) in the rejection region. Thus, the LR test involves rejection of \( H_0 \) if \( \Lambda < c \), where \( c \) is a threshold that can be determined from size conditions.

It is also a common practice to define the LR statistic as the inverse of (7.6), i.e.,

\[ \sup_{\theta \in \Theta} L(\theta) \]

\[ \sup_{\theta \in \Theta_0} L(\theta), \]

in which case higher values of the ratio lead to rejection of \( H_0 \).

**Example 7.5** (A two-sample test). Let \( X_1, \ldots, X_n \stackrel{iid}{\sim} \text{expo}(\theta) \) with pdf \( f_{\theta}(x) = \theta^{-1} \exp(-x/\theta), \ x \geq 0, \ \theta > 0 \); and \( Y_1, \ldots, Y_m \stackrel{iid}{\sim} \text{expo}(\mu) \) with pdf \( f_{\mu}(y) = \mu^{-1} \exp(-y/\mu), \ y \geq 0, \ \mu > 0 \).

Consider testing \( H_0 : \theta = \mu \) versus \( H_1 : \theta \neq \mu \). The log likelihood function is given by

\[ l(\theta, \mu) = -n \log \theta - \frac{\sum_{i=1}^{n} X_i}{\theta} - m \log \mu - \frac{\sum_{i=1}^{m} Y_i}{\mu}. \]

Maximizing the above log likelihood with respect to \( \theta \) and \( \mu \), their unrestricted MLEs are obtained as \( \bar{X} \) and \( \bar{Y} \) respectively. Now the log likelihood under \( H_0 \) is

\[ l(\theta) = -(n + m) \log \theta - \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}{\theta}, \]

maximizing which, we obtain the MLE of \( \theta \) under \( H_0 \) as

\[ \hat{\theta}_0 = \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}{n + m}. \]

Using definition (7.6), the LR statistic is \( L(\hat{\theta}_0)/L(\hat{\theta}, \hat{\mu}) \), where

\[ L(\hat{\theta}_0) = \frac{1}{\hat{\theta}_0^{n+m}} \exp \{ -(n + m) \}. \]
and
\[
L(\hat{\theta}, \hat{\mu}) = \frac{1}{\hat{\theta}^n} \exp \{-n\} \times \frac{1}{\hat{\mu}^m} \exp \{-m\}
\]
\[
= \frac{1}{\hat{\theta}^n \hat{\mu}^m} \exp \{-(n + m)\}.
\]
Consequently, the LR statistic is
\[
\Lambda = \frac{\hat{\theta}^n \hat{\mu}^m}{\hat{\theta}_0^{n+m}}
\]
\[
= \frac{(n + m)^{n+m}}{n^n m^m} T^n (1 - T)^m,
\]
where
\[
T = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i + \sum_{i=1}^m Y_i}.
\]

The logarithm of the LR statistic is
\[
\log \Lambda = \text{const} + n \log T + m \log(1 - T),
\]
which is a concave function of \(T\). Hence the rejection rule \(\Lambda < c\) is equivalent to the rejection rule \(T \leq a\) and \(T \geq b\) where \(a\) and \(b\) are constants that can be chosen in a manner that satisfies the size-\(\alpha\) condition (how?)

### 7.3. Asymptotic tests
We now discuss some asymptotic procedures for hypothesis testing. First, we state the following theorem on the asymptotic distribution of the LR statistic for a point null hypothesis about a scalar parameter

**Theorem 7.3** (LRT asymptotics one dimension). Consider testing the point null hypothesis \(H_0 : \theta = \theta_0\) versus the two-sided alternative \(H_0 : \theta \neq \theta_0\). Let

\[
\Lambda_n = \frac{L(\theta_0)}{\sup_{\theta \in \Theta} L_n(\theta)}
\]

denote the LR statistic for an iid sample of size \(n\) obtained from a distribution with density \(f_\theta(y)\), where \(\theta \in \Theta\). Then under the same regularity conditions as those assumed while deriving the asymptotic distribution of the MLE (Theorem 4.1),

\[
-2 \log \Lambda_n \xrightarrow{d} \chi_1^2, \quad \text{under } H_0.
\]

**Proof.** First, note that

\[
-2 \log \Lambda_n = -2 \left\{ l_n(\theta_0) - l_n(\hat{\theta}_n) \right\},
\]

where \(l_n(\theta) = \log L_n(\theta)\) and \(\hat{\theta}_n\) is the MLE of \(\theta\).

Now, expand \(l_n(\theta_0)\) around \(\hat{\theta}_n\) using Taylor’s theorem to obtain:

\[
l_n(\theta_0) \approx l_n(\hat{\theta}_n) + (\theta_0 - \hat{\theta}_n) l'_n(\hat{\theta}_n) + \frac{1}{2} (\theta_0 - \hat{\theta}_n)^2 l''_n(\hat{\theta}_n)
\]
\[ l_n(\hat{\theta}_n) + (\theta_0 - \hat{\theta}_n)^2 l''_n(\hat{\theta}_n), \]
since \( l'_n(\hat{\theta}_n) = 0 \). Thus, we have,

\[
-2 \left\{ l_n(\theta_0) - l_n(\hat{\theta}_n) \right\} = -(\theta_0 - \hat{\theta}_n)^2 l''_n(\hat{\theta}_n) \]

(7.7)

Now, under \( H_0 \), by Theorem 4.1, \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} N \left( 0, \left\{ I_1(\theta_0) \right\}^{-1} \right) \), which implies that

\[
\frac{n(\hat{\theta}_n - \theta_0)^2}{\left\{ I_1(\theta_0) \right\}^{-1}} \overset{d}{\rightarrow} \chi^2_1,
\]
or alternatively,

(7.8) \[ n(\hat{\theta}_n - \theta_0)^2 \overset{d}{\rightarrow} \left\{ I_1(\theta_0) \right\}^{-1} \chi^2_1. \]

Also, under \( H_0 \), by LLN,

(7.9) \[ \hat{\theta}_n \overset{P}{\rightarrow} I_1(\theta_0). \]

Substituting (7.8) and (7.9) in (7.7), and applying Slutsky’s theorem, the result immediately follows. □

**Remark 7.1.** Would the asymptotic distribution of the LR statistic as stated in Theorem 7.3 still hold if instead of the two-sided alternative, the one-sided alternative \( H_1 : \theta > \theta_0 \) was considered? The answer is NO. Consider the following counter example: let \( Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, 1) \). Suppose we are interested in testing \( H_0 : \mu = 0 \) versus \( H_1 : \mu > 0 \). Then,

\[
\sup_{\mu \in \Theta} L_n(\mu) = L_n \left\{ \max(\bar{Y}, 0) \right\},
\]
and therefore

\[
\Lambda_n = \frac{L_n(0)}{L_n \left\{ \max(\bar{Y}, 0) \right\}}.
\]

However, under \( H_0 \), \( P_\mu(\bar{Y} > 0) = P_{\mu=0}(\bar{Y} > 0) = 1/2 \). Thus, under \( H_0 \), for all \( n \), \( \Lambda_n \) equals one with probability 1/2. This means that under \( H_0 \), \( -2 \log \Lambda_n \) has a point mass of 0.5 at zero for all \( n \), and hence cannot converge to \( \chi^2_1 \). The result in Theorem 7.3 is generalized through the Wilks’ Theorem stated next.

**Theorem 7.4 (Wilks’ Theorem).** Suppose we are testing \( H_0 : \theta \in \Theta_0 \) vs. \( H_1 : \theta \in \Theta_1 \). Suppose also that \( \Theta_0 \) is contained in the interior of the closure of \( \Theta_1 \), where both are subspaces of some \( \mathbb{R}^m \).

Assume also that our model has enough regularity to be able to DUThIS and for asymptotics to work; for example, the support should not depend on \( \theta \).
Let
\[ \Lambda_n = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_1} L(\theta)} \]

Then,
\[ -2 \log \Lambda_n \overset{d}{\rightarrow} \chi^2_{\text{dim}(\Theta) - \text{dim}(\Theta_0)} \]

under the null hypothesis, where \( \text{dim}(\Theta_0) \) and \( \text{dim}(\Theta_1) \) denote the dimensions of \( \Theta_0 \) and \( \Theta_1 \) respectively, and \( \Theta = \Theta_0 \cup \Theta_1 \).

Remark 7.2. We now see from Wilks’ theorem why the asymptotic distribution of \( -2 \log \Lambda_n \) was not \( \chi^2_1 \) in the example of remark 7.1. While testing the hypothesis \( H_0 : \theta = \theta_0 \) versus the one-sided alternative \( H_0 : \theta > \theta_0 \), the closure (union of a set and its boundary points) of \( \Theta_1 = \{ \theta : \theta > \theta_0 \} \) is the set \( \{ \theta : \theta \geq \theta_0 \} \). The set \( \Theta_0 = \{ \theta : \theta = \theta_0 \} \) lies on the boundary of the closure, and not in the interior.

Two other popular asymptotic tests for testing the null hypothesis \( H_0 : \theta = \theta_0 \) versus \( H_1 : \theta \neq \theta_0 \) are the Score test and the Wald test described below:

• **Score Test (Rao Test)**
  Here we choose to use score function \( S(Y, \theta_0) \) as the test statistic. For large \( n \), \( S(Y, \theta_0) \approx N(0, nI_1(\theta_0)) \) under \( H_0 \) (why?). Thus we reject \( H_0 \) at level-\( \alpha \) if
  \[ \left| \frac{S(Y, \theta_0)}{\sqrt{nI_1(\theta_0)}} \right| > z_{\alpha/2}. \]

• **Wald Test**
  We can also directly use the maximum likelihood estimate (MLE) as the test statistic, since
  \[ \sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \overset{d}{\rightarrow} N \left( 0, I_1(\theta_0) \right) \text{ under } H_0. \]
  Thus we reject \( H_0 \) if
  \[ \left| \sqrt{nI_1(\theta_0)} \left( \hat{\theta}_n - \theta_0 \right) \right| > z_{\alpha/2}. \]

Asymptotically, all three tests are equivalent with probability 1 (under the null hypothesis), but the tests have different power. In the ideal case, the likelihood function is Normal. That is, the log likelihood is quadratic. (This does not necessarily follow from \( \hat{\theta}_n \overset{d}{\rightarrow} \text{Normal} \).) The three tests are related by the following diagram:
As labelled in the diagram, we use the:
• “width” \( w = \hat{\theta} - \theta_0 \) for the Wald test
• “length” \( l = l(\hat{\theta}) - l(\theta_0) = -\log \frac{L(\theta_0)}{L(\theta)} \) for the LRT
• “slope” \( s = l'(\theta_0) = S(Y, \theta_0) \) for the Score test

By the geometry of the parabola, \( s = 2l_w \). Thus \( 2l = sw \).

Assuming regularity conditions such that \( \hat{\theta} \to \theta_0 \), then \( sw = 2l \) gives the following (dropping the equality assumption of the ideal case):

\[
\left( \frac{1}{\sqrt{nI_1(\theta_0)S(Y, \theta_0)}} \left( \sqrt{nI_1(\theta_0)}(\hat{\theta} - \theta_0) \right) \right) \approx -2\log(LR) \xrightarrow{d} \chi^2_1.
\]

Note that the first term in parentheses converges to \( N(0, 1) \) and the second term also converges to a \( N(0, 1) \), but it is not obvious how these two Normals are related.

**Score test as the most powerful test against local alternatives**

**Theorem 7.5.** Suppose that we want to test \( H_0 : \theta = \theta_0 \), where \( \theta \) is a scalar, against the alternative \( H_1 : \theta = \theta_0 + \varepsilon \), for \( \varepsilon > 0 \) small. Then, the score test is the most powerful test.

**Proof.** By the N-P lemma, the MP test for \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta = \theta_0 + \varepsilon \) is of the form

\[ \frac{L(\theta_0 + \varepsilon)}{L(\theta_0)} > k. \]

Taking log of both sides yields

\[ l(\theta_0 + \varepsilon) - l(\theta_0) > \log k. \]
For small $\varepsilon$, the LHS of the above can be well approximated by replacing $l(\theta_0 + \varepsilon)$ by its first order Taylor expansion around $\varepsilon$, i.e., by $\varepsilon l'(\theta_0)$, which equals $\varepsilon S(Y, \theta_0)$. Thus, the rejection region based on the score function is the MP test. □

**LR, Score and Wald tests when $\theta$ is a vector of parameters**

(a) For the LR test, we have

$$-2 \log \Lambda_n \to \chi^2_{\text{dim}(\Theta) - \text{dim}(\Theta_0)}.$$

(b) For the Wald test, we have

$$(\hat{\theta} - \theta_0)' I(\theta_0)(\hat{\theta} - \theta_0) \to \chi^2_{\text{dim}(\Theta) - \text{dim}(\Theta_0)}.$$

(c) Finally, for the score test, we have as our test statistic

$$[\nabla l(Y, \theta_0)'] I^{-1}(\theta_0) [\nabla l(Y, \theta_0)] \to \chi^2_{\text{dim}(\Theta) - \text{dim}(\Theta_0)}.$$

8. **Interval Estimation**

**Definition 8.1** (Interval estimator). Let $L(Y)$ and $U(Y)$ be a pair of functions of the data $Y$ satisfying $L(Y) \leq U(Y)$ for all $Y$. The random interval $[L(Y), U(Y)]$ is called an interval estimator of a parameter $\theta$ if upon observing $Y = y$, the inference $L(y) \leq \theta \leq U(y)$ is made for parameter $\theta$.

The standard frequentist criterion for the evaluation of interval estimators is **coverage**. Consider an interval $[L(Y), U(Y)]$ (a statistic). The **coverage probability** of the interval is $P_{\theta}(L(Y) \leq \theta \leq U(Y))$, where this probability is taken over possible values of the data and with respect to a single fixed value of $\theta$. An interval $[L(Y), U(Y)]$ is said to be a $(1 - \alpha)$% **confidence interval** if it has a coverage probability of at least $(1 - \alpha)$ for all possible values of $\theta$. Thus, if we have an interval that covers with probability 0.95 for some values of $\theta$ and with probability 0.99 for other values, it is a 95% confidence interval. Below, we formalize these definitions.

**Definition 8.2** (Coverage probability). The probability $P_{\theta} [L(Y) \leq \theta \leq U(Y)]$, i.e., the probability that the random interval $[L(Y), U(Y)]$ contains the true value of $\theta$, is called the coverage probability. The coverage probability may depend on $\theta$.

**Definition 8.3** (Confidence coefficient). The probability

$$1 - \alpha = \inf_{\theta \in \Theta} P_{\theta} [L(Y) \leq \theta \leq U(Y)]$$

is known as the confidence coefficient. If the coverage probability does not depend on $\theta$, it is the confidence coefficient. However, if it does, one can only guarantee a coverage probability equal to the infimum, since the true value of $\theta$ is unknown.

**Definition 8.4** (Confidence Interval). An interval estimator with a confidence coefficient $1 - \alpha$ is referred to as a confidence interval (CI) with confidence coefficient $1 - \alpha$, or sometimes a $100(1 - \alpha)$% confidence interval.
Example 8.1. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Unif}[0, \theta]$, and let $Y_{(n)} = \max_{1 \leq i \leq n} Y_i$. Consider the following two interval estimators:

(i) $[aY_{(n)}, bY_{(n)}]$, where $1 \leq a < b$.

(ii) $[c + Y_{(n)}, d + Y_{(n)}]$, where $0 \leq c < d$.

In which of these cases does the coverage probability depend on $\theta$?

Example 8.2. $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$. We are interested in estimating a 95% confidence interval for $\mu$. If $\sigma^2$ is known, then a 100$(1 - \alpha)$% CI for $\mu$ is $\bar{Y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, whereas if $\sigma^2$ is unknown, then a 100$(1 - \alpha)$% CI for $\mu$ is $\bar{Y} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$, where $s^2$ is the sample standard deviation.

8.1. Method-I for constructing interval estimators: inverting tests of hypotheses. Many people think that interval estimation is more informative than hypothesis testing. The argument for it is that interval estimation gives an interval with coverage probability whereas hypothesis testing just gives two possible results, i.e., either rejecting or accepting the null hypothesis. In another sense, there is an equivalence between them, in that we can “invert” a hypothesis testing procedure to obtain an interval estimation procedure and vice versa. However, that does not mean inverting a “good” hypothesis testing procedure gives a “good” interval estimate! See [http://andrewgelman.com/2011/08/why_it_doesnt_m/](http://andrewgelman.com/2011/08/why_it_doesnt_m/) for some discussion of this issue.

Starting with the hypothesis testing procedure, we can construct a confidence interval. For each $\theta_0$ in the parameter space, consider testing the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative $H_1 : \theta \neq \theta_0$, and let $A(\theta_0)$ denote a level-$\alpha$ acceptance rejection (i.e., accept $H_0$ when $Y \in A(\theta_0)$), such that

$$P_{\theta_0}(Y \in A(\theta_0)) \geq 1 - \alpha. \tag{8.1}$$

Now, construct the set

$$C(Y) = \{ \theta : Y \in A(\theta) \}.$$ 

By construction, $P_{\theta_0}(\theta \in C(Y)) = P_{\theta_0}(Y \in A(\theta))$ for all $\theta$. Since inequality [8.1] holds for any $\theta$, we have

$$\inf_{\theta} P_{\theta}(\theta \in C(Y)) = \inf_{\theta} P_{\theta}(Y \in A(\theta)) \geq 1 - \alpha.$$ 

By definition, $C(Y)$ is a 100$(1 - \alpha)$% confidence interval for $\theta$.

Example 8.3. So how can we derive the confidence intervals of Example 8.2 by inverting a hypothesis test? Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ when $\sigma^2$ is unknown. A level $\alpha$ acceptance region is

$$A(\mu_0) = \left\{ Y : \left| \frac{\sqrt{n} (\bar{Y} - \mu_0)}{s} \right| \leq t_{n-1, \alpha/2} \right\}.$$ 

For any $\mu$, $Y \in A(\mu)$ is equivalent to $\bar{Y} - (s/\sqrt{n})t_{n-1, \alpha/2} \leq \mu \leq \bar{Y} + (s/\sqrt{n})t_{n-1, \alpha/2}$. Thus we arrive at the 100$(1 - \alpha)$% CI of Example 8.2.
Example 8.4. We now demonstrate how this method can be used to obtain CIs for more complicated problems. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Expo}(\theta)$ with pdf
\[ f_\theta(y) = \frac{1}{\theta} e^{-y/\theta}, \quad y \geq 0, \theta > 0, \]
and suppose we are interested in obtaining a $100(1 - \alpha)$% CI for $\theta$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Noting that the restricted (under $H_0$) and unrestricted MLEs of $\theta$ are $\hat{\theta}_0$ and $\bar{Y}$ respectively, the LR test statistic for this test is obtained as
\[ \Lambda = \frac{\theta_0^{-n} \exp (-T/\theta_0)}{(nT)^{-n} \exp (-n)} \propto \left( \frac{T}{\theta_0} \right)^n \exp \left( -\frac{T}{\theta_0} \right) \]
where $T = \sum_{i=1}^n Y_i$. The acceptance region is:
\[ A(\theta_0) = \left\{ Y : \left( \frac{T}{\theta_0} \right)^n \exp \left( -\frac{T}{\theta_0} \right) > k \right\}, \]
where $k$ is a constant to be determined from the size-$\alpha$ condition. Note that
\[ \left( \frac{T}{\theta_0} \right)^n \exp \left( -\frac{T}{\theta_0} \right) > k \Rightarrow k_1 \leq \frac{2T}{\theta_0} \leq k_2, \]
and thus the acceptance region can be written as
\[ A(\theta_0) = \left\{ Y : k_1 \leq \frac{2T}{\theta_0} \leq k_2 \right\}, \]
where $k_1$ and $k_2$ are constants to be determined from the size-$\alpha$ condition, i.e.,
\[ P_{\theta_0} \left\{ k_1 \leq \frac{2T}{\theta_0} \leq k_2 \right\} = 1 - \alpha. \]
Noting that $2T/\theta_0 \sim \text{Gamma}(n, 2) \equiv \chi^2_{2n}$, under $H_0$, $k_1$ and $k_2$ can be chosen as $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ respectively, leading to the final acceptance region as:
\[ A(\theta_0) = \left\{ Y : \chi^2_{2n,1-\alpha/2} \leq \frac{2T}{\theta_0} \leq \chi^2_{2n,\alpha/2} \right\}. \]
The interval obtained by inverting this test is thus
\[ C(Y) = \left\{ \theta : \chi^2_{2n,1-\alpha/2} \leq \frac{2T}{\theta} \leq \chi^2_{2n,\alpha/2} \right\}, \]
which yields the $100(1 - \alpha)$% interval
\[ \left[ \frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right]. \]

**Definition 8.5** (Pivotal quantity or Pivot). A pivotal quantity, or simply a “pivot” is a function $Q(Y, \theta)$ whose distribution does not depend on $\theta$. Note that a pivot is not a statistic as it involves $\theta$, and thus should not be confused with an ancillary statistic.

To find a $100(1 - \alpha)$% CI for $\theta$ using the pivot $Q(Y, \theta)$, find a set $\mathcal{A}$ that satisfies

$$P_\theta \{ Q(Y, \theta) \in \mathcal{A} \} \geq 1 - \alpha.$$ 

Thus the set

$$\{ \theta : Q(Y, \theta) \in \mathcal{A} \}$$

is a $100(1 - \alpha)$% CI for $\theta$, provided $Q(Y, \theta)$ is a monotone function of $\theta$ for fixed $Y$ (why is monotonicity needed?)

**Example 8.5.** In example 8.2 $Q(Y, \mu) \equiv \frac{(\bar{Y} - \mu)\sqrt{n}}{\sigma} \sim \mathcal{N}(0, 1)$ is a pivot if $\sigma^2$ is known. Now we can make obtain a frequentist confidence claim using

$$P(-z_{\alpha/2} < Q(Y, \mu) < z_{\alpha/2}) = 1 - \alpha.$$ 

Solving for $\mu$ then returns the same interval as before, $\bar{Y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. If $\sigma^2$ is unknown, for inference about $\mu$ it is very useful having a pivot which does not involve $\sigma^2$, and then a very widely-used choice is

$$Q(Y, \mu) \equiv \frac{(\bar{Y} - \mu)\sqrt{n}}{s} \sim t_{n-1},$$

where $s^2$ is the (unbiased) sample variance.

**Example 8.6.** Consider example 8.4 again. Note that, for $i = 1, \ldots, n$, $Y_i/\theta$ is Expo(1), and consequently $2T/\theta \sim \Gamma(n, 2) \equiv \chi^2_{2n}$. Thus, using the distribution of the pivot, we can derive the same confidence intervals as in 8.4. In fact the procedure of inverting the LRT leads us to a pivot.

8.3. Method-III for constructing interval estimators: inverting the CDF of a sufficient statistic. Let $T = T(Y)$ be a sufficient statistic having a continuous distribution with a cdf $F_\theta(t)$. Assume that $F_\theta(t)$ is a monotonically decreasing function of $\theta$ for every $t$.

Choose $0 < \alpha_1 < 1$ and $0 < \alpha_2 < 1$, such that $\alpha_1 + \alpha_2 = \alpha$. Invert the interval $\alpha_1 \leq F_\theta(T) \leq 1 - \alpha_2$ to obtain $L(T) \leq \theta \leq U(T)$, where $F_{U(T)}(t) = \alpha_1$ and $F_{L(T)}(t) = 1 - \alpha_2$ (see Figure 1). Then $[L(T), U(T)]$ is a $100(1 - \alpha)$% CI for $\theta$ because

$$P_\theta \{ L(T) \leq \theta \leq U(T) \} = P_\theta \{ \alpha_1 \leq F_\theta(T) \leq 1 - \alpha_2 \} = 1 - \alpha_2 - \alpha_1 = 1 - \alpha.$$ 

Note that $P_\theta \{ \alpha_1 \leq F_\theta(T) \leq 1 - \alpha_2 \} = 1 - \alpha_2 - \alpha_1$ because by the probability integral transform (PIT), $F_\theta(T)$ is Unif[0, 1].

If $F_\theta(t)$ is a monotonically increasing function of $\theta$ for every $t$, then the same procedure will lead to generation of the interval $L(T) \leq \theta \leq U(T)$, where $F_{U(T)}(t) = 1 - \alpha_2$ and $F_{L(T)}(t) = \alpha_1$. 

Example 8.7. Let $Y_1, \ldots, Y_n$ be an iid random sample from a location exponential family of distributions having pdf

$$f_\mu(y) = e^{-(y-\mu)}, \; y \geq \mu.$$ 

To find a CI for $\mu$ using the method described above, we first need to find a sufficient statistic for $\mu$. It is easy to see that $T = Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ is sufficient for $\mu$. Now, we have to find the CDF of $T$. Denoting this CDF by $F_\mu(t)$, for $t \geq \mu$,

$$F_\mu(t) = P_\mu \{T \leq t\} = P_\mu \{Y_{(1)} \leq t\} = 1 - P_\mu \{Y \geq t\}^n = 1 - e^{-n(t-\mu)},$$

where the last equality follows by noting that the CDF of the data-generating distribution is $1 - e^{-(y-\mu)}, \; y \geq \mu$. Now note that for fixed $t$, $F_\mu(t)$ is a decreasing function of $\mu$. Using the construction method described above, choose $\alpha_1 = \alpha_2 = \alpha/2$, where $\alpha \in (0, 1)$, and define $L(T)$ and $U(T)$ such that $F_{U(T)}(t) = \alpha_1 = \alpha/2$ and $F_{L(T)}(t) = 1 - \alpha_2 = 1 - \alpha/2$. This yields

$$L(Y) = T + \frac{1}{n} \log(\alpha/2),$$

$$U(Y) = T + \frac{1}{n} \log(1 - \alpha/2)$$

and $[L(T), U(T)]$ is a $100(1 - \alpha)$% CI for $\mu$ by the argument earlier in this subsection.
8.4. **Shortest length confidence intervals.** Recall that in Example 8.3 we derived a 100(1 − α)% CI for the normal mean \( \mu \) when the variance \( \sigma^2 \) is unknown as \( \bar{Y} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \), using the sampling distribution of the pivot \( \sqrt{n} (\bar{Y} - \mu) / s \). However, there can be many other choices for a 100(1 − α)% CI based on the distribution of the pivot - any pair of numbers \((a, b)\) that satisfies \( Pr[a \leq \sqrt{n} (\bar{Y} - \mu) / s \leq b] = 1 - \alpha \) yields a 100(1 − α)% CI \([\bar{Y} - bs/\sqrt{n}, \bar{Y} - as/\sqrt{n}]\). The width of such an interval is \((b - a)s/\sqrt{n}\).

For example, consider the following intervals, all based on the distribution of the pivot:

- \( a = t_{n-1,0.975}, b = t_{n-1,0.025} \)
- \( a = t_{n-1,0.985}, b = t_{n-1,0.035} \)
- \( a = t_{n-1,0.965}, b = t_{n-1,0.015} \).

All of these three intervals are 95% CI, but they have different widths. The “best” among these intervals will be the one with the smallest width. Because the width of intervals of the form \([\bar{Y} - bs/\sqrt{n}, \bar{Y} - a s/\sqrt{n}]\) is proportional to \((b - a)\), we can formulate the problem of finding the interval with smallest length as the following optimization problem:

\[
\text{maximize } (b - a) \\
\text{subject to } \int_{a}^{b} f(x) dx = 1 - \alpha
\]

where \( f(x) \) is the pdf of the pivot. For a symmetric distribution like \( t \), it may be a natural choice to pick \( t_{n-1,\alpha/2} \) and \( t_{n-1,1-\alpha/2} \) as \( b \) and \( a \) respectively. But what if we have an asymmetric distribution like \( \chi^2 \) or Gamma? The following theorem (Theorem 9.3.2 of [Casella and Berger (2002)]) provides a useful tool to solve such problems.

**Theorem 8.1.** If \( f(x) \) is a unimodal pdf (of a pivot), and the interval \([a, b]\) satisfies:

(i) \( \int_{a}^{b} f(x) dx = 1 - \alpha \),
(ii) \( f(a) = f(b) > 0 \),
(iii) \( a \leq x^* \leq b \), where \( x^* \) is the mode of \( f(x) \),

then \([a, b]\) is the shortest interval satisfying (i).

The proof of the theorem can be found in [Casella and Berger (2002)], and is fairly straightforward, although a pictorial visualization makes it easier.

8.5. **Asymptotic intervals.** We now consider some asymptotic approaches to construct CIs. Analogous to the two approaches for finding CIs described in Sections 8.1 and 8.2, the following two approaches can be adopted to find asymptotic confidence intervals:

(i) Find approximate or large sample distribution of pivots and use them to construct intervals.
(ii) Invert approximate level-\( \alpha \) tests (e.g., using asymptotic distribution of the likelihood ratio statistic).
Example 8.8. Consider the problem of finding an approximate $100(1-\alpha)\%$ interval for a Poisson parameter $\lambda$ using an iid random sample $Y_1, \ldots, Y_n$. The asymptotic distribution of the MLE $\hat{\lambda} = \bar{Y}$ can be used for this purpose. Recall that for the Pois($\lambda$) distribution, $I_1(\lambda) = 1/\lambda$ and hence

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda).$$

Thus,

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \sim \mathcal{N}(0, 1),$$

where $\sim$ means “approximately distributed as.” Thus a large sample $100(1-\alpha)\%$ can be obtained as:

$$C(Y) = \left\{ \lambda : \left| \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \right| \leq z_{\alpha/2} \right\}.$$

The LHS of the inequality $\lambda : n\lambda^2 - (2n\bar{Y} + z_{\alpha/2}^2)\lambda + n\bar{Y} \leq 0$ is a convex quadratic function of $\lambda$, and hence yields a CI of the form $L(Y) \leq \lambda \leq U(Y)$.

Another asymptotic CI for $\lambda$ can be obtained by using the fact that $\hat{\lambda} \xrightarrow{p} \lambda$ by consistency of MLE and (b) Slutsky’s theorem. Thus, an approximate $100(1-\alpha)\%$ can be obtained as:

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \sim \mathcal{N}(0, 1).$$

This follows by (a) $\hat{\lambda} \xrightarrow{p} \lambda$ by consistency of MLE and (b) Slutsky’s theorem. Thus, an approximate $100(1-\alpha)\%$ can be obtained as:

$$C(Y) = \left\{ \lambda : \left| \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \right| \leq z_{\alpha/2} \right\}.$$

A third approach is to use a variance stabilizing transformation discussed at the end of Section 4.2.2. The variance stabilizing transformation for the Poisson parameter $\lambda$ is $g(\lambda) \propto \sqrt{\lambda}$ (show this!). By invariance, the MLE of $g(\lambda)$ is $g(\hat{\lambda}) = \sqrt{\bar{Y}}$. By the Delta theorem,

$$\text{Var} \left( g(\hat{\lambda}) \right) = \text{Var} \left( \sqrt{\bar{Y}} \right) \approx (g'(\lambda))^2 \text{Var}(\hat{\lambda}) = \frac{1}{4\lambda} \times \lambda = \frac{1}{4}. $$

Consequently,

$$\sqrt{n} \left( \sqrt{\bar{Y}} - \sqrt{\lambda} \right) \xrightarrow{d} \mathcal{N} \left( 0, \frac{1}{4} \right).$$
9. **Bayesian Inference**

Unlike frequentist inference, in Bayesian inference the unknown parameter \( \theta \) is viewed as a random variable. Any information available about the parameter is captured through a so-called “prior” distribution denoted by \( \pi(\theta) \). In a Bayesian setting, we will denote the probability model by \( f(y|\theta) \) (instead of the frequentist notation of \( f_\theta(y) \)), to denote that the model is conditional on the random variable \( \theta \). Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} f(y|\theta) \) be the observed data from the model. Then the likelihood function \( L(\theta|Y) \) is the joint distribution of \( f(Y|\theta) \). By application of Baye’s theorem, the posterior distribution of \( \theta \) given the data \( Y \) is

\[
\pi(\theta|Y) = \frac{L(\theta|Y)\pi(\theta)}{\int L(\theta|Y)\pi(\theta)d\theta} \propto L(\theta|Y)p(\theta).
\] (9.1)

The denominator of (9.1) is known as the normalizing constant. Inference about \( \theta \) is made on the basis of the posterior distribution \( \pi(\theta|Y) \). The data is expected to “shrink” the prior distribution, thereby leading to more precise inference from the posterior distribution. Some of the issues related to Bayesian inference include: (i) how to choose a prior distribution? (ii) how to obtain the posterior? (However, it may be noted that obtaining closed-form solution for posterior distribution is no longer a requirement due to the incredible advancement of Markov chain Monte Carlo (MCMC) methods) (iii) how to use the posterior distribution to obtain point and interval estimators of \( \theta \), and how to do hypothesis testing?

9.1. **Conjugate Priors.** Loosely speaking, a prior is conjugate to a certain statistical model if it yields the same posterior distribution as the prior. This definition can be formalized as follows:

**Definition 9.1.** Let \( G = \{g(\theta|a) : a \in A\} \) be a collection of pdfs indexed by parameter \( a \in A \). Then \( G \) is conjugate to the model \( f(y|\theta) \) if \( \pi(\theta) = g(\theta|a) \) for some \( a \in A \) means for every \( y \), \( \pi(\theta|y) = g(\theta|a') \) for some \( a' \in A \).

**Example 9.1.** Let \( Y \sim \text{Bin}(n, \theta) \) with \( n \) known. The probability model is \( f(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \). Consider the prior \( \theta \sim \text{Beta}(\alpha, \beta) \), that is

\[
\pi(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}, \quad 0 < \theta < 1, \quad \alpha > 0, \beta > 0.
\]

Then

\[
\pi(\theta|y) \propto \pi(\theta)f(y|\theta) \propto \theta^{y+\alpha-1}(1 - \theta)^{n-y+\beta-1},
\]

which is the pdf of a \( \text{Beta}(y+\alpha, n-y+\beta) \) distribution.

Conjugate priors have the following advantage:

(i) Posterior distributions are tractable and can be obtained in closed-form. This is, however, no longer a strong advantage due to the extraordinary development of MCMC methods.
(ii) Natural appeal: Prior and posterior distributions are alike, and this leads to a clear visualization of shrinkage.

(iii) Historical significance.

(iv) The mathematical theory for exponential family of distributions is elegant, leading to nice interpretations of the results.

Consider the natural exponential family (NEF) with pdf
\[ f(y|\eta) = \exp(\eta y - \psi(\eta)) h(y), \]
where \( y \) is the natural observation, and let \( Y_1, \ldots, Y_n \overset{iid}{\sim} f(y|\eta) \). Then \( T = \sum_{i=1}^{n} Y_i \) is sufficient for \( \eta \). Suppose our interest lies in inferring about the mean \( \mu = \mathbb{E}(Y|\eta) = \psi'(\eta) \). Recall that \( \text{var}(Y|\eta) = \psi''(\eta) = V(\mu) \), say. The likelihood can be written in terms of the sufficient statistic \( T \) as
\[
(9.2) \quad L(\eta|Y) \propto \exp\left\{ \eta \sum_{i=1}^{n} Y_i - n\psi(\eta) \right\} \propto \left\{ n(\eta T - \psi(\eta)) \right\}.
\]

Consider the prior pdf
\[
(9.3) \quad \pi(\eta) \propto \exp\left\{ r(\eta \mu_0 - \psi(\eta)) \right\}.
\]

From (9.2) and (9.3), it follows that
\[
(9.4) \quad \pi(\eta|Y) \propto \exp\left\{ (nT + r\mu_0) \eta - (n + r) \psi(\eta) \right\}.
\]

Note that the prior (9.3) and the posterior (9.3) are of the same form. Let us study the properties of this family of PDFs more carefully. Note that, from (9.3),
\[
\pi(\eta) \propto \exp\left\{ r \int (\mu_0 - \psi'(\eta)) \, d\eta \right\}.
\]

Let us write this PDF in terms of the mean \( \mu = \psi'(\eta) \). Because \( d\mu/d\eta = \psi''(\eta) = V(\mu) \), \( d\eta = d\mu/V(\mu) \), and the Jacobian of transformation is \( 1/V(\mu) \). Thus,
\[
\pi(\mu) \propto \exp\left\{ r \int \frac{\mu_0 - \mu}{V(\mu)} \, d\mu \right\} \frac{1}{V(\mu)} \exp\left\{ -r \int \frac{\mu - \mu_0}{V(\mu)} \, d\mu \right\},
\]
which is a location-scale family with location parameter \( \mu_0 \) (the prior mean of \( \mu \)). We may want to interpret the prior as previously observed data with mean \( \mu_0 \) and sample size \( r \).

To obtain the posterior distribution, note that from (9.4), we can write
\[
\pi(\eta|Y) \propto \exp\left[ \int \{(nT + r\mu_0) - (n + r) \psi'(\eta)\} \, d\eta \right],
\]
Transforming \( \mu = \psi'(\eta) \), we get
\[
\pi(\mu|Y) \propto \frac{1}{V(\mu)} \exp\left[ \int -n + r \frac{\mu_0}{V(\mu)} \left\{ \mu - \frac{r\mu_0 + nT}{n + r} \right\} \, d\eta \right].
\]
Hence, the posterior mean is
\[
E(\mu|Y) = \frac{r\mu_0 + nT}{n + r} = \left(\frac{r}{n + r}\right)\mu_0 + \left(\frac{n}{n + r}\right)T
\]
\[
= B\mu_0 + (1 - B)T,
\]
where
\[
B = \frac{r}{r + n}.
\]
Note that the posterior mean is a convex combination of prior means $\mu_0$ and observed mean $T = \bar{Y}$. In fact the weighting represents the relative Fisher information of the ‘estimates’ $T$ and $\mu_0$ of the posterior mean i.e.
\[
B = \frac{r\sigma^2}{r\sigma^2 + n\sigma^2}
\]
where $\frac{r\sigma^2}{\sigma^2}$ is the Fisher information in the prior and $\frac{n\sigma^2}{\sigma^2}$ is the Fisher information in the data. Intuitively $B$ is the shrinkage factor and tells us how much the the prior mean $\mu_0$ contributes to the posterior mean compared to the data $Y$.

**Example 9.2.** In the case where $Y$ is normally distributed we know that $V(\mu) = c$, so we have
\[
\int \frac{\mu - \mu_0}{V(\mu)} d\mu = \frac{1}{2c} \mu^2 - \frac{\mu_0}{c} \mu
\]
and we see that after completing the square our conjugate prior on $\mu$ will be a Normal distribution.

**Example 9.3.** If $Y$ has a Poisson distribution then $V(\mu) = \mu$ and
\[
\int \frac{\mu - \mu_0}{V(\mu)} d\mu = \mu - \mu_0 \log \mu.
\]
Thus we can see the conjugate prior will be a Gamma distribution on $\mu$.

9.1. What is the conjugate prior distribution for a Binomial likelihood function?

9.2. **Jeffreys’ Prior.** Jeffreys’ prior on a parameter $\theta$ is
\[
\pi(\theta) d\theta \propto \sqrt{|I(\theta)|} d\theta,
\]
where $I(\theta)$ is the Fisher information matrix and $|I(\theta)|$ is the determinant of this matrix (why is this nonnegative?). Often, Jeffreys’ prior is called non-informative. This is misleading because there is no truly non-informative prior. Rather we think of it as a “reference” prior.

9.1. Jeffreys’ prior may or may not be proper (an improper prior is one that cannot be normalized since the integral diverges; not everyone agrees with the use of such priors, but in any case do not use an improper prior without knowing that the posterior is proper). In fact, it often isn’t proper. For this reason, it is important to check that the posterior is proper.
Jeffreys’ prior is motivated by an invariance argument: the goal is to find a method of constructing priors in such a way that finding Jeffreys’ prior on \( \theta \) and then reparametrizing to \( \tau \), a one-to-one function of \( \theta \), gives the same result as directly finding Jeffreys’ prior for \( \tau \).

\section*{9.2. Check that Jeffreys’ prior has this invariance property.}

For a location problem, a flat prior is a common choice of a (so-called) non-informative prior. Note that Jeffreys’ prior for a location parameter is indeed flat.

\begin{example}[Jeffreys’ prior in an NEF]
Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} \text{NEF}[\mu, V(\mu)] \). Then \( \bar{Y} \sim \text{NEF}[\mu, V(\mu)/n] \), with
\[
f(\bar{y}|\eta) \propto \exp(n(\eta \bar{y} - \psi(\eta))).
\]
Then \( I(\eta) = nV(\mu) \), so
\[
\text{Jeff}(\eta) = c\sqrt{V(\mu)},
\]
where \( n \) is absorbed into the proportionality constant \( c \). What if we are interested in Jeffreys’ prior on the \( \mu \) instead of on the natural parameter \( \eta \)? We must multiply by the Jacobian:
\[
\text{Jeff}(\mu) = \text{Jeff}(\eta) \left| \frac{d\eta}{d\mu} \right|^{-1} = c\sqrt{V(\mu)} \frac{1}{V(\mu)} = c \frac{1}{\sqrt{V(\mu)}}.
\]

\section*{9.3. Show that if you calculate Jeffreys’ prior on \( \mu \) directly, you get the same result as you do when you calculate the prior on \( \eta \) and then change variables.}

Five NEF’s with quadratic variance functions (NEF-QVF)s are summarized in Table 1:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\text{Elem.} & \text{Normal} & \text{Poisson} & \text{Gamma} & \text{Binomial} & \text{NBin} \\
\hline
\text{Y} & \mathcal{N}(\mu,1) & \text{Pois}(\mu) & \mu \text{Expo} & \text{Bern}(p) & \text{Geom}(p) \\
\hline
\text{Supp. of } Y & (\infty, \infty) & \{0,1,2,\ldots\} & (0,\infty) & \{0,1\} & \{0,1,\ldots\} \\
\hline
\mu = E(Y) & \mu & \mu & p & q/p \\
\hline
\eta & \log(\mu) & 1 - 1/\mu & \log(p/q) & \log(2q) \\
\hline
\psi(\eta) & \eta^2/2 & \exp - 1 & -\log(1 - \eta) & \log(2 - e^\eta) \\
\hline
\end{tabular}
\caption{Five NEF-QVF distributions}
\end{table}
Thus, the Jeffreys’ priors for the mean parameters of the five NEF-QVFs listed in Table 1 are shown in Table 2.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Distribution</th>
<th>Jeff(μ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>μ</td>
<td>1</td>
</tr>
<tr>
<td>Poisson</td>
<td>log(μ)</td>
<td>1/√μ</td>
</tr>
<tr>
<td>Gamma</td>
<td>1 − 1/μ</td>
<td>1/μ</td>
</tr>
<tr>
<td>Binomial</td>
<td>log(p/(1 − p))</td>
<td>1/√μ − μ²</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>log(2q)</td>
<td>1/√μ² + μ</td>
</tr>
</tbody>
</table>

Note that all of these are improper, except for the Binomial, for which the Jeffreys’ prior for p is Beta(1/2, 1/2).

9.3. Bayesian Point and Interval Estimation. The Bayesian approach focuses on $p(\theta | Y)$ instead of $p(Y | \theta)$; that is, on the probability of parameters given data instead of the probability of data given parameters. Thus, a point estimator of $\theta$ is a single quantity that “best represents” the posterior distribution. The expectation or the mode of the posterior distribution are two popular and common choices of point estimators of $\theta$.

A common Bayesian interval estimate is a probability interval (probability intervals are sometimes called by the weak name “credible intervals” or the wordy name “Bayesian confidence intervals”; we prefer to simply call them “probability intervals”). This means that our interval again includes $\theta$ with some set probability; however, we now take this probability over the posterior distribution of $\theta$ (holding $Y$ fixed), as opposed to the distribution of $Y | \theta$. Thus, while we still evaluate $P(\theta \in T(Y))$, this is now a probability statement about $\theta$ instead of $Y$. It should be noted that, while these intervals are constructed in a Bayesian way, they can still be evaluated using frequentist criteria.

In the Bayesian case, given the model $Y \sim \mathcal{N}(\mu, \sigma^2/n)$, we seek to construct an interval using the posterior $\mu | Y$. Letting the prior be flat (Jeffreys’ prior) the posterior $\mu | Y$ is distributed $\mathcal{N}(\bar{Y}, \sigma^2/n)$. Standardizing the posterior,

$$\frac{\mu - \bar{Y}}{\sigma \sqrt{n}} \sim \mathcal{N}(0, 1)$$

Therefore a 95% probability interval for $\mu$ is $\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

Example 9.5 (Numerical examples for point and interval estimation for the normal model). To illustrate the ideas of point and interval estimation, we consider a single observation $y = 120$ assumed to be obtained from a normal population with known variance $\sigma^2 = 100$ but unknown mean $\mu$ (to be inferred). The frequentist point estimator of $\mu$ is the observed
value \( y = 120 \), which is the same as the point estimate obtained by using Jeffreys’ prior. However, if one assumes a conjugate prior \( \mu \sim N(\mu_0, \tau^2) \) where \( \mu_0 = 100 \) and \( \tau^2 = 225 \), then the “prior sample size” \( r \) defined the context of conjugate priors for NEF distributions is \( \sigma^2/\tau^2 = 100/225 = 0.44 \). Consequently \( B = r/(r + n) = 0.44/1.44 \approx 0.31 \) and the point estimate of \( \mu \) equals \( B\mu_0 + (1 - B)y = 0.31 \times 100 + 0.69 \times 120 = 113.8 \). The variance of the posterior distribution of \( \mu \) is \( \sigma^2/(r + n) = 100/1.44 = 69.44 \) (notice the shrinkage from the prior variance of 225).

Moving to interval estimation, a 95% frequentist confidence interval for \( \mu \) is \( y \pm 1.96 \sigma \), i.e., \([100.4, 139.6]\). A probability interval or credible interval for \( \mu \) based on Jeffreys’ prior is exactly the same. However, with the conjugate prior, the posterior distribution of \( \mu \) given \( y \) is normal with mean 113.8 and variance 69.44 as seen in the previous paragraph, a 95% probability interval is \( 113.8 \pm 1.96 \times \sqrt{69.44} \) or \([97.56, 130.22]\).

Let us now evaluate the frequentist properties of the two Bayesian intervals obtained by using Jeffreys’ prior and the conjugate prior \( N(100, 225) \). Note that the interval obtained using Jeffreys’ prior has exactly the same coverage (95%) as the frequentist coverage because it is essentially the same interval. The coverage using the conjugate prior is calculated as follows:

\[
P_{\mu} \left[ B\mu_0 + (1 - B)y - 1.96 \frac{\sigma}{n + r} \leq \mu \leq B\mu_0 + (1 - B)y - 1.96 \frac{\sigma}{n + r} \right]
= P_{\mu} \left[ 31 + 0.69Y - 1.96 \frac{10}{1.44} \leq \mu \leq 31 + 0.69Y - 1.96 \frac{10}{1.44} \right]
= P_{\mu} \left[ 14.67 - \mu \leq -0.69Y \leq 47.33 - \mu \right]
= P_{\mu} \left[ \mu/0.69 - 68.59 \leq Y \leq \mu/0.69 + 21.26 \right]
= P_{\mu} \left[ \mu/0.69 - 68.59 - \mu \leq \frac{Y - \mu}{0.69} \leq \mu/0.69 + 21.26 - \mu \right]
= P_{\mu} \left[ \frac{-68.59 + 0.45\mu}{10} \leq Z \leq \frac{21.26 + 0.45\mu}{10} \right]
= \Phi(2.13 + 0.045\mu) - \Phi(-6.86 + 0.045\mu),
\]

where \( \Phi(\cdot) \) is the CDF of \( \mathcal{N}(0, 1) \). For \( \mu = 100 \), the above probability is approximately 0.99. The coverage probability for \( \mu \in [0, 130] \) is shown in Figure 2.

**Example 9.6** (Interval estimation for exponential parameter). Consider another problem, where \( Y_1 \ldots Y_n \) are i.i.d. from the Exponential distribution with mean \( \mu \). Let us construct the posterior distribution under the Jeffreys’ prior \( p(\mu) \propto \mu^{-1} \). Reparameterizing \( \lambda = \mu^{-1} \), we have

\[
p(\lambda|Y) = \lambda^n e^{-\lambda\bar{Y}} \frac{1}{\lambda},
\]
from which we obtain \( \lambda | \bar{Y} \sim \frac{\text{Gamma}(n)}{n \bar{Y}} \) and \( \mu | \bar{Y} \sim \frac{n \bar{Y}}{\text{Gamma}(n)} \).

Now approaching the problem from a frequentist perspective, note that \( \frac{n \bar{Y}}{\mu} \) is a pivot. Using the pivotal distribution, we obtain a frequentist confidence interval via

\[
P(a \leq \frac{n \bar{Y}}{\mu} \leq b) = 0.95.
\]

Note that this interval for \( \mu \) gives exactly the Bayesian solution. Therefore once again we find agreement between a frequentist solution and a Bayesian solution. However, the frequentist and the Bayesian may argue about the interpretation and philosophical underpinnings of these interval estimators.

9.4. **Prediction Intervals.** Consider \( Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2) \) with \( \sigma^2 \) known, and let \( \bar{Y} \) be their sample mean. We are interested in building a predictive interval for a future observation, \( Y_{n+1} \). We claim that the quantity \( Y_{n+1} - \bar{Y} \) is pivotal, with

\[
\frac{Y_{n+1} - \bar{Y}}{\sigma \sqrt{(1 + 1/n)}} \sim \mathcal{N}(0, 1).
\]

This gives the prediction interval

\[
\bar{Y} \pm 1.96 \sigma \sqrt{(1 + 1/n)}.
\]
This is the frequentist approach for obtaining predictive intervals.

In the Bayesian approach, predictive intervals are based on the posterior predictive distribution of the data, i.e., the probability distribution of a new observation $y_{\text{new}}$ given the observed data $y$. This distribution can be written as

$$p(y_{\text{new}}|y) = \int_{\Theta} p(y_{\text{new}}, \theta|y) d\theta = \int_{\Theta} p(y_{\text{new}}|\theta)\pi(\theta|y) d\theta,$$

where $\pi(\theta|y)$ is the posterior distribution of $\theta$.

### 10. Elementary Decision Theory

The general framework of Decision Analysis is based on Game Theory. Think about a game of “Nature” vs. “Statistician”. Nature picks the parameter $\theta$, which is unknown for the statistician, and then the statistician takes an action $a(Y)$, based on the data he has. An action is a map from the data to an action space. There are three important kinds of actions:

- **Point estimation.** Based on the data, the statistician derives a single numerical estimate for $\theta$. Possible estimators include MLE, UMVUE and posterior expectation.
- **Interval estimation.** The statistician provides an interval that contains the parameter $\theta$ with a certain level of confidence or probability.
- **Hypothesis testing.** The statistician either accepts or rejects a statement regarding hypothesized value(s) for $\theta$. As we have seen earlier, Hypothesis testing and interval estimation are very closely related.

The statistician faces a choice between a number of alternative actions. A natural question one may ask is what is a good action? To answer this question, we need to introduce the concept of a loss function $L(\theta, \delta(Y))$, which is a function of the truth $\theta$ and the action $\delta(Y)$. It represents the loss or cost due to making a specific decision. It is essentially the negative of utility. Once the loss function is chosen, the statistician’s job is to take the action to minimize the loss.

There are many choices for $L(\theta, \delta(Y))$, among which the squared error loss is the most widely used. However, there does not exist one single loss function that always outperforms others. Often, the context of the problem and the mathematical convenience may be the criteria for choosing the loss function.

If we consider the task of parameter estimation, loss functions are commonly functions of the difference between the estimator $\delta(Y)$ and the true value of the parameter $\theta$. For this task, it is often desirable for the loss to be monotonically increasing in the difference. Some common loss functions are

- Squared error loss: $L(\theta, \delta(Y)) = (\theta - \delta(Y))^2$ and
- Absolute loss: $L(\theta, \delta(Y)) = |\theta - \delta(Y)|$. 
• Asymmetric loss:
\[ L(\theta, \delta(Y)) = \begin{cases} p(\theta - \delta(Y)), & \theta \geq \delta(Y) \\ q(\delta(Y) - \theta), & \theta < \delta(Y). \end{cases} \]

• LinEx loss:
\[ L(\theta, \delta(Y)) = \exp \{ c(\theta - \delta(Y)) \} - c \{ \theta - \delta(Y) \} - 1. \]

Asymmetrical loss functions encourage us to err on one side of the true parameter because the cost of overestimating the parameter may be high compared to underestimating it by the same amount, or vice versa.

10.1. Risk: the frequentist approaches. In a frequentist framework, the first argument to the loss function – the true value of the parameter – is taken to be fixed, and the second argument – the estimator – is taken to be random. Taking an expectation across the data – and thus, across the estimator – we are able to map the estimator and for a given parameter value to a quantity with units of loss.

**Definition 10.1** (Risk function). The risk of an estimator \( \delta \) for a given parameter value \( \theta \) may be written as \( R(\theta, \delta) \), and is defined as the expectation of the loss holding \( \theta \) fixed and averaging over the data. That is,
\[ R(\theta, \delta) = E_{\theta} L(\theta, \delta(Y)). \]

**Remark 10.1.** Note that in the left hand side of (10.1), \( \delta \) is no longer a function of the observed data \( Y \). However, as a decision rule, it depends on data. For example, consider the problem of estimating \( \mu \), the mean of a distribution with known variance \( \sigma^2 \), where \( Y_1, \ldots, Y_n \) are iid observations from the distribution. Let \( \delta(Y) = \bar{Y} \) be a decision rule. Then, under the squared error loss, the risk is
\[ R(\theta, \delta) = R(\theta, \bar{Y}) = E_{\theta}(\theta - \bar{Y})^2 = E_{\theta}(\theta - \bar{Y})^2 = \sigma^2/n. \]

It is possible to denote the risk by \( R(\theta, \delta(Y')) \), where \( \delta(Y') \) is a function or a map. The dummy variable \( Y' \) can be used to emphasize the difference.

Since \( R(\theta, \delta) \) is a function of \( \theta \), we need additional definitions to compare decision rules. For example, how do we compare the risk functions associated with two decision rules \( \delta_1 \) and \( \delta_2 \)?

**Definition 10.2** (Dominating decision rule). A decision rule \( \delta_1 \) dominates decision rule \( \delta_2 \) if
\[ R(\theta, \delta_1) \leq R(\theta, \delta_2) \quad \forall \ \theta \in \Theta, \]

with strict inequality holding for some \( \theta \in \Theta \).

**Remark 10.2.** In Definition 10.2, if the strict inequality does not hold for any \( \theta \in \Theta \), then \( \delta_1 \) is as good as \( \delta_2 \).

**Definition 10.3** (Admissibility). A decision rule is admissible in the class of decision rules \( D \) if it is not dominated by any other decision rule in \( D \), i.e., there does not exist a decision rule \( \delta' \in D \) that dominates \( \delta \).
Remark 10.3. While admissibility is an important criterion, the following points should be kept in mind:

(a) Admissibility is typically a property of a decision rule with respect to a “class of decision rules,” and not a global property.

(b) Admissibility is a weak criterion of “goodness” of a decision rule. In fact, there are examples where a ridiculous decision rule can be admissible. Consider a single observation $Y \sim \text{Bin}(k, \theta)$ with $k$ known and consider the problem of estimating $\theta$ under a squared error loss from the single observation. Consider the decision rule $\delta(Y) = 1/2$ for $Y = 0, 1, \ldots, k$, i.e., irrespective of the outcome, $\theta$ is estimated as $1/2$. Then $\delta(Y)$ is an admissible estimator of $\theta$.

\textit{Ex} 10.1. Show that the $\delta(Y)$ defined in Remark 10.3(b) is an admissible estimator of $\theta$.

Definition 10.4 (Inadmissibility). A decision rule $\delta \in D$ is inadmissible if there exists a $\delta' \in D$ that dominates $\delta$.

To choose a rule $\delta$ from a class $D$, one can adopt the “play safe” strategy, and choose the rule that \textit{minimizes} the \textit{maximum} risk. Such a rule is called a minimax rule.

Definition 10.5 (Minimax Rule). The decision rule $\delta_M$ is minimax within a class of rules $D$ if

$$\delta_M = \arg\min_{\delta \in D} \max_{\theta \in \Theta} R(\theta, \delta).$$

Finding a minimax rule in a class of decision rules $D$ involves the following two steps:

(a) Fix a rule $\tilde{\delta} \in D$. Find $\theta^*(\tilde{\delta})$ that maximizes $R(\theta, \tilde{\delta})$ with respect to $\theta$.

(b) Search over all $\delta \in D$ and find $\delta_M$ that maximizes $R(\theta^*(\delta), \delta)$.

Example 10.1 (Finding a minimax rule). Consider a discrete parameter space $\Theta = \{1, 2, 3\}$, a class of decision rules

$$D = \{\delta_1, \delta_2, \delta_3, \delta_4\},$$

and the values of the risk function $R(\theta, \delta)$ for each $\theta \in \Theta$ and $\delta \in D$ shown in Table 3.

<table>
<thead>
<tr>
<th>Decision Rule</th>
<th>Parameter value $\theta_1$</th>
<th>Parameter value $\theta_2$</th>
<th>Parameter value $\theta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_1$</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$\delta_3$</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>$\delta_4$</td>
<td>12</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3. $R(\theta, \delta)$ for each $\theta \in \Theta$ and $\delta \in D$
• Which decision rules are admissible?
• Which is decision rule is minimax?

Now suppose one more decision rule $\delta_5$ ia added to Table 3, leading to Table 4.

<table>
<thead>
<tr>
<th>Decision Rule</th>
<th>Parameter value</th>
<th>Parameter value</th>
<th>Parameter value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_1$</td>
<td>12</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$\delta_3$</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>$\delta_4$</td>
<td>12</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>$\delta_5$</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

• Is $\delta_5$ admissible?
• Which decision rule is now minimax in the class $\{\delta_1, \ldots, \delta_5\}$?

**Proposition 10.1.** If a decision rule $\delta^*$ with a constant risk $R(\theta, \delta^*) = c(\delta^*)$ for all $\theta \in \Theta$ is admissible in the class $\mathcal{D}$, then it is minimax in $\mathcal{D}$.

**Proof.** Suppose $\delta^*$ has a constant risk and is admissible in the class $\mathcal{D}$, and suppose it is not minimax in $\mathcal{D}$. Then there exists a rule $\delta'$ that has a smaller maximum risk than that of $\delta^*$. This means for all $\theta \in \Theta$,

$$R(\theta, \delta') < \min_{\theta \in \Theta} R(\theta, \delta') < c(\delta^*),$$

which means $\delta^*$ is not admissible (contradiction).

10.2. **Bayes’ risk.** Minimax strategy is a conservative strategy (protection against the worst possible scenario). To see this, consider a slightly different variant of Table 4 obtained by adding 1000 to each entry in the column under $\theta_3$, as shown in Table 5.

<table>
<thead>
<tr>
<th>Decision Rule</th>
<th>Parameter value</th>
<th>Parameter value</th>
<th>Parameter value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_1$</td>
<td>12</td>
<td>7</td>
<td>1001</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>4</td>
<td>3</td>
<td>1002</td>
</tr>
<tr>
<td>$\delta_3$</td>
<td>2</td>
<td>5</td>
<td>1002</td>
</tr>
<tr>
<td>$\delta_4$</td>
<td>12</td>
<td>7</td>
<td>1002</td>
</tr>
<tr>
<td>$\delta_5$</td>
<td>3</td>
<td>3</td>
<td>1003</td>
</tr>
</tbody>
</table>

Then $\delta_1$ is minimax in $\{\delta_1, \ldots, \delta_5\}$. This is a strange choice, because clearly $\delta_2$ and $\delta_3$ are much better for $\theta = 1$ and $\theta = 2$, and $\delta_1$ is only
marginally better for $\theta = 3$. A better decision can be reached if we know a priori that $\theta$ has the following distribution:

$$\theta = \begin{cases} 
1 & \text{with probability 0.4} \\
2 & \text{with probability 0.5} \\
3 & \text{with probability 0.1}
\end{cases}$$

In this case, we can calculate the expected risk with respect to this prior distribution as shown in Table 6, and choose the rule $\delta_5$ as the decision rule with the smallest expected (Bayes’) risk.

<table>
<thead>
<tr>
<th>Decision Rule ($\delta$)</th>
<th>$\theta_1$ (.4)</th>
<th>$\theta_2$ (.5)</th>
<th>$\theta_3$ (.1)</th>
<th>Bayes’ Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_1$</td>
<td>12</td>
<td>7</td>
<td>1001</td>
<td>108.4</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>4</td>
<td>3</td>
<td>1002</td>
<td>103.3</td>
</tr>
<tr>
<td>$\delta_3$</td>
<td>2</td>
<td>5</td>
<td>1002</td>
<td>103.5</td>
</tr>
<tr>
<td>$\delta_4$</td>
<td>12</td>
<td>7</td>
<td>1002</td>
<td>108.5</td>
</tr>
<tr>
<td>$\delta_5$</td>
<td>3</td>
<td>3</td>
<td>1003</td>
<td>103.0</td>
</tr>
</tbody>
</table>

Thus, an easy way to translate the frequentist idea of risk to a Bayesian framework is to simply incorporate a prior, and to consider the expected risk taken with respect to the prior distribution. This quantity is called the Bayes’ risk.

**Definition 10.6 (Bayes’ risk).** Given a prior over the parameter $\pi(\theta)$, the Bayes risk is the expectation of the frequentist risk taken across $\theta$ with respect to $\pi(\theta)$. That is,

$$B(\pi, \delta) = \mathbb{E}_\pi \{R(\theta, \delta)\} = \mathbb{E}_\pi \{\mathbb{E}_\theta (\theta, \delta(Y))\}$$

**Definition 10.7 (Bayes’ rule).** For a given prior $\pi(\theta)$, the Bayes’ rule in a class $D$, is the decision rule $\delta^\pi \in D$ that minimizes the Bayes’ risk, i.e.,

$$\delta^\pi = \arg \min_{\delta \in D} B(\pi, \delta).$$

**10.2.** It seems unusual to define Bayes risk in terms of an expectation taken over the data. Rewrite the definition of Bayes risk so that it yields a quantity defined in terms of the expected posterior loss, and show how minimizing the Bayes risk corresponds to minimizing the expected posterior loss given any point in the sample space. *Hint: it may be clearer to do this by assuming that the model is continuous and using integral notation for the expectations.*

The above pencil problem leads to the following alternative definition of Bayes’ rule:
**Definition 10.8** (Bayes’ rule in terms of the posterior distribution). For a given prior \( \pi(\theta) \), the Bayes’ rule in a class \( \mathcal{D} \), is the decision rule \( \delta^\pi \in \mathcal{D} \) that minimizes the expectation of the loss function with respect to the posterior distribution of \( \theta \) given \( Y \), i.e.,

\[
\delta^\pi = \arg \min_{\delta \in \mathcal{D}} r(Y, \delta(Y)),
\]

where

\[
r(Y, \delta(Y)) = \int L(\theta, \delta(Y)) \pi(\theta|Y) d\theta.
\]

Definition 10.8 can be used to find Bayes’ rules for specific loss functions. For example, under the squared error loss, the objective function

\[
r(Y, \delta(Y)) = \int (\theta - \delta(Y))^2 \pi(\theta|Y) d\theta = E_{\theta|Y}[(\theta - \delta(Y))^2|Y],
\]

which is minimized by

\[
\delta^\pi = E(\theta|Y),
\]

the posterior expectation. The corresponding Bayes’ risk is

\[
B(\pi, \delta^\pi) = \int E_{\theta|y}[(\theta - \delta(y))^2|y] m(y) dy = \int \text{Var}(\theta|y)m(y) dy,
\]

where

\[
m(y) = \int f_\theta(y) \pi(\theta) d\theta
\]

is the marginal density of \( Y \).

10.3. What is the Bayes’ rule in terms of the posterior loss for the absolute error loss function?

10.3. **Bayes’ rule and admissibility.** Lehmann and Casella (1998) give a comprehensive discussion on this topic. Here we will discuss some of the main ideas and results on the connection between Bayes’ rules and admissibility. We shall see that under a wide array of conditions, Bayes’ estimators are admissible. Interestingly, a converse of this result is also true, and such a result is popularly known as the **complete class theorem.** First, we state a result on admissibility of unique Bayes’ rules.

**Theorem 10.1.** Any unique Bayes’ rule is admissible.

**Proof.** Let \( \delta^\pi \) be Bayes for prior \( \pi \). Consider a rule \( \delta' \) for which

\[
R(\theta, \delta') \leq R(\theta, \delta^\pi) \quad \forall \theta \in \Theta.
\]

Then, the Bayes’ risk associated with \( \delta' \) and prior \( \pi \) satisfies

\[
B(\pi, \delta') = \int R(\theta, \delta') \pi(\theta) d\theta
\]
Thus, $\delta'$ is also a Bayes rule. But by the condition of uniqueness of the Bayes’ rule, we must have $\delta^\pi(Y) = \delta'(Y)$ almost surely. \hfill \Box

**Theorem 10.2.** Let $\Theta$ be an open subset of $\mathbb{R}$ and suppose

(i) The prior distribution $\pi(\theta)$ has support $\Theta$.

(ii) the risk function $R(\theta, \delta)$ is continuous in $\theta$ for all $\delta \in D$.

If $\delta^\pi$ is Bayes’ rule for $\pi$ with finite Bayes’ risk, then $\delta^\pi$ is admissible.

**Proof.** Assume that $\delta^\pi$ is inadmissible and let $\delta'$ dominate $\delta^\pi$. Then,

$$ R(\theta, \delta') \leq R(\theta, \delta^\pi) \quad \forall \theta \in \Theta, $$

$$ R(\theta, \delta') < R(\theta, \delta^\pi) \quad \text{for some } \theta_0 \in \Theta. $$

Let $\eta = R(\theta_0, \delta^\pi) - R(\theta_0, \delta')$. Then, by continuity of $R(\cdot, \delta)$, we can find an $\varepsilon > 0$ such that

$$ R(\theta, \delta^\pi) - R(\theta, \delta') > \pi/2, \quad \forall \theta : |\theta - \theta_0| < \varepsilon. $$

Let $A = \{\theta : |\theta - \theta_0| < \varepsilon\}$ denote the set of all values of $\theta$ that lie in an $\varepsilon$-neighbourhood of $\theta_0$.

Let $B(\pi, \delta^\pi)$ and $B(\pi, \delta')$ denote the Bayes’ risks of $\delta^\pi$ and $\delta'$ with respect to prior $\pi(\theta)$. Then,

$$ B(\pi, \delta^\pi) - B(\pi, \delta') = \int \{R(\theta, \delta^\pi) - R(\theta, \delta')\} \pi(\theta)d\theta $$

$$ = \int_A \{R(\theta, \delta^\pi) - R(\theta, \delta')\} \pi(\theta)d\theta + \int_{A^c} \{R(\theta, \delta^\pi) - R(\theta, \delta')\} \pi(\theta)d\theta $$

$$ \geq \int_A \{R(\theta, \delta^\pi) - R(\theta, \delta')\} \pi(\theta)d\theta, \quad \text{since } R(\theta, \delta^\pi) \geq R(\theta, \delta') \quad \forall \theta \in A^c, $$

$$ > \frac{\eta}{2} \int_A \pi(\theta)d\theta $$

$$ > 0. $$

This contradicts that $\delta^\pi$ is Bayes’ rule with respect to prior $\pi$. \hfill \Box

**Theorem 10.3.** Consider a discrete parameter space $\Theta = \{\theta_1, \ldots, \theta_K\}$, and suppose $\delta^\pi$ is Bayes for $\pi$, which is discrete with $\pi(\theta = \theta_j) = \pi_j$ for $j = 1, \ldots, K$, such that $\pi_j > 0$ for all $j \in \{1, \ldots, K\}$. Then $\delta^\pi$ is admissible.

**Proof.** Assume that $\delta^\pi$ is inadmissible and let there exist a $\delta'$ that dominates $\delta^\pi$. Then,

$$ R(\theta, \delta') \leq R(\theta, \delta^\pi) \quad \forall \theta \in \{1, \ldots, K\}, $$
with strict inequality for at least one $\theta_0 \in \{\theta_1, \ldots, \theta_K\}$. Then
\[
B(\pi, \delta') = \sum_{j=1}^{K} \pi_j R(\theta_j, \delta') < \sum_{j=1}^{K} \pi_j R(\theta_j, \delta^\pi) = B(\pi, \delta^\pi).
\]
The inequality above occurs because $\pi_j > 0$ for all $j$. This contradicts the definition of Bayes’ rule. \hfill \Box

10.3.1. A geometric approach to visualize admissibility and connect it to the Bayes’ rule. Consider the situation where the parameter space is discrete and consists of only two possible values, i.e., $\Theta = \{\theta_1, \theta_2\}$. The risk of any decision rule $\delta$ is a two-dimensional vector $(R(\theta_1, \delta), R(\theta_2, \delta))$. Consider the class of decision rules $D = \{d_1, \ldots, d_8\}$, and their risk vectors plotted on a 2-D Cartesian plane as shown in Figure 3.

To determine if a given decision rule $\delta$ is admissible, draw a rectangle for the rule as in Figure 3 (green and red rectangles for rules $d_2$ and $d_5$ respectively); if other decision rules are within the area determined by the rectangle (excluding the boundary), $\delta$ is inadmissible. If no other decision rules are within the rectangle area (excluding the boundary), the rule is admissible. Thus, $d_2$ is admissible because nothing “beats” it in two dimensions, but $d_5$ is not admissible because it is dominated by $d_4$.

Another way to characterize admissible rules is to create a convex hull (smallest convex set) containing all the rules. Then all rules lying on the
south-west edge are admissible rules. If we there are only finitely many rules that are allowed, then there are finitely many points on the graph, and we can form the convex hull of these points, which will be a polygon. The interpretation of this polygon is that it is the set of possible rules if we allow convex combination of the existing rules (i.e. randomized rules).

Let us now extend the “rectangle rule” and the “south-west edge rule” to \( K \) dimensions, where \( K \) is any integer. Consider a set of decision rules \( D \) that includes a finite number of non-randomized rules, and the randomized rules that are convex combinations of these non-randomized rules. Let \( \Theta = \{\theta_1, \ldots, \theta_K\} \) be the \( K \)-dimensional discrete parameter space and for \( k = 1, \ldots, K \), let \( s_k = R(\theta_k, \delta) \) denote the risk associated with decision rule \( \delta \in D \) when \( \theta = \theta_k \). Define the risk set \( S \) of all risk functions as the set of all possible vectors \( s = (s_1, \ldots, s_K)^T \) that satisfy \( s_k = R(\theta_k, \delta) \), \( k = 1, \ldots, K \) for some decision rule \( \delta \in D \). In other words, each vector \( s \) in the risk set corresponds to a decision rule \( \delta \in D \), and represents the risk vector \( (R(\theta_1, \delta), \ldots, R(\theta_K, \delta)) \).

**Proposition 10.2.** The risk set \( S \) defined above is convex.

**Proof.** Consider \( s_a, s_b \in S \). Then the \( K \)-dimensional vectors \( s_a \) and \( s_b \) correspond to two decision rules \( \delta_a \in D \) and \( \delta_b \in D \) respectively. Let \( \delta_c \) be a randomized decision rule defined as

\[
\delta_c = \begin{cases} 
\delta_a & \text{with probability } p \\
\delta_b & \text{with probability } 1 - p
\end{cases},
\]

where \( 0 < p < 1 \). Then \( \delta_c \in D \). Consequently, \( s_c = ps_a + (1 - p)s_b \in S \). \( \Box \)

Let \( \delta \) be a rule with risk vector \( s = (s_1, \ldots, s_K)^T \). Consider the set

\[
Q(s) = \{(x_1, \ldots, x_K)^T \in \mathbb{R}^K : x_1 \leq s_1, \ldots, x_K \leq s_K\},
\]

called the lower quantant of the point \( s \) (see Figure 4 for a two dimensional representation). If \( \delta \) is admissible, then its risk vector \( s \) satisfies

\[Q(s) \cap S = \{s\}\] (10.2)

Define a prior distribution \( \pi(\theta_j) = \pi_j \) for \( j = 1, \ldots, K \), such that \( \sum_{j=1}^{K} \pi_j = 1 \) and \( \pi_j > 0 \) for \( j = 1, \ldots, K \). Then, the Bayes’ risk of decision rule with respect to prior \( \pi \) is

\[
B(\pi, \delta) = \pi_1 R(\theta_1, \delta) + \ldots + \pi_K R(\theta_K, \delta).
\]

Then minimizing Bayes’ risk boils down to the following linear programming problem:

\[\text{Minimize } \pi_1 s_1 + \ldots + \pi_K s_K \]
\[\text{subject to } (s_1, \ldots, s_K)^T \in S.\]
This linear programming problem can be solved by moving the hyperplane 
\[ \pi_1s_1 + \ldots + \pi_Ks_K = c \]
towards the origin till it touches a vertex of the convex set \( S \). Figure 5 shows a 2-D representation, where the hyperplane is a line. The vertex which is touched by the hyperplane represents a decision rule that satisfies (10.2) for admissibility, and hence is an admissible rule.

Now, we prove the converse, i.e., if \( \delta^* \) is an admissible rule, then it is Bayes’ with respect to some prior. Let \( \delta^* \) be an admissible decision rule in \( \mathcal{D} \). Then we shall show that it is Bayes’ with respect to some prior. If \( \delta^* \) is admissible, then it must satisfy (10.2), i.e.,
\[ Q(s^*) \cap S = \{s^*\}, \]
where
\[ s^* = (R(\theta_1, \delta^*), \ldots, R(\theta_K, \delta^*))^T. \]

Let \( \tilde{Q}(s^*) = Q(s^*) \setminus \{s^*\} \), the set \( Q(s^*) \) without the point \( s^* \). Then, \( \tilde{Q}(s^*) \) is convex, as is \( S \), and \( \tilde{Q}(s^*) \cap S = \emptyset \). Thus, \( \tilde{Q}(s^*) \) and \( S \) are two disjoint convex subsets of \( \mathbb{R}^K \).

By the separating hyperplane theorem, there exists a hyperplane separating \( \tilde{Q}(s^*) \) and \( S \), i.e., there exists a vector \( w = (w_1, \ldots, w_K)^T \) such that
\[ w^T x \leq w^T s, \quad \forall x \in \tilde{Q}(s^*), \ s \in S. \]
\[ (10.3) \]
Note that no co-ordinate of \( w \) can be negative. For example, if \( w_1 < 0 \), we can make \( x_1 \) arbitrarily negative within \( \tilde{Q}(s^*) \), so that \( w^T x \) is arbitrarily large and greater than \( w's \) for \( s \in S \). Hence \( w_j > 0 \) for all \( j = 1, \ldots, K \).
Define $\pi_j = w_j / \sum_{j=1}^{T} w_j$ for $j = 1, \ldots, K$. Then, from (10.3), it follows that

\[ \pi^T x \leq \pi^T s, \quad \forall x \in \tilde{Q}(s^*), \ s \in \mathcal{S}, \]

where $\pi = (\pi_1, \ldots, \pi_K)^T$ and $\pi_j > 0$ for all $j = 1, \ldots, K$. Thus, for any $s \in \mathcal{S}$, it follows from (10.4) that

\[ \pi^T s \geq \sup_{x \in \tilde{Q}(s^*)} \pi^T x = \pi^T s^*, \]

since $s^*$ is a limit point of $\tilde{Q}(s^*)$. Consequently,

\[ \pi^T s^* \geq \pi^T s \quad \forall s \in \mathcal{S}, \]

which means

\[ B(\pi, \delta^*) \leq B(\pi, \delta) \quad \forall \delta \in \mathcal{D}. \]

Noting that $B(\pi, \delta)$ is the Bayes' risk for rule $\delta$ with respect to prior $\pi(\theta_j) = \pi_j$ for $j = 1, \ldots, K$, (10.5) implies that $\delta^*$ is the Bayes' rule with respect to prior $\pi$.

10.4. Admissibility of sample mean in a univariate normal. Suppose we have an estimation problem with with squared error loss where the data are i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu$ unknown and $\sigma^2$ known. We also believe that $\mu \sim \mathcal{N}(\mu_0, \tau^2)$. 

\[ \pi_1 s_1 + \pi_2 s_2 = c \]
Normals are conjugate priors for normals, so Pearson conjugate theory tells us that the posterior distribution is given by
\[ \mu | \bar{Y} \sim N \left( (1 - B) \bar{Y} + B \mu_0, (1 - B) \frac{\sigma^2}{n} \right), \]
where \( B \) is given by
\[ B \equiv \frac{\sigma^2}{\frac{\sigma^2}{n} + \tau^2}, \]
and is called the shrinkage factor. The Bayes rule that minimizes posterior squared error loss is the posterior mean, which we will denote \( \tilde{\mu} \).

For the rest of the problem, we WeLG take \( \mu_0 = 0 \) and \( \frac{\sigma^2}{n} = 1 \). We call this the canonical form of the problem because it sets the known constants to their most convenient values. We don’t give up generality here because we note that we can transform any form of the problem into this canonical form by shifting or scaling.

Squared error loss provides a particularly convenient way to parameterize the risk function as a function of \( \tau^2 \), noted in figure 5. At one extreme, for \( \tau = 0 \), the prior is a point mass at \( \mu_0 = 0 \). The prior is dogmatic so it will not change regardless of the amount of data supplied, so it induces a loss function that is a parabola in the underlying parameter \( \mu \), with the loss exactly 0 for \( \mu = 0 \), and loss unbounded as the actual value of \( \mu \) moves away from 0.

On the other extreme, as we let \( \tau^2 \) tend toward infinity, we see that the posterior mean \( \tilde{\mu} \) converges to \( \bar{Y} \) by Slutsky’s Theorem. Squared error loss

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{risk_functions.png}
\caption{Risk functions for \( \mu \) as a function of \( \tau^2 \) in the canonical case. We see that risk functions for moderate values of \( \tau \) are mixtures of the risk functions for extreme values of \( \tau \). As \( \tau^2 \to \infty \), we see convergence with the sample mean, which has constant variance. This indicates that the sample mean is a Bayes rule for normal prior with variance that tends to infinity.}
\end{figure}
in this case is just the variance of $\bar{Y}$ which is constant for any true value of $\mu$. For values of $\tau^2$ in between, we get a mixture of these two risk functions, with parabolas that gradually flatten out and tend toward $R(\mu, \bar{\mu}) = 1$. See Figure [6]. The curvature for $\tau^2$ moderate is due to the bias introduced to the estimate for $\mu$ in order to shrink its variance.

\section*{10.4. Using the above expression for $\bar{\mu}$, derive an expression for the risk function as a function of $\mu$ and $\tau^2$, and show how the shape changes as $\tau^2$ goes from 0 to $\infty$.}

Using the fact that the sample mean is the limit of Bayes rules allows us to prove the theorem.

\textbf{Theorem 10.4.} For i.i.d. data $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ with $\sigma^2$ known, the sample mean $\bar{Y}$ is an admissible estimator for $\mu$ with respect to squared-error loss.

\textit{Proof.} We proceed by contradiction using the canonical form of the problem, where we assume $\frac{\sigma^2}{n} = 1$ WeLoG. Suppose that $\bar{Y}$ is inadmissible. This implies that there exists some estimator $t(\bar{Y})$ which dominates $\bar{Y}$, that is,

$$E_{\mu}(t(\bar{Y}) - \mu)^2 \leq E_{\mu}(\bar{Y} - \mu)^2$$

for all $\mu$, with strict inequality holding for some $\mu$.

Fix some $\mu$ at which strict inequality holds, and WeLoG, assume that this $\mu = 0$. Because the risk function is continuous, this implies that there exists a region of width $2\varepsilon$ where $R(t(\mu, Y)) - R(\mu, \bar{Y}) \geq \varepsilon$ for some $\varepsilon > 0$ (This is illustrated by inscribing squares on either side of 0 in figure [7]). Additionally, note that the risk of $\bar{Y}$ for the canonical case is 1 everywhere, so we can express this dominance with the inequality

$$E_{\mu}(t(\bar{Y}) - \mu)^2 \leq 1 - \varepsilon I_{\mu \in (-\varepsilon, \varepsilon)}.$$

Where $I$ is an indicator function that is 1 when $\mu$ is in the region where $t(\bar{Y})$ beats $\bar{Y}$.

Now consider Bayes rules for $\mu$ corresponding to priors of the form $\mu \sim \mathcal{N}(0, \tau^2)$. These estimators take the form $c\bar{Y}$, where $c \equiv \frac{\tau^2}{1 + \tau^2}$. Note that as $\tau^2 \to \infty$, $c\bar{Y} \to \bar{Y}$.

For all $\tau^2$, there is a corresponding Bayes rule that minimizes the Bayes risk. So for any $\tau$, we know the following about the estimator $t(\bar{Y})$ that we defined above:

$$E_{\pi_{\tau^2}}E_{\mu}(c\bar{Y} - \mu)^2 \leq E_{\pi_{\tau^2}}E_{\mu}(t(\bar{Y}) - \mu)^2 \leq E_{\pi_{\tau^2}}(1 - \varepsilon I_{\mu \in (-\varepsilon, \varepsilon)}))$$

where $\pi_{\tau^2}(\mu)$ is the prior defined with the $\tau^2$ corresponding to $C$, and the subscripting indicates that the outer expectations are taken given this prior. The second step follows from incorporating the result above.
Risk functions for $\bar{Y}$ and $t(\bar{Y})$

Figure 7. Risk functions for $\bar{Y}$ and a dominating estimator $t(\bar{Y})$ that is strictly better at 0. The continuity of risk functions allows us to inscribe a box between $\bar{Y}$ and $t(\bar{Y})$ with dimensions $\epsilon \times 2\epsilon$ for some $\epsilon > 0$. This defines a function between the two risk functions that is easily represented with an indicator.

By multiple bias-variance decompositions, it can be shown that the LHS of this equation becomes $C$ yielding:

$$
\frac{\tau^2}{1 + \tau^2} \leq E_{\pi, \tau}(1 - \epsilon I(\mu \in (-\epsilon, \epsilon)))
= 1 - \epsilon P(-\epsilon < \tau Z < \epsilon)
= 1 - \epsilon \left[ \Phi \left( \frac{\epsilon}{\tau} \right) - \Phi \left( -\frac{\epsilon}{\tau} \right) \right]
\approx 1 - \epsilon \left( \frac{2\epsilon}{\tau} \frac{1}{\sqrt{2\pi}} \right).
$$

We now rewrite both sides for comparison as we let $\tau \to \infty$, grouping all of the constant factors on the RHS into a constant $K$:

$$
1 - \frac{1}{1 + \tau^2} \leq 1 - \frac{k}{\tau}.
$$

As we allow $\tau \to \infty$ this is clearly wrong: the LHS goes as $1 - \tau^{-2}$ whereas the RHS only goes as $1 - \tau^{-1}$, so there exists some finite $\tau$ over which the LHS is strictly greater. Thus, the rule $c\bar{Y}$, does not minimize the Bayes risk for all $\tau$, and thus it is not a Bayes rule, which is a contradiction.

Therefore, the assumption that we began with, that there exists some $t(\bar{Y})$ that dominates $\bar{Y}$ in a frequentist sense must be false. Thus, there exists no such estimator, and $\bar{Y}$ is admissible.
10.5. Hierarchical Models, Stein’s Theorem and Empirical Theorem. We have proved in the previous section that in the univariate case, the sample average is an admissible estimator for estimating the mean of a Normal distribution. When it comes to $k \geq 3$ independent Normal populations, however, James and Stein proved in 1961 that the sample averages are no longer admissible under squared error loss.

Consider $K$ independent normal populations with mean $\mu_i$ and known variances $V_i$ for $i = 1, \ldots, K$. Assume that $V_i = 1$ for $i = 1, \ldots, K$. Let $Y_1, \ldots, Y_K$ be observations such that $Y_i \overset{iid}{\sim} N(\mu_i, 1)$ for $i = 1, \ldots, K$. The problem is to estimate the vector $\mu = (\mu_1, \ldots, \mu_K)^T$. Clearly the MLE of $\mu$ is $\mu^{MLE} = Y = (Y_1, \ldots, Y_K)^T$.

For $K = 1$, $\mu^{MLE}$ is an admissible estimator of $\mu$ under the squared error loss. For $k \geq 3$, consider the compound square error loss

$$L(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2 = \sum_{i=1}^{K} (\mu_i - \hat{\mu}_i)^2.$$  

The risk function for $\hat{\mu}$ under loss (10.6) is

$$R(\mu, \hat{\mu}) = E_{\mu} \left\{ \sum_{i=1}^{K} (\mu_i - \hat{\mu}_i)^2 \right\}.$$  

Substituting $\hat{\mu} = \mu^{MLE}$ in (10.7) yields the risk of the MLE as

$$R(\mu, \hat{\mu}) = E_{\mu} \left\{ \sum_{i=1}^{K} (\mu_i - Y_i)^2 \right\} = K.$$  

Is there any other estimator that can dominate the MLE? This seems unlikely, because observations are drawn from independent normal distributions. In 1961, James and Stein proved that the estimator

$$\hat{\mu}^{JS} = \left(1 - \hat{B}\right) Y,$$

dominates the MLE under loss function (10.6), where

$$\hat{B} = \frac{K - 2}{S},$$

and $S = \sum_{i=1}^{K} Y_i^2 = \|Y\|^2$ is the squared norm of $Y$.

To prove Stein’s theorem, we need a fundamental lemma, also known as Stein’s lemma, which is interesting in its own right because it provides a characterization of the normal distribution. Below, we state the lemma.

**Lemma 10.1** (Stein’s lemma). If $Y \sim N(\mu, V)$ and $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is a differentiable function satisfying $E |h'(Y)| < \infty$, then

$$E \{h(Y)(Y - \mu)\} = VE \{h'(Y)\}.$$
Proof of Stein’s Theorem. The risk of the J-S estimator with respect to loss (10.6) is:

\[
R(\mu, \hat{\mu}^{JS}) = E_\mu \left\{ \| (1 - \hat{B}) Y - \mu \|^2 \right\} \\
= E_\mu \left\{ \sum_{i=1}^{K} (Y_i - \mu_i)^2 \right\} + E_\mu \left\{ \hat{B}^2 \sum_{i=1}^{K} Y_i^2 \right\} - 2E_\mu \left\{ \hat{B} \sum_{i=1}^{K} Y_i (Y_i - \mu_i) \right\} ,
\]

(10.10)
after a few lines of straightforward algebra. The first term of (10.13) is the risk of MLE and equals \( K \) by (10.8). The second term equals

\[
E_\mu \left\{ \hat{B}^2 S \right\} = E_\mu \left\{ \frac{(k - 2)^2}{S^2} \times S \right\} = (k - 2)^2 E_\mu \left\{ \frac{1}{S} \right\} .
\]

Substituting \( \hat{B} = (K - 2)/S \) in the third term of (10.13) reduces to:

\[
-2E_\mu \left\{ \frac{K - 2}{S^2} \sum_{i=1}^{K} Y_i (Y_i - \mu_i) \right\} = -2(K - 2) \sum_{i=1}^{K} E_\mu \left\{ \frac{Y_i}{S} (Y_i - \mu_i) \right\} .
\]

Letting \( h(Y_i) = Y_i/S \) and applying Stein’s lemma, it follows that

\[
E_\mu \left\{ \frac{Y_i}{S} (Y_i - \mu_i) \right\} = E_\mu \left\{ \frac{\partial}{\partial Y_i} \left( \frac{Y_i}{S} \right) \right\} = E_\mu \left\{ \frac{S - 2Y_i}{S^2} \right\} .
\]

Substituting the above in the third term of (10.13) yields

\[
-2(K - 2) \sum_{i=1}^{K} E_\mu \left\{ \frac{S - 2Y_i}{S^2} \right\} = -2(K - 2)^2 E_\mu \left\{ \frac{1}{S} \right\} .
\]

Substituting the first, second and third terms into (10.13) gives the risk of the J-S estimator as

\[
(10.11) R(\mu, \hat{\mu}^{JS}) = K + (K - 2)^2 E_\mu \left\{ \frac{1}{S} \right\} - 2(K - 2)^2 E_\mu \left\{ \frac{1}{S} \right\} \\
= K - (K - 2)^2 E_\mu \left\{ \frac{1}{S} \right\} \\
< K = R(\mu, \hat{\mu}^{MLE}).
\]

This establishes Stein’s Theorem. \( \Box \)

10.5.1. Explaining Stein’s paradox. Stein’s theorem raises an important question: how can we explain the anomaly that pooling information from apparently unrelated populations “shrinks” the distribution of the estimator of the normal mean vector? Further, what is special about \( K \geq 2? \) There are two plausible explanations: (a) Stigler’s Galtonian perspective and (b) The empirical Bayes’ perspective by Efron and Morris. First we discuss Stigler’s Galtonian perspective of Stein’s theorem, as discussed in Stigler (1990).
Stigler (1990) notes that the estimation problem described in this Section involves pairs of values \((Y_i, \mu_i)\) for \(i = 1, \ldots, K\), where one element of each pair \((Y_i)\) is known and the other \((\mu_i)\) is unknown. Since the \(\mu_i\)'s are unknown, the pairs cannot be plotted, but one can imagine what such a plot should look like. Because the data generating model is \(Y_i \sim \mathcal{N}(\mu_i, 1)\) for \(i = 1, \ldots, K\), for fixed \(\mu_0\), one can find an estimator of \(Y\) from the regression line \(E(Y|\mu) = \mu\).

However, our goal is to estimate all the \(\mu_i\)'s given all the \(Y_i\)'s, without any assumption about a possible distribution structure for the \(\mu_i\)'s. In the context of the MLE, such an estimation is using the \(E(Y|\mu) = \mu\), which is \(\mu = Y\), to obtain

\[
\hat{\mu}^{\text{MLE}}_i = Y_i, \quad i = 1, \ldots, K.
\]

But Stigler (1990) points out that the MLE is using the “wrong regression line” of \(E(Y|\mu) = \mu\) for estimating \(\mu\) from \(Y\), instead of estimating it from the regression of \(\mu\) on \(Y\), i.e. \(E(\mu|Y)\). This observation suggests that the MLE can possibly be improved upon by attempting to approximate \(E(\mu|Y)\). But how this is even possible without making any distributional assumptions about \(\mu\)?

Consider the “best” estimator of \(\mu\) of the form

\[
(10.14) \quad \hat{\mu}_i = \hat{\beta}Y_i, \quad i = 1, \ldots, K,
\]

where \(\hat{\beta}\) can be obtained by minimizing the compound squared error loss

\[
\sum_{i=1}^{K} (\hat{\mu}_i - \mu_i)^2 = \sum_{i=1}^{K} (\hat{\beta}Y_i - \mu_i)^2.
\]

Minimizing the above expression, we obtain

\[
(10.15) \quad \hat{\beta} = \frac{\sum_{i=1}^{K} \mu_i Y_i}{\sum_{i=1}^{K} Y_i^2}.
\]

However, (10.15) cannot be used to estimate \(\beta\) because it involves the \(\mu_i\)'s in the numerator. One possible way to circumvent this problem is to find “something like an estimator” of the numerator \(\sum_{i=1}^{K} \mu_i Y_i\).

Note that

\[
E_{\mu} \left( \sum_{i=1}^{K} \mu_i Y_i \right) = \sum_{i=1}^{K} \mu_i^2
\]

\[
E_{\mu} \left( \sum_{i=1}^{K} Y_i^2 - K \right) = \sum_{i=1}^{K} \mu_i^2.
\]

Thus, \(\sum_{i=1}^{K} \mu_i Y_i\) and \(\sum_{i=1}^{K} Y_i^2 - K\), both have the same expectation. Substituting \(\sum_{i=1}^{K} Y_i^2 - K\) for \(\sum_{i=1}^{K} \mu_i Y_i\) in the numerator of the RHS of (10.15), we have

\[
\hat{\beta} = \frac{\sum_{i=1}^{K} Y_i^2 - K}{\sum_{i=1}^{K} Y_i^2} = 1 - \frac{K}{S}.
\]
Thus, we have a shrinkage estimator

\[ \hat{\mu}_i = \left(1 - \frac{K}{S}\right) Y_i, \quad i = 1, \ldots, K, \]

which is similar to the JS-estimator with the factor \( K - 2 \) replaced by \( K \).

This problem was resolved by Meng (2005) who presented the following elegant argument to obtain the correct JS-estimator through the Galtonian perspective. Instead of replacing the numerator of \( \hat{\beta} \) in (10.15), consider replacing the entire RHS of (10.15), i.e., \( \left( \sum_{i=1}^{K} \mu_i Y_i \right) / S \) by something that gives the same expectation. For this, use the following identity that was proved by Stigler (1990):

**Lemma 10.2.**

\[ E_\mu \left( \frac{\sum_{i=1}^{K} \mu_i Y_i}{S} \right) = E_\mu \left( 1 - \frac{K - 2}{S} \right). \]

**Proof.** Substitute \( h(Y_i) = Y_i / S \). Then \( h'(Y_i) = (S - 2Y_i^2)/S^2 \). Applying Stein’s lemma (Lemma 10.1), it follows that for \( i = 1, \ldots, K \),

\[
E_\mu \{ (Y_i - \mu_i) h(Y_i) \} = E_\mu \{ h'(Y_i) \}
\]

\[
\Rightarrow E_\mu \left\{ (Y_i - \mu_i) \frac{Y_i}{S} \right\} = E_\mu \left\{ \frac{S - 2Y_i^2}{S^2} \right\}
\]

\[
\Rightarrow E_\mu \left\{ \frac{Y_i^2 - \mu_i}{S} \right\} = E_\mu \left\{ \frac{S - 2Y_i^2}{S^2} \right\}
\]

Summing over \( i \), we get

\[ E_\mu \left\{ 1 - \frac{\sum_{i=1}^{K} \mu_i Y_i}{S} \right\} = E_\mu \left\{ \frac{K - 2}{S} \right\}, \]

and the result follows. \( \square \)

Using Lemma 10.2 to replace \( \hat{\beta} = \left( \sum_{i=1}^{K} \mu_i Y_i \right) / S \) by \( 1 - (K - 2)/S \) yields the JS-estimator.

**The empirical Bayes’ perspective**

Consider the following hierarchical model:

\[ Y_i \overset{\text{independent}}{\sim} \mathcal{N}(\mu_i, 1), \quad i = 1, \ldots, K, \]

\[ \mu_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad i = 1, \ldots, K. \]

(10.17)

Recall that the posterior distribution of \( \mu_i|Y_i \) is normal with mean and variance

\[ E(\mu_i|Y_i) = \frac{r \times 0 + n \times \bar{Y}}{r + n} = \frac{1}{r \tau^2 + 1} Y_i = \frac{\tau^2}{1 + \tau^2} Y_i, \]

\[ \text{Var}(\mu_i|Y_i) = \frac{1}{r + n} = \frac{1}{r \tau^2 + 1} = \frac{\tau^2}{1 + \tau^2}. \]
Thus, the Bayes’ estimator of $\mu_i$ under squared error loss is the expectation of its posterior distribution, and is given by

\begin{equation}
E(\mu_i | Y_i) = \left(1 - \frac{1}{1 + \tau^2}\right) Y_i = (1 - B) Y_i,
\end{equation}

where $B = (1 + \tau^2)^{-1}$. However, $\tau^2$ is unknown. Can we “learn” about it from the data? One way of doing this is to integrate out $\mu_i$, and consider the marginal distribution of $Y_i$ given $\tau^2$. By conjugacy, the marginal distribution of $Y_i$ given $\tau^2$ is normal, with respective expectation and variance

\begin{align*}
E(Y_i | \tau^2) &= E\{E(Y_i | \mu_i)\} = E(\mu_i) = 0, \\
\text{Var}(Y_i | \tau^2) &= E\{\text{Var}(Y_i | \mu_i)\} + \text{Var}\{E(Y_i | \mu_i)\} = 1 + \text{Var}(\mu_i) = 1 + \tau^2.
\end{align*}

Thus, $Y_i | \tau^2 \sim \mathcal{N}(0, 1 + \tau^2)$. The likelihood function of $\tau^2$ can therefore be written as

\begin{equation}
L(\tau^2) = \frac{1}{\{2\pi(1 + \tau^2)\}^{K/2}} \exp\left\{-\frac{1}{2(1 + \tau^2)} \sum_{i=1}^{K} Y_i^2\right\}.
\end{equation}

Thus, $S = \sum_{i=1}^{n} Y_i^2$ is CSS for $\tau^2$. Because $Y_i | \tau^2 \sim \mathcal{N}(0, 1 + \tau^2)$, it follows that

\begin{align*}
Y_i^2 &\sim \chi_1^2, \\
\Rightarrow \sum_{i=1}^{K} Y_i^2 &\sim \chi_K^2, \\
\Rightarrow E_{\tau^2} \left\{ \frac{S}{1 + \tau^2} \right\} &= K, \\
\Rightarrow E_{\tau^2} \left\{ \frac{S}{K} \right\} &= 1 + \tau^2.
\end{align*}

Thus, $S/K - 1$ is the UMVUE of $\tau^2$. A plug-in estimator for $B = (1 + \tau^2)^{-1}$ in (10.18) is thus $\hat{B}_{\text{plug-in}} = k/S$, which yields the following plug-in empirical Bayes’ estimator of $\mu_i$:

\begin{equation}
\hat{\mu}_{\text{plug-in}} = (1 - \hat{B}_{\text{plug-in}}) Y_i = \left(1 - \frac{k}{S}\right) Y_i.
\end{equation}

However, $\hat{B}_{\text{plug-in}}$ is a biased estimator of $B$, since

\begin{align*}
E_{\tau^2}(\hat{B}_{\text{plug-in}}) &= E_{\tau^2}(K/S) = KE(1/S) \\
&\geq K \frac{1}{E_{\tau^2}(S)} \text{ by Jensen’s inequality, since } 1/S \text{ is convex} \\
&= \frac{K}{K(1 + \tau^2)} \text{ by (10.20)} \\
&= \frac{1}{1 + \tau^2} = B.
\end{align*}
To find an unbiased estimator of $B$, note that $(1 + \tau^2)/S \sim 1/\chi^2_K$. Since the expectation of $1/\chi^2_\nu$ is $1/(\nu - 2)$, it follows that

$$E_{\tau^2} \left\{ \frac{1 + \tau^2}{S} \right\} = \frac{1}{K - 2},$$

$$\Rightarrow E_{\tau^2} \left\{ \frac{K - 2}{S} \right\} = \frac{1}{1 + \tau^2} = B.$$

This implies that

$$\hat{B}^{\text{UMVUE}} = \frac{K - 2}{S}$$

is the UMVUE of $B$. Substituting $\hat{B}^{\text{UMVUE}}$ in (10.18), the empirical Bayes’ estimator of $\mu_i$ is

$$(10.22) \quad \hat{\mu}_{i}^{\text{UMVUE}} = \left(1 - \hat{B}^{\text{UMVUE}}\right)Y_i = \left(1 - \frac{K - 2}{S}\right)Y_i,$$

which is the same as the JS-estimator.

Another possibility is to use the MLE of $B$ instead of $\hat{B}^{\text{plug-in}}$ or $\hat{B}^{\text{UMVUE}}$. Note that the likelihood function in (10.19) can be written as

$$L(B) \propto B^{K/2} \exp \left\{ -\frac{1}{2} SB \right\} \Rightarrow l(B) \propto \frac{K}{2} \log B - \frac{1}{2} SB \Rightarrow l'(B) = \frac{K}{2B} - \frac{S}{2},$$

equating which to zero, we obtain $\hat{B}^{\text{MLE}} = \hat{B}^{\text{plug-in}} = K/S$.

**Minimizing risk in the class of estimators** $(c/S)Y_i$.

Recall that the UMVUE for the shrinkage factor $B$ for $\mu_i$ from observed $Y_i$’s for $i = 1, 2, \ldots, K$ from the two-level Normal model is $\hat{B} = \frac{(k-2)}{S}$ where $S = \sum_{i=1}^{K} Y_i^2$. Now, the estimator for $B$ can be changed to $\hat{B} = \frac{c}{S}$ where $c > 0$ which is more general than the former assumption of $\hat{B} = \frac{(k-2)}{S}$ for shrinkage factor. Let

$$\tilde{\mu}_c = \left(1 - \frac{c}{S}\right)Y, \text{ i.e.}$$

$$\hat{\mu}_{i,c} = \left(1 - \frac{c}{S}\right)Y_i, \quad i = 1, \ldots, K,$$

be an estimator obtained by substituting $c/S$ as an estimator of $B$ in (10.18).

The risk of this estimator is given by

$$R(\hat{\mu}_c, \mu) = E_\mu \left[ \sum_{i=1}^{K} \left\{ \left(1 - \frac{c}{S}\right)Y_i - \mu_i \right\}^2 \right]$$

$$= E_\mu \left[ \sum_{i=1}^{K} \left\{ (Y_i - \mu_i) - \frac{c}{S} Y_i \right\}^2 \right]$$
\[ E\mu \left[ \sum_{i=1}^{K} (Y_i - \mu_i)^2 \right] + c^2 E\mu \left( \frac{\sum_{i=1}^{K} Y_i^2}{S^2} \right) - 2c E\mu \left\{ \sum_{i=1}^{K} (Y_i - \mu_i) \frac{Y_i}{S} \right\} \]

\[ = K + c^2 E\mu \left( \frac{1}{S} \right) - 2c(K-2) E\mu \left( \frac{1}{S} \right), \]

where the last step is obtained by proceeding along similar lines of the proof of Stein’s theorem. Equating the first derivative of the risk with respect to \( c \) to zero, we obtain the optimal choice of \( c \) as \( K - 2 \), which is the JS-estimator.

In one-level models if we observe one observation \((K = 1)\), regression towards the mean is not possible unless we have a mean to regress toward that. If there are a group of individuals, we can find the mean for all individuals, and thus regression toward the mean is a natural thing in two-level models where we have a distribution for individual means \( \mu_i \) as well. To further illustrate this fact suppose we have a two-level model for \( Y_1, Y_2, ..., Y_k \) as:

\[ Y_i|\mu_i \sim N(\mu_i, V) \]
\[ \mu_i \sim N(\mu_0, A = \tau^2) \]

Maximum likelihood is a natural approach to estimate the shrinkage estimator \( B \). With a Bayes calculation, regression towards the mean will be \( \mu_i|Y_i, V, A, \mu_0 \sim N[(1-B)Y_i + B\mu_0, (1-B)V] \) with positive \( B = \frac{V}{V+A} \). The shrinkage mean can be written as \( \tilde{\mu}_i = Y_i - B(Y_i - \mu_0) \), showing that if \( Y_i \) exceeds mean \( \mu_0 \), the estimator will be reduced by \( B(Y_i - \mu_0) \) and if \( Y_i \) is less than \( \mu_0 \), then the estimator will be increased by \( B(\mu_0 - Y_i) \).

**Multilevel modeling provides a natural way to think about and derive shrinkage estimation.**

As an extension, suppose the loss function is ordinary squared loss and \( V_i \)s depend on \( i \), because \( V_i = \frac{\sigma^2}{n_i} \). If large number of observed data are available for \( Y_i \) (\( n_i \) is big), there is no need to regress any more. This is also clear from the fact that for larger \( n_i \)’s, \( V_i = \frac{\sigma^2}{n_i} \) will be smaller and consequently \( B_i = \frac{V_i}{V_i+A} \) will be smaller.

For a further extension, suppose variances are unequal as before but the loss function is no longer an ordinary sum of squares and it is weighted by \( V_i \) meaningly \( Loss = \sum_i \frac{(\hat{\mu}_i - \mu_i)^2}{V_i} \). In other words, if you have more data, then \( V_i \) is small and we should have a better estimation of \( \mu_i \). This new loss function leads to an interesting risk function. What is the expected square error if \( \hat{\mu}_i = Y_i \)? This is an unbiased estimate and it means no shrinkage. The risk for this estimator will be

\[ R(\mu, \hat{\mu} = \bar{Y}) = E \sum_i \frac{(\hat{\mu}_i - \mu_i)^2}{V_i} = \sum_i \frac{V_i}{V_i} = K. \]
The question is whether you can do better than this. The answer is: SURE, with proper changes in scale and location of the canonical case with equal variances and shrinkage to 0.

In this case estimators for each \( \mu_i \) are going to shrink toward \( \mu_0 \) for each different coordinate. Here we have to set \( S = \sum_i \frac{(Y_i - \mu_0)^2}{V_i} \) and elements are divided by \( V_i \) since the variances are not equal. The final answer will look similar to before, with the difference that \( V_i \)s are incorporated in \( S \), and \( \hat{B} = \frac{k-2}{k} \) (which is the good news). The bad news is we can not definitely divide \( Y_i \)’s by \( \sqrt{V_i} \). For instance, if we are evaluating a hospital, the administrator should agree with it and this will depend on costs and etc for them as a price of this estimation, this procedure is weighting the errors in some statistical manner that might not be consistent with other incentives in the problem.

If the \( \mu_0 \) values are unknown but you now that they are equal, you can take the average of samples and estimate it by its mean as a weighted average which is the most natural approach. In case where only some samples are available, we will treat it as a random variable and estimate it by its mean. The same happens in case of equal variance, only keep in mind that one more regression coefficient is added because we estimate one more constant and thus one degree of freedom is lost. So, in the case of equal and unknown variances, the \( K - 2 \) will drop to \( K - 3 \) in \( \hat{B} \). There are more general approaches such as unequal variances and \( B_i \)’s depending on \( i \). If \( \mu_0 \) and \( A \)s are not known, marginal means and variances of \( Y_i \)s can be used to estimate them with \( Y_i \sim [\mu_0, V_i + A] \). However, there might be technical problems for some types of data, e.g. ML estimate for \( A \) can be negative.

Empirical Bayes was named and studied by [Robbins (1955)]. He proposed the following amazing (but not amazingly good!) estimator for a Poisson model. Suppose for \( i = 1, 2, \ldots, K \) we have:

\[
Y_i|\mu_i \sim \text{Pois}(\mu_i) \\
\mu_i \sim CDF \ G(\mu_0)
\]

Robbins assumed that \( K \to \infty \) and \( \mu_i \) is assumed to be i.i.d from an arbitrary CDF (with second moment and smoothness). If we had \( G(\mu) \), we would find \( E\mu_i|Y_i \) by integrating the following:

\[
E\mu_i|Y_i = \frac{\int_0^\infty \mu_i Y_i e^{-\mu_i} dG(\mu_i)}{\int_0^\infty Y_i e^{-\mu_i} dG(\mu_i)}
\]

Robbins multiplied and divided by \( (1 + Y_i) \) and wrote the expectation in terms of the marginal density of \( Y_i \):

\[
E\mu_i|Y_i = \frac{(1 + Y_i) \int_0^\infty \mu_i^{1+Y_i} e^{-\mu_i} dG(\mu_i)}{\int_0^\infty \mu_i^{Y_i} e^{-\mu_i} dG(\mu_i)} = \frac{(1 + Y_i) M(Y_i + 1)}{M(Y_i)}
\]

where \( M(Y_i) \) is the marginal density for \( Y_i \). Thus for an arbitrary prior, the estimation can be represented as the ratios of marginal densities and Robbins
called this empirical Bayes because the $Y_1, \ldots, Y_k$ are given and $k \to \infty$. Hence, these marginal densities can be estimated and as the samples are i.i.d. on the marginal density, with enough data, the marginal density can be learned. $\mu_i$'s are exchangeable since $G(\mu_0)$ is the same for all $\mu_i$'s. In the Stein case, all $\mu_i$'s had the same distribution and that was the reason we could learn something (exchangeability is a requirement). In both J-S and the Robbins case, the $Y_i$'s are not exchangeable. However, the $(Y_i, \mu_i)$ are pairwise exchangeable and they are i.i.d. samples from a two dimensional distribution, and this fact makes them suitable for the J-S estimator. In the Robbins case also $(Y_i, \mu_i)$ will be i.i.d. and independent of $i$ and it helps in estimating marginal densities.

However, the bad news is that there is a big difference between the asymptotic approximations case with $K \to \infty$ and the infinite sample case and this works only for the infinite sample case but not in $K \to \infty$.

To be more clear, we choose a prior 50%, 50% mixture of two different Gamma distributions and assuming that $K = 20$, we get 20 samples $Y_i$. Now let us see what happens with the Robbins estimator. From the 20 samples we know that $Y_{\text{max}} = Y_j = 11$ which means the largest observed $Y$ is $Y_j = 11$. Then

$$E\mu_j | Y_j = \frac{(1 + Y_j) \cdot 0}{1} = 0.$$  

This says that the hospital in which the most people died is the safest!

In practice, there will be a maximum $Y_j$ and we will estimate mean zero in the extreme case, which is why the estimator does not behave monotonically at the tails and rather has zigzag behavior.

### 11. EM Algorithm

Our goal is to find the MLE $\hat{\theta}$ from the likelihood $L(\theta | y) = g(y | \theta)$. We call this the “observed data” likelihood because $y$ is actually observed. This likelihood may be very complicated here making computation intractable. EM (Expectation Maximization) is an idea which suggests introducing a latent variable (augmenting the data) in order to simplify the likelihood. Suppose that if we had also observed the data $z$, then the complete data likelihood $L_{\text{comp}}(\theta | y, z) = f(y, z | \theta)$ may be much easier to work with in trying to obtain the MLE $\hat{\theta}$. Note that there are infinitely many proposals for what we might have wished to observe i.e. many choices for an expanded model. All these complete data models $(y, z) | \theta$ are “correct” as long as we have

$$g(y | \theta) = \int f(y, z | \theta) dz$$

That is to say that we can propose any complete data model for which marginalization over $z$ yields the observed data likelihood. However, some choices of latent variables may make the likelihood even more complicated. The “art” of EM is to introduce a latent variable in the way which simplifies computation the most. The following example illustrates this point.
Example. (World Series)
Suppose that teams A and B are playing in the World Series. Let $P(\text{Team A wins}) = p$. Games are considered independent and the winner of the series is the team who wins the best out of 7 games. If $N$ is the number of games played, then $N = \min(W_A, W_B)$ where $W_A, W_B \sim \text{NBin}(4, p)$ and are dependent. Obviously, the likelihood in this case is very complicated. However, we can simplify this model if we simply assume that all 7 games are always played even if a team has managed to win 4 games before then. We pretend that the additional games are played but that their results are not observed. By adding the extra unobserved games, the number of wins reduces to a Binomial model, $\text{Bin}(7, p)$ which is very simple to deal with.

Is this approach valid though? Ask yourself this: if someone told you that MLB had a great conspiracy in which world series teams were forced to play in secret any remaining games after the series had already been won, would you be able to disprove it? No, you wouldn’t be able to disprove it and more importantly, if this conspiracy is actually taking place in the real world, does it affect our inference about $p$?

The key insight is that it is hard to deal with a situation in which the number of games is variable, so we simply augment the data by adding extra games whose results are not observed. The augmented data then have a very nice likelihood.

\section{11.1.} Suppose that 5 games were played in which Team A won 4 out of the 5 games. Given $Z$, the number of wins by Team A in games 6 and games 7, what is the complete data likelihood? Having obtained the complete data likelihood, marginalize over $z$ to get the observed data likelihood.

It is nice to have simplified the likelihood by introducing a latent variable $Z$, however the fact still remains that we have not observed $Z$. Now, if we knew $\theta$, then we could estimate $Z$ by $E(Z|y, \theta)$. So let us assume that we knew $\theta$, say $\theta^{(0)}$. We initially estimate $Z$ by $Z^{(0)} = E(Z|y, \theta^{(0)})$. By obtaining $Z^{(0)}$, we can then perform MLE on the likelihood data likelihood $f(y, z|\theta)$.

\section{11.1.} Note that we unless $Z$ appears linearly in the complete data likelihood, we usually will not impute $Z^{(0)} = E(Z|y, \theta^{(0)})$. For example if $Z$ appears in $f(y, z|\theta)$ as $h(Z)$. Then we will impute $E(h(Z)|y, \theta^{(0)})$ rather than using $h(E(Z|y, \theta^{(0)}))$. EM is smart in this way in that we will directly impute for the function of $Z$ which appears in the likelihood.

Below, we will use the identity

$$f(y, z|\theta) = f(z|y, \theta) f(y|\theta) \Rightarrow f(y|\theta) = \frac{f(y, z|\theta)}{f(z|y, \theta)}$$

which relates the complete data likelihood $f(y, z|\theta)$ to the missing data likelihood $f(y|\theta)$. Taking logs gives

$$l(\theta|y) = l(\theta|y, z) - \log f(z|y, \theta)$$
Note that we cannot observe the complete data likelihood (i.e., we can’t obtain the curve). By formulating a missing data structure, however, we impose a relation between the missing values \( z_i \) and the observed values \( y_i \) and the parameters \( \theta \). Having observed \( y_i \), we will then wish to incorporate information about the \( z_i \) available in \( y_i \). The only reason \( y_i \) contains information about \( z_i \) is because of the formulated missing data structure. The method by which this information is incorporated is to take the expectation of this the above identity over \( f(z|y, \theta^{(t)}) \), for some initial value of \( \theta^{(t)} \).

\[
l(\theta|y) = E_{\theta^{(t)}} [l(\theta|y, z) | y] - E_{\theta^{(t)}} [f(z|y, \theta) | y]
\]

This constitutes the expectation step of the EM concept. Thus we see that the expectation step is simply an incorporation of information about the missing data contained in the observed data. This incorporation depends on our current guess of \( \theta \) and as we update \( \theta \), the incorporation should also change.

§ 11.2. Consider the case where \( Z \), the missing data, and \( Y \), the observed data are independent. How is there information about the missing data in the observed data when the two are independent?

Incorporating the information in the observed data, we obtain the expected log likelihood (in the sense that given the current information about the missing values \( z_i \) contained in the observed values \( y_i \) and our current guess of \( \theta \), \( \theta^{(t)} \), we would expect the complete data likelihood to be this)

\[
Q(\theta|\theta^{(t)}, y) \equiv E_{\theta^{(t)}} [l(\theta|y, z) | y]
\]

Keeping fixed \( \theta^{(t)} \), we then maximize \( Q(\theta|\theta^{(t)}, y) \) over \( \theta \) and call this value \( \theta^{(t+1)} \). Here we are encoding information in the observed data, \( y_i \), into \( \theta^{(t+1)} \), and this information is then incorporated into a best guess for \( Z \), in the E step, when we take the expectation \( E_{\theta^{(t+1)}} [l(\theta|y, z) | y] \). This constitutes the maximization step of the EM algorithm. The E and M steps are partially motivated by the following theorem which guarantees that the observed likelihood never decreases with successive iterations.

**Theorem 11.1.** The sequence \( \theta^{(t)} \) produced by the EM algorithm satisfies

\[
l(\theta^{(t+1)}|y) \geq l(\theta^{(t)}|y)
\]

**Proof.** Since \( l(\theta|y) = E_{\theta^{(t)}} [l(\theta|y, z) | y] - E_{\theta^{(t)}} [\log f(z|y, \theta) | y] \), it suffices to show the following two statements

\[
E_{\theta^{(t)}} [l(\theta^{(t+1)}|y, z) | y] \geq E_{\theta^{(t)}} [l(\theta^{(t)}|y, z) | y]
\]

\[
E_{\theta^{(t)}} [\log f(z|y, \theta^{(t+1)}) | y] \leq E_{\theta^{(t)}} [\log f(z|y, \theta^{(t)}) | y]
\]

We chose \( \theta^{(t+1)} = \text{arg max}_{\theta} E_{\theta^{(t)}} [l(\theta|y, z) | y] \). Then by construction, we have

\[
E_{\theta^{(t)}} [l(\theta^{(t+1)}|y, z) | y] \geq E_{\theta^{(t)}} [l(\theta^{(t)}|y, z) | y]
\]
To obtain the second inequality, we use Jensen’s inequality
\[
E_{\theta(t)} \left[ \log \frac{f(z|y,\theta^{(t+1)})}{f(z|y,\theta^{(t)})} \mid y \right] \leq \log E_{\theta(t)} \left[ \frac{f(z|y,\theta^{(t+1)})}{f(z|y,\theta^{(t)})} \mid y \right]
\]
\[
= \log \int \frac{f(z|y,\theta^{(t+1)})}{f(z|y,\theta^{(t)})} f \left( z \mid y, \theta^{(t)} \right) \, dz
\]
\[
= \log \int f \left( z \mid y, \theta^{(t+1)} \right) \, dz = \log 1
\]
\[
= 0
\]
Thus
\[
E_{\theta(t)} \left[ \log f(z|y,\theta^{(t+1)}) \mid y \right] \leq E_{\theta(t)} \left[ \log f(z|y,\theta^{(t)}) \mid y \right].
\]
Putting this with the first inequality gives the desired statement.

\[\square\]

**Example 11.1 (Censored data).** Suppose that we observe \(Y_1, \ldots, Y_n\) i.i.d. from \(f(y - \theta)\) and we have ordered the observations so that \(y_1, \ldots, y_m\) are uncensored and \((y_{m+1}, \ldots, y_n)\) are censored. The likelihood function is
\[
L(\theta \mid y) = [1 - F(a - \theta)]^{n-m} \prod_{i=1}^{m} f(y_i - \theta)
\]
where \(F\) is the cdf associated with \(f\). Suppose that we had observed all the censored data say \(z = (z_{m+1}, \ldots, z_n)\) with \(z_i > a\). The complete data likelihood would then be
\[
L(\theta \mid y, z) = \prod_{i=1}^{m} f(y_i - \theta) \prod_{i=m+1}^{n} f(z_i - \theta).
\]
The observed data likelihood is obtained from the complete data likelihood by integrating over all \(z_i\) for which \(z_i > a\)
\[
L(\theta \mid y) = \mathbb{E} \left[ L(\theta \mid y, z) \right] = \int_{z} L(\theta \mid y, z) \, dz = \prod_{i=1}^{m} f(y_i - \theta) \prod_{i=m+1}^{n} \int_{a}^{\infty} f(z_i - \theta) \, dz_i
\]
Now suppose that \(Y_i \sim N(\theta, 1)\). The complete data likelihood is then
\[
\prod_{i=1}^{m} \exp \left( -\frac{(y_i - \theta)^2}{2} \right) \prod_{i=m+1}^{n} \exp \left( -\frac{(z_i - \theta)^2}{2} \right)
\]
Note that each \(Z_i\) is a truncated Normal with support \([a, \infty)\). Thus we have
\[
f(z \mid \theta, y) = \frac{1}{1 - \Phi(a - \theta)} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(z - \theta)^2}{2} \right)
\]
Suppose that \(Z_i \sim \mathcal{N}(0, 1)\). Then it can be shown that \(\mathbb{E}(Z_i | Z_i \geq 0) = \frac{\phi(0)}{1 - \Phi(0)}\) (this is a Stat 210 problem). If \(Z_i \sim \mathcal{N}(\theta, 1)\), then it can be generalized that \(\mathbb{E}(Z_i | Z_i \geq a) = \theta + \frac{\phi(a - \theta)}{1 - \Phi(a - \theta)}\). Thus the expected complete data
The log likelihood is
\[
-\frac{1}{2} \sum_{i=1}^{m} (y_i - \theta)^2 - \frac{1}{2} \sum_{i=m+1}^{n} E_{\theta(t)} (Z_i - \theta)^2
\]

Maximizing this over \( \theta \) gives
\[
\sum_{i=1}^{m} (y_i - \theta) + \sum_{i=m+1}^{n} (E_{\theta(t)} (Z_i) - \theta) = 0 \Rightarrow \theta^{(t+1)} = \frac{m}{n} \bar{y} + \frac{n - m}{n} E_{\theta(t)} [Z_i]
\]

Then using the result from above
\[
\theta^{(t+1)} = \frac{m}{n} \bar{y} + \frac{n - m}{n} \left[ \theta^{(t)} + \frac{\phi(a - \theta^{(t)})}{1 - \Phi(a - \theta)} \right]
\]

11.3. Use Stein’s SURE lemma to show that \( E(Z_i | Z_i \geq a) = \theta + \frac{\phi(a - \theta)}{1 - \Phi(a - \theta)} \) if \( Z \sim \mathcal{N}(\theta, 1) \). Note that the derivative of an indicator function is a Dirac delta function.

Example 11.2 (Normal Mixture Model). Consider a 50-50 mixture of two Normals with means \( \mu_1, \mu_2 \) and variances \( \sigma_1^2, \sigma_2^2 \). Assume \( \sigma_1^2 = \sigma_2^2 = 1 \) for simplicity. How do we use EM to find the MLE for the means?

11.4. If \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown, show that the likelihood function is still bounded when \( \sigma_1^2 = \sigma_2^2 \), and it is unbounded when \( \sigma_1^2 \neq \sigma_2^2 \).

The likelihood function for one observation is:
\[
L(\mu_0, \mu_1 | y_j) = \frac{1}{2} \phi(y_j - \mu_0) + \frac{1}{2} \phi(y_j - \mu_1),
\]
where \( \phi(.) \) is the probability density function of the Normal distribution.

If we did not have defined groups, then the latent group membership can be viewed as missing data.

We introduce a “latent” variable \( z_j \) (binary), which is an indicator of which group \( y_j \) belongs to. Then, if we observed both \( y_j \) and \( z_j \), the likelihood function becomes
\[
L(\mu | y_j, z_j) = \frac{1}{2} \phi(y_j - \mu z_j)
\]
or
\[
= \frac{1}{2} \phi(y_j - \mu_0)^{1-z_j} \phi(y_j - \mu_1)^{z_j}
\]

The log-likelihood function in this case is:
\[
l(\mu | y, z) = -\frac{1}{2} \sum_{j} (1 - z_j)(y_j - \mu_0)^2 - \frac{1}{2} \sum_{j} z_j(y_j - \mu_1)^2
\]

We wish we could use the log-likelihood function above to find MLE, but it assumes that we know everything and we do not actually know \( z \). EM says: take the expectation of the log-likelihood, and maximize it.
Remark 11.1. Why do we use $E[\log\text{-likelihood}]$ rather than $E[\text{likelihood}]$? By Jensen’s inequality, we get the nice property that the log-likelihood increases at each iteration. Also, in general, it is easier to work with sums than products.

So we take $Q(\mu|\mu^{(t)})$, and maximize it, where

$$Q(\mu|\mu^{(t)}) = -\frac{1}{2} \sum_j (1 - E^{(t)} z_j)(y_j - \mu_0)^2 - \frac{1}{2} \sum_j E^{(t)} z_j(y_j - \mu_1)^2.$$ 

\section*{11.5.} To maximize $Q(\mu|\mu^{(t)})$, all we need to do is to find $\mu$ that maximizes $E^{(t)} z_j$ conditioning on $\mu^{(t)}$ and $y$. Why?

$$E^{(t)} z_j = P(z_j = 1|y_j, \mu^{(t)}) = \frac{1}{2} \phi(y_j - \mu_1^{(t)}) + \frac{1}{2} \phi(y_j - \mu_0^{(t)})$$

Note that we could treat 1 as a tiny interval around 1 here, so that we can treat $P(z_j = 1|y_j, \mu^{(t)})$ as a probability rather than as a density.

$$\mu^{(t+1)} = \frac{\sum_j (E^{(t)} z_j) y_j}{\sum_i E^{(t)} z_i},$$

and we can do the same for $\mu_0^{(t+1)}$.

\begin{example}[11.3 (EM in an EF)] Suppose that the complete data density $f$ is of the form

$$f(x, z|\theta) = h(x, z) \exp \{ \theta^T (x, z) - \psi(\theta) \}$$

The expected complete data log likelihood is then

$$E_{\theta^{(t)}} [l(\theta|x, z)|x] = E_{\theta^{(t)}} [\log(h(x, z))|x] + E_{\theta^{(t)}} [\theta^T (x, z)|x] - \psi(\theta)$$

Now we maximize this over $\theta$, note that $\psi'(\theta)$ returns a vector of the means of the sufficient statistics

$$\psi'(\theta^{(t+1)}) = E_{\theta^{(t)}} [T(x, z)|x]$$

So for example we solve for

$$\theta^{(t+1)}_i : \psi_i'(\theta^{(t+1)}) = E_{\theta^{(t)}} [T_i|x_i] \text{ for all } i$$

Thus in exponential families, EM imputes the natural sufficient statistics $T_i$ and sets it equal to its mean written in terms of current guess of $\theta$, $\theta^{(t)}$.

\begin{example}[11.4 (Probit Model)] The probit model specifies

$$Y|X = x \sim \text{Bern} \left( \Phi \left( x^T \beta \right) \right)$$


The likelihood associated with the sample \((x_1, y_1), \ldots, (x_n, y_n)\) is

\[
L(\beta|x, y) = \prod_{i=1}^{n} f(y_i|x_i, \beta) = \prod_{i=1}^{n} \Phi(x_i' \beta)^{y_i} (1 - \Phi(x_i' \beta))^{1-y_i}
\]

Now we associate with each observation \((x_i, y_i)\) a missing variable \(Z_i\) such that

\[
Z_i|X_i = x \sim N(x_i' \beta, 1) \quad \text{and} \quad Y_i = I(Z_i > 0)
\]

Now the complete data likelihood is

\[
L(\beta|x, y, z) = \prod_{i=1}^{n} f(y, z|x, \beta) = \prod_{i=1}^{n} f(y|z, x, \beta) f(z|x, \beta) = \prod_{i=1}^{n} f(z|x, \beta)
\]

\[
= \exp \left( -\sum_{i=1}^{n} \frac{(z_i - x_i' \beta)^2}{2} \right)
\]

The expected complete data log likelihood is

\[
l(\beta|x, y) = -\sum_{i=1}^{n} E_{\beta(t)} \left[ \frac{(z_i - x_i' \beta)^2}{2} | y, x \right]
\]

Maximizing this over \(\beta\) yields

\[
\beta(t+1) = (X'X)^{-1} X' E_{\beta(t)} [Z|x, y]
\]

where \(X = (x_1, \ldots, x_n)\). From above, we know that if \(Z \sim N(\theta, 1)\), then

\[
E(Z|Z \geq a) = \theta + \frac{\phi(a-\theta)}{1 - \Phi(a-\theta)}.
\]

Using symmetry properties, it can also be shown that \(E(Z|Z \leq a) = \theta - \frac{\phi(\theta-a)}{1 - \Phi(\theta-a)}\). Here \(Z_i|X_i = x \sim N(x_i' \beta, 1)\) and \(Y_i = I(Z_i > 0)\). Thus

\[
E_{\beta(t)} [Z|x, y = 1] = x' \beta + \frac{\phi(-x' \beta)}{1 - \Phi(-x' \beta)}
\]

\[
E_{\beta(t)} [Z|x, y = 0] = x' \beta - \frac{\phi(x' \beta)}{1 - \Phi(x' \beta)}
\]

REFERENCES


