STAT 460/560 + 461/561
STATISTICAL INFERENCE I & II

2017/2018, TERMs I & II

Jiahua Chen and Ruben Zamar

Department of Statistics

University of British Columbia

Contents

Chapter 1

Some basics

1.1 Discipline of Statistics

Statistics is a discipline that serves other scientific disciplines. Statistics is itself may not considered by many as a branch of science. A scientific discipline constantly develops theories to describe how the nature works. These theories are falsified whenever their prediction contradicts the observations. Based on these theories and hypotheses, scientists form a model for the natural world and the model is then utilized to predict what happens to the nature under new circumstances. Scientific experiments are constantly designed find evidences that may contradict the prediction of the proposed model and aim at DISPROVING hypotheses behind the model/theory. If a theory is able to make useful predictions and we fail to find contradicting evidences, it gains broad acceptance. We may then temporarily consider it as “the truth”. Even if a model/theory does not give a perfect prediction, but a prediction precise enough for practical purposes and it is much simpler than a more precise model/theory, we tend to retain it as a working model. I regard, for example, Newton's laws as such an example as compared to more elaborating Einstein's relativity.

If a theory does not provide any prediction that can potentially be disproved by some experiments, then it is not a scientific theory. Religious theories form a rich group of such examples.

Statistics in a way is a branch of mathematics. It does not model our nature. For example, it does not claim that when a fair die is rolled, the probability of observing 1 is 1/6. Rather, for example, it claims that if the probability of observing 1 is 1/6, and if the outcomes of two dice are independent, then the probability of observing $(1, 1)$ is 1/36, and the probability of observing either $(1, 2)$ and $(2, 1)$ is 2/36. If one applies a similar model to the spacial distribution of two electrons, the experimental outcomes may contradict the prediction of this probability model, yet the contradiction does not imply that the statistic theory is wrong. Rather, it implies that the statistical model does not apply to the distribution of the electrons. The moral of this example is, a statistical theory cannot be disproved by physical experiments. Its theories are of logical truth, and this makes it unqualified as a scientific discipline in the sense we mentioned earlier.

We should make a distinction of the inconsistency between a probability model and the real world, and the inconsistency within our logical derivations. If we err at proving a proposition, that proposition is very likely false within our logical system. It does not disprove the logical system. We call logically proved propositions as theorems. In comparison, the propositions regarded as temporary truth in science are named as laws. Of course, we sometimes abuse these terminologies such as “Law of Large Numbers”.

In a scientific investigation, one may not always be able to find clear-cut evidence against a hypothesis. For instance, genetic theory indicates that tall fathers have tall sons in general. Yet there are many factors behind the height of the son. Suppose we collect 1000 father-son pairs randomly from a human population. Let us measure their heights as (x_i, y_i), $i = 1, 2, \ldots, 1000$. A regression model in the form of

$$y_i = a + bx_i + \epsilon_i$$

with some regression coefficient (a, b) and random error ϵ, can be a useful summary of the data.

If the statistical analysis of the data supports the model with some $b > 0$, then the genetic theory survives the attack. If we have a strong evidence to suggest b is not very different from 0, or it may even be negative, then the genetic theory has to be abandoned. In this case, the genetic theory is not disproved by statistics, but by physical experiments (data collected on father-son heights) assisted by the statistical analysis. Whatever the outcome of the statistical analysis is, the statistic theory is not falsified. It is the genetic theory that is being tortured.

1.2 Probability and Statistics models

In scientific investigations, we often quantify the outcomes of an experiment in order to develop a useful model for the real world. An existing scientific theory can often give a precise prediction: the water boils at 100 degrees Celsius at the sea level on the Earth. In other cases, precise prediction is nearly impossible. For example, scientists still cannot predict when and where the next serious earthquake will be. There used to be beliefs that a yet to be discovered perfect scientific model exists which can explain away all randomness. In terms of earthquakes, it might be possible to have a precise prediction if we know the exact tensions between the geographic structures all around the world, the amount of heat being generated at the core of the earth, the positions of all heavenly bodies and a lot more.

In other words, the claim is that we study randomness only because we are incompetent in science or because a perfect model is too complicated to be practically useful. This is now believed not the case. The uncertainty principle in quantum theory indicates that the randomness might be more fundamental than many of us are willing to accept. It strongly justifies the study of statistics as an “academic discipline”.

A probability space is generally denoted as $(\Omega, \mathbb{B}, P)$. We call Ω the sample space, which is linked to all possible outcomes of an experiment under consideration. The notion of experiment becomes rough when the real world problem becomes complex. It is better off to take the mathematical convention to simply assume its existence. $\mathbb{B}$ is a σ-algebra. Mathematically, it stands for a collection of subsets of Ω with some desirable properties. We require that it is possible to assign a probability to each subset of Ω that is a member of $\mathbb{B}$ without violating some desired rules. How large a probability is assigned to a particular member of $\mathbb{B}$ is a rule denoted by P.

A random variable (vector) X is a measurable function on Ω. It takes values on $\mathcal{R}^n$ if X has length n. It induces a probability space $(\mathcal{R}^n, \mathbb{B}, F)$ where F is its distribution. In statistics, we consider problems of inferring about F within a set of distributions pre-specified. This set of distributions is called **statistical model**, and it is presented as a probability distribution family $\mathcal{F}$ sometime with additional structures. If vector X has n components and they are independent and identically distributed (i.i.d.), we use $\mathcal{F}$ for individual distribution, not for the joint distribution. This convention will be clear when we work with specific problems. In this case, we call it *population* F defined on $(\mathcal{R}, \mathbb{B})$. Components of X are samples from population F.

When the individual probability distributions in $\mathcal{F}$ is conveniently labelled by a subset of R^d, the Euclid space of dimension d, we say that $\mathcal{F}$ is a parametric distribution family. The label is often denoted as θ, and its all possible values Θ is called parameter space. In applications, we usually only consider parametric models whose probability distributions have a density function with respect to a common σ-finite measure. In such situations, we write

$$\mathcal{F} = \{f(x; \theta) : \theta \in \Theta\}.$$

The σ-finite measure is usually the Lebesgue which makes $f(x; \theta)$ the commonly referred density functions. When the σ-finite measure is the counting measure, the density functions are known as probability mass function.

If $\mathcal{F}$ is not parameterized, we have a non-parametric model.

Probability theory and statistics Probability theory studies the properties of stochastic systems. For instance, the convergence property of the empirical distribution based on an i.i.d. sample. Statistical theory aims at inferring about the stochastic system based on (often) an i.i.d. sample from this system. For instance, does the system (population) appear to be a mixture of two more homogeneous subpopulations? Probability theory is the foundation of statistical inference.

Given an inference goal, statisticians may propose many possible approaches. Some approaches may deem inferior and dismissed over the time. Most approaches have merits that are not completely shadowed by other approaches. Some statistical techniques used as standard methods in other disciplines yet most statisticians never heard of. As a statistician, I hope to have the knowledge to understand these approaches, not to have the knowledge of all statistical approaches.

1.3 Statistical inference

Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from a statistical model $\mathcal{F}$. That is, we assume that they are independent and identically distributed with a distribution which is a member of $\mathcal{F}$. Let their realized values be $x = (x_1, x_2, \ldots, x_n)$. A statistical inference is to infer about the specific member F of $\mathcal{F}$ based on the realized value x. If we take a single guess of F, the result is a point estimate; If we provide a collection of possible F, the result is an interval estimate (usually); If we make a judgement on whether a single or a subset of $\mathcal{F}$ contains the "true" distribution, the procedure is called hypothesis test. In general, in the last case, we are required to quantify the strength of the evidence based on which the judgement is made. If we partition the space of $\mathcal{F}$ into several submodels and infer which submodel F belongs, the procedure is called model selection. In general, for model selection, we do not quantify the evidence favouring the specific submodel. This is the difference between "hypothesis test" and "model selection".

Another general category of statistical inference is based on Bayesian paradigm. The Baysian approach does not identify any F or any set of F. Instead, it provides a probabilistic judgement on every member of subset of $\mathcal{F}$. The probabilistic judgement is obtained via conditional distribution by placing a prior distribution on $\mathcal{F}$ and conditional on observations in the form of $X = x$. We call it posterior distribution. The final decision will be made based on consideration such as minimizing expected lost.

Definition 1.1. *A statistic is a function of data which does not depend on any unknown parameters.*

The sample mean $\bar{x}_n = n^{-1}(x_1 + x_2 + \cdots + x_n)$ is a statistic. However, $\bar{x}_n - \mathbb{E}(X_1)$ is in general not a statistic because it is a function of both data, $\bar{x}_n$, and the usually unknown value, $\mathbb{E}(X_1)$. The value of $\mathbb{E}(X_1)$ often depends on parameter θ behind $\mathcal{F}$.

Let $T(x)$ be a statistic. We may also regard $T(x)$ as the realized value of T when the realized value of X is x. We may regard $T = T(X)$ as a quantity to be "realized". Since X is random, the outcome of T is also random. The distribution of $T(X)$ is called its sample distribution. Unfortunately, it is often hard to be completely consistent when we deal with $T(X)$ and

$T(x)$. We may have to read between lines to tell which one of the two is under discussion. Since the distribution of X is usually only known up to being a member of $\mathcal{F}$ which is often labeled by a parameter θ, the (sample) distribution of T is also only known up to the unknown parameter θ.

Definition 1.2. *Let $T(x)$ be a statistic. If the conditional distribution of X given T does not depend on unknown parameter values, we say T is a sufficient statistics.*

When T is sufficient, all information contained in X about θ is contained in T. In this case, one may choose to ignore X but work only on T without loss of any efficiency. Such a simplification is most helpful if T is much simpler than X or it is a substantial reduction of X.

Directly verifying the sufficiency of a statistic is often difficult. We generally use factorization theorem to identify sufficient statistics. If the density function of X can be written as

$$f(x;\theta) = h(x)g(T(x);\theta)$$

for some function $h(\cdot)$ and $g(\cdot;\cdot)$, then $T(x)$ is sufficient for θ.

In some situations, direct verification is not too complex. For example, if X_1, X_2 are independent Poisson distributed with mean parameter θ. Then the conditional distribution of X_1, X_2 given $T = X_1 + X_2$ are binomial (T, 1/2) which is free from the unknown parameter θ. Hence, T is sufficient for θ.

Definition 1.3. *Sufficient statistic $T(x)$ is minimum sufficient if T is the function of every other sufficient statistic.*

A minimum sufficient statistic may still contain some redundancy. If a statistic has the property that none of its non-zero function can have identically 0 expectation, this statistic is called complete. When the requirement is reduced to included only "bounded functions", then T is called bounded-complete. We have a few more such notions.

Definition 1.4. *Sufficient statistic $T(x)$ is complete if $\mathbb{E}(g(T)) = 0$ under every $F \in \mathcal{F}$ implies $g(\cdot) \equiv 0$ almost surely.*

In contrast, if the distribution of T does not depend on θ or equivalently on the specific distribution of X, we say that T is an ancillary statistic.

Definition 1.5. *If the distribution of the statistic $T(x)$ does not depend on any parameter values, it is an ancillary statistic.*

Example: Suppose $X = (X_1, \ldots, X_n)$ is a random sample from $N(\theta, 1)$ with $\theta \in R$. Recall that $T = \bar{X}$ is a complete and sufficient statistic of θ. At the same time, $X - T = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$ is an ancillary statistic. It does not contain any information about the value of θ. However, it is not completely useless. Under the normality assumption, $X - T$ is multivariate normal. We can study the realized value of $X - T$ to see whether it looks like a realized value from a multivariate normal. If the conclusion is negative, the normality assumption is in serious question. If the validity of a statistical inference heavily depends on normality, such a diagnostic procedure is very important.

Remark: In this example the probability model $\mathcal{F}$ is all normal distributions with mean θ and known variance $\sigma^2 = 1$. Notationally, $\mathcal{F} = \{N(\theta, 1) : \theta \in \mathcal{R}\}$.

Definition 1.6. *If T is a function of both data X and the parameter θ, but its distribution is not a function of θ, we call T a pivotal quantity.*

In the last example, $S = \bar{X} - \theta$ is a pivotal quantity. Note that this claim is made under the assumption that θ is the "true" parameter value of the distribution of X, it is not a dummy variable. This is another common practice in statistical literature: if not declared, notation θ is used both as a dummy variable and the "true" value of the distribution of the random sample X. This notion also applies to Bayes methods, θ is often regarded as a realized value from its prior distribution, and X is then a sample from the distribution labeled by this "true" value of θ.

Note that the parameter θ is a label of F that belongs to $\mathcal{F}$ in parametric models. It may as well be regarded as a function of F, call it **functional** if you please. Any function of F can be regarded as a parameter by the same token. For example, the median of F is a parameter. This works even if $\mathcal{F}$ is a popularly used parametric distribution family such as Poisson.

Chapter 2

Normal distributions

Let X be a random variable. Namely, it is a function on a probability space $(\Omega, \mathbb{B}, P)$. It randomness is inherited from probability measure P. By definition of random variable,

$$\{X \leq t\} = \{\omega : \omega \in \Omega, X(\omega) \leq t\}$$

is a member of $\mathbb{B}$ for any real value t. Hence, there is a definitive value

$$F_{\mathrm{X}}(t) = P(\{X \leq t\})$$

for any $t \in \mathcal{F}$. We refer $F_x(t)$ as the cumulative distribution function (c.d.f.) of X. Often, we omit the subscript and write it as $F(t)$. Note t itself is a dummy variable so it does not carry any specific meaning other than it stands for a real number. In most practices, we use $F(x)$ for the c.d.f. of X. This can lead to confusion: Once $F(x)$ is used as c.d.f. of X, $F(y)$ remains the c.d.f. of X, not necessarily that of another random variable called Y.

The c.d.f. of a random variable largely determines it randomness properties. This is the basis of forming distribution families: distributions whose c.d.f. having a specific algebraic form. Of course, there are often physical causes behind the algebraic form. For instance, success-failure experiment is behind the binomial distribution family.

Uni- and Multi-variate normal distribution families occupy a special space in the classical mathematical statistics. We provide a quick review as follows.

2.1 Uni- and Multivariate normal

A random variable has standard normal distribution if its density function is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2).$$

We generally use

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

to denote the corresponding c.d.f. . If X has probability density function

$$\phi(x;\mu,\sigma) = \sigma^{-1}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}x^2)$$

then it has normal distribution with mean μ and variance σ^2. We use $\Phi(x;\mu,\sigma)$ to denote the corresponding c.d.f.

If Z has standard normal distribution, then $X = \sigma Z + \mu$ has normal distribution with parameters (μ, σ^2) which represent mean and variance. The moment generating function of X is given by

$$M_x(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$$

which exists for all $t \in \mathcal{R}$. The moment of the standard normal Z are: $\mathbb{E}(Z) = 0$, $\mathbb{E}(Z^2) = 1$, $\mathbb{E}(Z^3) = 0$ and $\mathbb{E}(Z^4) = 3$.

Why is **the normal distribution** normal? The **central** limit theorem tells us that if $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of i.i.d. random variables with $E(X) = 0$ and $\text{VAR}(X) = 1$, then

$$P(n^{-1/2}\sum_{i=1}^{n} X_i \leq x) \to \int_{-\infty}^{x} \phi(t)dt$$

for all x, where $\phi(t)$ is the density function of the standard normal distribution (normal with mean 0 and variance 1).

Recall that many distributions we investigated can be viewed as distributions of sum of i.i.d. random variables, hence, when properly scaled as in the central limit theorem, their distributions are well approximated by normal. These examples include: binomial, Poisson, Negative binomial, Gamma.

In general, if the outcome of a random quantity is influenced by numerous factors and none of them play a determining role, then the sum of their effects is normally distributed. This reasoning is used to support the normality assumption on our "height" distribution, even though none of us ever had a negative height.

Multivariate normal. Let the vector $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_d\}'$ consist of independent, standard normally distributed components. Their joint density function is given by

$$f(\mathbf{z}) = \{2\pi\}^{-d/2} \exp\{-\frac{1}{2}\mathbf{z}^\tau \mathbf{z}\} = \{2\pi\}^{-d/2} \exp\{-\frac{1}{2}\sum_{j=1}^{d} z_i^2\}.$$

Easily, we have $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$ and $\textsc{var}(\mathbf{Z}) = \mathbb{I}_d$, the identity matrix. The moment generating function of $\mathbf{Z}$ (joint one) is given by

$$M_z(\mathbf{t}) = \exp\{\frac{1}{2}\mathbf{t}^\tau \mathbf{t}\}$$

which is in vector form.

Let $\mathbf{B}$ be a matrix of size $m \times d$ and $\boldsymbol{\mu}$ be a vector of length m. Then

$$\mathbf{X} = \mathbf{B}\mathbf{Z} + \boldsymbol{\mu}$$

is multivariate normally distributed with

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \textsc{var}(\mathbf{X}) = \mathbf{B}\mathbf{B}^\tau.$$

We will use notation $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\tau$. It is seen that if $\mathbf{X}$ is multivariate normally distributed, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then its linear function, $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ is also multivariate normally distributed: $N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\tau)$.

Note this claim does not require $\boldsymbol{\Sigma}$ nor $\mathbf{A}$ to have full rank. It also implies all marginal distributions of a multivariate normal random vector is normally distributed. The inverse is not completely true: if all marginal distributions of a random vector are normal, the random vector does not necessarily have multivariate normal distribution. However, if **all** linear combinations of $\mathbf{X}$ has normal distribution, then the random vector $\mathbf{X}$ has multivariate normal distribution.

When $\mathbf{\Sigma}$ has full rank, then $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ has a density function given by

$$\phi(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi)^{-d/2}\{\det(\mathbf{\Sigma})\}^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\tau \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

where $\det(\cdot)$ is the determinant of a matrix. We use $\Phi(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ for the multivariate c.d.f. .

Partition of X. Assume that a multivariate normal random vector is partitioned into two parts: $\mathbf{X}^\tau = (\mathbf{X}_1^\tau, \mathbf{X}_2^\tau)$. The mean vector, covariance matrix can be partitioned accordingly. In particular, we denote the partition of the mean vector as $\boldsymbol{\mu}^\tau = (\boldsymbol{\mu}_1^\tau, \boldsymbol{\mu}_2^\tau)$ and the covariance matrix as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Theorem 2.1. *Suppose* $\mathbf{X}^\tau = (\mathbf{X}_1^\tau, \mathbf{X}_2^\tau)$ *is multivariate normal,* $N(\boldsymbol{\mu}, \mathbf{\Sigma})$. *Then*

(1) $\mathbf{X}_1$ *is multivariate* $N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$.

(2) $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent if and only if* $\mathbf{\Sigma}_{12} = 0$.

(3) Assume $\mathbf{\Sigma}_{22}$ *has full rank. Then the conditional distribution of* $\mathbf{X}_1|\mathbf{X}_2$ *is normal with conditional mean* $\boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ *and variance matrix* $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.

That is, for multivariate normal random variables, zero-correlation is equivalent to independence. The above result for conditional distribution is given when $\mathbf{\Sigma}_{22}$ has full rank. The situation where $\mathbf{\Sigma}_{22}$ does not have full rank can be worked out by removing the redundancy in $\mathbf{X}_2$ before applying the above result.

2.2 Standard Chi-square distribution

We first fix the idea with a definition.

Definition 2.1. *Let* $Z_1, Z_2, \ldots, Z_d$ *be a set of i.i.d. standard normally distributed random variables. The sum of squares*

$$T = Z_1^2 + Z_2^2 + \cdots + Z_d^2$$

is said to have chi-square distribution with d *degrees of freedom.*

For convenience of future discussion, we first put down a simple result without a proof here.

Theorem 2.2. *Let $Z_1, Z_2, \ldots, Z_d$ be a set of i.i.d. standard normally distributed random variables. The sum of squares*

$$T = a_1 Z_1^2 + a_2 Z_2^2 + \cdots + a_d Z_d^2$$

has chi-square distribution if and only if $a_1, \ldots, a_d$ are either 0 or 1.

We use notation χ_d^2 as a symbol of the chi-square distribution with d degrees of freedom. The above definition is how we understand the chi-square distribution. Yet without seeing its probability density function and so on, we may only have superficial understanding

To obtain the density function of T, we may work on the density function of Z_1^2 first. It is seen that

$$P(Z_1^2 \le x) = P(-\sqrt{x} \le Z_1 \le \sqrt{x}). = \int_{-\sqrt{x}}^{\sqrt{x}} \phi(t)dt$$

Hence, by taking derivative with respect to x, we get its pdf as

$$f_{Z_1^2}(x) = \frac{1}{2\sqrt{\pi}} \left(\frac{x}{2}\right)^{1/2-1} \exp(-\frac{x}{2}).$$

This is the density function of a specific Gamma distribution with 1/2 degrees of freedom and scale parameter 2. Because of this and from the property of Gamma distribution, we conclude that T has Gamma distribution with $d/2$ degrees of freedom, and scale parameter 2. Its p.d.f. is given by

$$f_T(x) = \frac{1}{2\Gamma(d/2)} \left(\frac{x}{2}\right)^{d/2-1} \exp(-\frac{x}{2}).$$

Its moment generating function can also be obtained easily:

$$M_T(t) = \left(\frac{1}{1-2t}\right)^{d/2}.$$

Note that this function is defined only for $t < 1/2$. The mean of T is d, and its variance is $2d$.

Clearly, if $\mathbf{X}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of length d and that $\boldsymbol{\Sigma}$ has full rank, then $\mathbf{W} = (\mathbf{X} - \boldsymbol{\mu})^{\tau} \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$ has chi-square distribution with d degrees of freedom. The cumulative distribution function of standard chi-square distribution with (virtually) any degrees of freedom has been well investigated. There used to be detailed numerical tables for their quantiles and so on. We have easy-to-use R functions these days. Hence, whenever a statistic is found to have a chi-square distribution, we consider its distribution is **known**.

If $\mathbf{A}$ is a symmetric matrix such that $\mathbf{A}\mathbf{A} = \mathbf{A}$, we say that it is idempotent. In this case, the distribution of $\mathbf{Z}^{\tau} \mathbf{A} \mathbf{Z}$ is chisquare distribution with degrees of freedom equaling the trace of $\mathbf{A}$ when $\mathbf{Z}$ is $N(0, \mathbb{I})$.

2.3 Non-central chi-square distribution

We again first fix the idea with a definition.

Definition 2.2. *Let $Z_1, Z_2, \ldots, Z_d$ be a set of i.i.d. standard normally distributed random variables. The sum of squares*

$$T = (Z_1 + \gamma)^2 + Z_2^2 + \cdots + Z_d^2$$

is said to have non-central chi-square distribution with d degrees of freedom and non-centrality parameter γ^2.

Let

$$T' = (Z_1 - \gamma)^2 + Z_2^2 + \cdots + Z_d^2$$

with the same γ as in the definition. The distribution of T' is the same as the distribution of T. This can be proved as follows. Let $W_1 = -Z_1$ and $W_j = Z_j$ for $j = 2, \ldots, d$. Clearly,

$$T' = (W_1 + \gamma)^2 + W_2^2 + \cdots + W_d^2$$

and $W_1, W_2, \ldots, W_d$ remain i.i.d. standard normally distributed. Hence, T and T' must have the same distribution. However, $T \neq T'$ when they are regarded as random variables on the same probability space.

The second remark is about the stochastic order of two distributions. Without loss of generality, $\gamma > 0$. When $d = 1$, and for any $x > 0$, we find

$$P\{(Z_1 + \gamma)^2 \geq x^2\} = 1 - \Phi(x - \gamma) + \Phi(-x - \gamma).$$

Taking derivative with respect to γ, we get

$$\phi(x - \gamma) - \phi(-x - \gamma) = \phi(x - \gamma) - \phi(x + \gamma) > 0.$$

That is, the above probability increases with γ over the range of $\gamma > 0$. That is, $(Z_1 + \gamma)^2$ is always more likely to take larger values than Z_1^2 does.

For convenience, let χ_d^2 and $\chi_d^2(\gamma^2)$ be two random variables with respectively central and non-central chi-square distributions with the same degrees of freedom d. We can show that for any x,

$$P\{\chi_d^2(\gamma^2) \geq x^2\} \geq P\{\chi_d^2 \geq x^2\}.$$

This proof of this result will be left as an exercise.

In data analysis, a statistic or random quantity T often has central chisquare distribution under one model assumption, say A, but non-central chisquare distribution under another model assumption, say B. Which model assumption is better supported by the data? Due to the above result, a large observed value of T is supportive of B while a small observed value of T is supportive of A. This provides a basis for hypothesis test. We set up a threshold value for T so that we accept B when the observed value of T exceeds this value.

Let $\mathbf{X}$ be multivariate normal $N(\boldsymbol{\mu}, \mathbb{I}_d)$. Then $\mathbf{X}^\tau \mathbf{X}$ has non-central chisquare distribution with non-centrality parameter $\boldsymbol{\mu}^\tau \boldsymbol{\mu}$. This can be proved as follow. Without loss of generality, assume $\boldsymbol{\mu} \neq 0$. Let $\mathbf{A}$ be an orthogonal matrix so that its first row equals $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. Let

$$\mathbf{Y} = \mathbf{A}\mathbf{X}.$$

Write $\mathbf{Y}^\tau = (Y_1, Y_2, \ldots, Y_d)$. Then $Y_1' = Y_1 - \|\boldsymbol{\mu}\|, Y_2, \ldots, Y_d$ are i.i.d. standard normal random variables. Hence,

$$\mathbf{X}^\tau \mathbf{X} = \mathbf{Y}^\tau \mathbf{Y} = (Y_1' + \|\boldsymbol{\mu}\|)^2 + Y_2^2 + \cdots + Y_d^2$$

has non-central chi-square distribution with non-centrality parameter $\boldsymbol{\mu}^\tau \boldsymbol{\mu}$.

As an exercise, please show that if $\mathbf{X}$ is multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbf{Q} = \mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}$$

has non-central chi-square distribution with non-centrality parameter $\gamma^2 = \boldsymbol{\mu}^\tau \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

It can be verified that

$$\mathbb{E}(\mathbf{Q}) = d + \gamma^2; \quad \textsc{var}(\mathbf{Q}) = 2(d + 2\gamma^2).$$

When $\Sigma = \sigma^2 \mathbb{I}_d$, then $\mathbf{X}^\tau\mathbf{X}$ has non-central chi-square distribution with d degrees of freedom and non-centrality parameter $\gamma^2 = \|\boldsymbol{\mu}\|^2$.

Suppose $\mathbf{W}_1$ and $\mathbf{W}_2$ are two independent non-central chi-square distributed random variables with d_1 and d_2 degrees of freedome, and non-centrality parameters γ_1^2 and γ_2^2. Then $\mathbf{W}_1 + \mathbf{W}_2$ is also non-central chi-square distributed and its degree of freedom is $d_1 + d_2$ and non-centrality parameters $\gamma_1^2 + \gamma_2^2$.

2.4 Cochran Theorem

We first look into a simple case.

Theorem 2.3. *Suppose* $\mathbf{X}$ *is* $N(0, \mathbb{I}_d)$ *and that*

$$\mathbf{X}^\tau\mathbf{X} = \mathbf{X}^\tau\mathbf{A}\mathbf{X} + \mathbf{X}^\tau\mathbf{B}\mathbf{X} = \mathbf{Q}_A + \mathbf{Q}_B$$

such that both $\mathbf{A}$ *and* $\mathbf{B}$ *are symmetric with ranks* a *and* b *respectively.*

If $a + b = d$*, then* $\mathbf{Q}_A$ *and* $\mathbf{Q}_B$ *are independent and have* χ_a^2 *and* χ_b^2 *distributions.*

Proof: By standard linear algebra result, there exists an orthogonal matrix $\mathbf{R}$ and diagonal matrix $\boldsymbol{\Lambda}$ such that

$$\mathbf{A} = \mathbf{R}^\tau \boldsymbol{\Lambda} \mathbf{R}.$$

This implies

$$\mathbf{B} = \mathbb{I}_d - \mathbf{A} = \mathbf{R}^\tau(\mathbb{I}_d - \boldsymbol{\Lambda})\mathbf{R}$$

in which $(\mathbb{I}_d - \mathbf{\Lambda})$ is also diagonal.

The rank of $\mathbf{A}$ equals the number of non-zero entries of $\mathbf{\Lambda}$ and that of $\mathbf{B}$ is the number of entries of $\mathbf{\Lambda}$ not equalling 1. Since $a + b = d$, this necessitates all entries of $\mathbf{\Lambda}$ are either 0 or 1. Without loss of generality, $\mathbf{\Lambda} = \text{diag}(1, \cdots, 1, 0, \ldots, 0)$.

Note that orthogonal transformation $\mathbf{Y} = \mathbf{R}\mathbf{X}$ makes entries of $\mathbf{Y}$ i.i.d. standard normal. Therefore,

$$\mathbf{Q}_A = \mathbf{Y}^\tau \mathbf{\Lambda} \mathbf{Y} = Y_1^2 + \cdots + Y_a^2$$

which has χ_a^2 distribution. Similarly,

$$\mathbf{Q}_B = \mathbf{Y}^\tau (\mathbb{I}_d - \mathbf{\Lambda}) \mathbf{Y} = Y_{a+1}^2 + \cdots + Y_d^2$$

which has χ_b^2 distribution. In addition, they are quadratic forms of different segments of $\mathbf{Y}$. Therefore, they are independent. □

Remark: Since $\mathbf{X}^\tau \mathbf{A} \mathbf{X} = \mathbf{X}^\tau \mathbf{A}^\tau \mathbf{X}$, we have $\mathbf{Q}_A = \mathbf{X}^\tau \{(\mathbf{A} + \mathbf{A}^\tau)/2\} \mathbf{X}$ in which $\{(\mathbf{A} + \mathbf{A}^\tau)/2\}$ is symmetric. Hence, we do not loss much generality by assuming both $\mathbf{A}$ and $\mathbf{B}$ are symmetric. The result does not hold without symmetry assumption though I cannot find references: Try

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Under symmetry assumption, take it as a simple exercise to show that if

$$\mathbf{X}^\tau \mathbf{X} = \mathbf{X}^\tau \mathbf{A}_1 \mathbf{X} + \cdots + \mathbf{X}^\tau \mathbf{B}_p \mathbf{X} = \sum_{j=1}^{p} \mathbf{Q}_j$$

such that

$$\text{rank}(\mathbf{A}_1) + \cdots + \text{rank}(\mathbf{A}_p) = d$$

then $\mathbf{Q}_j$'s are independent, each has chisquare distribution of degrees $\text{rank}(\mathbf{A}_j)$.

2.5 F- and t-distributions

If X and Y have chisquare distributions with degrees of freedom m and n respectively, then the distribution of

$$F = \frac{X/m}{Y/n}$$

is called F with m and n degrees of freedom. Note that

$$X/(X+Y) = (1+Y/X)^{-1}$$

has Beta distribution. Thus, there is a very simple relationship between the F-distribution and the Beta distribution.

t-distribution. If X has standard normal distribution, and S^2 has chisquare distribution with n degrees of freedom. Further, when X and S^2 are independent,

$$t = \frac{X}{\sqrt{S^2/n}}$$

has t-distribution with n degrees of freedom.

When $n = 1$, this distribution reduces to the famous Cauchy distribution, none of its moments exist.

When n is large, S^2/n converges to 1. Thus, the t-distribution is not very different from the standard normal distribution. A general consensus is that when $n \geq 20$, it is good enough to regard t-distribution with n degrees of freedom as the standard normal in statistical inferences.

2.6 Examples

In this section, we give a few commonly used distributional results in mathematical statistics. Two examples are generally referred to as one-sample and two-sample problems.

Example 2.1. *Consider the normal location-scale model in which for $i = 1, \ldots, n$, we have*

$$Y_i = \mu + \sigma\epsilon_i$$

such that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. N(0, 1). Let $\mathbf{Y}$ be the corresponding Y vector which is multivariate normal with mean

$$\boldsymbol{\mu}^\tau = (1, 1, \ldots, 1) = \mu\mathbf{1}^\tau$$

and identity covariance matrix $\mathbb{I}$. Similarly, we use $\boldsymbol{\epsilon}$ for the vector of ϵ.

The sample variance can be written as

$$\begin{aligned} s_n^2 &= (n-1)^{-1}\mathbf{Y}^\tau(\mathbb{I} - n^{-1}\mathbf{1}\mathbf{1}^\tau)\mathbf{Y}^\tau \\ &= (n-1)^{-1}\sigma^2\boldsymbol{\epsilon}^\tau(\mathbb{I} - n^{-1}\mathbf{1}\mathbf{1}^\tau)\boldsymbol{\epsilon}. \end{aligned}$$

The key matrix $(\mathbb{I} - n^{-1}\mathbf{1}\mathbf{1}^\tau)$ *is idempotent. Hence, other than factor* $(n-1)^{-1}\sigma^2$*, the sample variance has chisquare distribution with* $n-1$ *degrees of freedom.*

In addition, the sample mean $\bar{Y}_n = n^{-1}\mathbf{1}^\tau\mathbf{Y}$ *is uncorrelated to* $(\mathbb{I} - n^{-1}\mathbf{1}\mathbf{1}^\tau)\mathbf{Y}^\tau$*. Hence, they are independent. This further implies that the sample mean and sample variance are independent.*

Example 2.2. *Consider the classical two-sample problem in which we have two i.i.d. samples from normal distribution:* $\mathbf{X}^\tau = (X_1, X_2, \ldots, X_m)$ *are i.i.d.* $N(\mu_1, \sigma^2)$ *and* $\mathbf{Y}^\tau = (Y_1, Y_2, \ldots, Y_n)$ *are i.i.d.* $N(\mu_2, \sigma^2)$*. We are often interested in examining the possibility whether* $\mu_1 = \mu_2$*.*

Let $\bar{X}_m$ *and* $\bar{Y}_n$ *be two sample means. It is seen that*

$$RSS_0 = \frac{mn}{m+n}\{\bar{X}_m - \bar{Y}_n\}^2$$

is a quadratic form that represents the variation between two samples. At the same time,

$$RSS_1 = \sum_{i=1}^m \{X_i - \bar{X}_m\}^2 + \sum_{j=1}^n \{Y_j - \bar{Y}_n\}^2$$

is a quadratic form that represents the internal variations within two populations. It is natural to compare the relative size of RSS_0 *against* RRS_1 *to decide whether two means are significantly different. For this purpose, it is useful to know their sample distributions and independence relationship.*

It is easy to directly verify that RSS_0 *and* RRS_1 *are independent and both have chisquare distributions. We may also find*

$$\mathbf{X}^\tau\mathbf{X} + \mathbf{Y}^\tau\mathbf{Y} = RSS_0 + RSS_1 + (m+n)^{-1}(\mathbf{X}^\tau\mathbf{1}_m + \mathbf{Y}^\tau\mathbf{1}_n)(\mathbf{1}_m^\tau\mathbf{X} + \mathbf{1}_n^\tau\mathbf{Y})$$

The ranks of three quadratic forms on the right hand side are 1, $m+n-2$ *and* 1 *which sum to* n*. The decomposition remains the same when we replace* $\mathbf{X}$ *by* $(\mathbf{X}-\mu)/\sigma$ *and* $\mathbf{Y}$ *by* $(\mathbf{Y}-\mu)/\sigma$*. Hence when* $\mu_1 = \mu_2 = \mu$ *and* $\sigma = 1$*,*

RSS_0 and RRS_1 independent and chisquare distributed by Cochran Theorem (after scaled by σ^2).

This further implies that

$$F = \frac{RSS_0}{RSS_1/(m+n-2)}$$

has F-distribution with degrees of freedom 1 and $m+n-2$.

The F-distribution conclusion is the basis for the analysis of variance, two-sample t-test and so on.

Chapter 3

Exponential distribution families

In mathematical statistics, the normal distribution family plays a very important role for its simplicity and for the reason that many distributions are well approximated by a normal distribution. We have also seen many useful other distributions are derived from normal distributions.

There are many other commonly used distribution families in mathematical statistics. Many of them have density functions conform to a specific algebraic structure. The algebraic structure further enables simple statistical conclusions in data analysis. Hence, it is often useful to have this structure discussed in mathematical statistics.

3.1 One parameter exponential distribution family

Consider a one parameter distribution family whose probability distributions have a density function with respect to a common σ-finite measure. That is, the family is made of

$$\{f(x;\theta) : \theta \in \Theta \subset \mathcal{R}\}$$

with Θ being its parameter space.

Definition 3.1. *Suppose there exist real valued functions $\eta(\theta)$, $T(x)$, $A(\theta)$ and $h(x)$ such that*

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - A(\theta)\}h(x). \tag{3.1}$$

We say $\{f(x;\theta) : \theta \in \Theta \subset \mathcal{R}\}$ is a one-parameter exponential family.

The definition does not give much insight on the specific algebraic form is of interest. Let us build some intuition from several examples.

Example 3.1. *Suppose $X_1, \ldots, X_n$ are i.i.d. from Binomial (m, θ). Their joint density (probability mass) function is given by*

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \left[\binom{m}{x_i} \theta^{x_i} (1-\theta)^{m-x_i} \right].$$

Let

$$T(X) = \sum X_i, \quad \textit{and} \quad T(x) = \sum x_i$$

and

$$h(x) = \prod_{i=1}^{n} \binom{m}{x_i}.$$

Then we find

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= \exp\{T(x)\log\theta + (nm - T(x))\log(1-\theta)\}h(x) \\ &= \exp\{\log\{\theta/(1-\theta)\}T(x) + nm\log(1-\theta)\}h(x). \end{aligned}$$

This conforms the definition of one parameter family with

$$\eta = \log\{\theta/(1-\theta)\}$$

and

$$A(\theta) = nm\log(1-\theta).$$

As an exercise, you can follow this example to show that both Negative Binomial, Poisson distributions are one-parameter exponential families.

In the above example, η is call log-odds because $\theta/(1-\theta)$ is the odds of success compared to failure in typical binary experiments. It is equally useful to "label" Binomial distribution family by log-odds. Note that

$$\theta = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

Hence, we may equivalently state that the joint density function of X is given by

$$g(x_1, \ldots, x_n; \theta) = \exp\{\eta T(x) - nm \log(1+\exp(\eta))\}h(x).$$

This form also confirms the definition of the one-parameter exponential family.

Definition 3.2. *Let X be a random variable or vector. The support of X of that of its distribution is the set of all x such that for any $\delta > 0$,*

$$P\{X \in (x-\delta, x+\delta)\} > 0.$$

For the sake of accuracy, a definition sometimes has to be abstract. The support of X is intuitively the set of x such that $X = x$ is a "possible event". When Z is $N(0,1)$, we have $P(Z = z) = 0$. Hence, we cannot interpret "possible event" as a positive probability event. The above definition first expands x and then judges its "possibility". Hence, the support contains all x at which the density function is positive and continuous.

We do not ask you to memorize this definition. Rather, we merely point out that if two distributions belong the same one-parameter exponential family, then they have the same support. In comparison, a standard exponential distribution has support $[0,\infty)$ and a standard normal distribution has support $\mathcal{R}$. Let us now show you another interesting property.

Example 3.2. *Let us now consider the natural form of the one-parameter exponential family:*

$$f(x_1, \ldots, x_n; \eta) = \exp\{\eta T(x) - A(\eta)\}h(x)$$

with η being a real value whose parameter space is an interval. The moment generating function of $T(x)$ is given by

$$M_T(s) = \mathbb{E}\exp\{sT(X)\} = \exp\{A(\eta+s) - A(\eta)\}.$$

This implies that

$$\mathbb{E}\{T\} = M_T'(0) = A'(\eta).$$

and

$$\mathbb{E}\{T^2\} = M_T''(0) = A''(\eta) + \{A'(\eta)\}^2.$$

Hence,

$$\text{VAR}(T) = A''(\eta).$$

This example shows that the exponential families have some neat properties which make them an interest object to study.

3.2 The multiparameter case

We can practically copy the previous definition without any changes.

Definition 3.3. *Suppose there exist real-vector valued functions $\eta(\boldsymbol{\theta})$, $\mathbf{T}(x)$, and real valued functions $A(\boldsymbol{\theta})$ and $h(x)$ such that*

$$f(x;\boldsymbol{\theta}) = \exp\{\eta^\tau(\boldsymbol{\theta})\mathbf{T}(x) - A(\boldsymbol{\theta})\}h(x). \tag{3.2}$$

We say $\{f(x;\theta) : \theta \in \Theta \subset \mathcal{R}^d\}$ is a multi-parameter exponential family.

Without the above expansion, the exponential family does not even include normal distribution.

Example 3.3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with distribution $N(\mu, \sigma^2)$. Their joint density function*

$$\begin{aligned}\phi(x_1, \ldots, x_n; \mu, \sigma^2) &= (2\pi)^{-n/2}\sigma^{-n}\exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}\\ &= (2\pi)^{-n/2}\exp\{\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - \frac{n\mu^2}{2\sigma^2} - n\log\sigma\}.\end{aligned}$$

We now regard $\boldsymbol{\theta}$ as a vector made of μ and σ. The above density function *fits into the definition* (3.2) *with the following functions:*

$$\begin{aligned}
\eta(\boldsymbol{\theta}) &= \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^\tau, \\
\mathbf{T}(x) &= (\sum x_i, \sum x_i^2)^\tau, \\
A(\boldsymbol{\theta}) &= -\frac{n\mu^2}{2\sigma^2} - n\log\sigma, \\
h(x) &= (2\pi)^{-n/2}.
\end{aligned}$$

Recall the Binomial distribution example. We had joint density function given by

$$f(x_1, \ldots, x_n; \theta) = \exp\{T(x)\log\theta + (nm - T(x))\log(1-\theta)\}h(x).$$

It can also be regarded as a multi-parameter exponential family with $d = 2$ and

$$\eta = (\log\theta, \log(1-\theta))^\tau; \quad \mathbf{T}_{new}(x) = (T(x), nm - T(x))^\tau.$$

The parameter space in terms of values of η is a curve in $\mathcal{R}^2$ which does not contain any open (non-empty) subset of $\mathcal{R}^2$. We generally avoid having a distribution families with degenerate parameter spaces.

As an exercise, one can verify that two-parameter Gamma distribution family is a multiple parameter exponential family.

3.3 Other properties

Suppose X_1 and X_2 both have distributions belonging to some exponential families and they are independent. Then their joint distribution also belongs to an exponential family.

By factorization theorem, $\mathbf{T}(X)$ in exponential family is a sufficient statistic. It is also a complete statistic when the family does not degenerate.

The distribution of $\mathbf{T}$ belongs to an exponential family.

Definition 3.4. *Let* $\mathbf{T}$ *be a k-dimensional vector valued function and h be a real value function. The canonical k-dimensional exponential family generated by* $\mathbf{T}$ *and h is*

$$g(x; \eta) = \exp\{\eta^\tau T(x) - A(\eta)\}h(x).$$

The parameter space for η is all $\eta \in \mathcal{R}^k$ such that $\exp\{\eta^\tau T(x)\}h(x)$ has finite integration with respect to the corresponding σ-finite measure.

We call the parameter space, $\mathcal{E}$, the natural parameter space. We call $\mathbf{T}$ *and h generators.*

Because the integration of a density function equals 1, the integration of $\exp\{\eta^\tau T(x)\}h(x)$ equals $\exp(A(\eta)$ if it is finite. Hence, the natural parameter space $\mathcal{E}$ contains all η at which $A(\cdot)$ is well-defined.

Definition 3.5. *We say that an exponential family $\mathcal{F}$ is of rank k if and only if the generating statistic* $\mathbf{T}$ *is k-dimensional and $1, T_1, \ldots, T_k$ are linearly independent with positive probability. That is,*

$$P(a_0 + \sum_{j=1}^{k} a_j T_j = 0; \eta) < 1$$

for some η unless all non-random coefficients $a_0 = a_1 = \cdots = a_k = 0$.

In the above definition, we only need to verify the probability inequality for one η value. If it is satisfied for one η value, then it is satisfied for any other η value.

Theorem 3.1. *Suppose $\mathcal{F} = \{g(x;\eta) : \eta \in \mathcal{E}\}$ is a canonical exponential family generated by* $(\mathbf{T}, h)$ *with natural parameter space $\mathcal{E}$ such that $\mathcal{E}$ is open. Then the following are equivalent:*

(a) $\mathcal{F}$ is of rank k.

(b) $\text{VAR}(\mathbf{T}; \eta)$ *is positive definite.*

(c) η is identifiable: $g(x;\eta_1) \equiv g(x;\eta_2)$ for all x implies $\eta_1 = \eta_2$.

These discussions on exponential family suffice for the moment so we move to the next topic.

Chapter 4

Criteria of point estimation

A general setting of the mathematical statistics is: we are given a data x believed to be the observed value of a random object X. The probability distribution of X will be denoted as F^* and F^* is believed to be a member of a distribution family $\mathcal{F}$. Based on the fact that X has an observed value x, identify a single or a set of F in $\mathcal{F}$ which might be the "true" F^* that describe the probability distribution of X.

There are many serious fallacies related to the above thinking. The first one I can tell is the specification of $\mathcal{F}$, which is referred as a model in this course. If a specific form of $\mathcal{F}$ is given, how certain are we on F^* is one of $\mathcal{F}$? Even if the distribution of X is a member of $\mathcal{F}$, X may not be accurately observed. What we have recorded may be $Y = X + \epsilon$. Hence, we may unknowingly working to the distribution of Y instead that of X.

In this course, we do not discuss these possible fallacies but leave them to other more applied courses. We take the approach that if the distribution of X is indeed a member of $\mathcal{F}$ and x is its accurate observed value, what can we say about F^*? Also, we often study the situation where X is an i.i.d. replication of some random system so that $X = (X_1, \ldots, X_n)$. The model of the distribution of X will be then taken over by the model for X_1 which is representative for every X_i, $i = 1, 2, \ldots, n$. We state that $X_1, \ldots, X_n$ is an random or an i.i.d. sample from population/distribution F of $\mathcal{F}$. In this case n is referred to as sample size. With many replications, or when $n \to \infty$, we should be able to learn a lot more about F^*.

4.1 Point estimator and some optimality criteria

Let θ be a parameter in the probability model $\mathcal{F}$ and suppose we have a random sample X. The parameter space is loosely $\Theta = \{\theta : \theta = g(F), F \in \mathcal{F}\}$ for some functional g. A point estimator of θ is a statistic T whose range is Θ. The realized value of T, $T(x)$, is an estimate of θ. We generally allow, for the least, T to take values on the smallest closed set containing Θ. That is, taking values on limiting points of Θ.

Definition 4.1. *A point estimator of θ is a statistic T whose range is Θ. The realized value of T, $T(x)$, is an estimate of θ.*

The definition implies that as an estimator, $T(X)$ is regarded as a mechanism/rule of mapping X to Θ; as an estimate, $T(x)$ is a value in Θ which corresponding to data x. In both cases, we may use $\hat{\theta}$ as their common notation.

One must realize $T(x) = 0$ is an estimator of θ as long as $0 \in \Theta$. Hence, we *always can* estimate the parameter in any statistical models, no matter how complex the model is. We may not be able to find an estimator with a satisfactory precision or certain desired properties.

Suppose the parameter space is a subset of $\mathcal{R}^d$ for some integer d. Hence, $T(X)$ takes values in $\mathcal{R}^d$. When the distribution of X is given by an $F \in \mathcal{F}$ or equivalently c.d.f. $F(x;\theta)$ or p.d.f. $f(x;\theta)$. Hence, $T(X)$ is a distribution induced by $F(x;\theta)$ or simply by θ. To fix the idea, we assume the "true" parameter value of F is θ, the generic θ. When $\hat{\theta} = T(X)$ has finite expectation under any θ, we define

$$\textsc{bias}(\hat{\theta}) = \mathbb{E}\{T(X);\theta\} - \theta$$

as the bias of $\hat{\theta} = T(X)$ when it is used as an estimator of θ and when the true parameter value is θ.

Definition 4.2. *Suppose X has a distribution $F \in \mathcal{F}$ which is parameterized by $\theta \in \Theta$. Suppose $T(X)$ is an estimator of θ such that*

$$\mathbb{E}\{T(X);\theta\} = \theta$$

for all $\theta \in \Theta$, then we say $T(X)$ is an unbiased estimator of θ.

For some reason, statisticians and others prefer estimators that are unbiased. This is not always well justified.

Example 4.1. *Suppose X has binomial distribution with parameters n and θ, n is known and θ is an unknown parameter.*

A commonly used estimator for θ is

$$\hat{\theta} = \frac{X}{n}.$$

An estimator motivated by Bayesian approach is

$$\tilde{\theta} = \frac{X+1}{n+2}.$$

It is seen $\mathbb{E}\{\hat{\theta}; \theta\} = \theta$. Hence, it is an unbiased estimator.

We find that other than $\theta = 0.5$,

$$\text{BIAS}(\tilde{\theta}) = \frac{1-2\theta}{n+2} \neq 0.$$

Hence, $\tilde{\theta}$ is a biased estimator.

Which estimator makes more sense to you?

In the above example, the bias of $\tilde{\theta}$ has a limit 0 when n goes to infinite. Often, we discuss situations where the data set contains n i.i.d. observations from a distribution F which is a member of $\mathcal{F}$. The above result indicates that even though $\tilde{\theta}$ is biased, the size of the bias diminishes when the sample size n gets large. Many of us tends to declare that $\tilde{\theta}$ is asymptotically unbiased when this happens.

While we do not feel such a notion of "asymptotically unbiased" is wrong, this terminology is often abused. In statistical literature, people may use this term when

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

has a limiting distribution whose mean is zero. In this case, the bias of $\hat{\theta}$ does not necessarily goes to zero.

To avoid such confusions, let us invent a formal definition.

Definition 4.3. *Suppose there is an index n such that X_n has a distribution in $\mathcal{F}_n$ and $a_n \to \infty$ as $n \to \infty$ while the parameter space Θ of $\mathcal{F}_n$ does not depend on n. Let θ be the true parameter value and $\hat{\theta}_n$ is an estimator (a sequence of estimators). If*

$$a_n(\hat{\theta}_n - \theta)$$

has a limiting distribution whose expectation is zero, for any $\theta \in \Theta$, then we say $\hat{\theta}_n$ is asymptotically rate-a_n unbiased.

Most often, we take $a_n = n^{1/2}$ in the above definition. We do not have good reasons to require an estimator unbiased. Yet we feel that being asymptotically unbiased for some a_n is a necessity. When $n \to \infty$ in common settings, the amount of information about which F is the right F becomes infinity. If we cannot make it right in this situation, the estimation method is likely very poor.

The variance of an estimator is as important a criterion in judging an estimator. Clearly, having a lower variance implies the estimator is more accurate. In fact, let $\varphi(\cdot)$ be a convex function. Then an estimator is judged superior if

$$\mathbb{E}\{\varphi(\hat{\theta} - \theta)\}$$

is smaller. When $\varphi(x) = x^2$, the above criterion becomes Mean Squared Error:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\{(\hat{\theta} - \theta)^2\}.$$

It is seen that

$$\text{MSE}(\hat{\theta}) = \text{BIAS}^2(\hat{\theta}) + \text{VAR}(\hat{\theta}).$$

To achieve lower MSE the estimator must balance the loss due to variation and bias.

Similar to asymptotic bias, it helps to give definite notions of asymptotic variance and MSE of an estimator.

Definition 4.4. *Suppose there is an index n such that X_n has a distribution in $\mathcal{F}_n$ and $a_n \to \infty$ as $n \to \infty$ while the parameter space Θ of $\mathcal{F}_n$ does not depend on n. Let θ be the true parameter value and $\hat{\theta}_n$ is an estimator (a sequence of estimators). Suppose*

$$a_n(\hat{\theta}_n - \theta)$$

has a limiting distribution with mean $\mathrm{B}(\theta)$ *and variance* $\sigma^2(\theta)$, *for* $\theta \in \Theta$.

We say $\hat{\theta}_n$ *has asymptotic bias* $\mathrm{B}(\theta)$ *and asymptotic variance* $\sigma^2(\theta)$ *at rate* a_n.

Further more, we define the asymptotic MSE *at rate* a_n *as the* $\sigma^2(\theta) + \mathrm{B}^2(\theta)$.

Unfortunately, the MSE is often a function of θ. In any specific application, the "true value" of θ behind X is not known. Hence, it is not possible to find an estimator which is a better estimator in terms of variance or MSE whichever value θ is the true value.

Example 4.2. *Suppose* $X_1, X_2, \ldots, X_n$ *form an i.i.d. sample from* $N(\theta; 1)$ *such that* $\Theta = \mathcal{R}$.

Define $\hat{\theta} = n^{-1} \sum X_i$ *and* $\tilde{\theta} = 0$.

It is seen that $\mathrm{VAR}(\hat{\theta}) = n^{-1} > \mathrm{VAR}(\tilde{\theta})$ *for any* $\theta \in \mathcal{R}$. *However, no one will be happy to use* $\tilde{\theta}$ *as his/her estimator.*

In addition, $\mathrm{MSE}(\hat{\theta}) = n^{-1} > \mathrm{MSE}(\tilde{\theta})$ *for all* $|\theta| < n^{-1/2}$. *Hence, even if we use a more sensible performance criterion, it still does not imply that our preferred sample mean is indisputably a superior estimator.*

4.2 Uniformly minimum variance unbiased estimator

This section contains some materials that most modern statisticians believe we should not have them included in statistical classes. Yet we feel a quick discussion is still a good idea.

Either BIAS, VAR, MSE can be used to separate the performance of estimators we can think of. Yet without any performance measure, how can statisticians recommend any method to scientists? This is the same problem when professors are asked to recommend their students. Everyone is unique. Simplistically declaring one of them is the best will draw more criticisms than praises. Yet at least, we can timidly say one of the students has the highest average mark on mathematics courses, in this term, among all students with green hair and so on.

Definition 4.5. *Suppose X is a random sample from $\mathcal{F}$ with parameter $\theta \in \Theta$.*

An unbiased estimator $\hat{\theta}$ is uniformly minimum variance estimator of θ, UMVUE, if for any other unbiased estimator $\tilde{\theta}$ of θ,

$$\textsc{var}_\theta(\hat{\theta}) \leq \textsc{var}_\theta(\tilde{\theta})$$

for all $\theta \in \Theta$.

In the above definition, we added a subscript θ to highlight the fact that the variance calculation is based on the assumption that the of X has true parameter value θ. We do not always do so in other part of the course note.

Upon the introduction of UMVUE, a urgent question to be answered is its existence. This answer is positive at least in textbook examples.

Example 4.3. *Suppose $X_1, X_2, \ldots, X_n$ form an i.i.d. sample from Poisson distribution with mean parameter θ and the parameter space is $\Theta = \mathcal{R}^+$.*

Let $\hat{\theta} = \bar{X}_n = n^{-1} \sum X_i$. It is easily seen that $\hat{\theta}$ is an unbiased estimator of θ.

Suppose that $\tilde{\theta}$ is another unbiased estimator of θ. Because $\bar{X}_n$ is complete and sufficient statistic, we find

$$\breve{\theta} = \mathbb{E}\{\tilde{\theta}|\bar{X}_n)$$

is a function of data only. Hence, it is an estimator of θ. Using a formula that for any two random variables, $\textsc{var}(Y) = \mathbb{E}\{\textsc{var}(Y|Z)\} + \textsc{var}\{\mathbb{E}(Y|Z)\}$, we find

$$\textsc{var}(\breve{\theta}) \leq \textsc{var}(\tilde{\theta}).$$

Furthermore, this estimator is also unbiased. Hence,

$$\mathbb{E}\{\hat{\theta} - \breve{\theta}\} = 0$$

for all $\theta \in \mathcal{R}^+$. Because both estimators are function of $\bar{X}_n$ and the completeness of $\bar{X}_n$, we have

$$\hat{\theta} = \breve{\theta}.$$

Hence,

$$\textsc{var}(\hat{\theta}) = \textsc{var}(\breve{\theta}) \leq \textsc{var}(\tilde{\theta}).$$

Therefore, $\bar{X}_n$ is the UMVUE.

Now, among all estimators of θ that are unbiased, the sample mean has the lowest possible variance. If UMVUE is a criterion we accept, then the sample mean is the best possible estimator under the Poisson model for the mean parameter θ.

Why is such a beautiful conclusion out of fashion these days? Some of the considerations are as follows. In real world applications, having a random sample strictly i.i.d. from a Poisson distribution is merely a fantasy. If so, why should we bother? Our defence is as follows. If the sample mean is optimal in the sense of UMVUE under the ideal situation, it is likely a superior one even if the situation is slightly different from the ideal. In addition, the optimality consideration is a good way of thinking.

Suppose $\lambda = 1/\theta$ which is called rate parameter under Poisson model assumption. How would you estimate λ? Many will suggest that $\bar{X}_n^{-1}$ is a good candidate estimator. Sadly, this estimator is biased and has infinite variance! Lastly, in modern applications, we rarely work with such simplistic models. In these cases, it is nearly impossible to have a UMVUE. If so, we probably should not bother our students with such technical notions.

4.3 Information inequality

At least in textbook examples, some estimators are fully justified as optimal. This implies that there is an intrinsic limit on how precise an estimator can achieve.

Let X be a random variable modelled by $\mathcal{F}$ or more specifically a parametric family $f(x;\theta)$. Let $T(X)$ be a statistic with finite variance given any $\theta \in \Theta$. Denote

$$\psi(\theta) = \mathbb{E}\{T(X);\theta\} = \int T(x)f(x;\theta)dx$$

where the Lebesgue measure can be replaced by any other suitable measures. Suppose some regularity conditions on $f(x;\theta)$ are satisfied so that our following manipulations are valid. Taking derivatives with respect to θ on two sides of the equality, we find

$$\begin{aligned}\psi'(\theta) &= \int T(x)f'(x;\theta)dx\\ &= \int T(x)s(x;\theta)f(x;\theta)dx\end{aligned}$$

where

$$s(x;\theta) = \frac{f'(x;\theta)}{f(x;\theta)} = \frac{\partial}{\partial\theta}\{\log f(x;\theta)\}.$$

It is seen that

$$\int s(x;\theta)f(x;\theta)dx = \int f'(x;\theta)dx = \frac{d}{d\theta}\int f(x;\theta)dx = 0.$$

We define the Fisher information

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

Hence,

$$\begin{aligned}\{\psi'(\theta)\}^2 &= \{\int\{T(x)-\psi(\theta)\}f(x;\theta)dx\}^2\\ &\leq \int\{T(x)-\psi(\theta)\}^2 s(x;\theta)f(x;\theta)dx \times \int\{s(x;\theta)\}^2 f(x;\theta)dx\\ &= \text{VAR}(T(x))\mathbb{I}(\theta).\end{aligned}$$

This leads to the following theorem.

Theorem 4.1. Cramér-Rao information inequality. *Let $T(X)$ be any statistic with finite variance for all $\theta \in \Theta$. Under some regularity conditions,*

$$\text{VAR}(T(X)) \geq \frac{\{\psi'(\theta)\}^2}{\mathbb{I}(\theta)}$$

where $\psi(\theta) = \mathbb{E}(T(X);\theta)$.

If $T(X)$ is unbiased for θ, then $\psi'(\theta) = 1$. Therefore, $\text{VAR}(T) \geq \mathbb{I}^{-1}(\theta)$. When $\mathbb{I}(\theta)$ is larger, the variance of T could be smaller. Hence, it indeed measures the information content in data X with respect to θ. For convenience of reference, we call $\mathbb{I}^{-1}(\theta)$ the information lower bound for estimating θ.

In assignment problems, X is often made of n i.i.d. observations from $f(x;\theta)$. Let X_1 be one component of X. It is a simple exercise to show that

$$\mathbb{I}(\theta; X) = n\mathbb{I}(\theta; X_1)$$

in the obvious notation. We need to pay attention to what $\mathbb{I}(\theta)$ stands for in many occasions. It could be the information contained in a single X_1, but also could be information contained in the i.i.d. sample $X_1, \ldots, X_n$.

Example 4.4. *Suppose $X_1, X_2, \ldots, X_n$ form an i.i.d. sample from Poisson distribution with mean parameter θ and the parameter space is $\Theta = \mathcal{R}^+$.*

The density function of X_1 is given by

$$f(x;\theta) = P(X_1 = x;\theta) = \frac{\theta^x}{x!}\exp(-\theta).$$

Hence,

$$s(x;\theta) = \frac{x}{\theta} - 1$$

and the information in X_1 is given by

$$\mathbb{I}(\theta) = \mathbb{E}\left\{\frac{X}{\theta} - 1\right\}^2 = \frac{1}{\theta}.$$

Therefore, for any unbiased estimator T_n of θ based on the whole sample, we have

$$\text{VAR}(T_n) \geq \frac{1}{n\mathbb{I}(\theta)} = \frac{\theta}{n}.$$

Since the sample mean is unbiased and has variance θ/n, it is an estimator that attains the information lower bound.

The definition of Fisher information depends on how the distribution family is parameterized. If η is a smooth function of θ, the Fisher information with respect to η is not the same as the Fisher information with respect to θ.

As an exercise, find the information lower bound for estimating $\eta = \exp(-\theta)$ under Poisson distribution model. Derive its UMVUE given n i.i.d. observations.

4.4 Other desired properties of a point estimator

Given a data set from an assumed model $\mathcal{F}$, we often ask or are asked whether certain aspect of $\mathcal{F}$ can be estimated. This can be the mean or median of F where F is any member of $\mathcal{F}$. In general, we may write the parameter as $\theta = \theta(F)$, a functional defined on $\mathcal{F}$.

Definition 4.6. Obsolete Concept of Estimability. *Suppose the data set X is a random sample from a model $\mathcal{F}$ and suppose $\theta = \theta(F)$ is a parameter. We say θ is estimable if there exists a function $T(\cdot)$ such that*

$$\mathbb{E}(T(X); F) = \theta(F)$$

for all $F \in \mathcal{F}$.

In other words, a parameter is estimable if we can find an unbiased estimator for this parameter. We can give many textbook examples of estimability. In contemporary applications, we are often asked to "train" a model given a data set with very complex structure. In this case, we do not even have a good description of $\mathcal{F}$. Because of this, being estimable for a useful functional on $\mathcal{F}$ is a luxury. We have to give up this concept but remain aware of such a definition.

It is not hard to give an example of un-estimable parameters according to the above definition though the example can overly technical. Instead, we show that there is a basic requirement for a parameter to be estimable.

Definition 4.7. Identifiability of a statistical model. *Let $\mathcal{F}$ be a parametric model in statistics and Θ be its parameter space. We say $\mathcal{F}$ is identifiable if for any $\theta_1, \theta_2 \in \Theta$,*

$$F(x; \theta_1) = F(x; \theta_2)$$

for all x implies $\theta_1 = \theta_2$.

A necessary condition for a parameter θ to be estimable is that θ is identifiable. Otherwise, suppose $F(x; \theta_1) = F(x; \theta_2)$ for all x, but $\theta_1 \neq \theta_2$. For any estimator $\hat{\theta}$, we cannot have both

$$\mathbb{E}\{\hat{\theta}; \theta_1\} = \theta_1; \quad \mathbb{E}\{\hat{\theta}; \theta_2\} = \theta_2$$

because two expectations are equal while $\theta_1 \neq \theta_2$.

Definition 4.8. Proposed notion of estimability. *Let $\mathcal{F}$ be a parametric model in statistics and Θ be its parameter space. Suppose the sample plan under consideration may be regarded as one of a sequence of sampling plans indexed by n with sample X_n from $\mathcal{F}$. If there exists an estimator T_n, a function of X_n, such that*

$$P(|T_n - \theta| \geq \epsilon; \theta) \to 0$$

for any $\theta \in \Theta$ and $\epsilon > 0$ as $n \to \infty$, then we say θ is (asymptotically) estimable.

The sampling plans in my mind include the plan of obtaining i.i.d. observations, obtaining observations of time series with extended length and so on. This definition makes sense but we will not be surprised to draw serious criticisms.

Example 4.5. *Suppose we have an i.i.d. sample of size n from Poisson distribution. Let λ be the rate parameter. It is seen that λ is asymptotically estimable because*

$$P\big(\big|\frac{1}{n^{-1} + \bar{X}_n} - \lambda\big| > \epsilon\big) \to 0$$

as $n \to \infty$, where $\bar{X}_n$ is the sample mean.

In this example, I have implicitly regarded "having i.i.d. sample of size n" as a sequence of sampling plan. If one cannot obtain more and more i.i.d. observations from this population, then the asymptotic estimability does not make a lot of sense.

If two random variables are related by $Y = (5/9)(X - 32)$ such as the case where Y and X are the temperatures measured in Celsius and Fahrenheit. Given measures $X_1, X_2, \ldots, X_n$ on a random sample from some population, it is most sensible to estimate the mean temperature as $\bar{X}_n$, the sample mean of X. If one measures the temperature in Celsius to get $Y_1, \ldots, Y_n$ on the same random sample, we should have estimated the mean by $\bar{Y}_n$, the sample mean of Y. Luckily, we have $\bar{Y}_n = (5/9)(\bar{X}_n - 32)$. Some internal consistency is maintained. Such a desirable property is termed as *equivariant.* and sometimes is also called *invariant.* See Lehmann for references.

In another occasion, one might be interested in estimating mean parameter μ in Poisson distribution. This parameter tells us the average number of events occuring in a time period of interest. At the same time, one might be interested in knowing the chance that nothing happens in the period which is $\exp(-\mu)$. Let $\bar{X}_n$ as the sample mean of the number of events over n distinct periods of time. We naturally estimate μ by $\bar{X}_n$ and $\exp(-\mu)$ by $\exp(-\bar{X}_n)$. If so, we find

$$\widehat{g(\mu)} = g(\hat{\mu})$$

with $g(x) = \exp(-x)$. This is a property most of us will find desirable. When an estimator satisfies above property, we say it is *invariant*.

Rigorous definitions of equivariance and invariance can be lengthy. We will be satisfied with a general discussion as above.

In the Poisson distribution example, it is seen that

$$\mathbb{E}\{\exp(-\bar{X}_n)\} = \exp\{n\mu[\exp(-1/n) - 1]\} \neq \exp(-\mu).$$

Hence, the most natural estimator of $\exp(-\mu)$ is not unbiased.

The UMVUE of $\exp(-\mu)$ is given by $\mathbb{E}\{\mathbb{1}(X_1 = 0)|\bar{X}_n\}$. The UMVUE of μ is given by $\bar{X}_n$. Thus, the UMVUE is not invariant when the population is the Poisson distribution family. As a helpful exercise for improving one's technical strength, work out the explicit expression of $\mathbb{E}\{\mathbb{1}(X_1 = 0)|\bar{X}_n\}$.

4.5 Consistency and asymptotic normality

A point estimator is a function of data and the data are a random sample from a distribution/population that is a member of distribution family. Hence, it is random in general: its does not take a value with probability one. In other words, we can never be completely sure about the unknown parameter. However, when the sample size increases, we gain more and more information about its underlying population. Hence, we should be able to decide what the "true" parameter value with higher and higher precision.

Definition 4.9. *Let θ_n be an estimator of θ based on a sample of size n from a distribution family $F(x;\theta) : \theta \in \Theta$. We say that θ_n is weakly consistent if,* *as $n \to \infty$, for any $\epsilon > 0$ and $\theta \in \Theta$*

$$P(|\hat{\theta}_n - \theta| \geq \epsilon; \theta) \to 0.$$

In comparison, we have a stronger version of consistency.

Definition 4.10. *Let θ_n be an estimator of θ based on a sample of size n from a distribution family $F(x; \theta) : \theta \in \Theta$. We say that θ_n is strongly consistent if, as $n \to \infty$, for any $\theta \in \Theta$*

$$P(\lim_{n\to\infty} \hat{\theta}_n = \theta; \theta) = 1.$$

Here are a few remarks one should not take them seriously but worth to point out. First, the i.i.d. structure in the above definitions is not essential. However, it is not easy to give a more general and rigorous definition without this structure. Second, the consistency is not really a property of **one** estimator, but a **sequence** of estimators. Unless $\hat{\theta}_n$ for all n are constructed based on the same principle, otherwise, the consistency is nothing relevant in applications: your n is far from infinity. For this reason, there is a more sensible definition called Fisher consistency. To avoid too much technicality, it is mentioned but not spelled out here. Lastly, when we say an estimator is consistent, we mean weakly consistent unless otherwise stated.

The next topic is asymptotic normality. It is in fact best to be called limiting distributions. Suppose $\hat{\theta}_n$ is an estimator of θ based on n i.i.d. observations from some distribution family. The precision of this estimator can be judged by its bias, variance, mean square error and so on. Ultimately, the precision of $\hat{\theta}_n$ is its **sample distribution**. Unfortunate, the sample distribution of $\hat{\theta}_n$ is often not easy to directly work with. At the same time, when n is very large, the distribution of its standardized version stabilizes. This is the limiting distribution. If we regard the limiting distribution as the sample distribution of $\hat{\theta}$, the difference is not so large. That is, the error diminishes when n increases. For this reason, statisticians are fond of finding limiting distributions.

Definition 4.11. *Let T_n be a sequence of random variables, we say its distribution converges to that of T if*

$$\lim_{n\to\infty} P(T_n \leq t) = P(T \leq t)$$

for all $t \in \mathcal{R}$ at which $F(t) = P(T \leq t)$ is continuous.

In this definition, T_n is just any sequence random variable, it may contain unknown parameters in specific examples. The index n need not be the sample size in typical set up. The multivariate case will not be given here. The typical applications, the limiting distribution is about asymptotic normality.

Example 4.6. *Suppose we have an i.i.d. sample $X_1, \ldots, X_n$ from a distribution family $\mathcal{F}$. A typical estimator for $F(t)$, the cumulative distribution function of X is the empirical distribution*

$$F_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t).$$

For each given t, the distribution of $F_n(t)$ is kind of binomial. At the same time,

$$\sqrt{n}\{F_n(t) - F(t)\} \xrightarrow{d} N(0, \sigma^2)$$

with $\sigma^2 = F(t)\{1 - F(t)\}$ as $n \to \infty$.

Remark: in this example, we have a random variable on one side but a distribution on the other side. It is interpreted as the distribution sequence of the random variables, indexed by n, converges to the distribution specified on the right hand side.

As an exercise, one can work out the following example.

Example 4.7. *Suppose we have an i.i.d. sample $X_1, \ldots, X_n$ from a uniform distribution family $\mathcal{F}$ such that $F(x;\theta)$ is uniform on $(0, \theta)$ and $\Theta = \mathcal{R}^+$. Define*

$$\hat{\theta}_n = \max\{X_1, X_2, \ldots, X_n\}$$

which is often denoted as $X_{(n)}$ and called order statistic. It is well known that

$$n\{\theta - \hat{\theta}\} \xrightarrow{d} \exp(\theta).$$

Namely, the limiting distribution is exponential.

Is $\hat{\theta}$ asymptotically unbiased at rate $\sqrt{n}$, at rate n?

Chapter 5

Approaches of point estimation

Even though any statistics with proper range is a point estimator, we generally prefer estimators derived based on some principles. This leads to a few common estimation procedures.

5.1 Method of moments

Suppose $\mathcal{F}$ is a parametric distribution family so that it permits a general expression

$$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$$

such that $\Theta \subset \mathcal{R}^d$ for some positive integer d. We assume the parameter is identifiable.

In most classical examples, the distributions are labeled smoothly by θ: two distributions having close parameter values are similar in some metric. In addition, the first d moments are smooth functions of θ. They map Θ to $\mathcal{R}^d$ in a one-to-one fashion: different θ value leads to different first d moments.

Suppose we have an i.i.d. sample $X_1, \ldots, X_n$ of size n from $\mathcal{F}$ and X is univariate. For $k = 1, 2, \ldots, d$, define equations with respect to θ as

$$n^{-1}\{X_1^k + X_2^k + \cdots + X_n^k\} = \mathbb{E}\{X^k; \theta\}.$$

The solution in θ, if exists and unique, are called moment estimator of θ.

Example 5.1. *If $X_1, \ldots, X_n$ is an i.i.d. sample from Negative binomial distribution whose probability mass function (p.m.f.) is given by*

$$f(x;\theta) = \binom{-m}{x}(\theta - 1)^x \theta^m$$

for $x = 0, 1, 2, \ldots$. It is known that $\mathbb{E}\{X;\theta\} = m/\theta$. Hence, the moment estimator of θ is given by

$$\hat{\theta} = m/\bar{X}_n.$$

If $X_1, \ldots, X_n$ is an i.i.d. sample from $N(\mu, \sigma^2)$ distribution. It is known that $\mathbb{E}\{X, X^2\} = (\mu, \mu^2 + \sigma^2)$. The moment equations are given by

$$\begin{aligned} n^{-1}\{X_1 + X_2 + \cdots + X_n\} &= \mu; \\ n^{-1}\{X_1^2 + X_2^2 + \cdots + X_n^2\} &= \mu^2 + \sigma^2. \end{aligned}$$

The moment estimators are found to be

$$\hat{\mu} = \bar{X}_n; \quad \hat{\sigma}^2 = n^{-1}\sum X_i^2 - \bar{X}_n^2.$$

Note that $\hat{\sigma}^2$ differs from the sample variance by a scale factor $n/(n-1)$.

Moment estimators are often easy to construct and have simple distributional properties. In classical examples, they are also easy to compute numerically.

The use of moment estimator depends on the existence and also uniqueness of the solutions to the corresponding equations. There seem to be little discussion on this topic. We suggest that moment estimators are estimators of ancient tradition in which era only simplistic models were considered. Such complications do not seem to occur too often for these models. We will provide an example based on exponential mixture as an exercise problem. One may find the classical example in Pearson (1904?) where a heroic effort was devoted to solve moment equations to fit a two-component normal mixture model. Other than it is a general convention, there exists nearly no theory to support the use of the first d moments for the method of moments rather than other moments. The method of moments also does not have to be restricted to situations where i.i.d. observations are available.

Example 5.2. *Suppose we have T observations from a simple linear regression model:*

$$Y_t = \beta X_t + \epsilon_t$$

for $t = 0, 1, \ldots, T$, such that $\epsilon_1, \ldots, \epsilon_T$ are i.i.d. N(0, 1) and $X_1, \ldots, X_T$ are non-random constants.

It is seen that

$$\mathbb{E}\{\sum Y_t\} = \beta \sum X_t.$$

Hence, a moment estimator of β is given by

$$\hat{\beta} = \frac{\sum Y_t}{\sum X_t}.$$

The method of moments makes sense based on our intuition. What statistical properties does it have? Under some conditions, we can show that it is consistent and asymptotically normal. Specifying exact conditions, however, is surprisingly more tedious than we may expect.

Consider the situation where an i.i.d. sample of size n from a parametric statistical model $\mathcal{F}$ is available. Let θ denote the parameter and $\Theta \subset \mathcal{R}^d$ be the parameter space. Let $\mu_k(\theta)$ be the kth moment of X, the random variable whose distribution is $F(x; \theta)$ which is a member of $\mathcal{F}$.

Assume that $\mu_k(\theta)$ exists and continuous in θ for $k = 1, 2, \ldots, d$. Assume also the moment estimator of θ, $\hat{\theta}$ is a unique solution to moment equations for large enough n. Recall the law of large numbers:

$$n^{-1}\{X_1^k + X_2^k + \cdots + X_n^k\} \to \mu_k(\theta)$$

almost surely when $n \to \infty$.

By the definition of moment estimates, we have

$$\mu_k(\hat{\theta}) \to \mu_k(\theta)$$

for $k = 1, 2, \ldots, d$ when $n \to \infty$, almost surely.

Assume that as a vector valued function made of first d moments, $\mu(\theta)$ "inversely continuous" a term we invent on spot: for any fixed θ^* and dynamic θ,

$$\|\mu(\theta) - \mu(\theta^*)\| \to 0$$

only if $\theta \to \theta^*$. Then, $\mu_k(\hat{\theta}) \to \mu_k(\theta)$ almost surely implies $\hat{\theta} \to \theta$ almost surely.

We omit the discussion of asymptotic normality here.

5.2 Maximum likelihood estimation

If one can find a σ-finite measure such that each distribution in $\mathcal{F}$ has a density function $f(x)$. Then the likelihood function is given by (not defined as)

$$L(F) = f(x)$$

which is a function of F on $\mathcal{F}$. To remove the mystic notion of $\mathcal{F}$, under parametric model, the likelihood becomes

$$L(\theta) = f(x; \theta)$$

because we can use θ to represent each F in $\mathcal{F}$. If $\hat{\theta}$ is a value in Θ such that

$$L(\hat{\theta}) = \sup_{\theta} L(\theta)$$

then it is a maximum likelihood estimate (estimator) of θ. If we can find a sequence $\{\theta_m\}_{m=1}^{\infty}$ such that

$$\lim_{m\to\infty} L(\theta_m) = \sup_{\theta} L(\theta)$$

and $\lim \theta_m = \hat{\theta}$ exists, then we also call $\hat{\theta}$ a maximum likelihood estimate (estimator) of θ.

The observation x includes the situation where it is a vector. The common i.i.d. situation is a special case where x is made of n i.i.d. observations from a distribution family $\mathcal{F}$. In this case, the likelihood function is given by the product of n densities evaluated at $x_1, \ldots, x_n$ respectively. It remains a function of parameter θ.

The probability mass function, when x is discrete, is also regarded as a density function. This remark looks after discrete models. In general, the likelihood function is defined as follows.

Definition 5.1. *The likelihood function on a model $\mathcal{F}$ based on observed values of X is proportional to*

$$P(X = x; F)$$

where the probability is computed when X has distribution F.

When F is a continuous distribution, the probability is computed as the probability of the event "when X belongs to a small neighbourhood of x". The argument of "proportionality" leads to the joint density function $f(x)$ or $f(x;\theta)$ in general. *The proportionality is a property in terms of F. The likelihood function is a function of F.*

The phrase "proportional to" in the definition implies the likelihood function is not unique. If $L(\theta)$ is a likelihood function based on some data, then $cL(\theta)$ for any $c > 0$ is also a likelihood function based on the same data.

5.3 Estimating equation

The MLE of a parameter is often obtained by solving a score equation:

$$\frac{\partial L_n(\theta)}{\partial \theta} = 0.$$

It is generally true that

$$\mathbb{E}\left[\frac{\partial \log L_n(\theta)}{\partial \theta}; \theta\right] = 0$$

where the expectation is computed when the parameter value (of the distribution of the data) is given by θ. Because of this, the MLE is often regarded as a solution to

$$\frac{\partial \log L_n(\theta)}{\partial \theta} = 0.$$

It appears that whether or not $\partial \log L_n(\theta)/\partial\theta$ is the derivative function of the log likelihood function matters very little. This leads to the following consideration.

In applications, we have reasons to justify that a parameter θ solves equation

$$\mathbb{E}\{g(X;\theta)\} = 0.$$

Given an set of i.i.d. observations in X, we may solve

$$\sum_{i=1}^{n} g(x_i;\theta) = 0$$

and use its solution as an estimate of θ (or estimator if x_i's are replaced by X_i).

Clearly, such estimators are sensible and may be preferred when completely specifying a model for X is at great risk of misspecification.

Example 5.3. *Suppose $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$ is a set of i.i.d. observations from some $\mathcal{F}$ such that $\mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \mathbf{x}_i^\tau \boldsymbol{\beta}$.*

We may estimate $\boldsymbol{\beta}$ by the solution to

$$\sum_{i=1}^{n} \mathbf{x}_i^\tau (y_i - \mathbf{x}_i^\tau \boldsymbol{\beta}) = 0.$$

The solution is given by

$$\hat{\boldsymbol{\beta}} = \{\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\tau\}^{-1} \{\sum_{i=1}^{n} \mathbf{x}_i y_i\}$$

which is the well known least squares estimator.

The spirit of this example is: we do not explicitly spell out any distributional assumptions on $(\mathbf{X}, Y)$ other than the form of the conditional expectation.

5.4 M-Estimation

Motivated from a similar consideration, one may replace $L_n(\theta)$ by some other functions in some applications. Let $\varphi(x;\theta)$ be a function of data and θ but we mostly interested in its functional side in θ after x is given. In the i.i.d. case, we may maximize

$$M_n(\theta) = \sum_{i=1}^{n} \varphi(x_i;\theta)$$

use its solution as an estimate of θ (or estimator if x_i's are replaced by X_i). In this situation, parameter θ is defined as the solution to the minimum point of $\mathbb{E}\{\varphi(X;\xi);F\}$ in ξ where F is the true distribution of X.

Example 5.4. *Suppose* $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$ *is a set of i.i.d. observations from some* $\mathcal{F}$ *such that* $\mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \mathbf{x}_i^{\tau}\boldsymbol{\beta}$.

We may estimate $\boldsymbol{\beta}$ *by the solution to the minimization/optimization problem:*

$$\min_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\tau}\boldsymbol{\beta})^2 \Big\}.$$

In this case, $\varphi(x, y; \boldsymbol{\beta}) = (y - \mathbf{x}^{\tau}\boldsymbol{\beta})^2$. *The solution is again given by*

$$\hat{\boldsymbol{\beta}} = \{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\tau}\}^{-1}\{\sum_{i=1}^{n} \mathbf{x}_i y_i\}$$

which is the well known least squares estimator.

In some applications, the data set may contain a few observations whose y *values are much much larger than the rest of observations. Their presence makes the other observed values have almost no influence on the fitted regression coefficient* $\hat{\boldsymbol{\beta}}$. *Hence, Huber suggested to use*

$$\varphi(x, y; \boldsymbol{\beta}) = \begin{cases} (y - \mathbf{x}^{\tau}\boldsymbol{\beta})^2 & |y - \mathbf{x}^{\tau}\boldsymbol{\beta}| \le k \\ k(y - \mathbf{x}^{\tau}\boldsymbol{\beta}) & y - \mathbf{x}^{\tau}\boldsymbol{\beta} > k \\ -k(y - \mathbf{x}^{\tau}\boldsymbol{\beta}) & y - \mathbf{x}^{\tau}\boldsymbol{\beta} < -k \end{cases}$$

for some selected constant k *instead.*

This choice limits the influences of observations with huge values. Sometimes, such abnormal values, often referred to as outliers, are caused by recording errors.

5.5 L-estimator

Suppose we have a set of univariate i.i.d. observations and and it is simple to record them in terms of sizes such that $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. We call them order statistics. To avoid the influence of outliers, one may estimate the population mean by a trimmed mean:

$$(n-2)^{-1}\sum_{i=2}^{n-1} X_{(i)}.$$

This practice is used on Olympic games though theirs are not estimators. One can certainly remove more observations from consideration and make the estimator more robust. The extreme case is to use the sample median to estimate the population mean. In this case, the estimator makes sense only if the mean and median are the same parameters under the model assumption.

In general, an L-estimator is any linear combination of these order statistics. The coefficients are required to be non-random and do not depend on unknown parameters.

Chapter 6

Maximum likelihood estimation

In textbooks such as here, we have plenty of examples where the solutions to MLE are easy to obtain. We now give some examples where the routine approaches do not work.

6.1 MLE examples

The simplest example is when we have i.i.d. data of size n from $N(\mu, \sigma^2)$ distribution (family). In this case, the log-likelihood function is given by

$$\ell_n(\mu, \sigma^2) = -n \log \sigma - \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Note that I have omitted the constant that does not depend on parameters. Regardless of the value of σ^2, the maximum point in μ is $\hat{\mu} = \bar{X}_n$, the sample mean. Let $\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (x_i - \mu)^2$ and do not regard it as an estimator but a statistic for the moment. Then, we find

$$\ell_n(\hat{\mu}, \sigma^2) = -n \log \sigma - \frac{n\tilde{\sigma}^2}{\sigma^2}.$$

This function is maximized at $\sigma^2 = \tilde{\sigma}^2$. Hence, the MLE of σ^2 is given by $\hat{\sigma}^2 = \tilde{\sigma}^2$.

Type I censor. The next example is a bit unusual. In industry, it is vital to ensure that components in a product will last for a long time. Hence, we need to have a clear idea on their survival distributions. Such information can be obtained by collecting complete failure time data on a random sample of the components. When the average survival time is very long, one has to terminate the experiment at some point, likely before all samples fail. Let the life time of a component be X and the termination time be nonrandom T. Then, the observation may be censored and we only observe $\min(X, T)$. This type of censorship is commonly referred to as type I censor.

Suppose the failure time data can be properly modelled by exponential distribution $f(x; \theta) = \theta^{-1} \exp(-x/\theta), \quad x > 0$. Let $x_1, x_2, \cdots, x_m$ be the observed failure times of m out of n components. The rest of $n-m$ components have not experienced failure at time T (which is not random). In this case, the likelihood function would be given by

$$L_n(\theta) = \theta^{-m} \exp\left\{ -\theta^{-1}[\sum_{i=1}^{m} x_i + (n-m)T] \right\}.$$

Interpreting likelihood function based on the above definition makes it easier to obtained the above expression.

Some mathematics behind this likelihood is as follows. To observe that $n-m$ components lasted longer than T, the probability of this event is given by

$$\binom{n}{n-m} \{\exp(-\theta^{-1}T)\}^{n-m} \{1 - \exp(-\theta^{-1}T)\}^{m}.$$

Given m components failed before time T, the joint distribution is equivalent to an i.i.d. conditional exponential distribution whose density is given by

$$\frac{\theta^{-1} \exp(-\theta^{-1}x)}{1 - \exp(-\theta^{-1}T)}.$$

Hence, the joint density of $x_1, \ldots, x_m$ is given by

$$\prod_{i=1}^{m} \left[\frac{\theta^{-1} \exp(-\theta^{-1}x_i)}{1 - \exp(-\theta^{-1}T)} \right].$$

The product of two factors gives us the algebraic expression of $L_n(\theta)$. Once the likelihood function is obtained, we can find the explicit solution to the MLE of θ easily.

There are more than one way to arrive at the above likelihood function.

Discrete parameter space. Suppose a finite population is made of two types of units, A and B. The population size $N = A + B$ units where A and B also denote the number of types A and B units. Assume the value of B is known which occurs in capture-recapture experiment. A sample of size n is obtained by "simple random sample without replacement" and x of the sampled units are of type B.

Based on this observation, what is the MLE of A?

To answer this question, we notice that the likelihood function is given by

$$L(A) = \frac{\binom{A}{n-x}\binom{B}{x}}{\binom{A+B}{n}}.$$

Our task is to find an expression of the MLE of A. Note that "find the MLE" is not very rigorous statement.

Let us leave this problem to classroom discussion.

Non-smooth density functions. Suppose we have an i.i.d. sample of size n from uniform distribution on $(0, \theta)$ and the parameter space is $\Theta = \mathcal{R}^+$. Find the MLE of θ.

6.2 Newton Raphson algorithm

Other than textbook examples, most applied problems do not permit an analytical solutions to the maximum likelihood estimation. In this case, we resort to any optimization algorithms that work. For illustration, we still resort to "textbook examples."

Example 6.1. *Let $X_1, \ldots, X_n$ be i.i.d. random variables from Weibull distribution with fixed scale parameter:*

$$f(x; \theta) = \theta x^{\theta-1} \exp(-x^\theta)$$

with parameter space $\Theta = \mathcal{R}^+$ on support $x > 0$.

Clearly, the log likelihood function of θ is given by

$$\ell_n(\theta) = n \log \theta + (\theta - 1) \sum_{i=1}^n \log x_i - \sum_{i=1}^n x_i^\theta.$$

It is seen that

$$\begin{aligned} \ell_n'(\theta) &= \frac{n}{\theta} + \sum_{i=1}^n \log x_i - \sum_{i=1}^n x_i^\theta \log x_i; \\ \ell_n''(\theta) &= -\frac{n}{\theta^2} - \sum_{i=1}^n x_i^\theta \log^2 x_i < 0. \end{aligned}$$

Therefore, the likelihood function is convex and hence has unique maximum in θ. Either when $\theta \to 0_+$ and when $\theta \to \infty$, we have $\ell_n(\alpha) \to -\infty$.

For numerical computation, we can easily locate $\theta_1 < \theta_2$ such that the maximum point of $\ell_n(\theta)$ is within the interval $[\theta_1, \theta_2]$.

Following the above example, a bisection algorithm can be applied to locate the maximum point of $\ell_n(\theta)$.

1. Compute $y_1 = \ell_n'(\theta_1), y_2 = \ell_n'(\theta_2)$ and $\theta = (\theta_1 + \theta_2)/2$;

2. If $\ell_n'(\theta) > 0$, let $\theta_1 = \theta$; otherwise, $\theta_2 = \theta$;

3. Repeat the last step until $|\theta_1 - \theta_2| < \epsilon$ for a pre-specified precision constant $\epsilon > 0$.

4. Report θ as the numerical value of the MLE $\hat{\theta}$.

It will be an exercise to numerically find an upper and lower bounds and the MLE of θ given a data set.

The bisection method is easy to understand. Its convergence rate, in terms of how many steps it must take to get the final result is judged not high enough. When θ is one dimensional, our experience shows the criticism is not well founded. Nevertheless, it is useful to understand another standard method in numerical data analysis.

Suppose one has an initial guess of the maximum point of the likelihood function, say $\theta^{(0)}$. For any θ close to this point, we have

$$\ell_n'(\theta) \approx \ell_n'(\theta^{(0)}) + \ell_n''(\theta^{(0)})(\theta - \theta^{(0)}).$$

If the initial guess is pretty close to the maximum point, then the value of the second derivative $\ell''_n(\theta^{(0)}) < 0$. From the above approximation, we would guess that

$$\theta^{(1)} = \theta^{(0)} - \ell'_n(\theta^{(0)})/\ell''_n(\theta^{(0)})$$

is closer to the solution of $\ell'_n(\theta) = 0$. This consideration leads to repeated updating:

$$\theta^{(k+1)} = \theta^{(k)} - \ell'_n(\theta^{(k)})/\ell''_n(\theta^{(k)}).$$

Starting from $\theta^{(0)}$, we therefore obtain a sequence $\theta^{(k)}$. If the problem is not tricky, this sequence converges to the maximum point of $\ell_n(\theta)$. Once it stabilizes, we regard the outcome as the numerical value of the MLE.

The iterative scheme is called Newton-Raphson method. Its success depends on a good choice of $\theta^{(0)}$ and the property of $\ell_n(\theta)$ as a function of θ. If the likelihood has many local maxima, then the outcome of the algorithm can be one of these local maxima. For complex models and multi-dimensional θ, the convergence is far from guaranteed. The good/lucky choice of $\theta^{(0)}$ is crucial.

Although in theory, each iteration moves $\theta^{(k+1)}$ toward true maximum faster by using Newton-Raphson method, we pay extra cost on computing the second derivation. For multi-dimensional θ, we need to invert a matrix which is not always a pleasant task. The implementation of this method is not always so simple.

Implementing Newton-Raphson for a simple data example will be an exercise.

Example 6.2. Logistic distribution. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with density function*

$$f(x;\theta) = \frac{\exp\{-(x-\theta)\}}{[1+\exp\{-(x-\theta)\}]^2}.$$

The support of the distribution is the whole line, and parameter space is $\mathcal{R}$. We usually call it a location distribution family.

The log-likelihood function is give by

$$\ell_n(\theta) = n\theta - n\bar{x}_n - 2\sum_{i=1}^{n} \log[1+\exp\{-(x_i-\theta)\}].$$

Its score function is

$$\ell_n'(\theta) = s(\theta) = n - 2\sum_{i=1}^{n} \frac{\exp\{-(x-\theta)\}}{1+\exp\{-(x-\theta)\}}.$$

The MLE is a solution to $s(\theta) = 0$.

One may easily find that

$$\ell_n''(\theta) = s'(\theta) = -2\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{[1+\exp\{-(x_i-\theta)\}]^2} < 0.$$

Thus, the score function is monotone in θ, which implies the solution to $s(\theta) = 0$ is unique. It also implies that the solution is the maximum point of the likelihood, not minimum nor stationary points.

It is also evident that there is no analytical solution to this equation, Newton-Raphson algorithm can be a good choice for numerically evaluate the MLE in applications.

6.3 EM-algorithm

Suppose we have n observations from a tri-nomial distribution. That is, there are n independent and independent trials each has 3 possible outcomes. The corresponding parameters are p_1, p_2, p_3. We summarize these observations into n_1, n_2, n_3. The log-likelihood function is

$$\ell_n(p_1, p_2, p_3) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3.$$

Using Lagrange method, we can easily show that the MLEs are

$$\hat{p}_j = n_j/n$$

for $j = 1, 2, 3$.

If, however, another m trials were carried out but we know only their outcomes are not of the third kind. In some words, the data contains some missing information.

The log-likelihood function when the additional data are included becomes

$$\ell_n(p_1,p_2,p_3) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + m \log(p_1+p_2).$$

Working out the MLE is no longer straightforward now. Given specific values, there are many numerical algorithms can be used to compute MLE. We recommand EM-algorithm in this case.

If we knew which of these m observations were of type I, we would have obtained the complete data log-likelihood as:

$$\ell_c(p_1,p_2,p_3) = (n_1+m_1)\log p_1 + (n_2+m_2)\log p_2 + n_3 \log p_3$$

where c stands for "complete data". Since we do not know what these m_1 and m_2 are, we replace them with some predictions based on what we know already. In this case, we use conditional expectations.

E-step: If the current estimates $\hat{p}_1 = n_1/n$ and $\hat{p}_2 = n_2/n$ are relevant. Then, we might expect that out of m non-type III observations, $\hat{m}_1 = m\hat{p}_1/(\hat{p}_1+\hat{p}_2)$ are of type I and $\hat{m}_2 = m\hat{p}_2/(\hat{p}_1+\hat{p}_2)$ are of type II. That is, the conditional expectation (given data, and the current estimates of the parameter values) of m_1 and m_2 are given by $\hat{m}_1$ and $\hat{m}_2$. When m_1 and m_2 are replaced by their conditional expectations, we get a function

$$Q(p_1,p_2,p_3) = (n_1+\hat{m}_1)\log p_1 + (n_2+\hat{m}_2)\log p_2 + n_3 \log p_3.$$

This is called E-stap because we Replace the unobserved values by their conditional expectations.

M-step: In this step, we update unknown parameters by the maximizer of $Q(p_1,p_2,p_3)$. The updated estimator values are

$$\tilde{p}_1 = (n_1+m_1)/(n+m) \;\; \tilde{p}_2 = (n_2+m_2)/(n+m), \;\; \tilde{p}_3 = n_3/(n+m).$$

If they represent a better guess of the MLE, then we should update the Q-function accordingly. After which, we should carry out the M-step again to obtain more satisfactory approximation to the MLE. We therefore iterate between the E and M steps until some notion of convergence.

These idea is particularly useful when the data structure is complex. In most cases, the EM iteration is guaranteed to increase the likelihood. Thus, it should converge, and converge to a local maximum for the least.

6.4 EM-algorithm for finite mixture models

Let envisage a population made of a finite number of subpopulations, each is governed by a specific distribution from some distribution family. Taking a random sample from a finite mixture model, we obtain a set of units without knowing their subpopulation identities. The resulting random variable has density function

$$f(x; G) = \sum_{j=1}^{m} \pi_j f(x; \theta_j)$$

with G denoting a mixing distribution on parameter space of θ, Θ, by assigning probability π_j on θ_j.

Given a random sample of size n, $x_1, x_2, \ldots, x_n$, from this distribution, the log likelihood function is given by

$$\ell_n(G) = \sum_{i=1}^{n} \log f(x_i; G). \tag{6.1}$$

Other than order m, we regard π_j, θ_j as parameters to be estimated. Computing the maximum likelihood estimate of G is to find the values of m pairs of π_j and θ_j such that $\ell_n(G)$ is maximized.

Taking advantage of the mixture model structure, EM-algorithm can often be effectively implemented to locate the location of the maximum point of the likelihood function.

Conceptually, each observation x from a mixture model is part of a complete vector observation $(x, \mathbf{z}^\tau)$ where $\mathbf{z}$ is a vector of mostly 0 and a single 1 of length m. The position of 1 is its subpopulation identity. Suppose we have a set of complete observations in the form of $(x_i, \mathbf{z}_i^\tau)$: $i = 1, 2, \ldots, n$. The log likelihood function of the mixing distribution G is given by

$$\ell_c(G) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log\{\pi_j f(x_i; \theta_j)\}. \tag{6.2}$$

Since for each i, z_{ij} equals 0 except for a specific j value, only one $\log\{\pi_j f(x_i; \theta_j)\}$ actually enters the log likelihood function.

We use $\mathbf{x}$ for the vector of the x_i and $\mathbf{X}$ as its corresponding random vector and start the EM-algorithm with an initial mixing distribution with

m support points:

$$G^{(0)}(\theta) = \sum_{j=1}^{m} \pi_j^{(0)} \mathbb{1}(\theta_j^{(0)} \leq \theta).$$

E-Step. This step is to find the expected values of the missing data in the full data likelihood function. They are $\mathbf{z}_i$ in the context of the finite mixture model. If the mixing distribution G is given by $G^{(0)}$, its corresponding random variable has conditional expectation given by

$$\begin{aligned}
\mathbb{E}\{\mathbf{Z}_{ij} | \mathbf{X} = \mathbf{x}; G^{(0)}\} &= \text{PR}(\mathbf{Z}_{ij} = 1 | X_i = x_i; G^{(0)}) \\
&= \frac{f(x_i; \theta_j^{(0)}) \text{PR}(\mathbf{Z}_{ij} = 1; G^{(0)})}{\sum_{k=1}^{m} f(x_i; \theta_k^{(0)}) \text{PR}(\mathbf{Z}_{ik} = 1; G^{(0)})} \\
&= \frac{\pi_j^{(0)} f(x_i; \theta_j^{(0)})}{\sum_{k=1}^{m} \pi_k^{(0)} f(x_i; \theta_k^{(0)})}.
\end{aligned}$$

The first equality has utilized two facts: the expectation of an indicator random variable equals the probability of "success"; only the ith observation is relevant to the subpopulation identity of the ith unit. The second equality comes from the standard Bayes formula. The third one spells out the probability of "success" if $G^{(0)}$ is the true mixing distribution. The superscript (0) reminds us that the corresponding quantities are from $G^{(0)}$, the initial mixing distribution. One should also note the expression is explicit and numerically easy to compute as long as the density function itself can be easily computed.

We use notation $w_{ij}^{(0)}$ for $\mathbb{E}\{\mathbf{Z}_{ij} | \mathbf{X} = \mathbf{x}; G^{(0)}\}$. Replacing z_{ij} by $w_i^{(0)}$ in $\ell^c(G)$, we obtain a function which is usually denoted as

$$Q(G; G^{(0)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^{(0)} \log\{\pi_j f(x_i; \theta_j)\}. \tag{6.3}$$

In this expression, Q is a function of G, and its functional form is determined by $G^{(0)}$. The E-Step ends at producing this function.

M-Step. Given this Q function, it is often simple to find a mixing distribution G having it maximized. Note that Q has the following decomposition:

$$Q(G; G^{(0)}) = \sum_{j=1}^{m} \big\{ \sum_{i=1}^{n} w_{ij}^{(0)} \big\} \log(\pi_j) + \sum_{j=1}^{m} \big\{ \sum_{i=1}^{n} w_{ij}^{(0)} \log f(x_i; \theta_j) \big\}.$$

In this decomposition, two additive terms are functions of two separate parts of G. The first term is a function of mixing probabilities only. The second term is a function of subpopulation parameters only. Hence, we can search for the maxima of these two functions separately to find the overall solution.

The algebraic form of the first term is identical to the log likelihood of a multinomial distribution. The maximization solution is given by

$$\pi_j^{(1)} = n^{-1} \sum_{i=1}^{n} w_{ij}^{(0)}$$

for $j = 1, 2, \ldots, m$.

The second term is further decomposed into the sum of m log likelihood functions, one for each subpopulation. When $f(x;\theta)$ is a member of classical parametric distribution family, then the maximization with respect to θ often has an explicit analytical solution. With a generic $f(x;\theta)$, we cannot give an explicit expression but an abstract one:

$$\theta_j^{(1)} = \arg\sup_{\theta} \{ \sum_{i=1}^{n} w_{ij}^{(0)} \log f(x_i; \theta_j) \}.$$

The mixing distribution

$$G^{(1)}(\theta) = \sum_{j=1}^{m} \pi_j^{(1)} \mathbb{1}(\theta_j^{(1)} \le \theta)$$

then replaces the role of $G^{(0)}$ and we go back to E-step.

Iterating between E-step and M-step leads to a sequence of intermediate estimates of the mixing distribution: $G^{(k)}$. Often, this sequence converges to at least a local maximum of $\ell_n(G)$.

With some luck, the outcome of this limit is the global maximum. In most applications, one would try a number of $G^{(0)}$ and compare the values of $\ell_n(G^{(k)})$ the EM-algorithm leads to. The one with the highest value will have its $G^{(k)}$ regarded as the maximum likelihood estimate of G.

The algorithm stops after many iterations when the difference between $G^{(k)}$ and $G^{(k-1)}$ is considered too small to continue. Other convergence criteria may also be used.

6.4.1 Data Examples

Leroux and Puterman (1992) and Chen and Kalbfleisch (1996) analyze data on the movements of a fetal lamb in each of 240 consecutive 5-second intervals and propose a mixture of Poisson distributions. The observations can be summarized by the following table.

x	0	1	2	3	4	5	6	7
freq	182	41	12	2	2	0	0	1

It is easily seen that the distribution of the counts is over-dispersed. The sample mean is 0.358 which is significantly smaller than the sample variance which is 0.658 given that the sample size is 240.

A finite mixture model is very effective at explaining the over-dispersion. There is a general agreement that a finite Poisson mixture model with order $m = 2$ is most suitable. We use this example to demonstrate the use of EM-algorithm for computing the MLE of the mixing distribution given $m = 2$.

Since the sample mean is 0.358 and the data contains a lot of zeros. Let us choose an initial mixing distribution with

$$(\pi_1^{(0)}, \pi_2^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}) = (0.7, 0.3, 0.1, 4.0).$$

We do not have more specific reasons behind the above choice.

A simplistic implementation of EM-algorithm for this data set is as follows.

```
pp = 0.7;
theta = c(0.1, 4.0)
xx = c(rep(0, 182), rep(1, 41), rep(2, 12), 3, 3, 4, 4, 7)
   #data inputted, initial mixing distribution chosen

last = c(pp, theta)
dd= 1
while(dd > 0.000001) {
temp1 = pp*dpois(xx, theta[1])
temp2 = (1-pp)*dpois(xx, theta[2])
w1 = temp1/(temp1+temp2)
w2 = 1 - w1
#E-step completed
pp = mean(w1)
theta[1] = sum(w1*xx)/sum(w1)
theta[2] = sum(w2*xx)/sum(w2)
#M-step completed
updated = c(pp, theta)
dd = sum((last - updated)^2)
last = updated
}
print(updated)
```

When the EM-algorithm converges, we get $\hat{\pi}_1 = 0.938$ and $\hat{\theta}_1 = 0.229$, $\hat{\theta}_2 = 2.307$. The likelihood value at this $\hat{G}$ equals -186.99 (based on the usual expression of the Poisson probability mass function). The fitted frequency vector is given by

x	0	1	2	3	4	5	6	7
freq	182	41	12	2	2	0	0	1
fitted freq	180.4	44.5	8.6	3.4	1.8	0.8	0.3	0.1

6.5 EM-algorithm for finite mixture models repeated

Let envisage a population made of a finite number of subpopulations, each is governed by a specific distribution from some distribution family. Taking a random sample from a finite mixture model, we obtain a set of units without knowing their subpopulation identities. The resulting random variable has density function

$$f(x; G) = \sum_{j=1}^{m} \pi_j f(x; \theta_j)$$

with G denoting a mixing distribution on parameter space of θ, Θ, by assigning probability π_j on θ_j.

Given a random sample of size n, $x_1, x_2, \ldots, x_n$, from this distribution, the log likelihood function is given by

$$\ell_n(G) = \sum_{i=1}^{n} \log f(x_i; G). \tag{6.4}$$

Other than order m, we regard π_j, θ_j as parameters to be estimated. Computing the maximum likelihood estimate of G is to find the values of m pairs of π_j and θ_j such that $\ell_n(G)$ is maximized.

Taking advantage of the mixture model structure, EM-algorithm can often be effectively implemented to locate the location of the maximum point of the likelihood function.

Conceptually, each observation x from a mixture model is part of a complete vector observation (x, z) where z takes values j with probability π_j for $j = 1, 2, \ldots, m$.

Suppose we have a set of complete observations in the form of (x_i, z_i): $i = 1, 2, \ldots, n$. The log likelihood function of the mixing distribution G is given by

$$\ell_c(G) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}(z_i = j) \log\{\pi_j f(x_i; \theta_j)\}. \tag{6.5}$$

Clearly, only one $\log\{\pi_j f(x_i; \theta_j)\}$ actually enters the log likelihood function.

We use $\mathbf{x}$ for the vector of the x_i and $\mathbf{X}$ as its corresponding random vector and start the EM-algorithm with an initial mixing distribution with m support points:

$$G^{(0)}(\theta) = \sum_{j=1}^{m} \pi_j^{(0)} \mathbb{1}(\theta_j^{(0)} \le \theta).$$

E-Step. This step is to find the expected values of the missing data in the full data likelihood function. If the mixing distribution G is given by $G^{(0)}$, its corresponding random variable has conditional expectation given by

$$\begin{aligned}
\mathbb{E}\{\mathbb{1}(Z_i = j) | \mathbf{X} = \mathbf{x}; G^{(0)}\} &= \frac{f(x_i; \theta_j^{(0)}) \text{PR}(Z_i = j; G^{(0)})}{\sum_{k=1}^{m} f(x_i; \theta_k^{(0)}) \text{PR}(Z_i = j; G^{(0)})} \\
&= \frac{\pi_j^{(0)} f(x_i; \theta_j^{(0)})}{\sum_{k=1}^{m} \pi_k^{(0)} f(x_i; \theta_k^{(0)})}.
\end{aligned}$$

The first equality has utilized two facts: the expectation of an indicator random variable equals the probability of "success"; only the ith observation is relevant to the subpopulation identity of the ith unit. The second equality comes from the standard Bayes formula. The third one spells out the probability of "success" if $G^{(0)}$ is the true mixing distribution. The superscript (0) reminds us that the corresponding quantities are from $G^{(0)}$, the initial mixing distribution. One should also note the expression is explicit and numerically easy to compute as long as the density function itself can be easily computed.

We use notation $w_{ij}^{(0)}$ for $\mathbb{E}\{\mathbb{1}(Z_i = j)|\mathbf{X} = \mathbf{x}; G^{(0)}\}$. Replacing $\mathbb{1}(Z_i = j)$ by $w_i^{(0)}$ in $\ell^c(G)$, we obtain a function which is usually denoted as

$$Q(G; G^{(0)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^{(0)} \log\{\pi_j f(x_i; \theta_j)\}. \tag{6.6}$$

In this expression, Q is a function of G, and its functional form is determined by $G^{(0)}$. The E-Step ends at producing this function. In other words, $Q(G; G^{(0)})$ is the conditional expectation of $\ell_c(G)$ when $\mathbf{X} = \mathbf{x}$ are given, and $G^{(0)}$ is regarded as the true mixing distribution behind $\mathbf{X}$.

M-Step. Given this Q function, it is often simple to find a mixing distribution G having it maximized. Note that Q has the following decomposition:

$$Q(G; G^{(0)}) = \sum_{j=1}^{m} \Big\{ \sum_{i=1}^{n} w_{ij}^{(0)} \Big\} \log(\pi_j) + \sum_{j=1}^{m} \Big\{ \sum_{i=1}^{n} w_{ij}^{(0)} \log f(x_i; \theta_j) \Big\}.$$

In this decomposition, two additive terms are functions of two separate parts of G. The first term is a function of mixing probabilities only. The second term is a function of subpopulation parameters only. Hence, we can search for the maxima of these two functions separately to find the overall solution.

The algebraic form of the first term is identical to the log likelihood of a multinomial distribution. The maximization solution is given by

$$\pi_j^{(1)} = n^{-1} \sum_{i=1}^{n} w_{ij}^{(0)}$$

for $j = 1, 2, \ldots, m$.

The second term is further decomposed into the sum of m log likelihood functions, one for each subpopulation. When $f(x;\theta)$ is a member of a classical parametric distribution family, then the maximization with respect to θ often has an explicit analytical solution. With a generic $f(x;\theta)$, we cannot give an explicit expression but an abstract one:

$$\theta_j^{(1)} = \arg\sup_{\theta}\{\sum_{i=1}^{n} w_{ij}^{(0)} \log f(x_i;\theta_j)\}$$

for $j = 1, 2, \ldots, m$.

The mixing distribution

$$G^{(1)}(\theta) = \sum_{j=1}^{m} \pi_j^{(1)} \mathbb{1}(\theta_j^{(1)} \le \theta)$$

is an updated estimate of G from $G^{(0)}$ based on data. We then replace the role of $G^{(0)}$ by $G^{(1)}$ and go back to E-step.

Iterating between E-step and M-step leads to a sequence of intermediate estimates of the mixing distribution: $G^{(k)}$. Often, this sequence converges to at least a local maximum of $\ell_n(G)$.

With some luck, the outcome of this limit is the global maximum. In most applications, one would try a number of $G^{(0)}$ and compare the values of $\ell_n(G^{(k)})$ the EM-algorithm leads to. The one with the highest value will have its $G^{(k)}$ regarded as the maximum likelihood estimate of G.

The algorithm stops after many iterations when the difference between $G^{(k)}$ and $G^{(k-1)}$ is considered too small to continue. Other convergence criteria may also be used.

Chapter 7

Properties of MLE

Consider the situation we have have a data set x whose joint density function is a member of distribution family specified by density functions $\{f(x;\theta) : \theta \in \Theta\}$.

Suppose $\eta = g(\theta)$ is an invertible parameter transformation and denote the inverse transformation by $\theta = h(\eta)$ and the parameter space of η be Υ. Clearly, for each θ, there is an η such that

$$f(x;\theta) = f(x;h(\eta)) = \tilde{f}(x;\eta)$$

where we have introduced $\tilde{f}(x;\eta)$ for the function under the new parameterization. In other words,

$$\{f(x;\theta) : \theta \in \Theta\} = \{\tilde{f}(x;\eta) : \eta \in \Upsilon\}.$$

The likelihood functions in these two systems are related by

$$\ell(\theta) = \tilde{\ell}(\eta)$$

for $\eta = g(\theta)$. If $\hat{\theta}$ is a value such that

$$\ell(\hat{\theta}) = \sup_{\theta\in\Theta} \ell(\theta)\},$$

we must also have

$$\tilde{\ell}(g(\hat{\theta})) = \ell(\hat{\theta}) = \sup_{\theta\in\Theta} \ell(\theta) = \sup_{\eta\in\Upsilon} \tilde{\ell}(\eta).$$

Hence, $h(\hat{\theta})$ is the MLE of $\eta = h(\theta)$.

In conclusion, the MLE as a general method for point estimation, is equivariant. If we estimate μ by $\bar{x}$, then we estimate μ^2 by $\bar{x}^2$ in common notation.

Next, we give results to motivate the use of MLE. The following inequality plays an important role.

Jensen's inequality. Let X be a random variable with finite mean and g be a convex function. Then

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

Proof: We give a heuristic proof. Function g is convex if and only if for every set of $x_1, x_2, \ldots, x_n$ and positive numbers $p_1, p_2, \ldots, p_n$ such that $\sum_{i=1}^n p_i = 1$, we have

$$\sum_{i=1}^n p_i g(x_i) \geq g(\sum_{i=1}^n p_i x_i).$$

This essentially proves the inequality when X is a discrete random variable of finite number of possible values. Since every random variable can be approximated by such random variables, we can take a limit to get the general case. This is always possible when X has finite first moment. □

Kulback-Leibler divergence. Suppose $f(x)$ and $g(x)$ are two density functions with respect to some σ-finite measure. The Kulback-Leibler divergence between f and g is defined to be

$$K(f, g) = \mathbb{E}\{\log[f(X)/g(X)]; f\}$$

where the expectation is computed when X has distribution f.

Let $Y = g(X)/f(X)$ and $h(y) = -\log(y)$. It is seen that $h(y)$ is a convex function. It is easily seen that

$$\mathbb{E}\{Y\} \leq 1$$

where the inequality can occur if the support of $f(x)$ is a true subset of that of $g(x)$. In any case, by Jensen's inequality, we have

$$\mathbb{E}\{h(Y)\} \geq h(\mathbb{E}\{Y\}) \geq 0.$$

This implies that

$$K(f, g) \geq 0$$

for any f and g. Clearly, $K(f, f) = 0$.

Because $K(f, g)$ is positive unless $f = g$, it serves as a metric to measure how different g is from f. At the same time, the KL divergence is not a distance in mathematical sense because $K(f, g) \neq K(g, f)$ in general.

Let $\mathcal{F}$ be a parametric distribution family possessing densities $f(x; \theta)$ and parameter space Θ. Let $f(x)$ be simply a density function may or may not be a member of $\mathcal{F}$. If we wish to find a density in $\mathcal{F}$ that is the best approximation to $f(x)$ in KL-divergence sense, a sensible choice is $f(x; \hat{\theta})$ such that

$$\hat{\theta} = \arg\min_{\theta \in \Theta} K(f(x), f(x; \theta)).$$

In most applications, $f(x)$ is not known but we have an i.i.d. sample $X_1, \ldots, X_n$ from it. In this case, we may approximate $K(f(x), f(x; \theta))$ as follows:

$$\begin{aligned} K(f(x), f(x; \theta)) &= \int \log\{f(x)/f(x; \theta)\} f(x) dx \\ &\approx n^{-1} \sum_{i=1}^{n} \log\{f(x_i)/f(x_i; \theta)\} \\ &= n^{-1} \sum_{i=1}^{n} \log\{f(x_i)\} - n^{-1} \ell_n(\theta) \end{aligned}$$

where the second term is the usual log likelihood function. Hence, minimizing KL-divergence is approximately the same as maximizing the likelihood function. The analog goes further to situations where non-i.i.d. observations are available.

Unlike UMVUE or other estimators, MLE does not aim at most precisely determining the best possible value of “true” θ. One may wonder if it measures up if it is critically examined from different angles. This will be the topic of the next section.

7.1 Trivial consistency

Under very general conditions, the MLE is strongly consistent. We work out a simple case her. Consider the situation where $\Theta = \{\theta_j : j = 1, \ldots, k\}$ for some finite k. Assume that

$$F(x; \theta_j) \neq F(x; \theta_l)$$

for at least one x value when $j \neq l$, where $F(x; \theta)$ is the cumulative distribution function of $f(x; \theta)$. The condition means that the model is identifiable by its parameters. We assume an i.i.d. sample from $F(x; \theta_0)$ has been obtained but pretend that we do not know θ_0. Instead, we want to estimate it by the MLE.

Let $\ell_n(\theta)$ be the likelihood function based on the i.i.d. sample of size n. By the strong law of large numbers, we have

$$n^{-1}\{\ell_n(\theta) - \ell_n(\theta_0)\} \to -K(f(x; \theta_0), f(x; \theta))$$

almost surely for any $\theta \in \Theta$. The identifiability condition implies that

$$K(f(x; \theta_0), f(x; \theta)) > 0$$

for any $\theta \neq \theta_0$. Therefore, we have

$$\ell_n(\theta) < \ell_n(\theta_0)$$

almost surely as $n \to \infty$. When there are only finite many choices of θ in Θ, we must have

$$\max\{\ell_n(\theta) : \theta \neq \theta_0\} < \ell_n(\theta_0)$$

almost surely. Hence, the MLE $\hat{\theta}_n = \theta_0$ almost surely.

Let us summarize the result as follows.

Theorem 7.1. *Let $X_1, \ldots, X_n$ be a set of iid sample from the distribution family $\{f(x; \theta) : \theta \in \Theta\}$ and the true value of the parameter is $\theta = \theta_0$.*

Assume the identifiability condition that

$$F(x; \theta^{'}) \neq F(x; \theta^{''}) \tag{7.1}$$

for at least one x whenever $\theta^{'} \neq \theta^{''}$.

Assume also that

$$E|\log f(X;\theta)| < \infty \tag{7.2}$$

for any $\theta \in \Theta$, where the expectation is computed under θ_0.

Then, the MLE $\hat{\theta} \to \theta_0$ almost surely when $\Theta = \{\theta_j : j = 0, 1, \ldots, k\}$ for some finite K.

Although the above proof is very simple. The idea behind it can be applied to prove the general result. For any subset B of Θ, define

$$f(x;B) = \sup_{\theta \in B} f(x;\theta).$$

We assume that $f(x;B)$ is a measurable function of x for all B under consideration. We can generalize the above theorem as follows.

Theorem 7.2. *Let $X_1, \ldots, X_n$ be a set of i.i.d. sample from the distribution family $\{f(x;\theta) : \theta \in \Theta\}$ and that $\Theta = \cup_{j=0}^{k} B_j$ for some finite k. Assume that the true value of the parameter is $\theta = \theta_0 \in B_0$ and that*

$$E|\log f(X;B_j)| < E[\log f(X;\theta_0)] \tag{7.3}$$

for $j = 1, 2, \ldots, k$. Then, the MLE $\hat{\theta} \in B_0$ almost surely.

7.2 Trivial consistency for one-dimensional θ

Consider the situation where we have a set of i.i.d. observations from a one-dimensional parametric family $\{f(x;\theta) : \theta \in \Theta \subset R\}$. The log likelihood function remains the same as

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

We likely have defined score function earlier, which is, given i.i.d. observations

$$S_n(\theta;x) = \sum_{i=1}^{n} \frac{\partial\{\log f(x_i;\theta)\}}{\partial\theta}.$$

We will use plain $S(\theta; x)$ if when x is regarded as a single observation. We can be sloppy by using notation $\mathbb{E}\{S(\theta)\}$ in which x has to be interpreted as the random variable X whose distribution is $f(x;\theta)$, with the same θ in S and f.

Let us put done a few regularity conditions. They are not most general but suffice in the current situation.

R0 The parameter space of θ is an open set of $\mathbb{R}$.

R1 $f(x;\theta)$ is differentiable to order three with respect to θ at all x.

R2 For each $\theta_0 \in \Theta$, there exist functions $g(x)$, $H(x)$ such that for all θ in a neighborhood $N(\theta_0)$,

$$\begin{aligned} &(i) && \left|\frac{\partial f(x;\theta)}{\partial \theta}\right| \leq g(x);\\ &(ii) && \left|\frac{\partial^2 f(x;\theta)}{\partial \theta^2}\right| \leq g(x);\\ &(iii) && \left|\frac{\partial^3 \log f(x;\theta)}{\partial \theta^3}\right| \leq H(x) \end{aligned}$$

hold for all x, and

$$\int g(x)dx < \infty; \quad \mathbb{E}_0\{H(X)\} < \infty.$$

R3 For each $\theta \in \Theta$,

$$0 < \mathbb{E}_\theta \left\{\frac{\partial \log f(x;\theta)}{\partial \theta}\right\}^2 < \infty.$$

Although the integration is stated as with respect to dx, the results we are going to state remain valid if it is replace by some σ-finite measure. For instance, the result is applicable to MLE under Poisson model where dx must be replaced by summation over non-negative integers. All conditions are stated as they are required for all x. An exception over a 0-measure set of x is allowed, as long as this 0-measure set is the same for all $\theta \in \Theta$.

Lemma 7.1. *(1) Under regularity conditions, we have*

$$\mathbb{E}\left\{\frac{\partial \log f(X;\theta)}{\partial\theta};\theta\right\} = 0.$$

(2) Under regularity conditions, we have

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

Proof. We first remark that the first result is the same as stating $\mathbb{E}\{S(\theta)\} = 0$. The proof of one is based on the fact that

$$\int f(x;\theta)dx = 1.$$

Taking derivative with respect to θ on both sizes, permitting the exchange of derivative and integration under regularity condition R2, and expressing the resultant properly, we get result (1).

To prove (2), notice that

$$\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2} = \left\{\frac{f''(X;\theta)}{f(X;\theta)}\right\} - \left\{\frac{f'(X;\theta)}{f(X;\theta)}\right\}^2.$$

The result is obtained by taking expectation on both sizes and the fact

$$\mathbb{E}\left\{\frac{f''(X;\theta)}{f(X;\theta)}\right\} = \int f''(x;\theta)dx = 0.$$

This completes the proof. □

We now give a simple consistency proof when θ is one-dimensional.

Theorem 7.3. *Given an i.i.d. sample of size n from some one-parameter model $\{f(x;\theta) : \theta \in \Theta \subset \mathcal{R}\}$. Suppose θ^* is the true parameter value. Under Conditions R0-R3, there exists an $\hat{\theta}_n$ sequence such that*

(i) $S_n(\hat{\theta}_n) = 0$ almost surely;

(ii) $\hat{\theta}_n \to \theta^*$ almost surely.

Proof. (i) As a function of θ, $\mathbb{E}\{S(\theta)\}$ has derivative equaling $-\mathbb{I}(\theta^*)$ at $\theta = \theta^*$. Hence, it is a decreasing function at θ^*. This implies the existence of sufficiently small $\epsilon > 0$, such that

$$\mathbb{E}\{S(\theta^* + \epsilon)\} < 0 < \mathbb{E}\{S(\theta^* - \epsilon)\}.$$

By the law of large numbers, we have

$$n^{-1}S_n(\theta^* \pm \epsilon) \xrightarrow{a.s.} \mathbb{E}\{S((\theta^* \pm \epsilon)\}.$$

Hence, almost surely, we have

$$S_n(\theta^* + \epsilon) < 0 < S_n((\theta^* - \epsilon).$$

By intermediate value theorem, there exists a $\hat{\theta} \in (\theta^* - \epsilon, \theta^* + \epsilon)$ such that

$$S_n(\hat{\theta}) = 0.$$

This proves (i).

(ii) is a direct consequence of (i) as ϵ can be made arbitrarily small. □

7.3 Asymptotic normality of MLE after the consistency is established

Under the assumption that $f(x;\theta)$ is smooth, and the MLE $\hat{\theta}$ is a consistent estimator of θ, we must have

$$S_n(\hat{\theta}) = 0.$$

By the mean-value theorem in mathematical analysis, we have

$$S_n(\theta^*) = S_n(\hat{\theta}) + S_n'(\tilde{\theta})(\theta^* - \hat{\theta})$$

where $\tilde{\theta}$ is a parameter value between θ^* and $\hat{\theta}$.

By the result in the last lemma, we have

$$n^{-1}S_n'(\tilde{\theta}) \to -\mathbb{I}(\theta^*),$$

the Fisher information almost surely. In addition, the classical central limit theorem implies

$$n^{-1/2}S_n(\theta^*) \to N(0, \mathbb{I}(\theta^*)).$$

Thus, by Slutzky's theorem, we find

$$\sqrt{n}(\hat{\theta} - \theta^*) = n^{-1/2}\mathbb{I}^{-1}(\theta^*)S_n(\theta^*) + o_p(1) \to N(0, \mathbb{I}^{-1}(\theta^*))$$

in distribution as $n \to \infty$.

Many users including statisticians ignore the regularity conditions. Indeed, they are satisfied by most commonly used models. If one does not bother with the full rigour, he or she should at least make sure that the parameter value in consideration is an interior point, the likelihood function is smooth enough. If the data set does not have i.i.d. structure, one should make sure that some form of uniformity hold.

7.4 Asymptotic efficiency, super-efficient, one-step update scheme

By Cramer-Rao information inequality, for any estimator of θ given i.i.d. data and sufficiently regular model, we have

$$\text{VAR}(\hat{\theta}_n) \geq \mathbb{I}_n^{-1}(\theta^*)$$

for any estimator $\hat{\theta}_n$ assuming unbiasedness. The MLE under regularity conditions has asymptotic variance $\mathbb{I}(\theta^*)$ at rate $\sqrt{n}$. Loosely speaking, the above inequality becomes equality for MLE. Hence, the MLE is "efficient": no other estimators can achieve lower asymptotic variance.

Let us point out the strict interpretation of asymptotic efficiency is not correct. Suppose we have a set of i.i.d. observations from $N(\theta, 1)$. The MLE of θ is $\bar{X}_n$. Clearly, if θ^* is the true value, we have

$$\sqrt{n}(\bar{X}_n - \theta^*) \xrightarrow{d} N(0, 1).$$

Can we do better than the MLE? Let

$$\tilde{\theta}_n = \begin{cases} 0 & \text{if } |\bar{X}_n| \leq n^{-1/4} \\ \bar{X}_n & \text{otherwise.} \end{cases}$$

When the true value $\theta^* = 0$, then

$$\text{PR}(|\bar{X}_n| \leq n^{-1/4}) \to 1$$

as $n \to 0$. Hence,

$$\sqrt{n}(\bar{X}_n - \theta^*) \xrightarrow{d} N(0, 0)$$

with asymptotic variance 0 at rate $\sqrt{n}$.

When the true value $\theta^* \neq 0$, then

$$\text{PR}(|\bar{X}_n| \leq n^{-1/4}) \to 0$$

which implies

$$\text{PR}(\tilde{\theta}_n = \bar{X}_n) \to 1.$$

Consequently,

$$\sqrt{n}(\tilde{\theta}_n - \theta^*) \xrightarrow{d} N(0, 1).$$

What have we seen? If $\theta^* \neq 0$, then $\tilde{\theta}_n$ has the same limiting distribution as that of $\bar{X}_n$ at the same rate. So they have the same asymptotic efficiency. When $\theta^* = 0$, the asymptotic variance of $\tilde{\theta}_n$ is 0 which is smaller than that of $\bar{X}_n$ (at rate $\sqrt{n}$). It appears that the unattractive $\tilde{\theta}_n$ is superior than the MLE in this example.

Is there any way to discredit $\tilde{\theta}_n$? Statisticians find that if $\theta^* = n^{-1/4}$, namely changes with n, then the variance of $\sqrt{n}\tilde{\theta}_n$ goes to infinity while that of $\sqrt{n}\bar{X}_n$ remains the same. It is a good exercise to compute its variance in this specific case.

If some performance uniformity in θ is required, the MLE is the one with the lowest asymptotic variance. Hence, the MLE is generally referred to as **asymptotically efficient under regularity conditions**, or simply **asymptotically optimal**.

Estimators such as $\tilde{\theta}_n$ are called super-efficient estimators. Their existence makes us think harder. We do not recommend these estimators.

If one estimator has asymptotic variance σ_1^2 and the other one has asymptotic variance σ_2^2 at the same rate and both asymptotically unbiased, then the relative efficiency of $\hat{\theta}_1$ against $\hat{\theta}_2$ is defined as σ_2^2/σ_1^2. A higher ratio implies higher relative efficiency. This definition is no longer emphasized in contemporary textbooks.

Suppose $\tilde{\theta}$ is not asymptotically efficient. However, it is good enough such that for any $\epsilon > 0$, we have

$$\text{PR}\{n^{1/4}|\tilde{\theta} - \theta| \geq \epsilon\} \to 0$$

as $n \to \infty$. Let

$$\hat{\theta}_n = \tilde{\theta}_n - \ell'_n(\tilde{\theta}_n)/\ell''_n(\tilde{\theta}_n)$$

in apparent notation. Under regularity conditions, it can be shown that

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \mathbb{I}^{-1}(\theta^*)).$$

Namely, the Newton-Raphson update formula can turn an ordinary estimator into an asymptotically efficient estimator easily.

Suppose we have a set of i.i.d. observations from Cauchy distribution with location parameter θ. Under this setting, the score function has multiple solutions. It is not straightforward to obtain the MLE in applications. One way to avoid this problem is to estimate θ by the sample median which is not optimal. The above updating formula can then be used to get an asymptotically efficient (optimal) estimator. Let us leave it as an exercise problem.

Chapter 8

Analysis of regression models

In this chapter, we investigate the estimation problems when data are provided in the form

$$(y_i; \mathbf{x}_i): \quad i = 1, 2, \ldots, n. \tag{8.1}$$

The range of y is $\mathcal{R}$ and the range of $\mathbf{x}$ is $\mathcal{R}^p$. We call then response variable and explanatory variables. In many applications, such data are collected because the users believe a large proportion of the variability in y from independent trials can be explained away from the variation in $\mathbf{x}$. Often, we feel that they are linked via a regression relationship with additive error:

$$y_i = g(\mathbf{x}_i; \boldsymbol{\theta}) + \sigma\epsilon_i \tag{8.2}$$

such that the error terms ϵ_i are uncorrelated with mean 0 and variance 1. In this setting, the analytical form of $g(x; \boldsymbol{\theta})$ is specified in general. Yet we are left to decide what is the most "appropriate" value of $\boldsymbol{\theta}$ for the specific occasion. The distributional information about ϵ may or may not be specified depending on specific circumstances. Factoring out σ in the error term may not always be most convenient.

The observations on the explanatory variable, $\mathbf{x}_i$, are either regarded as chosen by scientists (users) so that their values are not random, or they are independent samples from some population whose distribution is not related to $g(\cdot)$ nor $\boldsymbol{\theta}$. In addition, they are independent of ϵ.

The appropriateness of a regression model in specific applications will not be discussed in this course. We continue our discussion under the assumption that all promises for (8.2) are solid.

It is generally convenient to use matrix notation here. We define and denote the covariate matrix as

$$\begin{aligned}\mathbf{X}_n &= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\tau \\ \mathbf{x}_2^\tau \\ \cdots \\ \mathbf{x}_n^\tau \end{pmatrix} \\ &= (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p).\end{aligned}$$

We define design matrix as

$$\mathbf{Z}_n = (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$$

which is the covariate matrix supplemented by a column vector made of 1.

We also use bold faced $\mathbf{y}$ and $\boldsymbol{\epsilon}$ for column vectors of length n for response values and error terms. When necessary, we use $\mathbf{y}_n, \mathbf{X}_n$ with subindex n to highlight the sample size n. Be cautious that $\mathbf{X}_3$ stands for the column vector of the third explanatory variable, not the covariate matrix when $n = 3$. We trust that such abuses will not cause much confusion though mathematically ir-rigorous.

8.1 Least absolution deviation and least squares estimators

Suppose we are given a data set in the form of (8.1) and we are asked to use the data to fit model (8.2). Let us look into the problem of how to best estimate $\boldsymbol{\theta}$ and σ. We do not discuss the issues such that the fitness of function $g(\cdot)$ and the distribution of ϵ.

There are many potential approaches for estimating θ. One way is to select θ value such that the average difference between y_i and $g(x_i; \theta)$ is minimized. If so, one may come up with many potential distances. The absolute difference is one. If so, we would let

$$M_n(\theta) = \sum_{i=1}^{n} |y_i - g(\mathbf{x}_i; \theta)|$$

and find the corresponding M-estimator. This estimator is generally called the least absolute deviation estimator. A disadvantage of this approach is the inconvenience of working with absolute value function both analytically and numerically.

A more convenient choice is

$$M_n(\theta) = \sum_{i=1}^{n} \{y_i - g(\mathbf{x}_i; \theta)\}^2.$$

The resultant estimator is called the least squares estimator.

We may place a parametric distribution assumption on that of ϵ. If ϵ has standard normal N(0, 1) distribution, then the MLE of θ equals the least squares estimator. If ϵ has double exponential distribution with density function

$$f(u) = \frac{1}{2} \exp\{-|u|\}$$

then, the least absolute deviation estimator is also the MLE. Note the variance of this distribution equals 2, which is against model assumption in (8.2) but does not lead to any other issues.

Here is a likely mission-impossible task for many students at this moment. Find the asymptotic efficiency of the least absolute deviation estimator when the data are i.i.d. samples from normal distribution, and the asymptotic efficiency of the least squares estimator when the data are i.i.d. samples from double exponential.

8.2 Linear regression model

Linear regression model is a special signal plus error model. In this case, the **regression function** $\mathbb{E}(Y|X = x)$ has a specific form:

$$\mathbb{E}(Y|X = x) = g(\mathbf{x}; \theta) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

We can write it in vector form with $\mathbf{z}^\tau = (1, \mathbf{x}^\tau)$ as

$$g(\mathbf{x}; \theta) = \mathbf{z}^\tau \boldsymbol{\beta} \tag{8.3}$$

which is linear in **regression coefficient** $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\tau$. While we generally prefer to include β_0 in most applications, this is not a mathematical necessity. In some applications, the scientific principle may seriously demand a model with $\beta_0 = 0$. Luckily, even though the subsequent developments will be based on $\mathbf{z}$ which implies β_0 is part of the model, all of them remain valid when $\mathbf{z}$ is reduced to $\mathbf{x}$ so that $\beta_0 = 0$ is enforced. We will not rewrite the same result twice for this reason.

We have boldfaced two terminologies without formally defining them. It is worth to emphasize here that model is linear not because the regression function $g(\mathbf{x}; \theta)$ is linear in $\mathbf{x}$, but it is linear in θ which is denoted as $\boldsymbol{\beta}$ here. In applications, we may use x_1 for some explanatory variables such as dosage and include $x_2 = \log(x_1)$ as another explanatory variable in the linear model. If so, a linear regression model has a regression function $g(\mathbf{x}, \theta)$ not linear in x_1.

Suppose we have n independent observations from regression model (8.2) with linear regression function (8.3), one way to estimate the regression coefficient vector is by the least squares. The M-function now has form

$$M_n(\boldsymbol{\beta}) = (\mathbf{y}_n - \mathbf{Z}_n\boldsymbol{\beta})^\tau(\mathbf{y}_n - \mathbf{Z}_n\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{z}_i^\tau \boldsymbol{\beta})^2. \tag{8.4}$$

For linear regression model, there is an explicit solution to the least squares problem in a neat matrix notation.

Theorem 8.1. *Suppose* $(y_i, \mathbf{x}_i)$ *are observations from linear regression model* (8.2) *with* $g(\mathbf{x}, \theta)$ *given by* (8.3). *The solution to the least squares problem as defined in* (8.4) *is given by*

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{Z}_n^\tau \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\tau \mathbf{y}_n \tag{8.5}$$

if $\mathbf{Z}_n^\tau \mathbf{Z}_n$ *has full rank.*

If $\mathbf{Z}_n^\tau \mathbf{Z}_n$ *does not have full rank, one solution to the least squares problem is given by*

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{Z}_n^\tau \mathbf{Z}_n)^{-}\mathbf{Z}_n^\tau \mathbf{y}_n$$

where $\mathbf{A}^-$ *here denotes a specific generalize inversion.*

Remark: the statement hints that if $\mathbf{Z}_n^\tau \mathbf{Z}_n$ does not have full rank, the solution is not unique. However, we will not discuss it in details.

Proof. We only give a proof when $\mathbf{Z}_n^\tau \mathbf{Z}_n$ has full rank. It is seen that

$$\begin{aligned} M_n(\boldsymbol{\beta}) &= \{(\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\beta}}) + \mathbf{Z}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^\tau \{(\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\beta}}) + \mathbf{Z}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\} \\ &= (\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\beta}})^\tau(\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\tau(\mathbf{Z}_n^\tau\mathbf{Z}_n)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\geq (\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\beta}})^\tau(\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\beta}}). \end{aligned}$$

The lower bound implied by the above inequality is attained when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Hence, $\hat{\boldsymbol{\beta}}$ is the solution to the least squares problem. □

Let $\hat{\boldsymbol{\beta}}_n$ be the least squares estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ be the true value of the parameter without giving it a special notation. We find

$$\mathbb{E}\{\hat{\boldsymbol{\beta}}_n | X_n\} = (\mathbf{Z}_n^\tau\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\tau\{\mathbf{Z}_n\boldsymbol{\beta}\} = \boldsymbol{\beta}.$$

Hence, $\hat{\boldsymbol{\beta}}_n$ is an unbiased estimator of the regression coefficient vector. Notice that this conclusion is obtained under the assumption that $\mathbf{x}$ and ϵ are independent. Also notice that we assumed ϵ has zero mean and variance 1, but placed no assumption on its distributions. Next, it is seen that

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \sigma(\mathbf{Z}_n^\tau\mathbf{Z}_n)^{-1}\mathbf{Z}_n\boldsymbol{\epsilon}_n.$$

Hence,

$$\text{VAR}(\hat{\boldsymbol{\beta}}_n) = (\mathbf{Z}_n^\tau\mathbf{Z}_n)^{-1}\sigma^2.$$

Because we made a distinction between the covariate matrix $\mathbf{X}_n$ and the design matrix $\mathbf{Z}_n$, the above expression may appear a bit strange.

With $\boldsymbol{\beta}$ estimated by $\hat{\boldsymbol{\beta}}$, it is naturally to regard

$$\hat{\mathbf{y}}_n = \mathbf{Z}_n\hat{\boldsymbol{\beta}}_n = \mathbf{H}_n\mathbf{y}_n$$

as the estimated value of $\mathbf{y}_n$, where the hat matrix

$$\mathbf{H}_n = \mathbf{Z}_n(\mathbf{Z}_n^\tau\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\tau.$$

In fact, we call $\hat{\mathbf{y}}_n$ fitted value(s). How closely does $\hat{\mathbf{y}}_n$ match $\mathbf{y}_n$? The residual of the fit is given by

$$\hat{\epsilon}_n = (\mathbb{I}_n - \mathbf{H}_n)\mathbf{y}_n = \sigma(\mathbb{I}_n - \mathbf{H}_n)\boldsymbol{\epsilon}_n.$$

One can easily verify that $\mathbf{H}_n$ and $\mathbb{I}_n - \mathbf{H}_n$ are symmetric and idempotent, and $(\mathbb{I}_n - \mathbf{H}_n)\mathbf{Z}_n = 0$. From geometric angle, $\mathbf{H}_n$ is a projection matrix. The operation $\mathbf{H}_n\mathbf{y}_n$ projects $\mathbf{y}_n$ into the linear space spun by $\mathbf{Z}_n$. Naturally, $(\mathbb{I}_n - \mathbf{H}_n)\mathbf{y}_n$ is the projection of $\mathbf{y}_n$ into the linear space orthogonal to $\mathbf{Z}_n$. This leads to a decomposition of the sum of squares:

$$\mathbf{y}_n^\tau \mathbf{y}_n = \mathbf{y}_n^\tau \mathbf{H}_n \mathbf{y}_n + \mathbf{y}_n^\tau (\mathbb{I}_n - \mathbf{H}_n)\mathbf{y}_n.$$

The second term is the "residual sum of squares". It is an easy exercise to prove that

$$\mathbf{y}_n^\tau (\mathbb{I}_n - \mathbf{H}_n)\mathbf{y}_n = \hat{\boldsymbol{\epsilon}}_n^\tau \hat{\boldsymbol{\epsilon}}_n.$$

We directly verified that $\hat{\boldsymbol{\beta}}$ solves the least squares problem. One may derive this result by searching for solutions to

$$\frac{\partial M_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0.$$

This leads to normal equation

$$\mathbf{Z}_n^\tau \{\mathbf{y}_n - \mathbf{Z}_n \boldsymbol{\beta}\} = 0.$$

We again leave it as an easy exercise.

We have seen that the least squares estimator $\hat{\boldsymbol{\beta}}_n$ has a few neat properties. Yet we cannot help to ask: can we find other superior estimators? The answer is no at least in one respect. The least squares estimator has the lowest variance among all unbiased linear estimators of $\boldsymbol{\beta}$. A linear estimator is defined as one that can be written as a linear combinations of y_i. It must be able to be written in the form of $\mathbf{A}\mathbf{y}_n$ for some matrix $\mathbf{A}$ not dependent on $\mathbf{y}_n$.

Theorem 8.2. Gauss-Markov Theorem. *Let $\hat{\boldsymbol{\beta}}_n$ be the least squares estimator and*

$$\tilde{\boldsymbol{\beta}}_n = \mathbf{A}\mathbf{y}_n$$

for some nonrandom matrix $\mathbf{A}$ (may depend on $\mathbf{X}_n$) be an unbiased linear estimator of $\boldsymbol{\beta}$ under the linear regression model with n independent observations. Then

$$\textsc{var}(\tilde{\boldsymbol{\beta}}) - \textsc{var}(\hat{\boldsymbol{\beta}}) \geq 0.$$

Proof. Suppose $\mathbf{A}\mathbf{y}_n$ is unbiased for $\boldsymbol{\beta}$. We must have

$$\mathbb{E}(\mathbf{A}\mathbf{y}_n) = \mathbf{A}\mathbf{Z}_n\boldsymbol{\beta} = \boldsymbol{\beta}$$

for any $\boldsymbol{\beta}$. Hence, we must have $\mathbf{A}\mathbf{Z} = \mathbb{I}_{p+1}$. This implies

$$\text{VAR}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = \sigma^2\{\mathbf{A} - (\mathbf{Z}_n^\tau\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\tau\}\{\mathbf{A}^\tau - \mathbf{Z}_n(\mathbf{Z}_n^\tau\mathbf{Z}_n)^{-1}\} = \text{VAR}(\tilde{\boldsymbol{\beta}}) - \text{VAR}(\hat{\boldsymbol{\beta}}).$$

Because the variance matrix for any random variable is non-negative definite. Hence, we must have

$$\text{VAR}(\tilde{\boldsymbol{\beta}}) - \text{VAR}(\hat{\boldsymbol{\beta}}) \geq 0.$$

□

An estimator which is linear in data and unbiased for the target parameter is called **best linear unbiased estimator** (BLUE) if it has the lowest possible variance matrix.

Not only the least squares estimator $\hat{\boldsymbol{\beta}}$ is BLUE for $\boldsymbol{\beta}$, but $\mathbf{b}^\tau\hat{\boldsymbol{\beta}}$ is BLUE for $\mathbf{b}^\tau\boldsymbol{\beta}$ for any non-random vector $\mathbf{b}$.

At the same time, be aware that if we have additional information about the distribution of ϵ_n in the linear model, then we may obtain more efficient estimator for $\boldsymbol{\beta}$, but that estimator is either not linear or not unbiased.

8.3 Local kernel polynomial method

Naturally, a linear regression model is not always appropriate in applications, but we may still believe a signal plus noise relationship is sound. In this section, we consider the situation where the regression function $g(x)$ is smooth in x, but we are unwilling to place more restrictions on it. At the same time, we only study the simple situation where x is a univariate covariate.

Suppose we wish to estimate $g(x)$ at some specific x^* value. By definition, $g(x) = \mathbb{E}(Y|X = x^*)$. If among n observations $\{(y_i, x_i)\}$, $i = 1, \ldots, n$ we collected, there are many x_i such that $x_i = x^*$. Then the average of their corresponding y_i would be a good estimate of $g(x^*)$. In reality, there may not be any x_i equalling x^* exactly. Hence, this idea does not work. On the other hand, when n is very large, there might be many x_i which are very close to x^*. Hence, the average of their corresponding y_i should be a sensible estimate of $g(x^*)$. To make use of this idea, one must decide how close is close enough. Even within the small neighbourhood, should we merely use constant, rather than some other smooth functions of x to approximate $g(x)$?

For any u in close enough to x (rather than x^* for notation simplicity) and some positive integer p, when $g(x)$ is sufficiently smooth at x, we have

$$g(u) \approx f(x) + f'(x)(u-x) + \ldots + (1/p!)f^{(p)}(x)(u-x)^p.$$

Let

$$\beta_0 = f(x),\ \beta_1 = f'(x),\ \ \ldots, \beta_p = (1/p!)f^{(p)}(x).$$

Then the approximation can be written as

$$g(u) \approx \beta_0 + \beta_1(u-x) + \ldots + \beta_p(u-x)^p.$$

Note that at $u = x$, we have $g(x) \approx \beta_0$.

Suppose that for some $h > 0$, $f(u)$ perfectly coincides with the above polynomial function for $x \in [x-h, x+h]$. If so, within this region, we have a linear regression model with regression coefficient $\boldsymbol{\beta}_x$. A natural approach of estimating this *local* $\boldsymbol{\beta}_x$ is the least squares:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \mathbb{1}(|x_i - x| \le h)\{y_i - \mathbf{z}_i^{\tau}\boldsymbol{\beta}\}^2$$

where

$$\mathbf{z}_i = \{1, (x_i - x), (x_i - x)^2, \ldots, (x_i - x)^p\}^{\tau}.$$

Note again that $\mathbf{z}_i$ is defined dependent on x-value, the location at which $g(x)$ is being estimate.

Note that we have added a subindex x to $\boldsymbol{\beta}$. This is helpful because this vector is specific to the regression function $g(u)$ at $u = x$. When we change target from $u = x_1$ to $u = x_2 \ne x_2$, we must refit the data and obtain the $\boldsymbol{\beta}$ specific for $u = x_2$. We repeatedly state this to emphasize the local nature of the current approach.

The above formulation implies that ith observation will be excluded even if $|x_i - x|$ is only slightly larger than h. At the same time, any observations with $|x_i - x| \le h$ are treated equally. This does not seem right in our intuition. One way to avoid this problem is to replace the indicator function by a general kernel function $K(x)$ often selected to satisfy the following properties:

1. $K(x) \geq 0$;
2. $\int_{-\infty}^{\infty} K(x) dy = 1$;
3. $K(x) = K(-x)$, That is, $K(x)$ is a symmetric function.

For instance, the density function $\phi(x)$ of N(0, 1) has these properties. In fact, any symmetric density function does.

Let $K_h(x) = h^{-1}K(x/h)$. We now define the local polynomial kernel estimator of $\boldsymbol{\beta}_x$ as

$$\hat{\boldsymbol{\beta}}_x = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} K_h(x_i - x)\{y_i - \mathbf{z}_i^\tau \boldsymbol{\beta}\}^2$$

An explicit solution to the above optimization problem is readily available using matrix notation. Let $\mathbf{y}_m$ be the response vector, define design matrix

$$Z_x = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{pmatrix}$$

and weight matrix

$$W_x = \text{diag}\{K_h(x_1 - x), K_h(x_2 - x), \cdots, K_h(x_n - x)\}.$$

The M-function can then be written as

$$M_n(\beta) = (\mathbf{y} - \mathbf{Z}_x\boldsymbol{\beta})^\tau \mathbf{W}_x(\mathbf{y} - \mathbf{Z}_x\boldsymbol{\beta}).$$

It is an easy exercise to show that the solution is given by

$$\hat{\boldsymbol{\beta}}_x = (\mathbf{Z}_x^\tau W_x \mathbf{Z}_x)^{-1}\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{y}_n$$

Let $\mathbf{e}_j$ be a $(p+1) \times 1$ vector such that the jth element being 1 and all other elements being 0, $j = 1, \ldots, p+1$. Then we estimate $g(x)$ by

$$\hat{g}(x) = \hat{\beta}_0 = \mathbf{e}_1^\tau (\mathbf{Z}_x^\tau W_x \mathbf{Z}_x)^{-1}\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{y}_n$$

where $\hat{\beta}_0$ is the first element of $\hat{\boldsymbol{\beta}}_x$.

Remark: Notationally, the above locally kernel polynomial estimator remains the same for any choice of p.

Suppose $g(x)$ is differentiable up to order p. Then, for $k = 1, \ldots, p$, we estimate the kth derivative $g^{(k)}(x)$ by

$$\hat{g}^{(k)}(x) = k!\hat{\beta}_k = k!\mathbf{e}^{\tau}_{k+1}(\mathbf{Z}^{\tau}_x W_x \mathbf{Z}_x)^{-1}\mathbf{Z}^{\tau}_x \mathbf{W}_x \mathbf{y}_n.$$

When we decide to use $p = 0$ in this approach, the estimator $\hat{g}(x)$ becomes

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x)y_i}{\sum_{i=1}^{n} K_h(x_i - x)},$$

which is known as the local constant kernel estimator, kernel regression estimator and Nadaraya-Watson estimator. This estimator can be motivated by the fact that $g(u)$ is a constant function in a small neighborhood of x: $u \in [x-h, x+h]$ for some sufficiently small h. The estimator is the weighted average of the corresponding response values whose x is within small neighbourhood of x.

When we decide to use $p = 1$ in this approach, the estimator is called the local linear kernel estimator of $g(x)$.

Before this estimator is applied to any specific data, we must make a choice on the kernel function K, the degree of the polynomial p and the bandwidth h. We now go over these issues.

Choice of $K(y)$.

The choice of kernel function $K(x)$ is not crucial. Other than it should have a few desired properties, its specific form does not markedly change the variance or bias of $\hat{g}(x)$. In our future examples, we will mostly use normal density function. Clearly, the normal density function has the listed three properties.

Choice of p.

For the given bandwidth h and kernel $K(x)$, a large value of p would expectedly reduce the bias of the estimator because the local approximation becomes more and more accurate as p increases. At the same time, when p is large, we have more parameters to estimate as reflected in the dimension of $\boldsymbol{\beta}$. Hence, the variance of the estimator will increase and there will be a larger computational cost.

Fan and Gijbels (1996) showed that when the degree of the polynomial employed increases from $p = k + 2q$ to $p = k + 2q + 1$ for estimating $g^{(k)}(x)$, the variance does not increase. However, if we increase the degree from $p = k + 2q + 1$ to $p = k + 2q + 2$, the variance increases. Therefore for estimating $g^{(k)}(x)$, it is beneficial to use a degree p such that $p - k$ is odd. Since bandwidth h also controls the bias and variance trade-off of $g^{(k)}(x)$, they recommended the lowest odd order for $p - k$, namely $p = k + 1$, or occasionally $p = k + 3$. For the regression function itself, they recommended local linear kernel estimator (i.e. $p = 1$) instead of the Nadaraya-Watson estimator (i.e. $p = 0$).

To have a better understanding of the above information, we summarize some theoretical results about the local linear kernel estimator and Nadaraya-Watson estimator here. Let them be denoted as $\hat{g}_{\text{LL}}(x)$ and $\hat{g}_{\text{NW}}(x)$, respectively. We have

$$\begin{aligned}
\hat{g}_{\text{NW}}(x) &= \frac{\sum_{i=1}^n K_h(x_i - x) y_i}{\sum_{i=1}^n K_h(x_i - x)} \\
\hat{g}_{\text{LL}}(x) &= \hat{\beta}_0 = \arg\min_{\beta_0}\{\min_{\beta_1} \sum_{i=1}^n K_h(x_i - x)\{y_i - \beta_0 - \beta_1(x_i - x)\}^2\}.
\end{aligned}$$

Under the regression model assumption that

$$y_i = g(x_i) + \sigma\epsilon_i$$

and for random x_i such that its density function is given by $f(x)$, and under many conditions regulating $f(x)$, $g(x)$ and distribution of ϵ, we have

$$\begin{aligned}
\mathbb{E}\{\hat{g}_{\text{NW}}(x)|\mathbf{x}\} &\approx g(x) + 0.5h^2\mu_2(K)\left\{g''(x) + \frac{2f'(x)g'(x)}{f(x)}\right\}; \\
\mathbb{E}\{\hat{g}_{\text{LL}}|\mathbf{x}\} &\approx g(x) + 0.5h^2 g''(x)\mu_2(K); \\
\text{VAR}\{\hat{g}_{\text{NW}}(x)|\mathbf{x}\} &\approx \frac{\sigma^2}{nhf(x)}R(K); \\
\text{VAR}\{\hat{g}_{\text{LL}}(x)|\mathbf{x}\} &\approx \frac{\sigma^2}{nhf(x)}R(K)
\end{aligned}$$

where $\mu_2(K)$ and $R(K)$ are some positive constants depending on kernel function K.

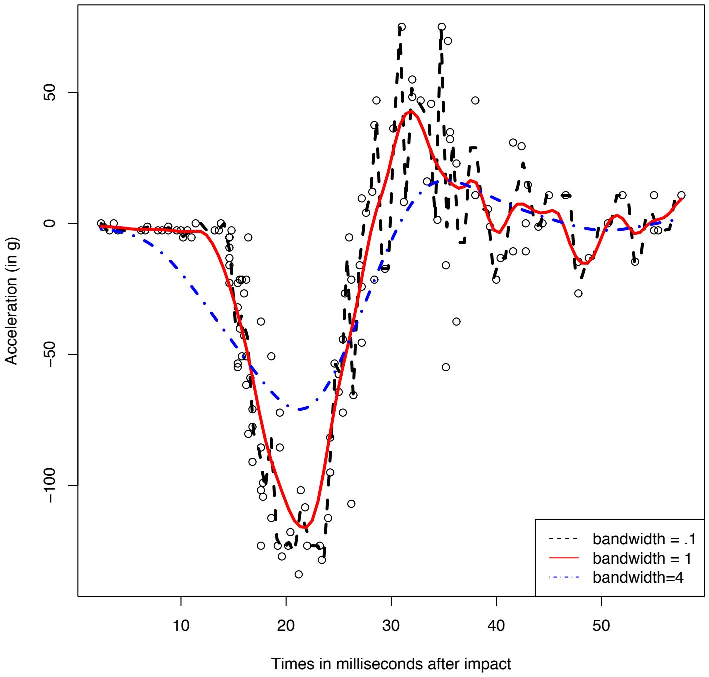
The above results show that the local linear kernel estimator $\hat{g}_{\text{LL}}(x)$ and Nadaraya-Watson estimator $\hat{g}_{\text{NW}}(x)$ have the same asymptotic variance conditional on $\mathbf{x}$. which is the conclusion that we discussed before. The asymptotic bias of $\hat{g}_{\text{NW}}(x)$ has an extra bias term $2f'(x)g'(x)\mu_2(K)h^2/f(x)$. The coefficient $2f'(x)g'(x)/g(x)$ is also called design bias because it depends on the design, namely, the distribution of x. This implies that the bias is sensitive to the positions of design point x_i's. Note that $\frac{f'(x)}{f(x)}$ can have high influence on the bias when x is close to the boundary. For example, when the density points x_i have standard normal distribution, $|f'(x)/f(x)| = |x|$, which is very large when x approaches to ∞. Hence $2f'(x)g'(x)/f(x)$ is also known as boundary bias. These two biases are reduced by using the local linear kernel estimator. In summary, local linear kernel estimator is free from the design and boundary biases, but Nadaraya-Watson estimator is not.

Choice of bandwidth h

Suppose we have made choice of the kernel function $K(x)$ and p. We now discuss the choice of bandwidth h. Bandwidth plays a very important role in estimating the regression function $g(x)$.

First, as h increases, the local approximation becomes worse and worse and hence the bias of local polynomial kernel estimator increases. On the other hand, more and more observations will be included in estimating $g(x)$. Hence the variance of local polynomial kernel estimator decreases. A good choice of a bandwidth helps to balance the bias and variance. Second, as h increases, the local polynomial kernel estimate becomes smoother and smoother. This can be observed in Figure 8.1, in which we compare the Nadaraya-Watson estimates of $g(x)$ constructed when the bandwidth h takes three values, 0.1, 1, and 4, respectively. Conceptually, the number of parameters required to describe the curve decreases. In this sense, h controls the model complexity. We should choose a bandwidth to balance the modelling fitting and model complexity.

Figure 8.1: Motorcycle data: Nadaraya-Watson estimates of $g(x)$ with normal kernel



We introduce two bandwidth selection methods here: l eave-one-out cross-validation (CV) and generalized cross-validation (GCV). These two methods are also widely used in studying other regression problems.

The idea of leave-one-out CV is as follows. Recall that one purpose of fitting a regression model is to predict the response value in a new trial. So a reasonable choice of h should result in a small prediction error. Unfortunately, we do not know the true response, and therefore we cannot know how good is the prediction $\hat{f}(x)$ given h. The idea of cross-validation is to first delete one observation from the data set, and treat the remaining $n-1$ observations as the training data set and the deleted observations as testing data. We then test the goodness of prediction for the testing observation by using the training data set. We repeat the process for all observations and get the prediction errors for all observations. We choose h by minimizing the sum of prediction errors. Mathematically, let $\hat{g}_{-i}(x_i)$ be the estimate of $g(x_i)$ based on the $n-1$ observations without x_i. For the given h, the CV score is defined as

$$\textsc{cv}(h) = \sum_{i=1}^{n} \{y_i - \hat{g}_{-i}(x_i)\}^2.$$

The optimal h based on the leave-one-out cross-validation idea is

$$h_{cv} = \arg\min \textsc{cv}(h).$$

It seems that it might be time consuming to evaluate $\textsc{cv}(h)$ since we apparently need to recompute the estimate after dropping out each observation. Fortunately, there is a shortcut formula for computing $\textsc{cv}(h)$.

Let

$$l(x) = \Big(l_1(x), \ldots, l_n(x)\Big) = \mathbf{e}_1^\tau (\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{Z}_x)^{-1} \mathbf{Z}_x^\tau \mathbf{W}_x.$$

Then

$$\hat{g}(x) = \sum_{j=1}^{n} l_j(x) y_j \text{ and } \hat{g}(x_i) = \sum_{j=1}^{n} l_j(x_i) y_j.$$

Define the fitted value vector

$$\widehat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_n)^\tau = (\hat{g}(x_1), \cdots, \hat{g}(x_n))^\tau.$$

It then follows that

$$\widehat{\mathbf{y}} = \mathbf{L}\mathbf{y}$$

where $\mathbf{L}$ is an $n \times n$ matrix whose ith row is $l(x_i)$; thus $\mathbf{L}_{ij} = l_j(x_i)$ and $\mathbf{L}_{ii} = l_i(x_i)$. It can be shown that

$$\textsc{cv}(h) = \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}(x_i)}{1 - \mathbf{L}_{ii}} \right\}^2 .$$

We can minimize the above $\textsc{cv}(h)$ to get the h_{cv}.

The second method for choosing h is called the generalized cross-validation. For this method, rather than minimizing $\textsc{cv}(h)$, an alternative is to use an approximation called generalized cross-validation (GCV) score in which each $\mathbf{L}_{ii}$ is replaced with its average v/n, where $v = \mathrm{tr}(\mathbf{L}) = \sum_{i=1}^{n} \mathbf{L}_{ii}$ is called the effective degrees of freedom. Thus, we would minimize GCVscore

$$\textsc{gcv}(h) = \sum_{i=1}^{n} \left\{ \frac{Y_i - \hat{f}(x_i)}{1 - v/n} \right\}^2$$

to obtain the bandwidth h_{gcv}. That is,

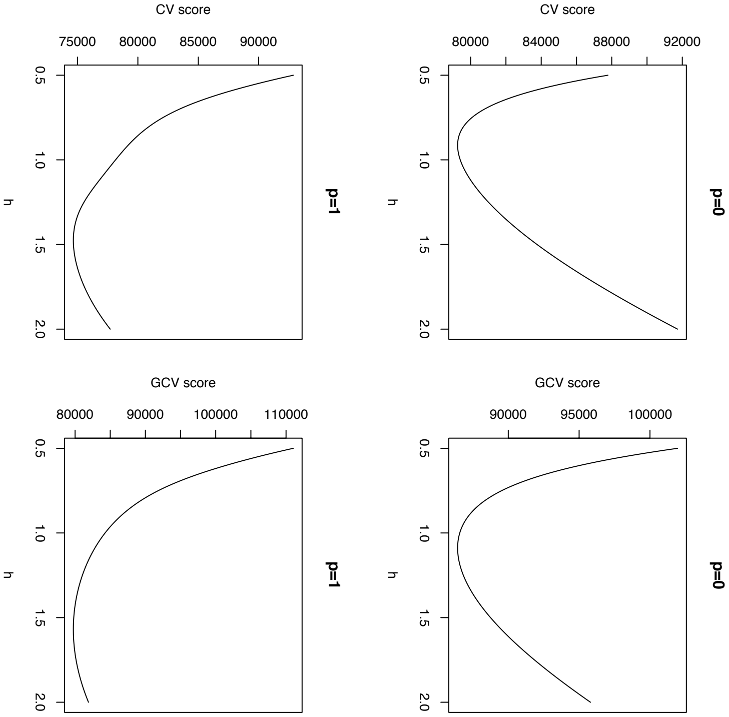
$$h_{gcv} = \arg\min_{h} \textsc{gcv}(h).$$

Usually h_{cv} is quite close to h_{gcv}.

In Appendix I, we include the R function *bw.cv()* to choose the bandwidth for the local polynomial kernel estimate for continuous response. The source code is saved in *bw_cv.R*. In this function, if the option *cv=T*, then the CVmethod is used; if the option *cv=F*, then the GCVmethod is used. The R function *regCVBwSelC()* in the R package *locpol* can also be used to obtain h_{cv} for the continuous response. The R function *regCVBwSelC()* gives the same result as the R function *bw.cv()* with *cv=T*. Further it is much faster. Figure 8.2 gives the $\textsc{cv}(h)$ and $\textsc{gcv}(h)$ for $p = 0, 1$. Here the normal kernel is used. (Remark by your instructor: these programs are not included).

Similar to kernel density estimation, Wand and Jones (1995) applied the idea of direct plug-in methods for bandwidth selection for local linear kernel estimate. This idea is implemented in R function *dpill()* in the package *KernSmooth*. I did not cover this idea because it is only applicable for local linear kernel estimate. Further it is more complicated to implement compared with CV and GCV methods.

Figure 8.2: Motorcycle data: CV(h) and GCV(h) for $p = 0, 1$ with normal kernel



Applying the above mentioned R functions, for $p = 0$, $h_{cv} = 0.914$ and $h_{gcv} = 1.089$; for $p = 1$, $h_{cv} = 1.476$, $h_{gcv} = 1.570$, and the direct plug-in gives $h_{DPI} = 1.445$. Figure 8.3 gives the fitted curves of $f(x)$ with $p = 0, 1$, in which the bandwidth is selected by CV or GCV. Here the normal kernel is used. The two curves for $p = 0$ are almost the same. The fitted curves for $p = 1$ with the bandwidths h_{cv}, h_{gcv}, and h_{DPI} are almost the same. Hence we only plot the curves with the bandwidths selected by CV and GCV. The four fitted curves are very close to each. They do not show too much difference when they are plotted in the same panel.

Properties of $\hat{f}(x)$

Let h be given. We have

$$\mathbb{E}\{\hat{g}(x)|\mathbf{x}\} \approx f(x)$$

and

$$\text{VAR}\{\hat{g}(x)|\mathbf{x}\} = \sigma^2 \mathbf{e}_1{}^\tau (\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{Z}_x)^{-1} (\mathbf{Z}_x^\tau \mathbf{W}_x^2 \mathbf{Z}_x)(\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{Z}_x)^{-1} \mathbf{e}_1.$$

Therefore the standard error is given by

$$\text{se}\{\hat{f}(x)\} = \sqrt{\hat{\sigma}^2 \mathbf{e}_1{}^\tau (\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{Z}_x)^{-1} (\mathbf{Z}_x^\tau \mathbf{W}_x^2 \mathbf{Z}_x)(\mathbf{Z}_x^\tau \mathbf{W}_x \mathbf{Z}_x)^{-1} \mathbf{e}_1},$$

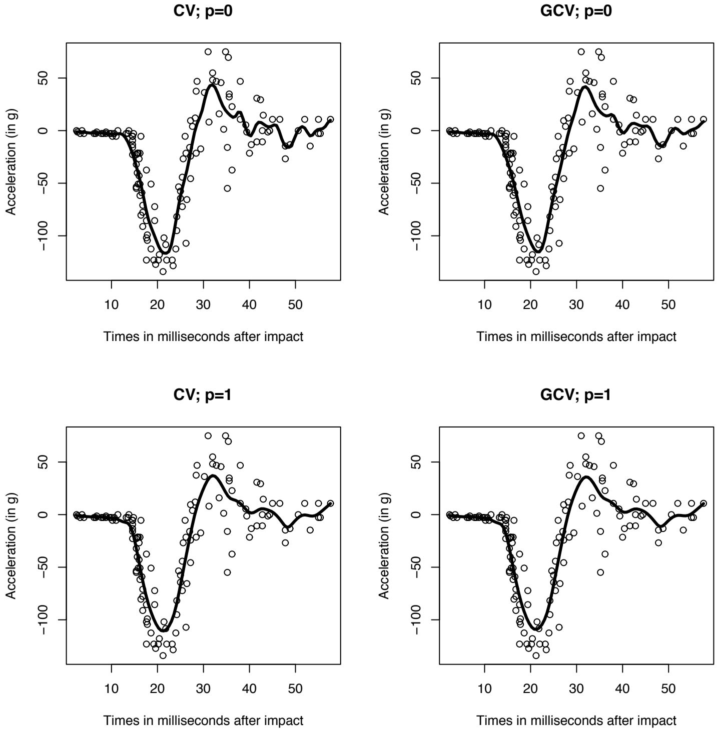
where $\hat{\sigma}^2$ is an estimator of σ^2. Wand and Jones (1995) suggested the following form for $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = n - 2v + \tilde{v}$$

with

$$v = \text{tr}(\mathbf{L}) = \sum_{i=1}^n \mathbf{L}_{ii}, \quad \tilde{v} = \text{tr}(\mathbf{L}^\tau \mathbf{L}) = \sum_{i=1}^n \sum_{j=1}^n \mathbf{L}_{ij}^2.$$

Figure 8.3: Motorcycle data: fitted curves for $p = 0, 1$ with normal kernel, in which the bandwidth is selected by CVor GCV



8.4 Spline method

Let us again go back to model (8.2) but do not assume a parametric regression function $g(\mathbf{x}; \boldsymbol{\theta})$. Instead, we only postulate that $\mathbb{E}(Y|X = \mathbf{x}) = g(\mathbf{x})$ for some smooth function $g(\cdot)$. Suppose we try to estimate $g(\cdot)$ by simplistic least squares estimator without a careful deliberation. The solution will be regarded as the solution to the minimization problem to

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2.$$

If all x_i values are different, the solution is given by any function $\hat{g}$ such that $\hat{g}(x_i) = y_i$. Such a perfect fitting clearly does not have any prediction power for a new observation whose covariate value is not equal to the existing covariate values. Furthermore, if $\hat{g}(x)$ just connects all points formed by observations, it lacks some smoothness we may expect.

If we require $g(x)$ to be a linear function of x, then it is a very smooth function, but the fitting is unsatisfactory if $\mathbb{E}(Y|X = \mathbf{x})$ is not far from linear in $\mathbf{x}$. One way to balance the need of smoothness and fitness is to use smoothing spline. Among all functions with first two continuous derivatives, let us find the one that minimizes the penalized L_2-loss function

$$\hat{g}_\lambda(\mathbf{x}) = \arg\min_{g(\mathbf{X})} \left[\sum_{i=1}^{n} \{y_i - g(\mathbf{x}_i)\}^2 + \lambda \int \{g''(x)\}^2 d\mathbf{x} \right], \qquad (8.6)$$

for some positive tuning or smoothing parameter λ. which is called smoothing parameter. In the penalized L_2-loss function, the first term measures the goodness of model fitting, while the second term penalizes the curvature in the function. We will remain vague on the range of x.

> When we use $\lambda = 0$: $\hat{g}_\lambda(\mathbf{x})$ becomes the ordinary least squares estimator. The solution is not unique and has little prediction power.
>
> When we use $\lambda = \infty$, then the optimal solution must be $g''(x) = 0$ for all x. The solution must be linear in $\mathbf{x}$. We are back to use linear regression model and the associated least squares estimator.

Clear, a good fit is possible by choose a λ value in between 0 to ∞ to get a smooth function with reasonable fitting. Note that the above minimization is taken over all possible function $g(\mathbf{x})$, and such functions form an infinite dimensional space. Remarkably, it can be shown that solution $\hat{g}_\lambda(x)$ to the penalized least squares problem is a *natural cubic spline with knots at the unique values of* $\{x_i\}_{i=1}^n$. Here we consider the case when x is one-dimensional.

8.5 Cubic spline

We now need a brief introduction to the cubic spline. A cubic spline is a function which is piece-wisely cubic polynomial. Namely, we partition the real line into finite number of intervals and a cubic spline is a polynomial of x of degree 3 which has continuous derivative.

More precisely, suppose $-\infty = t_0 < t_1 < t_2 < \ldots < t_k < t_{k+1} = \infty$ are k distinct real values, then $s(x)$ is a cubic spline if

1. It is a cubic function on each interval $[t_i, t_{i+1}]$:

$$\begin{aligned} s_i(x) &= \{a_i + b_i x + c_i x^2 + d_i x^3\} \\ s(x) &= \sum_{i=0}^{k} s_i(x) \mathbb{1}(t_i < x \le t_{i+1}). \end{aligned}$$

2. $s(x)$ and its first and second derivatives are continuous:

$$\begin{aligned} s_i(t_{i+1}) &= s_{i+1}(t_{i+1}), \\ s_i'(t_{i+1}) &= s_{i+1}'(t_{i+1}), \\ s_i''(t_{i+1}) &= s_{i+1}''(t_{i+1}). \end{aligned}$$

The connection values $t_1, \ldots, t_k$ are called the knots of the cubic spline. In particular, t_1 and t_k are called the boundary knots, and $t_2, \ldots, t_{k-1}$ are called the interior knots.

Furthermore, if

3. $s(x)$ is linear outside the interval $[t_1, t_k]$; that is,

   $$s(x)\mathbb{1}(x \leq t_1) = (a_0+b_0x)\mathbb{1}(x \leq t_1); \quad s(x)\mathbb{1}(x \geq t_k) = (a_k+b_kx)\mathbb{1}(x \geq t_k)$$

   for some a_0, b_0, a_k, b_k,

we call $s(x)$ a **natural** cubic spline with knots at $t_1, \dots, t_k$. Note that this also means $c_0 = c_k = 0$.

The following result shows that there is a simpler way to express a cubic spline.

Theorem 8.3. *Any cubic spline $s(x)$ with knots at $\{t_1, \dots, t_k\}$ can be written as:*

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \beta_{j+3}(x - t_j)_+^3, \tag{8.7}$$

where $(x)_+ = \max(0, x)$ for some coefficients $\beta_0, \dots, \beta_{k+3}$.

In other words, the cubic spline is a member of the linear space with basis functions

$$1, x, x^2, x^3, (x - t_1)_+^3, \dots, (x - t_k)_+^3.$$

Proof. The function defined by (8.7) is clearly a cubic function on every interval $[t_0, t_{i+1}]$. We can also easily verify that its first two derivatives are continuous. This shows that such functions are cubic splines.

To prove this theorem, we need further show that every cubic spline with knots at $\{t_1, \dots, t_k\}$ can be written in the form specified by (8.7).

Let $g(x)$ be a cubic spline with knots at $\{t_1, \dots, t_k\}$. Denote $\gamma_i = g''(t_i)$ for $i = 1, 2, \dots, k$. We show that there exists a function $s(x)$ in the form of (8.7) such that

$$\beta_3 = 0, \beta_{k+3} = 0,$$

and $s''(t_i) = \gamma_i$ for $i = 1, \dots, k$.

If such a function exists, we must have, for other β values

$$\begin{aligned}
&\beta_2/3 = \gamma_1/6;\\
&\beta_2/3 + \beta_4(t_2 - t_1) = \gamma_2/6;\\
&\beta_2/3 + \beta_4(t_3 - t_1) + \beta_5(t_3 - t_2) = \gamma_3/6;\\
&\cdots\\
&\beta_2/3 + \beta_4(t_{k-1} - t_1) + \cdots + \beta_{k+1}(t_{k-1} - t_{k-2}) = \gamma_{k-1}/6;\\
&\beta_2/3 + \beta_4(t_k - t_1) + \cdots + \beta_{k+1}(t_k - t_{k-2}) + \beta_{k+2}(t_k - t_{k-1}) = \gamma_k/6;
\end{aligned}$$

Taking differences, we find another set of equations whose solutions clearly exist:

$$\begin{aligned}
\beta_4 &= (1/6)(\gamma_2 - \gamma_1)/(t_2 - t_1);\\
\beta_4 + \beta_5 &= (1/6)(\gamma_3 - \gamma_2)/(t_3 - t_2);\\
\beta_4 + \beta_5 + \beta_6 &= (1/6)(\gamma_4 - \gamma_3)/(t_4 - t_3);\\
&\cdots\\
\beta_4 + \beta_5 + \cdots + \beta_{k+2} &= (1/6)(\gamma_k - \gamma_{k-1})/(t_k - t_{k-1}).
\end{aligned}$$

The solution $s(x)$ with any choice of β_0 and β_1 we have just obtained, has the same second derivatives with the cubic spline $g(x)$ at $\{t_1 = 0, t_2, \ldots, t_k\}$. Now we can select β_0 and β_1 values such that $s(t_1) = g(t_1)$ and $s'(t_1) = g'(t_1)$. Together with $s''(t_1) = g''(t_1)$, $s''(t_2) = g''(t_2)$, and they are both cubic functions, we must have $s(x) = g(x)$ for all $x \in [t_1, t_2]$. Applying the same argument, they must be identical over $[t_1, t_k]$. This proves the existence. □

As a remark, there can be multiple cubic splines identical on $[t_1, t_k]$ but different outside this interval.

Suppose

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \beta_{j+3}(x - t_j)_+^3$$

is a **natural** cubic spline with knots $\{t_1, t_2, \ldots, t_k\}$. Since it is linear below t_1, we must have

$$\beta_2 = \beta_3 = 0.$$

At the same time, being linear beyond t_k implies we must have

$$\sum_{j=1}^{k} \beta_{j+1}(x - t_j)_+ = 0$$

for all $x \geq t_k$. This is possible only if both

$$\sum_{j=1}^{k} \beta_{j+3} = 0, \quad \sum_{j=1}^{k} t_j \beta_{j+3} = 0.$$

In conclusion, out of $k+4$ entries of $\boldsymbol{\beta}$, only k of them are free for a natural cubic spline. For this reason, we need to think a bit about how to fit a natural cubic spline when data and knots are given.

One approach is as follows. Define functions for $j = 1, \ldots, k$

$$d_j(x) = \frac{(x - t_j)_+^3 - (x - t_k)_+^3}{t_k - t_j}.$$

Further, let $N_1(x) = 1,\ N_2(x) = x$, and for $j = 3, \ldots, k$, let

$$N_j(x) = d_{j-1}(x) - d_1(x).$$

The following theorem says that every natural cubic spline is a linear combination of $N_j(x)$.

Theorem 8.4. *Let $t_1 < t_2 < \ldots < t_k$ be k knots and $\{N_1(x), \ldots, N_k(x)\}$ be functions defined above. Then all natural cubic splines $s(x)$ with knots in $\{t_1, \ldots, t_k\}$ can be expressed as:*

$$s(x) = \sum_{j=1}^{k} \beta_j N_j(x),$$

for some coefficients $\beta_1, \ldots, \beta_k$.

Proof. Note that

$$(t_k - t_j) d_j(x) = (x - t_j)_+^3 - (x - t_k)_+^3.$$

Equivalently,

$$(x - t_j)_+^3 = (t_k - t_j)d_j(x) + (x - t_k)_+^3.$$

Substituting this expression into generic form of cubic spline, and activating the constrains on β_j implied by *natural* cubic spline, we find

$$s(x) = \beta_0 N_1(x) + \beta_1 N_2(x) + \sum_{j=1}^{k} \beta_{j+3}(t_k - t_j)N_{j+1}(x).$$

Note that the kth term is zero. The conclusion is therefore true. □

In general, a natural cubic spline can give very good approximation to any function in a finite interval. This makes it useful to fit nonparametric signal plus noise regression models. Given data $\{y_i; x_i\}$ and the k knots, $t_1, \ldots, t_k$, we may suggest that

$$g(x) \approx \sum_{j=1}^{k} \beta_j N_j(x).$$

For the ith observation, we have

$$g(x_i) \approx \sum_{j=1}^{k} \beta_l N_j(x_i),$$

which is now a linear combination of k derived covariates. Let $\mathbf{y}$ be the response vector, $\boldsymbol{\beta}$ the regression coefficient vector and $\boldsymbol{\epsilon}$ the error vector. Define design matrix

$$\mathbf{Z}_n = \begin{pmatrix} N_1(x_1) & \cdots & N_k(x_1) \\ \vdots & \vdots & \vdots \\ N_1(x_n) & \cdots & N_k(x_n) \end{pmatrix}.$$

The approximate regression model becomes

$$\mathbf{y} \approx \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{8.8}$$

We may use least squares estimator of $\boldsymbol{\beta}$ given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\tau \mathbf{Z})^{-1}\mathbf{Z}^\tau \mathbf{y}.$$

Let $\mathbf{N}(x) = \{N_1(x), \dots, N_k(x)\}^\tau$. Once $\hat{\boldsymbol{\beta}}$ is obtained, we estimate the regression function by

$$\hat{g}(x) = \mathbf{N}^\tau(x)\hat{\boldsymbol{\beta}}.$$

Suppose (8.8) is in fact exact, then the properties of least squares estimator are applicable. We summarize them as follows:

(a) $\mathbb{E}\{\hat{\boldsymbol{\beta}}\} = \boldsymbol{\beta}$ and $\mathbb{E}\{\hat{g}(x)\} = g(x)$;

(b) $\text{VAR}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$

(c) $\text{VAR}\{\hat{g}(x)\} = \sigma^2\mathbf{N}^\tau(x)(\mathbf{Z}^\tau\mathbf{Z})^{-1}\mathbf{N}(x)$.

If (8.8) is merely approximate, then the above equalities are approximate. The approximation errors will not be discussed here.

The above idea is known as *regression spline*, which is a large research topic in nonparametric regression. This approach is very widely used in many applications to model a nonlinear and unknown function $g(x)$. To apply this method, we must decide the number of knots and choose the knots $t_1, \dots, t_k$ after the number of knots (k) is decided.

8.6 Smoothing spline

Smoothing spline addresses the knot-selection problem of regression spline by taking all different covariate values as the knots. It uses the size of penalty to determine the level of smoothness.

Recall that we claim that the numeric solution of smoothing spline to (8.6) is a natural cubic spline with knots at all distinct values ($t_1 < \cdots < t_k$) of $\{x_i\}_{i=1}^n$. This conclusion is implied by the following two claims.

Suppose $\hat{g}_\lambda(x)$ is the solution to the penalized sum of squares. Two claims about this function is as follows.

1. Given $\{t_i; \hat{g}_\lambda(t_i)\}$, based on the discussion in the last section there is a unique natural cubic spline $s(x)$ with knots in $\{t_1, \dots, t_k\}$ such that

$$s(t_i) = \hat{g}_\lambda(t_i), \quad i = 1, \dots, k.$$

Because of the above, we have

$$\sum_{i=1}^{n} \{y_i - s(x_i)\}^2 = \sum_{i=1}^{n} \{y_i - \hat{g}_\lambda(x_i)\}^2.$$

2 For the $s(x)$ defined above, we have

$$\int \{\hat{g}''_\lambda(x)\}^2 dx \geq \int \{s''(x)\}^2 dx$$

with the equality holds if and only if $\hat{g}_\lambda(x) = s(x)$ for all x. If this is true, we must have $\hat{g}_\lambda(x) = s(x)$, a natural cubic spline.

A serious proof is needed for the second claim. Here is the proof.

Let $\gamma_i = s''(t_i)$ for $i = 1, \ldots, k$ with $s(x)$ being a cubic spline with knots on $t_1, \ldots, t_k$. Being "natural", we have $\gamma_1 = \gamma_k = 0$.

Let $g(x)$ be another function with finite second derivatives such that $g(t_i) = s(t_i)$ for $i = 1, 2, \ldots, t_k$. It is seen that

$$\begin{aligned} \int_{t_i}^{t_{i+1}} g''(x)s''(x)dx &= \int_{t_i}^{t_{i+1}} s''(x)dg'(x) \\ &= [s''(t_{i+1})g'(t_{i+1}) - s''(t_i)g'(t_i)] - \int_{t_i}^{t_{i+1}} g'(x)s'''(x)dx, \end{aligned}$$

Note that

$$\sum_{i=1}^{k-1} [s''(t_{i+1})g'(t_{i+1}) - s''(t_i)g'(t_i)] = \gamma_k g'(t_k) - \gamma_1 g'(t_1) = 0.$$

Being linear on every interval $[t_i, t_{i=1}]$, we have

$$s'''(x) = \frac{\gamma_{i+1} - \gamma_i}{t_{i+1} - t_i} = \alpha_i$$

where we have used α_i for the slope. With this, we find

$$\int_{t_i}^{t_{i+1}} g'(x)s'''(x)dx = \alpha_i\{g(t_{i+1}) - g(t_i)\} = \alpha_i\{s(t_{i+1}) - s(t_i)\}$$

where the last equality is from the fact that $g(x)$ and $s(x)$ are equal at knots. Hence, we arrive at the conclusion that

$$\int_{t_1}^{t_k} g''(x)s''(x)dx = -\sum_{i=1}^{k} \alpha_i\{s(t_{i+1}) - s(t_i)\}.$$

This result is applicable when $g''(x) = s''(x)$. Hence, we also have

$$\int_{t_1}^{t_k} s''(x)s''(x)dx = -\sum_{i=1}^{k} \alpha_i\{s(t_{i+1}) - s(t_i)\}.$$

This implies that

$$\int_{t_1}^{t_k} g''(x)s''(x)dx = \int_{t_1}^{t_k} s''(x)s''(x)dx.$$

Making use of this result, we get

$$\int_{t_1}^{t_k} \{g''(x) - s''(x)\}^2 dx = \int_{t_1}^{t_k} \{g''(x)\}^2 dx - \int_{t_1}^{t_k} \{s''(x)\}^2 dx \geq 0.$$

This equality holds only if $g''(x) = s''(x)$ for all $x \in [t_1, t_k]$. Hence the overall conclusion is proved.

Consider the problem of searching for a natural cubic splines that minimizes the penalized optimization problem (within this class of functions). Given a function

$$g(x) = \sum_{j=1}^{k} \beta_j N_j(x)$$

for some constants $\beta_1, \ldots, \beta_k$, its sum of squared residuals is given by

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 = (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^\tau(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})$$

where

$$\mathbf{Z} = \begin{pmatrix} N_1(x_1) & \cdots & N_k(x_1) \\ \vdots & \vdots & \vdots \\ N_1(x_n) & \ldots & N_k(x_n) \end{pmatrix}.$$

The penalty term over interval $[t_1, t_k]$ for this $g(x)$ becomes

$$\int \{g''(x)\}^2 dx = \int \sum_{j=1}^{k} \sum_{l=1}^{k} \beta_j \beta_l N_j''(x) N_l''(x) dx = \boldsymbol{\beta}^T N \boldsymbol{\beta}$$

with

$$\mathbf{N} = (N_{jl})_{k \times k} \text{ and } N_{jl} = \int_{t_1}^{t_k} N_j''(x) N_l''(x) dx.$$

The penalized sum of squares of $g(x)$ is given by

$$(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^\tau (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\tau \mathbf{N} \boldsymbol{\beta}.$$

It is minimized, given λ, at

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{Z}^\tau \mathbf{Z} + \lambda \mathbf{N})^{-1} \mathbf{Z}^\tau \mathbf{y}$$

and the fitted regression function is

$$\hat{g}_\lambda(x) = \sum_{j=1}^{k} \hat{\beta}_{\lambda,j} N_j(x).$$

8.7 Effective number of parameters and the choice of λ

If we regard $\hat{g}_\lambda(x)$ as a fit based on a linear regression, then we seem to have employed k independent parameters. Due to regularization induced by penalty, the effective number of parameters is lower than k. Note that the fitted value of response vector is given by

$$\hat{\mathbf{y}}_\lambda = \mathbf{Z}(\mathbf{Z}^\tau \mathbf{Z} + \lambda \mathbf{N})^{-1} \mathbf{Z}^\tau \mathbf{y} = \mathbf{A}_\lambda \mathbf{y}.$$

We call $\mathbf{A}_\lambda$ smoother matrix. Similar to local polynomial kernel method, we define the effective degrees of freedom (dfs) or effective number of parameters to be

$$df_\lambda = \text{trace}(A_\lambda).$$

As λ increases, the effective number of parameters (df_λ) decreases and $\hat{g}_\lambda(x)$ becomes smoother and smoother. We can hence try out a range of λ values and examine the resulting $\hat{g}_\lambda(x)$ and select the most satisfactory one. However, this procedure needs human interference and cannot be automated.

To overcome this deficiency, one may choose λ using CV or GCV criteria. Similar to local polynomial kernel method, we define the GCV score as a function of λ to be

$$\text{GCV}(\lambda) = \frac{(\mathbf{y} - \hat{\mathbf{y}}_\lambda)^\tau (\mathbf{y} - \widehat{\mathbf{y}}_\lambda)}{\{1 - \text{trace}(\mathbf{A}_\lambda)/n\}^2}.$$

The GCVmethod chooses λ as the minimizer of GCV(λ).

The CV approach is similar. Let $\hat{g}_{-i}(x_i)$ be the estimate of $g(x_i)$ based on $n-1$ observations without the ith observation. We define the CV score as a function of λ to be

$$\text{CV}(\lambda) = \sum_{i=1}^n \{y_i - \hat{g}_{-i}(x_i)\}^2.$$

It turns out that

$$\text{CV}(\lambda) = \sum_{i=1}^n \left(\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \text{trace}(\mathbf{A}_{\lambda,i,i})} \right)^2.$$

This expression enable us to only fit the model once for each λ in order to compute CV(λ). The CV method chooses λ value as the minimizer of CV(λ).

Remark: The so-called R-functions are not included.

Chapter 9

Bayes method

Most of the data analysis methods we have discussed so far are regarded as frequentist methods. More precisely, these methods are devised based on the conviction that the data are generated from a fixed system which is a member of a family of systems. While the system is chosen by nature, the outcomes are random. By analyzing the data obtained/generated/sampled from this system, we infer the properties of **this** system. The methods devised subsequently are judged by their average performances when they are repeated applied to all possible realized data from **this system**. For instance, we regard sample mean as an optimal estimator for the population mean under normal model in some sense. Whichever $N(\theta, \sigma^2)$ is the true, on average, $(\bar{x} - \theta)^2$ has the lowest average among all $\hat{\theta}$ whose average equals θ. A procedure is judged optimal only if this optimality holds at each and every possible θ, σ^2 value.

When considered from such a frequentist point of view, the statisticians do not play favours to any specific system against the rest of them in this family. Simplistically, each system in the family is regarded of equal likelihood before hand. This view is subject to dispute. In some applications, we may actual have some preference between such systems. What is the chance that a patient entering a clinic with fever actually has a simple flu? If this occurs at a flu season, the doctor would immediately look for more signs of flu. If it is not a flu season, the doctor will cast a bigger net to the cause of the fever. The conclusion arrived by the doctor is not completely dependent on the evidence: having fever. This example shows that most of human being act on their prior belief.

The famous Bayes theorem provides one way to formally utilize prior information. Let A and B be two events in the context of probability theory. It is seen that the conditional probability of B given A

$$\textsc{pr}(B|A) = \frac{\textsc{pr}(A|B)\textsc{pr}(B)}{\textsc{pr}(A|B)\textsc{pr}(B) + \textsc{pr}(A|B^c)\textsc{pr}(B^c)}$$

where B^c is the complement of B, or the event that B does not occur. This formula is useful to compute the conditional probability of B after A is known to have occurred when all probabilities on the right hand side are known. The comparison between $\textsc{pr}(B|A)$ and $\textsc{pr}(B)$ reflects what we learn from event A about the likeliness of event B.

9.1 An artifical example

Suppose one of two students is randomly selected to write a typical exam. Their historical averages are 70 and 80 percent. After we are told the mark of this exam is 85%, which student has been selected in the first place?

Clearly, both are possible but most of us will bet on the one who has historical average of 80%. It turns out that Bayes theorem gives us a quantitative way to justify our decision if we are willing to accept some model assumptions.

Suppose the outcome of the exam results have distributions who densities are given by

$$\begin{aligned} f_a(x) &= \frac{x^{7-1}(1-x)^{3-1}}{\mathcal{B}(7,3)}\mathbb{1}(0<x<1); \\ f_b(x) &= \frac{x^{8-1}(1-x)^{2-1}}{\mathcal{B}(8,2)}\mathbb{1}(0<x<1) \end{aligned}$$

for students A and B with beta function defined to be

$$\mathcal{B}(a,b) = \int_0^1 x^a(1-x)^{b-1}dx$$

for $a, b, > 0$. The probability that they are selected to write the exam is

$$\text{PR}(A) = \text{PR}(B) = 0.5$$

which is our prior belief that reflects the random selection very well. Let X denote the outcome of the exam. It is seen that

$$\text{PR}(A|X = x) = \frac{0.5 f_a(x)}{0.5 f_a(x) + 0.5 f_b(x)}.$$

If $X = 85\%$, we find

$$\text{PR}(A|X = 85) = 0.3818.$$

If $X = 60\%$, we find

$$\text{PR}(A|X = 60) = 0.7000.$$

Based on these calculations, we seem to know what to do next.

To use the frequentist approach discussed earlier, we re-state this experiment as follows. One observation X has been obtained from a Beta distribution family with parameter space

$$\Theta = \{(7, 3); (8, 2)\}.$$

If $X = 0.85$, what is your estimate of θ?

The likelihood values at these two parameter points are given by

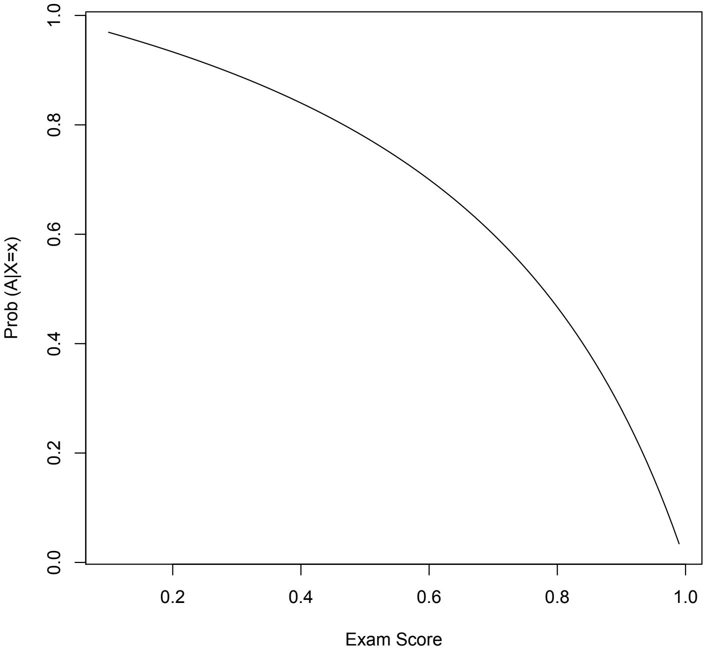
$$\begin{aligned} \ell((7, 3)) &= f_a(0.85) = 2.138; \\ \ell((8, 2)) &= f_b(0.85) = 3.462. \end{aligned}$$

Hence, the MLE is given by $\hat{\theta} = (8, 2)$ corresponding to student B.

Based on frequentist approach which ignores the prior information, we are told it is more likely that student B wrote the exam. If the MLE has been chosen as the frequentist method to be used, then student B is our conclusion, even though we know it is not certain.

Using Bayes analysis together with the prior information provided, we claim that there is a 82% chance that student B wrote the exam. At this moment, we have yet to make a decision. The calculation of the posterior probability itself does not directly provide one. Suppose wrongfully concluding it was written by student B may result in a loss of a million dollars, while wrongfully concluding it was student A may result in a loss of a single dollar, then we may still claim/act that it was student A who wrote the exam.

Figure 9.1: Posterior probability as a function of x



9.2 Classical issues related to Bayes analysis

We suggested that a statistical model is a family of distributions often represented as a collection of parameterized density functions. We use $\{f(x;\theta) : \theta \in \Theta\}$ as a generic notation. Often Θ is a subset of $\mathcal{R}^d$.

When a set of observations $\mathbf{X}$ are obtained and a statistical model is assumed, a frequentist would regard $\mathbf{X}$ is generated from **ONE** member of $\{f(x;\theta) : \theta \in \Theta\}$ but usually we do not know which one. The information contains in $\mathbf{X}$ helps us to decide which one is most likely, or a close proximate of this **ONE**.

In comparison, a Baysian may also regard $\mathbf{X}$ is generated from **ONE** member of $\{f(x;\theta) : \theta \in \Theta\}$. However, **this one** θ value itself is generated from another distribution called prior distribution, $\Pi(\theta)$. Hence, it is a realized value of a random variable whose distribution is given by $\Pi(\theta)$. If we have full knowledge of $\Pi(\theta)$, then it should be combined with $\mathbf{X}$ to infer which θ has been the θ in $\{f(x;\theta) : \theta \in \Theta\}$ that generated $\mathbf{X}$. We generally cannot nail down to a single θ value given $\mathbf{X}$ and $\Pi(\theta)$. With the help of Bayes theorem, we are able to compute the conditional distribution of θ given $\mathbf{X}$, which is called posterior. That is, we retain the random nature of θ but update our knowledge about its distributions when $\mathbf{X}$ becomes available. Statistical inference about θ will then be made based on the updated knowledge.

From the above discussion, it is seen that the a preliminary step in Bayes analysis is to obtain posterior distribution of θ, assuming the model itself has been given and the data have been collected. That is, we have already decided on the statistical model $f(x;\theta)$, prior distribution $\Pi(\theta)$ and data $\mathbf{X}$. Note that this $\mathbf{X}$ can be a vector of i.i.d. observations **given** θ. The notion of **GIVEN** θ is important because θ is a random variable in the context of Bayes analysis.

Particularly in early days, the Bayes analysis is possible only if some kind of neat analytical expression of the posterior is available. Indeed, I can give you many such examples when things lineup nicely.

Example 9.1. *Suppose we have an observation X from a binomial distribution $f(x;\theta) = C(n,x)\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n$. Suppose we set the* *prior distribution with density function*

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\mathcal{B}(a,b)} \mathbb{1}(0 < \theta < 1).$$

By Bayes rule, the density function of the posterior distribution of θ *is given by*

$$f_p(\theta | X = x) = \frac{f(x;\theta)\pi(\theta)}{\int f(x;\theta)\pi(\theta) d\theta}.$$

It appears to get explicit expression, we must find the outcome of the integration. However, this can often be avoided. Note that

$$f(x;\theta)\pi(\theta) = C(n,x)\theta^{a+x-1}(1-\theta)^{b+n-x-1}\mathbb{1}(0 < \theta < 1).$$

Hence, we must have

$$f_p(\theta | X = x) = \frac{\theta^{a+x-1}(1-\theta)^{b+n-x-1}\mathbb{1}(0 < \theta < 1)}{c(n,a,b,x)}$$

for some constant $c(a,b,x)$ *not depending on* θ*. As a function of* θ*, it matches the density function of Beta distribution with degrees of freedom* $a+x, b+n-x$*. At the same time, its integration must be 1. This shows that we must have*

$$c(n,a,b,x) = \mathcal{B}(a+1, n+b-x).$$

The posterior distribution is Beta with $a+x, n+b-x$ *degrees of freedom:*

$$f_p(\theta | X = x) = \frac{\theta^{a+x-1}(1-\theta)^{b+n-x-1}\mathbb{1}(0 < \theta < 1)}{\mathcal{B}(a+1, n+b-x)}$$

This will be the posterior distribution used for Bayes decision. □

You may notice that Binomial distribution and the Beta distribution are perfectly paired up to permit an easy conclusion on the posterior distribution. There are many such pairs. For instance, if X has Poisson distribution with mean θ, and θ has prior one parameter Gamma distribution, then the posterior distribution of θ is also Gamma. We leave this case as an exercise. Such prior distributions are call conjugate priors. Another good exercise problem is to draw the density function of many beta distributions. It helps to get an intuition on what you have assumed if a beta prior is applied.

Definition 9.1. *Let $\{f(x;\theta) : \theta \in \Theta\}$ be a statistical model. Namely, it is a family of distributions. Suppose for any prior distribution $\pi(\theta)$ as a member of distribution family $\{\pi(\theta;\xi) : \xi \in \Xi\}$, the posterior distribution of θ given a set of i.i.d. observations from $f(x;\theta)$ is a member of $\{\pi(\theta;\xi) : \xi \in \Xi\}$, then we say that $\{\pi(\theta;\xi) : \xi \in \Xi\}$ is a conjugate prior distribution family of $\{f(x;\theta) : \theta \in \Theta\}$.*

Remark: We have used

$$f_p(\theta|X = x) = \frac{f(x;\theta)\pi(\theta)}{\int f(x;\theta)\pi(\theta)d\theta}$$

in the above example. This formula is generally applicable. In addition, one should take note that the denominator in this formula does not depend on θ. Hence, the denominator merely serves as a scale factor in $f_p(\theta|X = x)$. In classical examples, its value can be inferred from the analytical form of the numerator. In complex examples, its value does not play a rule in Bayes analysis.

Example 9.2. *Suppose that given μ, $X_1, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma_0^2)$ with known σ_0^2. Namely, σ_0^2 is not regarded as random. The prior distribution of μ is $N(\mu_0, \tau_0^2)$ with both parameter values are known. The posterior distribution of μ given the sample is still normal with parameters*

$$\mu_B = \frac{n\bar{x}/\sigma_0^2 + \mu_0/\tau_0^2}{1/\sigma_0^2 + 1/\tau_0^2};$$

and

$$\sigma_B^2 = \left[\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right]^{-1}.$$

The philosophy behind the Bayes data analysis is to accommodate our prior information/belief about the parameter in statistical inference. Sometime, prior information naturally exists. For instance, we have a good idea on the prevalence of human sex ratio. In other applications, we may have some idea on certain parameters. For example, the score distribution of a typical course. Even if we cannot perfectly summarize our belief with a prior distribution, one of the distributions in the beta distribution family can be good enough.

It is probably not unusual that we do not have much idea about the parameter value under a statistical model assumption. Yet one may be attracted to the easiness of the Bayesian approach and would like to use Bayes analysis anyway. She may decide to use something called non-informative prior. Yet there seem to be no regular definition on what a prior is a non-informative prior.

In the normal distribution example, one may not have much idea about the mean of the distribution in a specific application. If one insists on use Bayesian approach, he or she may simply use a prior density function

$$\pi(\mu) = 1$$

for all $\mu \in \mathcal{R}$. This prior seems to reflect the lack of any idea on which μ value is more likely than any other μ values. In this case, $\pi(\mu)$ is not even a proper density function with respect to Lebesgue measure. Yet one may obtain a proper posterior density following the rule of Bayes theorem.

It appears to me that Bayes analysis makes sense when prior information about the parameter truly exists. In some occasions, it does not hurt to employ this tool even if we do not have much prior information. If so, the Bayes inference conclusion should be critically examined just likely any other inference conclusions.

9.3 Decision theory

Let us back to the position that a statistical model $f(x;\theta)$ is given, prior distribution $\Pi(\theta)$ is chosen and data $\mathbf{X}$ have been collected. At least in principle, the Bayes theorem has enabled us to obtain posterior distribution of θ: $f_p(\theta|\mathbf{X})$. At this point, we need to decide how to estimate θ, the value generated from $\Pi(\theta)$, and $\mathbf{X}$ is a random sample from $f(x;\theta)$ with **this** θ. With $f_p(\theta|\mathbf{X})$ at hand, how do you estimate θ?

First of all, you may pick any function of $\mathbf{X}$ as your estimator of θ. This has not changed.

Second, if you wish to find a superior estimator, then you must provide a criterion to judge superiority. In the content of Bayes data analysis, the criteria for point estimation is through loss functions.

Definition 9.2. *Assume a probability model with parameter space Θ. A loss function $\ell(\cdot, \cdot)$ is a non-negative valued function on $\Theta \times \Theta$ such that $\ell(\theta_1, \theta_2) = 0$ when $\theta_1 = \theta_2$.*

Finally, since we do not know what the true θ value is, with the posterior distribution, we can only hope to minimize the average loss. Hence, the decision based on the bayes rule is to look for $\hat{\theta}$ such that the expected loss is minimized:

$$\int L(\hat{\theta}, \theta) f_p(\theta|\mathbf{X}) d\theta = \min.$$

A naturally choice of the loss function is

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2.$$

The solution to this loss function is clearly the **posterior mean** of θ for one-dimension θ.. This extends to the situation where θ is multidimensional.

One may use the loss function

$$L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|.$$

If so, the solution is the **posterior median** for one-dimension θ. The extension to the multidimensional θ is possible.

Example 9.3. *Suppose we have an observation X from a binomial distribution $f(x;\theta) = C(n,x)\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n$. Suppose we set the prior distribution with density function*

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\mathcal{B}(a,b)} \mathbb{1}(0 < \theta < 1).$$

By Bayes rule, the density function of the posterior distribution of θ is given by

$$f_p(\theta|X = x) = \frac{f(x;\theta)\pi(\theta)}{\int f(x;\theta)\pi(\theta)d\theta}.$$

The posterior distribution is Beta with $a + x, n + b - x$ degrees of freedom:

$$f_p(\theta|X = x) = \frac{\theta^{a+x-1}(1-\theta)^{b+n-x-1}\mathbb{1}(0 < \theta < 1)}{\mathcal{B}(a+1, n+b-x)}$$

If the square loss is employed, then the Bayes estimator of θ is given by

$$\int \theta f_p(\theta|X = x)d\theta = \frac{a+x}{a+b+n}.$$

When $a = b = 1$, the prior distribution of θ is uniform on (0, 1). This is regarded as a non-informative prior. With this prior, we find

$$\hat{\theta} = \frac{x+1}{n+2}$$

which seems to make more sense than the MLE x/n. □

Since Bayes estimator is generally chosen as the minimizer of some expected posterior loss, it is optimal in this sense by definition. However, the optimality is judged with respect to the specific loss function and under the assumed prior. Blindly claiming a Bayes estimator is optimal out of content is not recommended here. If this logic is applicable, then we would as rightfully claim that the MLE is optimal, because it maximizes a criterion function called likelihood. Such a claim would be ridiculous because we have many examples where the MLEs are not even consistent.

We will have an exercise problem to work out Bayes estimators under square loss under normal model with some conjugate prior distribution on both mean and variance.

Once the posterior distribution is ready, we are not restricted to merely give a point estimation. These issues will be discussed in other parts of this course. At the same time, we may get some sense that being able to precisely describing the posterior distribution is one of the most important topic in Bayes data analysis.

9.4 Some comments

There are two major schools on how the statistical data analysis should be carried out: frequentist and Bayesian. If some prior information exists and can be reasonably well summarized by some prior distribution, then I feel the inference based on Bayes analysis is fully justified. If one does not have much sensible prior information on the statistical model appropriate to the data at hand, it is still acceptable to use the formality of the Bayes analysis. Yet blindly claiming the superiority of a Bayesian approach is not of my taste. Particularly in the later case, the Bayes conclusion should be critically examined as much as any data analysis methods.

To make things worse, many statisticians seem to regard themselves doing research on Bayesian methods, yet they do not aware the principle of the Bayes analysis. Probably, they merely feel that this is an easy topic to publish papers (not true if one is a serious Bayesian). To be more strict, a Bayesian should have a strong conviction that the model parameters are invariably realized values from some distribution. There is an interest and very valid question, is/was Bayes a Bayesian?

Chapter 10

Monte Carlo and MCMC

Recall that a statistical model is a distribution family, at least this is what we believe. Let us first focus on parametric models: $\{f(x;\theta) : \theta \in \Theta\}$. In this case, θ is generally a real valued vector and Θ is a subset of Euclidean space with nice properties such as convex, open and so on. After placing a prior distribution on θ, we have created a Bayes model. We do not seem to be a consensus on a definition of and a notation for Bayes model, even though statisticians are not shy at using this terminology. Based on my understanding, I define a Bayes model as a system with two important components: a family of distributions, and a prior distribution on distribution on this family:

$$\text{Bayes Model} = [\{f(x;\theta) : \theta \in \Theta\}, \pi(\theta)].$$

Logically, a Bayes model is not the same as Bayes analysis. When Θ is a subset of Euclidean space, we generally regard $\pi(\cdot)$ a density function with respect to Lesbesgue measure on Θ.

Suppose a θ value is generated according to $\pi(\cdot)$, and subsequently, a data set $\mathbf{X}$ is generated from **THIS** $f(x;\theta)$. Here we implicitly assume that $\mathbf{X}$ is accurately measured and available to use for the purpose of inference. The inference target is θ based on data from this experiment. Any decision about the possible value of θ in Bayes analysis will be based on the posterior density of θ given $\mathbf{X}$. We use notation $f_p(\theta|\mathbf{X})$ for posterior distribution (density). It is conceptually straightforward to define and derive the posterior distribution.

Hence, there are not much left for a statistician to do.

Bayes analysis makes a decision based on posterior distribution. Research on Bayesis methods includes: (a) most suitable prior distributions in specific applications; (b) the influence of the choice of prior distribution to the final decision; (c) numerical or theoretical methods for posterior distribution; (d) properties of the posterior distribution; (e) decision rule. There might be more topics out there. This chapter is about topic (c).

For some well paired up $f(x;\theta)$ and $\pi(\theta)$ (when $\pi(\cdot)$ is a conjugate prior for $f(x;\theta)$), it is simple to work out the analytical form of the posterior density function. A Bayesian needs only decide the best choices of $\pi(\theta)$ and the subsequent decision rule. In many real world problems, the posterior density is on high dimensional space and does not have an simple form. The Bayes analysis before the contemporary computing power has been a serious challenge. This problem becomes less and less an issue today. We discuss a number of commonly used techniques in this chapter.

10.1 Monte Carlo Simulation

The content of this section is related but not limited to Bayes analysis. Suppose in some applications, we wish to compute $\mathbb{E}\{g(X)\}$ and X is known to have certain distribution. This is certainly a simple task in many textbook examples. For instance, if X has Poisson distribution with mean θ and $g(x) = x(x-1)(x-2)(x-3)$, then

$$\mathbb{E}\{g(X)\} = \theta^4.$$

However, if $g(x) = x\log(x+1)$, the answer to $\mathbb{E}\{g(X)\}$ is not analytically available.

Suppose we have an i.i.d. sample $x_1, \ldots, x_n$ with sufficiently large n from this distribution, then by the law of large numbers,

$$\mathbb{E}\{g(X)\} \approx n^{-1}\sum_{i=1}^{n} x_i \log(x_i).$$

Let us generate $n = 100$ values from Poisson distribution with $\theta = 2$. Using a function in R-package, we get 100 values

```
5 2 3 4 1 2 1 2 1 1 2 3 2 2 2 3 1 2 0 4 1 2 5 1 1
2 3 1 1 1 2 0 2 1 1 3 0 5 1 5 1 2 1 0 2 3 5 2 6 3
2 4 3 1 1 2 2 1 1 2 2 5 0 2 1 3 3 1 3 1 1 2 2 3 1
2 1 4 0 4 2 3 0 0 2 1 3 1 0 2 1 0 3 1 3 6 1 3 3 3
```

Based on this sample, we get an approximated value

$$\mathbb{E}\{G(X)\} \approx 2.691.$$

I can just as easily use $n = 10,000$ and find $\mathbb{E}\{g(X)\} \approx 2.648$ in one try. With contemporary computer, we can afford to repeat it as many times as we like: $\mathbb{E}\{g(X)\} \approx 2.642, 2.641, 2.648$. It appears $\mathbb{E}\{g(X)\} = 2.645$ would be a very accurate approximations. Computation based on simulated data is generally called Monte Carlo method.

We must answer two questions before we continue. The first is why do not we use a numerical approach if we need to compute $\mathbb{E}\{g(X)\}$. Indeed, we can put up a quick R-code

```
{ii= 0:50; sum(ii*log(1+ii)*dpois(ii, 2))}
```

and get a value 2.647645. This is a very accurate answer to this specific problem. Yet if we wish to compute

$$\mathbb{E}\{(X_1 + \sqrt{X_2})^2 \log(X_1 + X_3 X_4)\},$$

where X_1, X_2, X_3, X_4 may have a not very simple joint distribution, a neat numerical solution becomes hard. Since the contemporary computers are so powerful, The above problem is only “slightly” harder. Yet there are real world problem of this nature, but involves hundreds or more random variables. For these problems, the numerical problem quickly becomes infeasible even for contemporary computers. In comparison, the complexity of the Monte Carlo method remains the same even when $g(X)$ is a function of vector X with high dimension.

The second question is how easy is it to generate quality “random samples” from a given distribution by computer? There are two issues related to this question. First, the computer does not have an efficient way to generate random numbers. However, with some well designed algorithms, it can produce massive amount of data which appear purely random. We call them pseudo random number generators. We do not discuss this part of the problem in this course. The other issue is how to make sure these random numbers behave like samples from the desired distributions.

Our starting point is that it is easy to generate i.i.d. observations (pseudo numbers) from uniform distribution $[0,1]$. We investigate the techniques for generating i.i.d. observations from other distributions.

Theorem 10.1. *Let $F(x)$ be any univariate continuous distribution function and U be a standard uniformly distributed random variable. Let*

$$Y = \inf\{x : F(x) \geq U\}.$$

Then the distribution function of Y is given $F(\cdot)$.

As an exercise, show that $F(Y)$ has uniform distribution when the distribution of Y is continuous and given by $F(\cdot)$.

Proof. We only need to work out the c.d.f. of Y. If it is the same as $F(\cdot)$, then the theorem is proved.

Routinely, we have

$$\text{PR}(Y \leq t) = \text{PR}(\inf\{x : F(x) \geq U\} \leq t) = \text{PR}(F(t) \geq U) = F(t)$$

because $\text{PR}(U \leq u) = u$ for any $u \in (0,1)$. This completes the proof. □

Since we generally only have pseudo numbers in U, applying this too will only lead to "pseudo numbers" in Y.

Example 10.1. *Let $g(u) = -\log u$. Then, $Y = g(U)$ has exponential distribution if U has standard uniform distribution.*

Let $g(u) = (-\log u)^a$ for some positive constant a. Then $Y = g(U)$ has Weilbull distribution.

As an exercise problem, find the function $g(\cdot)$ which makes $g(U)$ standard Cauchy distributed.

Here is another useful exercise problem for knowledge. If Z_1, Z_2 are independent standard normally distributed random variables, then $r^2 = Z_1^2 + Z_2^2$ are exponentially distributed. One should certainly know that r^2 is also chisquare distributed with 2 degrees of freedom.

Example 10.2. *Let U_1, U_2 be two independent standard uniform random variables. Let*

$$\begin{aligned} g_1(s,t) &= \sqrt{-2\log s}\cos(2\pi t); \\ g_2(s,t) &= \sqrt{-2\log s}\sin(2\pi t). \end{aligned}$$

Then, $g_1(U_1, U_2), g_2(U_1, U_2)$ are two independent standard normal random variables.

If we can efficiently generate pseudo numbers from uniform distribution, then the above result enables us to efficiently generate pseudo numbers from standard normal distributions. Since general normal distributed random variables are merely location-scale shifted standard normal random variables, their generation can hence also be efficiently generated this way.

Due to well established relationship between various distributions, pseudo numbers from many many classical distributions can be efficiently generated. Here are a few well-known results which were also given in the chapter about normal distributions.

Example 10.3. *Let $Z_1, Z_2, \ldots$ be i.i.d. standard normally distributed random variables.*

(a) $X_n^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ has chisquare distribution with n degrees of freedom.

(b) $F_{n,m} = (X_n^2/n)/(Y_m^2/m)$ has F distribution with n, m degrees of freedom when X_n^2, Y_m^2 are independent.

(c) $B_n = (X_n^2)/(X_n^2 + Y_m^2)$ has Beta distribution with n, m degrees of freedom when X_n^2, Y_m^2 are independent.

We can also generate multinomial pseudo numbers with any probabilities: $p_1, p_2, \ldots, p_m$: generate U from uniform, then let $X = k$ for k such that

$$p_1 + \cdots + p_{k-1} < U \leq p_1 + \cdots + p_{k-1} + p_{k-1}.$$

The left hand side is regarded as zero for $k = 1$.

10.2 Biased or importance sampling

: Back to the problem of computing $\mathbb{E}\{$g(X)$\}$ when X has a distribution with density or probability mass function $f(x)$. If generating pseudo numbers from $f(x)$ is efficient, then it is a good idea to approximate this expectation by

$$n^{-1}\sum_{i=1}^{n} g(x_i).$$

If it is more convenient to generate law of large numbers which recommends pseudo numbers from a different distribution $f_0(x)$ which has the same support as $f(x)$, then it is easier to approximate this expectation by

$$n^{-1}\sum_{i=1}^{n}\{g(y_i)f(y_i)/f_0(y_i)\}$$

where $y_1, \ldots, n$ observations are generated from $f_0(x)$.

If Y has distribution given by density $f_0(x)$, we have

$$\begin{aligned}\mathbb{E}\{g(Y)f(Y)/g_0(Y)\} &= \int \{g(y)f(y)/f_0(y)\}f_0(y)dy \\ &= \int g(y)f(y)dy = \mathbb{E}\{g(X)\}\end{aligned}$$

where X has distribution $f(x)$. Note that it is important that f and f_0 have the same support so that the range of integrations remains the same. If X has discrete distribution, the integration will be changed to summation. The conclusion is not affected.

In sample survey, the units in the finite population often have different probabilities to be included in the sample due to various considerations. The population total

$$Y = \sum_{i=1}^{N} y_i,$$

where N is the number of sample units in the finite population and y_i is the response value of the ith unit, is often estimated by Horvath-Thompson estimator:

$$\hat{Y} = \sum_{i\in s} y_i/\pi_i$$

where s is the set of units sampled and π_i is the probability that the unit i is in the sample. The role of π_i is the same as $f_0(x)$ in the importance sampling content.

In sampling practice, some units with specific properties of particular interest are hard to obtain in an ordinary sampling plan. Specific measures are often taken so that these units have higher probability to be included than otherwise when all units are treated equally. The practice may also be regarded as finding a specific $f_0(x)$ to replace $f(x)$ even though the expectation of $g(X)$ under $f(x)$ distribution is the final target. One such example is to obtain the proportional of HIV+ person in Vancouver population. A simple random sample may end up with a sample of all HIV- individuals giving lower accurate estimation of the rate of HIV+. The same motivation is used in numerical computation. If $f(x)$ has lower values in certain region of x, then a straightforward random number generator will have very few values generated from that region. This problem makes such numerical approximations inefficient. Searching for some $f_0(x)$ can be a good remedy to address this shortcoming.

Here is another example. To estimate the survival time of cancer patient. Let us a random sample from all cancer patients at a specific time point. If their survive times are denoted as $Y_1, Y_2, \ldots, Y_n$ whose distribution is denoted as $f_0(y)$. The actually survival distribution would be different if every cancer patient is counted equally. This is because $f_0(y) \propto y f(y)$ where $f(y)$ is the "true" survival time distribution. This may also be regarded as importance sampling created by nature.

10.3 Rejective sampling

Instead of generating data from an original target distribution $f(x)$, we may generate data from $f_0(x)$ and obtain more effective numerical approximation of $\mathbb{E}\{g(X)\}$. This is what we have seen in the last section. The same idea is at work in rejective sampling. The target of this game is to obtain pseudo numbers which may be regarded as random samples from $f(x)$. Of course, to make it a good tool, we must select an $f_0(x)$ which is easy to handle.

Let $f(x)$ be the density function from which we wish to get random samples. Let $f_0(x)$ be a density function with the same support and further

$$\sup_x \frac{f(x)}{f_0(x)} = \textsc{u} < \infty$$

Denote

$$\pi(x) = \frac{f(x)}{\textsc{u} f_0(x)}.$$

Apparently, $\pi(x) \leq 1$ for any x. In addition, if $f(x)$ is known up to a constant multiplication, the above calculations remain feasible. One potential example of such an $f(x)$ is when

$$f(x) = \frac{C \exp(-x^4)}{1 + x^2 + \sin^2(x)}.$$

Since $f(x) > 0$ and its integration converges, we are sure that

$$C^{-1} = \int \frac{\exp(-x^4)}{1 + x^2 + \sin^2(x)} dx$$

is well defined. Yet we do not have its exact value. In this example, an accurate approximate value of C is not hard to get. Yet if $f(\cdot)$ is the joint density of many variables, even a numerical approximation is not feasible. Particularly in Bayes analysis, this can occur. If an effective way to generate "random" samples from $f(x)$ is possible, then we do not need to know C any more in many applications.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. samples from $f_0(x)$ and $U_1, U_2, \ldots$ be i.i.d. samples from uniform distribution. For $i = 1, 2, \ldots$, if $U_i \leq \pi(X_i)$, let $Y_i = X_i$; otherwise, we leave Y_i undefined. Hence, $Y_1, Y_2, \ldots$ is a subsequence of $X_1, X_2, \ldots$ after some X_i rejected. Hence, this procedure is called rejective sampling.

The output of the rejective sampling, Y_i, has distribution $F(x)$ with density function $f(x)$ for any i. This is demonstrated as follows. First, we consider the case for $i = 1$. It is seen that

$$\textsc{pr}\{U > \pi(X)\} = \mathbb{E}\{1 - \pi(X)\} = 1 - \int \pi(x) f_0(x) dx = 1 - \textsc{u}^{-1}.$$

Hence, the distribution of Y_1 is given by

$$\begin{aligned}
\text{PR}(Y_1 \le y) &= \sum_{k=1}^{\infty} \text{PR}(U_1 > \pi(X_1), \ldots, U_{k-1} > \pi(X_{k-1}), U_k < \pi(X_k), X_k \le y) \\
&= \sum_{k=1}^{\infty} (1 - \text{U}^{-1})^{k-1} \text{PR}(U_k < \pi(X_k), X_k \le y) \\
&= \sum_{k=1}^{\infty} (1 - \text{U}^{-1})^{k-1} \text{PR}(U < \pi(X), X \le y) \\
&= \text{U}\mathbb{E}\{\text{PR}\big(X \le y, U \le \pi(X) | X\big)\} \\
&= \text{U}\mathbb{E}\{\pi(X)\mathbb{1}(X \le y)\}.
\end{aligned}$$

Taking the definition of $\pi(x)$ into consideration, we find

$$\text{PR}(Y_1 \le y) = \text{U} \int_{-\infty}^{y} \frac{f(x)}{\text{U} f_0(x)} f_0(x) dx = F(y).$$

This shows that the rejective sampling method indeed leads to random numbers from the target distribution.

Let us define the waiting time

$$T = \min\{i : U_i \le \pi(X_i)\}$$

which is the number of pairs of pseudo numbers in (X, U) it takes to get a pseudo observation Y. We find its probability mass function is given by

$$\begin{aligned}
\text{PR}(T = k) &= \text{PR}\big(U_1 > \pi(X_1), \ldots, U_{k-1} > \pi(X_{k-1}, U_k < \pi(X_k)\big) \\
&= (1 - \text{U}^{-1})^{k-1} \text{U}^{-1}.
\end{aligned}$$

That is, T has geometric distribution with mean U.

If we use an f_0 which leads to large u, the rejective sampling is numerically less efficient. It takes more tries on average to obtain one sample from the target distribution. The best choice is $f_0(\cdot) = f(\cdot)$. Of course, this means we are not using a rejective sampling tool at all.

Here is an exercise problem. Suppose we want to generate random numbers from standard normal distribution whose density is given by $\phi(x) =$

$(2\pi)^{-1/2}\exp(-x^2/2)$. Some how, we wish to generate data from double exponential:

$$f_0(x) = \frac{1}{2}\exp(-|x|).$$

Compute the constant U as defined above. Write a code in R to implement the rejective sampling method to generate $n = 1000$ observations from N(0, 1). Show the Q-Q plot of the data generated and report the number of pairs of (X, U) in rejective sampling required. How many pairs of (X, U) do you expect to be needed to generate $n = 1000$ normally distributed random numbers with this method?

10.4 Markov chain Monte Carlo

Not an expert myself, my comments here may not be accurate. The rejection sample approach appears to be effective for generating univariate random variables (pseudo numbers). In applications, we may wish to generate a large quantity of vector valued observations. Markov chain Monte Carlo seems to be one of the solutions to this problem. To introduce this method, we need a dose of Markov chain.

10.4.1 Discrete time Markov chain

A Markov chain is a special type of stochastic process. A stochastic process in turn is a collection of random variables. Yet we cannot pay equal amount of attention to all stochastic processes but the ones that behave themselves. Markov chain is one of them.

We narrow our focus even further on processes containing a sequence of random variables having a beginning but no end:

$$X_0, X_1, X_2, \ldots.$$

The subindices $\{0, 1, 2, \ldots\}$ are naturally called time. In addition, we consider the case where X_n takes values in the same space with countable members for all n. Without loss of generality, we assume the space is

$$\mathcal{S} = \{0, \pm 1, \pm 2, \ldots\}.$$

We call $\mathcal{S}$ state space. For such a stochastic process, we define transition probabilities for $s < t$ to be

$$p_{ij}(s,t) = \text{PR}(X_t = j | X_s = i).$$

Definition 10.1. *A discrete time Markov chain is an ordered sequence of random variables with discrete state space $\mathcal{S}$ and has Markov property:*

$$\text{PR}(X_{s+t} = j | X_s = i, X_{s-1} = i_1, \ldots, X_{s-k} = i_k) = p_{ij}(s, s+t)$$

for all $i, j \in \mathcal{S}$ and $s, t \geq 0$.

If further, all one-step transition probabilities $p_{ij}(s, s+1)$ do not depend on s, we say the Markov chain is time homogeneous.

The Markov property is often referred to as: given present, the future is independent of the past. In this section, we further restrict ourselves to homogeneous, discrete time Markov chain. We will work as if $\mathcal{S}$ is finite and

$$\mathcal{S} = \{1, 2, \ldots, N\}.$$

The subsequent discussion does not depend on this assumption. Yet most conclusions are simpler to understand under this assumption. We simplify the one step transition probability notation to $p_{ij} = \text{PR}(X_1 = j | X_0 = i)$.

Let $\mathbf{P}$ be a matrix formed by one step transition probabilities: $\mathbf{P} = (p_{ij})$. For finite state space Markov chain, its size is $N \times N$. We may also notice its row sums equal to 1. It is well known that the t-step transition matrix

$$\mathbf{P}^{(t)} = \{\text{PR}(X_t = j | X_0 = i)\} = \mathbf{P}^t$$

for any positive integer t. For convenience, we may take 0-step transition matrix as $\mathbf{P}^0 = \mathbb{I}$, the identity matrix. The relationship is so simple, we do not need a specific notation for t-step transition matrix.

Let Π_t be the column vector made of $\text{PR}(X_t = i), i = 1, 2, \ldots, N$ and $t = 0, 1, \ldots$. This vector fully characterizes the distribution of X_t. Hence, we simply call it the distribution of X_t. It is seen that

$$\Pi_t^\tau = \Pi_0^\tau \mathbf{P}^t.$$

Namely, the distribution of X_t in a homogeneous discrete time Markov chain is fully determined by the distribution of X_0 and the transition probability matrix $\mathbf{P}$.

Under some conditions, $\lim_{t\to\infty} \Pi_t$ always exists. The limit itself is unique and is a distribution on the state space $\mathcal{S}$. For a homogeneous discrete time Markov chain with finite state space, the following conditions are sufficient:

(a) irreducible: for any $(i, j) \in \mathcal{S}$, there exists a $t \geq 1$ such that $\text{PR}(X_t = j|X_0 = i) > 0$.

(b) aperiodic: the greatest common factor of $\{t : \text{PR}(X_t = i|X_0 = i) > 0\}$ is 1 for any $i \in \mathcal{S}$.

When a Markov chain is irreducible, all states in $\mathcal{S}$ have the same period which is defined as the greatest common factor of $\{t : \text{PR}(X_t = i|X_0 = i) > 0\}$.

Theorem 10.2. *If a homogeneous discrete time Markov chain has finite space and properties (a) and (b), then for any initial distribution* Π_0*,*

$$\lim_{t\to\infty} \Pi_t = \Pi$$

exists and is unique.

We call Π in the above theorem as equilibrium distribution and such a Markov chain ergodic. It can be shown further that when these conditions are satisfied, then for any $i, j \in \mathcal{S}$,

$$\lim_{t\to\infty} \text{PR}(X_t = j|X_0 = i) = \pi_j$$

where π_j is the jth entry of the equilibrium distribution Π.

Definition 10.2. *For any homogeneous discrete time Markov chain with transition matrix* $\mathbf{P}$ *and state space* $\mathcal{S}$*, if* Π *is a distribution on the state space such that*

$$\Pi^\tau = \Pi^\tau \mathbf{P}$$

when we call it a stationary distribution.

It is seen that the equilibrium distribution is a stationary distribution. However, there are examples where there exist many stationary distributions but there is no equilibrium distribution.

Finally, we comment on the relevance of this section to MCMC. If one wishes to generate observations from a distribution $f(x)$. It is always possible for us to find a discrete distribution Π whose c.d.f. is very close that that of $f(x)$. Suppose we can further create a Markov chain with proper state space and transition matrix with Π as its equilibrium distribution. If so, we may generate random numbers from this Markov chain: $x_1, x_2, \ldots$. When t is large enough, the distribution of X_t is nearly the same as the target distribution Π.

The Markov chain Monte Carlo also works for continuous distributions. However, the general theory cannot be presented without a full course on Markov chain. This section is helpful to provide some intuitive justification on the Markov chain Monte Carlo in the next section.

10.5 MCMC: Metropolis sampling algorithms

Sometime, direct generation of i.i.d. observations from a distribution $f(\cdot)$ is not feasible. Rejective sampling can also be difficult because to find a proper $f_0(\cdot)$ is not easy. These happen when $f(\cdot)$ is the distribution of a high-dimensional random vector, or it does not have an exact analytical form. Markov chain Monte Carlo is regarded as a way out in recent literature. Yet you will see that the solution is not to provide i.i.d. random numbers/vectors, but dependent with required marginal distributions.

Let $X_0, X_1, X_2, \ldots$ be random variables that form a time-homogeneous Markov **process**. We use process here instead of chain to allow the rang of X to be $\mathcal{R}^d$ or something generic. It has all the properties we mentioned in the last section "otherwise". We define the kernel function $K(x, y)$ be the conditional density function of X_1 given X_0. Roughly speaking,

$$K(x, y) = \text{PR}(X_1 = y | X_0 = x) = \frac{f(x, y)}{f_X(x)}$$

which is the transition probability when the process is in fact a chain. We may also use

$$K(x, y) = f_{1|0}(x_1 | x_0)$$

as the conditional density of X_1 given X_0 when the joint density is definitely needed.

One Metropolis sampling algorithm goes as follows.

1. Let $t = 0$ and choose a x_0 value.

2. Choose a *proposed kernel* $K_0(x, y)$ so that the corresponding Markov process is convenient to generate random numbers/vectors from the conditional density.

3. Choose a function $r(x, y)$ taking values in [0, 1] and $r(x, x) = 1$.

4. Generate a y value from conditional distribution $K_0(x_t, y)$ and a standard uniform random number u. If $u < r(x_t, y)$, let $x_{t+1} = y$; otherwise, let $x_{t+1} = x_t$. Update $t = t + 1$.

5. Repeat step 4 until sufficient number of random numbers are obtained.

In the above algorithm, we initially generate random numbers from a Markov chain with transition probability matrix specified by $K_0(x, y)$. Due to a rejective sampling step, the many outcomes are not accepted but the previous value x_t is retained. What have we obtained?

We can easily seen that $\{x_0, x_1, \ldots\}$ remains a Markov chain with the same state space in spite of rejecting many y values generated according to K_0. We use Markov **chain** to illustrate the point. The transition probability of this Markov chain is computed as follows. Consider the case when $X_0 = i$ and the subsequent Y is generated according to the conditional distribution $K(i, \cdot)$. Let U be i.i.d. uniform [0, 1] random variables. For any $j \neq i \in \mathcal{S}$, we have

$$K(i, j) = \text{PR}(X_1 = j | X_0 = i) = \text{PR}(U < r(i, Y), Y = j) = r(i, j) K_0(i, j).$$

Clearly, the chance of not making a move is

$$K(i, i) = 1 + K_0(i, i) - \sum_{j=1}^{\infty} K_0(i, j).$$

Suppose the target distribution has probability mass function Π. We hope to select $K_0(x, y)$ and $r(x, y)$ so that Π is the equilibrium distribution of the Markov chain with transition matrix $K(x, y)$. Consider the situation where the working transition matrix $K_0(x, y)$ is symmetric and we choose for all i, j,

$$r(i, j) = \min\{1, \Pi(j)/\Pi(i)\}$$

in the above so called Metropolis algorithm. One important property of this choice is that we need not know individual values of $\Pi(i)$ for each i but their ratios. This is a useful property in Bayes method where the posterior density function is often known up to a constant factor. Computing the value of the constant factor is not a pleasant task. The above choice of $r(i, j)$ makes the computation unnecessary which is a big relief.

With this choice of $r(x, y)$, we find

$$\begin{aligned}\Pi(i)K(i,j) &= \min\{\Pi(i), \Pi(j)\}K_0(i,j)\\ &= \min\{\Pi(i), \Pi(j)\}K_0(j,i)\\ &= \Pi(j)K(j,i).\end{aligned}$$

This property is a sufficient condition for Π to be the equilibrium distribution of the Markov chain with transition probabilities given by $K(i, j)$. Note that the existence of the equilibrium distribution is assumed and can be ensured by the choice of an appropriate $K_0(i, j)$.

Although Step 4 in the Metropolis algorithm is very similar to the rejective sampling, they are not the same. In rejective sampling, if a proposed value is rejected, this value will be thrown out and a new candidate will be generated. In current Step 4, if a proposed value is rejected, the previous value in the Markov chain will be adopted.

We presented the result for discrete time homogeneous Markov chain with countable state space. The symbolical derivation for general state space is the same.

The symmetry requirement on $K_0(x, y)$ is not absolutely needed to ensure the limiting distribution is given by Π. When $K_0(x, y)$ is not symmetric, we may instead choose

$$r(x, y) = \min\big\{1, \frac{f(y)K_0(y, x)}{f(x)K_0(x, y)}\big\}.$$

We use x, y here to reinforce the impression that both x, y can be real values, not just integers.

A toy exercise is to show that this choice also leads to $f(x)$ satisfying the balance equation:

$$f(x)K(x, y) = f(y)K(y, x).$$

Finally, because $f(x)$ is the density function of the equilibrium distribution, when $t \to \infty$, the distribution of X_t generated from the Metropolis algorithm is $f(x)$. At the same time, the distribution of X_t for any finite t is not $f(x)$ unless that of X_0 is. However, for large enough t, we may regard the distribution of X_t as $f(x)$. This is the reason why a burning period is needed before we use X_t as random samples from $f(x)$ in many applications.

Obviously, X_t, X_{t+1} generated by this algorithm are not independent except for very special cases. However, in many applications, a non-i.i.d. sequence suffices. For instance, when the Markov chain is ergodic,

$$n^{-1} \sum_{t=1}^{n} g(X_t) \to \mathbb{E} g(X)$$

where $\mathbb{E}$ is computed with respect to the limiting distribution.

10.6 The Gibbs samplers

Gibbs samplers are another class of algorithms to generate random numbers based on a Markov chain. Suppose $X = (U, V)$ has joint distribution $f(u, v)$ with both u and v can be real valued vectors. Suppose that given $U = u$ for any u, it is easy to generate a value v from conditional distribution of $V|(U = u)$; and the opposite is also true. The goal is to generate number vectors with distribution of U, with distribution of V, or with distribution of (U, V).

A Gibbs sampler as follows leads to a Markov chain/process whose equilibrium distribution is that of U.

1. Pick a value u_0 for U_0. Let $t = 0$.

2. Generate a value v_t from the conditional distribution $V|(U = u_t)$.

3. Generate a value u_{t+1} from the conditional distribution $U|(V = v_t)$.

4. Let $t = t + 1$ and go back to Step 2.

Theorem 10.3. *The random numbers generated from the above sampler with joint distribution/density $f(u, v)$ form an observed sequence of a Markov chain/process $\{U_0, U_1, \ldots\}$.*

The limiting distribution of U_t is the marginal distribution of $f(u, v)$.

Proof. This is only a proof for discrete case. Let $p_{u|v}(u, v)$ be the conditional probability mass function of U given V and similarly define $p_{v|u}(v, u)$. The transition probability of the Markov chain is given by

$$p_{ij} = \text{PR}(U_{t+1} = j|U_t = i) = \sum_k p_{u|v}(j|k)p_{v|u}(k|i).$$

Let $g_u(u)$ and $g_v(v)$ be the marginal distributions of U and V. We have

$$\begin{aligned}
\sum_i g_u(i)p_{ij} &= \sum_i \big\{ \sum_k p_{u|v}(j|k)p_{v|u}(k|i)g_u(i)\big\} \\
&= \sum_k p_{u|v}(j|k)\big\{ \sum_i p_{v|u}(k|i)g_u(i)\big\} \\
&= \sum_k p_{u|v}(j|k)g_v(k) \\
&= g_u(j).
\end{aligned}$$

This implies that the distribution of U satisfies the relationship

$$\Pi = \Pi\mathbf{P}$$

for the discrete Markov chain. □

Since the limiting distribution of U_t is $g_u(\cdot)$ and the conditional distribution of V_t is $p_{v|u}(\cdot)$. It is immediately clear that the marginal distribution of V_t in the limit is $g_v(v)$. Their joint limiting distribution is $f(u, v) = p_{v|u}(v|u)g_u(u)$ as desired.

There are clearly many other problems with the use of Gibbs sampling. Not an expertise myself, it is best for me to not say too much here.

10.7 Relevance to Bayes analysis

As we pointed out, the basis of Bayes data analysis is the posterior distribution of the model parameters. However, we often only have the analytical form of the posterior distribution up to a multiplicative constant. It is seen that in Metropolis sampling algorithm, this is all we need to generate random numbers from such distributions.

In the case of Gibbs samplers, the idea can be extended. Suppose $U = (U_1, U_2, \ldots, U_k)$ and we wish to obtain samples whose marginal distribution is that of U. Let U_{-i} be subvector of U with U_i removed. Suppose it is efficient to generate data from the conditional distribution of U_i given U_{-i} for all i. Then one may iteratively generate U_i to obtain sample from the distribution f U using Gibbs samplers.

10.8 See you next term

You are welcome to Stat461/561 next term. We will cover some basics such as hypothesis test and confidence interval. The rest of time, if any, will be used on selective topics that you are interested and I am capable to handle.

Chapter 11

More on asymptotic theory

Various approaches to point estimation has been discussed so far. An estimator is recommended when it has certain desirable properties. Among many things, we like to know its bias and variance which can be derived from its sampling distribution. Characterizing exact sampling distributions is difficult in most cases. Fortunately, in most cases, an estimator based on a large number of observations has a limiting distribution when the sample size increases. The limiting distribution approximate the finite sample distribution and enables us to make further inferences. In this chapter, we provide additional discussions on asymptotic theories.

11.1 Modes of convergence

Let $X, X_1, X_2, \dots$ be a sequence of random variables defined on some probability space $(\Omega, \mathbb{B}, P)$.

Definition 11.1. *We say $\{X_n\}_{n=1}^{\infty}$ or simply X_n converges in probability to random variable X, if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \text{PR}(|X_n - X| > \epsilon) = 0.$$

We use notation $X_n \xrightarrow{p} X$.

Here is an example in which the convergence in probability can be directly verified.

Example 11.1. *Let $Y_1, Y_2, \ldots,$ be a sequence of i.i.d. random variables each has exponential distribution with rate $\lambda > 0$. Let*

$$X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}.$$

Then $X_{(1)} \xrightarrow{p} 0$.

Proof: Here 0 is considered as a random variable which takes value 0 with probability 1. Note that for every $\epsilon > 0$,

$$\begin{aligned} \text{PR}(|X_{(1)} - 0| > \epsilon) &= \text{PR}(X_{(1)} > \epsilon) \\ &= \text{PR}(X_1 > \epsilon, \ldots, X_n > \epsilon) \\ &= \text{PR}(X_1 > \epsilon) \cdots P(X_n > \epsilon) \\ &= \exp(-n\lambda\epsilon) \to 0 \end{aligned}$$

as $n \to 0$. Hence, by Definition 11.1, $X_{(1)} \xrightarrow{p} 0$. □

Definition 11.2. *We say X_n converges to X almost surely (or with probability 1) if and only if*

$$P\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1.$$

We use notation $X_n \xrightarrow{a.s.} X$.

Here is a quick example for the mode of almost sure convergence.

Example 11.2. *Let Y be a random variable and let $X_n = n^{-1}Y$ for $n = 1, 2, \ldots$. For any sample point $\omega \in \Omega$, as $n \to \infty$, we have*

$$X_n(\omega) = n^{-1}Y(\omega) \to 0.$$

Hence,

$$\text{PR}(\omega : \lim X_n(\omega) = 0) = 1.$$

Therefore $X_n \to 0$ almost surely.

It is natural to ask whether the two modes of convergence defined so far are equivalent. The following example explains that the convergence in probability does not imply the almost sure convergence. The construction is somewhat involved. Please do not spend a lot of time on it.

Example 11.3. *Consider a probability space $(\Omega, \mathbb{B}, P)$ where $\Omega = [0,1]$, $\mathbb{B}$ is the usual Borel σ-algebra, and the probability measure* PR *is the Lesbesgue measure. For any event $A \in \mathbb{B}$, $\mathbb{1}(A)$ is an indicator random variable. Define, for $k = 1, 2, \ldots, 2^n$ and $n = 1, 2, \ldots$,*

$$X_{2^{n-1}+k} = \mathbb{1}([\frac{k-1}{2^n}, \frac{k}{2^n}]).$$

Since any positive integer m can be uniquely written as $2^{n-1} + k$ for some n and k between 0 and $2^{n-1} - 1$, we have well defined X_m for all positive integer m.

On one hand, for every $\epsilon > 0$, it is seen that

$$\text{PR}(|X_m - 0| > \epsilon) \leq 2^{-n} \to 0.$$

Hence, $X_m \xrightarrow{p} 0$.

On the other hand, for each$\omega \in \Omega$ and any given n, there is an k such that

$$\frac{k-1}{2^n} \leq \omega < \frac{k}{2^n}.$$

Hence, no matter how large N is, we can always find an $m = 2^{n-1} + k > N$ for which $X_m(\omega) = 1$, and $X_{m+1}(\omega) = 0$. Therefore, $X_m(\omega)$ does not have a limit. This claim is true for any sample point in Ω. Hence, X_m does not almost surely converge to anything. □

The following theorem shows that the mode of almost sure convergence is a stronger mode of convergence.

Theorem 11.1. *If X_n converges almost surely to X, then $X_n \xrightarrow{p} X$.*

Let B_n, $n = 1, 2, \ldots$ be a sequence of events. That is, they are subsets of sample space Ω and members of $\mathbb{B}$. If a sample point belongs to infinite many B_n, for example it belongs to all B_{2n}, we say it occurs infinitely often. The subset which consists of sample points that occur infinitely often is denoted as

$$\{B_n \ i.o.\} = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} B_i.$$

Theorem 11.2 (Borel-Cantelli Lemma). *1. Let $\{B_n\}$ be a sequence of events. Then*

$$\sum_{i=1}^{\infty} \mathrm{PR}(B_n) < \infty$$

implies

$$\mathrm{PR}(\{B_n \text{ i.o.}\}) = 0;$$

2. If B_n, $n = 1, 2, \ldots$ are mutually independent, then

$$\sum_{i=1}^{\infty} \mathrm{PR}(B_n) = \infty$$

implies

$$\mathrm{PR}(\{B_n \text{ i.o.}\}) = 1.$$

The proof of this lemma relies on the expression $\{B_n \text{ i.o.}\} = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} B_i$. We now introduce other modes of convergence.

11.2 Convergence in distribution

The convergence in distribution is usually discussed together with the modes of convergence for a sequence of random variables. Although they are connected, convergence in distribution is very different from other modes of convergence in nature.

Definition 11.3. *Let $G_1, G_2, \ldots,$ be a sequence of (univariate) cumulative distribution functions. Let G be another cumulative distribution function. We say G_n converges to G in distribution, denoted as $G_n \xrightarrow{d} G$ if*

$$\lim_{n \to \infty} G_n(x) = G(x)$$

for all points x at which $G(x)$ is continuous.

This definite is not based on a sequence of random variables. If there is a sequence of random variables $X_1, X_2, \ldots$ and X whose distributions are given by $G_1, G_2, \ldots$ and G, we also say that $X_n \xrightarrow{d} X$. These random variables may not be defined on the same probability space. When we state that $X_n \xrightarrow{d} X$, it means that the distributions of X_n converges to the distribution of X as $n \to \infty$.

Theorem 11.3. *If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.*

Suppose c is a non-random constant. If $X_n \xrightarrow{d} c$, then $X_n \xrightarrow{p} c$.

A probability space is generally irrelevant to the convergence in distribution. Yet we can create a shadow probability space for the corresponding random variables.

Theorem 11.4 (Skorokhod's representation theorem). *If $G_n \xrightarrow{d} G$, then there exists a probability space $(\Omega, \mathbb{B}, P)$ and random variables $Y_1, Y_2, \ldots$ and Y, such that*

1. *Y_n has distribution G_n for $n = 1, 2, \ldots$ and Y has distribution G.*
2. *$Y_n \xrightarrow{a.s.} Y$.*

The following result is intuitive right but hard to prove unless the above theorem is applied.

Example 11.4. *If $X_n \xrightarrow{d} X$ and g is a real, continuous function, then $g(X_n) \xrightarrow{d} g(X)$.*

This is a simple exercise problem. There is an equivalent definition of the mode of convergence in distribution. We state here as a theorem.

Theorem 11.5. *Let $X_1, X_2, \ldots$ be a sequence of random variables. Then, $X_n \xrightarrow{d} X$ if and only if $\mathbb{E}\{g(X_n)\} \to \mathbb{E}\{g(X)\}$ for all bounded, uniformly continuous real valued function g.*

11.3 Stochastic Orders

Random variables come with different sizes. When a number of random variable sequences are involved in a problem, it is helpful to know their relative sizes. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. If $X_n \xrightarrow{p} 0$, we say $X_n = o_p(1)$. That is, compared with constant 1, the size of X_n becomes less and less noticeable. Naturally, we may also want to compare X_n with other sequences of numbers.

Definition 11.4. *Let $\{a_n\}$ be a sequence of positive constants. We say $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$ as $n \to \infty$.*

Let $\{Y_n\}_{n=1}^{\infty}$ be another sequence of random variables. We say $X_n = o_p(Y_n)$ if and only if

$$X_n/Y_n = o_p(1).$$

How do we describe that X_n and a_n are about the same magnitude? Intuitively, this should be the case when $\frac{X_n}{a_n}$ stays clear from both 0 and infinity. In common practice, we only exclude the latter. A rigorous mathematical definition is as follows:

Definition 11.5. *We say $X_n = O_p(a_n)$ if and only if for every $\epsilon > 0$, there exist M_ϵ such that for all n,*

$$\text{PR}(|X_n/a_n| \geq M_\epsilon) < \epsilon.$$

Note that $X_n = O_p(a_n)$ only reveals that $|X_n|$ is not larger compared with a_n. The size of $|X_n|$ can, however, be much smaller than a_n.

Example 11.5. *Assume $X_1, X_2, \ldots$ is a sequence of i.i.d. Poisson random variables. Then*

$$\max\{X_1, X_2, \ldots, X_n\} = O_p(\log n).$$

This is a nice exercise.

11.3.1 Application of stochastic orders

Stochastic order enables us to ignore irrelevant details above X_n and Y_n in asymptotic derivations. Some useful facts are as follows.

Lemma 11.1.

1. *If $X_n = O_p(1)$ and $Y_n = o_p(1)$, then $-X_n = O_p(1)$, $-Y_n = o_p(1)$.*

2. *If $X_n = O_p(1)$ and $Y_n = O_p(1)$, then $X_nY_n = O_p(1)$, $X_n + Y_n = O_p(1)$.*

3. *If $X_n = o_p(1)$ and $Y_n = o_p(1)$, then $X_nY_n = o_p(1)$, $X_n + Y_n = o_p(1)$.*

4. *If $X_n = o_p(1)$ and $Y_n = O_p(1)$, then $X_nY_n = o_p(1)$, $X_n + Y_n = O_p(1)$.*

If X_n converges to X in distribution and Y_n differs from X_n by a random amount of size $o_p(1)$, we expect that Y_n also converges to X in distribution. This is a building block to for more complex approximation theorems.

Lemma 11.2. *Assume $X_n \xrightarrow{d} X$ and $Y_n = X_n + o_p(1)$. Then $Y_n \xrightarrow{d} X$.*

Proof: Let x be a continuous point of the c.d.f. of X. Let $\epsilon > 0$ such that $x + \epsilon$ is also a continuous point of the c.d.f. of X. Then

$$\begin{aligned} \text{PR}(Y_n \le x) &= \text{PR}(Y_n \le x, |Y_n - X_n| \le \epsilon) + \text{PR}(|Y_n - X_n| > \epsilon, Y_n < x) \\ &\le \text{PR}(X_n \le x + \epsilon) + \text{PR}(|Y_n - X_n| > \epsilon) \\ &\to \text{PR}(X \le x + \epsilon). \end{aligned}$$

The second term goes to zero because $Y_n - X_n = o_p(1)$.

For any given x, ϵ can be chosen arbitrarily small due to the property of the monotonicity of distribution functions. Thus we must have

$$\limsup_{n\to\infty} \text{PR}(Y_n \le x) \le \text{PR}(X \le x).$$

Similarly, we can show

$$\liminf_{n\to\infty} \text{PR}(Y_n \le x) \ge \text{PR}(X \le x).$$

The two inequalities together imply

$$\text{PR}(Y_n \le x) \to \text{PR}(X \le x)$$

for all x at which the c.d.f. of X is continuous. Hence $Y_n \xrightarrow{d} Y$. □

The above result makes the next lemma obvious.

Lemma 11.3. *If $a_n \to a$, $b_n \to b$, and $X_n \xrightarrow{d} X$, then $a_nX_n + b_n \xrightarrow{d} aX + b$. If $Y_n \xrightarrow{p} a$ and $Z_n \xrightarrow{p} b$, and $X_n \xrightarrow{d} X$, then $Y_nX_n + Z_n \xrightarrow{d} aX + b$.*

The following well-known theorem becomes a simple implication.

Theorem 11.6 (Slutsky's Theorem). *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ where c is a finite constant. Then*

1. *$X_n + Y_n \xrightarrow{d} X + c$;*

2. $X_nY_n \xrightarrow{d} cX$*;*

3. $X_n/Y_n \xrightarrow{d} X_n/c$ *when* $c \neq 0$*.*

Here is another theorem that is convenient.

Theorem 11.7. *Let* a_n *be a sequence of real values and* X_n *be a sequence of random variables. Suppose* $a_n \to \infty$ *and* $a_n(X_n - \mu) \xrightarrow{d} Y$*. If* $g(x)$ *is a function which has continuous derivative at* $x = \mu$*, then*

$$a_n\{g(X_n) - g(\mu)\} \xrightarrow{d} g'(\mu)Y.$$

The most useful result for convergence in distribution is the central limit theorem.

Theorem 11.8 (Central Limit Theorem). *Assume* $X_1, X_2, \dots$ *are i.i.d. . random variables with* $\mathbb{E}(X) = 0$ *and* $\text{VAR}(X) = 1$*. Then as* $n \to \infty$*,*

$$\sqrt{n}\bar{X}_n \xrightarrow{d} N(0, 1).$$

If, instead, $\mathbb{E}(X) = \mu$ *and* $\text{VAR}(X) = \sigma^2$*, then*

1. $\sqrt{n}\sigma^{-1}(\bar{X}_n - \mu) \xrightarrow{d} N(0, 1)$*;*

2. $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$*;*

3. $n^{-1/2}\sum_{i=1}^{n}\{(X_i - \mu)/\sigma\} \xrightarrow{d} N(0, 1)$*;*

4. $n^{-1/2}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow{d} N(0, \sigma^2)$*.*

It is not advised to state

$$\bar{X}_n - \mu \xrightarrow{d} N(0, \sigma^2/n).$$

The righthand side is not a limit at all.

Example 11.6. *Let* X_n, Y_n*, be a pair of independent Poisson distributed random variables with mean* $n\lambda_1$ *and* $n\lambda_2$*. Define*

$$T_n = (Y_n/X_n)\mathbb{1}(X_n > 0).$$

Then T_n *is asymptotically normal.*

Chapter 12

Hypothesis test

Recall again that a statistics model is a family of distributions. When they are parameterized, the model is parametric. Otherwise, the model is non-parametric. One may notice that the regression models are not exceptions to this definition. Suppose a random sample from a distribution F is obtained/observed. A statistical model assumption is to specify a distribution family $\mathcal{F}$ such that F is believed to be a member of it.

Often, we are interested in a special subfamily $\mathcal{F}_0$ of $\mathcal{F}$. The statistical problem is to decide whether or not F is a member of $\mathcal{F}_0$ based on a random sample from this unknown F. There might be situations where the question can be answered with certainty. Most often, statistics are used to quantify the strength of the evidence against $\mathcal{F}_0$ from chosen angles. Hypothesis test is an approach which recommends whether or not $\mathcal{F}_0$ should be rejected. It also implicitly recommends a distribution in the complement of $\mathcal{F}_0$ if $\mathcal{F}_0$ is rejected. We consider $\mathcal{F}_0$ as null hypothesis and also denote it as H_0. Its complement in $\mathcal{F}$ forms alternative hypothesis and is denoted as H_a or H_1.

The specification of $\mathcal{F}$ is based on our knowledge on the subject matter and the property of probability distributions. For instance, a binomial distribution family is used when the number of passengers show up for a specific flight, the number of students show up for a class and so on. The choice of $\mathcal{F}_0$ often relates to the background of the application. We provide a number of scenarios in the next section.

12.1 Null hypothesis.

Where is $\mathcal{F}_0$ from? The question is more complicated than we may believe. Here are some examples motivated from various classical books.

(a) The null hypothesis may correspond to the prediction out of some scientific curiosity. One wishes to use data to examine its validity.

We suspect that the sex ratio of new babies is 50%. In this case, one may collect data to critically examine how well this belief approximates the real world.

(b) In genetics, when two genes are located in two different chromosomes, their recombination rate is exactly $\theta = 0.5$ according to Mendel's law. Rejection of a null hypothesis of $\theta = 0.5$ based on experimental or observational data leads to meaningful scientific claims.

Scientists or geneticists in this and similar cases must bear the burden of proof. The null hypothesis stands on the opposite side of their convictions.

(c) Some statistical methods are developed under certain distributional assumptions on the data such as the analysis of variance. If the normality assumption is severely violated, the related statistical conclusions become dubious. A test of normality as the null hypothesis is often conducted. We are alarmed only if there is a serious departure from normality. Otherwise, we will go ahead to analyze the data under normality assumption.

(d) H_0 may assert complete absence of structure in some sense. So long as the data are consistent with H_0 it is not justified to claim that data provide clear evidence in favour of some particular kind of structure.

Does living near hydro power line make children more likely to have leukaemia? The null hypothesis would suggest the cases to be distributed geographically randomly.

(e) The quality of products from a production-line fluctuates randomly within some range over the time. One may set up a null hypothesis that the system is in normal status characterized by some key specific parameter values. The rejection of the null hypothesis sets off an alarm that the system is out of control.

(f) When a new medical treatment is developed, its superiority over the standard treatment must be established in order to be approved. Naturally, we will set the null hypothesis to be "there is no difference between two treatments".

(g) There are situations where we wish to show a new medicine is not inferior than the existing one. This is often motivated by the desire to produce a specific medicine at a lower cost. One needs to be careful to think about what the null hypothesis should be here.

(i) In linear regression models, we are often interested to test whether a regression coefficient has a value differs from zero. We put zero-value as the null hypothesis. Rejection of which implies the corresponding explanatory has no-nil influence on the response value.

In all examples, we do not reject H_0 unless the evidence against it is mounting. Often, H_0 is not rejected not because it holds true perfectly, but because the data set does not contain sufficient information, or the departure is too mild to matter in scientific sense, or the departure from H_0 is not in the direction of concern. It is hard to distinguish these causes. We will come to this issue again after introduction of the alternative hypothesis.

12.2 Alternative hypothesis

.

In the last section, we discussed the motivation of choosing a subset $\mathcal{F}_0$ of $\mathcal{F}$ to form H_0. It is naturally to form the alternative hypothesis H_a or H_1 as the remaining distributions in $\mathcal{F}$. If so, the alternative hypothesis is heavily dependent on our choice of $\mathcal{F}$. Since any data set is extreme in some respects, severe departure from $\mathcal{F}_0$ can always be established. Thus, it can be meaningless to ask absolutely whether $\mathcal{F}_0$ is true, by allowing $\mathcal{F}$ to contain all imaginable distributions. The question becomes meaningful only when a proper alternative hypothesis is proposed.

The alternative hypothesis serves the purpose of specifying the direction of the departure the true model from the null hypothesis that we care! In the example when a new medicine is introduced, the ultimate goal is to show that it extends our lives. We put down a null hypothesis that the new medicine is not better than the existing one. The goal of the experiment and hence the statistical significance test is to show the contrary: the new medicine is better. Thus, the alternative hypothesis specified the direction of the departure we intend to detect.

In regression analysis, we may want to test the normality assumption on the error term to ensure the suitability of the least sum of squares approach. In this case, we often worry whether the true distribution has a heavier tail probability than the normal distribution. Thus, we want to detect departures toward "having a heavy tail". If the error distribution is not normal but uniform on a finite interval, for instance, we may not care at all. Therefore, if H_1 is not rejected based on a hypothesis test, we have not provided any evidence to claim H_0 is true. All we have shown is that the error distribution does not seem to have a heavy tail.

According to genetic theory, the recombination rate θ of two genes on the same chromosome is lower than 0.5. Hence, if the data lead to an observed very high recombination rate, we may have evidence to reject the null hypothesis of $\theta = 0.5$. However, it does not support the sometimes sacred genetic claim that two genes are linked. To establish linkage, $\mathcal{F}$ would be chosen as all binomial distributions with probability of success no more than 0.5.

In many social sciences, theories are developed in which the response of interest is related to some explanatory variable. When one can afford to collect a very large data set, such a connection is always confirmed by rejecting the null hypothesis that the correlation is nil. As long as the theory is not completely nonsense, a lower level of connection inevitably exists. When the data size is large, even a practically meaningless connection will be detected with statistical significance.

In summary, specifying alternative hypothesis is more than simply putting done the possible distributions of the data in addition to these included in the null already. It specifies the direction of the departure from the null model which we hope to detect or to declare its non-fitness. We generally investigate the hypothesis test problem under the assumption that the data are generated from a distribution inside H_0 and what happens if this distribution is a member of H_1. This practice is convenient for statistical research. We should not take it as truth in applications. It could happen that the data suggest the truth is not in H_0, H_1 is slightly a better choice, yet the truth is not in H_0 nor H_1. Hence, by rejecting H_0, the hypothesis test itself does not prove that H_1 contains the truth.

12.3 Pure significance test and p-value

Suppose a random sample $X = x$ is obtained from a distribution F_0 and the statistics model is $\mathcal{F}$. We hope to test the null hypothesis $H_0 : F_0 \in \mathcal{F}_0$. Let $T(x)$ be a statistic to be used for statistical significance test. Hence, we call it test statistic. Ideally, it is chosen to has two desirable properties:

(a) the specific sample distribution of T when H_0 is true is known (not merely up to a distribution family but a specific distribution) at least approximately. If H_0 contains many distributions, this property implies that the sample distribution of T remains the same whichever distribution in $\mathcal{F}_0$ that X may have, or at least approximately. In other words, it is an auxiliary statistic under H_0.

(b) the larger the observed value of T, the stronger the evidence of departure from H_0, in the direction of H_1.

If a statistic has these two properties, we are justified to reject the null hypothesis when the realized value of T is large. Let $t_0 = T(x)$ be its realized/observed value and

$$p_0 = \text{PR}(T(X) \geq t_0; H_0)$$

which is the probability that $T(X)$ is larger than the observed value when the null hypothesis is true. When $P(T(X) = t_0; H_0) > 0$, a continuity correction may be applied. That is, we may revise the definition to

$$p_0 = P(T(X) > t_0; H_0) + 0.5P(T(X) = t_0; H_0).$$

In general, this is just a convention, not an issue of "correctness". The smaller the value of p_0, the stronger is the evidence that the null hypothesis is false. We call p_0 the p-value of the significance test.

Remark: the definition of p-value is most sensible when a test statistic has been introduced and it has the above two desired properties. With known-distribution assumption, $\text{PR}(T(X) \geq t_0; H_0)$ does not have an definite answer. Without the other property, we are not justified to be exclusively concerned on the choice of $T(X) \geq t_0$, rather than other possible values of $T(X)$.

If T is a test statistic with properties (a) and (b), and that g is a monotone strictly increasing function, the $g(T)$ makes an another test statistic, and the p-value based on $g(T)$ will be the same as the p-value based on T.

Since there is no standard choice of $T(x)$, there is not a definite p-value for a specific pair of hypothesis even if the test statistic $T(x)$ has these two properties. Because of this, the definition of p-value has been illusive in many books.

Assume issues mentioned above have been fixed. If magically, $p_0 = 0$, then H_0 cannot be true or something impossible would have been observed. When p_0 is very small, then either we have observed an unlikely event under H_0, or the rare event is much better explained by a distribution in H_1. Hence, we are justified to reject H_0 in favour of H_1. Take notice that a larger $T(x)$ value is more likely if the distribution F is a member of H_1.

How small p_0 should be in order for us to reject H_0. A statistical practice is to set up a standard, say 5%, so we commonly reject H_0 when $p_0 < 5\%$. The choice of 5% is merely a convention. There is no scientific truth behind this magic cut-off point. There is a joke related to this number: scientists tell their students that 5% is found to be optimal by statisticians, and statisticians tell their students that the 5% is chosen based on some scientific principles. Incidentally, the Federal Food and Drug administration in the United States uses 5% as its golden standard. If a new medicine beats the existing one by a pre-specified margin, and it is demonstrated by significance test at 5%

level, then the new medicine will be approved. Of course, we assume that all other requirements have been met. Most research journals accept results established via statistical significance test at 5% level. You will pretty soon be under pressure to find a statistical method that results in a p-value smaller than 5% for a scientist.

Not all test statistics we recommend have both properties (a) and (b). There are practical reasons behind the use of statistics without these properties. When their usage leads to controversies, it is helpful to review the reasons why properties (a) and (b) are desirable and interpret the data analysis outcomes accordingly.

12.4 Issues related to p-value

After one has seen the data, he can easily find the data are extreme in some way. One may select a null hypothesis accordingly and most likely, the p-value will be small enough to declare significance. This problem is well–known but hard to prevent. After you have seen the final exam results of stat460/560, you may compare the average marks between under and graduate students, between male and female students, foreign and domestic students, younger and older students and many more ways. If 5% standard on p-value is applied to each test, pretty soon we will find one that is significant. This is statistically invalid. To find one out of 20 tests with its p-value below 5% is much more likely than to find a p-value below 5% of a pre-decided test.

A pharmaceutical company must provide a detailed protocol before a clinical trial is carried out. If the data fail to reject the null hypothesis, but point to an other meaningful phenomenon, the FDR will not accept the result based on analysis if the current data. They must conduct another clinical trial to establish the new claim. For example, if they try to show that eating carrots reduces the rate of stomach cancer, yet the data collected imply a reduction in the rate of liver cancer, the conclusion will not be accepted. One could have examined the rates of a thousand cancers: liver cancer happened to produce a low p-value. By this standard, Columbus did not discover America because he did not put discovering America into his protocol. Rather, he aimed to find a short cut to India.

Another issue is the difference between **Statistical significance and the Scientific significance**. Consider a problem in lottery business, each ball, numbered from 1 to 49, should be equally likely to be selected. Suppose I claim that the odd numbers are more likely to be sampled than the even numbers. The rightful probability of a odd ball is selected should be $p = 25/49$. In the real world, nothing is perfect. Assume that the truth is $p = 25/49 + 10^{-6}$. It is not hard to show that if we conduct 10^{24} trials, the chance that the null hypothesis $p = 25/49$ being rejected is practically 1, at 5% level or any reasonable level based on a reasonable test. Yet such a statistical significant result is nonsensical to a lottery company. They need not be alarmed unless the departure from $p = 25/49$ is more than 10^{-3}, presumably. In a more practical example, if a drug extends the average life expectancy by one-day, it is not significant no matter how small the p-value of the significance test is.

There are abundant discussions on the usefulness of p-value. There has been suggestions of not teaching the concept of the p-value which I beg to differ. The key is to make everyone understand what it presents, rather than frantically searching for a test (analysis) that gives a p-value smaller than 0.05.

Here is an example suggested by students. It is not as meaningful to be 100% sure that someone stole 10 dollars from a store. It is a serious claim if we are 50% sure that someone killed the store owner.

In regression analysis, a regression coefficient is often declared highly significant. It generally refers to a very small p-value is obtained when testing for its value being zero. This is unfortunate: the regression coefficient may be scientifically indifferent from zero, but its effect is magnified by a microscope created by a big data set.

12.5 General notion of statistical significance test

Suppose a random sample of X from $\mathcal{F}$ is taken. The null hypothesis H_0 as a subset of $\mathcal{F}$ is specified and H_1 is made of the rest of distributions in $\mathcal{F}$. No matter how a test statistic is constructed, in the end, one divides the range of X into two, potentially three non-overlap regions: C and its complement C^c. We will come back to the potential third region.

The procedure of the significance test then rejects H_0 when the observed value of X, $x \in C$. Thus, C is called the critical region. When $x \notin C$, we retain the null hypothesis. However, I do not advocate the terminology of "**Accept** H_0". Such a statement can be misleading. When we fail to prove an accused guilty, it does not imply its innocence.

Once C is given, we define

$$\alpha = \sup_{F \in H_0} \text{PR}(X \in C; F)$$

as the size of the test. When the true distribution $F \in H_0$ yet $x \in C$ occurs, the null hypothesis H_0 is erroneously rejected. The probability $\text{PR}(X \in C)$ is called **Type I** error. Type I error is not the same as the size of the test because H_0 may contain many distributions. The size of a test is determined by the "least favourable distribution" which is the one that maximizes the probability of $X \in C$. Under simple models, it is easy to identify such a least favourable distribution. In a general context, we have long given up the effort of doing so.

If $x \notin C$ yet $F \in H_1$, we fail to reject H_0, the corresponding probability is called **Type II** error. For each distribution $F \in H_1$, we call

$$\text{PR}(X \in C; F)$$

the power function of F on H_1. If $\mathcal{F}$ is a parametric model with parameter θ, it makes sense to rewrite it as

$$\gamma_C(\theta) = \text{PR}(X \in C; \theta), \quad \theta \in H_1.$$

The type II error is also a function of θ: $\beta(\theta) = 1 - \gamma(\theta)$.

We do not usually discuss the situation where $F \notin \mathcal{F}$. If this happens, a "third type" of error has occurred. One should take this possibility into serious consideration in real world applications. It will not be discussed further here.

Example 12.1. *(One-sample t-test). Assume we have a random sample from $\mathcal{F} = \{N(\theta, \sigma^2)\}$ distribution. We test the null hypothesis $H_0 : \theta = 0$.*

Let

$$T(x) = \frac{\sqrt{n}\bar{x}}{s}$$

where $\bar{x} = n^{-1}(x_1 + x_2 + \cdots + x_n)$ is the realized value of $\bar{X}$ and s^2 is the realized value of the sample variance. It is seen that $T(X)$ has t-distribution regardless of which distribution in H_0 is the true distribution of X. Thus, it has property (a). At the same time, the larger is the value of $|T|$, the more obvious that the null hypothesis is inconsistent with the data. Thus, $|T|$ also has property (b). In other words, $|T|$ rather than T makes a desirable test statistic.

Let $t_{0.975,n-1}$ be the 97.5% quantile of the t-distribution with $n-1$ degrees of freedom. We may put

$$C = \{x : |T(x)| \geq t_{0.975,n-1}\}$$

as the critical region of a test. If so, its size is

$$\alpha = \text{PR}(|T(X)| \geq t_{0.975,n-1}; H_0) = 0.05.$$

It is less convenient to write down its power function.

The p-value of this test is

$$p_0 = \text{PR}(|T(X)| \geq T(x); H_0)$$

where $T(x)$ is the realized value of T. Rejecting H_0 whenever $p_0 < 0.05$ is equivalent to rejecting H_0 whenever $x \in C$. Providing a p-value has added benefit: we know whether H_0 is rejected with barely sufficient evidence or very strong evidence.

Again, p-value should be read with a pinch of salt. Even if the true θ-value is only slightly different from 0, the evidence against H_0 can be made very strong with a large sample size n. Hence, small p-value shows how strong the evidence is against H_0, it does not necessarily indicate H_0 is an extremely poor model for the data.

To avoid the dilemma implied by overly relying on small p-value, it might be better to specify H_1 as $|\theta| > 0.1$ and put H_0 as $|\theta| < 0.1$ instead. We have placed an arbitrary value 0.1 here, it is not hard to come up with a sensible small value in a real world application.

12.6 Randomized test

Particularly in theoretical development, we often hope to construct a test with exactly the pre-given size. The above approach may not be feasible in some circumstances.

Example 12.2. *Suppose we observe X from a binomial model with $n = 2$ and the probability of success $\theta \in (0, 1)$. Let the desired size of the test be $\alpha = 0.05$ for the null hypothesis $\theta = 0.5$. In this case, we have only 8 candidates for the critical region C. None of them result in a test of the exact size $\alpha = 0.05$.*

An artificial approach to find a test with the pre-specified size is as follows. We do not reject H_0 if $X = 1$. When $X = 0, 2$, we toss a biased coin and reject H_0 when the outcome is a head. By selecting a coin such that PR(Head) = 0.1, the probability of rejecting H_0 based on this approach is exactly 0.05 when $\theta = 0.5$. Thus, we have artificially attained the required size 0.05.

The region $\{0, 2\}$ is the third region in the range of X mentioned previously.

Abstractly, a statistical significance test is represented as a function $\phi(x)$ such that $0 \leq \phi(x) \leq 1$. We reject H_0 with probability $\phi(x)$ when $X = x$. When $\phi(x) = 0$ or 1 only, the sample space is neatly divided into the critical region and its complement. Otherwise, the region of $0 < \phi(x) < 1$ is a randomization region. When x falls into that region, we randomize the decision.

Defining a significance test by a function $\phi(x)$ is mathematically convenient. Note that its size

$$\alpha = \sup_{F \in H_0} \mathbb{E}\{\phi(X); F\}$$

and its power function on $F \in H_1$ is given by

$$\gamma(F) = \mathbb{E}\{\phi(X); F\}.$$

The type I error is defined for $F \in H_0$ and given by

$$\alpha(F) = \mathbb{E}\{\phi(X); F\}.$$

We do not place many restrictions on $\phi(x)$ to use it as a test function. Instead, we ask when $\phi(x)$ is a good test. This question leads to the call for optimality definitions. We will come to this issue later.

12.7 Three ways to characterize a test

Discussions in previous section have presented three hypothesis test procedures.

1. Define a test statistic, T, such that we reject H_0 when T is large. Preferably, T has two specific properties: known and same sample distribution under whichever distribution in H_0; larger observed value of T indicates more extreme departure of F from H_0 toward the direction we try to capture. We compute p-value as

   $$p = \text{PR}(T \geq t_{obs}; H_0)$$

   where t_{obs} is the observed value. When T has discrete distribution, we may apply a continuity correction

   $$p = \text{PR}(T > t_{obs}; H_0) + 0.5\text{PR}(T = t_{obs}; H_0).$$

   We reject H_0 if p is below some pre-decided level, usually 5%.

2. Define a critical region C in terms of the range of X. When the realized value $x \in C$, we reject H_0. The region C is often required to have a given size α:

   $$\sup_{H_0} \text{PR}(X \in C) = \alpha.$$

3. When X is discrete, we may get into situation where no critical region has a pre-specified size α. This is not problematic in applications, but is problematic for theoretical discussions. Hence, we define a test as a function $\phi(x)$ taking values between 0 and 1. We reject H_0 with probability $\phi(x)$ where x is the realized/observed value of X. The size of this test is calculated as $\sup_{H_0} \mathbb{E}\{\phi(X)\}$.

Method 1 is a special case of method 2 by letting $C = \{x : T(x) > k\}$ for some k. Both methods 1 and 2 can be regarded as special cases of method 3: by letting $\phi(x) = \mathbb{1}(x \in C)$. We reject H_0 with probability 1 when $x \in C$, and do not reject H_0 otherwise.

Clearly, a trivial test $\phi(x) = \alpha$ has size α. Its existence ensures that a test with any specific size between 0 and 1 is possible. The statistical issue is on finding one with good properties.