STAT 461/561 - STATISTICAL INFERENCE II
2014/15, TERM II

Jiahua Chen
Department of Statistics
University of British Columbia

Course Outline

Course description: Detailed development of the theory of testing hypotheses and confidence regions, Bayesian models and inference, elements of decision theory and additional topics. Any contemporary topics we come up with (e.g. Bootstrap, FDR, Lasso, Empirical likelihood). Intended for honours students and graduate students.

Pre-requisites: Stat460/Stat560.

Topics covered in year 2014/2015

1. General discussion: Discipline of Statistics, Probability and Statistics model, Statistics inference. Point estimation.

2. Statistical significance test: Null hypothesis, Alternative hypothesis, Pure significance test, General notion of statistical significance test.

3. Optimality discussions on hypothesis tests: Neyman-Pearson Lemma, Uniformly most powerful for one-sided alternative, Monotone likelihood ratio, Existence of UMPU tests, Locally most powerful test.

4. Likelihood based hypothesis test: Consistency of MLE for one-dimensional θ and as a local maximum, Likelihood ratio test, Score test, Wald test.

5. Inferences for data with normal distribution: One-sample problem, Test for equal variance, Test for equal mean under equal variance assumption.

6. Non-parametric test: One-sample sign test. Two-sample permutation test, Wilcoxon two-sample rank test, Kolmogorov-Smirnov and Cramér-von Mises tests.

7. Confidence intervals or confidence regions: Confidence interval via hypothesis test, Confidence interval via pivotal quantities, Likelihood intervals, Prediction intervals.

8. Empirical likelihood: Likelihood ratio function and profile likelihood, Numerical problem, Hypothesis test and confidence region, Adjusted empirical likelihood.

9. Resampling: Estimating Variance estimation, Estimating cumulative distribution function, Bootstrap Confidence Intervals.

10. Multiple comparison: Analysis of variance for one-way layout, The Bonferroni Method, Turkey Method, False discovery rate,

11. Variable selection/Model selection problem: Bayesian information criterion, Consistency of BIC, Extended BIC. (regularization methods such as Lasso and Scad will be added this year).

Assignments, Midterm and Final:

There will be one in-class midterm and one regular final exam.

We aim at giving 50 assignment problems for the whole semester (reductions for undergraduate students in Stat461). Due to the difficulty of controlling the progress of the lectures, the due dates of the assignment problems will be very flexible. Students are encouraged to hand in problems every Friday. Deadlines will be announced from time to time.

Please do not use pencil. Use regular lined papers. Use double space. Start a new page when you start a new problem. Skip two lines when you start a new part of a problem. Explain your steps to ensure that the TA and/or myself can understand your logic.

Marking will emphasize the logical flow in addition to the correctness. A smooth answer with generally correct answer is sufficient for a mark of 5. Correct answer alone worths a mark of 3. Illogical or ultimately fail to hand in assignment problems will loss all marks. The marks for lengthy questions will worth multiple '5 marks'. Students are encouraged to re-do the assignment problems if they fail to get at least 3 marks. The other side of the story is: we will refuse to mark a problem if it is not well presented.

TA will be instructed to provide as much comments as possible. Do ask the instructor if you do not understand his/her comments or do not agree. It can also be helpful to explain your solution to TA or myself face to face.

Final grade:

40% assignment + 40 % midterm + 40% final exam - 20% of the worst of midterm/final.

Contents

Chapter 1

Some basics

1.1 Discipline of Statistics

Statistics is a discipline that serves other scientific disciplines. Statistics is itself not considered by many as a branch of science. A scientific discipline constantly develops theories to explain the mechanisms behind the nature. These theories are falsified whenever their prediction contradicts the observations. Based on these theory and hypothesis, we may form a model for the natural world and the model is further utilized to predict what happens to the nature under new circumstances. Scientific experiments are constantly designed to contradict the prediction and aim at DISPROVING the hypothesis behind the model/theory. If a theory is able to make useful predictions and we fail to find contradicting evidences, the theory gains wide acceptance. We may then consider it temporarily 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 successful 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 it does not imply that the statistic theory is wrong. Rather, it implies that this 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 of 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 stated as laws. Unfortunately, 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 in disproving 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 model

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 sometimes give a precise prediction. The water boils at 100 degrees at the sea level. 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 there might be an unknown perfect scientific model which can explain away all randomness. In terms of earthquake, 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, 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 properties. We generally assume that it is possible to assign a probability to each subset of Ω that is a member of $\mathbb{B}$. 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 kind of 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 $\mathcal{F}$. 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 will be clear when we work with specific problems. In this case, we work with *population* F defined on $(\mathcal{R}, \mathbb{B})$. Components of X are samples from this population F, repeated.

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 often 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 system. 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? Thus, 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 Statistics 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 from $\mathcal{F}$. 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 select a submodel, the procedure is called model selection. In general, we do not quantify the evidence favouring the specific submodel. This is the difference between "hypothesis test" and "model selection".

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, $\bar{x}_n - E(X_1)$ is in general not a statistic because it is a function of both data, $\bar{x}_n$ and the usually unknown value $E(X_1)$, often depends on θ.

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. Otherwise, we may regard $T = T(X)$ as a quantity to be "realized". Since X is random, the final outcome of T is also random. The distribution of T is called its sample distribution. Unfortunately, it is often hard to be completely consistent when we deal with T 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 be a member of $\mathcal{F}$ which is often labeled by a parameter θ, the distribution of T is 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 any efficiency. Such a simplification is most useful if T is much simpler than X or it is a substantial reduction of X.

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

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 did not discuss this notion in Stat 460/560. We have a few more such notions.

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

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. statistics.*

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.

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 our 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.

Definition 1.6 Point estimation

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.

So 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.

Method of moments: If θ has dimension k and the first k moments of X exist, we can construct k equations forcing the first k moments of X equal to their corresponding sample moments. The solution in θ are called moment estimates of θ. Moment estimators are often easy to obtain and have simple distributional properties. One may use other types of moment relationships to construct equations in an attempt to get an estimate of θ. The corresponding estimators are just as sensible, but they are not called moment estimators.

Maximum likelihood estimator: If one can find a σ-finite measure such that all distributions in $\mathcal{F}$ has a density function $f(x)$. Then the likelihood function is 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

$$L(\hat{\theta}) = \sup_{\theta} f(x; \theta),$$

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.

Additional discussions given in the lecture:

The observation x includes the situation where it is a vector. The common i.i.d. situation is a special case. The probability mass function, when x is discrete, is also regarded as a density function. This looks after discrete models. In general, the likelihood function is defined as follows.

Definition 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 random distribution, the probability is computed as the probability of the even "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.* There are situations where "joint density" is not the best approach to define the likelihood function.

In industry, it is vital to ensure that the 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 components. When the average survival time is very long, one has to terminate the experiment at some point. Let a life time of a component be X and the termination time be T. Then, the observation is censored and we only observe $\min(X, T)$. This type of censor 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. *a type θ^{-n} is corrected to θ^{-m}.*

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 them 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 $L_n(\theta)$.

Chapter 2

Hypothesis test

Suppose a random sample from a distribution F has been obtained. In addition, F is believed to be a member of distribution family $\mathcal{F}$. Let $\mathcal{F}_0$ be a subset of $\mathcal{F}$. A statistical problem is to decide whether or not F is a member of this special subset of the distribution family.

We generally do not have direct information about the true distribution F but a random sample from it. There can be situations where the question can be answered without uncertainty. Most often, statistics is used provide some quantified evidence for or against $\mathcal{F}_0$ from various angles. Hypothesis test is an approach which recommends whether or not $\mathcal{F}_0$ should be rejected. We consider $\mathcal{F}_0$ as null hypothesis and also denote it as H_0.

2.1 Null hypothesis.

Where is $\mathcal{F}_0$ from? The following discussions are modified from Cox and Hinkley.

(a) H_0 may correspond to the prediction of some scientific theory or some model of the system though quite likely to be true or nearly so.

An example might be: it might be believed that the sex of a new baby has 50% chance to be a boy. The sexes of babies are independent of each other. Thus, a null hypothesis H_0 claiming the number of boys in a family is binomially distributed fits well into this category.

(b) H_0 divides the possible distribution into two qualitatively different

types. If the data are reasonably consistently with H_0, it is not possible nor necessary to establish which is the true type.

An example might be normality check in the analysis of variance. We are alarmed only if there is a serious departure from normality. Otherwise, we are happy enough to analyze the data under normality assumption.

(c) 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.

Are there evidences that the quality of wood product deteriorated this year compared to the previous years? Do the new batch of a product attain the quality standard?

(d) This is my suggestion. H_0 often represents a specific scientific claim. Rejection of the null hypothesis is in support of the correctness of a specific claim.

When a new treatment is developed, one would like to know if it truly provides some benefits to patients. The null hypothesis would be "no benefit" in this context.

In genomic studies, it is often interest to identify genetic variations connected to diseases. Rejection of no-connection hypothesis implies the identification of a likely culprit of a disease.

My understanding on the difference between (a) and (d): In (a), we are interested in truthfulness of H_0 itself; in (d), we are interested in whether a better theory is likely.

2.2 Alternative hypothesis

. The set of distributions that belong to the null hypothesis is denoted as H_0. The remaining distributions in $\mathcal{F}$ are denoted as H_a or H_1 which is called the alternative hypothesis. The specification of H_1 is also very important. 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, unless 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 hope to detect.

In regression analysis, we may want to test the normality assumption to ensure the suitability of the least sum of squares approach. In this case, we often worry whether the true distribution has 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, for instance, we may not care at all.

Once more, specifying alternative hypothesis is more than simply putting done what are the possible models 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-existence.

2.3 Pure significance test and p-value

Suppose a random sample $X = x$ is obtained from a probability model $\mathcal{F}$. We hope to test the null hypothesis $H_0 : F_0 \in \mathcal{F}_0$ where F_0 is the "true" distribution of X.

Let $T(x)$ be a statistic to be used for statistical significance test. Hence, we call it test statistic. Ideally, it has two desirable properties:

(a) the sample distribution of T when H_0 is true is known, at least approximately. If H_0 contains many distributions, we hope that the sample distribution of T remains the same which distribution in $\mathcal{F}_0$ X has, or at least approximately.

(b) the larger the observed value of T, the stronger the evidence of departure from H_0, in the direction of H_a.

If a statistic has these two properties, we are justified to reject the null

hypothesis when the realized value of T is large. Let $t_0 = T(x)$ and

$$p_0 = P(T(X) \geq t_0; H_0)$$

which is the probability that $T(X)$ is larger than the observed value when the null hypothesis is true. 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. As long as $p_0 > 0$, we are not completely sure that the H_0 is false. Having $p_0 = 0$, however, is impossible in most experiments involve randomness. A statistical practice is to set up a standard, say 5%, so that the H_0 is rejected when $p_0 < 5\%$.

When $P(T(X) = t_0; H_0) > 0$, a continuity correction may be used for p-value calculation. That is, we may revise the definition to define

$$p_0 = P(T(X) > t_0; H_0) + 0.5P(T(X) = t_0; H_0).$$

In general, this is just a convention. It is an issue of "correctness".

The choice of 5% is a convention in most applications. 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 highest scientific principle. Incidentally, the Federal Food and Drug administration in the United States use 5% as their golden standard. If a new medicine beats the existing one by a pre-specified margin, and it is demonstrated by statistical 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 under pressure to find a statistical method that results in a p-value smaller than 5% for a scientist.

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.

2.4 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 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 p-value below 5% is much more likely than to find a p-value of pre-determined test below 5%.

For a pharmaceutical company, they must provide a detailed protocol before a clinical trial is carried out. If the data fail to reject the null hypothesis, but point to some other direction, the FDR will not accept their analysis. They must conduct another clinical trial to establish their case. For example, if they try to show that eating carrots reduces the rate of stomach cancer, yet the data collected imply a reduction in rate of liver cancer, the conclusion will not be accepted. One could have examined the rates of thousands kinds of 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 odd numbers are more likely to be sampled than even numbers. The rightful probability of odd balls being selected should be $p = 25/49$. In real world, there is nothing perfect. Let us assume that the truth is $p = 25/49 + 10^{-6}$. It is not hard to show that if we collect data of 10^{24} number of trials, the chance that the null hypothesis $p = 25/49$ being rejected is practically 1, at 5% level or any reasonable level. Yet such a statistical significant result is nonsense to a lottery company. They should not be alarmed unless the departure from $p = 25/49$ is more than 10^{-3}, I guess. 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 discussion on the usefulness of p-value. There has been suggestions of not teaching the concept of p-value. I personally feel

that it is a useful concept. 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.

2.5 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 $\mathcal{F}$. Now 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 C^c which is the complement of C. I 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 the innocence.

Once C is given, we define

$$\alpha = \sup_{F \in H_0} P(X \in C; F)$$

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 $P(X \in C)$ is called **Type I** error. It is not completely the same as the size of the test because H_0 may contain many distributions. 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

$$P(X \in C; F)$$

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

$$\gamma_C(\theta) = P(X \in C; \theta), \quad \theta \in H_1.$$

Example 2.1 *(One-sample t-test). Assume we have a random sample from* $\mathcal{F} = \{N(\theta, \sigma^2)\}$ *distribution. We hope to 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).*

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 our test. If so, we find its size

$$\alpha = P(|T(X)| \geq t_{0.975,n-1} : H_0) = 0.05.$$

It is not as easy to write down its power function.

The p-value of this test is

$$p_0 = P(|T(X)| \geq T(x) : H_0)$$

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 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 dose 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, every small p-value shows how strong the evidence is against H_0. Small p-value does not necessarily indicate H_0 is an extremely poor model for the data. One way to avoid such issue might be to specify H_1 as $|\theta| > 0.1$ put H_0 as $|\theta| < 0.1$ instead.

2.6 Randomized test

Particularly in theoretical development, we often hope to construct a test with pre-given size. The above approach may not be feasible under some models.

Example 2.2 *Suppose we observe X from a binomial model with $n = 2$ and probability of success $\theta \in (0, 1)$. Let the desired size of the test be $\alpha = 0.05$ on the null hypothesis of $\theta = 0.5$. In this case, we have only 8 candidates for the critical region C. None of them result in a test of size $\alpha = 0.05$.*

One somewhat artificial approach to find a test with any pre-specified size is as follows. We do not reject H_0 if $X = 1$. If $X = 0, 2$, we toss a biased coin and reject H_0 when the outcome is a head. By selecting a coin such that $P(Head) = 0.1$, our probability of rejecting H_0 is 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 I mentioned earlier.

In abstract, a statistical significance test is represented as a function $\phi(x)$ such that $0 \le \phi(x) \le 1$. We reject H_0 with probability $\phi(x)$ when $X = x$. If $\phi(x) = 0$ or 1 only, then the sample space is neatly divided into the critical region and its complement. Otherwise, the regions corresponding to $0 < \phi(x) < 1$ are randomized region. When x falls into that region, we randomize our decision.

The notion that a significance test is described 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 is, for $F \in H_1$,

$$\beta(F) = \mathbb{E}\{\phi(X) : F\}.$$

There are few requirements on $\phi(x)$. Hence, it is natural to ask which $\phi(x)$ is the best. This calls for a definition of optimality. Historically, there was heated debate on these issues. I am simply not qualified/capable to teach on these issues.

2.7 Three ways to characterize a test

1. Define a test statistic, T, such that we reject H_0 when T is large. More specifically, we require T to have two specific properties: known and same distribution under whichever model in H_0; larger observed value of T indicates more extreme departure from H_0 toward the direction we try to capture. We compute p-value as

$$p = P(T \geq t_{obs} : H_0)$$

where t_0 is the observed value. When p is below some significance lever, we reject H_0. Note that when T has discrete distribution, we may use a continuity correction

$$p = P(T > t_{obs} : H_0) + 0.5P(T = t_{obs} : H_0).$$

2. Define a critical region C in terms of X. When the realized value $x \in C$, we reject H_0. The region C is often required to have given size α, which is

$$\sup_{H_0} P(X \in C) = \alpha.$$

Note that method 1 is a special case of method 2 by letting $C = \{x; T(x) > k\}$ for some k.

3. When X is discrete, we may get into situation where no critical region has given size α. This is usually not problematic in applications, problematic for theoretical discussion. Hence, we define a test as function $\phi(x)$ taking values between 0 and 1. We reject H_0 with probability $\phi(x)$ when x is the realized value of X.

Both methods 1 and 2 can be regarded as special case of method 3: by letting $\phi(x) = I(x \in C)$, we have a test that reject H_0 with probability 1 when $x \in C$, and do not reject H_0 otherwise.

Clearly, the specific test $\phi(x) = \alpha$ is a test of size α. Its existence ensures that a test with specific size is always possible. The statistical issue is always on finding one with some more desirable properties.

Chapter 3

Uniformly most powerful test

Let $\phi(x)$ be a test of size α for some H_0 and H_1. If for any size α test $\phi_1(x)$ and $F \in H_1$, we have

$$E\{\phi(X); F\} \geq E\{\phi_1(X); F\}$$

then $\phi(x)$ is the uniformly most powerful test.

The task of finding Uniformly Most Powerful (UMP) tests is often difficult or even impossible. Some may argue that such a result is not meaningful/useful. Nonetheless, There are special cases where UMP tests exist. We now start with the simplest case. Knowing which builds up our understanding of general problem.

3.1 Simple null hypothesis and simple alternative hypothesis

When a null hypothesis is identified, the task of statistical significance test is to see if the data suggest a departure from the null models in a specific direction. The simplest situation is where the statistical model $\mathcal{F}$ contains only two distinct distributions. The null hypothesis contains one, and the alternative hypothesis contains another. More specifically, we may present them as two density functions:

$$H_0 : f_0(x), \quad H_1 : f_1(x).$$

Note that if X is a set of i.i.d. random variables, the above setting still applies.

Based on measure theory, for any given two distributions, it is possible to find a σ-finite measure, with respect to which, the density functions of two distributions exist. This justifies the above general assumption.

Lemma 3.1 Neyman-Pearson Lemma*: Consider the simple null and alternative hypothesis test problem as specified.*

(1) For any size α between 0 and 1, there exists a test ϕ and a constant k such that

$$E_0\{\phi(X)\} = \alpha \tag{3.1}$$

and

$$\phi(x) = \begin{cases} 1 & \text{when } f_1(x) > kf_0(x) \\ 0 & \text{when } f_1(x) < kf_0(x) \end{cases} \tag{3.2}$$

(2) If a test has the properties (3.1) and (3.2), then it is the most powerful for testing H_0 against H_1.

(3) If ϕ is most powerful with size no more than α, then it satisfies (3.2) for some k. It also satisfies (3.1) unless there exists a test of size smaller than α and with power 1.

Proof and discussion.

Property (1) is for existence. A likelihood ratio test of size α exists. To prove the existence, let

$$\alpha(t) = P(f_1(X) > tf_0(X); H_0)$$

It is seen that $1 - \alpha(t)$ is a cumulative distribution function. Hence, there exists a t_0 such that

$$\alpha(t_0) \leq \alpha \leq \alpha(t_0-).$$

Let

$$\phi(x) = I(f_1(X) > t_0 f_0(X)) + cI(f_1(X) = t_0 f_0(X))$$

with $c = \frac{\alpha - \alpha(t_0)}{\alpha(t_0-) - \alpha(t_0)}$ if needed. Then this $\phi(x)$ is our test.

Remark: The seemly overly complex proof is caused due to of covering the discrete situation when $P\{f_1(X) = t_0 f_0(X)\} \neq 0$. Otherwise, the truthfulness is trivial.

Proof of (2):

Suppose $\phi(x)$ is the test given in (1), and $\tilde{\phi}$ is another test of size α. Then

$$\{\phi(x) - \tilde{\phi}(x)\}\{f_1(x) - kf_0(x)\} \geq 0$$

This implies, by integrating both sizes, with respect to the measure the density is defined,

$$\mathbb{E}_1\{\phi(X) - \tilde{\phi}(X)\} \geq k\mathbb{E}_0\{\phi(X) - \tilde{\phi}(X)\} = 0.$$

where we used $\mathbb{E}_1$ and $\mathbb{E}_0$ for expectations under the alternative and the null models. The right hand side equals 0 because two tests have the same size. Hence, ϕ has better power.

Proof of (3):

If $\tilde{\phi}(X)$ is also UMP, then we should have

$$P[\{\phi(x) - \tilde{\phi}(x)\}\{f_1(x) - kf_0(x)\} > 0] = 0.$$

Otherwise, the derivation in the proof of (2) would implies $\tilde{\phi}(X)$ has lower power which is in contradiction of the assumption hat $\tilde{\phi}(X)$ is also UMP. $\diamondsuit$

Property claims that the most powerful test has to be the likelihood ratio test. At the same time, this property leaves rooms for non-uniqueness. This is due to the flexibity of make decisions on the set of x such that $f_1(x)/f_0(x) = k$. The randomization test can be used to achieve the right size of the test. It may also be possible to split this set in other ways and obtain a non-randomization test with the right size. These test are all MP. Hence, MP test is not necessarily unique.

Example 3.1 *Let $X = (X_1, \ldots, X_n)$ be a random sample from $N(\theta, 1)$. Let us test $H_0 : \theta = 0$ against $H_1 : \theta = 1$.*

By Neyman-Pearson Lemma, the most powerful test has the form

$$\phi(x) = I(f_n(x; \theta = 1) > kf_n(x; \theta = 0)\}$$

where I use f_n for the n-variate density, and use $\theta = 1$ and $\theta = 0$ to highlight the parameter values under the alternative and null hypotheses. The constant k is to be chosen such that the test has given size.

Note that the critical region can be represented equivalently in many forms. Clearly,

$$\begin{aligned}
&\{f_n(x;\theta=1) > k f_n(x;\theta=0)\} \\
&\quad = \{\log f_n(x;\theta=1) > \log f_n(x;\theta=0) + \log k\} \\
&\quad = \{-\frac{1}{2}\sum_{i=1}^{n}(X_i-1)^2 > -\frac{1}{2}\sum_{i=1}^{n}X_i^2 + k'\} \\
&\quad = \{\sum_{i=1}^{n}(X_i-1)^2 < \sum_{i=1}^{n}X_i^2 - k''\} \\
&\quad = \{-2\sum_{i=1}^{n}X_i + n < -k''\} \\
&\quad = \{\sum_{i=1}^{n}X_i > k'''\}
\end{aligned}$$

In other words, there exists an k''' such that

$$\phi(x) = I(f_n(x;\theta=1) > k f_n(x;\theta=0)\} = I\{\sum_{i=1}^{n}X_i > k'''\}.$$

Since all we care is the size of the test, there is no need to find exactly how k''' is related to k. We need only work out the critical value k''' each time a size of the test is specified.

Suppose we want the test to have size $\alpha = 0.05$. This requires us to pick a specific value of k'''. Because the size is computed under the null hypothesis which has only a single distribution, we need only solve the equation

$$P(\sum_{i=1}^{n}X_i > c;\theta=0) = 0.05$$

which implies that $c = 1.645\sqrt{n}$ is the solution. If we set $\alpha = 0.025$, then $c = 1.960\sqrt{n}$ is the solution.

Suppose in addition to require the size of the test being 0.05, we also want to have power of the test $\beta(1) = 80\%$. This can be achieved by selecting an appropriate sample size n:

$$P(\sum_{i=1}^{n}X_i > 1.645\sqrt{n};\theta=1) \geq 0.8.$$

Because n is discrete, the problem should be interpreted as finding the smallest n such that the power is at least 0.8.

When $\theta = 1$, we have

$$\begin{aligned} P(\sum_{i=1}^{n} X_i > 1.645\sqrt{n}; \theta = 1) &= P(n^{-1/2}\sum_{i=1}^{n}(X_i - 1) > 1.645 - n^{1/2}; \theta = 1) \\ &= P(Z > 1.645 - n^{1/2}). \end{aligned}$$

with Z being a standard normal random variable. The 20% quantile of the standard normal is -0.842. Thus, we require $1.645 - n^{1/2} \leq -0.842$ or $n \geq (1.645 + 0.842)^2 = 6.18$. Thus, $n = 7$ meet the requirement.

Remark: It is seen that if the alternative hypothesis H_1 is replaced by $\theta = \theta_1$ for any $\theta_1 > 0$, the most powerful test itself remains the same. That is, the test is most powerful for any alternative such that $\theta_1 > 0$. In other words, the above test is also a UMP test against $H_1 : \theta > 0$. However, to attain the power of 80% at $\theta_1 = 0.5$, the required sample size will be increased.

Remark: It is easy to verify that the critical region of the most powerful test when H_1 becomes $\theta = \theta_1 < 0$ has the form

$$\sum X_i < c.$$

Clearly, a most powerful test for $\theta > 0$ cannot also be a most powerful for $\theta < 0$. Hence, the notion of most powerful is in general "alternative hypothesis" specific. It is often impossible to have a test that is uniformly most powerful against composite alternative hypothesis. Here, composite means the alternative hypothesis $\mathcal{F} - \mathcal{F}_0$ contains more than a single distribution.

Remark: Point of the example: the UMP test based on Neyman-Pearson Lemma is just the test we recommend in other courses.

3.2 Making more from N-P lemma

Theorem 3.1 *Suppose that there is a test $\phi(X)$ of size α such that for every $F_1 \in H_1$, $\phi(X)$ for testing H_0 against $\tilde{H}_1 : F = F_1$ is uniformly most powerful. Then it is UMP for H_0 versus H_1.*

Proof: Suppose $\tilde{\phi}(X)$ is another test of size α for testing H_0 versus H_1.

For any $F_1 \in H_1$, by the assumption on $\phi(X)$, we have

$$E\{\phi(X) : F_1\} \geq E\{\tilde{\phi}(X) : F_1\}.$$

This trivially shows that $\phi(X)$ is UMP against H_1. $\diamondsuit$

Example 3.2 *Suppose $X_1, \ldots, X_n$ is an iid sample from Poisson distribution. We test $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$.*

Consider testing $\tilde{H}_0 : \theta = 1$ versus $\tilde{H}_1 : \theta = 2$. The likelihood ratio $f(x;2)/f(x;1) = c\exp\{(\log 2)\sum x_i\}$. By Neyman–Pearson Lemma, one UMP test has the form of

$$\phi(X) = \begin{cases} 1 & \sum x_i > k; \\ c & \sum x_i = k; \\ 0 & \sum x_i < k \end{cases}$$

for some k and c to get the size of the test equaling α. That is, they are chosen so that

$$E\{\phi(X) : \theta = 1\} = \alpha.$$

Thus, the choice of k and c does not depend on $\tilde{H}_1$. Hence, it is UMP for $\tilde{H}_0$ versus H_1.

Next, we hope to retain the same proposition with $\tilde{H}_0$ replaced by H_0.

It is clear that $E\{\phi(X) : \theta\} < \alpha$ when $\alpha < 1$. Hence, $\phi(X)$ remains a size α test for H_0. Therefore, there cannot be any other test of size α having greater power at any $\theta > 1$.

The above result is more generally applicable.

3.3 Monotone likelihood ratio

Definition 3.1 *Suppose that the distribution of X belongs to a parameter family with density functions $\{f(x;\theta) : \theta \in \Theta \subset \mathbb{R}\}$.*

The family is said to have monotone likelihood ratio in $T(x)$ if and only if, for any $\theta_1 < \theta_2$,

$$\frac{f(x;\theta_2)}{f(x;\theta_1)}$$

is a nondecreasing function of $T(x)$ for values x at which at least one of $f(x;\theta_1)$ and $f(x;\theta_2)$ is positive.

It is seen that $T(x)$ is a useful statistic for the purpose of hypothesis test because it is a stochastically increasing function of θ.

Lemma 3.2 Monotonicity of $E\{T(x)\}$ *Suppose X has a distribution from a monotone likelihood ratio family. Then $E\{T;\theta\}$ is nondecreasing in θ.*

Proof: Because when $\theta_2 > \theta_1$,

$$\frac{f(x;\theta_2)}{f(x;\theta_1)}$$

is non-decreasing in T. Two random variables, $T(X)$ and $f(X;\theta_2)/f(X;\theta_1)$ are positively correlated when the distribution of X is $f(x;\theta_1)$. Let $\mu_1 = E_1\{T(X)\}$, the expectation under θ_1. We have

$$E_1\{[T(X) - \mu_1]f(X;\theta_2)/f(X;\theta_1)\} \geq 0.$$

Expanding this inequality gives us the conclusion. ♢

Extension This conclusion is applicable to any nondecreasing function $g(T)$.

Example 3.3 *One parameter exponential family*

$$f(x;\theta) = \exp(\eta(\theta)T(x) - \xi(\theta))h(x)$$

has monotone likelihood ratio in $T(x)$ when $\eta(\theta)$ is a nondecreasing function in θ.

Example 3.4 *Let $X_1, \ldots, X_n$ be an iid sample from*

$$f(x;\theta) = \theta^{-1} I(0 < x < \theta).$$

Then the distribution family of $X = (X_1, \ldots, X_n)$ has monotone likelihood ratio in $X_{(n)}$, the largest order statistic.

Theorem 3.2 *Suppose the distribution of X is in a parametric family with real valued parameter θ and has monotone likelihood ratio in $T(X)$.*

Consider $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$.

(i) There exists a UMP test of size α, given by

$$\phi(X) = \begin{cases} 1 & T(X) > k; \\ c & T(X) = k; \\ 0 & T(X) < k. \end{cases}$$

(ii) For any $\theta < \theta_0$, $\phi(X)$ minimizes $\beta(\theta)$ (the type I error at θ) among all $\tilde{\phi}$ such that $E\{\tilde{\phi}(X); \theta_0\} = \alpha$.

Proof

(i) By Neyman–Pearson lemma, this test is one of the Most Powerful tests for $\tilde{H}_0 : \theta = \theta_0$ against $\tilde{H}_1 : \theta = \theta_1$ for any $\theta_1 > \theta_0$ because the density ratio is an increasing function of T. Hence, $\phi(X)$ is Uniformly Most Powerful for $\tilde{H}_0$ against $H_1 : \theta > \theta_0$.

By Lemma on the Monotonicity of $E\{T(x)\}$, $E\{\phi(X); \theta\}$ is a nondecreasing function of θ. Therefore $E\{\phi(X); \theta\} \leq \alpha$ for all $\alpha \in H_0$. Thus, $\phi(X)$ is a size-α test for H_0 versus H_1. Subsequently, it is UMP H_0 versus H_1 by the extended N-P lemma.

(ii) Let us define $\xi = -\theta$ so that we have a density function

$$g(x; \xi) = f(x; -\theta).$$

In terms of ξ, the family has monotone density ratio for ξ in $\tilde{T} = -T(x)$.

Consider testing for $H_0^* : \xi \leq \xi_0 = -\theta_0$ versus $H_1^* : \xi > \xi_0 = -\theta_0$ with size $\alpha^* = 1 - \alpha$.

Hence, the UMP tests will have the following form

$$\phi^*(X) = 1 - \phi(X) = \begin{cases} 1 & T(X) < k; \\ c & T(X) = k; \\ 0 & T(X) > k. \end{cases}$$

with k and c chosen such that the test has size α^*. We remark here that the middle part is not unique but it does not invalidate our claim. This uniformly most power test has its power maximized is the same as having

$$E\{\phi(X) : \theta\} = 1 - E\{\phi^*(X) : \xi\}$$

minimized when $\xi \in H_1^*$ which is the same as $\theta \in H_0$. This completes the proof. ♢

Example 3.5 Uniform distribution *Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on $(0, \theta)$. Then the distribution family of $X = (X_1, \ldots, X_n)$ has monotone likelihood ratio in $X_{(n)}$.*

For any $\theta_1 < \theta_2$, the density ratio

$$f(x; \theta_2)/f(x; \theta_1) = (\theta_1/\theta_2)^n \frac{I(0 < x_{(n)} < \theta_2)}{I(0 < x_{(n)} < \theta_1)}.$$

Other than the constant factor $(\theta_1/\theta_2)^n$, the ratio takes three values: 1, ∞ and undefined. The last case does not matter as both densities are zero excluded in the definition. This is clearly an increasing function of $X_{(n)}$.

Consider the hypothesis $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$. By the theorem we have just proved, the UMP test can be written as

$$\phi(X) = \begin{cases} 1 & X_{(n)} > k; \\ c & X_{(n)} = k; \\ 0 & X_{(n)} < k. \end{cases}$$

for some k and c.

Because the distribution of $X_{(n)}$ is continuous. $P(X_{(n)} = k) = 0$ for any k. Hence, it can be simplified into

$$\phi(X) = \begin{cases} 1 & X_{(n)} > k; \\ c & X_{(n)} = k; \\ 0 & X_{(n)} < k. \end{cases}$$

The c.d.f. of $X_{(n)}$ is given by $(x/\theta_0)^n$ under null for $0 < x < \theta_0$. Hence, the choice of k is determined by

$$\alpha = 1 - (k/\theta_0)^n$$

and $k = \theta_0(1 - \alpha)^{1/n}$ is the solution.

The power at $\theta > \theta_0$ is

$$\beta(\theta) = 1 - (1 - \alpha)(\theta_0/\theta)^n.$$

Remark The UMP is not unique as the density ratio is a discrete random variable.

Chapter 4

Generalizing Neyman–Pearson Lemma

Theorem 4.1 *Consider the situation where* $H_0 = \{f_1, f_2\}$ *and* $H_1 = \{f_3\}$. *Let* α_1, α_2 *be constants taking values between 0 and 1.*

Let $\mathcal{T}$ *be the class of tests such that*

$$E\{\phi(X); f_j\} \leq \alpha_j; \quad j = 1, 2.$$

Let $\mathcal{T}_0$ *be a subset of* $\mathcal{T}$ *such that the above inequalities replaced by equalities.*

Suppose there are constants k_1 *and* k_2 *such that*

$$\phi_*(x) = \begin{cases} 1 & f_3 > k_1 f_1 + k_2 f_2 \\ 0 & f_3 < k_1 f_1 + k_2 f_2 \end{cases}$$

is a member of $\mathcal{T}_0$.

We have two conclusions:

(i) $E\{\phi_*(X); f_3\} \geq E\{\phi(X); f_3\}$ *for any* $\phi(x) \in \mathcal{T}_0$.

(ii) If both $k_1 \geq 0$ *and* $k_2 \geq 0$, *then* $E\{\phi_*(X); f_3\} \geq E\{\phi(X); f_3\}$ *for any* $\phi(x) \in \mathcal{T}$.

Proof

(i) Simply construct function

$$\{\phi_*(x) - \phi(x)\}\{f_3 - (k_1 f_1 + k_2 f_2)\}$$

which is non-negative at all x. If both $\phi_*(x), \phi(x) \in \mathcal{T}_0$, we find $E\{\phi_*(X); f_3\} \geq E\{\phi(X); f_3\}$ right away by integrating the above function.

(ii) If $\phi_*(x) \in \mathcal{T}_0$, it means that $E\{\phi^*(x); f_1\} = \alpha_1$ and $E\{\phi^*(x); f_2\} = \alpha_2$. When $\phi(x) \in \mathcal{T}$, it means we have $E\{\phi(X); f_j 1\} \leq \alpha_1$; $E\{\phi(X); f_2\} \leq \alpha_2$.

Integrating $\{\phi_*(x) - \phi(x)\}\{f_3 - (k_1 f_1 + k_2 f_2)\}$ with respect the corresponding σ-finite measure, we find

$$E\{\phi_*(X); f_3\} - E\{\phi(X); f_3\} \geq k_1[\alpha_1 - E\{\phi(X); f_1\}] + k_2[\alpha_2 - E\{\phi(X); f_2\}] \geq 0$$

when k_1 and k_2 are nonnegative.. Hence, the conclusion is true. $\diamond$

This proposition works if there are many but finite number of density functions in H_0. One shortcoming is how do we know whether such k_1 and k_2 exist. Answering this question is involved. So I only copy the following result below for your reference.

Theorem 4.2 *Let f_1, f_2, f_3 be three density functions with respect to the same σ-finite measure.*

The set $M = \{(E\{\phi(X); f_1\}, E\{\phi(X); f_2\}) : \phi$ is a test$\}$ is convex and closed.

If (α_1, α_2) is an interior point of M, then there exist constants k_1, k_2 such that $\phi_*(x)$ as given above exists.

Discussion: The N-P lemma gives us UMP when both H_0 and H_1 contains a single distribution. We have generalized N-P lemma to the situation where H_1 contains many distribution. This result expands the N-P lemma further: it allows H_0 to contain two distributions.

When H_0 is given in the form of $1 \leq \theta \leq 2$, say under $N(\theta, 1)$ model assumption, the distributions in H_0 that matter are the ones with $\theta = 1$ and $\theta = 2$. Once a UMP is obtained for $\tilde{H}_0 : \{\theta = 1, \theta = 2\}$, it might be possible to show this test is also a UMP for H_0 itself.

4.1 One parameter exponential family

The generalized N-P lemma has its targeted application to problems related to one parameter exponential family.

Theorem 4.3 *Suppose we have a sample from a one parameter exponential family with density function given by*

$$f(x;\theta) = \exp(\theta Y(x) - A(\theta))h(x).$$

This family has monotone density ratio in $T_n(x) = \sum Y(x_i)$.

Suppose we want to test for $H_0 : \theta \notin (\theta_1, \theta_2)$ *versus* $H_1 : \theta \in [\theta_1, \theta_2]$ *for some* $\theta_1 \neq \theta_2$.

(i) A UMP test of size α *is given by*

$$\phi(T) = \begin{cases} 1 & k_1 < T_n(x) < k_2; \\ c_j & T_n(x) = k_j, \quad j = 1, 2; \\ 0 & T_n(x) < k_1 \text{ or } T_n(x) > k_2 \end{cases}$$

where k_1, k_2, c_1, c_2 *are chosen such that*

$$E\{\phi(X); \theta_j\} = \alpha, \quad j = 1, 2.$$

(Note $0 < c_1, c_2 < 1$*).*

(ii) The test given in (i) minimizes type I error at every $\theta \in H_0$ *among the tests satisfying* $E\{\phi(T); \theta_j\} = \alpha$, $j = 1, 2$.

Proof of this proposition

Since $T_n(x) = \sum Y(x_i)$ is sufficient for θ. We need only work on a test defined as function of $T_n(x)$. Otherwise, $E\{\phi(X)|T\}$ is a test with the same size and power function.

(i) Next, we first work on a UMP for testing $\tilde{H}_0 : \{\theta_1, \theta_2\}$ versus $\tilde{H}_1 : \{\theta_3\}$ for some $\theta_3 \in (\theta_1, \theta_2)$. Note the structure: the alternative model is a single distribution within the interval; while the null models are two distributions at two ends.

According to the generalized Neyman–Pearson lemma in the form of proposition, such a UMP may exist.

For any test $\phi(T)$, we denote its rejection probability by $\beta(\theta; \phi) = E\{\phi(T); \theta\}$. One candidate test for having UMP property is proposed to be

$$\phi(T) = \begin{cases} 1 & f(x;\theta_3) > k_1 f(x;\theta_1) + k_2 f(x;\theta_2); \\ c & f(x;\theta_3) = k_1 f(x;\theta_1) + k_2 f(x;\theta_2); \\ 0 & f(x;\theta_3) < k_1 f(x;\theta_1) + k_2 f(x;\theta_2). \end{cases}$$

It is possible to find the corresponding c, k_1 and k_2 such that

$$\beta(\theta_1;\phi) = \beta(\theta_2;\phi) = \alpha.$$

We do not elaborate on the existence of c, k_1 and k_2 but assume so.

The inequality

$$f(x;\theta_3) > k_1 f(x;\theta_1) + k_2 f(x;\theta_2)$$

used in defining the above $\phi(T)$ under the exponential family can be written as

$$a_1 \exp(b_1 T) + a_2 \exp(b_2 T) < 1.$$

Due to the relative sizes of θ_1 and θ_2, we must have $b_1 b_2 < 0$.

We find the sign information about a_1 and a_2 is helpful and give it a careful discussion as follows:

(1) If both a_1, a_2 are smaller than 0, then the inequality holds with probability 1. That is, the size of the text would be 1. This is disallowed.

(2) If $a_1 \leq 0$ but $a_2 > 0$, together with $b_1 b_2 < 0$, it implies that $a_1 \exp(b_1 T) + a_2 \exp(b_2 T)$ is monotone in T. That is, the inequality in the form of

$$a_1 \exp(b_1 T) + a_2 \exp(b_2 T) < 1$$

is equivalent to one of $T < t$ or $T > t$ for some constant t. If so, the rejection probability $\beta(\theta;\phi)$ would be an monotone function in θ. This contradicts $\beta(\theta_1;\phi) = \beta(\theta_2;\phi) = \alpha$.

(3) The only choice left is $a_1 > 0$ and $a_2 > 0$. Note that $a_1 \exp(b_1 T) + a_2 \exp(b_2 T)$ is now convex in T. The inequality in the form of

$$a_1 \exp(b_1 T) + a_2 \exp(b_2 T) < 1$$

is equivalent to the one in the form of

$$k_1 < T < k_2$$

for another set of k_1 and k_2.

In summary, our discussion leads to conclusion that the test is to reject H_0 when $k_1 < T < k_2$. This is in good agreement with our intuition. Based on the generalized Neyman-Pearson together with $a_1 > 0$ and $a_2 > 0$, this $\phi(T)$ is UMP for testing $\tilde{H}_0 : \{\theta_1, \theta_2\}$ versus $\tilde{H}_1 : \{\theta_3\}$.

Because this $\phi(T)$ does not depend on the specific choice of θ_3, the UMP conclusion extends to $\tilde{H}_0 : \{\theta_1, \theta_2\}$ versus H_1.

To get the full generality that $\phi(T)$ is UMP for testing $H_0 : \theta \notin [\theta_1, \theta_2]$ versus H_1, we only need to verify that

$$\beta(\theta; \phi) \leq \alpha$$

at every $\theta \in [\theta_1, \theta_2]$. *This is from the convexity of β: This original writing is wrong. The correct proof is given below.*

Consider the test problem with $\tilde{H}_0 : \{\theta_1, \theta_2\}$ against $\tilde{H}_1 : \{\theta_3\}$ for some θ_3 in the original H_0. Consider the test $\phi^*(T) = 1 - \phi(T)$. It can be verified (similar to what have been done) that this $\phi^*(T)$ has the form specified in the generalized N-P lemma. Therefore, $\phi^*(T)$ has the best power at θ_3 (among those with $\beta(\theta_1) \leq 1 - \alpha, \beta(\theta_2) \leq 1 - \alpha$). This implies that $\phi(T)$ has the lowest type I error possible, which makes it at least as low as α.

(ii) Essentially, it has been proved by the above revised proof. $\diamondsuit$

Remark: The result itself is mathematically interesting. Its usefulness will be fully seen in the next topic.

4.2 Two-sided alternatives

Consider the hypothesis $\theta = 1$ versus the alternative $H_1 : \theta \neq 1$ given observations from exponential distribution with mean θ. Let us separate H_1 into $H_{11} : \theta > 1$ and $H_{12} : \theta < 1$. It is not hard to see H_1 is called two-sided alternative. Assume the size of the test is required to be α.

The UMP for H_0 versus H_{11}, according to our discussion, is given by

$$\phi_1(x) = I(\sum x_i > k_1)$$

for so that $E_0\{\phi_1(x)\} = \alpha$.

The UMP for H_0 versus H_{12}, according to our discussion, is given by

$$\phi_2(x) = I(\sum x_i < k_2)$$

for so that $E_0\{\phi_2(x)\} = \alpha$.

Suppose a UMP test $\phi(x)$ exists for H_0 versus H_1. This test remains a UMP for H_0 versus H_{11}. Hence, we must have $\phi(x) = \phi_1(x)$ except for a zero-measure set of x. For the same reason, we must also have $\phi(x) = \phi_2(x)$ except for a zero-measure set of x. Such a $\phi(x)$ clearly does not exist. Hence, there exist no UMP for this problem.

This example is not restrict to the exponential distribution but true in general. We may provide a sensible test based on the idea of "pure significance test". If we define

$$T_n = \max\{\bar{x}, 1/\bar{x}\}.$$

A large value of T_n (deviating from 1) is a good indication that $\theta = 1$ is violated. Thus, we may compute

$$p_0 = P(T_n \geq t_{obs}; \quad \theta = 1)$$

as the p-value and reject, say when $p_0 < 0.05$. We can all agree that this is a sensible test. However, we cannot help to ask whether this is the best we can do. Furthermore, in what sense that this test is best? We could have defined the test statistics as

$$T_n' = \max\{\bar{x}, 2/\bar{x}\}.$$

A test best on T_n' has the same properties.

In some situations, it is possible to set up a useful standard. This is our next topic.

4.3 Unbiased test

A great person is not necessary the best in a big population, he/she might be the best in a small community. Maybe the above sensible test is optimal in a more restricted class. One way to narrow down the community is to require a test to be unbiased.

Definition 4.1 *A test is unbiased if for some α, we have*

$$\sup_{F\in H_0} E\{\phi(X); F\} \le \alpha; \qquad \inf_{F\in H_1} E\{\phi(X); F\} \ge \alpha.$$

Justification of unbiasedness. Every guilty party should be more likely to be sent to prison than every innocent party in a court. Be aware of the wording: merely more likely.

Definition 4.2 *If a test is most power at every $F \in H_1$ within the class of unbiased tests of size α, it is the Uniformly Most Powerful Unbiased (UMPU) test of level α.*

4.3.1 Existence of UMPU tests

The idea of unbiasedness is helpful in some typical situations. We only discuss this topic for a one-parameter exponential family with density function given by

$$f(x;\theta) = \exp(\theta Y(x) - A(\theta))h(x).$$

This family has monotone density ratio in $T = \sum Y(x_i)$. Of course, T is complete and sufficient for θ. The above parameterization is a natural one.

Theorem 4.4 *Suppose we want to test for $H_0 : \theta \in [\theta_1, \theta_2]$ versus $H_1 : \theta \notin [\theta_1, \theta_2]$ for some $\theta_1 \ne \theta_2$.*

A UMPU test of size α is given by

$$\phi(T) = \begin{cases} 1 & T < k_1 \text{ or } T > k_2 \\ c_j & T = k_j, \quad j = 1, 2. \\ 0 & k_1 < T < k_2. \end{cases}$$

where k_1, k_2, c_1, c_2 are chosen such that

$$E\{\phi(T); \theta_j\} = \alpha, \quad j = 1, 2.$$

(Note $0 < c_1, c_2 < 1$).

Proof: The test should clearly be based on T as it is complete and sufficient for θ.

According to a theorem we proved earlier, $1-\phi(T)$ for the $\phi(T)$ defined above is a UMP for $\tilde{H}_0 : \theta \notin [\theta_1, \theta_2]$ versus $\tilde{H}_1 : \theta \in [\theta_1, \theta_2]$ of size $\tilde{\alpha} = 1-\alpha$.

We put **a side proposition** with a proof here. Under exponential family, $E\{\phi(T); \theta\}$ is continuous in θ for any test $\phi(T)$. Because of this proposition, if $\phi(T)$ is an unbiased test for H_0 versus H_1, we must have

$$E\{\phi(T); \theta_j\} = \alpha, \quad j = 1, 2.$$

If another unbiased test $\phi_*(T)$ is of size α for H_0 versus H_1 but has higher power at some $\theta_3 \in H_1$, we would have

$$E\{\phi_*(T); \theta_1\} = E\{\phi_*(T); \theta_2\} = \alpha$$

and

$$E\{\phi_*(T); \theta_3\} > E\{\phi(T); \theta_3\}.$$

In terms of $\tilde{H}_0$ and $\tilde{H}_1$, we find a pair of tests: $1-\phi^*(T)$ and $1-\phi(T)$ both of size $1-\alpha$, unbiased, but the type I error of $1-\phi^*(T)$ is lower than that of $1-\phi(T)$ at $\theta_3 \in \tilde{H}_0$. This contradicts the UMP result (ii) given earlier.

4.4 UMPU for normal models

The normal distribution has two parameters. Thus, what we have discussed do not allow us even to show the optimality of t-test. We will have this topic picked up later.

Chapter 5

Locally most powerful test

While the UMP theorems seem impressive mathematically, they are not broad enough. Other than being hard to find them, they often do not exist unless the data are from some classical well behaved parametric models. We have not choice but to relax the optimality requirements if we need some method for hypothesis test.

Definition 5.1 *Consider the simple null hypothesis* $H_0 : \{\theta_0\}$ *against* $H_1 : \theta > \theta_0$ *in a one parameter setting. Let* $\beta(\theta)$ *be the power function of a test* $\phi(x)$ *of size* α*. Suppose for any other test* $\phi^*(x)$ *of size* α*, there exists an* $\epsilon > 0$ *such that*

$$E\{\phi^*(X); \theta\} \leq \beta(\theta)$$

for all $\theta \in (\theta_0, \theta_0 + \epsilon)$*. Then we say* $\phi(X)$ *is locally most powerful.*

5.1 Score test and its local optimality

Let $\{f(x;\theta) : \theta \in \Theta\}$ be a regular statistical model with score function defined as

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

We consider the case where Θ is an interval of real number. A test defined by

$$\phi(x) = I(S(\theta_0; x) > k)$$

is a locally most powerful test for $H_0 : \{\theta_0\}$ against $H_1 : \theta > \theta_0$ among the tests with size $\alpha = E\{\phi(X); \theta_0\}$.

Remark: For mathematical simplicity, we have totally ignored the request of having pre-specified size α. It is a test of whatever the size itself ends up.

Being regular for a model here means that for any $T(X)$ integrable,

$$E\{T(X); \theta\} = \int T(x) f(x; \theta) d\nu(x)$$

is differentiable with respect to θ and the derivative can be taken within the integration size. In simple words, the order of derivative and integration can be exchanged without alter the outcome.

One drawback of this test is: the null hypothesis must be a simple one. Namely, it contains only one distribution. Apparently, otherwise, the score function would contain unknown parameter values.

Proof: Being locally most power is the same as to require

$$\beta(\theta) = E\{\phi(X); \theta\}$$

has the largest possible derivative at $\theta = \theta_0$. Thus, we show the test defined by $\phi(x) = I(S(\theta_0; x) > k)$ makes $\beta(\theta)$ having the largest derivative.

Let $\phi_*(x)$ be another test of the same size. Then

$$\{\phi(x) - \phi_*(x)\}\{S(x) - k\} \geq 0.$$

Taking expectation under distribution $f(x; \theta_0)$, and noticing $E\{\phi(x) - \phi_*(x); \theta_0\} = 0$, we get

$$E\{[\phi(x) - \phi_*(x)]S(x)\} \geq 0.$$

Under regularity conditions, the left hand side are difference of derivatives of two power functions. ♢

Example 5.1 *Let $X_1, \ldots, X_n$ be an iid sample from Cauchy distribution with*

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

Consider the test for $H_0 : \theta = 0$ against $H_1 : \theta > 0$.

The locally most powerful test is

$$\phi(x) = I(2\sum x_i/(1+x_i^2) > k)$$

for some k such that the test has the require size.

The distribution of $\sum X_i/(1+X_i^2)$ is not well investigated. There is no simple way to compute k value with which the size requirement is meet. However, it is easy to show that $\sum X_i/(1+X_i^2)$ is asymptotically $N(0, n/8)$. Thus, when n is large (say larger than 20), we may use normal approximation to get an approximate k value.

5.2 General score test

The locally most powerful test we find is a score test. When the model assumption $f(x;\theta)$ is plausible, we have

$$E\{S(x;\theta);\theta\} = 0$$

for any θ under regularities conditions. Thus, if a statistician is asked to judge when $\theta = \theta_0$ is a plausible value, he or she could take a look of the value

$$\sum S(x_i;\theta_0)$$

where the summation is needed under the i.i.d. setting. From pure significance test point of view, this is an informative statistic about whether θ_0 is an acceptable value.

Note that this idea works when θ is a vector parameter. Suppose we can split a vector parameter $\theta = (\xi, \eta)$ and wishes to test $H_0 : \xi = \xi_0$. One may work out

$$S(x, \xi_0, \eta) = \left.\frac{\partial \log f(x;\xi,\eta)}{\partial \xi}\right|_{\xi=\xi_0}$$

and build a test statistics based on

$$\sum_{i=1}^{n} S(x_i, \xi_0, \hat{\eta}_0)$$

where $\hat{\eta}_0$ is the MLE of η given $\xi = \xi_0$. Apparently, there are more issues before a test of this nature can be concretely spelled out. The most obvious difficulty is to find out its distribution, may be the asymptotic one.

In general, if there is a function $g(x;\theta)$ such that $E\{g(X;\theta);\theta\} = 0$ for all $\theta \in H_0$. Then, the value of

$$T = \inf_{\theta \in H_0} |\sum g(x_i;\theta)|$$

could be used as some kind of statistic for "pure significance hypothesis test". Among all such choices, the score test is optimal in some sense.

Chapter 6

Likelihood ratio test

The Neyman–Pearson lemma may not be too useful when we work with more complex models. However, it tells us that the "optimal metric" in testing for a model θ_0 against another θ_1 is their relative likelihood size. This motivates the likelihood ratio test.

Let us consider the situation where we have a random sample from a distribution that belongs to a parametric distribution family:

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

Let H_0 and H_1 be subsets of Θ. Let $L_n(\theta) = \prod_{i=1}^n f(X_i;\theta)$ be the likelihood function under i.i.d. setting.

The model in H_0 that fits the data best is the one with θ value that maximizes $L_n(\theta)$ within H_0. Similarly, the best value under the alternative is the one that maximizes $L_n(\theta)$ for $\theta \in H_1$. Let $\hat{\theta}_0$ be the maximizer of the former. However, let $\hat{\theta}_1$ to be the maximizer of $L_n(\theta)$ over Θ, the entire parameter space.

The likelihood ratio is then defined as

$$\Lambda_n = \frac{L_n(\hat{\theta}_0)}{L_n(\hat{\theta}_1)} = \exp\{\sup_{\theta\in H_0} \ell_n(\theta) - \sup_{\theta\in\Theta} \ell_n(\theta)\}.$$

The likelihood ratio statistic is defined as

$$R_n = -2\log\Lambda_n = 2\{\ell_n(\hat{\theta}_1) - \ell_n(\hat{\theta}_0)\}$$

where we have used the log likelihood function

$$\ell_n(\theta) = \log L_n(\theta) = \sum_i \log f(x_i; \theta).$$

We define the likelihood ratio test as

$$\phi(x) = I\{R_n(x) \geq c\}$$

for some c such that the test has pre-specified size. From now on, we will not pay attention to the situation where $R_n(x)$ is discrete. More precisely, the test will be regarded as if randomization is never needed so that the size of the test is exactly the same as pre-specified. One reason for this convention is that to find the precise critical value c is generally difficult, even without this complication. In addition, when n is large, we have the following result due to Wilks that works nicely. This result asks us to use an approximate critical value. If it is an approximation already, it is pointless to have another layer of approximation.

Theorem 6.1 *Suppose H_0 is a subset in a m-dimensional subspace of Θ and Θ is an open subset of R^{m+d}. Under some regularity conditions and the i.i.d. setting,*

$$P\{R_n \leq t) \to P(Z_1^2 + Z_2^2 + \cdots Z_d^2 \leq t\}$$

as the sample size $n \to \infty$, and under the true model $\theta = \theta_0$.

We have used $Z_1, \ldots, Z_d$ as a set of i.i.d. standard normal random variables. Based on the above theorem, when n is large, a test with approximate size α is obtained by choosing the critical value $c = \chi_d^2(1-\alpha)$, the $1-\alpha$th quantile of the chisquare distribution with d degrees of freedom.

We will not give the proof and conditions for the moment. Let us examine a few examples of likelihood ratio test.

Example 6.1 *Let us go back to exponential distribution in which $H_0 : \theta = 1$ and $H_1 : \theta \neq 1$. Given a random sample of observations, we find $\hat{\theta}_1 = \bar{X}_n$. Since we do not have any choices under H_0 but $\theta = 1$, the likelihood ratio statistic is given by*

$$R_n = 2n\{\log \bar{X}_n - (\bar{X}_n - 1)\}$$

Under the null hypothesis, it is known that $\bar{X}_n \to 1$. *Thus, we have, approximately,*

$$2n\{(\bar{X}_n - 1) - \log \bar{X}_n\} = n(\bar{X}_n - 1)^2 + o_p(1)$$

where $o_p(1)$ *is an asymptotically zero random quantity.*

By the central limit theorem, $\sqrt{n}(\bar{X}_n - 1)$ *is asymptotically* $N(0,1)$ *under the null hypothesis:* $\theta = 1$. *This implies that* R_n *is asymptotically* χ^2_1.

Because of this, an asymptotical critical region for a size 0.05 *likelihood ratio test is approximately*

$$C = \{R_n \geq 3.841\}.$$

In the form of test function, $\phi(x) = I\{R_n \geq 3.841\}$.

Suppose we put H_1 as the set of θ-values larger than 1. If so, the MLE of θ under H_1 is no longer always $\bar{X}_n$. In this case, the limiting distribution of R_n is not χ^2_1. We will see that the regularity condition is not satisfied with this H_1. That is, the theorem does not apply to this problem.

Example 6.2 *Consider the test problem where an iid sample is from* $N(\theta, \sigma^2)$ *and* $H_0 : \theta = 0$ *against* $H_1 : \theta \neq 0$.

The MLE under H_1 *is given by* $\hat{\theta} = \bar{X}_n$ *and* $\hat{\sigma}^2_n = n^{-1}\sum(X_i - \bar{X}_n)^2$. *Under the null hypothesis, the MLE of* σ^2 *is* $\hat{\sigma}^2_0 = n^{-1}\sum X_i^2$. *It is not too hard to find that*

$$R_n \approx n \log\left[1 + \frac{\hat{\sigma}^2_0 - \hat{\sigma}^2_n}{\hat{\sigma}^2_0}\right] \approx n\bar{X}^2_n.$$

Thus, its limiting distribution is χ^2_1.

There are many reasons why the likelihood ratio test is preferred by statisticians and practitioners. Let me try to give you a list that I am aware of.

a) Because the limiting distribution of the likelihood ratio statistic under regularity conditions is chisquare, it does not depend on unknown parameters. We say that it is **asymptotically pivotal**. One may recall that one of the two preferred properties of a test statistic is that the statistic has a sample distribution free from unknown parameters.

b) Due to Neyman-Pearson Lemma, we believe that the LRT is nearly "most powerful". The claim is unproven, and likely false, yet when lack of evidence to the contrary, we like to believe the power of the LRT is superior.

c) Whether a limiting distribution is useful or not depend on how closely it approximates the finite sample distribution when the sample size is in the range occurs in applications. For example, if a clinical trial typically recruits 200 patients, then the limiting distribution is useful when it provides a good approximation when $n = 200$. It would be not so useful if the approximation is poor until $n = 2000$. There is a general belief that the chisquare approximation for LRT is often good for moderate n.

d) The LRT is invariant to parameter transformation. If a one-to-one transformation is applied to θ to get $\xi = g(\theta)$. The LRT remains equal when testing $g(H_0)$ against $g(H_1)$. Note that I am regarding H_0 and H_1 as subsets of parameters.

At the same time, I would like to point out that the LRT is often abused. The asymptotic chisquare distribution is valid only if (a) the true value of the parameter is an interior point of the parameter space; (b) the distribution family is regular; (c) the observations are i.i.d. . The result may still be valid when (c) is violated. Yet the validity depends on the structure of the model which should be examined before the LRT is used. If (a) is violated, the result is almost surely void. If (b) is violated, we probably should not use LRT although there are examples, I believe, that the asymptotic result remains valid. Yet there is no reason to assume so in general.

Example 6.3 *Suppose we have an i.i.d. sample from*

$$f(x;\pi) = (1-\pi)N(0,1) + \pi N(1,1).$$

The parameter space is $[0,1]$. *Suppose we want to test* $H_0 : \pi = 0$ *against* $H_1 : \pi > 1$.

Under the null model, that is, assume the true value of $\pi = 0$, the MLE $\hat{\pi} = 0$ with probability approximately 0.5. Because of this, the limiting distribution of the likelihood ratio statistic equals 0 with probability 0.5. Hence, the chisquare limiting distribution does not apply. The reason for the failure is that $\pi = 0$ is on the boundary of the parameter space.

6.1 Review of likelihood related results

Suppose we have an iid sample of size n from a parametric distribution family $\{f(x;\theta) : \theta \in \Theta\}$. The observations will be denoted as $x_1, \ldots, x_n$. The corresponding random variables are X, X_i and so on.

Under some conditions, the maximum likelihood estimator is known to be consistent and asymptotically normal as $n \to \infty$.

Yet fewer and fewer statistics students know these "conditions" under which these claims are valid. I am particularly despised the practice of stating "the result is true by assuming the conditions are satisfied". It would much better to state that "the result is true if these conditions are satisfied"

The consistency proof in general is technical and tedious. I will provide something not as general but useful.

6.1.1 Consistency of MLE for one-dimensional θ and as a local maximum

Even for this simple case, we still need to list conditions carefully. Let the log likelihood function and score function be

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

$$S_n(\theta) = \sum_{i=1}^{n} \frac{\partial\{\log f(x_i;\theta)\}}{\partial\theta}. \tag{6.2}$$

Regularity conditions for this special case are:

R0 the parameter space of θ is an open interval in $\mathbb{R}$

R1 $f(x;\theta)$ is differential 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 $N(\theta_0)$,

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

hold for all x, and

$$\int g(x)dx < \infty; \quad E_\theta\{H(X)\} < \infty.$$

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

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

Although the integration in regularity conditions 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 Poisson distribution in which the integration is interpreted as summation. All conditions are stated as for all x. An except over a 0-measure set of x is allowed (with respect to the corresponding σ-finite measure), as long as it is the same for all θ.

Lemma 6.1 *Under regularity conditions, we have*

$$E_\theta \left\{\frac{\partial \log f(x;\theta)}{\partial \theta}\right\} = 0.$$

Proof. It is seen that

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

for all θ. Hence, for any δ, we have

$$\int \frac{f(x;\theta+\delta) - f(x;\theta)}{\delta} dx = 0.$$

Condition R2 (i) makes it okay to let $\delta \to 0$ within the integration, we hence get

$$\int f'(x;\theta)dx = E_\theta \left\{ \frac{\partial \log f(x;\theta)}{\partial \theta} \right\} = 0.$$

$\diamondsuit$

Sometimes, the regularity conditions are regarded as conditions to allow the exchange the order of derivative and integration.

Lemma 6.2 *Under regularity conditions, we have*

$$E_\theta \left\{ \frac{\partial \log f(x;\theta)}{\partial \theta} \right\}^2 = -E_\theta \left\{ \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} \right\}$$

Proof. Last lemma shows that

$$\int f'(x;\theta)dx = 0.$$

The Condition R2 (ii) now allows us to get

$$\int f''(x;\theta)dx = 0.$$

Note that

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

Multiplying $f(x;\theta)$ and integrating with respect to x, we get the conclusion. $\diamondsuit$

Lemma 6.3 *Under regularity conditions,*

$$E_0 \left\{ \frac{\partial \log f(x;\theta)}{\partial \theta} \right\}$$

is a strictly decreasing function of θ in a neighborhood of θ_0.

Proof. Condition $R3$ topped with $R2(iii)$ implies

$$E_0\left\{\frac{\partial \log f(x;\theta)}{\partial\theta}\right\}^2 > 0$$

at θ in a neighborhood of θ_0. Thus,

$$\begin{aligned} E_0\left\{\frac{\partial \log f(x;\theta)}{\partial\theta}\right\} - E_0\left\{\frac{\partial \log f(x;\theta_0)}{\partial\theta}\right\} &= (\theta-\theta_0)E_0\left\{\frac{\partial^2 \log f(x;\tilde{\theta})}{\partial\theta^2}\right\} \\ &= -(\theta-\theta_0)E_0\left\{\frac{\partial \log f(x;\tilde{\theta})}{\partial\theta}\right\}^2 \end{aligned}$$

for some $\tilde{\theta}$ between θ and θ_0. Clearly, the different has the opposite sign of $\theta-\theta_0$. Hence the claim. $\diamondsuit$

Remark: the proof has assumed the mean value theorem can be applied to inside the expectation. This is allowed due to R2(iii).

Lemma 6.4 *Suppose θ_0 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) $\hat{\theta}_n \to \theta_0$ almost surely.

Proof

(i) Let ϵ be an arbitrarily small positive value satisfying

$$[\theta_0-\epsilon, \theta_0+\epsilon] \subset N(\theta_0).$$

By the law of large numbers, and condition R2 first bound, we have

$$\begin{aligned} n^{-1}S_n(\theta_0-\epsilon) &\to E_0\left\{\frac{\partial\{\log f(X;\theta_0-\epsilon)\}}{\partial\theta}\right\} > 0 \\ n^{-1}S_n(\theta_0+\epsilon) &\to E_0\left\{\frac{\partial\{\log f(X;\theta_0+\epsilon)\}}{\partial\theta}\right\} < 0. \end{aligned}$$

Thus, as $n\to\infty$, we have that

$$S_n(\theta_0-\epsilon) < 0 < S_n(\theta_0+\epsilon)$$

almost surely. By the intermediate value theorem, there must a $\hat{\theta}_n \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$ at which,

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

That is, there is a solution in any small enough neighborhood of θ_0 in the sense of almost surely as $n \to \infty$.

(ii) Because we can choose an arbitrarily small ϵ, it implies that this particular estimator sequence $\hat{\theta}_n \to \theta_0$ almost surely. $\diamondsuit$

This result is not equivalent to the consistency of MLE even for this special case. The MLE is defined as the global maximum, not merely the solution to the score function. However, the result implies the consistency of MLE for some special distribution families.

Corollary 6.1 . *If $E_0[\partial\{\log f(X;\theta)\}/\partial\theta]$ is a monotone function of θ for all θ_0, then the MLE is strongly consistent.*

Remarks In some applications, the parameter is defined by

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

The function $g(x;\theta)$ may be vector valued and θ is also a vector. The estimator is then defined as the solution to

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

The consistency is somehow implied in special cases using the same proof.

6.2 Asymptotic Normality of MLE after consistency

Our discussion remains focused on the one-dim situation.

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

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

Note that this conclusion also relies on the true parameter value is an interior point. This is because the parameter space is an open interval.

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

$$S_n(\theta_0) = S_n(\hat{\theta}) + S_n'(\tilde{\theta})(\theta_0 - \hat{\theta})$$

where $\tilde{\theta}$ is a parameter value between θ_0 and $\hat{\theta}$.

By one of the lemmas, we have

$$n^{-1} S_n'(\tilde{\theta}) \to -I(\theta_0)$$

the Fisher information almost surely. In addition, the classical central limit theorem can be applied to obtain

$$n^{-1/2} S_n(\theta_0) \to N(0, I(\theta_0)).$$

Thus, by Slutzky's theorem, we find

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} S_n(\theta_0)/I(\theta_0) + o_p(1) \to N(0, I^{-1}(\theta_0))$$

in distribution as $n \to \infty$.

6.3 Asymptotic chisquare of LRT

Let consider the simplest case when $H_0 = \{\theta_0\}$. In this case, the LRT statistic

$$R_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\}.$$

Remember, the MLE is assumed consistent. Thus, it is within an infinitesimal neighborhood of θ_0.

Applying Taylor's expansion, we have

$$\ell_n(\theta_0) = \ell_n(\hat{\theta}) + \ell_n'(\hat{\theta})(\theta_0 - \hat{\theta}) + (1/2)\ell_n''(\tilde{\theta})(\theta_0 - \hat{\theta})^2.$$

However, being MLE, $\hat{\theta}$ makes $\ell_n'(\hat{\theta}) = 0$. In addition, being consistent, we find

$$n^{-1}\ell_n''(\tilde{\theta}) = n^{-1}\ell_n''(\theta_0) + o_p(1) = -I(\theta_0) + o_p(1).$$

Hence, we find

$$R_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} = \{nI(\theta_0) + o_p(n)\}(\theta_0 - \hat{\theta})^2.$$

Recall that

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} S_n(\theta_0)/I(\theta_0) + o_p(1)$$

we get

$$R_n = \{I(\theta_0)\}^{-1}\{n^{-1/2} S_n(\theta_0)\}^2 + o_p(1).$$

Because $n^{-1/2} S_n(\theta_0) \to N(0, I(\theta_0))$, we find

$$R_n \to \chi^2_1$$

in distribution.

Chapter 7

Likelihood with multi-dimensional parameters

Consider the situation where we have a set of iid observations from a parametric family $\{f(x;\theta) : \theta \in \Theta \subset R^d\}$ for some positive integer d. The log likelihood function remains the same as

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

Note that the dimension of X is not an issue here. The score function is still

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

but we should regard it as a vector.

The regularity conditions are the same though sometimes we should interpret them as "element wise":

R0 the parameter space of θ is an open set of $\mathbb{R}^d$

R1 $f(x;\theta)$ is differential to order thee 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 $N(\theta_0)$,

$$(i) \quad \left|\frac{\partial f(x;\theta)}{\partial \theta}\right| \le g(x);$$
$$(ii) \quad \left|\frac{\partial^2 f(x;\theta)}{\partial \theta^2}\right| \le g(x);$$
$$(iii) \quad \left|\frac{\partial^3 \log f(x;\theta)}{\partial \theta^3}\right| \le H(x)$$

hold for all x, and

$$\int g(x)dx < \infty; \quad E_\theta\{H(X)\} < \infty.$$

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

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

This one must be interpreted as positive-definited.

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.

All conditions are stated as for all x. An except over a 0-measure set of x is allowed, as long as it is the same for all θ.

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

$$E_\theta \left\{\frac{\partial \log f(x;\theta)}{\partial \theta}\right\} = 0.$$

(2) Under regularity conditions, we have

$$E_\theta \left\{\frac{\partial \log f(x;\theta)}{\partial \theta}\right\}^2 = -E_\theta \left\{\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right\}$$

Use matrix interpretation.

The proof remains the same as the one for one-dim θ.

Theorem 7.1 *Suppose θ_0 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) $\hat{\theta}_n \to \theta_0$ almost surely.

Proof. The high dimensional case is slightly different.

(i) Let ϵ be a small enough positive number. Consider a θ^* value such that $\|\theta^* - \theta_0\| = \epsilon$. That is, θ^* is on the ball centred at θ_0 with radius ϵ. We aim to show that almost surely,

$$\ell_n(\theta^*) < \ell_n(\theta_0) \qquad (*)$$

simultaneously for all such θ.

If (*) is true, it implies that $\ell_n(\theta)$ has a local maximum within this ball. Because the likelihood function is smooth, the derivative at this local maximum is 0. Hence, conclusion (i) is true.

Is (*) true? By Taylor's series, we have

$$\ell_n(\theta^*) = \ell_n(\theta_0) + \{\ell_n'(\theta_0)\}^T(\theta^* - \theta_0) + \frac{1}{2}(\theta^* - \theta_0)^T\ell_n''(\tilde{\theta})(\theta^* - \theta_0)$$

for some $\tilde{\theta}$ in the ϵ-ball.

It is known that $\ell_n'(\theta_0) = O_p(n^{1/2})$. In addition, we have

$$n^{-1}\ell_n''(\theta_0) \to -I(\theta_0)$$

almost surely. Here $I(\theta_0)$ is the Fisher Information which is positive definite by R3. We claim without a proof that

$$\sup_{\theta^*} n^{-1}|\ell_n''(\tilde{\theta}) - \ell_n''(\theta_0)| = \epsilon C$$

for some C not random nor depend on θ^* and so on.

These assessments lead to

$$\ell_n(\theta^*) - \ell_n(\theta_0) = \{\ell_n'(\theta_0)\}^T(\theta^* - \theta_0) - \frac{1}{2}(\theta^* - \theta_0)^T I(\theta_0)(\theta^* - \theta_0) + O(n\epsilon^3).$$

Roughly, the first term is of size $n^{1/2}$, the second is $-n\epsilon^2$ and the remainder is $n\epsilon^3$. Thus, the over all size is determined by $-n\epsilon^2$ which is negative. This completes the proof of (i).

(ii) is a direct consequence of (i). ♢

The same remark is valid. This result is not equivalent to the consistency of MLE even for this special case. There exists a proof of the consistency of MLE based on much more relaxed conditions. However, the proof takes too much time to present.

7.1 Asymptotic normality of MLE after the consistency is established

The asymptotic normality for multidimensional θ stays the same.

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

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

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

$$S_n(\theta_0) = S_n(\hat{\theta}) + S_n'(\tilde{\theta})(\theta_0 - \hat{\theta})$$

where $\tilde{\theta}$ is a parameter value between θ_0 and $\hat{\theta}$. (This is not exactly true but somehow accepted by most. A valid proof takes too long).

By one of the lemmas, we have

$$n^{-1}S_n'(\tilde{\theta}) \to -I(\theta_0)$$

the Fisher information almost surely. In addition, the classical multivariate central limit theorem can be applied to obtain

$$n^{-1/2}S_n(\theta_0) \to N(0, I(\theta_0)).$$

Thus, by Slutzky's theorem, we find

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2}S_n(\theta_0)/I(\theta_0) + o_p(1) \to N(0, I^{-1}(\theta_0))$$

in distribution as $n \to \infty$.

7.2 Asymptotic chisquare of LRT for composite hypotheses

Let us still consider the simplest case when $H_0 = \{\theta_0\}$ that is an interior point of Θ. The alternative is $\theta \neq \theta_0$.

In this case, the LRT statistic

$$R_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\}.$$

Remember, the MLE is assumed consistent. Thus, it is within an infinitesimal neighborhood of θ_0.

Applying Taylor's expansion, we have

$$\ell_n(\theta_0) = \ell_n(\hat{\theta}) + \{\ell_n'(\hat{\theta})\}^T(\theta_0 - \hat{\theta}) + (1/2)(\theta_0 - \hat{\theta})^T\{\ell_n''(\tilde{\theta})\}(\theta_0 - \hat{\theta}).$$

However, being MLE, $\hat{\theta}$ makes $\ell_n'(\hat{\theta}) = 0$. In addition, being consistent, we find

$$n^{-1}\ell_n''(\tilde{\theta}) = n^{-1}\ell_n''(\theta_0) + o_p(1) = I(\theta_0) + o_p(1).$$

Hence, we find

$$R_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} = n(\theta_0 - \hat{\theta})^T\{I(\theta_0) + o_p(1)\}(\theta_0 - \hat{\theta}).$$

Recall that

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2}I^{-1}(\theta_0)S_n(\theta_0) + o_p(1)$$

we get

$$R_n = n^{-1}S_n^T(\theta_0)I^{-1}(\theta_0)S_n(\theta_0) + o_p(1).$$

Because $n^{-1/2}S_n(\theta_0) \to N(0, I(\theta_0))$, we find

$$R_n \to \chi_d^2$$

in distribution.

Remark: d is the dimension difference between H_0 and H_1.

Counter Example (Skipped, to be introduced later) Suppose that we have an iid sample of size n from

$$(1-\gamma)N(0,1) + \gamma N(2,1)$$

where γ is the mixing proportion.

We would like to test the hypothesis $H_0 : \gamma = 0$ versus $H_1 : \gamma > 0$.

The log likelihood function is given by

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

We have

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

At $\gamma = 0$, we find

$$\ell_n'(0) = \sum_{i=1}^{n} \{\exp(-2(x_i - 2)) - 1\}$$

which has 0-expectation under H_0. According to CLT, we find

$$P(\ell_n'(0) > 0) \to 0.5$$

as $n \to \infty$. It is clear that $\ell_n'(\gamma)$ is a decreasing function over $\gamma > 0$. Thus, when $\ell_n'(0) < 0$, we get $\ell_n'(\gamma) < 0$.

Two facts imply that if the data are generated from H_0, and we look for MLE in general, we would find

$$P(\hat{\gamma} = 0) \to 0.5.$$

Case I: when $\ell_n'(0) \leq 0$, we have $\hat{\gamma} = 0$. This further leads to

$$R_n = 2\{\ell_n(\hat{\gamma}) - \ell_n(0)\} = 0.$$

Case II: when $\ell_n'(0) > 0$, we have $\hat{\gamma} > 0$. It solves the equation

$$\sum_{i=1}^{n} \frac{\exp(-2(x_i - 2)) - 1}{1 + \gamma[\exp(-2(x_i - 2)) - 1]} = 0.$$

For brevity, let us assume the solution is at a small neighborhood of $\gamma = 0$. Thus, the above equation is approximated by

$$\sum_{i=1}^{n} \{\exp(-2(x_i - 2)) - 1\} - \gamma \sum_{i=1}^{n} \{\exp(-2(x_i - 2)) - 1\}^2 + o_p(n) = 0.$$

This leads to

$$\hat{\gamma} = \frac{\sum_{i=1}^n \{\exp(-2(x_i-2))-1\}}{\gamma \sum_{i=1}^n \{\exp(-2(x_i-2))-1\}^2} + o_p(n^{-1/2}).$$

Consequently,

$$\ell_n(\hat{\gamma}) = \frac{[\sum_{i=1}^n \{\exp(-2(x_i-2))-1\}]^2}{\gamma \sum_{i=1}^n \{\exp(-2(x_i-2))-1\}^2} + o_p(1).$$

Combining two cases, we can unify the expansion to

$$\ell_n(\hat{\gamma}) = \frac{\{[\sum_{i=1}^n \{\exp(-2(x_i-2))-1\}]^+\}^2}{\gamma \sum_{i=1}^n \{\exp(-2(x_i-2))-1\}^2} + o_p(1).$$

As $n \to \infty$, the limiting distribution is given by that of

$$(Z^+)^2$$

which is often denoted as

$$0.5\chi_0^2 + 0.5\chi_1^2.$$

Morale of this example: The null model H_0 is not an interior point of the parameter space. This invalidates the result obtained under regularity condition.

In many applications, the i.i.d. assumption is violated. The regularity conditions are no-longer sensible. Yet particularly in biostatistics applications, the users still regard the MLEs asymptotically normal, and the likelihood ratio statistics asymptotically chisquare. Often, they are not so wrong. At the same time, it is a worry-some trend that our scientific claims are built on a less and less solid foundation.

I hope that these lectures help you to get a sense of when the "chisquare" distribution is valid. In addition, you are able to rigorously establish whatever conclusions needed in various applications.

7.3 Asymptotic chisquare of LRT: one-step further

Write $\theta^T = (\theta_1^T, \theta_2^T)$ so that θ_1 is a vector of length d and θ_2 is a vector of length k. The superscript T is to make all vectors column vector.

Consider the composite null hypothesis H_0 that

$$\theta_1 = 0$$

in vector sense. The alternative is $H_1 : \theta_1 \neq 0$.

This time, we denote $\theta_0^T = (\theta_{10}^T, \theta_{20}^T)$ as the true vector value of the parameter that generated the data $x_1, \ldots, x_n$. In addition, this θ_0 is one of the parameter vectors in H_0. We assume that θ_0 is an interior point of the parameter space as usual. **This is part of the regularity conditions** to ensure the validity of the asymptotic result to be introduced.

We use $\hat{\theta}$ as the MLE of θ without placing any restrictions on the range of θ. We use $\hat{\theta}_0$ as the MLE or the maximum point of θ in the space of H_0. The consistency results discussed before ensure that both $\hat{\theta}$ and $\hat{\theta}_0$ almost surely converge to θ_0. When notationally necessary, they will be partitioned into $(\hat{\theta}_1^T, \hat{\theta}_2^T)^T$ and $(\hat{\theta}_{01}^T, \hat{\theta}_{02}^T)^T$ respectively. Of course, we have $\hat{\theta}_{01} = 0$.

7.3.1 Some notational preparations

The Fisher information with respect to θ is now a matrix. We denote

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

This matrix can be partitioned into 4 blocks:

$$\mathbb{I}_{ij}(\theta) = E\left[\left\{\frac{\partial \log f(X;\theta)}{\partial \theta_i}\right\}\left\{\frac{\partial \log f(X;\theta)}{\partial \theta_j}\right\}^T\right] = -E\left\{\frac{\partial^2 \log f(X;\theta)}{\partial \theta_i \partial \theta_j^T}\right\}.$$

for $i.j = 1, 2$. In other words, we have

$$\mathbb{I}(\theta) = \left\{\begin{array}{cc} \mathbb{I}_{11}(\theta) & \mathbb{I}_{12}(\theta) \\ \mathbb{I}_{21}(\theta) & \mathbb{I}_{22}(\theta) \end{array}\right\}$$

The regularity conditions make $\mathbb{I}(\theta)$ positive definite which implies both $\mathbb{I}_{11}$ and $\mathbb{I}_{22}$ are positive definite. The expectations are understood as taken with the distribution of X is given by $f(x;\theta)$. Namely, the same parameter value for operation E and the subject.

The score function is now also a vector. Let us write

$$S_n^T(\theta) = (S_{n1}^T, S_{n2}^T) = \sum_{i=1}^n \left(\frac{\partial \log f(X;\theta)}{\partial \theta_1^T}, \frac{\partial \log f(X;\theta)}{\partial \theta_2^T} \right).$$

The subscripts stand for transpose and they make every vector a low vector. They do not have other practical purposes.

Matrix result. Let $\mathbb{I}_{11,2} = \mathbb{I}_{11} - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}\mathbb{I}_{21}$. It is laborious to verify that

$$\mathbb{I}^{-1}(\theta) = \begin{pmatrix} I & 0 \\ -\mathbb{I}_{22}^{-1}\mathbb{I}_{21} & I \end{pmatrix} \begin{pmatrix} \mathbb{I}_{11,2}^{-1} & 0 \\ 0 & \mathbb{I}_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\mathbb{I}_{12}\mathbb{I}_{22}^{-1} \\ 0 & I \end{pmatrix}$$

where I itself is an identity matrix of proper size.

Based on matrix theory, or direct verification, we have

$$x^T\mathbb{I}^{-1}x = (x_1^T - x_2^T\mathbb{I}_{22}^{-1}\mathbb{I}_{21})\mathbb{I}_{11,2}^{-1}(x_1 - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}x_2) + x_2^T\mathbb{I}_{22}^{-1}x_2$$

for any vector x of proper length and partition.

Applying the matrix result to S_n and $\mathbb{I}$, we find

$$S_n^T\mathbb{I}^{-1}(\theta)S_n = (S_{n1}^T - S_{n2}^T\mathbb{I}_{22}^{-1}\mathbb{I}_{21})\mathbb{I}_{11,2}^{-1}(S_{n1} - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}S_{n2}) + S_{n2}^T\mathbb{I}_{22}^{-1}S_{n2}.$$

It is known that $n^{-1/2}S_n$ is asymptotically normal with covariance matrix given by $\mathbb{I}(\theta)$. This implies that

$$n^{-1/2}(S_{n1} - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}S_{n2})$$

is asymptotically normal with covariance matrix $\mathbb{I}_{11,2}$. Hence, the first term

$$n^{-1}(S_{n1}^T - S_{n2}^T\mathbb{I}_{22}^{-1}\mathbb{I}_{12})\mathbb{I}_{11,2}^{-1}(S_{n1} - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}S_{n2}) \to \chi_d^2$$

where d is the dimension of θ_1.

Let us now use these results. The LRT statistic now becomes

$$R_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\hat{\theta}_0)\} = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} - 2\{\ell_n(\hat{\theta}_0) - \ell_n(\theta_0)\}.$$

For the first one, we apparently have

$$R_{n1} = n^{-1}S_n^T(\theta_0)\{\mathbb{I}^{-1}(\theta_0)\}S_n(\theta_0) + o_p(1).$$

Based on the same principle, we have

$$R_{n2} = n^{-1} S_{n2}^T(\theta_0)\{\mathbb{I}_{22}^{-1}(\theta_0)\}S_{n2}(\theta_0) + o_p(1).$$

Combining two expansions, we find

$$R_n = n^{-1}[S_n^T(\theta_0)\{\mathbb{I}^{-1}(\theta_0)\}S_n(\theta_0) - S_{n2}^T(\theta_0)\{\mathbb{I}_{22}^{-1}(\theta_0)\}S_{n2}(\theta_0)] + o_p(1).$$

With the preparational results, we have

$$R_n \to \chi_d^2$$

in distribution as $n \to \infty$. ♢

7.4 The most general case: final step

The null hypothesis discussed already can be expressed as

$$H_0 : A\theta = 0$$

with specific matrix $A = \text{diag}\{1, 1, \ldots, 1, 0, 0, \ldots, 0\}$. Denote the number of 1's as k and number of 0's as d.

We can easily generalize this result to be applicable to any matrix A. It is well known in linear algebra that the solution set of $A\theta = 0$ is a linear space. That is, having $A\theta = 0$ is the same as

$$\theta = \lambda_1\xi_1 + \cdots \lambda_d\xi_d$$

where ξ_j are a basis for solutions to $A\theta = 0$.

In the space of λ, H_0 becomes

$$\lambda = (\lambda_1, \cdots, \lambda_d, \lambda_{d+1} = 0, \ldots, \lambda_{k+d} = 0)$$

which is the same as the special case we have discussed. Namely, the $R_n \to \chi_d^2$ remains solid.

Most generally, assume the parameter space is a subset of R^{d+k}. The composite hypothesis is either expressed as

$$R(\theta) = 0 \qquad \textit{Hypothesis form I}$$

for a continuously differentiable vector valued function R or as

$$\theta = g(\lambda) \quad \textit{Hypothesis form II}$$

for a continuously differentiable $g(\cdot)$.

When it is in form I, denote the rank of the differential matrix at θ_0 as d. Hence, it puts d constraints on the parameter in a small neighborhood of θ_0. After which, based on inverse function theorem, there exists a smooth function g such that the solution to $R(\theta) = 0$ can be written as $\theta = g(\lambda)$ where the dimension of λ is k, in a neighborhood of θ_0.

In both cases, we may interpret that the null hypothesis sets k elements in θ to 0 and leave d of them free. The same proof presented earlier makes $R_n \to \chi^2_d$ in distribution.

The regularity condition in part R0 has to be revised.

- true parameter value θ_0 is an interior point of Θ.
- There is a neighborhood of θ_0, over which $R(\theta) = 0$ admits a smooth solution $\theta = g(\lambda)$.
- There are neighborhoods of λ_0 and θ_0 respectively such that $g(\lambda)$ is differentiable with full rank derivative matrix.

7.5 Statistical application of these results

The whole purpose of proving $R_n \to \chi^2_d$ is to test hypothesis in applications.

As the size of R_n represents the departure from the null model, the test based on likelihood ratio is mathematically given by

$$\phi(x) = I(R_n \geq c)$$

and this c will be chosen as $\chi_d(1-\alpha)$ for a size-α test.

Example 7.1 *Suppose we have an i.i.d. sample from a trinomial distribution. That is, each outcome of a trial is one of three types. Let the corresponding probabilities of occurrence be p_1, p_2, p_3. Clearly, $p_1 + p_2 + p_3 = 1$.*

After n trials, we have n_1, n_2, n_3 observations of three types. The log likelihood function is given by

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

The maximum likelihood estimator of these parameters are given by

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

for $j = 1, 2, 3$.

(i) Consider the test for $p_j = p_{j0} \neq 0$, $j = 1, 2, 3$ versus $p_j \neq p_{j0}$ for at least one of $j = 1, 2, 3$. The likelihood ratio test statistic is apparently given by

$$\begin{aligned} R_n &= 2n_1 \log(\hat{p}_1/p_{10}) + 2n_2 \log(\hat{p}_2/p_{20}) + 2n_3 \log(\hat{p}_3/p_{30}) \\ &= 2n \sum \hat{p}_j \log(\hat{p}_j/p_{j0}). \end{aligned}$$

According to our theorem on LRT, when $n \to \infty$, R_n is approximately χ^2_2 distributed under the null model.

The MLEs under this model are consistent and asymptotically normal. We have $\hat{p}_j = p_{j0} + O_p(n^{-1/2})$. Therefore, we have

$$\begin{aligned} \log(\hat{p}_j/p_{j0}) &= -\log\{1 - (\hat{p}_j - p_{j0})/\hat{p}_j\} \\ &= (\hat{p}_j - p_{j0})/\hat{p}_j + (1/2)(\hat{p}_j - p_{j0})^2/\hat{p}_j^2 + O_p(n^{-3/2}). \end{aligned}$$

Hence,

$$\begin{aligned} R_n &= n \sum_j (\hat{p}_j - p_{j0})^2/\hat{p}_j + O_p(n^{-1/2}) \\ &= n \sum_j (\hat{p}_j - p_{j0})^2/p_{j0} + O_p(n^{-1/2}). \end{aligned}$$

The leading term is the famous Pearson's chisquare test statistics. It is often used for "goodness of fit" test.

Another version of this test is an assignment problem. The result remains similar if there are more than 3 categories.

Chapter 8

Wald and Score tests

8.1 Wald test

Under regularity conditions, we have show that the MLE of θ is asymptotically normal. That is,

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

as the sample size $n \to \infty$ based on i.i.d. observations. When the MLE is consistent, we also have

$$\mathbb{I}(\hat{\theta}) \to \mathbb{I}(\theta)$$

in probability. These two results imply

$$n(\hat{\theta}_n - \theta)^\tau \mathbb{I}(\hat{\theta})(\hat{\theta}_n - \theta) \to \chi^2_d$$

where d is the dimension of θ. Based on this result, a test for the simple null hypothesis of $H_0 : \theta = \theta_0$ can be based on

$$W_n(\theta_0) = n(\hat{\theta}_n - \theta_0)^\tau \mathbb{I}(\theta_0)(\hat{\theta}_n - \theta_0).$$

We reject H_0 when $W_n(\theta_0) \geq \chi^2_d(1-\alpha)$ in favour of the generic alternative $H_1 : \theta \neq \theta_0$.

Because replacing $\mathbb{I}(\theta_0)$ with any of its consistent estimator does not change the limiting distribution, it is possible to construct many asymptotically equivalent tests:

1. We may replace $\mathbb{I}(\theta_0)$ by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i;\theta)}{\partial\theta^2}\bigg|_{\theta=\theta_0}.$$

   This is nice if we are too lazy to find the analytical form of the Fisher information matrix.

2. Another related choice is

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i;\theta)}{\partial\theta^2}\bigg|_{\theta=\hat{\theta}}.$$

   This quantity is often computed when the MLE is obtained by iterative methods such as Newton-Raphson.

3. The third choice is $\mathbb{I}(\hat{\theta})$.

When the regularity conditions are satisfied, it also makes sense to replace $\mathbb{I}(\theta_0)$ by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial \log f(x_i;\theta)}{\partial\theta}\right\}\left\{\frac{\partial \log f(x_i;\theta)}{\partial\theta}\right\}^{\tau}\bigg|_{\theta=\theta_0}.$$

Unlike the earlier choice, this quantity is always positive (or non-negative) definite.

The above discussion works for the simple null hypothesis. Suppose the vector θ can be written as $\theta^\tau = (\theta_1^\tau, \theta_2^\tau)$. To fix the idea, we denote the dimension of θ as $d+k$ and the partition for θ_1 and θ_2 are d and k. Consider the problem of testing $H_0 : \theta_1 = \theta_{10}$ against $H_0 : \theta_1 \neq \theta_{10}$.

Let $\hat{\theta}_n$ be the MLE over the whole parameter space Θ and $\hat{\theta}_n^\tau = (\hat{\theta}_{1n}^\tau, \theta_{2n}^\tau)$ be the corresponding decomposition. Because $\hat{\theta}_n$ is asymptotically normal, so is any of its sub-vector (or linear combination). This implies

$$\sqrt{n}(\hat{\theta}_{1n} - \theta_{10}) \xrightarrow{d} N(0, \mathbb{I}^{11}(\theta))$$

where $\mathbb{I}^{11}$ is upper-left corner block sub matrix of $\mathbb{I}^{-1}$ corresponding the θ_1. This leads to a sensible test statistic

$$W_n(\theta_{10}) = n(\hat{\theta}_{1n} - \theta_{10})^\tau \{\mathbb{I}^{11}(\hat{\theta})\}^{-1}(\hat{\theta}_{1n} - \theta_{10}).$$

Clearly, we have

$$W_n(\theta_{10}) \to \chi^2_d$$

with d being the dimension of θ_1. A test of approximate size α is therefore given by

$$\phi(X) = \begin{cases} 1 & \text{when } W_n(\theta_{10}) \le \chi^2_d(1-\alpha) \\ 0 & \text{otherwise} \end{cases}$$

More generally, suppose the null hypothesis is specified in the form of

$$H_0 : \quad \varphi(\theta) = 0$$

where $\varphi(\cdot)$ take vector values of dimension d. The overall dimension of θ is $d + k$. Note that when $\varphi(\theta) = \theta_1 - \theta_{10}$ with some known value θ_{10}, then this H_0 is the same as the one we have just discussed. Naturally, let the alternative hypothesis be $H_1 : \varphi(\theta) \neq 0$.

For this problem, we may define

$$W_n = n\varphi^\tau(\hat{\theta})\{\varphi'(\hat{\theta})\mathbb{I}^{-1}(\hat{\theta})\varphi'(\hat{\theta})^\tau\}^{-1}\varphi(\hat{\theta}).$$

It can be shown that

$$W_n \to \chi^2_d$$

and d is the rank of $\varphi'(\theta)$. Clearly, an approximate size α can be similarly constructed based on this W_n.

8.2 Score Test

We have seen that under regularity conditions,

$$E\{\partial \log f(X;\theta)/\partial\theta\} = 0$$

where the expectation is taken under the assumption that the distribution of X is given by $f(x;\theta)$.

Thus, when we test for $H_0 : \theta = \theta_0$, the value of the score function

$$S_n(\theta_0) = \sum_{i=1}^{n} \partial \log f(X_i;\theta_0)/\partial\theta\}$$

is indicative of the validity of H_0.

Recall that $n^{-1/2}S_n(\theta_0)$ is asymptotically multivariate normal with asymptotic variance $\mathbb{I}(\theta_0)$. Let us define a test statistic to be

$$T_n = S_n^\tau(\theta_0)\{n\mathbb{I}(\theta_0)\}^{-1}S_n(\theta_0).$$

The limiting distribution of T_n is chisquare with d degrees of freedom where d is the dimension of θ.

Base on this result, a score test of approximate size α is given by

$$\phi(X) = \begin{cases} 1 & \text{when } T_n \leq \chi_d^2(1-\alpha) \\ 0 & \text{otherwise} \end{cases}$$

Unlike the likelihood ratio test or the Wald test, this statistic does not ask us to compute MLE of θ. We do need to compute the Fisher information matrix and its inversion.

Similar to Wald test, let us now consider the null hypothesis $H_0 : \theta_1 = \theta_{10}$. The dimension of θ is $d + k$. This means the second part of θ vector is unspecified under H_0. Let $\hat{\theta}_0$ be the MLE under H_0. Let

$$S_{n1}(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta_1}$$

which was defined earlier too. This is a column vector of length d. If the same asymptotic techniques are used here, we will find that

$$n^{-1/2}S_{n1}(\hat{\theta}_0)$$

is asymptotically multivariate normal with mean 0 and variance matrix $\mathbb{I}_{11,2}(\theta^*)$ under the null hypothesis. This θ^* stands for the true parameter value and $\mathbb{I}_{11,2} = \mathbb{I}_{11} - \mathbb{I}_{12}\mathbb{I}_{22}^{-1}\mathbb{I}_{21}$ was also defined before.

Apparently, these notes on asymptotic results lead to the conclusion

$$T_n = S_{n1}^\tau(\hat{\theta}_0)\{n\mathbb{I}_{11,2}(\hat{\theta}_0)\}^{-1}S_{n1}(\hat{\theta}_0) \to \chi_d^2$$

as $n \to \infty$. This time, d is the dimension of θ_1. A test can be constructed the same way as earlier.

Finally, consider the null hypothesis specified by $H_0 : \varphi(\theta) = 0$ where φ is a smooth function. Under regularity conditions, and in most of applied

problems, we may equivalently write H_0 as $\theta = g(\lambda)$ for some smooth function g with new parameter setting λ.

In this case, we may obtain the MLE of λ as the maximum point of $\ell(g(\lambda))$. Denote it as $\hat{\lambda}$. Next, we redefine the score statistic to be

$$T_n = S_n^T(g(\hat{\lambda}))\{n\mathbb{I}(g(\hat{\lambda}))\}^{-1}S_n(g(\hat{\lambda})).$$

Under regularity conditions, we still have

$$R_n \to \chi_d^2.$$

This d is the dimension of θ minus the dimension of λ.

8.3 Remarks

Three tests are asymptotically equivalent.

The discussions on the use of $\mathbb{I}(\hat{\theta})$, $\mathbb{I}(\theta)$ or their estimators in various forms in Wald test section is generally applicable.

Chapter 9

Tests under normality

9.1 One-sample problem under Normality assumption

Suppose we have a random sample $x_1, \ldots, x_n$ of size n from $N(\theta, \sigma^2)$. We adopt common notations: sample mean $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and sample variance $s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$; Let

$$T_n = \frac{\sqrt{n}(\bar{x} - \theta_0)}{s_n}$$

for some given θ_0 value. The well known test for $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ of size α is given by

$$\phi(x) = \begin{cases} 1 & \text{when } T_n \geq t_{n-1,1-\alpha}; \\ 0 & \text{otherwise} \end{cases}$$

where $t_{n-1,1-\alpha}$ is the upper $1-\alpha$ quantile of the t-distribution with $n-1$ degrees of freedom.

The well know test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ of size α is given by

$$\phi(x) = \begin{cases} 1 & \text{when } |T_n| \geq t_{n-1,1-\alpha/2}; \\ 0 & \text{otherwise} \end{cases}$$

This is the famous two-sided t-test.

Both tests are convenient to use, have nice properties. Yet after having studied the "UMP" theory, we may question their "optimality". We will not prove anything but put a few classical theorems here. Their truthfulness cannot be lectured here. These interested are referred to our reference textbooks. Even better, you may practice your technical skill through an attempt to prove to disprove these results.

Theorem 9.1 *Suppose we have an iid sample from a distribution with density function from an exponential family*

$$f(x;\theta,\lambda) = \exp\{\theta U(x) + \lambda T(x) + A(\theta,\lambda)\}$$

with respect to some σ-finite measure. The parameter θ is one-dimensional, while λ can be multi-dimensional.

(i) Suppose that $V = h(U,T)$ is independent of T when $\theta = \theta_0$. In addition, for each t, h is an increasing function in u. Then the test defined as follows

$$\phi(v) = \begin{cases} 1 & \text{when } v > k; \\ c & \text{when } v = k \\ 0 & \text{otherwise} \end{cases}$$

satisfying $E\{\phi(V);\theta_0\} = \alpha$ is an UMP unbiased test for $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$.

(ii) Assume the same conditions as in (i) and that

$$h(u,t) = a(t)u + b(t) \quad \text{with} \quad a(t) > 0.$$

Let us define

$$\phi(v) = \begin{cases} 1 & \text{when } v > k_1 \text{ or } \quad v < k_2; \\ c_j & \text{when } v = k_j, \quad j = 1,2; \\ 0 & \text{otherwise} \end{cases}$$

for some constants such that $E\{\phi(V);\theta_0\} = \alpha$ and $E\{V\phi(V);\theta_0\} = \alpha E\{V;\theta_0\}$. This test is an UMP unbiased test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

According to this result, the UMPU tests have the same format to the ones we obtained in the absence of nuisance parameters. This is the case under the "exponential family" setting.

In the next example, we use the above theorem to construct a test for hypotheses about the variance σ^2 under normal model.

Example 9.1 *Suppose we have an iid sample from $N(\xi, \sigma^2)$. The joint function can be written as*

$$f(x; \xi, \sigma^2) = \exp\{\theta U(x) + \lambda T(x) + A(\theta, \lambda)\}$$

with $\theta = -1/(2\sigma^2)$, $\lambda = (n\xi)/\sigma^2$; and $U(x) = \sum x_i^2$, $T(x) = \bar{x}$.

Let

$$h(U, T) = U - nT^2 = \sum (x_i - \bar{x})^2.$$

It is seen that for any given value of σ^2, $h(U, T)$ is independent of T. Thus, an UMP test of size α for $H_0 : \sigma \leq \sigma_0$ against $H_1 : \sigma > \sigma_0$ is given by

$$\phi(U) = \begin{cases} 1 & \text{when } U > k; \\ 0 & \text{otherwise} \end{cases}$$

Because U/σ_0^2 has chisquare distribution with $n - 1$ degrees of freedom. The size of k is therefore the $1 - \alpha$th quantile of this distribution.

In the next example, we change the roles of σ and μ (in terms of notation) and therefore U and T to construct a UMPU about the size of population mean.

Example 9.2 *Suppose we have an iid sample from $N(\xi, \sigma^2)$. The joint function can be written as*

$$f(x; \xi, \sigma^2) = \exp\{\theta U(x) + \lambda T(x) + A(\theta, \lambda)\}$$

with $\lambda = -1/(2\sigma^2)$, $\theta = (n\xi)/\sigma^2$; and $T(x) = \sum x_i^2$, $U(x) = \bar{x}$.

This time, we find

$$V = h(U, T) = \frac{U}{\sqrt{T - nU^2}} = \frac{\bar{X}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

is independent of $T(x)$ when $\xi = 0$. See the remark after this example.

It is easily seen that V is an increasing function of U given T. However, it is not linear in U. Another step is needed to give the result which will be left as an assignment problem.

Thus, the UMPU test for $H_0 : \xi = 0$ versus $H_0 : \xi \neq 0$ is given by

$$\phi(V) = I(|V| > k)$$

and make $E\{\phi(V); \theta_0\} = \alpha$ and $E\{V\phi(V); \theta_0\} = \alpha E\{V; \theta_0\}$. The solution is certainly the famous t-test.

Some normalization steps are needed but omitted in order to put everything into the exact format of t-test.

Remark: When $\xi = 0$ is given, $T(x)$ is complete and sufficient for σ^2. At the same time, the distribution of V is not dependent on σ. Thus, the classical theorem implies that they are independent.

In one of our assignment problem, we asked whether or not V is independent of the sample variance which is not $T(x)$. The answer is negative.

9.2 Two-sample problem under normality assumption

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be iid samples from $N(\xi, \sigma^2)$ and $N(\eta, \tau^2)$ respectively. Their joint density function is given by

$$f(X, Y; \xi, \eta, \sigma, \tau) = \exp\left\{-\frac{1}{2\sigma^2}\sum x_i^2 - \frac{1}{2\tau^2}\sum y_j^2 + \frac{m\xi}{\sigma^2}\bar{x} + \frac{n\eta}{\tau^2}\bar{y} - A(\xi, \eta, \sigma, \tau)\right\}.$$

Next, let us transform the parameter by

$$\theta = -\frac{1}{2\tau^2} + \frac{1}{2\Delta\sigma^2}$$

and

$$\lambda_1 = -\frac{1}{2\sigma^2}; \quad \lambda_2 = \frac{m\xi}{\sigma^2}; \quad \lambda_3 = \frac{n\eta}{\tau^2}$$

for some constant $\Delta > 0$. Let the corresponding sufficient statistics be

$$U = \sum_{j=1}^{n} Y_j^2; \quad T_1 = \sum_{i=1}^{m} X_i^2 + \frac{1}{\Delta}\sum_{j=1}^{n} Y_j^2; \quad T_2 = \bar{X}; \quad T_3 = \bar{Y}.$$

Test for equal variance Consider the test for $H_0 : \tau^2 \leq \sigma^2$ versus $H_1 : \tau^2 > \sigma^2$ This is the same as, with $\Delta = 1$,

$$H_0 : \theta \leq 0; \quad \text{versus} \quad H_1 : \theta > 0.$$

Define

$$V = h(U, T_1, T_2, T_3) = \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}{\sum_{i=1}^{m}(X_i - \bar{X})^2}.$$

It is seen that given $\theta = 0$, V has F-distribution which does not depend on any parameters. Thus, it is independent of the sufficient and complete statistic (T_1, T_2, T_3). One needs to pay special attention to the fact that when $\theta = 0$, the distribution family does not have other parameters anymore.

The independence makes the test based on V UMPU. That is, a UMPU test for $H_0 : \tau^2 \geq \sigma^2$ versus $H_1 : \tau^2 < \sigma^2$ is given by

$$\phi(V) = I(V > k)$$

and this k is chosen according to the F-distribution to make the size right.

Extension. By put Δ equaling other values, we obtain various variations. One may also easily get the F-test for the two-sided test. I am not so sure whether the UMPU test assigns $\alpha/2$ probability on two sides of the distribution of V.

9.3 Test for equal mean under equal variance assumption

We certainly know that the two-sample t-test will show up here. Consider the case where $\tau^2 = \sigma^2$. Under this assumption, the joint density of two samples is given by

$$f(X, Y; \xi, \eta, \sigma, \tau) = \exp\left\{-\frac{1}{2\sigma^2}\{\sum x_i^2 + \sum y_j^2\} + \frac{m\xi}{\sigma^2}\bar{x} + \frac{n\eta}{\sigma^2}\bar{y} - A(\xi, \eta, \sigma)\right\}.$$

Let

$$\theta = \frac{\eta - \xi}{(m^{-1} + n^{-1})\sigma^2};$$

and

$$\lambda_1 = \frac{m\xi + n\eta}{(m+n)\sigma^2}; \quad \lambda_2 = -\frac{1}{2\sigma^2}.$$

The sufficient statistics are

$$U = \bar{Y} - \bar{X}; \quad T_1 = m\bar{X} + n\bar{Y}; \quad T_2 = \sum_{i=1}^{m} X_i^2 + \sum_{j=1}^{n} Y_j^2.$$

Consider a test for $H_0 : \xi = \eta$ versus $H_1 : \xi \neq \theta$, we construct a statistic

$$V = \frac{\bar{Y} - \bar{X}}{\sqrt{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2}}$$

which is a function of U, T_1, T_2 and T_3. Its distribution, when $\xi = \eta$ does not depend on λ_j, $j = 1, 2, 3$. Thus, it serves the proper statistic for constructing UMPU.

A UMPU test is given by

$$\phi(V) = I(|V| > k)$$

with k satisfying $E\{\phi(V); \eta = \xi\} = \alpha$ and $E\{V\phi(V); \eta = \xi\} = \alpha E\{V; \eta = \xi\}$.

This is apparently the two-sample t-test. Here are a few missing technical steps. First, the denominator in V can be written as

$$T_2 - \frac{1}{m+n}T_1^2 - \frac{mn}{m+n}U^2.$$

This ensures that V is indeed a function of the required format.

The second is linearity of V in U given T. The linearity is not exactly true. However, V is a monotone function of W:

$$W = \frac{\bar{Y} - \bar{X}}{\sqrt{\sum x_i^2 + \sum y_j^2 - \frac{(\sum x_i + \sum y_j)^2}{m+n}}}.$$

So a test based on W would satisfy all conditions specified in the theorem. Two tests are, however, equivalent. The reason for using V instead of W lies in the fact that the distribution of V is clearly related to t, while the distribution of W is not "standard".

9.4 Test for equal mean without equal variance assumption

If $\sigma^2 = \tau^2$ is not assumed (or not known to be equal), there is no such a simple solution as UMPU test. This is so-called *Behrens-Fisher problem*.

In terms of searching for "optimal tests", one usually starts to place restrictions on the test: we require the test is "unbiased", "invariant", "similar" and so on.

With some considerations, it appears that a good test should reject the null hypothesis when

$$\frac{\bar{Y} - \bar{X}}{\sqrt{S_x^2 + S_Y^2}} \geq g(S_Y^2/S_X^2)$$

for some suitable function g. If the test is required to be unbiased, then only "pathological functions g" can have this property.

"Approximate solutions are available which provide tests that are satisfactory for all practical purposes". Among them, we probably know Welch's approximate t-test.

My summary: if one attempts to have an "optimal" test for $\xi = \eta$ without knowing $\sigma = \tau$ in two-sample problem, there may not be such a solution. If the "optimality" is not strictly observed, there are many sensible methods available.

Summary We have gone over most famous tests based on data from normal models. Due to time constrains, we purposely skipped several important cases. We did not go over the theorem based on which these tests are justified to have various optimality properties.

The optimality will go away if the data are not from distribution well approximated by normal. My experience indicates, however, that the two-sample t-test works really nicely even in very extreme situations.

Tests without making use of normality will be discussed next.

Chapter 10

Non-parametric tests

The methods we have discussed so far are based on the assumption that the data are generated i.i.d. from some regular parametric models. These methods become either inapplicable, or inferior if the data do not listen to our command: "assume they have this or that distribution".

One minor side–effect of "parametric assumption" is that these tests do not have the pre-scribed type-I error. For instant, the one-sample t-test may have a much larger type-I error when the data have Cauchy distribution (for testing the location parameter value). Sometimes, the performance of a test can still be very respectable. For instance, I find the two-sample t-test is very hard to beat in terms of having both accurate type-I error and good power. One has to subject this test to very weird data sets to make it look bad.

Even though some parametric tests are rather robust, there is a need of tests whose validity are not dependent on the correctness of the model assumption.

10.1 One-sample sign test.

Suppose we have an i.i.d. sample from some distribution whose c.d.f. is given by $F(x)$. The family that F belongs is not very important so we do not carefully specify it. In some applications, we may not have much information about it.

The hypothesis under consideration is

$$H_0 : F(u) \leq p_0$$

versus $H_1 : F(u) > p_0$ for some user-specified u and p_0.

Apparently, the key information from a single observation in this problem is whether $x_i > u$ or $x_i \leq u$. Consequently, we define

$$\Delta_i = I(x_i - u \leq 0)$$

for $i = 1, \ldots, n$.

If Δ_i, $i = 1, \ldots, n$ are the only data we observe, then $Y = \sum_{i=1}^{n} \Delta_i$ is sufficient for p, the unknown value of $F(u)$. The UMP test for H_0 versus H_1 has the form

$$\phi(Y) = I(Y > k) + cI(Y = k)$$

with proper choices of k and c for the sake of the test to have a pre-given size.

This test does not depend on the specific form of F, hence we call it a non-parametric test. The statistic Y is the number of observations with $x_i - u \leq 0$, which is the number of observations where the quantity has negative/positive sign. This test $\phi(Y)$ is subsequently called sign-test.

This test is UMP in general, rather than merely in the restricted sense as specified above. However, it may not be so interesting to seriously prove it.

10.2 Two-sample permutation test.

Consider a situation where we have one random sample $x_1, \ldots, x_m$ from F and another random sample $y_1, \ldots, y_n$ from G. It is of interest to test for $H_0 : F = G$ versus $H_1 : F \neq G$.

To have a meaningful discussion, assume both F and G are continuous. That is, the model space contains all continuous distribution functions. Denote the pooled sample as $z = \{x_1, \ldots, x_m, \} \cup \{y_1, \ldots, y_n\}$. Define a set of vectors to be

$$\pi(z) = \{(z_{i_1}, z_{i_2}, \ldots, z_{i_{m+n}}) : (i_1, \ldots, i_{m+n}) \text{ is a permutation of } (1, 2, \ldots, m+n)\}.$$

That is, $\pi(z)$ contains all vectors of dimension $m+n$ that are permutations of each other.

Let $\phi(X, Y)$ be a test such that

$$\frac{1}{(m+n)!} \sum_{(x,y)\in\pi(z)} \phi(x; y) = \alpha.$$

Then it is called a permutation test for a given significance level α.

At a first look, this definition does not seem to make much sense. Let us assume that the observed values are all different. That is, $x_i \neq x_j$, $y_i \neq y_j$ for any $i \neq j$; and $x_i \neq y_j$ for any i and j. In this case, once the pooled sample z is specified, $\pi(z)$ contains $(m+n)!$ vectors. Suppose the test function $\phi(x, y)$ does not involves randomization. Then this function decides which of $(m+n)!$ vectors are in the rejection region. Under the null hypothesis of $F = G$ and both distributions are continuous, given **the set z**, every member of $\pi(z)$ has probability $1/(m+n)!$ to occur. Hence, if no randomization is involved, and say $(m+n)! = 1000$, the permutation test selects 50 of them to form the rejection region of a test of size $\alpha = 0.05$. Note that there are no $m+n$ which make $(m+n)! = 1000$, so this assumption is for convenient illustration only.

The name is now sensible as the rejection region is formed by permuted observed vectors.

The key issue left in a permutation test is: which 50 out of 1000?

One intuitive proposal.

The question of which 50 depends on the "optimality requirement" and the potential alternative hypothesis. What direction of the departure do we care? Without such a direction, we can always find two samples differ significantly in one way or in another.

Consider the situation where the alternative is $H_1 : G(x) = F(x - \delta)$ for some $\delta > 0$. In statistics, we say $G(x)$ is obtained from F by a location shift. Under this alternative hypothesis, the samples from G are stochastically larger than the samples from F. Any statistics which tend to take larger values under H_1 is a suitable candidate.

Suppose $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ are two random samples respectively.

Let

$$T_{m+n} = n^{-1}\sum_{j=1}^{n} y_j - m^{-1}\sum_{i=1}^{m} x_i = \bar{y}_n - \bar{x}_m.$$

For each permuted $x_1, \ldots, x_m;\ y_1, \ldots, y_n.$,

$$x'_1, \ldots, x'_m;\quad y'_1, \ldots, y'_n$$

we compute

$$T'_{m+n} = n^{-1}\sum_{j=1}^{n} y'_j - m^{-1}\sum_{i=1}^{m} x'_i = \bar{y}'_n - \bar{x}'_m.$$

The observed T_{m+n} is one of $\binom{(m+n)}{n}$ possible outcomes denoted as T'_{m+n}. It makes sense to select the permutations which results in the largest values of T'_{m+n} to form the rejection region.

To carry out this test, we compute all $\binom{(m+n)}{n}$ possible values of T'_{m+n}. One of them is the observed value T_{m+n}. If the observed value is among the top $100\alpha\%$, we reject H_0 in favour of $H_1 : G(x) = F(x - \delta)$ for some $\delta > 0$.

In applications, if $(m + n)$ is large, computing all possible values is not feasible. Computer simulation may be used to compute only a random subset of them and get an accurate enough rank of T_{m+n}.

If $(m + n)$ is small, some T'_{m+n} may equal T_{m+n}. Continuity correction is often used. That is, each equaling T'_{m+n} value is counted as half is larger, another half is smaller than T_{m+n}.

Under some conditions, this test is asymptotically equivalent to t-test.

One may check against the definition to verify that this test is a permutation test.

Another intuitive proposal.

Consider the same alternative $H_1 : G(x) = F(x - \delta)$ for some $\delta > 0$. Instead of examining the size of the difference in sample means $\bar{y} - \bar{x}$, we may replace each observed value by its rank.

Define

$$r(x) = \sum_{j=1}^{m} I(x_j \le x) + \sum_{k=1}^{n} I(y_k \le x).$$

Thus, $r(y_j)$ equals the number of observations in the pooled sample that are smaller than to equal to y_j. When both F and G are continuous, we do not

need to look into the possibility of tied observations. Let

$$T_{m+n} = \sum_{j=1}^{n} r(y_j).$$

The largest possible value of T_{m+n} is when $x_i \leq y_j$ for every pair of (i, j). A large observed value of T_{m+n} is indicative of departure from H_0 in favour of H_1. Thus, a rank based permutation test is to reject H_0 when the observed T_{m+n} is among the top $100\alpha\%$ values.

If H_0 holds, then T_{m+n} has same distribution as the sample total of a simple random sample of size n without replacement from a population made of $\{1, 2, \ldots, N\}$ with $N = m + n$. Hence, by some simple calculations, we have

$$E\{T_{m+n}\} = \frac{1}{2}n(m + n + 1)$$

and

$$\text{VAR}(T_{m+n}) = \frac{1}{12}nm(n + m + 1).$$

It can be proved that

$$\frac{T_{n+m} - E\{T_{m+n}\}}{\sqrt{\text{VAR}(T_{m+n})}} \to N(0, 1)$$

in distribution, as both $n, m \to \infty$ and $n/(n + m)$ has a limit in (0, 1). An approximate one-sided rejection region can be determined by using this limiting distribution.

This test is called Wilcoxon two-sample rank test.

None of the above two tests are Uniformly Most Powerful. The test based on ranks are nonparametric. Such tests are valued because their validity is free from model mis-specifications. We will not teach more elaborative theories.

Another additional remark is about the alternative model. The formulation is clearly geared for one-sided alternative. However, a two-sided Wilconxon two sample rank test can be built based on the same principle. We reject the null hypothesis when T_{m+n} is extremely large or extremely smaller among T'_{m+n}. I leave it to you to decide a way to define the p-value.

10.3 Kolmogorov-Smirnov and Cramér-von Mises tests

Let $x_1, x_2, \ldots, x_n$ be a set of i.i.d. observations from a continuous distribution F. The model under consideration is $\mathcal{F}$: all continuous univariate distributions

Without any additional knowledge about the specific F from which we obtained the sample, one estimator of the cumulative distribution function F is given by

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I(x_i \leq x).$$

When x_i's are all different, it is uniform on $x_1, \ldots, x_n$. We may not be too happy as this estimator is not a continuous cdf while the model is made of continuous distributions. Nevertheless, F_n is a good estimator of F in many ways.

Let

$$D_n(F) = \sup_x |F_n(x) - F(x)|.$$

It is well known that $D_n(F) \to 0$ almost surely as $n \to \infty$ when F is the true distribution.

Suppose we want to test for $H_0 : F = F_0$ versus $H_1 : F \neq F_0$. It is sensible to reject H_0 when $D_n(F_0)$ is large. The test in the form of

$$\phi(x) = I(D_n(F_0) > k)$$

for some $k > 0$ is called Kolmogorov-Smirnov test.

In application, we would like to choose k so that the test has some pre-specified size. This is possible only if we have an easy to computer expression of

$$P\{D_n(F_0) > k\}.$$

This is likely a mission impossible. However, Kolmogorov proved that

$$P\{\sqrt{n} D_n(F_0) \leq t\} \to 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 t^2)$$

as $n \to \infty$. Thus, when n is large, we may use the right hand size to pick a value of t so that

$$2\sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2j^2t^2) = \alpha$$

and reject H_0 when $\sqrt{n}D_n(F_0) > t$. The expression is certainly easy to use to compute an approximate P-value.

How large this n has to be in order for the approximation with satisfactory accuracy? I do not have an answer but it exists somewhere. I will not try to give a proof. All I can say that this large sample result is crazily elegant!

Kolmogorov-Smirnov test measures the maximum discrepancy between F_n and F. It might be more helpful to examine the average difference. The Cramér-von Mises test works in this fashion:

$$C_n(F) = \int \{F_n(x) - F(x)\}^2 dF(x).$$

Under null distribution F_0, it has been shown that

$$nC_n(F_0) \to \sum_{j=1}^{\infty} \lambda_j \chi^2_{1j},$$

where $\lambda_j = j^{-2}\pi^{-2}$.

There can certainly be many other ways to examine the difference between F_n and $F(x)$.

10.4 Pearson's goodness-of-fit test

Suppose the observations are naturally categories into K groups. At the same time, these n observations are believed i.i.d. . Let p_k be the probability of an observation falls into category k, $k = 1, 2, \ldots, K$. One simple question is: does the data support the hypothesis that $p_k = p_{k0}$, $k = 1, 2, \ldots, K$. One possible approach of addressing such a concern is Pearson's goodness-of-fit test. We phrase the question from an opposite angle: is there a significant evident against the null hypothesis $H_0 : p_k = p_{k0}$?

Let o_k be the number of observations out of total n fall into category k. Let $e_k = np_{k0}$ denote the expected value of o_k under the null model. Pearson's statistic for this test problem is defined to be

$$W_n = \sum_{k=1}^{K} \frac{(o_k - e_k)^2}{e_k}.$$

This statistic clearly has one desired property for a test: when the true model deviate from the null hypothesis, we expect to have larger differences between o_k and e_k. Thus, W_n is stochastically larger when H_0 is severely violated. Naturally, we reject H_0 if W_n value is large.

The next desired property for a test statistics is to have a known distribution under H_0. This is not completely true. However, when $n \to \infty$ while K is a fixed value, it can be shown that

$$W_n \xrightarrow{d} \chi^2_{K-1}.$$

Since the chisquare distribution is well documented, we may use its upper $1 - \alpha$ quantile as the critical region for this test. Namely, the test would be

$$\text{Reject } H_0 \text{ when } W_n > \chi^2_{K-1}(1 - \alpha).$$

Of course, this writing has assumed a size-α test is desired in the first place.

In a more realistic situation, for instance, these K categories means the number of boys in a family with $K - 1$ children. Is this number truly binomially distributed as it would be under the assumption that there is no correlation between siblings and the population is homogeneous? In this case, we do not have p_{k0}'s completely specified but

$$p_{0k}(\theta) = \binom{K-1}{k-1} \theta^{k-1}(1-\theta)^{K-k}.$$

Namely, they are specified by a single parameter, the probability of success.

In this case, let $\hat{\theta}$ be the maximum likelihood estimate of θ and compute

$$\hat{e}_k = np_{0k}(\hat{\theta}).$$

Let us revise the definition of W_n and get

$$W_n = \sum_{k=1}^{K} \frac{(o_k - \hat{e}_k)^2}{\hat{e}_k}.$$

Although we have to estimate θ, the limiting distribution of W_n is only altered slightly:

$$W_n \xrightarrow{d} \chi^2_{K-2}.$$

In general, if p_{0k} are function of θ and θ has dimension d, the same approach is applicable. The limiting distribution remains chi-square with degrees of freedom being $K - d - 1$.

Being a course in mathematical statistics, one may ask how to establish the asymptotic result. One approach is to connect W_n with the likelihood ratio test. This will be left as an assignment problem.

The applied aspect of this test can be more troublesome. The biggest concern is when the chi-square approximation kicks in? The rumour is: do not use the goodness-of-fit test unless $\min\{o_k\} \geq 5$. In other applications, the observations are not "naturally categorized". The step of creating K categories in order to examine the goodness-of-fit can be controversial.

10.5 Other tests

Fisher's exact test may be added next year.

Chapter 11

Confidence intervals or confidence regions

Suppose we have a sample X from a distribution that belongs to $\mathcal{F}$. Under parametric setup, the distributions in $\mathcal{F}$ are labeled by θ and the "true" distribution of X is the distribution with label θ_0, a value we do not know.

We often do not bother to use a special single θ_0 to denote the true parameter value unless it becomes ambiguous otherwise. Most often, we simply declare that the parameter of the distribution of X is θ and its range or the parameter space is Θ. Furthermore, we implicitly assume that Θ is a subset of R^d with all mathematical properties needed (such as open, convex and so on). In addition, the distribution of X depends on the value of θ in a continuous fashion.

Based on the realized value of X, one can estimate of θ using any preferred methods. This is called point estimation. One may also make a judgement on whether or not θ is a member of an elite subset H_0 by conducting a hypothesis test. The third option is to specify a subset Θ_0 of Θ so that we are confident that $\theta \in \Theta_0$. When $d = 1$, we usually prefer that Θ_0 is an interval in Θ. When $d > 1$, Θ_0 will be called a confidence region. It is preferable that Θ is connected and it does not contain holes. More often than not, convex set is more appealing.

Do remember that Θ_0 is decided by the value of X, and it cannot depend on any unknown parameters. Thus, it is a **random** set. Its randomness is

dependent on the distribution of X. Recall that a statistic is a function of exclusively data; the same is true for the confidence region.

As usual, constructing an interval is easy. The real task is to construct an interval with desirable properties. Our task is hence to determine what properties such an interval estimation should have, and how do we construct intervals with these properties.

Definition 11.1 *An interval $C(x)$ as a function of the realized value of X is a confidence interval of θ at level $1-\alpha$ for some $\alpha \in (0,1)$, if*

$$\inf_{\theta} P\{\theta \in C(X); \theta\} = 1 - \alpha.$$

In the above probability calculation, θ is the putative true parameter value of the distribution of X and it is not random. The interval $C(X)$ is random. In comparison, when θ is regarded as random in Bayes analysis, the above probability would be computed conditional on X. We will not call it as confidence region but "credential region". There is no specific shape requirement on a confidence region in general. Yet we have preferences.

The probability that $C(X)$ covers θ generally depends on the parameter value θ from which the data are observed. It is desirable to have the coverage probability not dependent on the specific value of θ. If this is achieved, the infimum operation in the above definition would be unnecessary.

Similar to the optimality notion in hypothesis test, comparison between different confidence regions are possible only if they have been lined them up so that they have the same confidence level. After which, it is most sensible to require the interval to have the shortest average/expected length. On top of that, one may hope that the variation in the length of the interval as low as possible.

Suppose $C(X) = [1, 4]$ is a confidence interval for the population mean. If this interval is sensibly constructed, it is generally true that the most likely value of θ is located at the centre of this interval. The chance that the first quarter, $[1, 2]$, contains the true θ value is not as high as that of $[2, 3]$. Yet the frequentist notation of confidence interval as given in the above definition has it completely ignored. We may notice that this notion is somewhat accommodated in procedures of the construction of confidence intervals.

Unlike the theory for hypothesis test, there seem to be fewer solid mathematical criteria for the optimality of confidence regions. Confidence regions are often derived from other well known procedures. If these procedures have optimal properties in some sense, statisticians seem to feel comfortable to recommend the corresponding confidence regions. Having remarked as this, I do aware one optimality criterion which can be found in Classical textbooks. We will not have it discussed here.

11.1 Constructing confidence intervals via hypothesis test

Assume a sensible hypothesis test $\phi(x)$ of size α is possible for all simple null hypothesis $H_0 : \theta = \theta_0$ against a composite alternative hypothesis $H_1 : \theta \neq \theta_0$. To be more specific, let the test be $\phi(x;\theta)$. Thus, θ_0 is rejected when $\phi(x;\theta_0) = 1$, ignoring the randomization trouble.

Based on $\phi(x;\theta)$, let us define

$$C(x) = \{\theta : \phi(x;\theta) < 1\}.$$

It is easy to see that

$$P\{\theta \in C(x);\theta\} = P\{\phi(X;\theta) < 1\} \geq 1 - E\{\phi(X;\theta) : \theta\} \geq 1 - \alpha.$$

for all $\theta \in \Theta$. Thus, $C(x)$ is a $1-\alpha$ level confidence interval.

Example 11.1 *Suppose we have a random sample from $N(\theta, \sigma^2)$. We hope to construct a confidence interval for θ. One approach is to figure out the likelihood ratio test for each $H_0 : \theta = \theta_0$. In this case, the test statistic simplifies to*

$$T(x;\theta_0) = \frac{\sqrt{n}|\bar{X} - \theta_0|}{s_n}.$$

The critical region for each θ_0 is defined as

$$\{x : T(x;\theta_0) \geq t_{n-1}(1-\alpha/2)\}.$$

Consequently, the confidence interval based on this test is

$$\{\theta_0 : T(x;\theta_0) \le t_{n-1}(1-\alpha/2)\}$$

or

$$[\bar{x} - t_{n-1}(1-\alpha/2)s_n/\sqrt{n}, \bar{x} + t_{n-1}(1-\alpha/2)s_n/\sqrt{n}].$$

It is nice to see that the outcome is indeed an interval.

We will discuss more on procedures of hypothesis test, each will imply a way of constructing confidence intervals (regions).

Confidence interval by pivotal quantities. Recall that if a function of both data and unknown parameter has a distribution which does not depend on unknown parameters, we call it a pivotal quantity.

Suppose $q(x;\theta)$ is a pivotal quantity. It is hence conceptually possible to find a quantity, say q_α such that

$$P\{q(X;\theta) > q_\alpha;\theta\} = \alpha$$

by ignoring the situation where $q(X;\theta)$ has a discrete distribution. The solution q_α exists (barring discreteness) because the distribution does not depend on the unknown value θ.

Let

$$C(x) = \{\theta : q(x;\theta) < q_\alpha\}.$$

It is easily seen that $C(x)$ is a $1-\alpha$-level confidence region of θ.

Examples of pivotal quantity are most readily available in location-scale families.

Example 11.2 *Suppose we have a random sample from $N(\theta,\sigma^2)$. Let us try to find a confidence interval for σ^2.*

It is well known that

$$q(x;\sigma^2) = \frac{\sum(x_i - \bar{x})^2}{\sigma^2}$$

has chisquare distribution with $n-1$ degrees of freedom. Thus, it is a pivotal quantity.

Let $\chi^2_{n-1}(0.95)$ *be the 95th percentile of the chisquare distribution with* $n-1$ *degrees of freedom. Then,*

$$\{\sigma^2 : \frac{\sum(x_i - \bar{x})^2}{\sigma^2} < \chi^2_{n-1}(0.95)\} = [\frac{\sum(x_i - \bar{x})^2}{\chi^2_{n-1}(0.95)}, \infty)$$

is a 95% confidence interval for σ^2*.*

It is clearly possible for us to construct a two-sided confidence interval based on this pivotal.

11.2 Likelihood intervals.

By the definition, a confidence region must be characterized by its level of confidence. Yet the interval makes more sense if a parameter value within the region is more "likely" than a parameter value outside of the region, to be the "true" value of the parameter. This is particularly the case in the confidence interval for σ^2 in the last example. The notation of likelihood interval or related Bayesian approach seem to improve in this direction.

Suppose we have a random sample of size n from a parametric family $\{f(x;\theta) : \theta \in R^d\}$. Consider the problem of constructing a confidence interval/region for θ. Since by "definition", the maximum likelihood estimator is the most "likely" value of the parameter, the interval should contain the MLE $\hat{\theta}$. In addition, if the likelihood value at θ' is almost as large as the likelihood value as $\hat{\theta}$, it is also a good candidate to be included into the interval.

This notion quickly deduces to a likelihood region/interval in the form of

$$C(X) = \{\theta : L(\theta)/L(\hat{\theta}) \geq c\}$$

where $\hat{\theta}$ is the MLE, and c is a positive constant to be chosen.

To make a likelihood interval into a confidence interval, all we need is to choose c such that

$$P_\theta\{\theta \in C(X)\} = 1 - \alpha$$

is true for any θ, when the pre-specified level is $1 - \alpha$.

Usually, there does not exist such a constant c such that the coverage probability is $1-\alpha$ under all θ. However, when the sample size n is large and the model is regular, it is possible to find an c_n such that the coverage probability is approximately $1-\alpha$ for each θ. That is, the difference is an quantity converges to 0 when $n \to \infty$, whichever θ is the true value. We call such confidence region has asymptotic $(1-\alpha)$-level.

To students with rigorous mathematics background, you may notice that the asymptotic notion is not uniformly in θ. We only required the convergence point-wise, not uniformly over the parameter space.

Example 11.3 *Find an example here. Say based on Binomial distribution.*

11.3 Intervals based on asymptotic distribution of $\hat{\theta}$

It is arguable whether or not this is a new method. We might call it Wald's method, yet it has too many moving parts to be solidly called as this method. Often, $\sqrt{n}(\hat{\theta}-\theta)$ is asymptotic normal with limiting variance σ^2. When σ^2 is known, then

$$q(X;\theta) = \frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma}$$

is an approximate pivotal quantity. Because of this, an approximate $1-\alpha$ confidence interval of region of θ is given by

$$\hat{\theta} \pm z_{1-\alpha/2}\sigma/\sqrt{n}.$$

If σ is unknown but a consistent estimator $\hat{\sigma}$ is available, then a substitute is given by

$$\hat{\theta} \pm z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}.$$

It might be more convenient to write the above as

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\mathrm{VAR}}(\hat{\theta})/n}.$$

The meaning of the above notation is obvious.

Example 11.4 *Let $X_1, \ldots, X_n$ be an i.i.d. sample from Poisson distribution with mean parameter denoted as θ. The MLE of θ is given by $\hat{\theta} = \bar{X}_n$, the sample mean. Construct a 95% CI for θ.*

Solutions: It is well known that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta)$. Thus, a 95% CI for θ is given by

$$\bar{X}_n \pm 1.96\sqrt{\bar{X}_n/n}.$$

When $1.96\sqrt{\bar{X}_n/n} > \bar{X}_n$, one must set lower confidence limit(bound) to 0.

It is equally appropriate to notice that

$$\sqrt{n}(\sqrt{\hat{\theta}} - \sqrt{\theta}) \xrightarrow{d} N(0, 1/4).$$

Hence, one may construct a 95% CI based on

$$\sqrt{n}|\sqrt{\hat{\theta}} - \sqrt{\theta}| \leq 1.96/2.$$

Solving this inequality, we get

$$[\{\sqrt{\bar{X}_n} - 1.96/2\sqrt{n}\}^+]^2 < \theta < \{\sqrt{\bar{X}_n} + 1.96/2\sqrt{n}\}^2.$$

The third choice is to work with

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\theta}} \xrightarrow{d} N(0, 1).$$

With this pivotal quantity, the CI has lower and upper limits

$$\bar{X}_n + \frac{1.96^2}{2n} - \sqrt{\frac{1.96^4}{4n^2} + \frac{1.96^2\bar{X}}{n}}$$

and

$$\bar{X}_n + \frac{1.96^2}{2n} + \sqrt{\frac{1.96^4}{4n^2} + \frac{1.96^2\bar{X}}{n}}.$$

♢

Many students have natural tendency to ask the following question. Which of the above confidence intervals is **correct**? The answer is: none of them. The reason is: the critical value 1.96 is based on the limiting distribution of

$\hat{\theta}$ in every case. Hence, none of them have exact 95% coverage probability (even if the rounding-off is not counted).

If the above answer is unsatisfactory to you, then you need to think hard about what it means by "correct". If approximate 95% CIs are acceptable, all three are **fine**.

The real question in your mind might be: which one is the best? Answering this question needs an optimality criterion. We do not have one at the moment. We will put up one later but I am reluctant to use an optimality criterion anyway. Now it boils down to a weak question: what are their relative merits?

The first one is analytically simple. If the sample size n is not very large, the normal approximation can be poor. The CI may even have a negative lower bound. Chop-off the segment of the CI containing the negative values is a mathematical must but somewhat unnatural. The interval is always symmetric with respect to $\hat{\theta} = \bar{X}_n$. This is somewhat unattractive. I would use this one when n and $\bar{X}_n$ are both large. How large is large? I do not have an absolute standard.

The second one is nice in one way: after transforming θ into $g(\hat{\theta}) = \sqrt{\theta}$, the limiting distribution of $g(\hat{\theta})$ has a constant limiting distribution. For this reason, this type of transformation is called *variance stabilization transformation*. Since the limiting distribution does not depend on unknown parameter values, this interval is truly based on approximate pivotal. If n is not large, this is a good choice.

The third one has its own merit. Scaling $\hat{\theta} - \theta$ by a function of θ creates a more complex pivotal. This often leads to more naturally shaped confidence regions (intervals). While I have intuitions for this approach, I cannot come up with concrete evidences for this preference.

Recall that testing hypothesis on θ value based on the limiting distribution of the MLE $\hat{\theta}$ is called Wald's method. I am not sure if this group of intervals should be credited to Wald, but I feel it is natural to call it Wald's interval/region.

General case; Binomial example; Odds ratio. Intervals for quantiles.

11.4 Bayes Interval

Under Bayesian setup, the parameter θ is a sample from some prior distribution. Thus, its value itself is a realization of random variable. Constructing a confidence interval for a random quantity does not make sense. However, in the presence of data from $f(x;\theta)$, it helps us to guess that this realized value of θ. The information about θ is completely summarized in the posterior distribution of θ. If one must take a guess on a region in which this θ has been located based on Bayesian setup, she would select the region with the highest posterior density.

Definition 11.2 *Let $\pi(\cdot|x)$ denote the posterior density function of parameter θ given $X = x$. Then*

$$C_k = \{\theta : \ \pi(\cdot|x) \geq k\}$$

is called a **level $1-\alpha$ credible region** *for θ if $P(\theta \in C_k|x) \geq 1-\alpha$.*

Note that $P(\cdot|x)$ is used for posterior distribution of θ. If one can credibly regard θ as an outcome from a prior distribution, then the above credible region has a very strong appealing.

If θ is not a vector but a real value, then we may look for a confidence interval instead.

Definition 11.3 *Let $\Pi(\cdot|x)$ denote the posterior cumulative distribution function of parameter θ given $X = x$. Suppose $\bar{\theta}$ and $\underline{\theta}$ satisfy*

$$\Pi(\underline{\theta} \leq \theta|x) \geq 1-\alpha; \quad \Pi(\bar{\theta} \geq \theta|x) \geq 1-\alpha.$$

Then, $\underline{\theta}$ and $\bar{\theta}$ are **level $1-\alpha$ lower and upper credible bounds** *for θ.*

In both definitions, it appears more sensible to replace $\geq$ to $=$. My source of these two definitions are from Bickel and Doksum (2001). There might some advantage to keep the first one as is, and I will not take the lead to change the second definition. The following is an example directly copied from Bickel and Doksum (2001).

Example 11.5 *Suppose that given μ, $X_1, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma_0^2)$ with known σ_0^2. The prior distribution of μ is $N(\mu_0, \tau_0^2)$ with both parameter values are known. Find the credible bounds and regions according to the above definitions*

Solution. 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 lower and upper $1-\alpha$ bounds are simply

$$\mu_B \pm z_{1-\alpha}\frac{\sigma_0}{\sqrt{n + \sigma_0^2/\tau_0^2}}.$$

The $1-\alpha$ credible region is also an interval with lower and upper limits given by

$$\mu_B \pm z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n + \sigma_0^2/\tau_0^2}}.$$

♢

Note that the centre of the credible interval is shift toward μ_0 compared with usual confidence intervals. The length is shortened too.

11.5 Prediction intervals

In general, the notion of confidence region is defined for unknown parameters of a distribution family. There are cases where we hope to predict the outcome of a future trial from the same probability model.

Suppose we have a set of iid sample $X_1, X_2, \ldots, X_n$ from $\{f(x;\theta) : \theta \in \Theta\}$. Based on this sample, we might have an estimate of θ. The question is: if another independent sample is to be taken from the same distribution, what are the possible values of this future X?

If the value of θ for this experiment were known, we could use the high density region of $f(x;\theta)$ as our prediction region. That is, let

$$C(\theta) = \{x : f(x;\theta) > c\}$$

with the **known value** θ. We may choose c such that

$$P(X \in C(\theta)) = 1 - \alpha$$

for $1-\alpha$ coverage probability. Note that this region is not dependent on the random sample $X_1, \ldots, X_n$.

If θ is unknown as usual, it is natural to replace θ by its estimator, say $\hat{\theta}$. Although $C(\hat{\theta})$ is a very sensible prediction region for X, its coverage probability is likely lower than $1-\alpha$. This is because the event

$$X \in C(\hat{\theta})$$

contains two random components: X and $\hat{\theta}$. The randomness in X is unaffected by how well θ is estimated, while the precision of $\hat{\theta}$ usually improves with sample size n. The limit of the improvement is $C(\theta)$. Due to the build-in randomness in X, one cannot do anything better than $C(\theta)$ no matter what.

In comparison, the precision of the confidence region for θ usually improves with n. When $n \to \infty$, the size of the confidence region with fixed confidence level shrinks to 0.

Example Suppose we have a random sample $X_1, \ldots, X_n$ from $N(\theta, 1)$. It is well known that $\hat{\theta} = \bar{X}_n$ is the MLE of θ.

If X is the outcome of a future experiment, then $X - \bar{X}_n$ has normal distribution with mean 0 and variance $1 + n^{-1}$. Thus, a 95% prediction interval of X is given by

$$(\bar{X}_n - 1.96\sqrt{1+n^{-1}}, \bar{X}_n + 1.96\sqrt{1+n^{-1}}).$$

Clearly, increasing the sample size does not have much impact on reducing the length of the prediction interval.

In general, a prediction interval can be obtained via some "pivotal" quantities. That is, we look for a function of the random quantity to be predicted,

and a statistic based on observations such that the resulting random quantity has a distribution free from unknown parameters. Thus, it is possible to find a subset of its range such that its probability equals $1 - \alpha$, the confidence level we hope to get. The range can often be converted to obtain a prediction interval or region.

Chapter 12

Empirical likelihood

Likelihood method for regular parametric models has many nice properties. One potential problem, though, is the risk of model mis-specification. If a data set is a random sample from Cauchy distribution, but we use normal model in the analysis, the statistical claims could be grossly false.

Of course, the problem is not always so serious. If the data set is a sample from a double exponential model, but we use normal model as the basis for data analysis, most statistical claims will still be asymptotically valid. For instance, the sample mean remains a good estimator of the population mean. The efficiency of the point estimator, however, is compromised.

To avoid the risk of model mis-specification, non-parametric methods are sensible alternatives. The empirical likelihood methodology is a systematic non-parametric approach to the statistical analysis. I will take materials in this chapter from Professors Owen, Qin and others including myself.

12.1 Definition of the empirical likelihood

Suppose we have a set of i.i.d. observations $X_1, X_2, \ldots, X_n$. We hope to make statistical inferences without placing restrictive assumptions on their common distribution F. Can we still make meaningful and effective inferences on F? The answer is positive because we already know that the empirical distribution $F_n(x)$ is a good estimate of F. This is an estimator based on no parametric assumptions at all. The empirical distribution is a non-

parametric maximum likelihood estimator of F and it has various "optimal" properties.

Let $F(\{x_i\}) = P(X = x_i)$, where x_i is the observed value of X_i, $i = 1, 2, \ldots, n$. When all x_i's are distinct, the likelihood function becomes

$$L_n(F) = \prod_{i=1}^{n} F(\{x_i\}).$$

Denote $p_i = F(\{x_i\})$. This likelihood can also be written as

$$L_n(F) = \prod_{i=1}^{n} p_i.$$

Clearly, we have $0 \le p_i \le 1$ and $\sum_{i=1}^{n} p_i \le 1$. It is often convenient to work with the log-empirical likelihood function

$$\ell_n(F) = \sum_{i=1}^{n} \log p_i.$$

If F is a continuous distribution, we have $L_n(F) = 0$. Because of this, the empirical likelihood appears insensible. In its eyes, no continuous distributions are likely at all. Yet we will find the empirical likelihood is not boggled down by this problem.

When there are ties in the data, that is when some x_i are equal, $L_n(F)$ defined in terms of p_i can have a technical problem. Namely, the requirement of $\sum p_i \le 1$ is no longer valid. To justify the continued use of this $L_n(F)$ via p_i, we may add a set of independent and very small continuous noises to these observed values. After which, $L_n(F)$ remains a valid likelihood function but is constructed on a slightly different data set and of a different F. We can then proceed to whatever analysis first, and then let this amount of noise go to zero. In most situations, the analysis conclusion on original F will remain valid. Owen (2001) contains a more rigorous justification to resolve the technicalities caused by tied observations. Mine justification might be regarded as a lazy-man's approach.

It is easy to see that the likelihood is maximized when $F(x) = F_n(x)$. Hence, empirical distribution $F_n(x)$ based on an i.i.d. sample is also the nonparametric MLE. One may note that this conclusion does not depend on whether or not there are nay ties in the sample.

12.2 Likelihood ratio function and profile likelihood

Since $L_n(F_n) > L_n(F)$ for any $F \neq F_n$, it is useful to introduce the empirical likelihood ratio function

$$R_n(F) = L_n(F)/L_n(F_n) = \prod_{i=1}^{n}(np_i).$$

This function has the maximum value of 1.

Suppose we are interested in a parameter $\theta = T(F)$ for some functional of the distribution F. We also assume that the class of distributions under consideration is $\mathcal{F}$ which could be all possible distributions or any smaller set of distributions. An example of $\mathcal{F}$ is all distributions with finite variance. An example of θ is the first moment of F. The profile empirical likelihood ratio function is defined as

$$R_n(\theta) = \sup\{R_n(F)|T(F) = \theta, F \in \mathcal{F}\}. \tag{12.1}$$

That is, for each given θ, we search for distributions in $\mathcal{F}$ such that $T(F) = \theta$. We then determine the supremum of the empirical likelihood ratio over this set, and define it as the value of the empirical likelihood ratio function at θ. When the set of $T(F) = \theta$ is empty, $R_n(\theta)$ is defined to be 0 suggested by some researchers. In some applications, this convention may not be satisfactory. We will introduce a different convention at the end of this chapter.

The above consideration leads us to the definition of the profile empirical likelihood function

$$L_n(\theta) = \sup\{L_n(F)|T(F) = \theta, F \in \mathcal{F}\}$$

and in its logarithm form $\ell_n(\theta) = \log L_n(\theta)$. We may also use $r_n(\theta) = \log R_n(\theta)$. There are some issues related to this definition. For this reason, we do not frame it as a definition. A formal definition will be introduced later.

In parametric inference we may base hypothesis test and confidence regions on the size of the likelihood ratio function. When $R_n(\theta)$ is large, then

θ is a likely value of the true parameter. A confidence region hence is made of θ's such that $R_n(\theta)$ is larger than a threshold value. With this statistical idea un-disputed, we consider the problem of how large this threshold should be. In the non-parametric inference, we adopt the same idea. The empirical likelihood confident regions are defined as

$$\{\theta : R_n(\theta) \geq r_0\}$$

for some r_0 which is determined in accordance to the confidence level.

12.3 Confidence region for means

Let $\theta = E(X) = \int x dF(x)$. We assume that X is a random vector of dimension d in this section. To use the idea discussed in the last section, the distribution family $\mathcal{F}$ has to be restricted to certain class carefully. One natural choice of $\mathcal{F}$ is the set of all F such that $E|X| < \infty$.

Had $R_n(\theta)$ be defined as in (12.1), we would have difficulties to use it for confidence region construction for the population mean. For this $\mathcal{F}$, it can be shown that

$$\{\theta : R_n(\theta) \geq r_0\}$$

contains all real values (vectors) for any choice of $r_0 < 1$. This is given as an assignment problem. The assignment problem is made based on $R_n(\theta)$ defined by (12.1). Be aware of this difference when you work on the problem.

To avoid this difficulty, one may restrict the range of F to be the convex hall formed by n observed values. Mathematical, the likelihood is then maximized on F in the form of

$$F(x) = \sum_{i=1}^{n} p_i I(x_i \leq x).$$

In other words, we would strictly require $\sum_{i=1}^{n} p_i = 1$ in defining the profile empirical likelihood as compared with the conceptually correct restriction $\sum_{i=1}^{n} p_i \leq 1$.

Definition 12.1 *Suppose we are given a set of i.i.d. observations $x_1, x_2, \ldots, x_n$ from distribution F in the family $\mathcal{F}$. Let $\theta = T(F)$ be a functional on $\mathcal{F}$.*

We define the profile empirical likelihood function of θ as

$$L_n(\theta) = \sup\{\prod_{i=1}^{n} p_i |\ T(F) = \theta, F(x) = \sum_{i=1}^{n} p_i I(x_i \le x); T(F) = \theta, F \in \mathcal{F}\}. \tag{12.2}$$

The corresponding profile likelihood ratio function of θ as

$$R_n(\theta) = \sup\{\prod_{i=1}^{n} (np_i) |\ T(F) = \theta, F(x) = \sum_{i=1}^{n} p_i I(x_i \le x); T(F) = \theta, F \in \mathcal{F}\}. \tag{12.3}$$

Although $F \in \mathcal{F}$ is in the above definition, it does not truly impose any restrictions because we generally make $\mathcal{F}$ as broad as possible. I do not aware of any other writings to have it included. Hopefully, this seemingly unnecessary requirement serves as a reminder that the likelihood is defined on some distribution space.

It could happen that there exist no distribution in the form of $F(x) = \sum_{i=1}^{n} p_i I(x_i \le x)$ such that $T(F) = \theta$ for a value of θ. When that happens, the convention is to define $L_n(\theta) = 0$. If such a solution does not exist for any θ, which could happen in more general setup to be introduced, we say the empirical likelihood has encountered an empty-set problem. When the model is "correctly specified", the probability of having an empty-set problem reduces to 0 when the sample size n goes to infinite under i.i.d. setting. This result will be presented later.

Theorem 12.1 *Let $X_1, X_2, \ldots, X_n$ be a set of iid random vectors with common distribution F_0. Let $\theta_0 = E[X_1]$, and suppose $0 < \text{VAR}(X_1) < \infty$. Then*

$$-2\log[R_n(\theta_0)] \to \chi^2_d$$

in distribution as $n \to \infty$.

Because of the above Wilks type result, an effective empirical likelihood based hypothesis test procedure is possible by rejecting $H_0 : E(X) = \theta_0$ in favour of $H_0 : E(X) \neq \theta_0$ when $T_n = -2\log[R_n(\theta_0)] \ge \chi^2_d(1-\alpha)$. Note that this d is the dimension of X.

12.4 Lagrange multiplier

One practical problem is how to compute the confidence region numerically given a random sample. At least for the computation of profile likelihood, the solution is simpler than what we may expect. We have n observed values or vectors $x_1, \ldots, x_n$. For each given θ, $\ell_n(\theta)$ is the maximum of $\ell_n(F)$ for all F such that $T(F) = \theta$. Hence, to compute $\ell_n(\theta)$, the numerical problem is:

$$\begin{aligned} \text{maximize :} \quad & \sum_{i=1}^{n} \log p_i \\ \text{subject to:} \quad & 0 < p_i < 1; \quad i = 1, 2, \ldots, n \\ & \sum_{i=1}^{n} p_i = 1, \\ & \sum_{i=1}^{n} p_i x_i = \theta. \end{aligned}$$

The method of Lagrange multiplier is very effective in solving this maximization problem with restrictions. Suppose that θ is **an interior point of the convex hull formed by the n observed values**. Define

$$g(s, \lambda) = \sum_{i=1}^{n} \log p_i + s(\sum_{i=1}^{n} p_i - 1) - n\lambda(\sum_{i=1}^{n} p_i x_i - \theta)$$

where s and λ are Lagrange multipliers. When x's are vectors, λ is also a vector and the multiplication is interpreted as the dot product. The method of Lagrange multiplier requires us to find the stationary points of $g(s, \lambda)$ with respect to s and λ. After some routine derivations, we find

$$p_i = \frac{1}{n(1 + \lambda\{x_i - \theta\})}$$

with λ satisfying

$$\sum_{i=1}^{n} \frac{x_i - \theta}{1 + \lambda\{x_i - \theta\}} = 0. \tag{12.4}$$

In the univariate case, since all $0 < p_i < 1$, we find

$$\frac{1 - n^{-1}}{\theta - x_{(n)}} < \lambda < \frac{1 - n^{-1}}{\theta - x_{(1)}}$$

where $x_{(1)}$ and $x_{(n)}$ are the minimum and maximum observed values. In addition, the function on the left hand side of (12.4) is monotone decreasing function in λ. One may verify this claim by finding its derivative with respect to λ. Hence, numerical value of λ may be easily computed. In the vector case, the function in (12.4) is the derivative of a convex function. A revised Newton's method may be designed to ensure the numerical solution being obtained.

Once the value of λ is obtained numerically, we have

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

and

$$r_n(\theta) = \log R_n(\theta) = -\sum_{i=1}^{n} \log\{1 + \lambda(x_i - \theta)\}. \tag{12.5}$$

The computer implementation will be further discussed.

12.5 Some technical results and proofs

The discussion in the last section is also useful to the proof of the theorem. For this purpose, we first show that the true parameter value θ_0 will fall into the convex hull of the data with probability approaching 1 as $n \to \infty$. Mathematically, this means that

$$\inf\{\max\{a(x_i - \theta_0) : i = 1, \ldots, n\} : a \text{ is a unit vector}\} > 0.$$

The reason is: viewed from θ_0 to which ever direction, there should always be data located in that direction. In case you do not know, a unit vector is a vector of length one. If not specifically declared, we use Euclidean norm to define length. A simple mathematical result presented in Owen (2001) is very helpful here.

Lemma 12.1 *Let X be a random vector with mean 0 and finite variance-covariance matrix V of full rank. We have*

$$\inf_a P(a^\tau X > 0) > 0$$

where the infimum is taken over all unit d-dimensional vectors.

.

PROOF: If the conclusion is not true, then there must exist a sequence a_m such that $P(a_m^\tau X > 0) \to 0$ as $m \to \infty$. Since the set of all unit d-dimensional vectors is compact, we must be able to find a sub-sequence of a_m such that $a_m \to a_0$ for some a_0. Without loss of generality, assume $a_m \to a_0$ as $m \to \infty$. Consequently,

$$\lim_{m\to\infty} I(a_m^\tau X > 0) = I(a_0^\tau X > 0).$$

Hence, due to continuity of the probability measure, we have

$$P(a_0^\tau X > 0) \le \lim_{m\to\infty} P(a_m^\tau X > 0) = 0.$$

This contradicts the assumption that V has full rank. That is, the conclusion of the Lemma is true. ♢

Recall that the empirical measure approximates the true probability measure uniformly over the half space $\{a^\tau X > 0\}$. This implies that the solution exists with probability converging to 1.

By Slutsky's theorem, the limiting distribution will not be affected by the event with probability going to zero. We now pretend that the solution exists for all data sets observed. This is acceptable for deriving asymptotic results, though one should not used it for other purposes. At least, you should be very cautious on activating this "assumption".

Our next lemma is to show that $\max_{1\le i\le n} \|X_i\| = o_p(n^{1/2})$. This fact is helpful to determine the closeness of $\hat{p}_i$ to $1/n$ as $n \to \infty$.

Lemma 12.2 *Assume $Y_1, \ldots, Y_n$ be a set of i.i.d. positive random variables with $E[Y_1]^2 < \infty$, then $Y_{(n)} = \max_i Y_i = o(n^{1/2})$.*

PROOF There is a simple inequality for positive valued random variables:

$$\sum_{j=1}^{\infty} P(Y_1^2 > j) \le E[Y_1^2].$$

Due to the i.i.d. assumption, it can also be read as

$$\sum_{j=1}^{\infty} P(Y_j^2 > j) \le E[Y_1^2] < \infty.$$

The inequality in the above line is the assumption of the lemma. The inequality can then be easily refined to

$$\sum_{j=1}^{\infty} P(Y_j^2 > \epsilon j) < \infty$$

for any $\epsilon > 0$. By Borel-Cantelli Lemma, it implies that $Y_j^2 > \epsilon j$ for $j = 1, 2, \ldots$ almost surely does not occur infinitely often. That is, there exists an event A_ϵ, $P(A_\epsilon) = 1$ and for each $\omega \in A_\epsilon$, $Y_n^2(\omega) > \epsilon n$ for only finite number of n. For this ω, let us assume that $Y_n^2(\omega) > \epsilon n$ does not occur when $n > M$ for some large M. Let

$$N = \max\{Y_n^2(\omega) : n \leq M\}/\epsilon.$$

For all $n \geq max\{M, N\}$,

$$Y_{(n)}^2 \leq \max[\max\{Y_n^2(\omega) : n \leq M\}, \epsilon n] \leq \max\{\epsilon N, \epsilon n\} = \epsilon n.$$

That is, $Y_{(n)}^2 \leq \epsilon n$ almost surely for all $\epsilon > 0$. This is the conclusion of the lemma. ♢

The above result is then use to show that Lagrange multiplier $\lambda = O_p(n^{-1/2})$. After which, the Wilks type theorem will be proved.

Lemma 12.3 *Under the conditions of Theorem 12.1, we have*

$$\lambda = O_p(n^{-1/2}).$$

Further, we have

$$\lambda = [\sum_{i=1}^{n}(x_i - \theta)(x_i - \theta)^\tau]^{-1} \sum_{i=1}^{n}(x_i - \theta) + o_p(n^{-1/2})$$

and $\max_i |\lambda^\tau(X_i - \theta)| = o_p(1)$.

PROOF: Let $\rho = \|\lambda\|$ and denote $\xi = \lambda/\rho$. For brevity, assume $\theta = 0$ so that the equation for λ becomes

$$\sum_{i=1}^{n} \frac{x_i}{1 + \rho\xi^\tau x_i} = 0.$$

We have

$$0 = \sum_{i=1}^{n}(\xi^\tau x_i) - \rho \sum_{i=1}^{n} \frac{\{\xi^\tau x_i\}^2}{1+\rho\xi^\tau x_i}.$$

This implies

$$\sum_{i=1}^{n}(\xi^\tau x_i) = \rho \sum_{i=1}^{n} \frac{\{\xi^\tau x_i\}^2}{1+\rho\xi^\tau x_i} \geq 0.$$

Let $t_i = \xi^\tau x_i$ and $\delta_n = \max_i |t_i|$. It is known $1+\rho t_i > 0$ for all i and therefore $1+\rho\delta_n \geq 0$. Further, by the finiteness of the second moment of x_i and Lemma 2, $\delta_n = o(n^{1/2})$. Hence,

$$\begin{aligned} \sum_{i=1}^{n} \xi^\tau x_i &= \rho \sum_{i=1}^{n} \frac{\{\xi^\tau x_i\}^2}{1+\rho\xi^\tau x_i} \\ &\geq \rho \frac{[\sum_{i=1}^{n}\{\xi^\tau x_i\}^2}{1+\rho\delta_n}. \end{aligned}$$

Multiply the positive constant $1+\rho\delta_n$ on both sides, and some simple algebra, we get

$$\sum_{i=1}^{n}(\xi^\tau x_i) \geq n\rho[n^{-1}\sum_{i=1}^{n}(\xi^\tau x_i)^2 - n^{-1}\delta_n \sum_{i=1}^{n}(\xi^\tau x_i)].$$

By the law of large numbers,

$$n^{-1}\sum_{i=1}^{n} x_i x_i^\tau \to \text{VAR}(X_1)$$

which is a positive definite matrix. Hence, $n^{-1}\sum_{i=1}^{n}(\xi^\tau x_i)^2 \geq \sigma_1^2 > 0$ almost surely with σ_1^2 being the smallest eigenvalue of the covariance matrix. At the same time, it is clear that

$$n^{-1}\delta_n \sum_{i=1}^{n}(\xi^\tau x_i) = o_p(1).$$

Consequently, we have shown

$$\rho \leq [\sum_{i=1}^{n}(\xi^\tau x_i)^2]^{-1} \sum_{i=1}^{n} \xi^\tau x_i(1+o_p(1)) = o_p(n^{-1/2}).$$

This implies $\max_i |\lambda X_i| = o_p(1)$. Substituting back to

$$\sum_{i=1}^{n} \frac{x_i}{1+\lambda x_i} = 0,$$

we get the expression for λ. This concludes the proof. $\diamondsuit$

Now we come to the final step of the proof.

PROOF OF THEOREM: Because $\max \|\lambda^\tau X_i\| = o_p(1)$, let us focus on events such that it is no more than $1/10$ in absolute value. For $|t| \leq 1/10$, it is simple to see that

$$|\log(1+t) - \{t - \frac{1}{2}t^2\}| \leq |t|^3/2.$$

We in fact have given a big margin for the error.

We again set $\theta_0 = 0$. Using this fact, we have

$$\begin{aligned} -2\log R_n(\theta_0) &= 2\sum_{i=1}^{n} \log\{1+\lambda^\tau x_i\} \\ &= 2\lambda^\tau \sum_{i=1}^{n} x_i - \lambda^\tau \{\sum_{i=1}^{n} x_i x_i^\tau\}\lambda + \epsilon_n \\ &= \{\sum_{i=1}^{n} x_i\}^\tau \{\sum_{i=1}^{n} x_i x_i^\tau\}^{-1} \{\sum_{i=1}^{n} x_i\} + o_p(1) + \epsilon_n. \end{aligned}$$

The leading term has chisquare limiting distribution. We need only verify that $\epsilon_n = o_p(1)$. This is true as

$$|\epsilon_n| \leq \sum_{i=1}^{n} |\lambda^\tau x_i|^3 \leq \max_i |\lambda^\tau x_i| \sum_{i=1}^{n} |\lambda^\tau x_i|^2 = o_p(1).$$

This completes the proof. $\diamondsuit$

12.6 Numerical computation

The numerical computation appears to be problematic initially. We have to maximize a function with respect to n variables under various linear constraints. As seen in the last a few subsections, it turns out that once the

Lagrange multiplier λ is known, the remaining computation is very simple. We illustrate the numerical computation in this section.

Consider the problem of computing the profile likelihood for the mean. The computation is particularly simple when x is a scale. In this case, we need to solve

$$g(\lambda) = \sum_{i=1}^{n} \frac{x_i - \theta}{1 + \lambda(x_i - \theta)} = 0$$

for a given set of data, and value θ. Our first step is to subtract θ from x_i and call them y_i. Namely define $y_i = x_i - \theta$ whenever a θ value is selected. We then sort y_i to increase order and obtain $y_{(1)}$ and $y_{(n)}$. If they have the same sign, there will be no solution. The numerical solution is mission impossible.

Otherwise, the sign of λ is the same as $\bar{y}_n$. If $\bar{y}_n > 0$, we search in the interval of $[0,\ (n^{-1} - 1)/y_{(1)})$. Otherwise, we search in the interval $(\ (n^{-1} - 1)/y_{(n)}, 0]$. We also note that $g(\lambda)$ is a decreasing function. Let us provide the following pseudo code for computing λ:

1. Compute $y_i = x_i - \theta$;
2. Sort y_i to get $y_{(i)}$;
3. If $y_{(1)}y_{(n)} \geq 0$, stop and report "not solution". Otherwise, continue;
4. Compute $\bar{y}$. If $\bar{y} > 0$, set $L = 0$, $U = (n^{-1} - 1)/y_{(1)}$, otherwise set $L = (n^{-1} - 1)/y_{(n)}$, $U = 0$.
5. Set $\lambda = (L + U)/2$.
6. If $g(\lambda) < 0$, set $U = \lambda$ otherwise set $L = \lambda$.
7. If $U - L < \epsilon$, stop and report $\lambda = (U + L)/2$. Otherwise, go to Step 5.

This algorithm is guaranteed to stop. The constant ϵ is the tolerance level set by the user or by default. Often, it is chosen to be 10^{-8} or so. In applications, we should take the scale of x_i's into consideration. If all of them are small in absolute values (after subtracting θ), λ will be larger hence the above tolerance is fine. If $x_i - \theta$ are in the order of 10^8, then to tolerance for λ must be reduce substantially, say to 10^{-16}.

To find the upper and lower limits of the confidence interval of the mean, we first note that $\bar{x}_n$ is always included in the interval. The upper and lower limits cannot exceed the smallest and the largest observed values. A simple method is to bisection the interval between $\bar{x}_n$ and $x_{(n)}$ iteratively until we find the location θ_U at which the profile likelihood ratio function equals some quantile of the chisquare distribution set according to the confidence level suggested by the user. The typical value is of course 3.84 for one-dim problem at 95% confidence level.

When X_i's are vector valued, Chen, Sitter and Wu (2002, Biometrika) showed that a revised Newton-Raphson method can be used for computing the profile likelihood ratio function for the mean. The algorithm is guaranteed to converge when the solution exists.

12.7 Empirical likelihood applied to estimating functions

We only provide very brief discussion here.

In some applications, particularly in econometrics, the parameter of interest is defined through estimating functions. Namely, if X is a sample from a population of interest, the parameter vector θ is the unique solution to

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

for some vector valued and smooth function g. In this setting, the distribution of X is left completely unspecified. Some restrictions will be introduced to permit meaningful discussion of some large sample properties.

Let the dimension of g be denoted as m and the dimension of θ be denoted as d. When $m < d$, the solution to equation $E\{g(X;\theta)\} = 0$ is likely not unique given a hypothetical distribution F of X. In this case, θ is underdefined. When $m = d$, the same equation usually has a unique solution. The parameter is then just-defined. When $m > d$, solution to $E\{g(X;\theta)\} = 0$ exists only for special F. More concretely, if an i.i.d. sample from a distribution

F is available, the corresponding estimating equation

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

may not have any solution in θ.

Given an i.i.d. sample and the corresponding model specification, how may we make inference about θ? In econometrics, researchers propose to find $\tilde{\theta}$ as the minimization solution of

$$\arg\min_{\theta} S_n^{\tau}(\theta) A S_n(\theta)$$

for some positive definite matrix A with $S_n(\theta) = \sum_{i=1}^{n} g(x_i;\theta)$. There have been a lot of discussions on the optimal choice of A so that the resulting estimator $\tilde{\theta}$ has "optimal properties'. Often, the asymptotic efficiency is a concern, and A is chosen as inverse of the variance function of S_n. This approach is called Generalized Method of Moments (GMM). Constructing the confidence interval for θ must rely on deriving the limiting distribution of θ and provide a consistence variance estimator of $\tilde{\theta}$.

In comparison, for each given value of θ, one may define a profile empirical likelihood function as

$$L_n(\theta) = \sup\{\prod p_i : \sum_{i=1}^{n} p_i g(x_i;\theta) = 0\}.$$

We omitted the requirements of $p_i > 0$ and $\sum p_i = 1$ in the writing but they are there.

For each given θ, the computation of $L_n(\theta)$ in the current case is not different from the case where θ is the population mean. We may also notice that the dimensions of g and d do not really matter. The optimal solution to the maximization problem is given by

$$p_i = \frac{1}{n[1 + \lambda^{\tau} g(x_i;\theta)]}$$

with the Lagrange multiplier λ being the solution to

$$\sum_{i=1}^{n} \frac{g(x_i;\theta)}{n[1 + \lambda^{\tau} g(x_i;\theta)]} = 0.$$

The profile empirical likelihood defined here works almost the same way as the parametric likelihood. Let $\hat{\theta}$ be the maximum empirical likelihood estimator and θ_0 be the true value of the parameter. It is known that

$$R_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} \to \chi^2_{m-d}$$

as $n \to \infty$ under some conditions. It provides a simple way to construct a likelihood interval/region and hence confidence interval/region.

The maximum empirical likelihood estimator is in general asymptotically normally distributed. Among certain type of estimators, it is also known to be "optimal". That is, it has the lowest asymptotic variance.

A much liked advantage of EL method, compared with GMM, is that one does not need to estimate the variance of $\hat{\theta}$ in order to construct confidence interval or regions of θ.

12.8 Adjusted empirical likelihood

One problem with the EL for estimating function setup is that the solution to the maximization problem may not exist. That is,

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

may not have a solution in p_i such that $p_i > 0$ and $\sum p_i = 1$. The Lagrange multiplier λ is well defined only if 0 is in the convex hull of $\{g(x_i; \theta), i = 1, \ldots, n\}$.

Thus, for each θ value given, one must first make sure $L_n(\theta)$ is actually defined. Looking for $\hat{\theta}$ is in the second step. If the set of θ, on which $L_n(\theta)$ is well defined, is empty, the rest of inference strategies falls apart.

In theory, if the model is correct, g has finite second moment, then $L_n(\theta_0)$ is well defined with probability approaching 1 as $n \to \infty$. In applications, there is no guarantee we can locate a θ-value at which $L_n(\theta)$ is well defined. In fact, it can be an issue to merely determine whether or not it is well defined.

There have been a few remedies proposed in the literature. One of them is by myself. Let us define

$$g(x_{n+1};\theta) = -a_n \bar{g}_n$$

where $\bar{g}_n = n^{-1}\sum_{i=1}^{n} g(x_i;\theta)$, for any θ, with a positive constant a_n. In this definition, we do not look for a x_{n+1} value at which the above relationship holds. We only need $g(x_{n+1};\theta)$ value.

Next, we define profile empirical likelihood as

$$L_N(\theta) = \sup\{\prod p_i : \sum_{i=1}^{N} p_i g(x_i;\theta) = 0\}.$$

with $N = n+1$. Namely, we have added a pseudo observation $g(x_{n+1};\theta)$ into the above definition of the original empirical likelihood. Note that the restrictions $p_i > 0$ and $\sum p_i = 1$ are satisfied by $p_i = a_n/c$ for $i = 1, 2, \ldots, n$ and $p_{n+1} = n/c$ and $c = na_n + n$ for the expanded data set $g_1, \ldots, g_N$. Hence, $L_N(\theta)$ is well defined for any value of θ.

Under mild conditions, the first order asymptotic properties of $L_n(\theta)$ remain valid for $L_N(\theta)$. This so-called adjusted empirical likelihood is getting a lot of attention. Read related papers yourself if you are interested.

Chapter 13

Resampling methods

13.1 Problems addressed by resampling

Discussion of statistical inference often starts with "let $x_1, \ldots, x_n$ be a set of i.i.d. observations from a parametric model $f(x;\theta)$". Ultimately, the goal of the inference to make an educated guess on the value of θ based on the belief that $f(x;\theta)$ is a member of some distribution family $\mathcal{F}$. It is believed that once the possible value of θ is well specified, the stochastic system (population) from which the data are generated would be satisfactorily characterized. The parametric assumption $f(x;\theta)$ makes the subsequent analysis simple: the stochastic system is "fully" determined once the value of θ is determined. At the same time, when we apply an inference method to a data set under a parametric assumption, there is no reason to trust the outcome of the analysis simply because we have **assumed** a parametric model for the data. Likely, we do not have much assurance that the data will behave like a sample from a normal distribution simply because we assume a normal model before the data analysis.

This discussion does not mean we should totally give up the inference methods developed for data under parametric model assumptions. My experience shows that two sample t-test, for instance, works nicely even if the data are generated from distributions far from normal. That is, the test determined by the t-distribution has accurate size and respectable power in even very wild situations. At the same time, some parametric inference methods

may have rather different behaviours than we would expect when the data are a sample from a population where the model assumption is grossly violated. In these situations, one should first determine whether there is reasonable evidence that the model assumption fits the population under investigation. If not, a nonparametric approach can be an attractive alternative. The new dilemma, however, is the complexity usually associated with nonparametric methods. It is often very difficult to determined the distribution of the statistics employed in these methods. The technically-challenging asymptotic distribution, even available, may not be a good approximation to the finite sample distribution.

To study the finite sample properties of any statistical methods, designed under parametric or nonparametric model assumptions, the resampling methods can be very effective. There might be many varieties of resampling methods, jackknife and bootstrap are among the most popular ones. These methods can often be employed with little technical work at potentially an extra cost on computation. In the age of advanced information technology, the computational cost is less and less an issue.

Let θ be some aspects of F, namely a functional. For example, the mean of F. Write $\theta = T(F)$. A natural estimator of θ is $\hat{\theta}_n = T(F_n)$ where $F_n(x)$ is the empirical distribution based on the i.i.d. sample. One should easily tell that in this case, the estimator is the sample mean.

A point estimation of θ is generally only a starting point of the statistical inference. An immediate question might be: what is the distribution of $\hat{\theta}_n$.

If F is a member of Poisson distribution and $T(F)$ is the population mean, then $\hat{\theta} = \bar{X}_n$ whose distribution is a scaled Poisson. If F is a member of normal distribution, then $\hat{\theta} = \bar{X}_n$ has normal distribution with some mean and variance.

If $T(\cdot)$ is more complex than the usual sample mean, and F is a member of a generic distribution family, the answer to "what is the distribution of $\hat{\theta}_n$" is much more complex.

Typically, if n is very large, $\hat{\theta}_n = T(F_n)$ is asymptotic normal. This partly answers the above question. Yet even if so, we are burdened at analytically obtaining the mean and variance of the asymptotic distribution.

Bootstrap and other resampling methods provide some alternative solu-

tions which are labor intensive in terms of computation, but simple in terms of mathematical derivation. Because of these properties, this line of approach is admired by many applied statisticians. At the same time, it can be abused by many who know very little on its limitations.

13.2 Resampling procedures

Since $x_1, x_2, \ldots, x_n$ are i.i.d. observations, the empirical distribution function $F_n(x)$ is a good estimate of their common distribution. Note that $F_n(x)$ is the uniform distribution on these n observed values. If these observation has ties, this interpretation is harmless. Let X^* denote a random variable with distribution F_n. That is,

$$P(X^* = x_i) = \frac{1}{n} \quad \text{for } i = 1, 2, \ldots, n.$$

In addition, let $X_1^*, \ldots, X_n^*$ be i.i.d. random variables with the same distribution as that of X^*. Let $F_n^*(x)$ be the empirical distribution based on observed values of $X_1^*, \ldots, X_n^*$. We regard F_n^* and its related entities as mirror image of F_n in the bootstrapping world.

For each parameter of interest $\theta = T(F)$, we may estimate its value by $\hat{\theta} = T(F_n)$. In the bootstrapping world, we find their images as $\theta^* = T(F_n)$ and $\hat{\theta}^* = T(F_n^*)$. The distribution of $T(F_n)$ has its bootstrap world image as distribution of $T(F_n^*)$. When $F \approx F_n$ as it is the case when n is large, we anticipate that the distribution of $T(F_n)$ is well approximated by the distribution of $T(F_n^*)$ in conditional sense. We should take note that such claims are meaningful when $T(F)$ is smooth in F in some sense.

Alone this line of thinking, the distributions of the sample mean and sample variance $\bar{X}_n$ and s_n^2 should be approximately the same as the conditional distribution of $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$ and $s_n^{2*} = \frac{1}{n-1}\sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2$. In fact, these claims extend to their joint distribution and their functions.

Instead of working hard mathematically on the distributions of the sample mean and variance and so on, we approximate them by these of their bootstrap images.

At this point, one should question: if deriving the distribution of $\bar{X}_n$ is hard, deriving the distribution of $\bar{X}_n^*$ is likely harder. This is true. However, if

we can generate a million of independent and distributionally identical copies of $\bar{X}_n^*$, the distribution of $\bar{X}_n^*$ can then be numerically determined accurately. If this idea works, we would have successfully unloaded our technical burden to computers. We will be able to work on many problems without placing restrictive normality assumption or other parametric assumptions on the populations. Remember, placing assumptions on the population does not validate these assumptions.

Bootstrapping or other resampling procedures are generally portrayed as a non-parametric method. They are used for many purposes far more than merely approximating the distributions of $\bar{X}_n$ and s_n^2. For example, the bootstrap method can be used to approximate the distribution of sample median under very general conditions. Such universality makes it a popular choice.

In some applications, a parametric model $f(x;\theta)$ itself is an acceptable assumption. Suppose $\hat{\theta}$ is the maximum likelihood estimator of θ. What is the distribution of $\hat{\theta}$? The answer based on $n \to \infty$ can also be difficulty to find based on mathematical derivation. In this case, one may study the distribution of $\hat{\theta}$, when the data are a sample from $f(x;\hat{\theta})$. To implement this idea by resampling method, one may generate samples from $f(x;\hat{\theta})$, and obtain a large number of $\hat{\theta}^*$, the MLE based on generated data sets. The empirical distribution based on $\hat{\theta}^*$ can be an accurate approximation of the distribution of $\hat{\theta}$. When the resampled data are drawn from a parametric distribution, the bootstrapping method becomes parametric bootstrap.

13.3 Bias correction

Let $\theta = T(F)$ be a parameter. Since the empirical distribution $F_n(x)$ is a good estimator of F, we have proposed to use $\hat{\theta} = T(F_n)$ to estimate θ. At the same time, the bias of $T(F_n)$ itself is a functional of F. Thus, bootstrap can be used to estimate the bias of $T(F_n)$ and subsequently reduce the bias.

Although $\mathcal{E}\{F_n(x)\} = F(x)$ for all x, it is not necessarily true that $\mathcal{E}\{T(F_n)\} = T(F)$. If so, how large is the bias of $\hat{\theta} = T(F_n)$? Let us denote the bias by $\xi = E[T(F_n)] - T(F)$.

Let $X_1^*, \ldots, X_n^*$ be i.i.d. observations from the empirical distribution F_n.

Let $F_n^*(x)$ be the corresponding empirical distribution. Subsequently, $\hat{\theta}^* = T(F_n^*)$ is then a bootstrap estimator of $\hat{\theta}$. Its (conditional) bias is given by

$$\hat{\xi}^* = E_*\{T(F_n^*)\} - T(F_n),$$

where E_* is the expectation conditional on $X_1, \ldots, X_n$. If $\hat{\xi}^*$ cannot be evaluated theoretically, we can evaluate it by simulation. Is $\hat{\xi}^*$ an good estimator of ξ?

Example 13.1 *(a) Assume $\theta = \int x dF(x)$ and $E|X| < \infty$. Consequently, the bias $\xi = 0$. At the same time,*

$$\mathcal{E}_* T(F_n^*) = \mathcal{E}_*\{n^{-1}\sum_{i=1}^n X_i^*\} = n^{-1}\sum_{i=1}^n X_i = T(F_n).$$

Thus, we also have $\hat{\xi}^* = 0$. This result shows that $\hat{\xi}^*$ works fine as an estimator of ξ. Of course, in this example, the exercise does not lead to any useful results.

(b) Let us tonsider the parameter estimation of

$$\theta = T(F) = [\int x dF(x)]^2.$$

Assume that $\sigma^2 = var(X_1) < \infty$. We have $T(F_n) = [\bar{X}_n]^2$ and its bias is given by $\xi = n^{-1}\sigma^2$. The conditional expectation of $T(F_n^*)$ given F_n is given by

$$\mathcal{E}_* T(F_n^*) = \mathcal{E}_*\{n^{-1}\sum_{i=1}^n X_i^*\}^2 = \{\mathcal{E}_* X_1^*\}^2 + n^{-2}\mathrm{VAR}_*(X_1^*) = T(F_n) + (n-1)s_n^2/n^2.$$

Thus, if we estimate ξ by $\xi_i^* = E_* T(F_n^*) - T(F_n)$, we have $\hat{\xi}^* = (n-1)s_n^2/n^2$. This is a very reasonable estimator of ξ though we certainly do not have to go over bootstrap resampling procedure to find out.

13.4 Variance estimation

Consider the problem of assessing the variance of $T(F_n)$. The bootstrap method estimates the variance of $T(F_n)$ by the conditional variance of $T(F_n^*)$,

where F_n^* is the empirical distribution based on an i.i.d. sample from the distribution F_n.

Example 13.2 *(a) Let the parameter of interest be* $\theta = T(F) = \int x dF$ *again. It is seen that* $\hat{\theta} = T(F_n) = \bar{X}_n$. *Let us work as if we do not have a good idea on its variance. Consequently, we use resampling method to estimate its variance. Take an i.i.d. samples from the empirical distribution* F_n. *Let* $\bar{X}_n^*$ *be the resulting sample mean. We now use the conditional variance of* $\bar{X}_n^*$ *to estimate the variance of* $\bar{x}_n$.

We can easily calculate the conditional variance as

$$\text{VAR}_*(\bar{X}_n^*) = n^{-1}\text{VAR}_*(X_1^*) = n^{-2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Recall the true variance of $\bar{X}_n$ *is* $n^{-1}\sigma^2$ *where* $\sigma^2 = \text{VAR}(X_1)$. *The bootstrap variance estimation is* $n^{-1}s_n^2 + O(n^{-2})$. *Clearly, we have*

$$\frac{\text{VAR}_*(\bar{X}_n^*)}{\text{VAR}(\bar{X}_n)} \to 1$$

almost surely as $n \to \infty$. *This result shows that the bootstrap variance estimator is well justified.*

It is important to realize that $\text{VAR}(\bar{X}_n) \to 0$ *as* $n \to \infty$. *Hence, even if a variance estimator* $\hat{\text{VAR}}(\bar{X}_n)$ *makes*

$$\hat{\text{VAR}}(\bar{X}_n) - \text{VAR}(\bar{X}_n) \to 0$$

almost surely, this property alone does not make it a good estimator.

(b) Let the parameter of interest be $\theta = \{\int x dF\}^2$. *Its natural estimator is* $\hat{\theta} = \bar{X}_n^2$. *How large is the variance of* $\hat{\theta}$*? Let us again pretend that we do not have a good idea on how to estimate its variance. Consequently, we resort to bootstrap.*

For the bootstrap method, it is easy to get

$$\begin{aligned}\text{VAR}_*(\{\bar{X}_n^*\}^2) &= 4\bar{X}_n^2 E_*\{\bar{X}_n^* - \bar{X}_n\}^2 + E_*\{\bar{X}_n^* - \bar{X}_n\}^4 \\ &\quad -\{E_*[\bar{X}_n^* - \bar{X}_n]^2\}^2 + 4\bar{X}_n E_*\{\bar{X}_n^* - \bar{X}_n\}^3.\end{aligned}$$

The order of the last three terms are $O_p(n^{-2})$. The order of the first one is $O_p(n^{-1})$ when the true mean is not zero. Thus, the leading term in this bootstrap variance estimator is $(4\bar{X}_n^2/n^2)\sum_{i=1}^n (X_i - \bar{X}_n)^2$. This marches the approximate variance of $\bar{X}_n^2$ which equals $(4\mu^2\sigma^2)/n$.

In both examples, we analytically obtained the properties of the bootstrap method for bias and variance estimation of estimators in the form of $T(F_n)$ for parameter $T(F)$. Analytical derivation is not always feasible. For instance, suppose θ is the location parameter in Cauchy distribution, we will not be able to find $\text{VAR}_*(T(F_n^*))$ by theoretical computation. Instead, computer simulation is likely the only option which can be carried out as follows.

First, draw an i.i.d. sample of size n $x_1^*, \ldots, x_n^*$ from F_n based on some computer package. Compute, based on bth sample,

$$\hat{\theta}_b^* = T(F_n^*)$$

where $F_n^*(x) = n^{-1}\sum_{i=1}^n I(x_i^* \le x)$.

Next, define the simulated $\text{VAR}_*(T(F_n^*))$ value to be

$$v_*^2 = \frac{1}{B-1}\sum_{b=1}^B \{\hat{\theta}_b^* - \bar{\theta}^*\}^2$$

where $\bar{\theta}^* = B^{-1}\sum_{b=1}^B \hat{\theta}_b^*$. If θ is a vector, we need to modify the above formula for the variance-covariance matrix.

Under some conditions, v_*^2 is a consistent estimator of $\text{VAR}(T(F_n))$. Yet we must be more cautious on the meaning of consistency:

$$v_*^2/\text{VAR}(T(F_n)) \to 1$$

in some modes.

One many define the bootstrap variance estimator to be

$$\tilde{v}_*^2 = \frac{1}{B-1}\sum_{b=1}^B \{\hat{\theta}_b^* - \hat{\theta}\}^2.$$

Since the difference between $\hat{\theta}$ and $\bar{\theta}^*$ is likely very small in asymptotic argument, both of them are well justified. None of them can be judged as "wrong" as many would like to ask.

In addition, simulation study will likely find situations where v_*^2 is more accurate and other situations where $\tilde{v}_*^2$ is superior.

In summary, being a statistician does not make you an authority to decide between these estimators. We do notice that $\tilde{v}_*^2$ resembles mean square error. It therefore takes a larger value. If one likes to have a more conservative statistical procedure, using $\tilde{v}_*^2$ a good choice.

13.5 The cumulative distribution function

Consider the problem of approximating the distribution of $T(F_n)$ by that of $T(F_n^*)$. The idea here is the same as the one for variance estimation. We hope that the conditional distribution of $T(F_n^*)$ is close to the distribution of $T(F_n)$.

Consider the simplest situation where the parameter to be estimated is $\theta = T(F) == \int x dF$. The estimator of θ is $\bar{X}_n$, and we aim at estimating the cumulative distribution function of $\bar{X}_n$. Assume without loss of generality that the true values $\theta = 0$ and $\sigma^2 = \text{VAR}(X_1) = 1$.

Under the assumption of the finite second moment, $\sqrt{n}\bar{X}_n$ is asymptotic normal. This fact pretty much tells us to not bother at estimating its distribution. Nevertheless, if we insists on using bootstrap to estimate the distribution of $\bar{X}_n$, we should have, as $n \to \infty$,

$$P(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \le x | F_n) \to P(Z \le x) \quad \text{almost surely.}$$

Note that this is a limit where both the event under investigation and the condition are changing when n increases. As n increases, the central limit theorem for triangular array can be used to obtain the above result.

To prove the asymptotic normality, Berry-Esseen bound is most simple though at a relatively stronger conditions. For any sample sizes, applying this bound gives

$$\sup_x |P(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \le x | F_n) - P(Z \le x)| \le c \frac{\mathcal{E}_* |X_1^* - \bar{X}_n|^3}{\sqrt{n}[\mathcal{E}_* |X_1 - \bar{X}_n|^2]^{3/2}}.$$

Thus, the asymptotic normality is valid when

$$\frac{\mathcal{E}_* |X_1^* - \bar{X}_n|^3}{[\mathcal{E}_* |X_1 - \bar{X}_n|^2]^{3/2}} = o(n^{1/2})$$

almost surely.

Suppose the model satisfies $\mathcal{E}|X_1|^3 < \infty$. In this case, we have

$$\mathcal{E}_*|X_1^* - \bar{X}_n|^3 = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_n|^3 \to \mathcal{E}|X_1|^3, \quad \text{almost surely.}$$

$$E_*|X_1^* - \bar{X}_n|^2 = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_n|^2 \to \sigma^2 = 1 \quad \text{almost surely.}$$

Thus, it is trivial to find that

$$\frac{\mathcal{E}_*|X_1^* - \bar{X}_n|^3}{[\mathcal{E}_*|X_1 - \bar{X}_n|^2]^{3/2}} \to \mathcal{E}|X_1|^3.$$

Hence, when $\mathcal{E}|X_1|^3 < \infty$, the (conditional) asymptotic normality is proved. The simple proof is benefitted from an unnecessarily strong assumption on the finiteness of the third moment.

A generalization can be easily made. If $g(\bar{X}_n)$ is a smooth function of $\bar{X}_n$, then $g(\bar{X}_n)$ is asymptotic normal. By the same logic, $g(\bar{X}_n^*)$ is also asymptotically normal conditional on F_n. Thus, the conditional distribution of $g(\bar{X}_n^*)$ still marches that of $g(\bar{X}_n)$.

Although the above example is very supportive on the usefulness of the bootstrap method, it is not without its limitations. For the sample mean, its asymptotic normality can be established easily. The calculation of the limiting distribution is also very simple. Why should we bootstrap in these simple situations? In situations where the asymptotic become complex, do we have a good theory to support the bootstrap?

One crucial justification of using bootstrap method comes from Singh (1981). There are many results contained in this paper. Here I only pick up a relatively simple case.

Theorem 13.1 *(Singh, 1981). Assume $X_1, \ldots, X_n$ are i.i.d. samples from F. Assume $\mathcal{E}X_1 = 0$, $\sigma^2 = \text{VAR}(X_1) > 0$, and $\mathcal{E}|X_1|^3 < \infty$. Let $\bar{X}_n$ be the sample mean and s_n^2 be the sample variance. In addition, let $\bar{X}_n^*$ be the bootstrap sample mean. Then*

$$\sup_x |P(\frac{\sqrt{n}\bar{X}_n}{\sigma} \le x) - P(\frac{\sqrt{n}[\bar{X}_n^* - \bar{X}_n]}{s_n} \le x|F_n)| = O(n^{-1/2})$$

almost surely. If F is a continuous distribution, then

$$\sup_x |P(\frac{\sqrt{n}\bar{X}_n}{\sigma} \leq x) - P(\frac{\sqrt{n}[\bar{X}_n^* - \bar{X}_n]}{s_n} \leq x|F_n)| = o(n^{-1/2}).$$

We will not go over its proof. The result shows that the bootstrapping approximation has better precision than the normal approximation. This is a surprising good news.

The sampling procedure for approximating c.d.f. is very simple. First, we draw an i.i.d. sample of size n $X_1^*, \ldots, X_n^*$ from F_n based on some computer software package. This will be repeated B times. Next we compute, based on bth sample,

$$\hat{\theta}_b^* = T(F_n^*)$$

where $F_n^*(x) = n^{-1}\sum_{i=1}^n I(x_i^* \leq x)$. The last step is to define the estimated cumulative distribution function to be

$$\hat{H}_n(t) = B^{-1}\sum_{b=1}^B I(\hat{\theta}_b^* \leq t).$$

Needless to say, under some conditions, $\hat{H}_n(t)$ is consistent for $H(t) = P(\hat{\theta} \leq t)$.

13.6 A few recipes of confidence limits

Let $\hat{\theta}$ be an estimator of parameter θ.

Percentile method Consider the case when $\hat{\theta} - \theta$ is more or less a pivotal quantity. Suppose that its distribution is given by $H(x)$, namely,

$$P(\hat{\theta} - \theta \leq x) = H(x).$$

If so, a lower confidence limit for θ of size $1-\alpha$ is given by $H^{-1}(\alpha)$, the αth quantile of $H(x)$. This is seen by

$$P(\hat{\theta} - \theta \geq H^{-1}(\alpha)) = 1 - \alpha$$

when $H(\cdot)$ is a strictly increasing function.

Let $\hat{H}(x)$ be an estimator of $H(x)$. Define

$$\underline{\theta}_{BP} = \hat{\theta} + \inf_t\{\hat{H}_n(t) \geq \alpha\} = \hat{\theta} + \hat{H}^{-1}(\alpha).$$

This is an approximate lower confidence bound for θ because

$$P(\theta > \underline{\theta}_{BP}) = P(\hat{\theta} - \theta \geq \hat{H}_n^{-1}(\theta)) \approx P(\hat{\theta} - \theta \geq H^{-1}(\theta)) = 1 - \alpha.$$

Computing confidence lower limit based on the above approach is generally called percentile method. The subscript BP is used for "bootstrap percentile" though we motivated this lower limit without bootstrapping procedure.

Ordinary and studentized methods Consider the case when some i.i.d. observations are obtained from some distribution and we have estimators $\hat{\theta}$ and $\hat{\sigma}$ for θ and the standard error of $\sqrt{n}\hat{\theta}$. Note that the later is not the population standard error. In many cases, it might be more realistic that

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma}; \quad \frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\sigma}}$$

are approximately pivotal quantities. If they are, without approximations, we may define

$$H(x) = P\{\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma} \leq x\}$$

and

$$K(y) = P\{\frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\sigma}} \leq y\}.$$

With complete information of $H(x)$ and $K(y)$, constructing confidence intervals for θ is a simple task.

We further notice that the task is reduced to find upper and lower confidence limits. Let $x_\alpha = H^{-1}(\alpha)$ and $y_\alpha = K^{-1}(\alpha)$ so that $1-\alpha$ is the targeted level of confidence. Depending on whether we have knowledge on H or on K, the lower confidence limits are respectively

$$\hat{\theta}_{ord}(\alpha) = \hat{\theta} - x_{1-\alpha}\sigma/\sqrt{n},$$

and

$$\hat{\theta}_{stud}(\alpha) = \hat{\theta} - y_{1-\alpha}\hat{\sigma}/\sqrt{n}.$$

Both of them have the format we presented in an early chapter. The subscripts, ord and stud, are abbreviations for *ordinary* and *studentized*. They would have been $z_{1-\alpha}$ or $t_{1-\alpha}$ when H and K are c.d.f. of normal and t distributions.

Hybrid and backward methods. As it is well known, when the sample size is large, $\hat{\sigma} \approx \sigma$ and hence $x_\alpha \approx y_\alpha$. One may therefore use a hybrid lower confidence limit:

$$\hat{\theta}_{hyb} = \hat{\theta} - x_{1-\alpha}\hat{\sigma}/\sqrt{n}.$$

This can be compared to the situation where quantile of t-distribution should be used, yet we mistakenly use the quantile of the normal distribution.

Under normal distribution, $z_\alpha = -z_{1-\alpha}$ because the normal distribution is symmetric. If H is symmetric, then $x_\alpha = -x_{1-\alpha}$ for the same reason. Hence, when H is believed symmetric, we may use another lower confidence limit:

$$\hat{\theta}_{back}(\alpha) = \hat{\theta} + \hat{\sigma}x_\alpha\hat{\sigma}/\sqrt{n}.$$

It is clearly confusing to present so many possibilities. Which one is correct? The answer depends on what we mean by "correct". If any (random) interval which covers the true value of θ with probability $1-\alpha+o(1)$, we feel that they are okay, or "correct". When the sizes of these intervals are not taken into consideration, we may want to examine $o(1)$ term in the coverage probability. Let leave this issue to the next section.

13.7 Implementation based on resampling

Having complete knowledge of $H(x)$ and $K(x)$ is not possible. More often than not, they are also dependent on unknown parameter values. Nonetheless, bootstrap simulation can be used to find estimate of H and K, when the population distribution is given by F and the parameter θ is a functional of F.

We will see what it means by "can be estimated". The distribution H is estimated by

$$\hat{H}(x) = P(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \le \hat{\sigma}x|F_n).$$

The distribution K is estimated by

$$\hat{K}(x) = P(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq \hat{\sigma}^* x | F_n).$$

Once they are obtained via bootstrap simulation, we define $\hat{x}_\alpha = \hat{H}^{-1}(\alpha)$ and $\hat{y}_\alpha = \hat{K}^{-1}(\alpha)$. Every lower confidence limits proposed in the last section is transformed to bootstrap lower confidence limits by putting a hat on either x_α or y_α.

Now, which one makes $o(1)$ inside the coverage probability $1 - \alpha + o(1)$ the smallest? Peter Hall (AOS some year) had a discussion paper specifically for this problem. The technical discussion is too complex for this course. The results are not that insightful either. Without going back to the paper itself, I put down my unverified memory here: the studentized approach together with bootstrap resampling has this $o(1)$ reduced to $O(n^{-1})$. Without studentization, this $o(1)$ is $o(n^{-1/2})$. Both conclusions are obtained under the assumption that $\hat{\theta}$ is a smooth function of the sample mean $\bar{X}_n$, after being broadly interpreted. For instance,

$$\theta = \frac{\mu}{\sigma}$$

has its estimator given by

$$\hat{\theta} = \frac{\bar{x}_n}{\sqrt{\overline{x^2} - (\bar{x})^2}}.$$

This estimator is a smooth function of the sample mean in (x_1, x_i^2).

13.8 A word of caution

The bootstrap method is generally used to simulate the variance and the distribution of a point estimator. Based on bootstrap simulation, we can often subsequently make inference on various parameters. Most noticeably, the results are subsequently used to construct confidence interval for parameter θ and subsequently test the hypothesis such as $\theta = \theta_0$. We are often freed from complex technical issues.

At the same time, one has to have a good point estimator $\hat{\theta}$ before the resampling procedure can even start. The statistical properties of the corresponding data analysis is largely determined by that of $\hat{\theta}$. The resampling methods help to determine these properties. They do not induce good properties into these procedures.

There is no guarantee that the resampling methods always lead to valid statistical inferences. By this statement, for instance, a $1-\alpha$ level confidence interval may have far lower coverage probability and the under-coverage problem does not go away when the sample size increases. The theory in mathematical statistics cannot be thrown out simply because the resampling procedure is powerful at freeing us from the task of a lot of technical derivations.

Chapter 14

Multiple comparison

One-way ANOVA is a typical method to compare a number of treatments in terms of a specific measurement of some experimental outcomes. For example, an experiment might be designed to compare the volumes of harvest when different fertilizers are used.

Let the number of treatments be k. Let $N = n_1+n_2+\cdots+n_k$ experimental units randomly assigned to k treatments with $n_1, n_2, \ldots, n_k$ units each. Let the response variable be denoted as y. Suppose the jth treatment is replicated n_j times. The outcomes can be displayed as

$$\begin{gathered} y_{11}, y_{12}, \ldots, y_{1n_1}; \\ y_{21}, y_{22}, \ldots, y_{2n_2}; \\ \ldots, \ldots \\ y_{k1}, y_{k2}, \ldots, y_{kn_k}. \end{gathered}$$

The outcome y_{ij} is the reading of the unit assigned to the ith treatment and the jth replication.

A linear model for this set up is

$$y_{ij} = \eta + \tau_i + \epsilon_{ij}$$

for $i = 1, 2, \ldots, k$, and $j = 1, 2, \ldots, n_i$. We assume η is the overall mean, τ_i is the mean response from the ith treatment after subtracting the overall

mean. The error term ϵ_{ij} is what cannot be explained by **the treatment effect** τ_i. The statistical analysis is often done based on the assumption that

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

and they are assumed independent of each other. The normality assumption and the equal variance assumption are the ones that may be violated in the real world. The decomposition of the treatment means is always feasible.

14.1 Analysis of variance for one-way layout.

Let

$$\bar{y}_{\cdot\cdot} = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

be the over all sample mean. Let

$$\bar{y}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$$

be the sample mean restricted to samples from the ith treatment. In general, whenever an index is replaced by a dot, the resulting notation represents the sample mean over the corresponding index. For example, $\bar{y}_{\cdot 1}$ would be the average of $y_{11}, y_{21}, \ldots, y_{k1}$.

Now we may decompose the response as

$$y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot})$$

which will also be written as

$$y_{ij} = \hat{\eta} + \hat{\tau}_i + r_{ij}.$$

These quantities marked with hats are estimates/estimators of the corresponding parameters in the linear model.

The sum of squares in $(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$ represents the variation in the mean responses between different levels of the factor (or between treatments), while

$r_{ij} = (y_{ij} - \bar{y}_{i\cdot})$ represents the residual variations. The residual variation is the variation not explainable by the treatment effect.

The analysis of variance aims to compare the relative sizes of these two sources of variation. The resulting ANOVA table is as follows.

ANOVA for One-Way Layout

Source	D.F.	SS
Treatment	$k-1$	$\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$
Residual	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$

One may notice that each sum of squares contain N terms, sometimes obtained by duplicating entrances. Consider the hypothesis test problem on the null hypothesis:

$$H_0 : \ \tau_1 = \tau_2 = \cdots = \tau_k.$$

The alternative hypothesis is that not all population means are equal. The test statistic we commonly use is

$$F = \frac{(k-1)^{-1}\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{(N-k)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2}$$

which, under normality/equal-variance and H_0, has F-distribution with $k-1$ and $N-k$ degrees of freedom. This might be an opportunity for us to refresh the memory on the desired properties a test statistics. It is also a useful exercise to refresh the memory of what a UMPU test is.

If H_0 is true, the randomness of the test statistic F is completely determined and it does not depend on any external factors. We first compute the p-value:

$$p = P(F > F_{obs}).$$

Rejecting H_0 when $p < 0.05$ is a common practice.

14.2 Multiple comparison

Once (and if) the null model is rejected, it is natural to ask: which pair of treatments are the culprits that lead to the rejection? The rejection may be

caused by a single treatment that is substantially different from the rest. It may also be caused by smaller differences between all treatments. Of course, the rejection may be erroneous.

Regardless of these possibilities, let us ask the question on which pairs of treatments are significantly different? The technique used to addressing this question is called **multiple comparison**, because many pairs are being compared **simultaneously**.

Borrowing the idea of two-sample test, we may define

$$t_{ij} = \frac{\bar{y}_{j\cdot} - \bar{y}_{i\cdot}}{\sqrt{(1/n_i + 1/n_j)\hat{\sigma}^2}}$$

where $\hat{\sigma}^2$ is the variance estimator from the ANOVA table. The denominator is different from the usual two-sample t-test. It pools information from all k-treatments. Each t_{ij} has t-distribution with $N - k$ degrees of freedom. Hence, if a level-α test is desired, we may reject $H_0 : \mu_i = \mu_j$ when

$$|t_{ij}| > t(1 - \alpha/2; N - k).$$

This test has probability α to falsely reject the hypothesis that the corresponding pair of treatment means are equal.

Suppose we set $\alpha = 0.05$ as in common practice, and $k = 5$. There will be 10 such pairs of treatments. Even if all pairs of treatments are not different, there will a chance about 5% to declare any one of them significant. The chance of declaring one of them is significant by a simple t-test is likely much larger, approaching possibly 50%. Such analysis is clearly not acceptable.

14.3 The Bonferroni Method

. To address the problem of inflated type I error in multiple comparison, we could simply set up a high standard for every pair of i and j such that the overall type I error is guaranteed to be lower than pre-specified value α. Let $k' = k(k-1)/2$ be the number of possible treatment pairs. We may reject $H_{ij} : \mu_i = \mu_j$ only if

$$|t_{ij}| > t(1 - \alpha/2k'; N - k).$$

Since the probability that any pair of treatments wrongfully judged different is no more than α/k' (note this is two-sided test), and there are k' such pairs, it is simple to seen that the chance that at least one pair to be declared different, when none of them are difference, is controlled tightly by $100\alpha\%$.

14.4 Tukey Method

Particularly when k is large (5 or more), the Bonferroni method is too conservative. It means that the actual type I error can be far lower than the targeted level $100\alpha\%$. Having a small type I error is not strictly wrong in term of being a valid test. The real drawback of such a test is that this increases the type II error. When k is large, the statistical power of detecting any departure from the null hypothesis is too small if the conservative Bonferroni method is used. If such a method is used as standard, scientists have to work unjustifiably harder to prove their point.

Let us define

$$t^* = \sqrt{2}\max\{|t_{ij}|\}.$$

It is seen that t^* has a distribution which does not depend on any unknown parameters. However, it does depend on k and $N-k$, and in fact, also on how N units are divided and assigned to k treatments. It is a test statistic with almost all desirable properties we specified in an early chapter. Unlike t-distribution, however, the c.d.f. of this distribution is not as well documented. When all n_j are equal, the distribution might be named after Tukey. Let $\text{qtukey}(1-\alpha; k, N-k)$ be its upper quantile when all n_i are equal. That is, under that restriction,

$$P\{t^* > \text{qtukey}(1-\alpha; k, N-k)\} = \alpha$$

for any $\alpha \in (0,1)$. We may reject the hypothesis that the i and j pair of treatments have equal mean when $|t_{ij}| > \text{qtukey}(1-\alpha; k, N-k)/\sqrt{2}$.

The type I error of this approach is only approximate when $n_1, \cdots, n_k$ are not all equal. In fact, it is bounded by α which is proved by someone based on my memory.

My observation: Tukey's method is not so much as a new method. It simply requires us to use a critical value so that the probability that wrongfully reject any pair of $\tau_i = \tau_j$ is below $100\alpha\%$.

Pitfalls of Bonferroni and Tukey Methods In the case of Bonferroni, the adjustment is too conservative. If we hope to test 1000 hypotheses based on a single data set, then the significance level would be placed at 0.005%. If it is applied to a t-test with $n = 20$ degrees of freedom, the critical value is 5.134. This is to be compared to 2.09 if only one hypothesis is being tested. The actual type I error of the test is likely much lower (Assignment problem).

As a side remark, in statistical consulting practice, we often look into many aspects of data. Based on what we spot, various hypotheses are proposed and then tested. In the end, we report the p-value on the hypothesis that is below 0.05 (or several). The practice strictly violates the statistical principle we preach. Nonetheless, statisticians do it routinely and our collaborators will not be pleased otherwise.

In the case of Tukey Method, I feel that it is specifically designed for one-way anova. I am not sure if there is one for other situations. Regardless, my understanding is that it simply requires statisticians to make sure the probability of wrongfully rejecting even one of many hypotheses is below α, the pre-specified level. This principle leads to a technical issue: we may not be able to find even a well approximated critical value.

14.5 New problems and FDR

In modern statistical applications, we are confronted with a problem that is radically different from the one-way anova. Due to bio and info technical advances, we can now cost-effectively and timely taking measurements of thousands of genes expression levels from each subject. It is of interest to identify some genes whose expression levels are different on different groups of people. Typically, one group is made of health controls and another group is made of patients of a specific type of disease. The genes that are significantly differentially expressed might be related to the disease.

There are two aspects of this new problem.

First, if 500,000 genes are inspected on 50+50 subjects, even if we use $\alpha = 0.001$ to test for each hypothesis that a gene is significantly differentially expressed, and that none of them are differentially expressed, 500 of them will likely be found statistically significant. This is bad.

Second, suppose a handful of genes are indeed differentially expressed but the differences are not exceedingly large. Applying Bonferroni method likely results in none of them judged significant. The high standard set by Bonferroni method may fail the researchers for this wrong reason.

The dilemma seems solved by giving up the notion of type I error. When thousands and thousands hypotheses are examined simultaneously, we probably should not mind to have a larger probably of "wrongfully declare a few genes significantly differently expressed". Rather, we should probably ask: among many genes judged significantly differently expressed, what percentage of them are falsely significant?

Because "rejecting a null hypothesis" in such context is regarded as a scientific discovery, the percentage of false significant outcomes among all significant outcomes is called "false discovery rate". In such applications, controlling the false discovery rate is regarded as a better principle. A widely accepted standard is 5%.

In comparison, the classical practice of controlling the overall type I error is renamed as "family-wise error rate".

I feel that there is a need to be reminded about the difference between "statistical significance" and the "real world" significance here. How large a difference in the expression levels is scientifically significant should be judged by scientist. When two expression levels are judged statistically significantly different, it means are have sufficient statistical evidence to declare that difference is likely genuine. However, the magnitude of the difference could be so small that it is scientifically meaningless.

14.6 Method of Benjamini and Hochberg

We will only discuss the result of Benjamini and Hochberg (1995, JRSSB). There have been a lot of new developments and I have not followed them very closely.

False discovery rate. Suppose m hypotheses are being tested. Let m_0 denote the number of them that are true. Let R the be number of hypothesis rejected. Note that R is random.

We have decomposition

$$m_0 = U + V$$

with U of them are tested non-significant, and V of them are tested significant.

Similarly, $m - m_0 = T + S$: T of them are tested non-significant, and S of them are tested significant.

The total number of hypotheses tested significant is $R = V + S$. The total number of hypotheses tested non significant is $m - R = U + T$.

When $R > 0$, the percentage of false discovery is V/R. When $R = 0$, there cannot be any false discovery. Thus, they propose to define

$$Q = \frac{V}{V+S} I(V + S > 0).$$

Clearly, this value is not observed and is random in any applications.

The false discovery rate (FDR) is defined to be

$$Q_e = E(Q).$$

In comparison, the type I error in the current situation is also called family-wise error rate (FWER). It measures the probability of having at least one hypotheses rejected when all of them are truthful.

According to Benjamini and Hochberg (direct quote):

(a) If all null hypotheses are true, the FDR is equivalent to the FWER: in this case $s = 0$ and $v = r$, so if $v = 0$ then $Q = 0$, and if $v > 0$ then $Q = 1$, leading to

$$P(V \geq 1) = E(Q) = Q_e.$$

Therefore control of the FDR implies the control of the FWER in the weak sense.

(b) When $m_0 < m$, the FDR is smaller than or equal to the FWER: in this case, if $v > 0$ then $v/r \leq 1$, leading to $I(V \geq 1) \geq Q$. Taking expectations on both sides we obtain $P(V \geq 1) \geq Q_e$, and the two can be quite different.

As a result, any procedure that controls the FWER also controls the FDR. However, if a procedure controls the FDR only, it can be less stringent, and a gain in power may be expected. In particular, the larger the number of the false null hypotheses is, the larger S tends to be, and so is the difference between the error rates. As a result, the potential for increase in power is larger when more of the hypotheses are non-true.

14.7 How to apply this principle to applied problems?

Suppose the hypotheses to be tested are $H_1, H_2, \ldots, H_m$. Whatever methods are used, the outcome of each test is summarized by a p-value: $P_1, \ldots, P_m$. Sorting these values to get $P_0^{(1)} \leq P_0^{(2)} \leq \cdots \leq P_0^{(m)}$. Their corresponding hypotheses are denoted as $H_0^{(i)}$ accordingly. Select an upper bound for the false discovery rate and denote it as q^*.

The BH procedure:

Step I Let k be the largest i for which $P_{(i)} \leq (i/m)q^*$;

Step II Reject all $H_{(i)}$, $i = 1, 2, \ldots, k$.

Numerically, the BH procedure can be carried out as follows

- If $p_{(m)} \leq q^*$, reject all null hypotheses and stop;
- else if $p_{(m-1)} \leq \frac{m-1}{m}q^*$, reject these $(m-1)$ null hypotheses and stop;
- else if $p_{(m-2)} \leq \frac{m-2}{m}q^*$, reject these $(m-2)$ null hypotheses and stop;
- Continue the above process until the last step: if $p_{(1)} \leq (1/m)q^*$, reject $H_0^{(1)}$ step;
- else, reject none and terminate.

Moral of this procedure: for the targeted application, it is not a serious issue if one falsely declare 10 genes are differentially expressed for diabetes

patients when 2 of them are not. We can figure out the true set subsequently. The procedure is more effective than to declare none of them are significantly differentially expressed.

Suppose we choose $q^* = 0.05$. The procedure will have at least one gene declared significantly differentially expressed when

$$p_{(1)} \leq 0.05/m.$$

Thus, if the Bonferroni's method reject "all H_0's are true", then at least one of them is rejected by the Benjamini-Hochberg procedure. The new procedure may rejects many individual H_0's.

For instance, the Bonferroni's method rejects $H_0^{(2)}$ only if

$$p_{(1)} \leq p_{(2)} \leq 0.05/m$$

but the FDR method will do so when

$$p_{(1)} \leq p_{(2)} \leq (2/m) \times 0.05.$$

Hence, FDR method will have more hypotheses rejected in long run. Rejecting both $H_0^{(1)}$ and $H_0^{(2)}$ requires only $p_{(2)} \leq (2/m) * q^*$.

14.8 Theory and its proof

Theorem 14.1 *For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.*

Remark: "independent test statistics" implies that the p-values are independent of each other, when they are regarded as random variables. When a null hypothesis is true, its corresponding p-value, however it is obtained as long as it is valid, has uniform [0, 1] distribution.

Lemma 14.1 *For any $0 \leq m_0 \leq m$, independent p-values corresponding to true null hypotheses, and for any values that the $m_1 = m - m_0$ p-values corresponding to the false null hypotheses can take, the multiple-testing procedure defined by procedure satisfies the inequality*

$$\mathcal{E}\{Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq (m_0/m)q^*.$$

Interpreting this lemma: Suppose that m_1 of the hypotheses are false. Whatever the joint distribution of their corresponding p-values, integrating inequality in the lemma we obtain

$$\mathcal{E}(Q) \leq (m_0/m)q^* \leq q^*$$

and the FDR is controlled.

Namely, the theorem is implied by this lemma.

The independence of the test statistics corresponding to the false null hypotheses is not needed for the proof of the theorem.

Proof of the Lemma. Recall m is the number of hypotheses; m_0 is the number of true hypotheses.

Denote $H_0^{(i)}$ and $P'_{(i)}$, $i = 1, 2, \ldots, m_0$ the true null hypotheses and their p-values, with p-values in increasing order. $P'_{(i)}$, $i = 1, 2, \ldots, m_0$ are order statistics of m_0 iid uniform [0, 1] random variables.

Denote false null hypotheses as $H_f^{(i)}$: $i = m_0 + 1, m_0 + 2, \ldots, m$. Their p-values will be denoted as P_i, capitalized P and indexed by i. Their values are denoted as $p_1 \leq p_2 \leq \cdot \leq p_{m_1}$.

The proof is obtained by using mathematical induction. We work on a few simple cases first before truly starting the induction.

Case I: The case $m = 1$ is immediate.

(a) $m_1 = 1$ so that $m_0 = 0$. Hence, $Q \equiv 0$ and

$$E(Q|P_1) = 0 \leq \frac{m_0}{m}q^*.$$

(b) $m_1 = 0$ so that $m_0 = 1$. Hence

$$Q = I(P'_{(1)} < q^*).$$

Thus, there is nothing to condition on. We have

$$E(Q) = P(P'_{(1)} < q^*) = q^* = \frac{m_0}{m}q^*$$

Combining (a) and (b), we find the conclusion of the lemma is true for the case where $m = 1$.

Case II: The case $m = 2$.

(a) $m_1 = 2$ so that $m_0 = 0$. In this case, $Q \equiv 0$ and

$$\mathcal{E}(Q|P_1, P_2) = 0 \leq \frac{m_0}{m} q^*.$$

(b) $m_1 = 1$ so that $m_0 = 1$. In this case, Q can take values 0, 1/2, and 1.

When $P_1 > q^*$, $H_f^{(2)}$ is never rejected. $H_0^{(1)}$ is rejected when $P'_{(1)} < 0.5q^*$. Hence, we have

$$\mathcal{E}(Q|P_1 > q^*) = P(P'_{(1)} \leq 0.5q^*) = 0.5q^* = \frac{m_0}{m} q^*;$$

When $P_1 < q^*$, both $H_0^{(1)}$ and $H_f^{(2)}$ are rejected if $p_{(1)} < q^*$. When this happens, $Q = 0.5$; otherwise, $Q = 0$. Hence,

$$\mathcal{E}(Q|P_1 < q^*) = (0.5)P(P'_{(1)} < q^*|P_1 < q^*) = 0.5q^* = \frac{m_0}{m} q^*.$$

(c) $m_1 = 0$ so that $m_0 = 2$. In this case, any rejection leads to $Q = 1$. There is nothing to be conditioned on. Hence,

$$\begin{aligned}
\mathcal{E}(Q) &= P\{P'_{(1)} < (1/2)q^* \text{ or } P'_{(2)} < (2/2)q^*\} \\
&= P\{P'_{(1)} < (1/2)q^*\} + P\{P'_{(1)} > (1/2)q^*, P'_{(2)} < (2/2)q^*\} \\
&= 1 - (1 - .5q^*)^2 + (0.5q^*)^2 \\
&= q^* = (m_0/m)q^*.
\end{aligned}$$

Combining (a), (b) and (c), we find the conclusion of the lemma is true for the case where $m = 2$.

Induction assumption Assume that the lemma is true for any $m \leq N - 1$. We work on proving this lemma when $m = N$.

Suppose $m_0 = 0$ so that all null hypotheses are false. The false discovery rate $Q \equiv 0$. Hence,

$$\mathcal{E}\{Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) = 0 \leq (m_0/m)q^*.$$

That is, the lemma is true when $m = k$ and $m_0 = 0$.

Thus, we need only discuss the situation where $m = k$ and $m_0 > 0$.

Let j_0 be the largest $0 \leq j \leq m_1$ satisfying

$$p_j \leq \frac{m_0 + j}{N} q^*.$$

(These are p-value for false null hypotheses). Denote

$$p'' = \frac{m_0 + j_0}{N} q^*.$$

This value will be used as cut-off point.

The key steps of the proof start from here

Step 1 Conditioning on $P'_{(m_0)}$, the largest p-value in the group of true null hypotheses, we find

$$\begin{aligned}
&\mathcal{E}(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \\
&\quad = \textstyle\int_0^{p''} \mathcal{E}(Q|P'_{m_0} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{m_0}(p) dp \\
&\quad + \textstyle\int_{p''}^1 \mathcal{E}(Q|P'_{m_0} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{m_0}(p) dp
\end{aligned}$$

with $f_{m_0} = m_0 p^{(m_0-1)}$ being the density function of $P'_{(m_0)}$.

Step 2 Analyzing the integration in two intervals:

In the **first integral**, we are dealing with the situation where $p \leq p''$. If so, all m_0 true null, plus j_0 false hypotheses are rejected. The false discovery rate is hence

$$Q = m_0/(m_0 + j_0).$$

Substituting this value into the first integral,

$$\int_0^{p''} \{\cdot\} dp = \{m_0/(m_0 + j_0)\} \int_0^{p''} m_0 p^{(m_0-1)} dp = \{m_0/(m_0 + j_0)\}(p'')^{m_0}.$$

Recall $p'' = \frac{m_0+j_0}{N} q^*$, we get

$$\{m_0/(m_0+j_0)\}(p'')^{m_0} \leq \{m_0/(m_0+j_0)\}(p'')^{m_0-1} \left\{ \frac{m_0 + j_0}{N} q^* \right\} = \frac{m_0}{N} q^* (p'')^{m_0-1}.$$

Now keep this result and work on the second integral.

In the **the second integral**, by definition of j_0,

$$P'_{(m_0)} = p \geq p'' = \frac{m_0 + j_0}{N} q^*$$

and $p_{j_0} \leq p''$.

Consider separately each

$$p_{j_0} < p_j \leq P'_{(m_0)} = p < p_{j+1},$$

(the value p exceeds many more p-values of the false null hypothesis); along with

$$p_{j_0} \leq p'' < P'_{(m_0)} = p < p_{j_0+1}$$

(the value p barely larger than $p'' = \frac{m_0+j_0}{N} q^*$. **Now we regard j as fixed and satisfies one of the above inequalities**

Because of the way by which j_0 and p'' are defined, no hypothesis can be rejected as a result of the values of

$$p, p_{j+1}, \ldots, p_{m_1}.$$

That is, none of $H_0^{(m_0)}$, $H_f^{(m_0+j+1)}$, $H_f^{(m_0+j+2)}$, ..., $H_f^{(m_0+m_1)}$ are rejected. **Reminder: j is fixed value in this argument**. Hence, the pool of hypotheses that might be rejected is shrunk to

$$H_0^{(i)} : \ i = 1, 2, \ldots, m_0 - 1; \quad H_f^i : i = m_0 + 1, m_0 + 2, \ldots, m_0 + j.$$

There are $m_0 + j - 1 < N$ of hypotheses in this pool.

In this pool, whether or not a hypothesis is *true and false*, get their p-values ordered together to obtain hypotheses $\tilde{H}_0^{(i)}$, $i = 1, 2, \ldots, m + j - 1$. A hypothesis $\tilde{H}_0^{(i)}$ is rejected only if there exists k, $i \leq k \leq m_0 + j - 1$, for which $\tilde{p}_{(k)} \leq \{k/(m+1)\}q^*$. Namely, we look for the largest k such that

$$\frac{\tilde{p}_{(k)}}{p} \leq \frac{k}{m_0 + j - 1} \frac{m_0 + j - 1}{(N)p} q^*. \quad (7)$$

When conditioning on $P'_{(m_0)} = p$, P'_i/p are iid U(0, 1) random variables (before sorting). Also, p_i/p for $i = 1, 2, \ldots, j$ are numbers between 0 and 1 corresponding to false null hypotheses ($H_f^{(m_0+1)}$, ..., $H_f^{(m_0+j)}$).

Using inequality (7) to test the $m_0 + j - 1 = m' < N$ hypotheses is equivalent to **using the same procedure**, with the constant

$$\tilde{q}^* = \frac{(m_0 + j - 1)}{Np} q^*$$

taking the role of q^*.

Applying now the induction hypothesis to this procedure in which $m' < N$, we have

$$\begin{aligned}\mathcal{E}(Q|P'_{m_0} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) &\leq \frac{m_0 - 1}{m'}\tilde{q}^* \\ &= \frac{m_0 - 1}{m_0 + j - 1}\frac{m_0 + j - 1}{Np}q^* \\ &= \frac{m_0 - 1}{Np}q^*.\end{aligned}$$

The above bound depends on p, but not on the segment $p_j < p < p_{j+1}$ for which it was evaluated. (That is, whichever fixed j is under consideration).

Therefore

$$\int_{p''}^1 \mathcal{E}(Q|P'_{(m_0)} = p, P_{m_0+1}, \ldots, P_m) f_{m_0}(p)dp \leq \int_{p''}^1 \frac{m_0 - 1}{Np}q^*\ m_0 p^{(m_0-1)}dp$$

The outcome of the integration is

$$\frac{m_0}{N}q^* \int_{p''}^1 (m_0 - 1)p^{(m_0-2)}dp = \frac{m_0}{N}q^*(1 - \{p''\}^{(m_0-1)}).$$

Adding two upper bounds on integrations completes the proof of the lemma for the case of $m = N$.

Now the induction is completed and the lemma is fully proven.

Chapter 15

Variables/Model selection problem

We have generally regarded a distribution family as a model. We believe that one of them appropriately describes the randomness exhibited in the data set at hand. The only question is which one of the distributions in this family is the one governing the random mechanism behind the system? The specifics of this model provide evidences for or against the validity/plausibility of the scientific postulation.

The model specification may sometimes not so crucial. For instance, the normality assumption in two-sample problem is rarely an issue unless there are extreme values in the observations.

We general feel that there is a big difference between "Poisson distribution" and "negative binomial". An inference based on Poisson assumption may be too "optimistic" and therefore risky if a "negative binomial" model is more appropriate. How do we choose between Poisson and Negative Binomial? There are various discussions in the literature in this respect. I pick the one that I am most familiar with.

15.1 Nested model setup

We do not usually try to choose between two rather distinct distribution families. For instance, whether the data should be modelled as sample from

Normal distribution or Gamma distribution. Instead, we may proposal a model/family that includes both specific families as special cases.

Even more simplistically, we may first propose a broad enough distribution family which covers all potential situations of interest. After which, we investigate whether some aspects of this family can be set to take certain values to induce specific features.

For instance, in two-sample problem, allowing generic unequal variance normal model assumption but investigate whether an equal variance assumption is plausible.

In regression analysis, allowing the expectation of the response variable to be a linear function of covariates and some derived covariates, such as the squared values of the original covariates. After which, we may investigate whether the quadratic term can be omitted without hurting the fit between the model and the data.

The discussion in regression problem explains why we often use **variable selection** and **model selection** interchangeablly.

Let us temporarily forget the regression model, but work on a model from which we have i.i.d. observations, and it is represented as

$$\{f(x;\theta_1,\theta_2) : \Theta_1 \times \Theta_2\}.$$

We may regard both θ_1 and θ_2 as vectors of real numbers, and their corresponding parameter spaces are regular enough. The model itself is also regular. We interpret "regular" liberally here.

Let $\{f(x;\theta_1) : \Theta_1\}$ be a special case of the general model with $\theta_2 = 0$. The model selection problem is: given a set of i.i.d. observations, should we use the full $\{f(x;\theta_1,\theta_2)\}$ or sub model $f(x;\theta_1)$ for data analysis?

(a) If we use $\{f(x;\theta_1,\theta_2)\}$, the model space is broader. Thus, it is more likely "true". Yet if $\theta_2 = 0$ is true and we do not make use of this fact, the full model used in data analysis is unnecessarily complex. The efficiency of the inference can be low on the most pressing question of interest.

(b) If we use submodel $\{f(x;\theta_1)\}$ for statistical inference, the analysis can the mathematically simple, statistically efficient, and scientifically easier to interpret the conclusion. Yet a "strong claim" based a wrong model can do more damage than a weak one.

In conclusion, there is a demand on "variable/model selection".

15.2 One of many proposed procedures

Let $x_1, \ldots, n$ be the i.i.d. observations from $\{f(x; \theta_1, \theta_2)$. The log likelihood function is given by

$$\ell_n(\theta_1, \theta_2) = \sum \log f(x_i; \theta_1, \theta_2).$$

Let $\hat{\theta}_1, \hat{\theta}_2$ be the MLEs under the full model, and $\tilde{\theta}_1$ be the MLE under the reduced model in which $\theta_2 = 0$.

Apparently, we always have

$$\ell_n(\hat{\theta}_1, \hat{\theta}_2) \geq \ell_n(\tilde{\theta}_1, 0)$$

and most likely, the inequality is strict. Because of this, it is not sensible to select the model who can achieve a higher likelihood function. Otherwise, the full model will always be chosen.

Based on this consideration, we probably should choose the simpler model even if the full model has a slightly large likelihood value than that of the sub-model. This consideration leads to the question: where should we draw the line?

15.3 Bayesian information criterion

One of such lines is proposed by Schwartz (1978) from a Bayesian point of view. It is very popular in the classical model/variable selection context.

Suppose we have a collection of models subject to selection. Let them be $\mathcal{S}$. In the previous example, $\mathcal{S}$ consists of two models: the full model and the reduced model with $\theta_2 = 0$. Let $\pi(\theta(s))$ stand for the prior distribution of θ in the context of model s.

Let $f(Y; \theta)$ denote the joint density function of the observation Y, which usually represents a set of n i.i.d. random variables.

Given a model s, its joint density function is given by

$$m(Y|s) = \int f(Y; \theta(s))\pi(\theta(s))d\theta(s).$$

It is seen that this π may depend on s.

The posterior probability of model s is obtained as

$$p(s|Y) = \frac{m(Y|s)p(s)}{\sum_{s\in\mathcal{S}} m(Y|s)p(s)}.$$

The Bayesian procedure is to select the model which maximizes $p(s|Y)$. In addition, because the denominator does not dependent on s, the model of choice is simplified to

$$s^* = \arg\max\{m(Y|s)p(s)\}.$$

Some math. For the full model, let $\hat{\theta}$ be the maximum likelihood estimator. The likelihood function

$$L_n(\theta) = f(Y;\theta(s))$$

is maximized at $\theta = \hat{\theta}$. Here is another detail of importance: the dimension of θ has been confined by specification of s. For mathematical simplicity, assume the Fisher information matrix based on a single observation is an identity matrix of size 2×2. In this neighbourhood and under some regularity conditions, we have

$$L_n(\theta) \approx L_n(\hat{\theta})\exp\{-\frac{n}{2}\|\hat{\theta}-\theta\|^2\}.$$

The identity matrix assumption is used to get the above simpler quadratic form in the exponent. Note that the simple quadratic form is benefitted from the assumed simple Fisher information matrix.

It is seen that when θ is within $n^{-1/2}$ neighbourhood of $\hat{\theta}$, $\log L_n(\theta)$ is within constant distance from $\log L_n(\hat{\theta})$. By "constant distance", we mean it does not increase with n. For θ out of this distance, the distance of $\log L_n(\theta)$ from $\log L_n(\hat{\theta})$ diverges to infinity.

In addition, as long as $\pi(\theta)$ is a smooth function and non-zero at $\hat{\theta}$, it is virtually a constant function of θ in the $n^{-1/2}$-neighborhood. Hence, approximately, we have

$$\begin{aligned} m(Y|s) &= c(s)L_n(\hat{\theta})\int \exp\{-\frac{n}{2}\|\hat{\theta}-\theta\|^2\}d\theta \\ &= c(s)L_n(\hat{\theta})\{\sqrt{2\pi/n}\}^{\dim(s)} \end{aligned}$$

In logarithm, it becomes

$$\log m(Y|s) = \ell_n(\hat{\theta}) - (1/2)\dim(s)\log n + c(s).$$

Ignoring the term that does not depend on n, we arrive at

$$s^* = \arg\max\{m(Y|s)p(s)\} = \arg\max\{2\ell_n(\hat{\theta}) - \dim(s)\log n\}.$$

The Bayesian information criterion is therefore given by

$$\text{BIC}(s) = -2\ell_n(\hat{\theta}) + \dim(s)\log n.$$

The model s that minimizes $\text{BIC}(s)$ is the model of choice.

Abuse of BIC Although the Bayesian information criterion is derived under regularity conditions and other conditions, it has a form that is free from these details. Hence,

$$\text{BIC}(s) = 2\ell_n(\hat{\theta}) - \dim(s)\log n$$

is used in all occasions: regression, time series, mixture model and so on.

Because maximizing the posterior probability has inherited property of being "optimal", the BIC is often referred to as an optimal model selection procedure. Yet we should be aware of the context within which it is optimal.

The mathematical form of BIC has a natural interpretation. A model is not chosen purely based on how large its likelihood is. The gain in likelihood by adopting a fuller model is discounted according to its complexity. BIC measures the complexity in terms of the dimensionality of the model, and uses $(1/2)\log n$ as the scale of compensation/penalty.

While BIC has been advocated by a vast majority of users, the motivations of these users are not the same. Many use BIC simply because they heard that it is "optimal". Others use it because it makes good sense. The percentage of the users who have good understanding of BIC is not very high. There are also users who find other justifications for BIC (in computer science). It could be said that a large percentage of users are not Bayesian.

At the same time, many Bayesians do not fully advocate BIC. After all, it is merely an approximation to the authentic Bayesian procedure. A Bayesian

with principle should specify the prior distribution on model space and the prior densities for θ much more seriously. When the mathematical form is applied to non-iid case, the notion of sample size is also questionable. The "optimality" is criterion quoted by many becomes a joke to a large degree.

Regardless, BIC remains largely the most popular criterion for model/variable selection.

Selection consistency of BIC

When there are only handful number of models in the model space, BIC has probability 1 to choose the "true one" as the sample size $n \to \infty$.

Under the i.i.d. situation, the scale $(1/2)\log n$ can be modified to something slightly larger than $\log\log n$ and smaller than n in asymptotic "order" without invalidating the selection consistency.

15.4 Extended BIC

One hidden assumption in the derivation of BIC is that prior probability $p(s)$ is a constant across the model space, or they differ by a factor not depending on n.

In recent applications, a response variable is often regressed against a huge number of covariates. For instance, we may has one million measurements on one hundred patients together with a response variable. Which of these measurements can be combined to build a solid model to predict the response value?

In this example, there are 10^6 models based on a single covariate $\mathcal{S}_1$, about 5.0×10^{11} models based on a pair of covariates $\mathcal{S}_2$, and so on. If $p(s)$ is constant on each model in $\mathcal{S}_1 \cup \mathcal{S}_2$, it implies that collectively, the prior probability on $\mathcal{S}_2$ is half-million times that of $\mathcal{S}_1$. This is apparently not justifiable.

A more sensible prior is therefore to place an equal prior probability $1/m$ on each class of models $\mathcal{S}_1$, $\mathcal{S}_2$, ..., $\mathcal{S}_m$ for some pre-chosen m. Within each class, individual models divide the total prior probability evenly. If this idea is used, we would end up with

$$\mathrm{EBIC}(s) = 2\ell_n(\hat{\theta}) - \dim(s)\{\log n + 2\log p\}$$

where p is the number of candidate covariates.

When p is at the order of n^k which some $k > 0$, the EBIC is selection consistent for the linear model and the generalized linear model.

The constant 2 in $2 \log p$ can be reduced slightly to improve the finite sample performance while retaining the selection consistency.

15.5 Variable/model selection techniques

The variable selection ideas have to be implemented numerically. Particularly when p is large, the implementations of BIC or EBIC can be impossible.

LASSO, in comparison, provides another selection criterion. It is a procedure with a very efficient numerical solution. It has many other desired properties in addition to variable selection. SCAD is a serious competitor of LASSO. It is more difficult to implement. However, this procedure is shown to have better statistical performances from several angles. We can only afford to give a very brief introduction to these methods.

Suppose we have n observations from a linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

$i = 1, 2, \ldots, n$. Consider the most standard situation where ϵ_i are i.i.d. $\mathrm{N}(0, \sigma^2)$ random errors. The design points, $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^\tau$ are non-random. The least square estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\tau$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\tau \mathbb{X})^{-1} \mathbb{X}^\tau \mathbf{y}$$

where $\mathbb{X}$ is an $n \times p$ matrix formed by stacking up $\mathbf{x}_i$, and $\mathbf{y}$ is the vector of response values y_i. This estimator is obtained as the vector which minimizes the sum of squares:

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^\tau (\mathbf{y} - \mathbb{X}\boldsymbol{\beta}).$$

In traditional examples where the linear model is applied, we usually have $p << n$. That is, we have a lot more observations than the number of explanatory variables. If we suspect that an explanatory variable, say the

first entry of $\mathbf{x}$, does not help to explain the variations expressed in $\mathbf{y}$, then this variable will be examined to see whether removal of which helps.

One may certainly make use of t-test to test the hypothesis that $\beta_1 = 0$. If x_1 is the only variable being checked, this approach is very reasonable. If practically every one of p variables is examined, we are confronting the multiple comparison (more precisely multiple test) issue. To make the matter worse, the fitted value of $\boldsymbol{\beta}$ changes in every of its component in general when just one of $\mathbf{x}$ is removed from the model. It can happen that both $\beta_1 = 0$ and $\beta_2 = 0$ are not rejected based on the fit of the full model. Yet $\beta_2 = 0$ will be rejected after x_1 has been removed from the model. Strictly speaking, the hypothesis test based variable select procedures are illogical from mathematical statistics sense. They are all ad hoc procedures without principles. I should also add here that I have no objections on the use of these procedures such as forward or backward selections. However, one should make it clear that they are just some ad hoc procedures in lectures.

The next choice is BIC. As we already claimed, it is a procedure with a principle. We may not agree with this principle but it has one. In addition, the mathematical form is obtained based on i.i.d. assumptions. Hence, its use in regression analysis needs further scrutiny. Overall speaking, I still feel that this is the most sensible method when p is not very large. When p is very large, EBIC is recommended. Whether BIC or EBIC are recommended, there is a serious computational difficult. When $p = 100$, there are 1.73×10^{13} many linear models with exactly 10 explanatory variables selected. There is no computer in this world at the moment fit them all fast enough to allow us compare their BIC or EBIC values.

So far, our discussion has not touched the issue of LASSO. Why is it introduced here? It can be used as a variable method which does not require as much computation.

The initial motivation of LASSO is not just for variable selection. Once the regression coefficient $\boldsymbol{\beta}$ has been estimated, it is natural to predict the future outcome of y at design point $\mathbf{x}$ by

$$\hat{y} = \mathbf{x}^\tau \hat{\boldsymbol{\beta}}.$$

Because $\hat{\boldsymbol{\beta}}$ is the vector value which makes the $\mathbf{y}$ and $\hat{\mathbf{y}}$ the closest possible in terms of $S(\boldsymbol{\beta})$, by definition it overfits. One way to stop overfitting is to

restrict the size of $\boldsymbol{\beta}$. LASSO is a method which estimates $\boldsymbol{\beta}$ by the minimizer of

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^\tau(\mathbf{y} - \mathbb{X}\boldsymbol{\beta}) + \lambda|\boldsymbol{\beta}|$$

instead, where λ is a non-negative constant and

$$|\boldsymbol{\beta}| = \sum_{j=1}^{p} |\beta_j|.$$

Let us call this fit $\boldsymbol{\beta}_\lambda$. The extra term introduced is referred to as regularization term.

If the least sum of squares fit makes $\hat{\beta}_1 = 1$, the likely LASSO fitted value will be $\hat{\beta}_1 = 0.9$. So the LASSO fit leaves some room in the fit, rather than trying to match $\mathbf{y}$ with $\hat{\mathbf{y}}$ to the extreme. After all, $\mathbf{y}$ contains a random error term ϵ. Indeed, LASSO based prediction of the future observation

$$y_\lambda = \mathbf{x}^\tau \boldsymbol{\beta}_\lambda$$

is more accurate with a "proper choice" of λ.

It turns out that LASSO has another very useful property. Because $|\boldsymbol{\beta}|$ is not a smooth function of $\boldsymbol{\beta}$, the fitted vector value $\boldsymbol{\beta}_\lambda$ often contains 0 entries. When λ exceeds a certain value, we would have $\boldsymbol{\beta}_\lambda = \{\mathbf{0}\}$ (other than β_0 the intercept). When $\lambda = 0$, the LASSO fit of $\boldsymbol{\beta}$ is the same as the least sum of squares fit. When λ decreases from a very large value down to 0, the number of non-zero entries of $\boldsymbol{\beta}_\lambda$ gradually increases though not strictly monotonically.

Suppose at $\lambda = 4.3$, we find $\boldsymbol{\beta}_\lambda$ has 5 non-zero entries. For simplicity, suppose

$$\boldsymbol{\beta}_\lambda = (1.2, 0, 0, 1.3, 1.4, -0.3, -3.4, 0, 0, \ldots)^\tau.$$

We may regard that x_3, x_4, x_5, x_6 are selected to form a linear regression model. All other explanatory variables such as x_1, x_2, x_7 and so on are deselected. Clearly, the LASSO procedure has one variable selection and parameter estimation simultaneously.

Computationally, the inventor of LASSO jointly with others discovered a very efficient numerical algorithm to obtain $\boldsymbol{\beta}_\lambda$ for all values of λ. They proved that the entries of $\boldsymbol{\beta}_\lambda$ is piece-wise linear in λ. The total amount

of computation is at the same order as the least sum of squares fit. It is a miracle that LASSO has many of the properties one can dream of when used as a variable-selection method.

If one regards $\lambda|\boldsymbol{\beta}|$ as a prior, then the LASSO is an optimal variable select criterion with this fixed λ. However, which λ should one use? A technique called cross-validation is often recommended. We will not discuss it here.

Since LASSO can be used to provide a sequence of linear models with practically increasing number of explanatory variables selected, one may keep track on the BIC or EBIC values of these models. When the BIC or EBIC increases at some point, we stop reducing the value of λ and choose the model with the variables selected by LASSO at this λ.

When LASSO is used for variable selection in applications, one must make sure that all explanatory variables are at the same scale. Otherwise, if x_1 is replaced by $1000x_1$, the corresponding β_1 would be shrunk to $\beta_1/1000$. The effect of the regularization term $\lambda|\boldsymbol{\beta}|$ on x_1 will be reduced by a thousand folder for a very bad reason. A simple strategy to avoid this scale problem is to While this remark is very important conceptually, the R-program developed for LASSO is free from this issue. We must be thankful to those who did all for us already.

What is the role for SCAD? The regularization term employed by LASSO, $|\boldsymbol{\beta}|$, has the effect to lower the fitted value of $\hat{\boldsymbol{\beta}}$ indiscrimitively. We are told that this is bad: if $|\hat{\beta}_3|$ is very large, it implies x_3 has a significant effect on y. Reducing its fitted value lowers the statistical efficiency. The regularization term in SCAD is specifically designed to address this issue. In fact, it has been proved that LASSO is not selection consistent but SCAD is. There have been a lot of new developments in this area. I must stop at some point and the point is here.