Statistical Modelling I

Kexing Ying

January 11, 2021

Contents

1 Introduction

In this module, we will consider and analyse the relationship between measurements through the use of statistical models. This is realised in several ways including quantifying distributions, comparing distributions and predicting observations. We shall study these methods through deriving, evaluating and applying estimators, confidence intervals and hypothesis tests based, first on parametric models, and later based on the theory of linear models.

Definition 1.1 (Statistical Model)**.** A statistical model is a specification of the distribution of Y up to an unknown parameter θ.

Definition 1.2 (Parameter Space)**.** Given a statistical model Y up to some parameter θ, the set Θ of all possible parameter values is called the parameter space.

In this module we will assume $\Theta \subseteq \mathbb{R}^p$ for some $p \in \mathbb{N}$ so that we consider *parametric models*. A *semiparametric model* is a statistical model which parameters belong to a more general space, e.g. functions spaces.

As with last years **Probability and Statistics**, we will denote the data $\mathbf{y} = (y_1, \cdots, y_n) \in \mathbb{R}^n$ as a vector and $\mathbf{Y} = (Y_1, \cdots, Y_n)$ a random vector. In this case, the statistical model specifies the joint distribution of $Y_1, \cdots, Y_n$ up to some unknown parameter θ. If $Y, \cdots, Y_n$ are independent and identically distributed (iid.), then we call it a *random sample*.

Furthermore, in some situations, the random vector $\mathbf{Y} = (Y_1, \cdots, Y_n)$ might be dependent on random or nonrandom values $x_1, \cdots, x_n$. The x_i's are an example of covariates. An example of this could be that, in a clinical trial, some patients are given a treatment while others received a placebo. If we would like to model the outcome of the i-th patient by their survival time Y_i, it is clear that as covariate for the i-th patient, we may use the indicator function for whether or not i received the treatment as a covariate.

As another example, say we would like to ask whether or not taller people have a higher income. To answer this question we might create a statistical model in which Y_i is the income, x_i is the height and

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i$$

for $i = 1, \cdots, n$, $\epsilon \sim N(0, \sigma^2)$ iid. and $\theta = (\beta_0, \beta_1, \sigma^2)$, $\Theta = \mathbb{R}^2 \times [0, \infty)$.

Having formulated a model, we can draw inferences from the sample. By estimating the unknown parameters we attempts to "fit the model". Through this, we receive a model that can provide us with point estimates, or better yet, tools that can help us make decisions through some combination of hypothesis tests and confidence intervals.

However, with all statistical models, we have to accept that it will not perfectly reflect reality. But, that is not the point of statistical models anyway. Statistical models are meant to be useful and in general we would like a model to

- agree with the observed data reasonably;
- be relatively simple;
- interpretable, e.g. parameters have a physical interpretation.

With these aims in mind, we might conduct sensitivity analysis in which we discard models that are not adequate for the data through a iterative process.

2 Point Estimation

From the introduction, we seen that during the process of fitting the model, we need to estimate θ in the statistical model, and furthermore, during the inference process, we need to point estimate, interval estimate or hypothesis test to address our question. We recall from first year that this can be achieved through several methods and we shall quickly review them here.

2.1 Review

We recall the following definitions.

Definition 2.1 (Realisation, Statistic, Estimate, Estimator)**.**

- Data $y_1, \cdots, y_n$ is called a realisation of $Y_1, \cdots, Y_n$.
- A function t of observable random variables is called a statistic.
- An estimate of θ is $t(y_1, \cdots, y_n)$.
- An estimator of θ is $T = t(Y_1, \cdots, Y_n)$.

Example 1. Let $Y_1, \cdots, Y_n \sim N(\mu, 1)$ iid. for some unknown $\mu \in \mathbb{R}$. There are many methods for estimating μ.

- the sample mean $\hat{\mu} = \frac{1}{n}\sum y_i$;
- the sample median;
- the k-trimmed mean where we discard the highest and lowest k observed y_i before computing the mean;
- $\cdots$

For the sample mean estimate, the corresponding estimator is $T = \bar{Y} = \frac{1}{n}\sum Y_i$.

As we can see from the example, there are many possible estimations for the same parameter. To justify the use of a specific estimator, one might use a frequentist's perspective and generate many data and tabulate the results of each estimator. Through this process, one can justify a particular estimator through observed data.

As estimators are random variables, we can formalise this idea by considering properties of its sampling distribution (that is the distribution of the estimator), e.g.

$$\mathbb{P}_\theta(T \in \mathcal{A}), \quad E_\theta(T), \quad \mathrm{Var}_\theta(T), \cdots$$

We saw this idea last year in the form of *bias* and *mean square error*. We recall the definitions here.

Definition 2.2 (Bias)**.** Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then the bias of T is

$$\mathrm{bias}_\theta(T) = E_\theta(T) - \theta.$$

If $\mathrm{bias}_\theta(T) = 0$ for all $\theta \in \Theta$, then we say T is unbiased for θ.

If the parameter space is higher dimensional, say $\Theta \subseteq \mathbb{R}^k$, we may be instead be interested in the value of $g(\theta)$ for some $g : \Theta \to \mathbb{R}$. Then, we can naturally extend the definition of bias to this by

$$\text{bias}_\theta(T) = E_\theta(T) - g(\theta).$$

Example 2. Let $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ iid. $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Then, say if we are in μ, we may define $g : \Theta \to \mathbb{R} : (\mu, \sigma^2) \mapsto \mu$.

Definition 2.3 (Mean Square Error)**.** Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then the mean square error of T is

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

In addition to this, we have the standard error of a estimator

Definition 2.4 (Standard Error)**.** Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then the standard error of T is

$$\text{SE}_\theta(T) = \sqrt{\text{Var}_\theta(T)}.$$

From last year, we saw the following proposition.

Proposition 1. Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then

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

If we restrict out estimators to be unbiased, often times, we find that the remaining possible estimators well-behaved and we can often find the best estimators by minimising the MSE. However, a biased estimator might have a small MSE than an unbiased estimator (recall sample variance), and it is not necessarily true that such an estimator even exists.

2.2 Cramér-Rao Lower Bound

As the mean square error provide us with a method of quantifying how good an estimator is, we are motivated by minimising the mean square error for a family of estimators. That is, if $\theta \in \Theta$ is a parameter, then is there an estimator T of θ such that for all estimators of θ, S,

$$\text{MSE}_\theta(T) \leq \text{MSE}_\theta(S).$$

Unfortunately, the answer to this question is in general, no, however, for unbiased estimators, the answer is often yes. Indeed, if T is an unbiased estimator, then,

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

so it suffices to minimise the variance.

Theorem 1 (Cramér-Rao Lower Bound)**.** Suppose $T = T(X)$ is an unbiased estimator for $\theta \in \Theta \subseteq \mathbb{R}$ based on $X = (X_1, \cdots, X_n)$ with joint pdf $f_\theta(x)$. Then under mild regularity conditions (which is elaborated on in the proof below),

$$\text{Var}_\theta(T) \geq \frac{1}{I(\theta)},$$

where

$$I(\theta) = E_\theta \left[\left\{ \frac{\partial}{\partial \theta} \log f_\theta(X) \right\}^2 \right],$$

and we call $I(\theta)$ the *Fisher information* of the sample.

By computing, we find the Fisher information to equal the following.

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

Indeed,

$$\begin{aligned} E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right] &= E_\theta \left[\frac{\partial}{\partial \theta} \frac{f'_\theta(X)}{f_\theta(X)} \right] \\ &= E_\theta \left[-\frac{f'_\theta(X)}{f^2_\theta(X)} f'_\theta(X) + \frac{f''_\theta(X)}{f_\theta(X)} \right] \\ &= E_\theta \left[-\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] + E_\theta \left[\frac{f''_\theta(X)}{f_\theta(X)} \right]. \end{aligned}$$

So, the result follows as,

$$\begin{aligned} E_\theta \left[\frac{f''_\theta(X)}{f_\theta(X)} \right] &= \int_{x \in A} \frac{f''_\theta(x)}{f_\theta(x)} f_\theta(x) \mathrm{d}x \\ &= \int_{x \in A} f''_\theta(x) \mathrm{d}x = \frac{\partial^2}{\partial \theta^2} \int_{x \in A} f_\theta(x) \mathrm{d}x = 0, \end{aligned}$$

where we denoted A as the support of f_θ. This is a useful identity whenever the second derivative is easy to compute.

Corollary 1.1. Suppose $X_1, \cdots, X_n$ is a random sample. Then, is $f^{(1)}_\theta$ is the pdf of single observation, then

$$I_f(\theta) = n I_{f^{(1)}_\theta}(\theta).$$

Proof. Since a random sample is iid. $f_\theta(x) = \prod f^{(1)}_\theta$ and so

$$I_f(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right] = \sum_{i=1}^n -E_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f^{(1)}_\theta(X_i) \right) = n I_{f^{(1)}_\theta}(\theta).$$

□

From this, we can conclude that the Fisher information is proportional to the sample size.

Example 3. Let us find the Fisher information for the random sample $X_1, \cdots, X_n \sim \mathrm{Bern}(\theta)$.

By the above corollary, we have $I_f(\theta) = n I_{f^{(1)}_\theta}(\theta)$. So, since the pmf of a Bernoulli random variable is $f^{(1)}_\theta(x) = \theta^x (1-\theta)^{1-x}$, we have

$$\frac{\partial}{\partial \theta} \log f^{(1)}_\theta(x) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)},$$

hence,

$$I_{f_\theta^{(1)}}(\theta) = E\left[\left(\frac{x-\theta}{\theta(1-\theta)}\right)^2\right] = \frac{1}{\theta^2(1-\theta)^2}\operatorname{Var}(X) = \frac{1}{\theta(1-\theta)}.$$

Thus, the Fisher information of the random sample is just $I_f(\theta) = n/\theta(1-\theta)$.

With the Fisher information, we can apply the Cramér-Rao lower bound theorem allowing us to conclude that an unbiased estimator T for θ has variance $\operatorname{Var}(T) \geq \theta(1-\theta)/n = \operatorname{Var}(\bar{X})$. This allows us to conclude that the sample mean $\bar{X}$ minimises the mean square error among unbiased estimators for θ.

Let us now prove the Cramér-Rao lower bound theorem.

Proof. (Cramér-Rao lower bound theorem). Let us first specify the regularity conditions for the Cramér-Rao lower bound theorem.

- Assume that the set $A := \operatorname{supp} f_\theta = \{x \in \mathbb{R}^n \mid f_\theta(x) > 0\}$ is independent of θ.
- Θ is an open interval in $\mathbb{R}$.
- For all $\theta \in \Theta$ there exists $\frac{\partial f_\theta}{\partial \theta}$.
- Differentiation and integration commutes (for the specific cases where it is used).

As we saw last year, the space of random variables form an inner product space with the inner product

$$\langle X, Y\rangle = E[XY],$$

and so, the Cauchy-Schwarz inequality applies. That is for all random variables X, Y

$$[E(XY)]^2 \leq E(X^2)E(Y^2).$$

So, we have

$$\begin{aligned}\operatorname{Var}_\theta(T) I_f(\theta) &= E_\theta[(T - E_\theta T)^2] E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] \\ &\geq \left(E_\theta\left[(T - E_\theta(T))\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\right)^2.\end{aligned}$$

Thus, it suffices to show that the expectation on the right hand side evaluates to 1.

$$\begin{aligned}E_\theta\left[(T - E_\theta(T))\frac{\partial}{\partial\theta}\log f_\theta(X)\right] &= E_\theta\left[(T - E_\theta(T))\frac{\frac{\partial}{\partial\theta} f_\theta(X)}{f_\theta(X)}\right] \\ &= \int_{x\in A}(T(x) - E_\theta(T))\frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(X)} f_\theta(x)\mathrm{d}x \\ &= \int_{x\in A} T(x)\frac{\partial}{\partial\theta} f_\theta(x)\mathrm{d}x - \int_{x\in A} E_\theta(T)\frac{\partial}{\partial\theta} f_\theta(x)\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\int_{x\in A} T(x) f_\theta(x)\mathrm{d}x - E_\theta(T)\frac{\partial}{\partial\theta}\int_{x\in A} f_\theta(x)\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}E_\theta(T) - 0 = \frac{\partial}{\partial\theta}\theta = 1.\end{aligned}$$

□

2.3 Asymptotic Properties of Estimators

While the Cramér-Rao lower bound theorem provides us with a lower bound for the variance for non-biased estimators, as we have previously seen, it is not always true that there exists an unbiased estimator. So, rather than giving up, we instead study the estimators as the sample size becomes large.

As we have seen, evaluating an estimator $T = T(X_1, \cdots, X_n)$ of θ depends on its *sampling distribution*. From the sampling distribution, one can possibly find properties about the estimator such as is bias, mean square error and so on. However, it is not necessarily true that an estimator has a closed form. Indeed, often times, the estimator is defined as a solution to some equation.

To simplify this, one often consider $T_n = T_n(X_1, \cdots, X_n)$ as a sequence of random variables indexed by $n \in \mathbb{N}$ and consider the stochastic convergence of the variables in question. We recall from last term's probability course, there are three different notions of convergence for random variables,

- convergence in probability;
- convergence almost surely (almost everywhere);
- convergence in distribution.

Let us quickly define them here again.

Definition 2.5 (Convergence in Probability)**.** Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. Then, (X_n) converges to the random variable X in probability if for all $\epsilon > 0$,

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

Definition 2.6 (Convergence Almost Surely)**.** Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. Then, (X_n) converges to the random variable X almost surely if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1,$$

where $\lim_{n\to\infty} X_n = X$ is denoting the event

$$\{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}.$$

With other words, $X_n \to X$ almost surely, if the set of points ω such that $X_n(\omega)$ does not converge to $X(\omega)$ has measure 0.

Definition 2.7. Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. Then, (X_n) converges to the random variable X with cdf F_X in distribution if

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = F_X(x),$$

for all x at which F_X is continuous.

We also recall the following chain of implications,

$$X_n \to_{\text{as}} X \implies X_n \to_{\text{p}} X \implies X_n \to_{\text{d}} X.$$

If $X = c$ is a constant, then

$$X_n \to_{\text{p}} X \iff X_n \to_{\text{d}} X.$$

We apply this notion onto estimators.

Definition 2.8 (Consistency)**.** A sequence of estimators $(T_n)_{n=1}^{\infty}$ for $g(\theta)$ is called (weakly) consistent if for all $\theta \in \Theta$, $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|T_n - g(\theta)| > \epsilon) = 0.$$

While it is possible to prove consistency for certain estimators, it is often a non-trivial task. Instead, we often look at whether a sequence of estimators are asymptotically unbiased and prove the a special class of these estimators are consistent.

Definition 2.9 (Asymptotically Unbiased Estimators)**.** A sequence of estimators $(T_n)_{n=1}^{\infty}$ for $g(\theta)$ is called asymptotically unbiased if for all $\theta \in \Theta$,

$$E_\theta(T_n) \to g(\theta).$$

We see that $E_\theta(T_n)$ is simply a value and this is simply the convergence for real sequences.

Before moving on to prove results about estimators, let us quickly recall the Markov inequality.

Proposition 2. Let X be a random variable with $X(\Omega) \subseteq [0, \infty)$, then for all $a \in \mathbb{R}$,

$$\mathbb{P}(|X| \geq a) \leq \frac{E(|X|)}{a}.$$

Proof. See first year notes. □

Lemma 2.1 (MSE Consistency)**.** Let $(T_n)_{n=1}^{\infty}$ be asymptotically unbiased for $g(\theta)$ for all $\theta \in \Theta$. Then, if $\text{Var}_\theta(T_n) \to 0$ as $n \to \infty$, T_n is consistent for $g(\theta)$.

Proof. Let $\epsilon > 0$, then, by Markov's inequality,

$$\begin{aligned}\mathbb{P}_\theta(|T_n - g(\theta)| \geq \epsilon) = \mathbb{P}_\theta((T_n - g(\theta))^2 \geq \epsilon^2) &\leq \frac{1}{\epsilon^2} E_\theta(T_n - g(\theta))^2 \\ &= \frac{\text{MSE}_\theta(T_n)}{\epsilon^2} = \frac{1}{\epsilon}(\text{Var}_\theta(T_n) + (E_\theta(T_n) - g(\theta))^2).\end{aligned}$$

Since the right hand side tends to 0 as $n \to \infty$, so is the left hand side. □

Thus, to show that a sequence of estimators is consistent, it suffices to show that it is asymptotically unbiased and its variance tends to 0.

However, while consistency is a nice property for a sequence of estimators to have, it is a very minimal requirement. So, in order to derive hypothesis tests and confidence intervals,

we also need the sampling distribution of T_n. As we have seen previously, the sample mean estimators T_n for a normal distribution $N(\theta, 1)$ has distribution $T_n \sim N(\theta, 1/n)$, and so, by centring and scaling, we have

$$\sqrt{n}(T_n - \theta) \sim N(0, 1),$$

for all $n \geq 1$. This means we can work with the CDF of T_n allowing us to easily analyse the behaviours of these estimators. However, this is in general not the case and in most cases, we cannot derive easily the distributions of the estimators. Nonetheless, often, we may approximate their distribution with a normal distribution.

Definition 2.10 (Asymptotically Normal)**.** A sequence of estimators T_n for $\theta \in \mathbb{R}$ is asymptotically normal if, for some $\sigma^2(\theta)$,

$$\sqrt{n}(T_n - \theta) \to N(0, \sigma^2(\theta)),$$

in distribution.

From last term, we recall the central limit theorem (CLT).

Theorem 2 (Central Limit Theorem)**.** Let $Y_1, \cdots, Y_n$ be iid. random variables with $E(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2 < \infty$. Then the sequence $\sqrt{n}(\bar{Y} - \mu)$ converges in distribution to a $N(0, \sigma^2)$ distribution.

Proof. See the *probability for statistics* course. □

The central limit theorem allows us to conclude that a large class of estimators are asymptotically normal. Indeed, sample means and estimators which can be written as a combination of sample means under weak conditions are certainly asymptotically normal. However, we would also consider other estimators.

Lemma 2.2 (Slutsky's lemma)**.** Let X_n, X and Y_n be random variables (or random vectors). If $X_n \to X$ in distribution and $Y_n \to c$ in probability for some constant c, then

- $X_n + Y_n \to X + c$ in distribution;
- $Y_n X_n \to cX$ in distribution;
- $Y_n^{-1} X_n \to c^{-1} X$ in distribution if $c \neq 0$.

Proof. See the *probability for statistics* course. □

Another useful result for determining whether or not a sequence of estimators are asymptotically normal is the δ-method. The δ-method allows us to consider whether or not the transformation of a asymptotically normal estimator remains asymptotically normal.

Theorem 3 (δ-Method)**.** Suppose that T_n is an asymptotically normal estimator of θ with

$$\sqrt{n}(T_n - \theta) \to_{\mathrm{d}} N(0, \sigma^2(\theta)),$$

and $g : \Theta \subseteq \mathbb{R} \to \mathbb{R}$ is a differentiable function with $g'(\theta) \neq 0$. Then

$$\sqrt{n}(g(T_n) - g(\theta)) \to_{\mathrm{d}} N(0, g'(\theta)^2 \sigma^2(\theta)).$$

Proof. Since g is differentiable,

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + o((T_n - \theta)^2),$$

where R is the remainder. So

$$\sqrt{n}(g(T_n) - g(\theta)) = g'\sqrt{n}(\theta)(T_n - \theta) + o((T_n - \theta)^2).$$

Thus, assuming the remainder is negligible, we have

$$\sqrt{n}(g(T_n) - g(\theta)) \to_{\mathrm{d}} N(0, g'(\theta)^2\sigma^2(\theta)).$$

□

Lastly, a useful result we will often use (perhaps implicitly) is the continuous mapping theorem. Alike sequential continuity for metric spaces, the continuous mapping theorem will allow us to preserve stochastic convergence under continuous mappings.

Theorem 4 (Continuous Mapping Theorem)**.** Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. Then if $X_n \to X$, then $g(X_n) \to g(X)$ for all three notions of convergence, i.e. convergence in distribution, in probability and almost surely.

2.4 Maximum Likelihood Estimation

We recall from first year the maximum likelihood estimator, that is the estimator that maximises the probability of observing the given realisations.

Definition 2.11 (Likelihood Function)**.** Given the realisation $\mathbf{x}$ of the random object $\mathbf{X}$, the likelihood function for θ is

$$L(\theta) = L(\theta \mid \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x} \mid \theta).$$

Definition 2.12 (Maximum Likelihood Estimator)**.** The maximum likelihood estimator of $\theta \in \Theta^n$ is an estimator $\hat{\theta}$ such that

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

where L is the likelihood function.

The maximum likelihood estimator is often well defined. However, it is possible to construct situations in which the MLE does not exist or is not unique. We also recall that given a strictly increasing function f, the maximum likelihood estimator can also be obtained by maximising $f \circ L$. This is most commonly seen in the log-likelihood function where we maximise $\log L$.

Maximum likelihood estimators has some nice properties. In short, maximum likelihood estimators are functionally invariant, consistent and asymptotically normal.

Proposition 3. If g is a bijective function and if $\hat{\theta}$ is a MLE of θ, then $\hat{\phi} = g(\hat{\theta})$ is a MLE of $\phi = g(\theta)$.

Proof. Let us denote $\tilde{L}$ for the likelihood function of ϕ, then $\tilde{L} = L \circ g^{-1}$ and so,

$$\tilde{L}(\hat{\phi}) = L(g^{-1}(\hat{\phi})) = L(g^{-1}(g(\hat{\theta}))) = L(\hat{\theta}) \geq L(g^{-1}(\phi)) = \tilde{L}(\phi).$$

□

Suppose we now relax the bijective condition on g. If g is not surjective, then there exists $\phi \in \psi$ such that $\phi \notin g(\Theta)$, and so, for these values no model is defined. If this is the case it does not make sense to speak of the likelihood of these parameters and so, we define their likelihood to be 0. With that, we see that the original proposition remains true.

On the other hand, if g is not injective, then ϕ does not uniquely identify θ and so, there might exists multiply θ such that $g(\theta) = \phi$. However, by defining the induced likelihood for all θ,

$$\tilde{L} : \mathbb{R} \to \mathbb{R} : \phi \mapsto \sup\{L(\theta) \mid g(\theta) = \phi\},$$

we see that the invariance over functions is retained.

Proposition 4. Let $X_1, \cdots$ be iid. observations with pdf $f_\theta(x)$ where $\theta \in \Theta$ and Θ is an open interval. Furthermore, let $\theta_0 \in \Theta$ be some parameter. Then under regularity conditions,

- there exists a consistent sequence $(\hat{\theta}_n)_{n=1}^{\infty}$ of maximum likelihood estimators;
- if $(\hat{\theta}_n)_{n=1}^{\infty}$ is a consistent sequence of MLEs, then

  $$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_{\mathrm{d}} N(0, I_f(\theta_0)^{-1}),$$

  where $I_f(\theta)$ is the Fisher information of a sample of size 1.

We note that this proposition requires the Fisher information of a distribution which is often not known in practical situations. So, to use this result we need to estimate $I_f(\theta_0)$. In general, this can be approximated by

- $I_f(\hat{\theta})$;
- $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log(f(x_i \mid \theta))\mid_{\theta=\hat{\theta}}\right)^2$;
- $-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log(f(x_i \mid \theta))\mid_{\theta=\hat{\theta}}$.

Proof. (Sketch of the existence of consistent MLEs). Let us denote $L(\theta) := \prod_{i=1}^{n} f_\theta(X_i)$ and $S_n(\theta) = \frac{1}{n}\log L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f_\theta(X_i)$. Since log is strictly increasing, we see that $\hat{\theta}$ maximises $L(\theta)$ if and only if it maximises $S(\theta)$. Then, by the weak law of large numbers, given iid. $Z_1, \cdots, Z_n$ where $Z_i = \log f_\theta(X_i)$, $S_n(\theta) \to E_{\theta_0}(Z_1) = E_{\theta_0}(\log f_{\theta_0}(X_i))$ in probability. So, S_n is a consistent estimator for $E_{\theta_0}(\log f_{\theta_0}(X_i))$.

Let us now define $R(\theta) := E_{\theta_0}(\log f_\theta(X_1))$ and I claim that θ_0 maximises R. Indeed, by considering $z - 1 \geq \log z$ for all $z \in \mathbb{R}^+$, we have for all θ,

$$R(\theta) - R(\theta_0) = E_{\theta_0}\left[\log\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right] \leq E_{\theta_0}\left[\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} - 1\right] = \int\left[\frac{f_\theta(x)}{f_{\theta_0}(x)} - 1\right] f_{\theta_0}(x)\mathrm{d}x.$$

But this is simply,

$$\int f_\theta(x) - f_{\theta_0}(x)\mathrm{d}x = 1 - 1 = 0,$$

and hence, $R(\theta) \leq R(\theta_0)$ for all $\theta \in \Theta$.

With this in mind, we have $S_n(\theta) \to R(\theta)$ pointwise and $S_n(\hat{\theta}) \to R(\theta_0)$ in probability. Then, by using analysis techniques, we may show $\hat{\theta} \to \theta_0$ in probability, completing the proof of the first claim. □

This particular sketch of the proof (which classically Wald used) requires that the map $\theta \to R(\theta)$ to be continuous for some compact set $K \subseteq \Theta$ in which for all $\epsilon > 0$, $\mathbb{P}(|\hat{\theta} - \theta_0| > \epsilon, \hat{\theta} \in K) \to 0$. There is a modern approach in which one shows that

$$\sup_{\theta \in \Theta} |S_n(\theta) - R(\theta)| \to 0$$

in probability. This approach relaxes the condition some what and will be examined in the third year course *Statistical Theory*.

3 Confidence Regions & Hypothesis Testing

So far, we have being considering point estimators for single values. This does not reflect any uncertainty. Indeed, as this estimator is simply resulted from a random sample, it does not tell us how variable this estimate would be if we drew another sample. To account for this, we may, instead of estimating a single value, we provide an interval of values that contains the true parameter with a certain probability.

As an example, let us recall an example from first years statistics. Given a random sample $Y_1, \cdots, Y_n \sim N(\mu, \sigma_0^2)$ with σ_0^2 known, we would like to find the confidence intervale I such that $\mathbb{P}(\mu \in I) = 1 - \alpha$ for some $\alpha > 0$, e.g. $\alpha = 0.05$. By using the sample mean, we have $\bar{Y} = \frac{1}{n}\sum Y_i \sim N(\mu, \sigma_0^2/n)$. So, by standardising, we have

$$\frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1),$$

and so, $c_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ where $c_{\alpha/2}$ is the value such that,

$$1 - \alpha = \mathbb{P}\left(-c_{\alpha/2} < \frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1) < c_{\alpha/2}\right).$$

Hence, the $1 - \alpha$ confidence interval of μ is a realisation of the **random** interval

$$I(\mathbf{Y}) = (\bar{Y} - c_{\alpha/2}\sigma_0/\sqrt{n}, \bar{Y} + c_{\alpha/2}\sigma_0/\sqrt{n}).$$

We note that, this does not mean that given a realisation of the random interval I, $\mathbb{P}(\mu \in I) = 1 - \alpha$. Indeed, if I is realised, either $\mathbb{P}(\mu \in I) = 0$ or $\mathbb{P}(\mu \in I) = 1$.

Definition 3.1 ($1 - \alpha$ Confidence Interval)**.** A $1 - \alpha$ confidence interval for $\theta \in \Theta$ is a random interval I_θ that contains θ with probability $\geq 1 - \alpha$, that is,

$$\mathbb{P}_\theta(\theta \in I) \geq 1 - \alpha.$$

A confidence interval can be any types of interval including unbounded ones. Indeed, if the confidence interval is unbounded, we say the confidence interval is a one-sided confidence interval. An application of such an confidence interval could be that we would like to measure the pollutant in drinking water with a maximum percentage. Then, given a random sample of measurements $Y_1, \cdots, Y_n$, we would like to find a confidence interval for such that

$$\mathbb{P}(\theta \leq h(Y)) = 1 - \alpha.$$

So, the confidence interval in this case would be in the form $(-\infty, h(y)]$.

3.1 Construction of Confidence Intervals

Definition 3.2 (Pivotal Quantity)**.** A pivotal quantity for θ is a function $t(Y, \theta)$ of the data θ (and **not** any over parameters).

With the pivotal quantity for θ, $t(Y, \theta)$, we can find constants a_1, a_2 such that

$$\mathbb{P}(a_1 \leq t(Y, \theta) \leq a_2) \geq 1 - \alpha$$

since we know the distribution of t. In many cases, we may rearrange the terms to give

$$\mathbb{P}(h_1(Y) \leq \theta \leq h_2(Y)) \geq 1 - \alpha$$

where $[h_1(Y), h_2(Y)]$ is a random interval. This is a $1 - \alpha$ confidence interval for θ.

As an example of a pivotal quantity, suppose we would like to construct an confidence interval for the random sample $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Then, we may define the pivotal quantity $(Y, \mu) \mapsto \frac{\bar{Y}-\mu}{S/\sqrt{n}}$ where S is the sample standard deviation. This pivotal quantity follows the Student-t distribution with $n - 1$ degrees of freedom allowing us to construct a confidence interval according the the method above.

On the other hand, if we would like to construct an confidence interval for σ^2, we can use the pivotal quantity $(Y, \sigma^2) \mapsto \frac{1}{\sigma^2}\sum(Y_i - \bar{Y})^2$ which has χ^2-distribution with $n - 1$ degrees of freedom.

However, we see that these constructions are rather specialised to normal distributions and without justification, cannot be applied to other distributions. Nonetheless, as we have discussed asymptotic behaviours of estimators, in which many estimators are asymptotically normal, we can use this fact to extend our theory of confidence intervals.

Definition 3.3 (Asymptotic Confidence Interval)**.** A sequence of random intervals I_n is called an asymptotic $1 - \alpha$ confidence interval for θ if

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in I_n) \geq 1 - \alpha.$$

Suppose $\hat{\sigma}_n$ is consistent for $\sigma(\theta)$ and thus, $\hat{\sigma}_n \to \sigma(\theta)$ in probability for all θ. By Slutsky's lemma, we have

$$\sqrt{n}\frac{T_n - \theta}{\hat{\sigma}_n} \to N(0, 1)$$

in distribution. Using the left hand side as the pivotal quantity leads us to the approximate *confidence limits*

$$T \pm c_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$$

where $\Phi(c_{\alpha/2}) = 1 - \alpha/2$. In general, the easiest choice for $\hat{\sigma}_n$ is simply $\mathrm{SE}(T_n)$.

Lastly, we might be interested in constructing a confidence region for more than one parameters.

Definition 3.4 (Simultaneous Confidence Interval)**.** Suppose $\theta = (\theta_1, \cdots, \theta_k)^T \in \Theta \subseteq \mathbb{R}^k$ and we have $(L_i(Y), U_i(Y))$ such that for all θ,

$$\mathbb{P}(L_i(Y) < \theta_i < U_i(Y) \mid i = 1, \cdots, k) \geq 1 - \alpha.$$

Then, $(L_i(y), U_i(y))$ is a $1 - \alpha$ simultaneous confidence interval for $\theta_1, \cdots, \theta_k$.

Theorem 5 (Bonferroni Correction for Simultaneous Confidence Intervals)**.** Suppose $[L_i, U_i]$ is a $1 - \alpha/k$ confidence interval for θ_i. Then, $\prod[L_i, U_i]$ is a $1 - \alpha$ simultaneous confidence interval for $\theta = (\theta_1, \cdots, \theta_k)^T$.

Proof.

$$\mathbb{P}(\theta_i \in [L_i, U_i] \mid i = 1, \cdots, k) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{k} \{\theta_i \notin [L_i, U_i]\}\right) \geq 1 - \sum_{i=1}^{k} \mathbb{P}(\theta_i \notin [L_i, U_i]) \geq 1 - \alpha.$$

□

We note that the Bonferroni corrections are conservative and it is very possible that the resulting simultaneous confidence interval has a higher coverage probability that that is suggested by the Bonferroni correction. We see this in the example where we attempt to find the simultaneous confidence interval of two *independent* random samples $X_1, \cdots, X_n \sim N(\mu, 1)$ and $Y_1, \cdots, Y_n \sim N(\theta, 1)$. Then by the usual method, we find I, J the $1 - \alpha$-confidence intervals for μ and θ respectively. By the Bonferroni correction, $I \times J$ is a $1 - 2\alpha$-confidence region for (μ, θ) while in actuality,

$$\mathbb{P}_{\mu,\theta}((\mu, \theta) \in I \times J) = \mathbb{P}(\mu \in I)\mathbb{P}(\theta \in J) = (1 - \alpha)^2.$$

Choosing $\alpha = 0.1$, we see that Bonferroni guarantees the coverage probability to be above 0.8 while the actual probability is $0.9^2 = 0.81$.

3.2 Hypothesis Testing

Definition 3.5 (Null and Alternative Hypothesis)**.** Given a model f_θ where $\theta \in \Theta \subseteq \mathbb{R}^d$, the null-hypothesis H_0 and the alternative hypothesis H_1 are propositions that $\theta \in \Theta_0$ and $\theta \in \Theta_1$ respectively for some Θ_0, Θ_1 a partition of Θ such that $\Theta_0 \cap \Theta_1 = \varnothing$ and $\Theta_0 \cup \Theta_1 = \Theta$.

The goal of an hypothesis test is to determine whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, or with other words, whether to accept H_0 or reject H_0 and hence accept H_1. This is normally achieved through the observation of a particular subset of the sample space and we call this sample space for which H_0 is rejected the rejection region (or critical region).

In some literature, we might find some authors reframe from using the word *accept* (such as we were told in year one). In practice however, since we are acting based on the result of these tests, it makes some practical meaning to say we accept the null-hypothesis H_0 or we accept the alternative hypothesis H_1.

As the accuracy of the hypothesis tests is arbitrary, it is possible to make errors. The below table demonstrates the two types of errors.

	$\theta \in \Theta_0(H_0)$	$\theta \in \Theta_1(H_1)$
$\neg$ reject H_0	✓	Type II error
reject H_0	Type I error	✓

Definition 3.6 (Level of a Test)**.** A hypothesis test is of level α for $0 < \alpha < 1$ if

$$\mathbb{P}_\theta(\text{reject } H_0) \leq \alpha$$

for all $\theta \in \Theta$.

Usually we choose $\alpha << 1$ with common values being 0.05 and 0.01. However, it is not clear whether or not these values are optimal for general experiments and often times, α is chosen to be much smaller, e.g. $\alpha \sim 10^{-6}$.

Definition 3.7 (Power)**.** Let Θ be a parameter space and $\Theta_0 \subseteq \Theta$ and $\Theta_1 = \Theta \setminus \Theta_0$ so $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ are null and alternative hypothesis'. Suppose we can constructed some test for this hypothesis, then, the power function is the mapping

$$\beta : \Theta \to [0, 1] : \theta \mapsto P_\theta(\text{reject } H_0).$$

Conceptually, if $\theta \in \Theta_0$, we would like $\beta(\theta)$ to be small while if $\theta \in \Theta_1$, we would like $\beta(\theta)$ to be large.

Definition 3.8 (p-Value)**.** The p-value of a particular hypothesis test is

$$p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\text{observing something "at least as extreme" as the observation}).$$

That is, if a test is based on the statistic T with rejection for large values of T, then

$$p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T \geq t)$$

where t is the observed value.

In any case, we reject H_0 if and only if $p \leq \alpha$ and this results a α-level test.

The hypothesis tests are related to confidence intervals in that we can construct a test from any confidence regions.

Let Y be the random observation of the experiment and suppose $A(Y)$ is the $1-\alpha$ confidence region for the parameter $\theta \in \Theta$, i.e.

$$\mathbb{P}_\theta(\theta \in A(Y)) \geq 1 - \alpha,$$

for all $\theta \in \Theta$. Then, by defining $H_0 : \theta \in \Theta_0$ and $H_0 = \theta \notin \Theta_0$ where Θ_0 is some subset of Θ with level α such that we reject H_0 if $\Theta_0 \cap A(y) = \varnothing$. In this case, we see that

$$\mathbb{P}_\theta(\text{Type I error}) = \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(\Theta_0 \cap A(Y) = \varnothing) \leq \mathbb{P}_\theta(\theta \notin A(Y)) \leq \alpha.$$

The reverse is also possible – constructing a confidence region from a hypothesis test. Suppose that for all $\theta_0 \in \Theta$, we have a level α test ϕ_{θ_0} for $H_0^{\theta_0} : \theta = \theta_0$ and $H_1^{\theta_0} : \theta \neq \theta_0$ such that

$$\mathbb{P}_{\theta_0}(\phi_{\theta_0} \text{ reject } H_0) \leq \alpha.$$

Then, by defining

$$A := \{\theta_0 \in \Theta \mid \phi_{\theta_0} \text{ does not reject } H_0^{\theta_0}\},$$

we find A to be a $1 - \alpha$ confidence region for θ. Indeed, for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in A) = \mathbb{P}_\theta(\phi_\theta \text{ does reject }) = 1 - P_\theta(\phi_\theta \text{ rejects}) \geq 1 - \alpha.$$

Through this method, we may construct a test for multiple parameter test through the use of simultaneous confidence regions. Indeed, if $I \times J$ is a $1 - 2\alpha$ confidence region for (μ, θ), a level 2α test of $H_0 : (\mu, \theta) = (\mu_0, \theta_0)$ against $H_1 : (\mu, \theta) \neq (\mu_0, \theta_0)$ is given by

$$R = \{(\bar{X}, \bar{Y}) \mid (\mu_0, \theta_0) \notin I \times J\},$$

where $E(X_i) = \mu$ and $E(Y_i) = \theta$.