STAT 542: Multivariate Analysis **Spring 2021**

Lecture 1: Likelihood method and exponential family

Instructor: Yen-Chi Chen

1.1 The parametric model and likelihood approach

Consider the problem where we observe IID random variables $X_1, \cdots, X_n$ from an unknown CDF F. The parametric model assumes that the underlying CDF belongs to a parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^p$ is called the parameter space. Namely, there exists an element $\theta^* \in \Theta$ such that the true CDF $F = F_{\theta^*}$. When we assume a parametric model, the parameter of interest is often the underlying parameter θ indexing elements in $\mathcal{F}$. In many scenarios, the CDF leads to a PDF or a PMF so we often express the parametric model in terms of the PDF or PMF.

In a parametric model, we often estimate the parameter θ using the **maximum likelihood estimator (MLE)**. The idea is very simple. Suppose we observe only one observation X from a PDF/PMF $p(x)$. The parametric model assumes that such a PDF/PMF can be written as $p(x) = p(x;\theta)$, where θ is the parameter of the model (θ is often the parameter of interest) inside a parameter space Θ ($\theta \in \Theta$). The idea of MLE is to ask the following question: given the observation X, which θ is the *most likely* parameter that generates X? To answer this question, we can vary θ and examine the value of $p(X;\theta)$.

Because we are treating X as fixed and θ being something that we want to optimize, we can view the problem as finding the best θ such that the **likelihood function** $L(\theta|X) = p(X;\theta)$ is maximized. The MLE uses the θ that maximizes the likelihood value. Namely,

$$\widehat{\theta}_{MLE} = \mathsf{argmax}_\theta L(\theta|X).$$

When we have multiple observations $X_1, \cdots, X_n$, the likelihood function can be defined in a similar way – we use the joint PDF/PMF to define the likelihood function. Let $p(x_1, \cdots, x_n;\theta)$ be the joinr PDF/PMF. Then the likelihood function is

$$L_n(\theta) = L(\theta|X_1, \cdots, X_n) = p(X_1, \cdots, X_n;\theta).$$

Note that when we assume IID observations,

$$L_n(\theta) = \prod_{i=1}^n L(\theta|X_i) = \prod_{i=1}^n p(X_i;\theta).$$

In many cases, instead of using the likelihood function, we often work with the **log-likelihood function**

$$\ell_n(\theta) = \log L_n(\theta).$$

Because taking the logarithmic does not change the maximizer of a function, the maximizer of the log-likelihood function is the same as the maximizer of the likelihood function. There are both computational and mathematical advantages of using a log-likelihood function over likelihood function. To see this, we consider the case of IID sample. Computationally, the likelihood function often has a very small value due to the product form of PDF/PMFs. So it is very likely that the number if too small, making the computation very challenging. Mathematically, when we take log of the likelihood function, the product of PDF/PMFs becomes an additive form

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^n \log p(X_i;\theta).$$

Under IID assumption, each $\log p(X_i;\theta)$ is an IID random variable so the central limit theorem and the law of large number can be applied to the average, making it possible to analyze it asymptotic behavior.

Since under the IID assumptions, we have many advantages, we will assume IID from now on. Because MLE finds the maximum of $\ell_n(\theta)$, a common trick to find MLE is to study the gradient of the log-likelihood function, which is also known as the **score function**:

$$S_n(\theta) = \nabla \ell_n(\theta) = \sum_{i=1}^n s(\theta|X_i),$$

where $s(\theta|X_i) = \nabla \ell(\theta|X_i) = \nabla_\theta \log p(X_i;\theta)$, where ∇ is the gradient operator and ∇_θ refers to the gradient operator applied to variable θ. Under suitable conditions, the MLE satisfies the *score equation*:

$$S_n(\widehat{\theta}_{MLE}) = 0.$$

Note that if there are more than one parameter, say $\theta \in \mathbb{R}^p$, the score equation will be a system of p equations.

Because the MLE is at the maximal point of the likelihood function, the curvature of the likelihood function around the maximal will determine its stability. To measure the curvature, we use the **Fisher's information matrix**:

$$I_n(\theta) = -\mathbb{E}\left[\nabla\nabla \ell_n(\theta)\right] = n \cdot I_1(\theta) = n \cdot -\mathbb{E}\left[\nabla_\theta \nabla_\theta p(X_1;\theta)\right].$$

Example 1: Binomial Distribution. Assume that we obtain a single observation $Y \sim \mathsf{Bin}(n,p)$, and we assume that n is known. The goal is to estimate p. The log-likelihood function is

$$\ell(p) = Y \log p + (n-Y)\log(1-p) + C_n(Y),$$

where $C_n(Y) = \log \binom{n}{Y}$ is independent of p. The score function is

$$S(p) = \frac{Y}{p} - \frac{n-Y}{1-p}$$

so solving the score equation gives us $\widehat{p}_{MLE} = \frac{Y}{n}$. Moreover, the Fisher's information is

$$I(p) = \mathbb{E}\left\{\frac{\partial}{\partial p} S(p)\right\} = -\frac{\mathbb{E}(Y)}{p^2} - \frac{n-\mathbb{E}(Y)}{(1-p)^2} = \frac{n}{p(1-p)}.$$

Example 2: Multinomial Distribution. Let $X_1, \cdots, X_n$ be IID from a multinomial distribution such that $P(X_1 = j) = p_j$ for $j = 1, \cdots, s$ and $\sum_{j=1}^s p_j = 1$. Note that the parameter space is $\Theta = \{(p_1, \cdots, p_s) : 0 \leq p_j, \sum_{j=1}^s p_j = 1\}$. By setting $N_j = \sum_{i=1}^n I(X_i = j)$ for each $j = 1, \cdots, s$, we obtain the random vector $(N_1, \cdots, N_s) \sim \mathsf{Multinomial}(n, p)$, where $p = (p_1, \cdots, p_s)$. The parameters of interest are $p_1, \cdots, p_s$. In this case, the likelihood function is

$$L_n(p_1, \cdots, p_s) = \frac{n!}{N_1! \cdots N_s!} p_1^{N_1} \cdots p_s^{N_s}$$

and the log-likelihood function is

$$\ell_n(p_1, \cdots, p_s) = \sum_{j=1}^s N_j \log p_j + C_n,$$

where C_n is independent of p. Note that naively computing the score function and set it to be 0 will not grant us a solution (think about why) because we do not use the constraint of the parameter space – the parameters are summed to 1. To use this constraint in our analysis, we consider adding the Lagrange multipliers and optimize it:

$$F(p,\lambda) = \sum_{j=1}^{s} N_j \log p_j + \lambda\left(1 - \sum_{j=1}^{s} p_j\right).$$

Differentiating this function with respect to $p_1, \cdots, p_s$, and λ and set it to be 0 gives

$$\frac{\partial F}{\partial p_j} = \frac{N_j}{p_j} - \lambda 0 \Rightarrow N_j = \lambda \widehat{p}_{MLE,j}$$

and $1 - \sum_{j=1}^{s} p_j = 0$. Thus, $n = \sum_{j=1}^{s} N_j = \lambda \sum_{j=1}^{p} = \lambda$ so $\widehat{p}_{MLE,j} = \frac{N_j}{n}$.

1.1.1 A rough derivation of the asymptotic normality of the MLE

The MLE has the following asymptotic property:

$$\sqrt{n}\left(\widehat{\theta}_{MLE} - \theta^*\right) \overset{D}{\to} N(0, \Sigma(\theta^*))$$

for some matrix-valued function $\Sigma(\theta)$. Namely, the MLE is asymptotically normally distributed around some parameter θ^*. The above result does NOT require the parametric model to be correct. When the parametric model is correct, i.e., there exists a parameter $\theta_0 \in \Theta$ such that the data is generated from $p(x;\theta_0)$, $\theta^* = \theta_0$ and $\Sigma(\theta^*) = I_1(\theta_0)$.

Here is the usual way of deriving the asymptotic normality of the MLE. We assume that the score equation exists and the MLE solves the score equation. Recall that $S_n(\theta) = \sum_{i=1}^{n} s(\theta|X_i)$, where $s(\theta|X_i) = \nabla_\theta \log p(X_i;\theta)$. If we divide the score function by n, the law of large numbers implies

$$\frac{1}{n} S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} s(\theta|X_i) \overset{P}{\to} \mathbb{E}(s(\theta|X_1)) = S_0(\theta).$$

Thus, we define the *population MLE* θ^* to be the parameter that satisfies $S_0(\theta^*) = 0$. You can easily show that if the parametric model is correct, the population MLE is often the true parameter.

Thus, we now have two equations:

$$\frac{1}{n} S_n(\widehat{\theta}) = 0, \quad S_0(\theta^*) = 0.$$

We then consider the following quantity and perform a Tyler expansion:

$$\begin{aligned} \frac{1}{n} S_n(\theta^*) - S_0(\theta^*) &= \frac{1}{n} S_n(\theta^*) - \frac{1}{n} S_n(\widehat{\theta}) \\ &= \frac{1}{n}\left(\theta^* - \widehat{\theta})^T \nabla S_n(\theta^*) + O_P\left(\|\theta^* - \widehat{\theta}\|^2\right)\right). \end{aligned} \tag{1.1}$$

If we keep the leading term, the above equality implies

$$\begin{aligned} \widehat{\theta} - \theta^* &= \left(\frac{1}{n}\nabla S_n(\theta^*)\right)^{-1}\left(\frac{1}{n} S_n(\theta^*) - S_0(\theta^*)\right) \\ &= \left(\frac{1}{n}\nabla S_n(\theta^*)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} [s(\theta^*|X_i) - \mathbb{E}(s(\theta^*|X_i)]\right) \end{aligned}$$

By law of large numbers,

$$\frac{1}{n}\nabla S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\nabla_\theta \log p(X_i;\theta) \xrightarrow{P} \mathbb{E}\left(\nabla_\theta\nabla_\theta \log p(X_1;\theta)\right) = -I_1(\theta). \tag{1.2}$$

Moreover, central limit theorem implies that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}[s(\theta^*|X_i) - \mathbb{E}(s(\theta^*|X_i)]\right) \xrightarrow{d} N(0, \mathbb{E}(s(\theta^*|X_1)s(\theta^*|X_1)^T)).$$

Thus, the Slutsky's theorem further implies

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, I_1^{-1}(\theta^*)\mathbb{E}(s(\theta^*|X_1)s(\theta^*|X_1)^T)I_1^{-1}(\theta^*)). \tag{1.3}$$

The asymptotic normality in equation (1.3) does not require the parametric model to be correct! In fact, even if the parametric model is incorrect, the MLE still converges to θ^*. The corresponding model $p(x;\theta^*)$ can be viewed as a 'projection' of the true PDF onto the specified parametric family. We will discuss this in greater detail in Section 1.6.

When the model is correct, we can further deduce an elegant form of the covariance function. Note that when the model is correct, $\theta_0 = \theta^*$ and our observation is generated from $p(x;\theta^*)$. The key is to argue that $\mathbb{E}(s(\theta^*|X_1)s(\theta^*|X_1)^T) = I_1(\theta^*)$. Using the fact

$$\begin{aligned}
0 = \nabla_\theta\nabla_\theta 1 &= \nabla_\theta\nabla_\theta \int p(x;\theta)dx \\
&= \int \nabla_\theta\nabla_\theta p(x;\theta)dx \\
&= \int \nabla_\theta[p(x;\theta)\nabla_\theta \log p(x;\theta)dx] \\
&= \int p(x;\theta)\underbrace{[\nabla_\theta \log p(x;\theta)]}_{s(\theta|x)}[\nabla_\theta \log p(x;\theta)]^T + p(x;\theta)\nabla_\theta\nabla_\theta \log p(x;\theta)dx,
\end{aligned}$$

we conclude that

$$0 = \mathbb{E}(s(\theta^*|X_1)s(\theta^*|X_1)^T) - I_1(\theta^*).$$

As a result,

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, I_1^{-1}(\theta^*)), \tag{1.4}$$

which is the well-known asymptotic normality of the MLE under correct model.

Based on the above derivation, you can then deduce what are the assumptions we need to achieve this result. One small note: when the parameter space is univariate, equation (1.1) can be replaced by the mean value theorem so we only need the existence of the first order derivative. However, there is no multivariate mean value theorem, so we have to place additional assumptions to ensure that the second-order derivative is bounded.

1.2 Exponential family

The exponential family is a collection of parametric models that has very elegant properties. Specifically, a parametric model belongs to the exponential family if the underlying PDF/PMF $p(x;\theta)$ can be written as

$$p(x;\theta) = \frac{1}{C(\theta)}h(x)\exp(\theta^T t(x)), \tag{1.5}$$

for some functions $C(\theta), h(x) \in \mathbb{R}$ and $t(x) \in \mathbb{R}^p$. $t(x)$ is called the canonical statistic and the parameter $\theta \in \Theta$ is called the canonical parameter and Θ is the parameter space. Note that

$$C(\theta) = \int h(x) \exp(\theta^T t(x)) dx$$

and can be viewed as a normalizing constant.

Here are some technical terms related to an exponential family. The PDF in equation (1.5) may not be unique–we can have other representation of the same PDF. The *order* of an exponential family is the minimal dimension of $t(x)$ such that we can express the family using equation (1.5). If a representation in equation (1.5) reaches the order, we call it the *minimal* representation

Example: non-minimal representation. An example of non-minimal representation is the multinomial distribution (n, θ). Suppose we have k categories with proportion $\pi_1, \cdots, \pi_k$. We can write the PMF as

$$p(x;\theta) = \frac{n!}{x_1! \cdots x_k!} \prod_{j=1}^k \pi_j^{x_j} = \frac{n!}{x_1! \cdots x_k!} \exp\left(\sum_{j=1}^k x_j \log \pi_j\right)$$

so $\theta_j = \log \pi_j$. However, there is a hidden constraint that $1 = \sum_{j=1}^k \pi_j = \sum_{j=1}^k e^{\theta_j}$ so there is only $k-1$ parameters. Thus, the above representation is not minimal.

The family is called *full* if the dimension of θ equals the dimension of the parameter space Θ. An exponential family of order k is called *regular* if its parameter space Θ is an open set in $\mathbb{R}^k$. Note that the regularity of an exponential family is important in terms of the uniqueness and asymptotic normality of the MLE.

Example: non-full exponential family. One example that an exponential family is not full is the distribution $N(\mu, \mu^2)$, which is a curved-exponential family.

Example: non-regular exponential family. A non-regular exponential family is $p(x;\theta) = \frac{1}{x^2} C(\theta) e^{-\theta x} I(x > 1)$ with $\Theta = (-\infty, 0]$

You can easily show that popular distributions such as Gaussian, Exponential, Poisson, Binomial, Beta all belong to the exponential family. And you can easily find their minimal representation and show that they are full and regular.

The log-likelihood function of an exponential family has a very beautiful form:

$$\ell(\theta|x) = \log p(x;\theta) = \theta^T t(x) - \log C(\theta) + \log h(x).$$

One may notice that the last term $\log h(x)$ does not involve θ, so we can ignore it when computing the MLE.

The exponential family has many elegant properties. For instance, if we have IID observations $X_1, \cdots, X_n \sim p(x;\theta)$, where $p(x;\theta) = \frac{1}{C(\theta)} h(x) \exp(\theta^T t(x))$ belongs to an exponential family, then the joint PDF

$$p(x_1, \cdots, x_n; \theta) = \prod_{i=1}^n p(x_i;\theta) = \frac{1}{C^n(\theta)} \left[\prod_{i=1}^n h(x_i)\right] \exp\left(\theta^T \sum_{i=1}^n t(x_i)\right).$$

The log-likelihood function is

$$\ell_n(\theta|X_1, \cdots, X_n) = \theta^T \sum_{i=1}^n t(X_i) - n \log C(\theta) + \sum_{i=1}^n \log h(X_i),$$

so all information about θ is from the canonical statistic $\sum_{i=1}^n t(X_i)$. This property is related to the classical concept of *sufficient statistic.*

Moment property of $\log C(\theta)$. The derivative of $\log C(\theta)$ is related to the moments of $t(X)$. We denote $\mathbb{E}(t(X)) \equiv \mu_t(\theta)$ and $\mathsf{Cov}(t(X)) \equiv \Sigma_t(\theta)$.

Lemma 1.1 *When X is from an exponential family with $p(x;\theta) = \frac{1}{C(\theta)} h(x)\exp(\theta^T t(x))$. Then*

$$\mathbb{E}(t(X)) \equiv \mu_t(\theta) = \nabla \log C(\theta), \quad \mathsf{Cov}(t(X)) \equiv \Sigma_t(\theta) = \nabla\nabla \log C(\theta).$$

Proof:

To see the expectation, a direct computation of $\nabla \log C(\theta)$ shows that

$$\begin{aligned}\nabla \log C(\theta) &= \frac{1}{C(\theta)} \nabla C(\theta)\\ &= \frac{1}{C(\theta)} \int h(x) \nabla_\theta \exp(\theta^T t(x)) dx\\ &= \frac{1}{C(\theta)} \int h(x) t(x) \exp(\theta^T t(x)) dx\\ &= \mathbb{E}(t(X)).\end{aligned}$$

Similarly, when we take the derivative again, we obtain

$$\begin{aligned}\nabla\nabla \log C(\theta) &= \nabla_\theta \frac{1}{C(\theta)} \int h(x)t(x)\exp(\theta^T t(x))dx\\ &= \left[\nabla_\theta \frac{1}{C(\theta)}\right] \int h(x)t(x)\exp(\theta^T t(x))dx + \frac{1}{C(\theta)} \int h(x)t(x)\nabla_\theta \exp(\theta^T t(x))dx\\ &= -\frac{1}{C^2(\theta)} \left[\int h(x)t(x)\exp(\theta^T t(x))\right] \left[\int h(x)t(x)\exp(\theta^T t(x))\right]^T\\ &\quad + \frac{1}{C(\theta)} \int h(x)t(x)t(x)^T \nabla_\theta \exp(\theta^T t(x))dx\\ &= -\mathbb{E}(t(X))\mathbb{E}(t(X))^T + \mathbb{E}(t(X)t(X))\\ &= \mathsf{Cov}(t(X)).\end{aligned}$$

■

Not only the mean and variance, you can recover higher-order moments via taking derivatives of $\log C(\theta)$. Note that $C(\theta)$ is also called *partition function* in probability theory and statistical physics.

In addition to the moment property, the score of an exponential family also has a nice form. A direct computation shows

$$s(\theta|x) = \nabla \ell(\theta|x) = t(x) - \nabla_\theta \log C(\theta) = t(x) - \mu_t(\theta).$$

Thus, the score equation becomes

$$0 = S_n(\widehat{\theta}) = \sum_{i=1}^n t(X_i) - n\mu_t(\widehat{\theta})$$

or equivalently,

$$\mu_t(\widehat{\theta}) = \frac{1}{n} \sum_{i=1}^n t(X_i).$$

Namely, we find the MLE by tuning the parameter θ until the function $\mu_t(\theta)$ equals the value of $\frac{1}{n}\sum_{i=1}^n t(X_i)$. Thus, you can see that all information about the MLE is determined by the canonical statistic $\sum_{i=1}^n t(X_i)$. When μ_t has a well-defined and unique inverse, we may write the MLE as $\widehat{\theta} = \mu_t^{-1}\left(\frac{1}{n}\sum_{i=1}^n t(X_i)\right)$.

1.3 Confidence Intervals

In some analysis, we not only want to have just a point estimate of the parameter of interest, but also want to use an interval to infer the parameter of interest. And we also want to assign a level to this interval to describe how 'accurate' this interval is. Note that here the concept of accuracy is not well-defined – we will talk about it later. Ideally, given an accuracy level, we want an interval as small as possible.

The Frequentist and the Bayesian defines the accuracy differently so their construction of intervals are also different. In short, the Frequentists defines the accuracy as the *long term frequency coverage* of the underlying true parameter of interest whereas the Bayesian defines the accuracy in terms of covering the posterior probability. In this section, we will talk about the Frequentist approach and the interval is known as the **confidence interval**. The accuracy that Frequentists are using is called the *confidence level*.

Formally, given a confidence level $1-\alpha$, a confidence interval of θ_0 is a random interval $C_{n,\alpha}$ that can be constructed solely from the data (i.e., can be constructed using $X_1,\cdots,X_n$) such that

$$P(\theta \in C_{n,\alpha}) \geq 1-\alpha+o(1).$$

Beware, what is random is not θ but the interval $C_{n,\alpha}$. The quantity $P(\theta \in C_{n,\alpha})$ is also called the (Frequentist) coverage. Note that we allow the coverage to be *asymptotically* $1-\alpha$; when there is no $o(1)$ term, we will say that the confidence interval has a finite sample coverage. A confidence interval with the above property is also called a (asymptotically) valid confidence interval.

Normal confidence interval. A traditional approach to constructing a confidence interval of θ_0 is based on the asymptotic normality of the MLE:

$$\sqrt{n}(\widehat{\theta}_{MLE}-\theta_0) \overset{D}{\to} N(0, I_1^{-1}(\theta_0)).$$

When the dimension of the parameter is 1, a simple confidence interval is

$$\widehat{\theta}_{MLE} \pm z_{1-\alpha/2}\cdot\sigma_{\theta_0},$$

where $\sigma^2_{\theta_0} = \frac{1}{n}I_1^{-1}(\theta_0)$ Such interval is not a confidence interval because θ_0 is unknown. We can modify it using a plug-in estimate of the Fisher's information:

$$C_{n,\alpha} = [\widehat{\theta}_{MLE} - z_{1-\alpha/2}\cdot\sigma_{\widehat{\theta}_{MLE}}, \widehat{\theta}_{MLE} + z_{1-\alpha/2}\cdot\sigma_{\widehat{\theta}_{MLE}}],$$

where z_β the β-percentile of the standard normal distribution. Using the slutsky's theorem, you can easily show that this confidence interval has the asymptotic coverage.

When the dimension of the parameter is greater than 1, there are multiple ways we can construct a confidence interval. Note that in this case, the set $C_{n,\alpha}$ is no longer an interval but a region/set so it is often called a confidence region/set. A simple approach of constructing a confidence set is via an ellipse. Note that the asymptotic normality also implies that (using continuous mapping theorem)

$$n(\widehat{\theta}_{MLE}-\theta_0)^T I_1(\widehat{\theta}_{MLE})(\widehat{\theta}_{MLE}-\theta_0) \overset{D}{\to} \chi^2_p,$$

where χ^2_p denotes the χ^2 distribution with a degree of freedom p. So we construct the confidence set using

$$C_{n,\alpha} = \left\{\theta : n(\widehat{\theta}_{MLE}-\theta)^T I_1(\widehat{\theta}_{MLE})(\widehat{\theta}_{MLE}-\theta) \leq \chi^2_{p,1-\alpha}\right\},$$

where $\chi^2_{p,\beta}$ is the β-percentile of the χ^2 distribution with a degree of freedom p.

Bootstrap confidence interval. Bootstrap approach is an Monte Carlo method for assessing the uncertainty of an estimator. It can be used to compute the variance of an estimator (not necessarily the MLE) and construct a confidence interval. In the case of likelihood inference, the bootstrap approach has an advantage that we do not need to know the closed-form of $I_1(\theta)$ to construct the confidence interval or to approximate the variance of the MLE.

While there are many variant of bootstrap methods, we introduce the simplest one – the empirical bootstrap. For simplicity, we assume that the dimension of θ is 1 (the bootstrap works for higher dimensions as well). Let $X_1, \cdots, X_n$ be the original sample. We then *sample with replacement* from the original sample to obtain a new sample of each size $X^*_1, \cdots, X^*_n$. This new sample is called a bootstrap sample. We find the MLE using the bootstrap sample and let $\widehat{\theta}^*_{MLE}$ denote the bootstrap MLE. Now we repeat the bootstrap process B times, leading to B bootstrap MLEs

$$\widehat{\theta}^{*(1)}_{MLE}, \cdots, \widehat{\theta}^{*(B)}_{MLE}.$$

Let t_β denotes the β-percentile of these B values, i.e.,

$$\widehat{t}_\beta = \widehat{G}^{-1}(\beta), \quad \widehat{G}(t) = \frac{1}{B}\sum_{b=1}^{B} I(\widehat{\theta}^{*(b)}_{MLE} \leq t).$$

Then the bootstrap confidence interval of θ_0 is

$$C_{n,\alpha} = [\widehat{t}_{\alpha/2}, \widehat{t}_{1-\alpha/2}].$$

One can prove that under very mild conditions, the bootstrap confidence interval has asymptotic coverage.

The power of the bootstrap method is that *we do not use anything about the Fisher's information*! As long as we can compute the estimator, we can construct an asymptotically valid confidence interval. Note that if we do know the Fisher's information, the bootstrap method can be modified using the bootstrap t-distribution method, which provides a better asymptotic coverage (namely, the $o(1)$ decays faster to 0 than the above method and the normal confidence interval)[1].

1.4 Test of Significance

Statistical test is about how to design a procedure that allows us to make scientific discovery. Such a procedure has to be able to handle the uncertain nature of our data. In statistics, we model the data as random variables so the testing procedure needs to account for the randomness.

Let $\mathcal{D}_n = \{X_1, \cdots, X_n\}$ denotes our data. The testing procedure involves two competing hypotheses:

- *Null hypothesis* H_0: the hypothesis that we want to challenge. It is often related to the current scientific knowledge.
- *Alternative hypothesis* H_a: the hypothesis that complements to the null hypothesis. It is the hypothesis we would like to prove to be plausible using our data.

The goal is to see if we have strong enough evidence (from $\mathcal{D}_n$) that we can argue the alternative hypothesis is more reasonable than the null hypothesis. If we do have enough evidence, then we will reject the null

[1] see, e.g., Chapter 2 of *All of nonparametric statistics* by Larry Wasserman and Chapter 3.5 of *The bootstrap and Edgeworth expansion* by Peter Hall.

hypothesis. When the null hypothesis reflects the scenarios that can be explained by the current scientific knowledge, rejecting the null hypothesis means that we have discovered something new.

Here is a summary of hypothesis test.

1. Based on the model and null hypothesis, design a test statistic.
2. Compute the distribution of the test statistics under the null hypothesis.
3. Plug-in the data into the test statistic, compute the probability of observing a more extreme data against the null hypothesis. Such a probability is the p-value.
4. Compare p-value to the significance level. If p-value is less than the significance level, we reject the null hypothesis.

The central idea of hypothesis test is to control the *type-1 error*, the probability of falsely rejecting H_0 when H_0 is correct. Essentially, the p-value can be interpreted as *if we reject the null hypothesis (under this p-value), then our type-1 error is the same as the p-value*. The significance level reflects the amount of type-1 error we can tolerate so when p-value is less than the significance level, we can reject H_0. Due to the construction of p-value, a small p-value means that the null hypothesis does not fit to the data very well (so we are seeing an extreme event if H_0 is true). Thus, small p-value or rejecting H_0 under a small significance level means that we have more evidence against H_0.

Note that there is another quantity called *type-2 error*, the probability of not rejecting H_0 when H_0 is false. Namely, type-2 error is concerned with the case that we fail to reject H_0 when we should.

In statistics, we often control type-1 error first and the hope that the type-2 error is also small. When do we put more emphasis on type-1 error? This has something to do with the philosophy of scientific research. The scientific approach is a systematic way to acquire reliable knowledge. Thus, every discovery we made should be accompanied with sufficient evidences. In Frequentist approach, the measure of evidence against H_0 is the p-value – the smaller p-value, the more evidence. Thus, controlling type-1 error means that we put requirements on the amount of evidence we need to claim a scientific discovery.

1.4.1 Three popular tests for parameters

For a parametric model, the null hypothesis can often be expressed as

$$H_0 : \theta \in \Theta_0, \tag{1.6}$$

where $\Theta_0 \subset \Theta$. The alternative hypothesis will be $H_A : \theta \in \Theta \backslash \Theta_0$. We assume that $\mathsf{dim}(\Theta) = p$ and $\mathsf{dim}(\Theta_0) = s < p$.

While there are many possible ways to construct a test statistic of testing H_0 in equation (1.6), here we introduce four popular test and later we will argue that they are all asymptotically equivalent in simple case. Let the MLE be

$$\widehat{\theta} = \mathsf{argmin}_{\theta \in \Theta} L_n(\theta)$$

and the MLE under H_0 be

$$\widehat{\theta} = \mathsf{argmin}_{\theta \in \Theta_0} L_n(\theta).$$

Note that when the null hypothesis is not composite, we have to compute the MLE $\widehat{\theta}_0$ under the null because there are infinite number of feasible parameters in Θ_0.

Wald test. The Wald test is based on the insight from the asymptotic normality in equation (1.4):

$$\sqrt{n}(\widehat{\theta}-\theta^*) \xrightarrow{d} N(0, I_1^{-1}(\theta^*))$$

assuming that the parametric model is correct. When H_0 is correct, the best approximation to θ^* is $\widehat{\theta}_0$. So we use the test statistic

$$T_{n,\mathsf{Wald}} = n(\widehat{\theta}-\widehat{\theta}_0)^T I_1(\widehat{\theta}_0)(\widehat{\theta}-\theta^*).$$

Under H_0, this test statistic

$$T_{n,\mathsf{Wald}} \xrightarrow{d} \chi^2_{p-s}.$$

Later we will discuss why the degrees of freedom is $p-s$.

Score test (Rao's test). In our derivation of the asymptotic normality of MLE (Section 1.2.1), the 'normality' comes from the score function:

$$\frac{1}{n}S_n(\theta^*) = \frac{1}{n}\sum_{i=1}^n s(\theta^*|X_i).$$

One can easily show that

$$\frac{1}{\sqrt{n}}S_n(\theta^*) \xrightarrow{d} N(0, I_1(\theta^*)).$$

Thus, under H_0, the best approximation to θ^* is $\widehat{\theta}_0$, so the score test is based on the test statistic

$$T_{n,\mathsf{Score}} = nS_n(\widehat{\theta}_0)^T I_1^{-1}(\widehat{\theta}_0)S_n(\widehat{\theta}_0).$$

Note that while $S_n(\widehat{\theta})=0$, $S_n(\widehat{\theta}_0)\neq 0$ in general. Similar to the Wald test, the score test has an asymptotic distribution

$$T_{n,\mathsf{Score}} \xrightarrow{d} \chi^2_{p-s}$$

under H_0.

Likelihood ratio test (LRT). The LRT is one of the most popular test in this scenario. It uses the test statistic

$$T_{n,\mathsf{LRT}} = 2\log\frac{L_n(\widehat{\theta})}{L_n(\widehat{\theta}_0)}.$$

Similar to the above two tests, the LRT has the asymptotic distribution

$$T_{n,\mathsf{LRT}} \xrightarrow{d} \chi^2_{p-s},$$

which is also known as the *Wilk's theorem*. Here is a rough derivation of how the asymptotic distribution of the LRT comes from. Let $H_n(\theta)=\nabla_\theta S_n(\theta)=\nabla_\theta\nabla_\theta \ell_n(\theta)$ be the Hessian matrix of the log-likelihood function. By the Tyler's theorem, the LRT test statistic can be expanded as

$$\begin{aligned}
T_{n,\mathsf{LRT}} &= 2\log\frac{L_n(\widehat{\theta})}{L_n(\widehat{\theta}_0)}\\
&= -2(\ell_n(\widehat{\theta}_0)-\ell_n(\widehat{\theta}))\\
&= -2\left(S_n(\widehat{\theta}_0)^T(\widehat{\theta}_0-\widehat{\theta}) + \frac{1}{2}(\widehat{\theta}_0-\widehat{\theta})^T H_n(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}) + o_P(\|(\widehat{\theta}_0-\widehat{\theta})\|^2)\right).
\end{aligned}$$

The first term $S_n(\widehat{\theta}_0) \overset{P}{\to} 0$ so we can ignore it. The dominating term is the second quantity. Note that equation (1.2) implies that $\frac{1}{n}H_n(\theta) \overset{P}{\to} -I_1(\theta)$, so

$$\begin{aligned}(\widehat{\theta}_0-\widehat{\theta})^T H_n(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}) &= n(\widehat{\theta}_0-\widehat{\theta})^T \frac{1}{n} H_n(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta})\\ &\approx n(\widehat{\theta}_0-\widehat{\theta})^T I_1(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}).\end{aligned}$$

As a result,

$$T_{n,\mathsf{LRT}} \approx n(\widehat{\theta}_0-\widehat{\theta})^T I_1(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}). \tag{1.7}$$

In the next section, we will argue that $n(\widehat{\theta}_0-\widehat{\theta})^T I_1(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}) \overset{d}{\to} \chi^2_{p-s}$. So again, LRT has the same asymptotic distribution as the Wald and score tests.

1.4.2 A geometric interpretation of the $p-s$ degrees of freedom

Now we argue that under H_0,

$$n(\widehat{\theta}_0-\widehat{\theta})^T I_1(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}) \overset{d}{\to} \chi^2_{p-s}.$$

First, under H_0, both $\widehat{\theta}$ and $\widehat{\theta}_0$ are approaching the true parameter $\theta_0 \in \Theta_0$, so the information matrix $I_1(\widehat{\theta}_0) \overset{p}{\to} I_1(\theta_0)$. As a result, this quantity behaves like

$$n(\widehat{\theta}_0-\widehat{\theta})^T I_1(\widehat{\theta}_0)(\widehat{\theta}_0-\widehat{\theta}) \approx n(\widehat{\theta}_0-\widehat{\theta})^T I_1(\theta_0)(\widehat{\theta}_0-\widehat{\theta}).$$

Moreover, the above quantity is equivalent to

$$n(\widehat{\theta}-\widehat{\theta}_0)^T I_1(\theta_0)(\widehat{\theta}-\widehat{\theta}_0). \tag{1.8}$$

Now we decompose $(\widehat{\theta}_0-\widehat{\theta})$ as

$$\widehat{\theta}-\widehat{\theta}_0 = (\widehat{\theta}-\theta_0)-(\widehat{\theta}_0-\theta_0).$$

Now suppose the space Θ_0 is smooth around θ_0. Specifically, we need Θ_0 to be an s-dimensional manifold around θ_0. Namely, for region in Θ_0 around θ_0, this region behaves like an s-dimensional Euclidean space. Then the deviation $\widehat{\theta}_0-\theta_0$ is the in this local s-dimensional Euclidean space (technically, it belongs to the 'tangent' space of the manifold Θ_0 at θ_0) because we have a constraint $\widehat{\theta}_0 \in \Theta_0$. The other quantity $\widehat{\theta}-\theta_0$ is without the constraint, so it can be decomposed into

$$\widehat{\theta}-\theta_0 = V_N + V_T,$$

where V_N is the projection of vector $\widehat{\theta}-\theta_0$ onto the normal space of Θ_0 at θ_0 and V_T is the projection of $\widehat{\theta}-\theta_0$ onto the tangent space of Θ_0 at θ_0. Asymptotically, $V_T \approx \widehat{\theta}_0-\theta_0$, so

$$\widehat{\theta}-\widehat{\theta}_0 = (\widehat{\theta}-\theta_0)-(\widehat{\theta}_0-\theta_0) \approx V_N,$$

the normal component of $\widehat{\theta}-\theta_0$. The information matrix $I_1(\theta_0)$ in equation (1.8) has normalize $(\widehat{\theta}-\theta_0)$ and $(\widehat{\theta}_0-\theta_0)$ so that they behave like a standard normal vector. Thus,

$$n(\widehat{\theta}-\widehat{\theta}_0)^T I_1(\theta_0)(\widehat{\theta}-\widehat{\theta}_0)$$

only has the component from the normal direction of $\widehat{\theta}-\theta_0$. Given that Θ_0 has s dimensions, the normal space of Θ_0 has $p-s$ dimensions. Accordingly,

$$n(\widehat{\theta}-\widehat{\theta}_0)^T I_1(\theta_0)(\widehat{\theta}-\widehat{\theta}_0) \approx \|Z_{p-s}\|^2 \overset{d}{=} \chi^2_{p-s},$$

where $Z_{p-s} \in \mathbb{R}^{p-s}$ is a standard normal vector. This proves the asymptotic distribution of $T_{n,\mathsf{LRT}}$.

For the case of Wald and score tests, we can show that the test statistic is asymptotically the same as the LRT, so the asymptotic distribution of LRT also applies to the other two tests. While these three tests are asymptotically the same in this nice scenario, they may be very different under other scenarios; see https://arxiv.org/abs/1807.04431.

Based on the above derivation, you can see that the 'distance' between two parameter θ_1 and θ_2 in the LRT scenario is determined by the information matrix $I_1(\theta)$. This means that the parameter space Θ is not flat but curved. Using terms from differential geometry, the Riemannian metric of Θ is $I_1(\theta)$, which is the information matrix. This leads to the field of *information geometry*[2].

Moreover, the above analysis shows that the LRT and the algebraic/geometric structure of Θ_0, the parametric under the null hypothesis, is closely related. This insight leads to the field of algebraic statistics[3], a study of how the algebraic structure of the parameter space interacts with the behavior of a statistic.

1.5 Model Mis-specification

Many theory about the MLE assumes that the population distribution function belongs to our parametric family. However, this is a very strong assumption in reality. It is very likely that the population distribution function does not belong to our parametric family (e.g., the population PDF is not Gaussian but we fit a Gaussian to it). What will happen in this case for our MLE? Will it still converge to something? If so, what will be the quantity that it is converging?

Model mis-specification studies the situation like this – we assume a wrong model for the population distribution function. Let $p_0(x)$ be the population PDF and we assume that the population PDF can be written as $p(x;\theta)$. However, $p_0 \neq p(x;\theta)$ for every $\theta \in \Theta$. It turns out that the MLE $\widehat{\theta}_{MLE}$ still converges under mild assumptions to a quantity θ^* in probability. Moreover, the corresponding PDF/PMF $p(x;\theta^*)$ has an interesting relation with $p_0(x)$. Assume that the RV X has a PDF/PMF p_0. Then

$$\mathbb{E}\left\{\log\left(\frac{p_0(X)}{p(X;\theta^*)}\right)\right\} = \inf_{\theta\in\Theta} \mathbb{E}\left\{\log\left(\frac{p_0(X)}{p(X;\theta)}\right)\right\} = \inf_{\theta\in\Theta} \mathsf{KL}(p_0, p_\theta),$$

where KL is also known as the *Kullback-Liebler (KL)* divergence and $p_\theta(x) = p(x;\theta)$. Namely, the MLE corresponds to the parametric distribution in the specified family that minimizes the KL divergence to the population distribution. To see why the population model θ^* is the one that minimizes the KL divergence, we first recall the definition of population log-likelihood function:

$$\ell(\theta) = \mathbb{E}(\theta(\theta|X)) = \mathbb{E}(\log p(X;\theta)),$$

where $X \sim p_0$. As a result,

$$\ell(\theta) = \int p_0(x) \log p(x;\theta) dx.$$

Maximizing $\ell(\theta)$ is the same as maximizing $\ell(\theta) - \int p_0(x) \log p_0(x) dx$ since the later part does not involve

[2] See https://en.wikipedia.org/wiki/Information_geometry for more details.
[3] See https://en.wikipedia.org/wiki/Algebraic_statistics.

θ. So we conclude

$$\begin{aligned}
\theta^* &= \mathsf{argmax}_{\theta\in\Theta}\ell(\theta) \\
&= \mathsf{argmax}_{\theta\in\Theta}\ell(\theta) - \int p_0(x)\log p_0(x)dx \\
&= \mathsf{argmax}_{\theta\in\Theta}\int p_0(x)\log\left(\frac{p(x;\theta)}{p_0(x)}\right)dx \\
&= \mathsf{argmax}_{\theta\in\Theta} - \mathsf{KL}(p_0, p_\theta).
\end{aligned}$$

In model mis-specification case, the MLE still satisfies the score equation (under appropriate assumption) but the Fisher's information may not reflect the actual curvature of the likelihood function around θ^*. Equation (1.3) still holds, i.e.,

$$\sqrt{n}(\widehat{\theta}_{MLE} - \theta^*) \overset{D}{\to} N(0, \Sigma),$$

where $\Sigma = I_1^{-1}(\theta^*)\mathbb{E}(S_1(\theta^*)S_1^T(\theta^*))I_1^{-1}(\theta^*)$ is the covariance matrix.