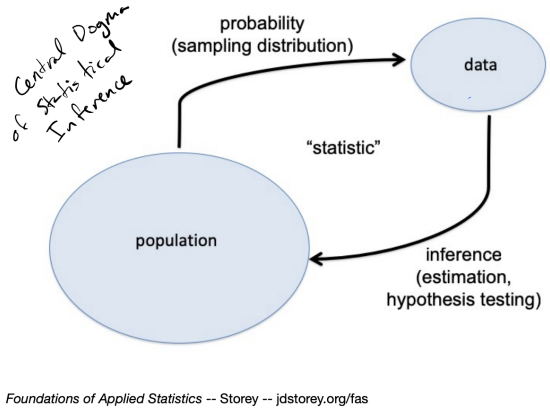
QCB 408 / 508 – Notes on Week 4

Kaiqian Zhang

2020-03-06

1 Summary

Recall the following diagram of "central dogma of statistical inference". Previously, we have talked about the probability part from population to data. Today we will discuss the inference part from data to population. We will begin statistical inference by introducing maximum likelihood estimation (MLE). Then we will learn about exponential family distributions (EFDs), whose special form has certain mathematical convenience in inference. Finally, We will study what pivotal statistic is and some inference strategies used by Frequentists.



Here is a list of topics for this week:

- Likelihood and maximum likelihood estimation (MLE),
- Exponential family distributions (EFDs),
- Frequentist inference from pivotal statistic.

2 Take-away This Week

- A *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. A statistic $T(\boldsymbol{X})$ is a *sufficient statistic* for θ if $f(\underline{\boldsymbol{x}};\boldsymbol{\theta}) = g(T(\underline{\boldsymbol{x}});\boldsymbol{\theta})h(\underline{\boldsymbol{x}})$.
- A *pivotal statistic* is a function of observations and unobservable parameters such that the function's probability distribution does not depend on the unknown parameters.
- The *MLE* is the value of parameter $\boldsymbol{\theta}$ that maximizes the likelihood. We usually work with log-liklihood rather than likelihood in the maximum likelihood estimation.

- MLE has three important properties: (1) It is consistent. (2) It has equivariance, i.e., if $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$, then $g(\hat{\boldsymbol{\theta}}_n)$ is the MLE of $g(\boldsymbol{\theta})$. (3) We can use Fisher information to approximate the standard error of MLE.
- We can construct an asymptotic pivotal statistic with MLE by using central limit theorem of MLE.
- Here is a useful table of MLEs, standard errors, and pivotal statistics for three common distributions.

Distribution	MLE	Std Err	Z Statistic
Binomial(n,p)	$\hat{p}=X/n$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
Normal(μ,σ^2)	$\hat{\mu}=\overline{X}$	$\frac{\hat{\sigma}}{\sqrt{n}}$	$\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$
Poisson(λ)	$\hat{\lambda}=\overline{X}$	$\sqrt{\frac{\hat{\lambda}}{n}}$	$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}}$

- The pdf or pmf of an *exponential family distribution (EFD)* parametrized on the observed scale by $\underline{\boldsymbol{\theta}}$ has the following form:

$$f(x;\underline{\boldsymbol{\theta}}) = h(x)\exp\{\sum_{k=1}^{d}\eta_k(\underline{\boldsymbol{\theta}})T_k(x) - A(\boldsymbol{\eta})\},$$

  where $\boldsymbol{\eta} = (\eta_1(\underline{\boldsymbol{\theta}}), \eta_2(\underline{\boldsymbol{\theta}}), \ldots, \eta_d(\underline{\boldsymbol{\theta}}))^T$ and $T_1(x), T_2(x), \ldots, T_d(x)$ are sufficient statistics for $\eta_1, \eta_2, \ldots, \eta_d$.
- There are three goals in the frequentist statistical inference: (1) point estimate of θ; (2) confidence interval of θ, which is uncertainty of point estimate; (3) hypothesis test to assess specific value(s) of θ.

3 Likelihood and Maximum Likelihood Estimation

3.1 Likelihood

Given a model with independent and identically distributed random variables

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$$

or jointly distributed random variables

$$(X_1, X_2, \ldots, X_n) \sim F_{\boldsymbol{\theta}},$$

where F is some distribution and $\boldsymbol{\theta}$ are parameters of F. Note that parameters $\boldsymbol{\theta}$ should be informative about what we want to know about the population. Note that there are two levels in a study:

- $x_1, x_2, \ldots, x_n$ are observed data from the study. We also write $\underline{\boldsymbol{x}}$ as a concise form of observed data.
- $X_1, X_2, \ldots, X_n$ are random variables that model observed data.

Definition: We further assume that we have a pdf or a joint pmf for the above model $f(.;\boldsymbol{\theta})$. If we are going to evaluate this on observed data $\underline{\boldsymbol{x}}$, we obtain a function of $\boldsymbol{\theta}$,

$$f(\underline{\boldsymbol{x}};\boldsymbol{\theta}),$$

which is also *likelihood*

$$L(\boldsymbol{\theta}; \underline{\boldsymbol{x}}).$$

Note that likelihood is a function of parameters $\boldsymbol{\theta}$ given observed data.

Example from Casella & Berger: (*Negative binomial likelihood*) Let X have a negative binomial distribution with $r = 3$ and success probability p. If $x = 2$ is observed, then the likelihood function is the fifth-degree polynomial on $0 \leq p \leq 1$ defined by

$$L(p|2) = P_p(X = 2) = \binom{4}{2} p^3(1-p)^2.$$

In general, if $X = x$ is observed, then the likelihood function is the polynomial of degree $3 + x$,

$$L(p|x) = \binom{3+x-1}{x} p^3(1-p)^x.$$

Extension from FAS: **Likelihood Principle** states that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $L(\theta|\boldsymbol{x})$ is proportional to $L(\theta|\boldsymbol{y})$, that is, there exists a constant $c(\boldsymbol{x}, \boldsymbol{y})$ such that

$$L(\theta|\boldsymbol{x}) = c(\boldsymbol{x}, \boldsymbol{y}) L(\theta|\boldsymbol{y}) \text{for all } \theta,$$

then the conclusion drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical.

3.2 Log-likelihood

Definition: *Log-likelihood* is

$$\log L(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) = l(\boldsymbol{\theta}; \underline{\boldsymbol{x}}).$$

We usually work with log-likelihood because log-likelihood has a sum form rather than a product form in likelihood. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f_{\boldsymbol{\theta}}$. Then log-likelihood of the observed data is

$$\begin{aligned} l(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) &= \log(L(\boldsymbol{\theta}; \underline{\boldsymbol{x}})) \\ &= \log(f(\underline{\boldsymbol{x}}; \boldsymbol{\theta})) \\ &= \log(\prod_{i=1}^{n} f(\underline{x}_i; \boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} \log f(\underline{x}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} l(\boldsymbol{\theta}; \underline{x}_i). \end{aligned}$$

3.3 Sufficient statistic

A *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

Definition: A statistic $T(\boldsymbol{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\underline{\boldsymbol{x}}$ given the value of $T(\underline{\boldsymbol{x}})$ does not depend on θ. In other words, if $f(\underline{\boldsymbol{x}};\boldsymbol{\theta}) = g(T(\underline{\boldsymbol{x}});\boldsymbol{\theta})h(\underline{\boldsymbol{x}})$, then $T(\underline{\boldsymbol{x}})$ is sufficient. Note that for the likelihood,

$$L(\boldsymbol{\theta};\underline{\boldsymbol{x}}) = g(T(\underline{\boldsymbol{x}});\boldsymbol{\theta})h(\underline{\boldsymbol{x}}) \propto L(\boldsymbol{\theta};T(\underline{\boldsymbol{x}})).$$

Example: (*Normal sufficient statistic*) Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We will show that $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is a sufficient statistic for μ. Hint: $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$.

Proof: The joint pdf of the sample $\underline{\boldsymbol{x}}$ is

$$\begin{aligned}
f(\underline{\boldsymbol{x}}|\mu) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2}\exp(-(x_i) - \mu)^2/(2\sigma^2) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2)\right) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2/(2\sigma^2)\right) && \text{(add and subtract } \bar{x}) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 + 2\sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \mu)\right)/(2\sigma^2)\right) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right) && \text{(since } \sum_{i=1}^n x_i - \bar{x} = 0).
\end{aligned}$$

Recall that the sample mean $\overline{X}_n$ follows a $\mathcal{N}(\mu, \sigma^2/n)$ distribution. Thus, the ratio of pdfs is

$$\begin{aligned}
\frac{f(\underline{\boldsymbol{x}}|\mu)}{g(T(\underline{\boldsymbol{x}}) = \overline{X}_n|\mu)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right)}{(2\pi\sigma^2)^{-1/2}\exp(-n(\bar{x} - \mu)^2/(2\sigma^2))} \\
&= n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2}\exp\left(-\sum_{i=1}^n (x_i - \bar{x})^2/(2\sigma^2)\right),
\end{aligned}$$

which does not depend on μ. Therefore the sample mean is a sufficient statistic for μ.

Example from Casella & Berger: (*Binomial sufficient statistic*) Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameters θ, $0 < \theta < 1$. We will show that $T(\boldsymbol{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ.

Proof: Note that $T(\boldsymbol{X})$ counts the number of X_is that equal 1, so $T(\boldsymbol{X})$ has a Binomial(n, θ) distribution. The ratio of pmfs is thus

$$
\begin{aligned}
\frac{p(\boldsymbol{x}|\theta)}{g(T(\boldsymbol{x})|\theta)} &= \frac{\prod \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && (\text{define } t = \sum x_i) \\
&= \frac{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && (\prod \theta^{x_i} = \theta^{\sum x_i}) \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{1}{\binom{n}{t}} \\
&= \frac{1}{\binom{n}{\sum x_i}}.
\end{aligned}
$$

Since this ratio does not depend on θ, $T(\boldsymbol{X})$ is a sufficient statistic for θ. The interpretation is this: The total number of 1s in this Bernoulli sample contains all the information about θ that is in the data. Other features of the data, such as the exact value of X_3, contain no additional information.

3.3.1 Extensions of sufficient statistic

- Minimal sufficient statistic: A sufficient statistic is *minimal sufficient* if it can be represented as a function of any other sufficient statistic. In other words, $S(X)$ is minimal sufficient if and only if
  - $S(X)$ is sufficient, and
  - if $T(X)$ is sufficient, then there exists a function f such that $S(X) = f(T(X))$.
- Complete statistic: A statistic T is *complete* if $\mathrm{E}[g(T)] = 0$ for all θ and some function g implies that $P(g(T) = 0; \theta) = 1$ for all θ.
- Ancillary statistic: An *ancillary statistic* is a measure of a sample whose distribution does not depend on the parameters of the model. An ancillary statistic is a pivotal quantity that is also a statistic.
- Basu's theorem: If $T(X)$ is complete and sufficient (for $\theta \in \Theta$), and $S(X)$ is ancillary, then $S(X)$ and $T(X)$ are independent for all $\theta \in \Theta$.

3.4 Maximum likelihood estimation

Definition: The *MLE* is the value of $\boldsymbol{\theta}$ that maximizes the likelihood. In math,

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{MLE} &= \mathrm{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) \\
&= \mathrm{argmax}_{\boldsymbol{\theta}} l(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) \\
&= \mathrm{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; T(\underline{\boldsymbol{x}})).
\end{aligned}
$$

Example: (*Binomial maximum likelihood estimation*) Given that $X \sim \mathrm{Binomial}(n, p)$. We will show the MLE for p is $\frac{x}{n}$.

Proof: The likelihood is

$$
L(p; x) = \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x}.
$$

And the log-likelihood is

$$l(p;x) \propto x\log p + (n-x)\log(1-p).$$

We take derivative of the log-likelihood $l(p;x)$ with respect to p and set it zero to obtain MLE:

$$\frac{\partial}{\partial p} l(p;x) = 0 \Rightarrow \hat{p}_{\text{MLE}} = \frac{x}{n}.$$

```
> set.seed(508)
> n <- 1000
> p <- 0.3
> x <- rbinom(1, n, p)
> log.likelihood <- dbinom(x, n, prob = seq(0,1, by=0.01), log=TRUE)
> df <- data.frame(p=seq(0,1, by=0.01), logLik = log.likelihood)
```

```
> ggplot(df, aes(x=p, y=logLik))+
+   geom_point(color="#B57865", alpha=0.7, size=2) +
+   theme_minimal() +
+   xlab("p") +
+   ylab("log-likelihood") +
+   geom_vline(xintercept = x/n, color="#B29082", size=1, alpha=0.5) +
+   geom_text(aes(x=x/n, label="\n p = x/n", y=-2500),
+             colour="darkgray", text=element_text(size=2))
```

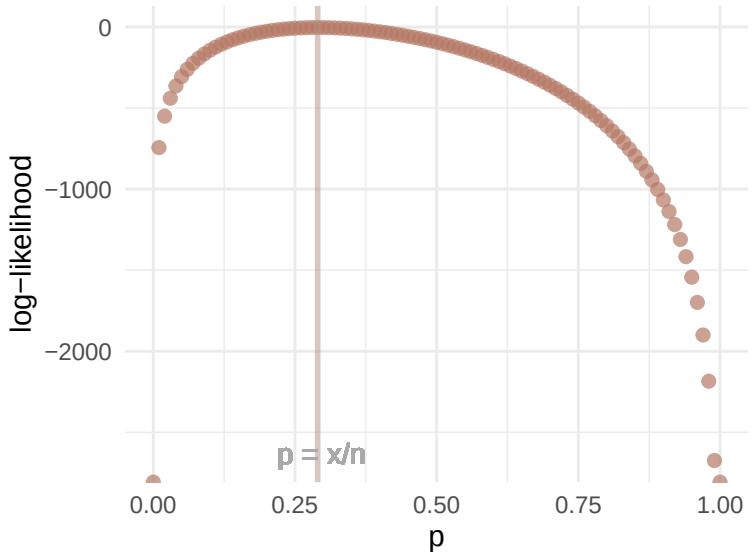


Figure 1: Binomial Log-likelihood versus Parameter p

Here is a visualization of a minimal simulation on Binomial MLE. We notice that the log-likelihood obtains maximum when p is around $\frac{x}{n}$, which is the MLE as we computed above.

3.5 Pivotal statistic

Definition: A *pivotal statistic* is a function of observations and unobservable parameters such that the function's probability distribution does not depend on the unknown parameters.

Example: (*Pivotal statistic for Binomial*) Using the example above, we have a pivotal statistic

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

for large n. Suppose we observe that $\hat{p} = 0.32$. We want to know the sampling distribution of $\hat{p}$. So we can consider the pivotal statistic for this purpose. Note that in this pivotal statistic, we observe $\hat{p}$ but p is unknown. The distribution of this pivotal statistic is $\mathcal{N}(0,1)$, which does not involve p.

Example from FAS: (*Pivotal statistic for Normal*) Given that $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. The pivotal statistic

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}(0,1).$$

Note that in general, for a random variable Y, it is the case that $(Y - \mathrm{E}[Y])/\sqrt{\mathrm{Var}(Y)}$ has population mean 0 and variance 1.

3.6 Important properties for MLE

3.6.1 Assumptions

We assume the following conditions:

- $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$,
- $\hat{\boldsymbol{\theta}}_n$ is the MLE from data,
- "regularity conditions" are met.

3.6.2 Three important properties

- MLE is "consistent". In other words,

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}.$$

- Equivariance: If $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$, then $g(\hat{\boldsymbol{\theta}}_n)$ is the MLE of $g(\boldsymbol{\theta})$.

Example: (*Binomial MLE*) Suppose $X \sim$ Binomial(n,p). The MLE for p is $\hat{p} = \frac{x}{p}$. Then the MLE of $\mathrm{Var}(X) = np(1-p)$ is $n\hat{p}(1-\hat{p})$.

- Fisher information is define as followed.

$$I_n(\theta) = \text{Var}(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \underline{\boldsymbol{x}})) \tag{1}$$

$$= \text{Var}(\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} l(\boldsymbol{\theta}; x_i)) \tag{2}$$

$$= \sum_{i=1}^{n} \text{Var}(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; x_i)) \tag{3}$$

$$= -\mathbb{E}[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta}; \underline{\boldsymbol{x}})] \tag{4}$$

$$= -\sum_{i=1}^{n} \mathbb{E}[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta}; x_i)] \tag{5}$$

Fisher information gives you standard error of MLE. In general, the standard error of an estimator is the standard deviation of the sampling distributions of the estimator. For MLEs, the standard error of $\hat{\boldsymbol{\theta}}_n$ is

$$\text{se}(\hat{\boldsymbol{\theta}}_n) \equiv \sqrt{\text{Var}(\hat{\boldsymbol{\theta}}_n)}.$$

And we can use Fisher information to approximate

$$\text{se}(\hat{\boldsymbol{\theta}}_n) \approx \frac{1}{\sqrt{I_n(\boldsymbol{\theta})}},$$

where high Fisher information $I_n(\boldsymbol{\theta})$ means low standard error. Note that $I_n(\boldsymbol{\theta})$ uses true $\boldsymbol{\theta}$. We usually use $\hat{\text{se}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{I_n(\hat{\boldsymbol{\theta}})}}$ as the standard error estimator of an MLE.

Example: (*Binomial Fisher information*) Suppose $X \sim \text{Binomial}(n, p)$. Then $I_n(p) = \frac{n}{p(1-p)}$.

Proof: The log-likelihood is

$$l(n, p; x) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p).$$

We compute the first and second derivative of log-likelihood with respect to p.

$$\frac{\partial}{\partial p} l(n, p; x) = xp^{-1} + (n-x)(p-1)^{-1}$$
$$\frac{\partial^2}{\partial p^2} l(n, p; x) = -xp^{-2} + (x-n)(p-1)^{-2}.$$

Note that

$$\begin{aligned}
\mathbb{I}(p) &= -\mathbb{E}[\frac{\partial^2}{\partial p^2} l(n, p; x)] \\
&= -\left(-p^{-2}\mathbb{E}[x] + (p-1)^{-2}\mathbb{E}[x-n]\right) \\
&= -\left(-\frac{np}{p^2} + \frac{np-n}{(p-1)^2}\right) && \text{(since } \mathbb{E}[x] = np\text{)} \\
&= \frac{n}{p(1-p)}.
\end{aligned}$$

3.7 Central limit theorem for MLE

Theorem: Let $\hat{\boldsymbol{\theta}}_n$ be the MLE of $\boldsymbol{\theta}$. As $n \to \infty$,

$$\frac{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}}{\text{se}(\hat{\boldsymbol{\theta}}_n)} \xrightarrow{D} \mathcal{N}(0,1),$$

$$\frac{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}}{\hat{\text{se}}(\hat{\boldsymbol{\theta}}_n)} \xrightarrow{D} \mathcal{N}(0,1).$$

3.7.1 Asymptotic pivotal statistic

We define

$$Z = \frac{\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}}{\hat{\text{se}}(\hat{\boldsymbol{\theta}}_n)}.$$

By Central Limit Theorem above, as $n \to \infty$

$$Z \xrightarrow{D} \mathcal{N}(0,1).$$

Thus, Z is approximately a pivotal statistic, which we call asymptotic pivotal statistic.

3.8 Optimality of MLE

The MLE $\hat{\boldsymbol{\theta}}_n$ is such that as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(0, \tau^2).$$

Suppose $\tilde{\boldsymbol{\theta}}_n$ is any other estimator where as $n \to \infty$,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

In follows that $\frac{\tau^2}{\sigma^2} \leq 1$, which suggests that MLE is asymptotically optimal.

3.9 Delta method

Definition *from Casella & Berger*: (*Delta method*) Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.}$$

If $g(.)$ is a differentiable function and $g'(\theta) \neq 0$. We have

$$g(t) \approx g(\theta) + g'(\theta)(t - \theta).$$

Suppose I have $\hat{\text{se}}(\hat{\boldsymbol{\theta}}_n)$. Using the approximation above, we obtain

$$\hat{\text{se}}(g(\hat{\boldsymbol{\theta}}_n)) = |g'(\hat{\boldsymbol{\theta}}_n)|\hat{\text{se}}(\hat{\boldsymbol{\theta}}_n).$$

Example: (*Standard error of Binomial variance MLE*)Suppose $X \sim \text{Binomial}(n, p)$. The MLE is $\hat{\theta}_n = \hat{p} = \frac{x}{n}$. Since $I_n(p) = \frac{n}{p(1-p)}$, $\hat{\text{se}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Our goal here is to compute $\hat{\text{se}}(\hat{p}(1-\hat{p}))$. Let $g(p) = p(1-p)$. Then $g'(p) = 1 - 2p$. According to the formula above, we obtain

$$\hat{\text{se}}(\hat{p}(1-\hat{p})) = |1 - 2\hat{p}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

4 Exponential Family Distributions (EFDs)

Why exponential family distributions? Exponential family distributions (EFDs) provide a generalized parameterization and form of a very large class of distributions used in inference. For example, Binomial, Poisson, Exponential, Normal, Multinomial, MVN, and Dirichlet are all EFDs. The generalized form provides generally applicable formulas for moments, estimators, etc. EFDs also facilitate developing general algorithms for model fitting. (*FAS*)

4.1 Natural single parameter EFD

A simple case of EFD is natural single parameter EFD such that

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

where η is a paramter, $h(.)$ is a function only with respect to x, and $A(.)$ is a "log-normalizer" to make sure pmf/pdf integrated to 1.

Example: (*Binomial exponential family*) Suppose $X \sim \text{Bernoulli}(p)$. We can write the pmf of X into the above form:

$$\begin{aligned} f(x;p) &= p^x(1-p)^{1-x} \\ &= \exp\{x\log p + (1-x)\log(1-p)\} \\ &= \exp\{\log(\frac{p}{1-p})x + \log(1-p)\}, \end{aligned}$$

where we let

$$\begin{aligned} \eta(p) &= \log(\frac{p}{1-p}) \\ A(\eta) &= \log(1+e^{\eta}). \end{aligned}$$

Then we obtain

$$f(x;p) = \exp\{\eta x - A(\eta)\}.$$

4.2 Generalized EFD

Here is a general definition for EFD. If a random varaible X follows an exponential family distribution parametrized on the observed scale by $\underline{\boldsymbol{\theta}}$, then its pdf or pmf has the following form:

$$f(x;\underline{\boldsymbol{\theta}}) = h(x)\exp\{\sum_{k=1}^{d}\eta_k(\underline{\boldsymbol{\theta}})T_k(x) - A(\boldsymbol{\eta})\},$$

where

$$\boldsymbol{\eta} = \begin{pmatrix}\eta_1(\underline{\boldsymbol{\theta}})\\ \eta_2(\underline{\boldsymbol{\theta}})\\ \vdots \\ \eta_d(\underline{\boldsymbol{\theta}})\end{pmatrix},$$

and $T_1(x), T_2(x), \ldots, T_d(x)$ are sufficient statistics for $\eta_1, \eta_2, \ldots, \eta_d$. Note that functions $\eta_k(\underline{\boldsymbol{\theta}})$ for $k = 1, 2, \ldots, d$ map the usual (observed) to the "natural parameter". $A(\boldsymbol{\eta})$ is sometimes called "log-normalizer":

$$A(\boldsymbol{\eta}) = \log\int h(x)\exp\{\sum_{k=1}^{d}\eta_k(\underline{\boldsymbol{\theta}})T_k(x)\}dx.$$

Example: (*Normal exponential family*) Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. We can write its pdf in an EFD form.

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \log(\sigma) - \frac{\mu^2}{2\sigma^2}\}. \end{aligned}$$

Then we have

$$\boldsymbol{\eta}(\mu,\sigma^2) = \begin{pmatrix}\eta_1(\mu,\sigma^2)\\ \eta_2(\mu,\sigma^2)\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix}$$

$$\boldsymbol{T}(x) = \begin{pmatrix}T_1(x)\\ T_2(x)\end{pmatrix} = \begin{pmatrix}x\\ x^2\end{pmatrix}$$

$$A(\boldsymbol{\eta}) = \log(\sigma) + \frac{\mu^2}{2\sigma^2} = -\frac{1}{2}\log(-2\eta_2) - \frac{\eta_1^2}{4\eta_2}.$$

4.3 Calculating moments

The first and second moment of $T_k(x)$ are

$$\begin{aligned} \frac{\partial}{\partial\eta_k}A(\boldsymbol{\eta}) &= \mathbb{E}[T_k(x)] \\ \frac{\partial^2}{\partial\eta_k^2}A(\boldsymbol{\eta}) &= \mathrm{Var}[T_k(x)]. \end{aligned}$$

Example from FAS: (*Calculate moments with Normal EFD*) Given that $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[X] = \frac{\partial}{\partial \eta_1} A(\boldsymbol{\eta}) = -\frac{\eta_1}{2\eta_2} = \mu,$$
$$\mathrm{Var}(X) = \frac{\partial^2}{\partial \eta_1^2} A(\boldsymbol{\eta}) = -\frac{1}{2\eta_2} = \sigma^2.$$

4.4 Maximum likelihood of an EFD

Suppose $x_1, x_2, \ldots, x_n$ are iid from an EFD. The log-likelihood is

$$l(\underline{\boldsymbol{\eta}}; \underline{\boldsymbol{x}}) = \sum_{i=1}^{n} [\log(h(x_i)) + \sum_{k=1}^{d} \eta_k(\underline{\boldsymbol{\theta}}) T_k(x_i) - A(\underline{\boldsymbol{\eta}})].$$

We take the derivative of log-likelihood with respect to η_k and set it to zero.

$$\frac{\partial}{\partial \eta_k} l(\underline{\boldsymbol{\eta}}; \underline{\boldsymbol{x}}) = \sum_{i=1}^{n} T_k(x_i) - n \frac{\partial}{\partial \eta_k} A(\underline{\boldsymbol{\eta}}) = 0.$$

Thus MLE of η_k is the solution to

$$\frac{1}{n} \sum_{i=1}^{n} T_k(x_i) = \frac{\partial}{\partial \eta_k} A(\underline{\boldsymbol{\eta}}) = \mathbb{E}[T_k(x)].$$

5 Frequentist Inference

5.1 Statistical inference

We have observed data that is modeled by a probablity generation process. The probability distribution has parameters informative about the population. Statistical inference reverse engineers this forward process to estimate parameters and provide measures of uncertainty about the estimates such as:

- parameters: a number that describes a population. A parameter is often a fixed number and we usually do not know its value.
- statistic: a number calculated from a sample of data. A statistic is used to estimate a parameter.
- sampling distribution: is the probability disribution of the statistic under repeated realizations of the data from the assumed data generating probability distribution. Note that sampling distribution is how we connect our calculated statistic to the population (i.e. probability model).

5.2 Inference goals and strategies

Given data $x_1, x_2, \ldots, x_n$ and model $X_1, X_2, \ldots, X_n \sim F_\theta$. We have the following three goals:

- point estimate of θ;
- confidence interval of θ, which is uncertainty of point estimate;
- hypothesis test to assess specific value(s) of θ.

5.2.1 Point estimation of θ

Example: MLE is a point estimation of θ.

Example: Method of moments estimator is a point estimation of θ.

5.2.2 Confidence interval of θ

Confidence interval has the form

$$(\hat{\theta} - C_l, \hat{\theta} + C_u),$$

where $C_l, C_u > 0$. Here $Pr(\hat{\theta} - C_l \leq \theta \leq \hat{\theta} + C_u; \theta)$ forms the "level" or coverage probability. Note that $\hat{\theta}, C_l, C_u$ are random variables. **Interpretation**: If we repeat the study many times, then the CI $(\hat{\theta} - C_l, \hat{\theta} + C_u)$ will contain the true θ with a long run frequency equal to $Pr(\hat{\theta} - C_l \leq \theta \leq \hat{\theta} + C_u; \theta)$.

Let's approximate a 95% confidence interval for MLEs.

$$\begin{aligned}
0.95 &\approx Pr(-1.96 \leq \frac{\hat{\theta} - \theta}{\hat{\text{se}}}(\hat{\theta}) \leq 1.96) \\
&= Pr(-1.96 \leq \frac{\theta - \hat{\theta}}{\hat{\text{se}}}(\hat{\theta}) \leq 1.96) \\
&= Pr(-1.96 \cdot \hat{\text{se}}(\hat{\theta}) \leq \theta - \hat{\theta} \leq 1.96 \cdot \hat{\text{se}}(\hat{\theta})) \\
&= Pr(\hat{\theta} - 1.96 \cdot \hat{\text{se}}(\hat{\theta}) \leq \theta \leq \hat{\theta} + 1.96 \cdot \hat{\text{se}}(\hat{\theta})).
\end{aligned}$$

Note that a $(1-\alpha)100\%$ approximate confidence interval is

$$(\hat{\theta} - |Z_{\frac{\alpha}{2}}|\hat{\text{se}}(\hat{\theta}), \hat{\theta} + |Z_{\frac{\alpha}{2}}|\hat{\text{se}}(\hat{\theta})),$$

where Z_α is the $\alpha-$percentile of $\mathcal{N}(0,1)$.

Example from FAS: (*Simulation*)

```
> mu <- 5
> n <- 20
> x <- replicate(10000, rnorm(n=n, mean=mu)) # 10000 studies
> m <- apply(x, 2, mean) # the estimate for each study
> ci <- cbind(m - 1.96/sqrt(n), m + 1.96/sqrt(n))
> head(ci)
         [,1]     [,2]
[1,] 4.354280 5.230818
[2,] 4.480737 5.357276
[3,] 4.543961 5.420500
[4,] 5.027599 5.904138
[5,] 4.481520 5.358059
[6,] 4.850693 5.727232
> cover <- (mu > ci[,1]) & (mu < ci[,2])
> mean(cover)
[1] 0.9496
```

Example: (*Sample survey*) A sample survey should satisfy the following inequality to control its confidecne interval

$$|Z_{\frac{\alpha}{2}}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq |Z_{\frac{\alpha}{2}}|\sqrt{\frac{0.5^2}{n}},$$

we can solve for n to determine how large we need to sample.

6 References

- Casella, G., and Berger, R. L. (2002). *Statistical inference*. Duxbury Press.
- Storey J. (2020). *Foundations of Applied Statistics*. https://jdstorey.org/fas/

7 Session Information

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Mojave 10.14.5

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gapminder_0.3.0 forcats_0.4.0   stringr_1.4.0   dplyr_0.8.4
 [5] purrr_0.3.3     readr_1.3.1     tidyr_1.0.2     tibble_2.1.3
 [9] ggplot2_3.2.1   tidyverse_1.3.0 knitr_1.27

loaded via a namespace (and not attached):
 [1] tidyselect_1.0.0 xfun_0.12        haven_2.2.0      lattice_0.20-38
 [5] colorspace_1.4-1 vctrs_0.2.2      generics_0.0.2   htmltools_0.4.0
 [9] yaml_2.2.1       rlang_0.4.4      pillar_1.4.3     withr_2.1.2
[13] glue_1.3.1       DBI_1.1.0        dbplyr_1.4.2     modelr_0.1.5
[17] readxl_1.3.1     lifecycle_0.1.0  munsell_0.5.0    gtable_0.3.0
[21] cellranger_1.1.0 rvest_0.3.5      evaluate_0.14    labeling_0.3
[25] fansi_0.4.1      broom_0.5.4      Rcpp_1.0.3       scales_1.1.0
[29] backports_1.1.5  jsonlite_1.6.1   farver_2.0.3     fs_1.3.1
[33] hms_0.5.3        digest_0.6.23    stringi_1.4.5    grid_3.6.0
[37] cli_2.0.1        tools_3.6.0      magrittr_1.5     lazyeval_0.2.2
[41] crayon_1.3.4     pkgconfig_2.0.3  xml2_1.2.2       reprex_0.3.0
[45] lubridate_1.7.4  assertthat_0.2.1 rmarkdown_2.1    httr_1.4.1
[49] rstudioapi_0.10  R6_2.4.1         nlme_3.1-143     compiler_3.6.0
```