Theory of Statistical Practice

Derek Li

Contents

1 Review

1.1 Basic Definition

Definition 1.1. ***Random experiment*** is a mechanism producing an outcome (result) perceived as random or uncertain.

Definition 1.2. ***Sample space*** is a set of all possible outcomes of the experiment:

$$\mathcal{S} = \{\omega_1, \omega_2, \cdots\}.$$

Example 1.1. Waiting time until the next bus arrives: $\mathcal{S} = \{t : t \geqslant 0\}$.

1.2 Probability Function/Measure

Definition 1.3. Given a sample space $\mathcal{S}$, define $\mathcal{A}$ to be a collection of subsets (events) of $\mathcal{S}$ satisfying the following conditions:

1. $\mathcal{S} \in \mathcal{A}$;
2. $A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$;
3. $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow A_1 \cup A_2 \cup \cdots \in \mathcal{A}$.

If $\mathcal{S}$ is finite or countably infinite, then $\mathcal{A}$ could consist of all subsets of $\mathcal{S}$ including $\varnothing$.

Definition 1.4. The ***probability function*** (***measure***) P on $\mathcal{A}$ satisfies the following conditions:

1. $P(A) \geqslant 0, \forall A \in \mathcal{A}$;
2. $P(\varnothing) = 0$ and $P(\mathcal{S}) = 1$;
3. If $A_1, A_2, \cdots$ are disjoint (mutually exclusive) events, i.e., $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Property 1.1. $P(A^C) = 1 - P(A)$.

Proof. $1 = P(\mathcal{S}) = P(A \cup A^C) = P(A) + P(A^C)$. □

Property 1.2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. $P(A) = P(A \cap B) + P(A \cap B^C)$ and $P(A \cup B) = P(B) + P(A \cap B^C)$. □

Property 1.3. $P(A \cup B) \leqslant P(A) + P(B)$.

Property 1.4. In general,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots - (-1)^n P(A_1 \cap \cdots \cap A_n).$$

Property 1.5 (Bonferroni's Inequality)**.** In general,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i).$$

1.3 Conditional Probability

Definition 1.5. The probability of A ***conditional*** on B is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

if $P(B) > 0$. Note that if $P(B) = 0$, we can still define $P(A|B)$ but we need to be more careful mathematically.

Theorem 1.1 (Bayes Theorem)**.** If $B_1, \cdots, B_k$ are disjoint events with $B_1 \cup \cdots \cup B_k = \mathcal{S}$, then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

1.4 Independence

Definition 1.6. Two events A and B are ***independent*** if

$$P(A \cap B) = P(A)P(B).$$

When $P(A), P(B) > 0$, we can also say

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

Events $A_1, \cdots, A_k$ are independent if

$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i).$$

1.5 Interpretation of Probability

- Long-Run frequencies: If we repeat the experiment many times, then $P(A)$ is the proportion of times the event A occurs.
- Degrees of belief (subjective probability): If $P(A) > P(B)$, then we believe that A is more likely to occur than B.
- Frequentist versus Bayesian statistical methods:
    - Frequentists: Pretend that an experiment is at least conceptually repeatable.
    - Bayesians: Use subjective probability to describe uncertainty in parameters and data.

1.6 Random Variable

Definition 1.7. ***Random variable*** is a real-valued function defined on a sample space $\mathcal{S}, X : \mathcal{S} \to \mathbb{R}$. In other words, for each outcome $\omega \in \mathcal{S}, X(\omega)$ is a real number.

Definition 1.8. The ***probability distribution*** of X depends on the probabilities assigned to the outcomes in $\mathcal{S}$.

Definition 1.9. The ***cumulative distribution function*** (CDF) of X is

$$F(x) = P(X \leqslant x) = P(\omega \in \mathcal{S} : X(\omega) \leqslant x).$$

We denote it $X \sim F$.

Property 1.6. CDF satisfies:

1. If $x_1 \leqslant x_2$, then $F(x_1) \leqslant F(x_2)$;
2. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$;
3. F is right-continuous with left-hand limits:

$$\lim_{y \to x^+} F(y) = F(x), \lim_{y \to x^-} F(y) = F(x-) = P(X < x);$$

4. $P(X = x) = F(x) - F(x-)$.

Definition 1.10. If $X \sim F$ where F is a continuous function, then X is a ***continuous r.v.***, and we can typically find a non-negative ***probability density function*** (PDF) f s.t.

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

Definition 1.11. If X takes only a finite or countably infinite number of possible values, then X is a ***discrete r.v.***, and F is a step function. We can define its ***probability mass function*** (PMF) by

$$f(x) = F(x) - F(x-) = P(X = x).$$

1.7 Expected Value

Definition 1.12. Suppose X with PDF $f(x)$ and Y with PMF $f(y)$. We can define the ***expected value*** of $h(X)$ and $h(Y)$ by

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x \text{ and } \mathbb{E}[h(Y)] = \sum_{y} h(y)f(y).$$

We can also write $h(x) = h^+(x) - h^-(x)$ where $h^+(x) = \max\{h(x), 0\}$ and $h^-(x) = \max\{-h(x), 0\}$, then $\mathbb{E}[h(X)] = \mathbb{E}[h^+(X)] - \mathbb{E}[h^-(X)]$:

1. If $\mathbb{E}[h^+(X)]$ and $\mathbb{E}[h^-(X)]$ are finite, then $\mathbb{E}[h(X)]$ is well defined.
2. If $\mathbb{E}[h^+(X)] = \infty$ and $\mathbb{E}[h^-(X)]$ is finite, then $\mathbb{E}[h(X)] = \infty$.
3. If $\mathbb{E}[h^+(X)]$ is finite and $\mathbb{E}[h^-(X)] = \infty$, then $\mathbb{E}[h(X)] = -\infty$.
4. If $\mathbb{E}[h^+(X)]$ and $\mathbb{E}[h^-(X)]$ are infinite, then $\mathbb{E}[h(X)]$ does not exist.

Example 1.2 (Expected Values of Cauchy Distribution)**.** X is a continuous r.v. with

$$f(x) = \frac{1}{\pi(1 + x^2)}, -\infty < x < \infty.$$

We have

$$\mathbb{E}[X^+] = \mathbb{E}[X^-] = \int_0^{\infty} \frac{x}{\pi(1 + x^2)}\mathrm{d}x = \lim_{x \to \infty} \frac{1}{2\pi} \ln(1 + x^2) = +\infty.$$

Thus, $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ does not exist.

1.8 Independent Random Variable

Definition 1.13. R.v.s. $X_1, X_2, \cdots$ are ***independent*** if the events $[X_1 \in A_1], [X_2 \in A_2], \cdots$ are independent events for any $A_1, A_2, \cdots$.

If $X_1, \cdots, X_n$ are independent r.v.s. with PDF or PMF $f_1, \cdots, f_n$, then the joint PDF or PMF of $(X_1, \cdots, X_n)$ is

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

Suppose $X_1, \cdots, X_n$ are independent r.v.s. with mean $\mu_1, \cdots, \mu_n$ and variance $\sigma_1^2, \cdots, \sigma_n^2$. Define $S = X_1 + \cdots + X_n$, then $\mathbb{E}[S] = \mu_1 + \cdots + \mu_n$ (which is true even if $X_1, \cdots, X_n$ are not independent) and $\mathrm{Var}[S] = \sigma_1^2 + \cdots + \sigma_n^2$.

1.9 Convergence of Random Variable

Theorem 1.2 (Markov's Inequality)**.** Suppose Y is a random variable with $\mathbb{E}[|Y|^r] < \infty$ for some $r > 0$, then

$$P(|Y| > \varepsilon) \leqslant \frac{\mathbb{E}[|Y|^r]}{\varepsilon^r}.$$

Proof. For any $\varepsilon > 0$,

$$\mathbb{E}[|Y|^r] = \mathbb{E}[|Y|^r I(|Y| \leqslant \varepsilon)] + \mathbb{E}[|Y|^r I(|Y| > \varepsilon)] \geqslant 0 + \varepsilon^r P(|Y| > \varepsilon),$$

then $P(|Y| > \varepsilon) \leqslant \frac{\mathbb{E}[|Y|^r]}{\varepsilon^r}$. □

Theorem 1.3 (Chebyshev's Inequality)**.**

$$P(|X - \mathbb{E}[X]| > \varepsilon) \leqslant \frac{\mathrm{Var}[X]}{\varepsilon^2}.$$

Proof. Take $r = 2, Y = X - \mathbb{E}[X]$ in Markov's Inequality. □

1.9.1 Convergence in Probability

Definition 1.14. A sequence of r.v.s. $\{Y_n\}$ ***converges in probability*** to a r.v. Y (denoted $Y_n \xrightarrow{p} Y$) if for each $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|Y_n - Y| > \varepsilon) = 0.$$

Typically, the limiting r.v. Y is a constant.

Theorem 1.4 (Weak Law of Large Numbers)**.** If $X_1, X_2, \cdots$ are independent r.v.s. with finite mean μ, then

$$\overline{X} = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

1.9.2 Convergence in Distribution

Definition 1.15. A sequence of r.v.s. $\{X_n\}$ ***converges in distribution*** to a r.v. X if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every $x \in \mathbb{R}$ at which F is continuous. F_n and F are CDF of X_n and X, respectively.

Let $S_n = \sqrt{n}(\overline{X}_n - \mu)$, then we have $\mathbb{E}[S_n] = 0$ and $\mathrm{Var}[S_n] = \sigma^2$. $\{S_n\}$ is bounded in probability since

$$P(|S_n| > M) \leqslant \frac{\mathbb{E}[S_n^2]}{\varepsilon^2} = \frac{\sigma^2}{M^2} \to 0 \text{ as } M \to \infty.$$

Theorem 1.5 (Basic Central Limit Theorem)**.** If $X_1, X_2, \cdots$ are independent r.v.s. with common CDF F with finite mean and variance μ and σ^2, then

$$\lim_{n\to\infty} P(S_n \leqslant x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t,$$

denoted as $S_n \xrightarrow{d} S \sim \mathcal{N}(0, \sigma^2)$.

As a consequence, the distribution of $\overline{X}_n$ is approximately $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, when n is sufficiently large, denoted as $\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

We can approximate $g(\overline{X}_n)$ by Taylor's Formula. We have

$$g(\overline{X}_n) = g(\mu) + g'(\mu)(\overline{X}_n - \mu) + o(\overline{X}_n - \mu)$$

and thus

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(\mu)\sqrt{n}(\overline{X}_n - \mu) + \sqrt{n}o(\overline{X}_n - \mu)$$

with $o(\overline{X}_n - \mu) \to 0$, suggesting that

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, [g'(\mu)]^2\sigma^2).$$

Theorem 1.6 (General Central Limit Theorem)**.** Suppose $X_1, X_2, \cdots$ are independent with $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}[X_i] = \sigma_i^2$ and let

$$S_n = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right),$$

then

$$S_n \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)$$

provided that $\sum_{i=1}^{n}\sigma_i^2$ is not dominated by a small number of terms and the tails of the distributions of $\{X_i\}$ are not too dissimilar.

1.9.3 Quality of Normal Approximation

Definition 1.16. Define ***skewness*** of X_i as

$$\mathrm{Skew}(X_i) = \frac{\mathbb{E}[(X_i - \mu)^3]}{\sigma^3}.$$

Definition 1.17. Define ***kurtosis*** of X_i as

$$\mathrm{Kurt}(X_i) = \frac{\mathbb{E}[(X_i - \mu)^4]}{\sigma^4}.$$

Let $S_n = \sqrt{n}(\overline{X}_n - \mu)$, where $\overline{X}_n$ is the sample mean of independent $X_1, \cdots, X_n$ with CDF F with mean μ and variance σ^2. The normal approximation works better for fixed n if $\mathrm{Skew}(X_i)$ and $\mathrm{Kurt}(X_i)$ are close to the values for normal distribution (0 and 3, respectively).

1.9.4 Distribution Approximation

Theorem 1.7 (Slutsky's Theorem)**.** Suppose $X_n \xrightarrow{d} X \sim G$ and $Y_n \xrightarrow{p} \theta$, where θ is a constant, then as $n \to \infty, \psi(X_n, Y_n) \xrightarrow{d} \psi(X, \theta)$, where ψ is continuous.

Theorem 1.8 (Delta Method)**.** Suppose $a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n \uparrow \infty$. If $g(x)$ is differentiable at $x = \theta$, then

$$a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z.$$

Proof. By Taylor's Formula,

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + \Delta(X_n)(X_n - \theta),$$

where $\Delta(X_n) \xrightarrow{p} 0$. Therefore,

$$a_n(g(X_n) - g(\theta)) = g'(\theta)a_n(X_n - \theta) + \Delta(X_n)a_n(X_n - \theta) \xrightarrow{d} g'(\theta)Z + 0 = g'(\theta)Z.$$

□

We can use the Delta Method to estimate standard errors of parameter estimators, and extend the Delta Method to functions of several sample means: $g(\overline{X}_n, \overline{Y}_n, \overline{Z}_n, \cdots)$.

Another application with the Delta Method is the variance stabilizing transformations for some distributions with $\text{Var}[X_i] = \phi(\mu)$.

Example 1.3. For the Poisson distribution, $\text{Var}[X_i] = \phi(\mu) = \mu$, then by CLT,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \mu).$$

We find g s.t.

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, 1),$$

i.e.,

$$[g'(\mu)]^2\mu = 1 \Rightarrow g'(\mu) = \frac{1}{\sqrt{\mu}}.$$

Thus, $g(x) = 2\sqrt{x} + C$ and

$$2\sqrt{\overline{X}_n} \dot{\sim} \mathcal{N}\left(2\sqrt{\mu}, \frac{1}{n}\right).$$

1.10 Mathematics

Theorem 1.9 (Jensen's Inequality)**.** If g is a strictly concave function ($g''(x) < 0$) then $\mathbb{E}[g(Y)] \leqslant g(\mathbb{E}[Y])$ with equality iff Y is constant.

2 Statistical Models

2.1 Probability versus Statistics

Suppose $X_1, \cdots, X_n$ are independent r.v.s., with some CDF F. For probability, F is known and we can calculate probabilities involving the r.v.s. $X_1, \cdots X_n$. Knowledge of the population F gives information about the nature of samples from the population. For statistics, F is unknown and we observe outcomes of $X_1, \cdots X_n : x_1, \cdots, x_n$ (data).

Definition 2.1. ***Statistical inference*** uses the information in the data to estimate of infer properties of the unknown F.

Definition 2.2. Assume that the data $x_1, \cdots, x_n$ are outcomes of r.v.s. $X_1, \cdots, X_n$ whose joint distribution is F (which is assumed to be unknown to some degree). A ***statistical model*** is a family $\mathcal{F}$ of probability distributions of $(X_1, \cdots, X_n)$.

We assume that true distribution $F \in \mathcal{F}$ but in practice, $\mathcal{F}$ typically represents only an approximation to the truth, i.e., $F \notin \mathcal{F}$ but F is close to some $F_0 \in \mathcal{F}$.

Definition 2.3. $\mathcal{F}$ is called a ***parametric model*** if

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^p,$$

where θ is the parameter and Θ is the parameter space. We can write $\theta = (\theta_1, \cdots, \theta_p)$.

Definition 2.4. The model is said to be ***non-parametric*** if the parameter space Θ is not finite dimensional.

In practice, we often approximate the infinite dimensional parameter by a finite dimensional parameter. For example, we assume $g(x) \approx \sum_{k=1}^{p} \beta_k \phi_k(x)$ for some functions ϕ_k's and unknown parameters β_k's.

Definition 2.5. The model is said to be ***semi-parametric*** if non-parametric model has a finite dimensional parametric component.

2.2 Bayesian Models

Assume we have a parametric model with parameter space $\Theta \subset \mathbb{R}^p$. For each $\theta \in \Theta$, the joint CDF F_θ is the conditional distribution of $(X_1, \cdots, X_n)$ given θ. ***Bayesian inference*** is the process that we put a probability distribution on Θ - ***prior distribution***, and then after observing $X_1 = x_1, \cdots, X_n = x_n$, we can use Bayes Theorem to obtain a ***posterior distribution*** of θ.

Note that we can take Bayesian inference for non-parametric models.

2.3 Empirical Distribution Function (EDF)

We are often interested in estimating characteristics of $F, \theta(F)$, called statistical functionals (mathematically, $\theta : \mathcal{F} \to \mathbb{R}$).

2.3.1 Substitution Principle

Given $X_1, \cdots, X_n$ from F, we first estimate F by $\widehat{F}$ and substitute $\widehat{F}$ for F into $\theta(F)$:

$$\widehat{\theta}(F) = \theta(\widehat{F}).$$

If $\theta(\cdot)$ is continuous and $\widehat{F} \approx F$, then $\theta(\widehat{F}) \approx \theta(F)$. A simple estimator of F is the ***empirical distribution function*** (EDF)

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x).$$

Note that $\widehat{F}$ is a discrete CDF putting mass $\frac{1}{n}$ at $X_1, \cdots, X_n$. In R, we can plot by `plot(ecdf(x))`.

2.3.2 Properties of EDF

Note that the EDF is a sample mean for each x and so the WLLN and CLT hold as $n \to \infty$, and $I(X_1 \leqslant x), \cdots, I(X_n \leqslant x)$ are independent Bernoulli random variables.

Property 2.1. EDF satisfies:

1. $\mathbb{E}[\widehat{F}(x)] = F(x), \mathrm{Var}[\widehat{F}(x)] = \frac{F(x)(1-F(x))}{n}$;
2. (WLLN)

$$\widehat{F}(x) = \widehat{F}_n(x) \xrightarrow{p} F(x) \text{ for each } x \Leftrightarrow \sup_{-\infty<x<\infty} |\widehat{F}(x) - F(x)| \xrightarrow{p} 0;$$

3. (CLT) $\sqrt{n}(\widehat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x))$.

2.3.3 Example and Counterexample

If $\theta(F)$ is based on expected values, the substitution principle using the EDF works well.

Example 2.1. $\theta(F) = \mathbb{E}_F[X_i]$:

$$\widehat{\theta}(F) = \theta(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Example 2.2. $\theta(F) = \mathbb{E}_F[h(X_i)]$:

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

Example 2.3 (Theil Index)**.** $\theta(F) = \mathbb{E}_F\left[\frac{X_i}{\mu(F)} \ln\left(\frac{X_i}{\mu(F)}\right)\right]$:

$$\widehat{\theta}(F) = \theta(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{\overline{X}} \ln\left(\frac{X_i}{\overline{X}}\right).$$

Example 2.4. Suppose F is continuous and $\theta(F) = f(x) = F'(x)$, and application of the substitution principle with EDF gives

$$\widehat{f}(x) = \widehat{F}'(x) = \begin{cases} 0, & x \neq X_i, \forall i \\ \text{undefined}, & x = X_i \end{cases}.$$

The substitution principle fails.

2.4 Order Statistics

Suppose $X_1, \cdots, X_n$ are independent with unknown CDF F, we order $X_1, \cdots, X_n$ from smallest to largest:

$$X_{(1)} \leqslant \cdots \leqslant X_{(n)}.$$

Due to the independence assumption, the order statistics carry the same information about F as the unordered data.

Order statistics can be used to estimate the quantiles $F^{-1}(\tau)$ of F.

Example 2.5 (Sample Median)**.**

$$M = \begin{cases} \frac{X_{(n/2)}+X_{(n/2+1)}}{2}, & n \text{ is even} \\ X_{((n+1)/2)}, & n \text{ is odd} \end{cases}.$$

M is an estimator of $F^{-1}(\frac{1}{2})$.

Likewise, we can estimate $F^{-1}(\tau)$ by $X_{(k)}$ where $k \approx \tau n$.

2.4.1 Distribution of $X_{(k)}$

Now suppose $X_1, \cdots, X_n$ are independent with continuous CDF F and PDF f.

We first find CDF of sample minimum and maximum. For the minimum,

$$P(X_{(1)} > x) = P(X_1 > x, \cdots, X_n > x) = [1 - F(x)]^n$$

and thus

$$P(X_{(1)} \leqslant x) = 1 - [1 - F(x)]^n.$$

For the maximum,

$$P(X_{(n)} \leqslant x) = P(X_1 \leqslant x, \cdots, X_n \leqslant x) = F(x)^n.$$

Therefore, the densities of $X_{(1)}$ and $X_{(n)}$ are

$$g_1(x) = n[1 - F(x)]^{n-1} f(x), g_n(x) = nF(x)^{n-1} f(x).$$

Now define

$$Z(x) = \sum_{i=1}^{n} I(X_i \leqslant x) \sim \text{Bin}(n, F(x)),$$

and note that $[X_{(k)} \leqslant x] = [Z(x) \geqslant k]$. Thus,

$$P(X_{(k)} \leqslant x) = P(Z(x) \geqslant k) = \sum_{i=k}^{n} \binom{n}{i} F(x)^i (1 - F(x))^{n-i}.$$

Wherefore, the PDF of $X_{(k)}$ is

$$g_k(x) = \frac{\mathrm{d}}{\mathrm{d}x} \sum_{i=k}^{n} \binom{n}{i} F(x)^i [1 - F(x)]^{n-i} = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} [1 - F(x)]^{n-k} f(x).$$

2.4.2 Convergence in Distribution of Central Order Statistics

Definition 2.6. Suppose $k = k_n \approx \tau n$ for some $\tau \in (0,1)$ (but not too close to 0 or 1). $X_{(k)}$ is called a ***central order statistic***.

For example, if $k \approx \frac{1}{2}$, then $X_{(k)}$ is the sample median. **Intuitively, $X_{(k)}, k \approx \tau n$ is an estimator of the τ-quantile $F^{-1}(\tau)$.**

Theorem 2.1. If $\{k_n\}$ is a sequence of integers with $\sqrt{n}\left(\frac{k_n}{n} - \tau\right) \to 0$ for some $\tau \in (0,1)$ and $f(F^{-1}(\tau)) > 0$, then

$$\sqrt{n}(X_{(k_n)} - F^{-1}(\tau)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{[f(F^{-1}(\tau))]^2}\right).$$

Proof. Suppose $U_1, \cdots, U_n$ are independent Unif$(0,1)$ r.v.s. and $U_{(1)} \leqslant \cdots \leqslant U_{(n)}$ are the corresponding order statistics.

Take $E_1, \cdots, E_{n+1}$ to be independent exponential r.v.s. with mean 1, then

$$(U_{(1)}, \cdots, U_{(n)}) \overset{d}{=} \left(\frac{E_1}{S}, \cdots, \frac{E_1 + \cdots + E_n}{S}\right),$$

where $S = E_1 + \cdots + E_{n+1}$. Note that

$$\frac{S}{n} = \underbrace{\frac{E_1 + \cdots + E_{n+1}}{n+1}}_{\xrightarrow{p} 1 \text{ (WLLN)}} \cdot \frac{n+1}{n} \xrightarrow{p} 1.$$

Therefore, we can approximate the distribution of $U_{(k)}$ by a distribution of a sum of exponential r.v.s.:

$$U_{(k)} = \frac{(E_1 + \cdots + E_k)/n}{(E_1 + \cdots + E_{n+1})/n} \approx \frac{1}{n}(E_1 + \cdots + E_k).$$

Assume $\sqrt{n}\left(\frac{k_n}{n} - \tau\right) \to 0$, then

$$\sqrt{n}(U_{(k_n)} - \tau) \overset{d}{=} \sqrt{n}\left(\frac{E_1 + \cdots + E_{k_n} - \tau S}{S}\right).$$

Now we need to show:

$$\sqrt{n}\left(\frac{1}{n}(E_1 + \cdots + E_{k_n} - \tau S)\right) \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)).$$

Let

$$E_1 + \cdots + E_{k_n} - \tau S = \sum_{i=1}^{n+1} a_i E_i,$$

where $a_i = 1 - \tau$ for $i = 1, \cdots, k_n$ and $a_i = -\tau$ for $i = k_n + 1, \cdots, n+1$.

We have

$$\mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n+1} a_i E_i\right] = \frac{1}{\sqrt{n}}(k_n(1-\tau) - (n - k_n + 1)\tau) \to 0$$

and

$$\mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n+1} a_i E_i\right] = \frac{1}{n}(k_n(1-\tau)^2 + (n - k_n + 1)\tau^2) \to \tau(1-\tau).$$

Recall that if $U \sim \text{Unif}(0,1)$ and F is a continuous CDF, $0 < F(x) < 1$ with PDF f, $f(x) > 0$ for all x, then $X = F^{-1}(U) \sim F$.

Thus if $U_{(1)} \leqslant \cdots \leqslant U_{(n)}$, then $F^{-1}(U_{(1)}) \leqslant \cdots \leqslant F^{-1}(U_{(n)})$ are order statistics from F. In other words,

$$\sqrt{n}(X_{(k_n)} - F^{-1}(\tau)) \stackrel{d}{=} \sqrt{n}(F^{-1}(U_{(k_n)}) - F^{-1}(\tau))$$

and we can use the Delta method with $g(\tau) = F^{-1}(\tau)$.

Note that $F(F^{-1}(\tau)) = \tau$ and

$$\frac{\mathrm{d}}{\mathrm{d}\tau} F(F^{-1}(\tau)) = \frac{\mathrm{d}}{\mathrm{d}\tau}\tau \Rightarrow f(F^{-1}(\tau))\frac{\mathrm{d}}{\mathrm{d}\tau}F^{-1}(\tau) = 1,$$

and thus

$$g'(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}F^{-1}(\tau) = \frac{1}{f(F^{-1}(\tau))}.$$

Applying Delta method, we have

$$\sqrt{n}(X_{(k_n)} - F^{-1}(\tau)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{[f(F^{-1}(\tau))]^2}\right)$$

□

Example 2.6 (Sample Median versus Sample Mean from Normal Distribution)**.** Suppose $X_i, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ r.v.s., and $\mathbb{E}[X_i] = F^{-1}(\frac{1}{2}) = \mu$. Both sample mean and sample median can be used to estimate μ. However, $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$, and

$$\text{Var}[X_{(k)}] \approx \frac{\tau(1-\tau)}{n[f(F^{-1}(\tau))]^2} = \frac{\pi\sigma^2}{2n}, k \approx \frac{n}{2}.$$

Thus,

$$\frac{\text{Var}[X_{(k)}]}{\text{Var}[\overline{X}]} = \frac{\pi}{2} > 1,$$

and $\overline{X}$ is a better estimator of μ.

Example 2.7 (Sample Median versus Sample Mean from Laplace Distribution)**.** Note that

$$f(x) = \frac{1}{2}e^{-|x-\mu|}.$$

We have

$$\text{Var}[\overline{X}] = \frac{2}{n} \text{ and } \text{Var}[X_{(k)}] \approx \frac{1}{(4n[f(\mu)]^2)} = \frac{1}{n}, k \approx \frac{n}{2}.$$

Therefore, $X_{(k)}$ is a better estimator of μ.

2.5 Plots

2.5.1 Boxplot

Boxplot is the simplified graphical representation of the data. R function is `boxplot`.

2.5.2 Quantile-Quantile Plot

Quantile-quantile plot is the graphical tool to check if data come from a particular location or scale family of distributions.

Suppose we check whether data $x_1, \cdots, x_n$ are well-modeled by some specified F_0, we plot ordered value $x_{(k)}$ against $F_0^{-1}(\tau_k)$ for $k = 1, \cdots, n$, where $\tau_k = \frac{k-1/2}{n}, \frac{k}{n+1}$ or $\frac{k-3/8}{n+1/4}$. If F_0 is a good model then the points should fall close to a straight line.

Suppose we check whether data $x_1, \cdots, x_n$ are well-modeled by $F(x) = F_0\left(\frac{x-\mu}{\sigma}\right)$, where F_0 is specified but μ and σ are unknown. We have $F^{-1}(\tau) = \mu + \sigma F_0^{-1}(\tau)$ and can plot $x_{(k)}$ against $F_0^{(-1)}(\tau_k)$ for $k = 1, \cdots, n$.

By theorem, if the data come from a distribution of this form, then

$$x_{(k)} = \mu + \sigma F_0^{-1}(\tau) + \varepsilon_k, k = 1, \cdots, n,$$

where

$$\varepsilon_k \dot{\sim} \mathcal{N}\left(0, \frac{\sigma^2 \tau_k(1-\tau_k)}{n[f_0(F_0^{-1}(\tau_k))]^2}\right).$$

For each fixed τ_k, $\mathrm{Var}[\varepsilon_k] \to 0$ as n increases. Behavior of $\mathrm{Var}[\varepsilon_k]$ as $\tau_k \to 0$ or 1 is less obvious.

Note that quantile-quantile plots are most useful for lighter tailed distributions (such as the normal distribution).

Example 2.8 (Normal Quantile-Quantile Plot)**.** $F_0 = \mathcal{N}(0, 1)$ and so

$$F_0(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

R function is `qqnorm` - uses $\tau_k = \frac{k-3/8}{n+1/4}$ for $n \leqslant 10$ and $\frac{k-1/2}{n}$ for $n > 10$.

Also, we can use Shapiro-Wilk test that is based on correlation between $\{x_{(k)}\}$ and normal scores $\{\Phi^{-1}(\tau_k)\}$: if the data are normal, then the correlation should be very close to 1. R function is `shapiro.test`. If p-value is close to 0, then a normal model is not good for the data.

2.5.3 Line-Up Plot

Line-up plot compares the Q-Q plot of $x_1, \cdots, x_n$ to other Q-Q plot constructed from normally distributed data.

2.5.4 Weibull Plot

Weibull distribution is

$$f(x; \alpha, \lambda) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{x}{\lambda}\right)^{\alpha}}, x \geqslant 0,$$

where $\lambda, \alpha > 0$ (α is the Weibull modulus). The quantiles of the Weibull distribution are

$$F^{-1}(\tau; \alpha, \lambda) = \lambda(-\ln(1-\tau))^{\frac{1}{\alpha}}$$

or

$$\ln(F^{-1}(\tau; \alpha, \lambda)) = \frac{1}{\alpha}\ln(-\ln(1-\tau)) + \ln(\lambda).$$

The Weibull plot is $\ln(x_{(k)})$ against $\ln(-\ln(1-\tau_k))$ for $k = 1, \cdots, n$.

2.5.5 Histogram

A histogram is a very simple density estimator. Given data $x_1, \cdots, x_n$ and bins $B_k = [u_{k-1}, u_k)$ for $k = 1, \cdots, m$, we define for $x \in B_k$,

$$\text{Hist}(x) = \frac{1}{n(u_k - u_{k-1})} \sum_{i=1}^{n} I(x_i \in B_k).$$

Note that $\text{Hist}(x)$ is constant for x in each B_k and Hist is a density function since $\text{Hist}(x) \geqslant 0, \forall x$ and

$$\int_{-\infty}^{\infty} \text{Hist}(x)\mathrm{d}x = \sum_{k=1}^{m} \int_{B_k} \text{Hist}(x)\mathrm{d}x = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} I(x_i \in B_k)}_{=1} \underbrace{\int_{B_k} \frac{1}{u_k - u_{k-1}} \mathrm{d}x}_{=1} = 1.$$

The appearance of the histogram depends on two factors:

1. Number of bins (m);
2. Boundaries of the bins ($u_0, \cdots, u_m$).

To find optimal bin width (i.e., $u_k - u_{k-1}$), we need the knowledge of the true f. If we assume f is clos to normal then one proposal is

$$u_k - u_{k-1} = 3.49 \cdot \text{SD} \cdot n^{-\frac{1}{3}},$$

where SD is the sample standard deviation of the data.

2.6 Spacings

Definition 2.7. Given $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$, we define $(n-1)$ ***spacings*** (or ***first-order spacings***) by

$$D_k = X_{(k+1)} - X_{(k)}, k = 1, \cdots, n-1.$$

Intuitively, the spacings should carry some information about the PDF f.

2.6.1 Exponential Spacings

Suppose $X_1, \cdots, X_n$ are independent exponential r.v.s. with

$$f(x; \lambda) = \lambda e^{-\lambda x}, x \geqslant 0.$$

Given the order statistics, define

$$\begin{aligned}
Y_1 &= nX_{(1)} \\
Y_2 &= (n-1)(X_{(2)} - X_{(1)}) = (n-1)D_1 \\
Y_3 &= (n-2)(X_{(3)} - X_{(2)}) = (n-2)D_2 \\
&\vdots \\
Y_n &= X_{(n)} - X_{(n-1)} = D_{n-1}
\end{aligned}$$

Theorem 2.2. $Y_1, \cdots, Y_n$ are independent exponential r.v.s. with $f(x; \lambda)$, i.e., spacings from an exponential sample are themselves exponential and independent.

Proof. The joint PDF of $(X_{(1)}, \cdots, X_{(n)})$ is

$$f(x_1, \cdots, x_n) = n!\lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

for $0 \leqslant x_1 < \cdots < x_n$.

By definition of Y_k, we have

$$X_{(1)} = \frac{Y_1}{n}, X_{(k)} = \frac{Y_1}{n} + \cdots + \frac{Y_k}{n-k+1}$$

for $k = 2, \cdots, n$. Thus,

$$g(y_1, \cdots, y_n) = f\left(\frac{y_1}{n}, \cdots, \frac{y_1}{n} + \cdots + y_n\right) |J(y_1, \cdots, y_n)|,$$

where

$$J(y_1, \cdots, y_n) = \begin{pmatrix} \frac{1}{n} & 0 & 0 & \cdots & 0 \\ \frac{1}{n} & \frac{1}{n-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n-1} & \frac{1}{n-2} & \cdots & 1 \end{pmatrix},$$

which is a lower triangular and thus $|J(y_1, \cdots, y_n)| = \frac{1}{n!}$. Therefore,

$$g(y_1, \cdots, y_n) = \lambda^n e^{-\lambda \sum_{i=1}^n y_i}$$

for $y_1, \cdots, y_n \geqslant 0$. □

2.6.2 Spacings from Continuous F

Now we assume that $\tau \approx \frac{k}{n} \approx \frac{k+1}{n}$ and thus if τ is not too close to 0 or 1, then $X_{(k+1)} \approx X_{(k)} \approx F^{-1}(\tau)$.

Theorem 2.3. If $\frac{k_n}{n} \to \tau, \tau \in (0,1)$ and $f(F^{-1}(\tau)) > 0$, then

$$nD_{k_n} \xrightarrow{d} \exp(f(F^{-1}(\tau))),$$

i.e.,

$$P(D_{k_n} \leqslant x) \approx 1 - e^{-nf(F^{-1}(\tau))x}$$

for $x \geqslant 0$.

Proof. Recall that

$$X_{(k_n+1)} \overset{d}{=} F^{-1}(U_{(k_n+1)}) \overset{d}{=} F^{-1}\left(\frac{E_1 + \cdots + E_{k_n+1}}{E_1 + \cdots + E_{n+1}}\right)$$

and

$$X_{(k_n)} \overset{d}{=} F^{-1}\left(\frac{E_1 + \cdots + E_{k_n}}{E_1 + \cdots + E_{n+1}}\right),$$

where $E_1, \cdots, E_{n+1}$ are independent exponential r.v.s. with mean 1. Thus

$$\begin{aligned} nD_{k_n} &\overset{d}{=} n\left[F^{-1}\left(\frac{E_1 + \cdots + E_{k_n+1}}{E_1 + \cdots + E_n}\right) - F^{-1}\left(\frac{E_1 + \cdots + E_{k_n}}{E_1 + \cdots + E_{n+1}}\right)\right] \\ &\approx \frac{1}{f(F^{-1}(\tau))} \cdot \frac{nE_{k_n+1}}{E_1 + \cdots + E_{n+1}} \\ &= \frac{1}{f(F^{-1}(\tau))} \cdot \frac{E_{k_n+1}}{\frac{1}{n}(E_1 + \cdots + E_{n+1})}. \end{aligned}$$

By WLLN, we have

$$\frac{E_1 + \cdots + E_{n+1}}{n} = \frac{E_1 + \cdots + E_{n+1}}{n+1} \cdot \frac{n+1}{n} \xrightarrow{p} 1,$$

and thus

$$nD_{k_n} \approx \frac{1}{f(F^{-1}(\tau))} E_{k_n+1} \sim \exp(f(F^{-1}(\tau))).$$

□

Note that if $\frac{k_n}{n} = \tau$, then D_{k_n} is approximately exponential with mean $\frac{1}{nf(F^{-1}(\tau))}$ and variance $\frac{1}{[nf(F^{-1}(\tau))]^2}$. Spacings have a variety of applications in statistics, such as goodness-of-fit.

2.7 Density Estimation

2.7.1 Spacings Estimation

We can use the spacings to estimate the density f. Suppose $D_1, \cdots, D_{n-1}$ are independent exponential r.v.s. with

$$\mathbb{E}[nD_k] = e^{g(V_k)}, V_k = \frac{X_{(k+1)} + X_{(k)}}{2}.$$

Note that $V_k \approx F^{-1}(\tau)$ for $\tau \approx \frac{k}{n} \approx \frac{k+1}{n}$. Then the density function is

$$f(x) = e^{-g(x)}.$$

We estimate $g(x)$ by B-spline function

$$g(x) = \beta_0 + \sum_{j=1}^{p} \beta_j \psi_j(x),$$

where $\psi_j(x)$ is the B-spline, and $\beta_0, \cdots, \beta_p$ are unknown parameters. The implementation below uses R function `glm`:

```
den.splines <- function(x, p = 5){
            library(splines)
            n <- length(x)
            x <- sort(x)
            x1 <- c(NA, x)
            x2 <- c(x, NA)
            sp <- (x2 - x1)[2:n]
            mid <- 0.5 * (x1 + x2)[2:n]
            y <- n * sp
            xx <- bs(mid, df = p)
            r <- glm(y ~ xx, family = quasi(link = "log", variance = "mu^2"))
            density <- exp(-r$linear.predictors)
            r <- list(x = mid, density =density)
            r
            }
```

Note that there are several issues with this method: $\widehat{f}(x)$ does not necessarily integrate to 1 and it has problems with discretized observations where 2 or more observations may be equal (e.g., due to rounding). Better density estimation methods exist such as kernel density estimation. Also note that the spacings method is similar in spirit to nearest neighbor density estimation.

2.7.2 Example: Hazard Functions

Definition 2.8. If X is a positive continuous r.v. with CDF F and PDF f, its ***hazard function*** is

$$h(x) = \frac{f(x)}{1 - F(x)}.$$

If X is a survival time,

$$\begin{aligned}\frac{1}{\delta}P(x < X \leqslant x + \delta | X > x) &= \frac{1}{\delta} \cdot \frac{P(x < X \leqslant x + \delta)}{P(X > x)} \\ &= \frac{1}{\delta} \cdot \frac{F(x + \delta) - F(x)}{1 - F(x)} \\ &\to \frac{f(x)}{1 - F(x)} = h(x)\end{aligned}$$

as $\delta \to 0$, i.e., $h(x)$ is the instantaneous death rate given survival to time x.

Since

$$h(x) = \frac{f(x)}{1 - F(x)} = -\frac{\mathrm{d}}{\mathrm{d}x} \ln(1 - F(x)),$$

then given $h(x)$, we can define CDF and PDF by

$$F(x) = 1 - \exp\left(-\int_0^x h(t)\mathrm{d}t\right)$$

and

$$f(x) = h(x) \exp\left(-\int_0^x h(t)\mathrm{d}t\right).$$

We also assume

$$\int_0^\infty h(x)\mathrm{d}x = \infty.$$

The shape of $h(x)$ gives information not immediately apparent in f and F:

- $h(x)$ is increasing: new better than used.
- $h(x)$ is decreasing: used better than new.

Suppose $X_1, \cdots, X_n$ are independent continuous positive r.v.s. with $h(x)$. Define normalized spacings $nX_{(1)}, (n-1)(X_{(2)} - X_{(1)}), \cdots, X_{(n)} - X_{(n-1)}$. If $X_1, \cdots, X_n$ are $\exp(h(x) = C)$, where C is a constant, then these normalized spacings are exponential.

In general, if $\frac{k}{n} \approx \tau \in (0, 1)$, then

$$(n - k)(X_{(k+1)} - X_{(k)}) \approx (1 - \tau)n(X_{(k+1)} - X_{(k)}).$$

Since

$$h(F^{-1}(\tau)) = \frac{f(F^{-1}(\tau))}{1 - \tau},$$

then $(n - k)(X_{(k+1)} - X_{(k)})$ is approximately exponential with mean $\frac{1}{h(F^{-1}(\tau))}$. We can use it to estimate the hazard function using a similar approach to that used to estimate the density.

2.7.3 Kernel Density Estimation

Definition 2.9. Let a (usually symmetric) density function $w(x)$ be a ***kernel***. Given the kernel w and a ***bandwidth*** parameter h, we defined the ***kernel density estimator*** as

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right).$$

In R, the function is `density` that scales its kernels to have variance 1 and mean 0. The bandwidth parameter h controls the amount of smoothing: as h increases, the estimator becomes smoother. The choice of kernel is much less important than the choice of bandwidth.

The choice of h depends on what we believe the underlying density looks like: if we believe f is smooth then we should take larger h; if we believe f has a number of modes then we should take smaller h. There are methods for choosing h but not always reliable. In R, the default is

$$h_0 = 0.9 \cdot \min\left\{\text{SD}, \frac{\text{IQR}}{1.34}\right\} \cdot n^{-\frac{1}{5}}.$$

There is a bias-variance trade-off:

$$\text{MSE}(\widehat{f}_h(x)) = \mathbb{E}[(\widehat{f}_h(x) - f(x))^2] = \text{Var}[\widehat{f}_h(x)] + (\mathbb{E}[\widehat{f}_h(x)] - f(x))^2 = \text{Var}[\widehat{f}_h(x)] + (\text{Bias}(\widehat{f}_h(x)))^2.$$

As h increases, bias increases and variance decreases; as h decreases, bias decreases and variance increases.

Example 2.9 (Gaussian Kernel)**.**

$$w(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

which is the default kernel in R.

Example 2.10 (Epanechnikov Kernel)**.**

$$w(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), |x| \leqslant \sqrt{5}.$$

Example 2.11 (Rectangular Kernel)**.**

$$w(x) = \frac{1}{2\sqrt{3}}, |x| \leqslant \sqrt{3}.$$

Example 2.12 (Triangular Kernel)**.**

$$w(x) = \frac{1}{\sqrt{6}}\left(1 - \frac{|x|}{\sqrt{6}}\right), |x| \leqslant \sqrt{6}.$$

Example 2.13 (Property of Rectangular Kernel)**.** Suppose n is large and h is small. Take $w(x) = \frac{1}{2}$ for $|x| \leqslant 1$, then

$$\widehat{f}_h(x) = \frac{1}{2nh}\sum_{i=1}^{n} I(x - h \leqslant X_i \leqslant x + h).$$

The mean of $\widehat{f}_h(x)$ is

$$\mathbb{E}[\widehat{f}_h(x)] = \frac{F(x+h) - F(x-h)}{2h} \approx f(x) + \frac{h^2}{6} f''(x)$$

and so the squared bias is

$$(\mathbb{E}[\widehat{f}_h(X)] - f(x))^2 \approx \frac{h^4}{36}[f''(x)]^2.$$

The variance is

$$\mathrm{Var}[\widehat{f}_h(x)] = \frac{1}{4h^2n}\mathrm{Var}[I(x-h \leqslant X_i \leqslant x+h)] \approx \frac{1}{4h^2n} \cdot 2hf(x) = \frac{f(x)}{2hn}$$

and thus the mean square error is

$$\mathrm{MSE}(\widehat{f}_h(x)) \approx \frac{f(x)}{2hn} + \frac{h^4}{36}[f''(x)]^2$$

and the latter term is minimized at $h^* = \gamma(x)n^{-\frac{1}{5}}$.

Here are two motivations: redistribution and convolution.

- **Redistribution**: The empirical distribution function $\widehat{F}$ puts probability mass $\frac{1}{n}$ at each of the points $X_1, \cdots, X_n$. We use the kernel with bandwidth h to redistribute the mass around each X_i :

  $$\frac{1}{nh}w\left(\frac{x-X_i}{h}\right),$$

  where

  $$\int_{-\infty}^{\infty} \frac{1}{nh}w\left(\frac{x-X_i}{h}\right)\mathrm{d}x = \frac{1}{n}\int_{-\infty}^{\infty} w(t)\mathrm{d}t = \frac{1}{n}.$$

  The density estimate is now simply the sum of these densities over all observations.

- **Convolution**: The distribution $Y_h = U + hV$ where $U \sim \widehat{F}$ and V has density w with $V \perp U$. Y_h is a continuous r.v. for each $h > 0$:

  $$P(Y_h \leqslant x) = \sum_{i=1}^{n} P(U + hV \leqslant x | U = X_i)P(U = X_i) = \frac{1}{n}\sum_{i=1}^{n} P\left(V \leqslant \frac{x-X_i}{h}\right).$$

  Differentiating we get the density estimate.

2.8 Uncertainty Estimation: Frequentist Approach

2.8.1 Unbiasedness

Definition 2.10. $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$ is an ***estimator*** of a parameter θ.

Definition 2.11. The ***sampling distribution*** of $\widehat{\theta}$ is its probability distribution, which will depend on θ (and possibly other unknown parameters).

Definition 2.12. The ***mean square error*** of $\widehat{\theta}$ is defined to be

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = \mathrm{Var}_\theta(\widehat{\theta}) + (\mathbb{E}_\theta(\widehat{\theta}) - \theta)^2 = \mathrm{Var}_\theta(\widehat{\theta}) + (\mathrm{Bias}_\theta(\widehat{\theta}))^2.$$

Definition 2.13. If $\mathrm{Bias}_\theta(\widehat{\theta}) = 0, \forall\theta \in \Theta$, then $\widehat{\theta}$ is said to be ***unbiased***.

Note that unbiasedness is considered to be a desirable property of an estimator but:

1. In many problems, unbiased estimators do not exist;

2. In some problems where they do exist, the estimator lies outside the parameter space with positive probability.

3. If $\widehat{\theta}$ is an unbiased estimator of θ and $g(x)$ is a non-linear function, then $\mathbb{E}_\theta[g(\widehat{\theta})] \neq g(\theta)$ unless $P_\theta(\widehat{\theta} = \theta) = 1$.

We worry about bias when:

1. $\widehat{\theta}$ is systematically larger or smaller than θ;

2. the squared bias is approximately equal to or greater than the variance.

Example 2.14 (Sample Variance)**.** Suppose $X_1, \cdots, X_n$ are independent with mean μ and variance σ^2. The sample variance (`var(x)` in R) is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

S^2 is an unbiased estimator of $\sigma^2 : \mathbb{E}[S^2] = \sigma^2$ but $S = \sqrt{S^2}$ is biased. If we assume $X_1, \cdots, X_n$ are normal, then

$$\mathbb{E}[S] = \sigma\left(\frac{2}{n-1}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \approx \sigma\left(1 - \frac{1}{4n} - \frac{7}{32n^2}\right).$$

Therefore

$$\mathbb{E}[S] - \sigma \approx -\frac{\sigma}{4n} - \frac{7\sigma}{32n^2}.$$

Note that the bias goes to 0 as $n \to \infty$.

Using $\mathbb{E}[S] \approx \sigma\left(1 - \frac{1}{4n}\right)$, variance can be approximated as follows:

$$\mathrm{Var}[S] = \mathbb{E}[S^2] - (\mathbb{E}[S])^2 \approx \frac{\sigma^2}{2n},$$

Therefore,

$$\mathrm{MSE}(S) \approx \frac{\sigma^2}{2n} + \frac{\sigma^2}{16n^2}.$$

Note that as n increases, the variance term is much larger than the squared bias term.

2.8.2 Consistency

Definition 2.14. Suppose that for each $n, \widehat{\theta}_n$ is based on $(X_1, \cdots, X_n)$. The sequence of estimators $\{\widehat{\theta}_n\}$ is ***consistent*** for θ if

$$\forall \varepsilon > 0, \theta \in \Theta, \lim_{n\to\infty} P_\theta(|\widehat{\theta}_n - \theta| > \varepsilon) = 0,$$

denoted by $\widehat{\theta}_n \xrightarrow{p} \theta$.

Property 2.2. If $\mathrm{MSE}_\theta(\widehat{\theta}_n) \to 0$ as $n \to \infty$, then $\{\widehat{\theta}_n\}$ is consistent.

Proof. We can prove by Chebyshev's Inequality. □

If we have enough information (i.e., n is large enough) then we can estimate θ arbitrarily precisely. For a finite n, consistency is not meaningful.

Example 2.15 (Sample Mean)**.** Suppose $X_1, \cdots, X_n$ are independent with mean μ and variance σ^2. We estimate μ by

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By WLLN, $\widehat{\mu}_n$ is a consistent estimator of μ (i.e., $\{\widehat{\mu}_n\}$ is consistent). Likewise, if we want to estimate $\theta = g(\mu)$ where g is a continuous function, then $\widehat{\theta}_n = g(\widehat{\mu}_n)$ is a consistent estimator of θ.

We can also approximate the sampling distributions of $\widehat{\mu}_n$ and $\widehat{\theta}_n$ by normal distribution:

$$\widehat{\mu}_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$
$$\widehat{\theta}_n \dot{\sim} \mathcal{N}\left(\theta, [g'(\mu)]^2\frac{\sigma^2}{n}\right).$$

Example 2.16 (Regression Design)**.** Suppose the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \cdots, n$ and ε_i are independent $\mathcal{N}(0, \sigma^2)$ r.v.s.. The least squares estimator of β_1 is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

Thus $\widehat{\beta}_1 = \widehat{\beta}_1^{(n)}$ will be consistent if

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \to \infty.$$

If $x_i = x_0$ for all but a finite small number of i, it might fail.

2.8.3 Sampling Distributions and Standard Errors

Definition 2.15. The ***standard error*** of $\widehat{\theta}, \mathrm{SE}(\widehat{\theta})$ is defined to be the standard deviation of the sampling distribution of $\widehat{\theta}$.

Example 2.17. If $\widehat{\mu} = \overline{X}$, where $\overline{X}$ is based on n independent observations with variance σ^2 then $\mathrm{SE}(\widehat{\mu}) = \frac{\sigma}{\sqrt{n}}$.

IF the sampling distribution is approximately normal then we can approximate the standard error by the standard deviation of the approximating normal distribution. To estimate the standard error, we need to estimate some additional unknown parameters.

Example 2.18. $\mathrm{SE}(\widehat{\theta}) = \frac{\sigma}{\sqrt{n}}$ and the estimated standard error is

$$\widehat{\mathrm{SE}}(\widehat{\mu}) = \frac{S}{\sqrt{n}}, S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Example 2.19 (The Delta Method Estimator)**.** Suppose $X_1, \cdots, X_n$ are independent with some unknown CDF F, and $\widehat{\theta} = g(\overline{X})$ (e.g., $\theta = g(\mu), \mu = \mathbb{E}[X_i]$). If g is differentiable then we can approximate the sampling distribution of $\widehat{\theta}$ by a normal distribution

$$\widehat{\theta} = g(\overline{X}) \dot{\sim} \mathcal{N}\left(g(\mu) = \theta, [g'(\mu)]^2\frac{\sigma^2}{n}\right),$$

where $\sigma^2 = \text{Var}[X_i]$. We can estimate $\text{SE}(\widehat{\theta})$ using the Delta method estimator

$$\widehat{\text{SE}}(\widehat{\theta}) = \frac{|g'(\overline{X})|S}{\sqrt{n}},$$

where S^2 is the sample variance of $X_1, \cdots, X_n$. Note that we use substitution principle here to estimate the unknown μ and σ^2.

Example 2.20 (Trimmed Mean)**.** Suppose $X_1, \cdots, X_n$ are independent continuous r.v.s. with density $f(x-\theta)$ where $f(x) = f(-x)$ and θ is unknown. We also assume that f is heavy-tailed. To minimize the effect of extreme observations, we estimate θ by a trimmed mean:

$$\widehat{\theta} = \frac{1}{n-2r}\sum_{k=r+1}^{n-r} X_{(k)}.$$

In general, the trimmed mean is a substitution principle estimator of

$$\theta(F) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(\tau)\mathrm{d}\tau,$$

where $\alpha = \frac{r}{n}$.

The sampling distribution of the trimmed mean is approximately normal:

$$\widehat{\theta} \dot{\sim} \mathcal{N}\left(\theta, \frac{v^2(F)}{n}\right),$$

where

$$v^2(F) = \frac{1}{(1-2\alpha)^2}\int_{\alpha}^{1-\alpha}\int_{\alpha}^{1-\alpha} \frac{\min(s,t)-st}{f(F^{-1}(t))f(F^{-1}(s))}\mathrm{d}s\mathrm{d}t.$$

Suppose that $\widehat{\theta}$ is some complicated estimator like a trimmed mean in the example. Noted that if we could approximate $\widehat{\theta}$ by some sort of sample mean, i.e.,

$$\widehat{\theta} \approx \frac{1}{n}\sum_{i=1}^{n}\phi(X_i),$$

then we could estimate the standard error of $\widehat{\theta}$ from the sample variance of $\phi(X_1), \cdots, \phi(X_n)$. For example, $\widehat{\text{SE}}(\widehat{\theta}) = \frac{S_\phi}{\sqrt{n}}$, where $S_\phi^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\phi(X_i)-\overline{\phi})^2$ with $\overline{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$.

Recall that the Delta method follows from the Taylor series approximation:

$$\widehat{\theta} - \theta = g(\overline{X}) - g(\mu) \approx g'(\mu)(\overline{X}-\mu) = \frac{1}{n}\sum_{i=1}^{n} g'(\mu)(X_i-\mu),$$

and thus

$$\widehat{\theta} \approx \frac{1}{n}\sum_{i=1}^{n}(g(\mu)+g'(\mu)(X_i-\mu)) := \frac{1}{n}\sum_{i=1}^{n}\phi(X_i).$$

We estimate $\phi(X_i)$ by pseudo-value:

$$\Phi_i = g(\overline{X}) + g'(\overline{X})(X_i-\overline{X}), i = 1, \cdots, n,$$

and

$$\widehat{\theta} = g(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\Phi_i = \overline{\Phi}.$$

The Delta method estimator can now be written as

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\Phi_i - \overline{\Phi})^2\right]^{\frac{1}{2}} = \frac{|g'(\overline{X})|S}{\sqrt{n}}.$$

2.8.4 Jackknife Standard Error Estimator

Definition 2.16. Suppose that $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$, define ***leave-one-out estimators***

$$\widehat{\theta}_{-i} = \widehat{\theta}(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n).$$

Example 2.21. If $\widehat{\theta} = \overline{X}$, then

$$\widehat{\theta}_{-i} = \frac{1}{n-1}\sum_{j\neq i} X_j.$$

Example 2.22 (Theil Index)**.** Define

$$\theta(F) = \mathbb{E}_F\left[\frac{X_i}{\mu(F)}\ln\left(\frac{X_i}{\mu(F)}\right)\right],$$

where $P(X_i > 0) = 1$. We estimate $\theta(F)$ by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\overline{X}}\ln\left(\frac{X_i}{\overline{X}}\right).$$

The leave-one-out estimators are

$$\widehat{\theta}_{-i} = \frac{1}{n-1}\sum_{j\neq i}\frac{X_j}{\overline{X}_{-i}}\ln\left(\frac{X_j}{\overline{X}_{-i}}\right),$$

where $\overline{X}_{-i} = \frac{1}{n-1}\sum_{j\neq i} X_j$.

Suppose we can approximate $\widehat{\theta}$ by a sample mean:

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

for some unknown function ϕ. For the leave-one-out estimators, we have

$$\widehat{\theta}_{-i} \approx \frac{1}{n-1}\sum_{j\neq i}\phi(X_j).$$

Then we can recover $\phi(X_i)$ by the pseudo-value

$$\Phi_i = n\widehat{\theta} - (n-1)\widehat{\theta}_{-i} \approx \phi(X_i).$$

Definition 2.17. Given the pseudo-values $\Phi_1, \cdots, \Phi_n$, define the ***jackknife estimator*** of $\mathrm{SE}(\widehat{\theta})$:

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \left[\frac{1}{n(n-1)} \sum_{i=1}^{n} (\Phi_i - \overline{\Phi})^2 \right]^{\frac{1}{2}}.$$

A more compact form of the jackknife estimator is

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \left[\frac{n-1}{n} \sum_{i=1}^{n} (\widehat{\theta}_{-i} - \widehat{\theta}_{\bullet})^2 \right]^{\frac{1}{2}},$$

where

$$\widehat{\theta}_{\bullet} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{-i}.$$

Note that for many estimators, we have

$$\mathbb{E}_{\theta}[\widehat{\theta}] = \theta + \frac{a_1(\theta)}{n} + \frac{a_2(\theta)}{n^2} + \cdots.$$

We can use the jackknife to remove the $\frac{1}{n}$ bias term. Define the bias-corrected $\widehat{\theta}$:

$$\widehat{\theta}_{\mathrm{bc}} = n\widehat{\theta} - (n-1)\widehat{\theta}_{\bullet} = \widehat{\theta} - \underbrace{(n-1)(\widehat{\theta}_{\bullet} - \widehat{\theta})}_{\text{Bias correction}} = \frac{1}{n} \sum_{i=1}^{n} \Phi_i.$$

For $\widehat{\theta}_{\mathrm{bc}}$, we have

$$\mathbb{E}_{\theta}[\widehat{\theta}_{\mathrm{bc}}] = \theta + \frac{a_2^*(\theta)}{n^2} + \cdots.$$

Example 2.23 (Theil Index)**.** Generate 100 observations from a Gamma distribution with shape parameter $\alpha = 2$ and estimate Theil index and standard error.

```
x = rgamma(100, 2)
y = x / mean(x)
theil = mean(y * log(y))

# Compute pseudo-values
pseud = NULL
for (i in 1:100){
   xi = x[-i]
   yi = xi / mean(xi)
   loo = mean(yi * log(yi))
   pseud = c(pseud, 100 * theil - 99 * loo)
   }
mean(pseud) # Mean of pseudo-values - bias-corrected estimate
sqrt(var(pseud) / 100) # Jackknife std error estiamte
```

If an estimator cannot be well-approximated by an average then the jackknife tends to fare poorly. For example, if we estimate quantiles by order statistics then the jackknife standard error estimator is too small. We can improve the performance of the jackknife in these cases but deleting $d > 1$ observations to estimate the standard error.

Related to the jackknife is the bootstrap: we try to estimate the sampling distribution of θ by sampling with replacement from the data and then computing estimates from these samples.

Example 2.24 (Delta Method versus Jackknife)**.** Sample 100 observations from a Gamma distribution with $\alpha = 2$ and $\lambda = 1$. Estimate $\theta = \ln(\mu) = g(\mu)$ with $g'(\mu) = \frac{1}{\mu}$.

For our sample $\overline{x} = 1.891$ and $s^2 = 1.911$, and $\widehat{\theta} = \ln(\overline{x}) = 0.637$.

Thus the Delta method standard error estimate is

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \frac{|g'(\overline{x})|s}{\sqrt{n}} = \frac{s}{\overline{x}\sqrt{n}} = 0.0731.$$

Now we compute the jackknife estimate:

```
x = rgamma(100, 2)
thetaloo = NULL
for (i in 1:100){
   xi = x[-i]
   thetaloo = c(thetaloo, log(mean(xi)))
   }
jackse = sqrt(99 * sum((thetaloo - mean(thetaloo))^2) / 100)
```

Example 2.25 (Lorenz Curve and Gini Index)**.** Suppose F is the CDF of a positive r.v. with finite mean $\mu(F)$. For example, F describes the income distribution within some population. For each such F, we can define its Lorenz curve

$$\mathcal{L}_F(\tau) = \frac{1}{\mu(F)} \int_0^\tau F^{-1}(s)\mathrm{d}s, 0 \leqslant \tau \leqslant 1,$$

which is the fraction of total income held by poorest $100\tau\%$. We have $\mathcal{L}_F(\tau) \leqslant \tau$ with $\mathcal{L}_F(0) = 0$ and $\mathcal{L}_F(1) = 1$. The difference between τ and $\mathcal{L}_F(\tau)$ can be used to measure income inequality.

One measure of income inequality is the Gini index defined by

$$\mathrm{Gini}(F) = 2\int_0^1 (\tau - \mathcal{L}_F(\tau))\mathrm{d}\tau = \frac{1}{\mu(F)} \int_0^1 (2\tau - 1)F^{-1}(\tau)\mathrm{d}\tau.$$

$\mathrm{Gini}(F) \in [0, 1]$: $\mathrm{Gini}(F) = 0$ represents perfect equality and $\mathrm{Gini}(F) = 1$ represents perfect inequality. Gini indices for countries range from around 0.2 to over 0.6.

We estimate the quantiles $F^{-1}(\tau)$ by order statistics. Given independent observations $X_1, \cdots, X_n$ from F, we have

$$\widehat{\mathrm{Gini}}(F) = \frac{1}{n\overline{X}} \sum_{k=1}^{n} \left(\frac{2k-1}{n} - 1 \right) X_{(k)}$$

and the leave-one-out estimates are easy to compute here.

```
gini = function(x){
   # Compute point estimate
   n = length(x)
   x = sort(x)
   wt = (2 * c(1:n) - 1) / n - 1
   g = sum(wt * x) / sum(x)

   # Compute leave-one-out estimates
```

```
wt1 = (2 * c(1:(n-1)) - 1) / (n - 1) -1
gi = NULL
for (i in 1:n){
   x1 = x[-i] # Data with x[i] deleted
   gi = c(gi, sum(wt1 * x1) / sum(x1))
   }

# Compute jackknife std error estimate
gbar = mean(gi)
se = sqrt((n-1) * sum((gi - gbar)^2) / n)
r = list(gini = g, se = se)
r
}
```

Note that in this example, our estimate of Gini(F) does not use the fact that the data were Gamma distributed. If F is a Gamma distribution with shape parameter α, then

$$\text{Gini}(F) = \frac{\Gamma\left(\alpha + \frac{1}{2}\right)}{\alpha\Gamma(\alpha)\sqrt{\pi}} = \text{Gini} - \text{Gamma}(\alpha).$$

3 Point and Interval Estimation

3.1 Parametric Models

We observe $(X_1, \cdots, X_n)$ whose joint PDF of PMF is

$$f(x_1. \cdots, x_n; \theta_1, \cdots, \theta_k)$$

for some unknown (real-valued) parameters $\theta_1, \cdots, \theta_k$.

Note that the distribution of $(X_1, \cdots, X_n)$ depends only on a finite number of unknowns. Some parameters are more interesting to us than others and the non-interesting parameters are called nuisance parameters.

Example 3.1 (Simple Linear Regression)**.** $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \cdots, n$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We are interested in β_0, β_1 with σ^2 being a nuisance parameter. However, estimating σ^2 is essential for estimating uncertainty of LS estimators of β_0 and β_1.

3.2 Point Estimation

We observe $X_1, \cdots, X_n$ assumed to have a distribution depending on some unknown parameter θ. We can estimate the value of θ by a ***point estimator*** $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$ and $\widehat{\theta}$ will have a sampling distribution, which will depend on θ as well as possibly other unknown parameters. The standard deviation of the sampling distribution is called the ***standard error***. We can often estimate the standard error using the Delta method or jackknife method.

3.3 Interval Estimation

We can use ***interval estimation***. Define an interval

$$\mathcal{I} = [l(X_1, \cdots, X_n), u(X_1, \cdots, X_n)]$$

that we believe will contain θ with probability close to 1.

- Confidence intervals: These are typically defined in terms of the sampling distribution of a point estimator $\widehat{\theta}$. We will often need to use approximations to the sampling distributions and the confidence level is defined in terms of repeated sampling.
- Credible intervals: These are based on the posterior distribution of θ given the observed data $x_1, \cdots, x_n$. If $\pi(\theta | x_1, \cdots, x_n)$ is the posterior density of θ, then $\mathcal{I}$ is a $100p\%$ credible interval if
$$\int_{\mathcal{I}} \pi(\theta | x_1, \cdots, x_n) \mathrm{d}\theta = p.$$
The credible level is defined in terms of the posterior distribution, which depends on the prior distribution and the data.

Definition 3.1. An interval $\mathcal{I} = [l(X_1, \cdots, X_n), u(X_1, \cdots, X_n)]$ is a ***confidence interval*** (CI) with coverage $100p\%$ (or a $100p\%$ CI) if

$$\underbrace{P_\theta[l(X_1, \cdots, X_n) \leqslant \theta \leqslant u(X_1, \cdots, X_n)]}_{P_\theta(\theta \in \mathcal{I})} = p, \forall \theta \in \Theta.$$

If the probability statement above holds approximately (e.g., if n is large) then we often say that the interval is an approximate $100p\%$ CI for θ. Typically, we have $\mathcal{I}_n = [l_n(X_1, \cdots, X_n), u_n(X_1, \cdots, X_n)]$ with

$$P_\theta(\theta \in \mathcal{I}_n) \to p$$

as $n \to \infty, \forall \theta \in \Theta$.

Note that (1) In the definition of a CI, the interval $\mathcal{I}$ is a random interval depending on the r.v.s. and a CI is defined in terms of the probability that the random interval contains θ;

(2) The data-based interval $[l(x_1, \cdots, x_n), u(x_1, \cdots, x_n)]$ cannot be interpreted in terms of the probability distribution of $(X_1, \cdots, X_n)$ since the interval either contains θ or does not;

(3) The length of the CI gives an idea about the uncertainty in the estimation of θ much like an estimate of the standard error (many CIs are formed in terms of an estimator and its standard error).

Example 3.2. Suppose $X_1, \cdots, X_{20}$ are independent $\mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown. Classic 95% CI for μ is

$$\left[\overline{X} \mp 2.093 \frac{S}{\sqrt{20}}\right]$$

where $\overline{X}$ and S^2 are the sample mean and variance respectively, 2.093 is the 0.975 quantile of Student's t distribution with 19 degrees of freedom. Note that $\widehat{\mathrm{SE}}(\overline{X}) = \frac{S}{\sqrt{n}}$.

Example 3.3 (Simulation Experiment)**.** Generate 100 samples of size 20 from a $\mathcal{N}(0, 1)$ distribution and compute 95% CIs for each sample. Given the theory, we would expect that approximately 95 of the 100 constructed CIs will contain the true mean 0.

3.3.1 Pivotal Method

We find a r.v. $g(X_1, \cdots, X_n, \theta)$ whose distribution is independent of θ and any other unknown parameters. $P_\theta[g(X_1, \cdots, X_n, \theta) \leqslant x] = G(x)$ where $G(x)$ is completely known and $g(X_1, \cdots, X_n, \theta)$ is called a ***pivot***. Given the pivot, we choose a and b s.t.

$$p = P_\theta[a \leqslant g(X_1, \cdots, X_n, \theta) \leqslant b] = \underbrace{G(b) - G(a-)}_{\text{Independent of } \theta}.$$

We get

$$p = P_\theta[a \leqslant g(X_1, \cdots, X_n, \theta) \leqslant b] = \cdots = P_\theta[l(X_1, \cdots, X_n) \leqslant \theta \leqslant u(X_1, \cdots, X_n)].$$

- Choice of pivot: If we have a point estimator $\widehat{\theta}$ then we can often define the pivot to be $g(\widehat{\theta}, \theta)$ where g is chosen to make its distribution independent of θ. If we cannot find an exact pivot, we can sometimes find an approximate pivot $g(\widehat{\theta}, \theta)$ whose distribution is approximately independent of θ. For example,

$$\frac{\widehat{\theta} - \theta}{\widehat{\mathrm{SE}}(\widehat{\theta})} \dot{\sim} \mathcal{N}(0, 1).$$

- Choice of a and b : Ideally, we want to choose a and b to make the CI as short as possible. However if G is the CDF of the pivot then a good default is to define a s.t. $G(a) = \frac{1-p}{2}$ and b s.t. $1 - G(b) = \frac{1-p}{2}$.

Example 3.4 (Normal Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$. Using pivot,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\widehat{\mu} - \mu}{\widehat{\mathrm{SE}}(\widehat{\mu})} \sim \mathcal{T}(n-1)$$

where $\mathcal{T}(n-1)$ is a Student's t distribution with $n-1$ degrees of freedom. Define t_p s.t.

$$P\left(-t_p \leqslant \frac{\overline{X} - \mu}{S/\sqrt{n}} \leqslant t_p\right) = p,$$

then the $100p\%$ CI for μ is

$$\left[\overline{X} \mp t_p \frac{S}{\sqrt{n}}\right].$$

We also have

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1).$$

Define a and b s.t.

$$P\left(a \leqslant \frac{(n-1)S^2}{\sigma} \leqslant b\right) = p$$

from which a $100p\%$ CI for σ^2 is

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right].$$

We choose a and b s.t. $G(a) = 1 - G(b)$ with $G(b) - G(a) = p$ and $\arg\min_{a,b} a^{-1} - b^{-1}$.

Example 3.5 (Exponential Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent exponential r.v.s. with

$$f(x; \lambda) = \lambda e^{-\lambda x}, x \geqslant 0,$$

where $\lambda > 0$ is unknown. We can estimate λ by $\widehat{\lambda} = \frac{1}{\overline{X}}$, and we can approximate the distribution of $\widehat{\lambda}$ as follows:

$$\sqrt{n}(\widehat{\lambda} - \lambda) \dot{\sim} \mathcal{N}(0, \lambda^2) \text{ or } \widehat{\lambda} \dot{\sim} \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right).$$

Thus,

$$\mathrm{SE}(\widehat{\lambda}) \approx \frac{\lambda}{\sqrt{n}} \Rightarrow \widehat{\mathrm{SE}}(\widehat{\lambda}) = \frac{\widehat{\lambda}}{\sqrt{n}}.$$

There are four possible pivots for λ :

- $\lambda \sum_{i=1}^{n} X_i \sim \mathrm{Gamma}(n, 1)$ (exact pivot).
- $\frac{\widehat{\lambda} - \lambda}{\lambda/\sqrt{n}} \dot{\sim} \mathcal{N}(0, 1)$ (approximate pivot).
- $\frac{\widehat{\lambda} - \lambda}{\widehat{\lambda}/\sqrt{n}} \dot{\sim} \mathcal{N}(0, 1)$ (approximate pivot).
- $\sqrt{n}(\ln(\widehat{\lambda}) - \ln(\lambda)) \dot{\sim} \mathcal{N}(0, 1)$(approximate pivot using variance stabilizing transformation).

The CI resulting from the four pivots are:

- $\left[\frac{a}{\sum_{i=1}^{n} X_i}, \frac{b}{\sum_{i=1}^{n} X_i}\right]$, where a and b are s.t. $P(a \leqslant \mathrm{Gamma}(n, 1) \leqslant b) = p$.

- $\left[\frac{\widehat{\lambda}}{1\pm z_p/\sqrt{n}}\right]$ where z_p satisfies $P(-z_p \leqslant \mathcal{N}(0,1) \leqslant z_p) = p$.
- $[\widehat{\lambda}(1 \mp z_p/\sqrt{n})]$.
- $[\widehat{\lambda}e^{\mp z_p/\sqrt{n}}]$.

Note that all four CIs have the form $[\widehat{\lambda}l_p, \widehat{\lambda}u_p]$ where l_p and u_p depend on the pivot. As n increases, l_p and u_p are quite similar for the four pivots:

$$\begin{aligned} 1 + \frac{z_p}{\sqrt{n}} - \left(1 - \frac{z_p}{\sqrt{n}}\right) &= \frac{2z_p}{\sqrt{n}} \\ \frac{1}{1 - z_p/\sqrt{n}} - \frac{1}{1 + z_p/\sqrt{n}} &= \frac{2z_p}{\sqrt{n}} + \frac{2z_p^3}{n^{3/2}} + \cdots \\ e^{z_p/\sqrt{n}} - e^{-z_p/\sqrt{n}} &= \frac{2z_p}{\sqrt{n}} + \frac{z_p^3}{3n^{3/2}} + \cdots \end{aligned}$$

and when n is large, a Gamma$(n, 1)$ distribution can be approximated by a $\mathcal{N}(n, n)$ so that $a \approx n - z_p\sqrt{n}$ and $b \approx n + z_p\sqrt{n}$.

Example 3.6 (CIs for Quantiles $F^{-1}(\tau)$)**.** Suppose $X_1, \cdots, X_n$ are independent r.v.s. with continuous CDF F and PDF f. We can use the fact that the order statistic $X_{(k)} = \widehat{\theta}$ where $\tau \approx \frac{k}{n}$ is approximately normal with mean θ and variance $\frac{\tau(1-\tau)}{n[f(\theta)]^2}$. We estimate $f(\theta)$ by $\widehat{f}(\widehat{\theta})$ (e.g., kernel density estimation) and obtain approximate pivot

$$\frac{\sqrt{n}\widehat{f}(\widehat{\theta})}{\sqrt{\tau(1-\tau)}}(\widehat{\theta} - \theta) \dot{\sim} \mathcal{N}(0, 1).$$

The approximate 95% CI is

$$\left[\widehat{\theta} \mp 1.96 \times \frac{\sqrt{\tau(1-\tau)}}{\sqrt{n}\widehat{f}(\widehat{\theta})}\right].$$

But the approach is very dependent on $\widehat{f}(\widehat{\theta})$ being a good estimator of $f(\theta)$.

We can also use the pivot

$$g(X_1, \cdots, X_n, \theta) = \sum_{i=1}^{n} I(X_i \leqslant \theta) \sim \text{Bin}(n, \tau).$$

Note tat if $a < b$ are integers between 1 and n then

$$\left\{\theta : a \leqslant \sum_{i=1}^{n} I(X_i \leqslant \theta) \leqslant b\right\} = \{\theta : X_{(a)} \leqslant \theta \leqslant X_{(b)}\}.$$

Thus $[X_{(a)}, X_{(b)}]$ is a $100p\%$ CI for $\theta = F^{-1}(\tau)$ where

$$p = \sum_{k=a}^{b} \binom{n}{k} \tau^k (1-\tau)^{n-k},$$

which is a distribution-free CI for θ and it works for any continuous distribution.

For a given p we can find a and b using a normal approximation to the Binomial distribution so that

$$p \approx \sum_{k=a}^{b} \binom{n}{k} \tau^k (1-\tau)^{n-k}.$$

If n is large enough then

$$\text{Binomial}(n, \tau) \dot{\sim} \mathcal{N}(n\tau, n\tau(1-\tau)).$$

Using the normal approximation with a continuity correction, we have

$$a = \left\lfloor n\tau + \frac{1}{2} - z_p\sqrt{n\tau(1-\tau)} \right\rfloor$$
$$b = \left\lceil n\tau - \frac{1}{2} + z_p\sqrt{n\tau(1-\tau)} \right\rceil,$$

where $\lfloor x \rfloor$ and $\lceil x \rceil$ round x respectively, down and up to the nearest integer.

Example 3.7 (Comparison of CIs for the Normal Mean)**.** Suppose $X_1, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ r.v.s., note that $\mu = \mathbb{E}[X_i] = F^{-1}\left(\frac{1}{2}\right)$. We want to compare the two CIs: $\left[\overline{X} \pm t_p \frac{S}{\sqrt{n}}\right]$ and $[X_{(a)}, X_{(b)}]$.

The lengths when n is large:

- $t_p \to z_p$ and $S \xrightarrow{p} \sigma$ so that length of parametric CI is approximately $\frac{2z_p\sigma}{\sqrt{n}}$.
- $X_{(b)} - X_{(a)} \approx \frac{z_p\sigma\sqrt{2\pi}}{\sqrt{n}}$.

Thus

$$\frac{\text{Length of distribution free } CI}{\text{Length of parametric } CI} \approx \frac{\sqrt{2\pi}}{2} \approx 1.253.$$

Hence, distribution free CI is always valid but parametric CI will be shorter if that model is correct.

For example, $n = 20$,

3.4 Methods of Estimation: Method of Moments

This is an application of the substitution principle and is useful for quick and dirty and robust estimators.

Start by assuming $k = 1$ (single unknown parameter). We want to find a statistic $T(X_1, \cdots, X_n)$ s.t. $\mathbb{E}_\theta[T(X_1, \cdots, X_n)] = h(\theta)$, where h has a well-defined inverse. Then we set $T(X_1, \cdots, X_n) = h(\widehat{\theta})$ so that

$$\widehat{\theta} = h^{-1}(T).$$

Example 3.8 (Special Case)**.** $X_1, \cdots, X_n$ are independent with PDF of PMF $f(x; \theta)$, where $\mathbb{E}_\theta[X_i] = h(\theta)$. By substitution principle, we estimate $\mathbb{E}_\theta[X_i]$ by $\overline{X}$, then $\overline{X} = h(\widehat{\theta})$ and so $\widehat{\theta} = h^{-1}(\overline{X})$.

If we have k unknown parameters, we need k moment conditions to implement the method and we need to be careful to ensure that we have a injective map between $\theta_1, \cdots, \theta_k$ and the moments. We can also use quantiles as moment conditions: If $X_1, \cdots, X_n$ are independent with CDF F_θ and $F_\theta^{-1}(\tau) = h(\theta)$, we can define $\widehat{\theta} = h^{-1}(X_{(k)})$ where $\frac{k}{n} \approx \tau$.

Method of moments (MoM) estimators are often disparaged as inefficient. But the approach to estimation is often very useful in practice: useful as initial estimates for computing more computationally complex estimators (e.g., maximum likelihood estimators); useful for assessing goodness-of-fit of a model.

Example 3.9 (Exponential Distribution)**.** $X_1, \cdots, X_n$ are independent with

$$f(x;\lambda) = \lambda e^{-\lambda x}, x \geqslant 0,$$

where $\lambda > 0$ is unknown. For $r > 0$, we have

$$\mathbb{E}_\lambda(X_i^r) = \lambda^{-r}\Gamma(r+1),$$

suggesting that

$$\frac{1}{n}\sum_{i=1}^n X_i^r = \frac{\Gamma(r+1)}{\widehat{\lambda}^r}$$

or

$$\widehat{\lambda}(r) = \left(\frac{1}{n\Gamma(r+1)}\sum_{i=1}^n X_i^r\right)^{-\frac{1}{r}}.$$

The CLT gives us

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i^r - \frac{\Gamma(r+1)}{\lambda^r}\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}_\lambda(X_i^r))$$

and we can apply the Delta method to get

$$\sqrt{n}(\widehat{\lambda}(r) - \lambda) \dot{\sim} \mathcal{N}\left(0, \sigma^2(r) = \frac{\lambda^2}{r^2}\left(\frac{\Gamma(2r+1)}{\Gamma(r+1)} - 1\right)\right).$$

$\sigma^2(r)$ is minimized at $r = 1$.

Example 3.10 (Goodness-of-Fit for the Exponential Distribution)**.** We can compare $\widehat{\lambda}(r)$ to $\widehat{\lambda}(1)$: if the data are exponential then $\frac{\widehat{\lambda}(r)}{\widehat{\lambda}(1)} \approx 1$ for all $r > 0$. We can use MoM plot for $\ln\left(\frac{\widehat{\lambda}(r)}{\widehat{\lambda}(1)}\right)$ against r. If the data are exponential, then

$$\ln\left(\frac{\widehat{\lambda}(r)}{\widehat{\lambda}(1)}\right) \dot{\sim} \mathcal{N}\left(0, \frac{\sigma^2(r) - 1}{n}\right)$$

for $r > 0$.

Example 3.11 (Gamma Distribution)**.** $X_1, \cdots, X_n$ are independent with

$$f(x;\lambda) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, x \geqslant 0,$$

where $\lambda, \alpha > 0$ are unknown. We have

$$\mathbb{E}[X_i] = \frac{\alpha}{\lambda}, \mathrm{Var}[X_i] = \frac{\alpha}{\lambda^2}.$$

So we set

$$\overline{X} = \frac{\widehat{\alpha}}{\widehat{\lambda}}, S^2 = \frac{\widehat{\alpha}}{\widehat{\lambda}^2}$$

and get

$$\widehat{\alpha} = \frac{\overline{X}^2}{S^2}, \widehat{\lambda} = \frac{\overline{X}}{S^2}.$$

Example 3.12 (Capture-Recapture Experiments)**.** The experiments are used to estimate the size of animal populations and census undercount. We have a finite population with unknown population size N, which is a discrete parameter (integer-valued). Simple capture-recapture experiment has two steps: (1) Capture m_0 individuals from the population, which are marked and released back to the population; (2) Capture m_1 individuals and define X to be the number of marked individuals. X has a hypergeometric distribution:

$$P_N(X = x) = f(x; N) = \frac{\binom{m_0}{x}\binom{N-m_0}{m_1-x}}{\binom{N}{m_1}}.$$

We can use method of moments to estimate N :

$$\mathbb{E}_N[X] = m_1\frac{m_0}{N} \Rightarrow X = m_1\frac{m_0}{\widehat{N}}.$$

The estimator $\widehat{N} = \frac{m_0 m_1}{X}$ is called the Lincoln-Petersen estimator and can be extremely unstable if X is small.

Example 3.13 (Colley Matrix)**.** Using the outcomes of games between pairs of teams (k teams in total), we want to rank the teams from weakest to strongest. Teams may not play each other an equal number of times and some teams may not play each other at all. The model with unknown strength parameters $s_1, \cdots, s_k$ is

$$P(i \text{ beats } j) = \frac{1}{2} + s_i - s_j, 1 \leqslant i \neq j \leqslant k.$$

For model identifiability, we put the following constraint on $s_1, \cdots, s_k$:

$$\frac{1}{k}\sum_{i=1}^{k} s_i = \frac{1}{2}.$$

We estimate $s_1, \cdots, s_k$ using method of moments.

We define W_i be the number of wins for team $i, n_{ij} = n_{ji}$ be the number of games between teams i and j, and $n_{i\bullet} = \sum_{j\neq i} n_{ij}$ be the total number of games played by team i. From our model

$$\mathbb{E}[W_i] = \sum_{j\neq i} n_{ij}\left(\frac{1}{2} + s_i - s_j\right), i = 1, \cdots, k.$$

Thus the method of moments estimator of $s_1, \cdots, s_k$ satisfy

$$W_i = \sum_{j\neq i} n_{ij}\left(\frac{1}{2} + \widehat{s}_i - \widehat{s}_j\right)$$

with $\frac{\widehat{s}_1+\cdots+\widehat{s}_k}{k} = \frac{1}{2}$. In matrix form, we have

$$\underbrace{\begin{pmatrix} n_{1\bullet} & -n_{12} & -n_{13} & \cdots & -n_{1k} \\ -n_{21} & n_{2\bullet} & -n_{23} & \cdots & -n_{2k} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ -n_{k1} & -n_{k2} & \cdots & -n_{k,k-1} & n_{k\bullet} \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}}_{(k+1)\times k} \begin{pmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \vdots \\ \widehat{s}_k \end{pmatrix} = \begin{pmatrix} W_1 - n_{1\bullet}/2 \\ W_2 - n_{2\bullet}/2 \\ \vdots \\ W_k - n_{k\bullet}/2 \\ k/2 \end{pmatrix}.$$

The Colley's variation is

$$\underbrace{\begin{pmatrix} n_{1\bullet}+2 & -n_{12} & \cdots & -n_{1k} \\ -n_{21} & n_{2\bullet}+2 & \cdots & -n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -n_{k1} & -n_{k2} & \cdots & n_{k\bullet}+2 \end{pmatrix}}_{k\times k} \begin{pmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \vdots \\ \widehat{s}_k \end{pmatrix} = \begin{pmatrix} W_1 + 1 - n_{1\bullet}/2 \\ W_2 + 1 - n_{2\bullet}/2 \\ \vdots \\ W_k + 1 - nk\bullet/2 \end{pmatrix}.$$

This method has other applications to social networks (ranking friends), animal social behavior (dominance) etc. There are some related methods: (1) Bradley-Terry model:

$$P(i \text{ beats } j) = \frac{e^{s_i - s_j}}{1 + e^{s_i - s_j}};$$

(2) Page-Rank: Markov chain $\{X_t\}$ with n states (teams) - team ranking based on stationary distribution of empirical transition matrix (estimated from the data).

3.5 Methods of Estimation: Maximum Likelihood

Definition 3.2. Suppose $(X_1, \cdots, X_n)$ are r.v.s. with joint PDF of PMF $f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_k)$ where $\theta_1, \cdots, \theta_k$ are unknown parameters. Given the data $x_1, \cdots, x_n$, we define the ***likelihood function***

$$\mathcal{L}(\theta_1, \cdots, \theta_k) = f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_k),$$

which is a function over the parameter space.

Definition 3.3. Suppose that for each $\mathbf{x} = (x_1, \cdots, x_n)$, $(T_1(\mathbf{x}), \cdots, T_k(\mathbf{x}))$ maximize $\mathcal{L}(\theta_1, \cdots, \theta_k)$, then ***maximum likelihood estimators*** (***MLEs***) of $\theta_1, \cdots, \theta_k$ are

$$\widehat{\theta}_j = T_j(X_1, \cdots, X_n) \text{ for } j = 1, \cdots, k.$$

Note that maximum likelihood estimation is essentially an ad hoc procedure albeit one that works very well in many problems. MLEs need not be unique although this is very rare in practice but a likelihood function can have many local maxima. MLEs need not exist, for example non-existence typically occurs if the sample size is very small. Even if an MLE does not exist, the likelihood function provides information about θ. All the information about θ in the data is contained in the likelihood function.

Definition 3.4. Suppose $X_1, \cdots, X_n$ are independent with PDF pr PMF $f(x; \theta)$ for some real-valued $\theta \in \Theta$. Assume Θ is an open set, $A = \{x : f(x; \theta) > 0\}$ does not depend on θ, and $f(x; \theta)$ is differentiable w.r.t. θ for each $x \in A$. We define the ***log-likelihood function***

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^{n} \ln f(X_i; \theta).$$

Example 3.14 (Non-Regular Model)**.** Suppose $X_1, \cdots, X_n$ are independent r.v.s. with PDF

$$f(x; \theta) = \frac{|x - \theta|^{-1/2}}{2\sqrt{\pi}} \exp(-|x - \theta|).$$

The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[\frac{|x - \theta|^{-1/2}}{2\sqrt{\pi}} \exp(-|x - \theta|) \right].$$

Note that as $\theta \to x_i, \mathcal{L}(\theta) \to \infty$, which suggests that either the MLE does not exist or that any X_i is an MLE of θ.

3.5.1 Sufficiency

Definition 3.5. A statistic $\mathbf{T}(T_1(\mathbf{X}), \cdots, T_m(\mathbf{X}))$ is a ***sufficient statistic*** for θ (or a model) if the conditional distribution of X given $\mathbf{T} = \mathbf{t}$ depends only on $\mathbf{t}$ and not θ.

Theorem 3.1 (Neyman Factorization Theorem)**.** Suppose that the joint PDF or PMF of $\mathbf{X} = (X_1, \cdots, X_n)$ is $f(\mathbf{x}; \theta)$, then the statistic $\mathbf{T} = (T_1(\mathbf{X}), \cdots, T_m(\mathbf{X}))$ is a sufficient statistic for θ iff

$$f(\mathbf{x}; \theta) = g(\mathbf{T}(\mathbf{x}); \theta)h(\mathbf{x}),$$

where the function h does not depend on θ.

By theorem, we know the likelihood function is

$$\mathcal{L}(\theta) = f(\mathbf{x}; \theta) = g(\mathbf{T}(\mathbf{x}); \theta)h(\mathbf{x}).$$

Since $h(\mathbf{x})$ does not depend on θ, it is just a multiplicative constant in the likelihood function and maximizing $\mathcal{L}(\theta)$ is equivalent to maximizing $g(\mathbf{T}(\mathbf{x}); \theta)$. Effectively, $\mathcal{L}(\theta)$ depends on the data $\mathbf{x}$ only through the value of $\mathbf{T}(\mathbf{x})$. If the MLE is unique, it depends on $\mathbf{X}$ only through $\mathbf{T}(\mathbf{X})$.

3.5.2 Maximum Likelihood Estimation

Two general scenarios for maximizing $\mathcal{L}(\theta)$:

- $\mathcal{L}(\theta)$ is differentiable and the parameter space Θ is an open set, then $\widehat{\theta}$ is s.t.

  $$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(\widehat{\theta}) = 0.$$

  In some cases, we can use the second derivative to estimate the standard error.

- $\widehat{\theta}$ occurs at a boundary and we need to directly maximize $\mathcal{L}(\theta)$:
  - Boundary of Θ (if Θ is not an open set);
  - An extreme of the data (e.g., $\widehat{\theta} = X_{(n)}$).

Definition 3.6. Given the MLE $\widehat{\theta}$, we define the ***observed Fisher information*** as

$$-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln \mathcal{L}(\widehat{\theta}).$$

Recall that this is simply the absolute curvature of the log-likelihood function at its maximum. The greater this curvature, the mode well-defined the maximizer $\widehat{\theta}$, i.e., as the observed Fisher information increases, the uncertainty in estimator decreases. We can approximate the log-likelihood function by a quadratic function

$$g(\theta) = C - \frac{1}{2}\left[-\frac{\mathrm{d}}{\mathrm{d}\theta^2} \ln \mathcal{L}(\widehat{\theta})\right](\theta - \widehat{\theta})^2.$$

Example 3.15 (Uniform distribution)**.** Suppose $X_1, \cdots, X_n$ are independent Unif$(0, \theta)$ r.v.s. with $\theta > 0$ unknown:

$$f(x; \theta) = \begin{cases} \theta^{-1}, & 0 \leqslant x \leqslant \theta \\ 0, & \text{otherwise} \end{cases} = \frac{1}{\theta} I(0 \leqslant x \leqslant \theta).$$

The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[\frac{1}{\theta} I(0 \leqslant x_i \leqslant \theta)\right] = \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \leqslant x_i \leqslant \theta) = \frac{1}{\theta^n} I(\theta \geqslant \max\{x_1, \cdots, x_n\}).$$

Example 3.16 (Geometric Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent Geometric(θ) r.v.s:

$$f(x;\theta) = \theta(1-\theta)^x \text{ for } x = 0, 1, \cdots,$$

where $0 < \theta < 1$. The likelihood and log-likelihood functions are

$$\mathcal{L}(\theta) = \prod_{i=1}^{n}[\theta(1-\theta)^{x_i}]$$

and

$$\ln \mathcal{L}(\theta) = n\ln(\theta) + \ln(1-\theta)\sum_{i=1}^{n} x_i.$$

Thus we get the likelihood equation

$$\frac{n}{\widehat{\theta}} - \frac{1}{1-\widehat{\theta}}\sum_{i=1}^{n} x_i = 0.$$

Since $\sum_{i=1}^{n} x_i > 0$ and the likelihood equation has a unique solution

$$\widehat{\theta} = \frac{1}{1+\overline{x}}.$$

Since

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln \mathcal{L}(\theta) = -\frac{n}{\theta^2} - \frac{1}{(1-\theta)^2}\sum_{i=1}^{n} x_i < 0,$$

the MLE is $\widehat{\theta} = (1+\overline{x})^{-1}$.

We can estimate the standard error of $\widehat{\theta}$ by

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln \mathcal{L}(\widehat{\theta})\right]^{-1/2} = \left[\frac{\overline{x}}{n(1+\overline{x})^3}\right]^{1/2}.$$

Example 3.17 (Exponential Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent exponential random variables with PDF

$$f(x;\lambda) = \lambda\exp(-\lambda x) \text{ for } x \geqslant 0,$$

where $\lambda > 0$ is unknown. The log-likelihood function is

$$\ln \mathcal{L}(\lambda) = n\ln(\lambda) - \lambda\sum_{i=1}^{n} x_i.$$

We get the likelihood equation

$$\frac{n}{\widehat{\lambda}} - \sum_{i=1}^{n} x_i = 0$$

and note that the second derivative is $-\frac{n}{\lambda^2} < 0$. Thus the MLE is $\widehat{\lambda} = \frac{1}{\overline{x}}$.

The observed Fisher information is $\frac{n}{\widehat{\lambda}^2}$ and we obtain the standard error estimator

$$\widehat{\mathrm{SE}}(\widehat{\lambda}) = \left(\frac{n}{\widehat{\lambda}^2}\right)^{-1/2} = \frac{\widehat{\lambda}}{\sqrt{n}}.$$

Example 3.18 (Cauchy Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent Cauchy r.v.s. with unknown location θ :

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

The log-likelihood function is

$$\ln \mathcal{L}(\theta) = -\sum_{i=1}^{n} \ln(1+(x_i-\theta)^2) - n\ln(\pi).$$

The likelihood equation is

$$\sum_{i=1}^{n} \frac{2(x_i-\widehat{\theta})}{1+(x_i-\widehat{\theta})^2} = 0.$$

We observe that $\widehat{\theta}$ cannot be computed in closed-form - we need to numerically evaluate it for a given sample and the likelihood equation can have multiple solutions.

Definition 3.7. We define a sequence $\{\widehat{\theta}_k\}$ s.t.

$$\widehat{\theta}_{k+1} = \widehat{\theta}_k + \frac{S(\widehat{\theta}_k)}{I(\widehat{\theta}_k)},$$

where $S(\theta)$ is the first derivative of the log-likelihood (called the ***score function***) and $I(\theta)$ is the negative second derivative of the log-likelihood. We call this ***Newton-Raphson algorithm***.

We can use N-R algorithm to converge to the solution maximizing the likelihood function: choose a good initial estimator $\widehat{\theta}_0$. If $\widehat{\theta}_0$ is good then $\widehat{\theta}_1 \approx \widehat{\theta}$ and we call it one-step N-R estimator.

3.5.3 Approximate MLEs Using MoM Estimators

MLEs are optimal in the sense of having minimal asymptotic variance. However in may cases (e.g., Cauchy distribution), we do not have a simple formula for the MLE: MLE can only be evaluated numerically via some iterative algorithm and we inevitably need initial estimates for algorithms.

MoM estimates are very useful as initial estimates. For example, the one-step N-R estimator

$$\widehat{\theta}_1 = \widehat{\theta}_0 + \frac{S(\widehat{\theta}_0)}{I(\widehat{\theta}_0)}$$

has the same properties as the exact MLE.

Property 3.1. The MLE $\widehat{\theta}_n$ maximizes

$$\phi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right),$$

and for each $\theta \in \Theta$,

$$\phi_n(\theta) \xrightarrow{p} \phi(\theta) = \mathbb{E}_{\theta_0}\left[\ln\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)\right],$$

which is maximized at $\theta = \theta_0$.

Proof. Consider the simple case where $X_1, \cdots, X_n$ are independent with PDF or PMF $f(x;\theta)$ and assume that the true value of θ generating the data is θ_0. The MLE $\widehat{\theta} = \widehat{\theta}_n$ maximizes the log-likelihood function $\ln \mathcal{L}(\theta)$ over $\theta \in \Theta$ and also maximizes

$$\frac{1}{n}\left[\ln \mathcal{L}(\theta) - \ln \mathcal{L}(\widehat{\theta}_0)\right] = \frac{1}{n}\sum_{i=1}^{n}[\ln f(X_i;\theta) - \ln f(X_i;\theta_0)] = \frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right).$$

Now we apply the WLLN for each $\theta \in \Theta$:

$$\phi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) \xrightarrow{p} \mathbb{E}_{\theta_0}\left[\ln\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)\right] = \phi(\theta).$$

We know $\widehat{\theta}_n$ maximizes $\phi_n(\theta)$.

We can use Jensen's Inequality to show that $0 = \phi(\theta_0) > \phi(\theta), \forall \theta \neq \theta_0$: (now suppose $f(x;\theta)$ is a PDF) since $g(x) = \ln(x)$ is strictly concave and for $\theta \neq \theta_0$, we have

$$\begin{aligned}\phi(\theta) &= \mathbb{E}_{\theta_0}\left[\ln\left(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right)\right] < \ln\left(\mathbb{E}_{\theta_0}\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right]\right) \\ &= \ln\left(\int_{-\infty}^{\infty}\frac{f(x;\theta)}{f(x;\theta_0)}f(x;\theta_0)\mathrm{d}x\right) = \ln\left(\int_{-\infty}^{\infty}f(x;\theta)\mathrm{d}x\right) = 0.\end{aligned}$$

□

Note that it does not imply that $\widehat{\theta}_n \xrightarrow{p} \theta_0$ and we need more regularity conditions on $f(x;\theta)$. The point-wise convergence of $\phi_n(\theta)$ to $\phi(\theta)$ is not strong enough. we need a more uniform convergence tu guarantee consistency, i.e., $\widehat{\theta}_n \xrightarrow{p} \theta_0$. The two-sided Gamma $\left(\frac{1}{2}\right)$ distribution is one such example.

3.5.4 Approximate/Asymptotic Normality of MLEs

If θ_0 is the true parameter value then we can prove consistency, i.e., $\widehat{\theta}_n \xrightarrow{p} \theta_0$. We can approximate the sampling distribution of $\widehat{\theta}_n$ by a normal distribution:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right) \text{ and } \widehat{\theta}_n \dot{\sim} \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right),$$

where $I(\theta)$ can be computed from $f(x;\theta)$.

3.5.4.1 One-Parameter Exponential Families

Assume that $f(x;\theta)$ has the form

$$f(x;\theta) = \exp[\theta T(x) - d(\theta) + h(x)] \text{ for } x \in A.$$

The log-likelihood function is

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^{n}[\theta T(x_i) - d(\theta)_h(x_i)]$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ln \mathcal{L}(\theta) = \sum_{i=1}^{n}[T(x_i) - d'(\theta)].$$

The MLE $\widehat{\theta}_n$ therefore satisfies

$$\frac{1}{n}\sum_{i=1}^{n} T(x_i) = d'(\widehat{\theta}_n),$$

where $\widehat{\theta}_n$ is a method of moments estimator.

Property 3.2. $\mathbb{E}_\theta[T(X_i)] = d'(\theta)$ and $\mathrm{Var}_\theta[T(X_i)] = d''(\theta)$.

Proof. Assume $f(x;\theta)$ is a PDF then we can use the moment generating function of $T(X_i)$:

$$\begin{aligned} m_\theta(s) &= \mathbb{E}_\theta[\exp(sT(X_i))] \\ &= \int_A \exp[sT(x)]\exp[\theta T(x) - d(\theta) + h(x)]\mathrm{d}x \\ &= \int_A \exp[(\theta+s)T(x) - d(\theta) + h(x)]\mathrm{d}x \\ &= \exp[d(\theta+s) - d(\theta)]. \end{aligned}$$

If $\theta + s \in \Theta$,

$$\int_A \exp[(\theta+s)T(x) - d(\theta+s) + h(x)]\mathrm{d}x = 1.$$

We have

$$\begin{aligned} \mathbb{E}_\theta[T(X_i)] &= m'_\theta(0) = d'(\theta) \\ \mathrm{Var}_\theta[T(X_i)] &= m''_\theta(0) - [m'_\theta(0)]^2 = d''(\theta). \end{aligned}$$

□

The CLT therefore gives

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T(X_i) - d'(\theta)\right) \xrightarrow{d} \mathcal{N}(0, d''(\theta)).$$

Define g to be the inverse of d' so that $g(d'(\theta)) = \theta$. The derivative of g is

$$g'(x) = \frac{1}{d''(g(x))} \Rightarrow g'(d'(\theta)) = \frac{1}{d''(\theta)}.$$

Applying the Delta method, we get

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \underbrace{[g'(d'(\theta))]^2 d''(\theta)}_{1/d''(\theta)}).$$

Define $l(x;\theta) = \ln f(x;\theta)$ and $l'(x)$ and $l''(x;\theta)$ to be first two partial derivatives w.r.t. θ, then

$$\begin{aligned} \mathrm{Var}_\theta[l'(X_i;\theta)] &= \mathrm{Var}_\theta[T(X_i)] = d''(\theta) \\ -\mathbb{E}_\theta[l''(X_i;\theta)] &= d''(\theta) = I(\theta). \end{aligned}$$

We have

$$\widehat{\theta} \dot{\sim} \mathcal{N}\left(\theta, \frac{1}{nd''(\theta)}\right)$$

and thus

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = [nd''(\widehat{\theta})]^{-1/2}.$$

Note that

$$nd''(\widehat{\theta}) = -\frac{\mathrm{d}}{\mathrm{d}\theta^2} \ln \mathcal{L}(\widehat{\theta}),$$

which is the observed Fisher information.

An alternative form for a one-parameter exponential family is

$$f(x;\theta) = \exp[c(\theta)T(x) - d(\theta) + h(x)] \text{ for } x \in A.$$

Under the parametrization we have

$$\begin{aligned}
\mathbb{E}_\theta[T(X_i)] &= \frac{d'(\theta)}{c'(\theta)} \\
\mathrm{Var}_\theta[T(X_i)] &= \frac{d''(\theta)c'(\theta) - c''(\theta)d'(\theta)}{[c'(\theta)]^2}.
\end{aligned}$$

As before we have

$$\begin{aligned}
\mathrm{Var}_\theta[l'(X_i;\theta)] &= [c'(\theta)]^2 \mathrm{Var}_\theta[T(X_i)] \\
&= \frac{d''(\theta)c'(\theta) - c''(\theta)d'(\theta)}{c'(\theta)} \\
&= d''(\theta) - c''(\theta)\mathbb{E}_\theta[T(X_i)] = -\mathbb{E}_\theta[l''(X_i;\theta)].
\end{aligned}$$

If $X_1, \cdots, X_n$ are independent from this one-parameter exponential family then the MLE $\widehat{\theta}_n$ satisfies

$$c'(\widehat{\theta}_n)\sum_{i=1}^n T(X_i) = nd'(\widehat{\theta}_n) \text{ or } \frac{1}{n}\sum_{i=1}^n T(X_i) = \frac{d'(\widehat{\theta}_n)}{c'(\widehat{\theta}_n)}.$$

Applying the Delta method, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right),$$

where

$$I(\theta) = -\mathbb{E}_\theta[l''(X_i;\theta)] = \mathrm{Var}_\theta[l'(X_i;\theta)] = \frac{d''(\theta)c'(\theta) - c''(\theta)d'(\theta)}{c'(\theta)}.$$

Example 3.19 (Poisson Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent Poisson r.v.s. with

$$f(x;\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!} = \exp[x\ln(\lambda) - \lambda - \ln(x!)]$$

for $x = 0, 1, 2, \cdots$. The MLE is $\widehat{\lambda} = \overline{X}$ with $\mathrm{Var}_\lambda(\widehat{\lambda}) = \frac{\lambda}{n}$, which gives $\widehat{\mathrm{SE}}(\widehat{\lambda}) = \sqrt{\widehat{\lambda}/n}$. Since $d''(\lambda) = 0$, we have

$$I(\lambda) = -\frac{c''(\lambda)d'(\lambda)}{c'(\lambda)} = \frac{1}{\lambda}.$$

Thus

$$\widehat{\mathrm{SE}}(\widehat{\lambda}) = [nI(\widehat{\lambda})]^{-1/2}.$$

3.5.4.2 Generalizing from Exponential Families

We can generalize from exponential families: We start by assuming that $\widehat{\theta}_n$ is close to the true parameter value (θ_0). We then expand the likelihood equation in a Taylor series around θ_0 and can approximate $\widehat{\theta}_n - n\theta_0$ by an average of n r.v.s. with mean 0 and finite variance.

Suppose $X_1, \cdots, X_n$ are independent r.v.s. with $f(x;\theta)$. We expand the likelihood equation around the true parameter value θ_0 :

$$\begin{aligned} 0 &= \sum_{i=1}^{n} l'(X_i; \widehat{\theta}_n) \\ &= \sum_{i=1}^{n} l'(X_i; \theta_0) + (\widehat{\theta}_n - \theta_0) \sum_{i=1}^{n} l''(X_i; \theta_0) + \cdots . \end{aligned}$$

Then

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(X_i; \theta_0) + [\sqrt{n}(\widehat{\theta}_n - \theta_0)] \frac{1}{n} \sum_{i=1}^{n} l''(X_i; \theta_0) + \underbrace{\sqrt{n}(\widehat{\theta}_n - \theta_0) R_n}_{\text{Remainder}},$$

where for some θ_n^* between $\widehat{\theta}_n$ and θ_0 s.t.

$$R_n = \frac{1}{2}(\widehat{\theta}_n - \theta_0) \frac{1}{n} \sum_{i=1}^{n} l'''(X_i; \theta_n^*).$$

Thus,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left(-\frac{1}{n} \sum_{i=1}^{n} l''(X_i; \theta_0) - R_n \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(X_i; \theta_0).$$

Suppose we know that $\widehat{\theta}_n \xrightarrow{p} \theta_0$ and $|l'''(x;\theta)| \leqslant M(x), \forall \theta$ with $|\theta - \theta_0| < \varepsilon$ for some $\varepsilon > 0$ and $\mathbb{E}_{\theta_0}[M(X_i)]$ finite, then

$$R_n = \frac{1}{2} \underbrace{(\widehat{\theta}_n - \theta_0)}_{\xrightarrow{p} 0} \underbrace{\frac{1}{n} \sum_{i=1}^{n} l'''(X_i; \theta_n^*)}_{\text{Bounded}} \xrightarrow{p} 0.$$

Also,

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(X_i; \theta_0) &\xrightarrow{d} \mathcal{N}(0, I(\theta_0)) \\ -\frac{1}{n} \sum_{i=1}^{n} l''(X_i; \theta_0) &\xrightarrow{p} I(\theta_0). \end{aligned}$$

Therefore,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{l'(X_i; \theta_0)}{I(\theta_0)} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right),$$

where $\text{Var}_\theta[l'(X_i;\theta)] = -\mathbb{E}_\theta[l''(X_i;\theta)] = I(\theta)$ and $\mathbb{E}_\theta[l'(X_i;\theta)] = 0$.

Hence, for model we assume, if it is approximately correct then the MLE $\widehat{\theta}$ is approximately normal with mean θ and variance $\frac{1}{nI(\theta_0)}$. We can estimate its standard error

$$\widehat{\text{SE}}(\widehat{\theta}) = \left(-\sum_{i=1}^{n} l''(X_i; \widehat{\theta}) \right)^{-1/2}.$$

We can base CIs for θ on the approximate pivot

$$\frac{\widehat{\theta}-\theta}{\widehat{\mathrm{SE}}(\widehat{\theta})} \dot{\sim} \mathcal{N}(0,1).$$

Example 3.20 (Fisher-von Mises Distribution)**.** Suppose $X_1, \cdots, X_n$ are r.v.s. on the interval $[0, 2\pi)$ with

$$f(x;\kappa,\mu) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa\cos(x-\mu)) \text{ for } 0 \leqslant x < 2\pi$$

where $I_0(\kappa)$ is a 0-th order modified Bessel function of the first kind. $X_1, \cdots, X_n$ are random directions: μ represents the mean direction, κ is a concentration parameter and we assume $\kappa > 0$ is known and μ is unknown. The log-likelihood function is

$$\ln \mathcal{L}(\mu) = \kappa \sum_{i=1}^{n} \cos(x_i - \mu) + C.$$

The likelihood equation is

$$\kappa \sum_{i=1}^{n} \sin(x_i - \widehat{\mu}) = 0.$$

There are always two solutions to the likelihood equation, corresponding to the global maximum and minimum. The MLE $\widehat{\mu}$ satisfies

$$\tan(\widehat{\mu}) = \frac{\sum_{i=1}^{n} \sin(X_i)}{\sum_{i=1}^{n} \cos(X_i)}.$$

$\widehat{\mu}$ maximizes the log-likelihood if $\cos(\widehat{\mu})$ and $\sin(\widehat{\mu})$ have the same signs as $\sum_{i=1}^{n} \cos(X_i)$ and $\sum_{i=1}^{n} \sin(X_i)$, respectively.

3.5.4.3 Bartlett Identities

Suppose that $f(x;\theta)$ is a PDF (same argument works for PMF) and define

$$A = \{x : f(x;\theta) > 0\},$$

which we assume to be independent of θ. Thus for all $\theta \in \Theta$,

$$\int_A f(x;\theta)\mathrm{d}x = 1$$

and so for any integer $k \geqslant 1$,

$$\frac{\mathrm{d}^k}{\mathrm{d}\theta^k} \int_A f(x;\theta)\mathrm{d}x = 0.$$

Then

$$\int_A \frac{\partial^k}{\partial\theta^k} f(x;\theta)\mathrm{d}x = 0.$$

Definition 3.8. The ***Bartlett identities*** follow from:

- For $k = 1$,

$$0 = \int_A \frac{\partial}{\partial\theta} f(x;\theta)\mathrm{d}x = \int_A \underbrace{\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}}_{l'(x;\theta)} f(x;\theta)\mathrm{d}x$$

or

$$\mathbb{E}_\theta[l'(X_i;\theta)] = 0.$$

- For $k = 2$,

$$0 = \int_A \frac{\partial}{\partial\theta}[l'(x;\theta)f(x;\theta)]\mathrm{d}x = \int_A l''(x;\theta)f(x;\theta)\mathrm{d}x + \int_A [l'(x;\theta)]^2 f(x;\theta)\mathrm{d}x$$

or

$$\mathrm{Var}_\theta[l'(X_i;\theta)] = -\mathbb{E}_\theta[l''(X_i;\theta)].$$

Example 3.21. Suppose that $\mathbb{E}_\theta[g(X_i)] = \psi(\theta)$ where $\psi(\theta)$ is continuous and differentiable. We have

$$\begin{aligned}
\psi'(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_A g(x)f(x;\theta)\mathrm{d}x \\
&= \int_A g(x)\frac{\partial}{\partial\theta} f(x;\theta)\mathrm{d}x \\
&= \int_A g(x)l'(x;\theta)f(x;\theta)\mathrm{d}x \\
&= \mathbb{E}_\theta[g(X_i)l'(X_i;\theta)].
\end{aligned}$$

Since $\mathbb{E}_\theta[l'(X_i;\theta)] = 0$, we have

$$\psi'(\theta) = \mathrm{Cov}_\theta(g(X_i), l'(X_i;\theta)).$$

3.5.4.4 Confidence Intervals from Log-Likelihood

A simple $100p\%$ CI for θ is $[\widehat{\theta} \mp z_p\widehat{\mathrm{SE}}(\widehat{\theta})]$, where $\widehat{\mathrm{SE}}(\widehat{\theta})$ is based on the observed Fisher information:

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \left[-\sum_{i=1}^n l''(X_i;\widehat{\theta})\right]^{-1/2}.$$

Alternatively, we can use the approximate pivot

$$2[\ln\mathcal{L}(\widehat{\theta}) - \ln\mathcal{L}(\theta)] \approx nI(\theta)(\widehat{\theta}-\theta)^2 \dot{\sim} \chi^2(1).$$

Thus an approximate $100p\%$ CI for θ is

$$\{\theta : 2[\ln\mathcal{L}(\widehat{\theta}) - \ln\mathcal{L}(\theta)] \leqslant q_p\},$$

where q_p is the p quantile of the $\chi^2(1)$ distribution.

Example 3.22 (Exponential Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent Exponential with

$$f(x;\lambda) = \lambda\exp(-\lambda x),$$

where $x \geqslant 0$ and $\lambda > 0$. The MLE $\widehat{\lambda} = \frac{1}{\overline{X}}$ and

$$\ln\mathcal{L}(\lambda) = n\ln(\lambda) - \lambda\sum_{i=1}^n X_i.$$

Thus an approximate 95% CI for λ is

$$\{\lambda : 2n[\lambda\overline{X} - \ln(\lambda\overline{X}) - 1] \leqslant 3.841\}.$$

Given the value of $\overline{X}$ we need to solve the equations $2n[\lambda\overline{X} - \ln(\lambda\overline{X}) - 1] = 3.841$ to determine the endpoints.

3.5.5 Misspecified Model

Suppose $X_1, \cdots, X_n$ are independent with PDF or PMF $f(x;\theta)$ for some real-valued $\theta \in \Theta$. But the true PDF or PMF is $g(x)$ where $g(x) \neq f(x;\theta), \forall\theta$, we call it misspecified model. Ideally $g(x) \approx f(x;\theta_0)$ for some θ_0. We define $\widehat{\theta}$ as the solution of the likelihood equation.

3.5.5.1 Defining θ_0

We define θ_0 to be the value of θ maximizing

$$\phi(\theta) = \mathbb{E}_g\left[\ln\left(\frac{f(X_i;\theta)}{g(X_i)}\right)\right].$$

Typically, θ_0 satisfies $\mathbb{E}_g[l'(X_i;\theta_0)] = 0$.

3.5.5.2 Approximate Normality

Expanding a first-order Taylor series around θ_0 and ignoring the remainder term gives

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx \left(-\frac{1}{n}\sum_{i=1}^{n} l''(X_i;\theta_0)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l'(X_i;\theta_0).$$

Now we have

$$-\frac{1}{n}\sum_{i=1}^{n} l''(X_i;\theta_0) \xrightarrow{p} I_g(\theta_0) = -\mathbb{E}_g[l''(X_i;\theta_0)]$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l'(X_i;\theta_0) \xrightarrow{d} \mathcal{N}(0, J_g(\theta_0) = \mathrm{Var}_g[l'(X_i;\theta_0)]),$$

which suggests that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \dot{\sim} \mathcal{N}\left(0, \frac{J_g(\theta_0)}{I_g^2(\theta_0)}\right),$$

i.e.,

$$\widehat{\theta} \dot{\sim} \mathcal{N}\left(\theta_0, \frac{J_g(\theta_0)}{nI_g^2(\theta_0)}\right).$$

3.5.5.3 Estimating Standard Errors

We can estimate $J_g(\theta_0)$ and $I_g(\theta_0)$:

$$\widehat{J}_g = \frac{1}{n-1}\sum_{i=1}^{n}[l'(X_i;\widehat{\theta})]^2$$

$$\widehat{I}_g = -\frac{1}{n}\sum_{i=1}^{n} l''(X_i;\widehat{\theta})$$

and thus we can estimate the standard error of $\widehat{\theta}$ by

$$\widehat{\mathrm{SE}}(\widehat{\theta}) = \left(\frac{\widehat{J}_g}{n\widehat{I}_g^2}\right)^{1/2},$$

which is called ***sandwich estimator*** that named after the case of k parameters, we sometimes estimate the variance-covariance matrix of an estimator by

$$\frac{1}{n}\widehat{I}_g^{-1}\widehat{J}_g\widehat{I}_g^{-1}$$

where $\widehat{I}_g^{-1}$ and $\widehat{J}_g$ are $k \times k$ matrices. The sandwich estimator turns out to be very closely related to the jackknife estimator. If the parametric model is correct or almost correct then $\widehat{J}_g \approx \widehat{I}_g$.

3.5.6 Limiting Distributions of MLEs

To summarize, suppose $X_1, \cdots, X_n$ are independent with PDF or PMF $f(x;\theta)$ for some real-valued $\theta \in \Theta$, where Θ is an open set, $A = \{x : f(x;\theta) > 0\}$ does not depend on θ, and $l(x;\theta)$ is three times differentiable w.r.t. θ for each $x \in A$. If θ_0 is the true parameter value then

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &\approx \left(-\frac{1}{n}\sum_{i=1}^{n} l''(X_i;\theta_0)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l'(X_i;\theta_0) \\ &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{l'(X_i;\theta_0)}{I(\theta_0)} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right),\end{aligned}$$

where $I(\theta_0) = \mathrm{Var}_{\theta_0}[l'(X_i;\theta_0)] = -\mathbb{E}_{\theta_0}[l''(X_i;\theta_0)]$.

Now suppose $\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$. If $\sigma^2(\theta) < \frac{1}{I(\theta)}$, it is easy to construct estimators $\widetilde{\theta}_n$ with $\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} Z(\theta)$ and $\mathrm{Var}_\theta[Z(\theta)] < \frac{1}{I(\theta)}$ for some $\theta \in \Theta$, which is called super-efficiency.

Example 3.23. Suppose $X_1, \cdots, X_n$ are independent $\mathcal{N}(\mu, 1)$ r.v.s., and the MLE is $\widehat{\mu}_n = \overline{X}_n \sim \mathcal{N}\left(\mu, \frac{1}{n}\right)$. Define

$$\widetilde{\mu}_n = \begin{cases}\overline{X}_n, & |\overline{X}_n| > \frac{1.96}{\sqrt{n}} \\ 0, & |\overline{X}_n| \leqslant \frac{1.96}{\sqrt{n}}\end{cases}.$$

The limiting distribution of $\sqrt{n}(\widetilde{\mu} - \mu)$ depends on μ :

$$\sqrt{n}(\widetilde{\mu} - \mu) \xrightarrow{d} \begin{cases} Z, & \mu \neq 0 \\ ZI(|Z| > 1.96), & \mu = 0\end{cases},$$

where $Z \sim \mathcal{N}(0,1)$. And $\mathrm{Var}[ZI(|Z| > 1.96)] = \mathbb{E}[Z^2 I(|Z| > 1.96)] = 0.279 < 1$.

We are able to reduce the variance of the limiting distribution for $\mu = 0$. For $\mu \neq 0$, the variance of the limiting distribution is the same as that of the MLE. The type of estimation is common in practice. For example, in a regression modeling $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$, we set certain $\{\beta_j\}$ equal to 0 based on t statistics. Another example is image denoising or compression:

$$\underbrace{X(i,j)}_{\text{Image}} = \sum_{k=1}^{mn} \alpha_k \underbrace{B_k(i,j)}_{\text{Basis}} \text{ for } i = 1, \cdots, m; j = 1, \cdots, n,$$

we set certain $\{\alpha_k\}$ to 0 based on some rules.

3.5.7 MoMs versus MLEs

Suppose $X_1, \cdots, X_n$ are independent with PDF or PMF $f(x;\theta), \theta$ is real-valued and $f(x;\theta)$ satisfies the regularity conditions for asymptotic normality of the MLE. Suppose that $\mathbb{E}_\theta[g(X_i)] = \psi(\theta)$ where ψ is a injective function (i.e., ψ^{-1} is well-defined).

With MoMs,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = \psi(\widetilde{\theta}_n) \Rightarrow \widetilde{\theta}_n = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right).$$

Property 3.3. $\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ with $\sigma^2(\theta) \geqslant \frac{1}{I(\theta)}$.

Proof. From the CLT, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \psi(\theta)\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}_\theta[g(X_i)]).$$

Applying the Delta method,

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(\theta) = \frac{\mathrm{Var}_\theta[g(X_i)]}{[\psi'(\theta)]^2}\right)$$

since the derivative of $\psi^{-1}(x)$ is $\frac{1}{\psi'(\psi^{-1}(x))}$ and thus

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \approx \frac{\sqrt{n}}{\psi'(\theta)}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \psi(\theta)\right)$$

For the MLE and MoM estimator, we now have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \approx \frac{1}{I(\theta)\sqrt{n}}\sum_{i=1}^{n} l'(X_i;\theta)$$
$$\sqrt{n}(\widetilde{\theta}_n - \theta) \approx \frac{1}{\psi'(\theta)\sqrt{n}}\sum_{i=1}^{n}[g(X_i) - \psi(\theta)],$$

which follows that

$$\sqrt{n}\begin{pmatrix}\widehat{\theta}_n - \theta \\ \widetilde{\theta}_n - \theta\end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, C(\theta))$$

where $C(\theta)$ is a 2×2 covariance matrix. We have

$$C(\theta) = \begin{pmatrix}\frac{1}{I(\theta)} & \eta(\theta) \\ \eta(\theta) & \sigma^2(\theta)\end{pmatrix},$$

where

$$\eta(\theta) = \mathrm{Cov}_\theta\left(\frac{l'(X_i;\theta)}{I(\theta)}, \frac{g(X_i) - \psi(\theta)}{\psi'(\theta)}\right) = \frac{1}{I(\theta)\psi'(\theta)}\mathrm{Cov}_\theta(g(X_i), l'(X_i;\theta)).$$

Note that $\psi(\theta) = \mathbb{E}_\theta[g(X_i)]$ has derivative

$$\psi'(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\int_A g(x)f(x;\theta)\mathrm{d}x = \int_A g(x)\frac{\partial}{\partial\theta}f(x;\theta)\mathrm{d}x$$
$$= \int_A g(x)l'(x;\theta)f(x;\theta)\mathrm{d}x = \mathbb{E}_\theta[g(X_i)l'(X_i;\theta)] = \mathrm{Cov}_\theta(g(X_i), l'(X_i;\theta))$$

and thus $\eta(\theta) = \frac{1}{I(\theta)}$.

By Cauchy-Schwarz inequality, we have $\text{Cov}(U, V)^2 \leqslant \text{Var}[U]\text{Var}[V]$, then

$$[\eta(\theta)]^2 \leqslant \frac{\sigma^2(\theta)}{I(\theta)}.$$

Since $\eta(\theta) = \frac{1}{I(\theta)}$ for any MoM estimator, it follows that

$$\frac{1}{I(\theta)^2} \leqslant \frac{\sigma^2(\theta)}{I(\theta)} \Leftrightarrow \frac{1}{I(\theta)} \leqslant \sigma^2(\theta).$$

□

To summarize, if $\widehat{\theta}_n$ is an MLE and $\widetilde{\theta}$ is a MoM estimator of θ then

$$\text{Var}_\theta(\widehat{\theta}_n) \approx \frac{1}{nI(\theta)} \leqslant \frac{\sigma^2(\theta)}{n} \approx \text{Var}_\theta(\widetilde{\theta}_n)$$

and

$$\text{Cov}_\theta(\widehat{\theta}_n, \widetilde{\theta}_n) \approx \frac{1}{nI(\theta)}.$$

The result can be extended beyond MoM estimators: If $\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ and $\sigma^2(\theta)$ is a continuous function of θ over Θ then $\sigma^2(\theta) \geqslant \frac{1}{I(\theta)}$ for all $\theta \in \Theta$. The lower bound $\frac{1}{I(\theta)}$ can be attained for non-MLEs, for example, one step Newton-Raphson estimators or linear combinations of order statistics.

Example 3.24 (Exponential Goodness-of-Fit)**.** Suppose we observe positive data $x_1, \cdots, x_n$ and we want to check if an exponential model for these data is appropriate. We can use exponential Q-Q plot, MoM approach, or compare the estimated CDF for the exponential model (using the MLE $\lambda = \frac{1}{\overline{X}}$) to the empirical distribution function

$$F(x; \widehat{\lambda}) = 1 - \exp(-\widehat{\lambda}x) \text{ versus } \widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x).$$

If the exponential model is appropriate, then $F(x; \widehat{\lambda})$ should be close to $\widehat{F}(x)$.

To compare $F(x; \widehat{\lambda})$ and $\widehat{F}(x)$, we first look at

$$\text{Var}_\lambda[\widehat{F}(x - F(x; \widehat{\lambda}))] = \text{Var}_\lambda[\widehat{F}(x)] + \text{Var}_\lambda[F(x; \widehat{\lambda})] - 2\text{Cov}_\lambda(\widehat{F}(x), F(x; \widehat{\lambda})).$$

Note that $\widehat{\lambda} \dot{\sim} \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$, $F(x; \widehat{\lambda})$ is a differentiable function of $\widehat{\lambda}$ for each x, and $\widehat{F}(x)$ is a sample mean for each x. We have

$$\text{Var}_\lambda[\widehat{F}(x)] = \frac{\exp(-\lambda x)(1 - \exp(-\lambda x))}{n}.$$

We can approximate $\text{Var}_\lambda[F(x; \widehat{\lambda})]$ using Delta method with $g(\lambda) = 1 - \exp(-\lambda x)$:

$$\text{Var}_\lambda[F(x; \widehat{\lambda})] \approx \frac{\lambda^2 x^2 \exp(-2\lambda x)}{n},$$

where $x^2 \exp(-2\lambda x) = [g'(\lambda)]^2$.

Since $F(x;\widehat{\lambda})$ is an MLE of $F(x;\lambda)$ and $\widehat{F}(x)$ is a MoM estimator, we have

$$\mathrm{Cov}_\lambda[\widehat{F}(x), F(x;\widehat{\lambda})] \approx \mathrm{Var}_\lambda[F(x;\widehat{\lambda})] \approx \frac{\lambda^2 x^2 \exp(-2\lambda x)}{n}.$$

Hence,

$$\begin{aligned}\mathrm{Var}_\lambda[\widehat{F}(x) - F(x;\widehat{\lambda})] &\approx \mathrm{Var}_\lambda[\widehat{F}(x)] - \mathrm{Var}_\lambda[F(x;\widehat{\lambda})] \\ &\approx \frac{1}{n}[\exp(-\lambda x)(1-\exp(-\lambda x)) - \lambda^2 x^2 \exp(-2\lambda x)],\end{aligned}$$

which can be estimated by substituting $\widehat{\lambda}$ for λ.

We now can plot $\widehat{F}(x) - F(x;\widehat{\lambda})$ versus x and put lines at

$$\pm 2\left[\frac{\exp(-\widehat{\lambda}x)(1-\exp(-\widehat{\lambda}x)) - \widehat{\lambda}^2 x^2 \exp(-2\widehat{\lambda}x)}{n}\right]^{1/2}.$$

3.5.8 Monte Carlo Simulation

Some properties of statistical procedures cannot be assessed analytically and we assess deterministic properties using random sampling: We use random sampling to evaluate integrals:

$$\int_{\mathbb{R}^n} h(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})\mathrm{d}\mathbf{x},$$

where $f(\mathbf{x})$ is a joint PDF and $g(\mathbf{x}) = \frac{h(\mathbf{x})}{f(\mathbf{x})}$.

With WLLN, we can approximate the integral by

$$\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{X}_i),$$

where $\mathbf{X}_1, \cdots, \mathbf{X}_N$ are sampled independently from f. The same idea works for sums:

$$\sum_{\mathbf{x}} h(\mathbf{x}) = \sum_{\mathbf{x}} \underbrace{\frac{h(\mathbf{x})}{f(\mathbf{x})}}_{g(\mathbf{x})} f(\mathbf{x}) \approx \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{X}_i),$$

where $f(\mathbf{x})$ is a PMF and $\mathbf{X}_1, \cdots, \mathbf{X}_N$ are sampled independently from f.

In statistical applications, the PDF or PMF $f(\mathbf{x})$ is typically the joint PDF or PMF of r.v.s. $(X_1, \cdots, X_n)$. For example, if $X_1, \cdots, X_n$ are independent r.v.s. from $f(x;\theta)$, then

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta).$$

We may be interested in estimating:

- MSE of an estimator $\widehat{\theta}$:

$$\int_{\mathbb{R}^n} (\widehat{\theta}(\mathbf{x}) - \theta)^2 f(\mathbf{x})\mathrm{d}\mathbf{x}.$$

- The CDF of an estimator $\widehat{\theta}$:
$$\int_{\mathbb{R}^n} I[\widehat{\theta}(\mathbf{x}) \leqslant t] f(\mathbf{x}) \mathrm{d}\mathbf{x}.$$
- The coverage of a CI:
$$\int_{\mathbb{R}^n} I[l(\mathbf{x}) \leqslant \theta \leqslant u(\mathbf{x})] f(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Note that when doing a Monte Carlo simulation, we are carrying out a designed experiment: we should employ principles of experimental design (e.g., blocking) in setting up the simulation; the results of the simulation can be analyzed statistically as for any other designed experiment. We can and should use known information about the model we are simulating in the design of the simulation and in the analysis of the results.

Example 3.25. We now estimate the variance of $\widehat{\theta}$. Suppose we have another r.v. T where the distribution of T is known exactly and T is independent of $T - \widehat{\theta}$. Then we have
$$\mathrm{Var}_\theta(\widehat{\theta}) = \underbrace{\mathrm{Var}_\theta(T)}_{\text{Known}} + \underbrace{\mathrm{Var}_\theta(T - \widehat{\theta})}_{\text{Estimate}}.$$

Example 3.26 (Normal Mean Estimation)**.** Suppose $X_1, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ r.v.s., the MLE of μ is $\widehat{\mu} = \overline{X}$. We can use the symmetry of the Normal distribution to propose other estimators: sample median or trimmed means.

Recall: Suppose that $T(\mathbf{X}) = T(X_1, \cdots, X_n)$ satisfies $T(X_1+a, \cdots, X_n+a) = T(X_1, \cdots, X_n)+a, \forall a$, then $\overline{X}$ and $T(\mathbf{X}) - \overline{X}$ are independent. Estimators such as the sample median and trimmed mean satisfy this condition.

We want to approximate the sampling distribution of $T(\mathbf{X})$ and assume $\mu = 0, \sigma^2$. Note that
$$\begin{aligned} P(T(\mathbf{X}) \leqslant x) &= P(T(\mathbf{X}) - \overline{X} + \overline{X} \leqslant x) = P(T(\mathbf{X}) - \overline{X} \leqslant x - \overline{X}) \\ &= \int_{-\infty}^{\infty} \underbrace{P(T(\mathbf{X}) - \overline{X} \leqslant x - t)}_{\text{Estimate by MC}} \underbrace{\sqrt{n}\phi(\sqrt{n}t)}_{\text{Density of } \overline{X}} \mathrm{d}t \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \Phi(\sqrt{n}(x - D_i)), \end{aligned}$$
where $D_i = T(\mathbf{X}_i) - \overline{X}_i$. The density approximation will be a kernel density estimate using the differences with a Gaussian kernel with known bandwidth $\frac{1}{\sqrt{n}}$.

Example 3.27 (Trimmed Mean with Cauchy Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent Cauchy r.v.s. with PDF
$$f(x; \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}.$$
We estimate θ by a trimmed mean
$$\widehat{\theta}(r) = \frac{1}{n - 2r} \sum_{k=r+1}^{n-r} X_{(k)}.$$
We know the distribution of $\widehat{\theta}(r)$ is symmetric around θ and the MSE depends on n and r but not θ. For a given value of n, we can find the value of r that minimizes the MSE $\mathbb{E}_\theta[(\widehat{\theta}(r) - \theta)^2]$:

```
means = NULL
for (i in 1:10000){
   x = rcauchy(100)
   m = NULL
   for (j in 1:10){
      m = c(m, mean(x, trim = j / 20))
      }
   means = rbind(means, m)
   }
boxplot(means, names = c(1:10) / 20, xlab = "r/n", col = "lightblue")
MSE = NULL
for (i in 1:10){
   MSE = c(MSE, mean(means[, i]^2))
   }
```

Example 3.28 (Estimating Gini Index)**.** For data coming from a Gamma distribution, the estimator based on the MLE of α in the Gamma model is the best estimator of the three although the non-parametric estimator performs very well: MLE should dominate all estimators when the Gamma model is true.

For data coming from a Weibull distribution, the non-parametric estimator is the best, the other two estimators have non-trivial bias (in opposite directions).

3.5.9 General MLE

The proof of asymptotic normality follows from approximating the score function by a linear function close to the true parameter value and the argument also applies more or less even if the model is misspecified with appropriate modifications to the variance of the normal distribution.

Suppose $(X_1, \cdots, X_n)$ have joint PDF or PMF $f(x_1, x_n; \theta)$, which can be written as a product of conditional PDFs or PMFs:

$$f(x_1, \cdots, x_n; \theta) = f(x_1; \theta) \prod_{i=2}^{n} f(x_i | x_1, \cdots, x_{i-1}; \theta).$$

The log-likelihood function is

$$\ln \mathcal{L}(\theta) = \ln f(x_1; \theta) + \sum_{i=2}^{n} \ln f(x_i | x_1, \cdots, x_{i-1}; \theta) = l(x_1; \theta) + \sum_{i=2}^{n} l(x_i | x_1, \cdots, x_{i-1}; \theta).$$

The likelihood equation is

$$l'(x_1; \widehat{\theta}) + \sum_{i=2}^{n} l'(x_i | x_1, \cdots, x_{i-1}; \widehat{\theta}) = 0.$$

Property 3.4. If θ_0 is the true parameter value, then the r.v.s. $l'(X_1; \theta_0)$ and $l'(X_i | X_1, \cdots, X_{i-1}; \theta_0)$ (for $i \geqslant 2$) have mean 0 and are uncorrelated.

The approximation for $\widehat{\theta} - \theta_0$ now becomes

$$\widehat{\theta} - \theta_0 \approx - \frac{\sum_{i=1}^{n} l'(X_i | X_1, \cdots, X_{i-1}; \theta_0)}{\sum_{i=1}^{n} l''(X_i | X_1, \cdots, X_{i-1}; \theta_0)}.$$

The numerator is a sum of uncorrelated r.v.s. with mean 0 and thus it should be approximately normal. The variance of the numerator is equal to the negative expected value of the dominator.

Thus

$$\operatorname{Var}(\widehat{\theta}) \approx \left[-\sum_{i=1}^{n} l''(X_i|X_1, \cdots, X_{i-1}; \theta_0) \right]^{-1}$$

and

$$\widehat{\operatorname{SE}}(\widehat{\theta}) = \left[-\sum_{i=1}^{n} l''(X_i|X_1, \cdots, X_{i-1}; \widehat{\theta}) \right]^{-1/2} = \left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln \mathcal{L}(\widehat{\theta}) \right]^{-1/2}.$$

Example 3.29 (Multiple Mark-Recapture Experiments)**.** Suppose finite population with unknown population size $N \in \mathbb{N}$. Recall that simple mark-recapture experiment has two steps: (1) Capture n_0 individuals from the population, which are then marked and released back to the population. (2) Capture n_1 individuals and define M to be the number of marked individuals. M has a hypergeometric distribution:

$$P_N(M = x) = f(x; N) = \frac{\binom{n_0}{x}\binom{N-x_0}{n_1-x}}{\binom{N}{n_1}}.$$

The Lincoln-Petersen estimator is

$$\widehat{N} = \frac{n_0 n_1}{M}.$$

We repeat the simple mark-recapture experiment several times, i.e., capture $n_0, n_1, \cdots, n_k$ individuals at stages $0, 1, \cdots k$ and at each stage, unmarked individuals are marked. Define M_i to be the number of marked individuals captured at stage i. Given $M_1 = m_1, \cdots, M_{i-1} = m_{i-1}$, the number of marked individuals in the population at stage i is

$$t_i = \sum_{j=0}^{i-1} (n_j - m_j) = t_{i-1} + \underbrace{(n_{i-1} - m_{i-1})}_{\text{Unmarked at stage } (i-1)}$$

where $m_0 = 0$.

The joint PMF of $(M_1, \cdots, M_k)$ can be written as a product of conditional PMFs: M_i depends on $M_1, \cdots, M_{i-1}$ for $i \geqslant 2$ and each of the conditional distribution is a hypergeometric distribution:

$$P(M_i = m_i | M_1 = m_1, \cdots, M_{i-1} = m_{i-1}; N) = \frac{\binom{t_i}{m_i}\binom{N-t_i}{n_i-m_i}}{\binom{N}{n_i}}.$$

The likelihood function for N is

$$\mathcal{L}(N) = \prod_{i=1}^{k} \frac{\binom{t_i}{m_i}\binom{N-t_i}{n_i-m_i}}{\binom{N}{n_i}}$$

where $N \geqslant \sum_{i=0}^{k} (n_i - m_i)$. We can estimate the population size by maximizing the likelihood function over N.

We also want to obtain estimates of the uncertainty of the MLE $\widehat{N}$ or CIs for N. Note that the likelihood function provides some information but the model does not satisfy the usual assumptions due to discreteness of N. We then can approximate the hypergeometric distribution by Poisson distribution and replace N by a pseudo-continuous parameter.

Recall that suppose finite population of size N consisting of two disjoint groups A and B consisting of K and $N-K$ elements respectively. We sample n elements at random without replacement and let X be the number of sampled elements in group A. The PMF of X is

$$P(X=k)\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

for $\max(0, n+K-N) \leqslant k \leqslant \min(K,n)$. If we sampled with replacement then X would have a Binomial distribution:

$$P(X=k) = \binom{n}{k}\left(\frac{K}{N}\right)^k\left(1-\frac{K}{N}\right)^{n-k}$$

for $k = 0, \cdots, n$. If N is large and n is small relative to N then sampling without replacement is essentially equivalent to sampling with replacement:

$$P(\text{Duplicates}) = 1 - \prod_{k=1}^{n}\left(1-\frac{1}{N-k+1}\right) \approx 1-\exp\left(-\frac{n}{N}\right) \approx 0.$$

Thus if $\frac{n}{N}$ is small,

$$X \dot{\sim} \text{Binomial}\left(n, \frac{K}{N}\right).$$

Now suppose $\frac{K}{N}$ is small and we can approximate the binomial by a Poisson with the same mean. Thus if $\frac{n}{N}$ and $\frac{K}{N}$ are small,

$$X \dot{\sim} \text{Poisson}\left(\frac{nK}{N}\right).$$

At each stage, the number of marked individuals t_i is small compared to N and thus

$$\frac{\binom{t_i}{m_i}\binom{N-t_i}{n_i-m_i}}{\binom{N}{n_i}} \approx \frac{\exp\left(-\frac{n_i t_i}{N}\right)\left(\frac{n_i t_i}{N}\right)^{m_i}}{m_i!}.$$

We define the parameter $\omega = \frac{1}{N}$, which is a pseudo-continuous parameter with $0 < \omega < 1$. Technically, the upper bound for ω depends on the data. The log-likelihood function is

$$\ln \mathcal{L}(\omega) = \sum_{i=1}^{k}[m_i \ln(n_i t_i \omega) - n_i t_i \omega - \ln(m_i!)].$$

The likelihood equation is

$$\frac{\mathrm{d}}{\mathrm{d}\omega}\ln\mathcal{L}(\omega) = \sum_{i=1}^{k}\left(\frac{m_i}{\widehat{\omega}} - n_i t_i\right) = 0$$

and thus the MLE of ω is

$$\widehat{\omega} = \frac{\sum_{i=1}^{k} M_i}{\sum_{i=1}^{k} n_i t_i}.$$

Hence we obtain the Schnabel estimator

$$\widehat{N} = \frac{1}{\widehat{\omega}} = \frac{\sum_{i=1}^{k} n_i t_i}{\sum_{i=1}^{k} M_i}.$$

We can obtain an estimator of the standard error of $\widehat{\omega}$ from the observed Fisher information:

$$-\frac{\mathrm{d}^2}{\mathrm{d}\omega^2}\ln\mathcal{L}(\widehat{\omega}) = \frac{1}{\widehat{\omega}}\sum_{i=1}^{k} M_i = \frac{\left(\sum\limits_{i=1}^{k} n_i t_i\right)^2}{\sum\limits_{i=1}^{k} M_i}.$$

The standard error estimator is

$$\widehat{\mathrm{SE}}(\widehat{\omega}) = \frac{\widehat{\omega}}{\left(\sum\limits_{i=1}^{k} M_i\right)^{1/2}}$$

and $\widehat{\mathrm{SE}}(\widehat{\omega})$ can be used to obtain CIs for ω and hence for $N = \frac{1}{\omega}$.

3.6 Methods of Estimation: Bayes

Bayesian approach is to quantify a priori information about θ by probability distributions. We can think of $f(x_1, \cdots, x_n; \theta)$ as representing the conditional PDF/PMF of $(X_1, \cdots, X_n)$ given the parameter θ where θ has some probability distribution on Θ :

$$f(x_1, \cdots, x_n; \theta) = f(x_1, \cdots, x_n|\theta).$$

The information about θ given via a ***prior density function*** $\pi(\theta)$ and thus for example, we can think of $\pi(\theta)f(x_1, \cdots, x_n; \theta)$ as the joint PDF of $(X_1, \cdots, X_n, \theta)$.

If the prior density $\pi(\theta)$ quantifies the a priori information about θ before observing the data, the ***posterior density function*** combines the information from the prior with the information from the data via the likelihood function.

Theorem 3.2 (Bayes Theorem)**.** Suppose θ is discrete-valued with prior PMF $\pi(\theta)$ on $\Theta = \{\theta_1, \cdots\}$, then

$$\pi(\theta_j|x_1, \cdots, x_n) = \frac{\pi(\theta_j)f(x_1, \cdots, x_n; \theta_j)}{\sum\limits_{k}\pi(\theta_k)f(x_1, \cdots, x_n; \theta_k)} = c(x_1, \cdots, x_n)\pi(\theta_j)\mathcal{L}(\theta_j)$$

where

$$c(x_1, \cdots, x_n) = \left[\sum_{k}\pi(\theta_k)f(x_1, \cdots, x_n; \theta_k)\right]^{-1}.$$

For a continuous parameter space, the posterior density is

$$\pi(\theta|x_1, \cdots, x_n) = \frac{\pi(\theta)f(x_1, \cdots, x_n; \theta)}{\int_\Theta \pi(s)f(x_1, \cdots, x_n; s)\mathrm{d}s} = c(x_1, \cdots, x_n)\pi(\theta)\mathcal{L}(\theta) \propto \pi(\theta)\mathcal{L}(\theta)$$

where

$$c(x_1, \cdots, x_n) = \left[\int_\Theta \pi(s)f(x_1, \cdots, x_n; s)\mathrm{d}s\right]^{-1}.$$

Note that the shape of the posterior density depends only on $\pi(\theta)\mathcal{L}(\theta)$, the normalizing constant $c(x_1, \cdots, x_n)$ depends on the data. In practice, $c(x_1, \cdots, x_n)$ may be difficult to evaluate.

Example 3.30 (Exponential Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent exponential r.v.s. with

$$f(x; \lambda) = \lambda\exp(-\lambda x), x \geqslant 0$$

where $\lambda > 0$. The likelihood function is

$$\mathcal{L}(\lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

We will put an exponential prior on λ :

$$\pi(\lambda) = \alpha \exp(-\alpha\lambda), \lambda > 0$$

where $\alpha > 0$ is called a ***hyperparameter***. α can be varied according to a priori information about λ.

For the posterior density, we have

$$\pi(\lambda|x_1, \cdots, x_n) \propto \lambda^n \exp\left[-\left(\alpha + \sum_{i=1}^{n} x_i\right)\lambda\right].$$

The form of the posterior density is a Gamma distribution with shape parameter $n + 1$ and rate parameter $\alpha + \sum_{i=1}^{n} x_i$. The mean and variance of the posterior distribution are

$$\frac{n+1}{\alpha + \sum_{i=1}^{n} x_i} \text{ and } \frac{n+1}{\left(\alpha + \sum_{i=1}^{n} x_i\right)^2}.$$

Note that as n and $\sum_{i=1}^{n} x_i$ increase, the influence of the hyperparameter α on the posterior decreases and the contribution of the likelihood to the posterior increases. Also, the mean and variance of the posterior distribution are approximately $\frac{1}{\bar{x}}$ and $\frac{1}{n\bar{x}^2}$, respectively ,when n is large. As the shape parameter of the Gamma distribution increases, the Gamma density approaches a normal density with the same mean and variance, which suggests that the posterior distribution is approximately $\mathcal{N}\left(\frac{1}{\bar{x}}, \frac{1}{n\bar{x}^2}\right)$. Recall that the MLE of λ is $\frac{1}{\bar{x}}$ and the observed Fisher information is $n\bar{x}^2$.

3.6.1 Multiparameter Model

If the joint PDF/PMF is $f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_k)$ and $\pi(\theta_1, \cdots, \theta_k)$ is a prior density on Θ, then the posterior density of $(\theta_1, \cdots, \theta_k)$ is

$$\pi(\theta_1, \cdots, \theta_k|x_1, \cdots, x_n) = c(\mathbf{x})\pi(\theta_1, \cdots, \theta_k)\mathcal{L}(\theta_1, \cdots, \theta_k)$$

where $c(\mathbf{x})$ is a normalizing constant depending on $x_1, \cdots, x_n$.

Given the posterior density of $(\theta_1, \cdots, \theta_k)$, we can determine the posterior density of a subset of parameters by integrating over the other parameters. For example,

$$\pi(\theta_1, \cdots, \theta_{k-1}|x_1, \cdots, x_n) = \int \pi(\theta_1, \cdots, \theta_k|x_1, \cdots, x_n)\mathrm{d}\theta_k$$

or

$$\pi(\theta_1|x_1, \cdots, x_n) = \int \pi(\theta_1, \cdots, \theta_k|x_1, \cdots, x_n)\mathrm{d}\theta_2 \cdots \mathrm{d}\theta_k.$$

3.6.2 Choice of Prior Distribution

The prior density $\pi(\theta)$ should reflect a priori information about θ, which can come from previous studies or expert opinion. If we have no information at all (which is rare in practice), there are typically regions of the parameter space that are implausible.

If the parameter space is bounded, then we can set $\pi(\theta) = C$ for $\theta \in \Theta$, which is called ***uniform prior***. Note that if we transform θ, i.e., set $\phi = g(\theta)$, then the prior for ϕ is no longer uniform on $g(\Theta)$ if g is a non-linear transformation.

3.6.2.1 Conjugate Prior

Given a model, i.e., a joint PDF/PMF $f(x_1, \cdots, x_n; \theta)$, we can choose a prior density (indexed by some hyperparameters) such that the posterior density has the same form as the prior:

$$\pi_\alpha(\theta) \xrightarrow{\text{Data}} \pi_{\alpha'}(\theta | x_1, \cdots, x_n)$$

where the hyperparameter in the posterior will depend on the data $x_1, \cdots, x_n$. In the example above, we use a Gamma prior with shape parameter 1 and the posterior was also Gamma.

The advantages are:

- Ease of computation: The data just change the values of the hyperparameters.
- In simple models, conjugate priors can provide a rich choice of prior densities.

Example 3.31 (One Parameter Exponential Families)**.** Suppose $(X_1, \cdots, X_n)$ has density

$$f(\mathbf{x}; \theta) = \exp[c(\theta)T(\mathbf{x}) - d(\theta) + h(\mathbf{x})] \text{ for } \mathbf{x} \in A.$$

Let a prior density of the form

$$\pi(\theta) = K(\alpha, \beta) \exp[\alpha c(\theta) - \beta d(\theta)]$$

for α, β in some set s.t. $\pi(\theta)$ is a density function. Note that

$$\pi(\theta)\mathcal{L}(\theta) \propto \exp[(\alpha + T(\mathbf{x}))c(\theta) - (\beta + 1)d(\theta)]$$

and therefore the posterior density is

$$\pi(\theta|\mathbf{x}) = K(\alpha + T(\mathbf{x}), \beta + 1) \exp[(\alpha + T(\mathbf{x}))c(\theta) - (\beta + 1)d(\theta)].$$

Example 3.32 (Binary Data)**.** Suppose $X_1, \cdots, X_n$ are independent binary r.v.s. with

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x} \text{ for } x = 0, 1$$

where $0 < \theta < 1$. The joint PMF of $(X_1, \cdots, X_n)$ is

$$f(x_1, \cdots, x_n; \theta) = \theta^{\sum x_i}(1 - \theta)^{n - \sum x_i} = \exp\left[\ln\left(\frac{\theta}{1 - \theta}\right) \sum_{i=1}^{n} x_i + n \ln(1 - \theta)\right].$$

The form of the joint PMF suggests a Beta prior:

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1} \text{ for } 0 < \theta < 1$$

where $\alpha, \beta > 0$ are hyperparameters. By varying α, β, we can approximate a wide variety of prior densities on $(0, 1)$. We have

$$\pi(\theta)\mathcal{L}(\theta) \propto \theta^{\alpha-1+\sum x_i}(1-\theta)^{\beta-1+n-\sum x_i}.$$

Thus the posterior distribution is Beta with hyperparameters

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i$$
$$\beta' = \beta + n - \sum_{i=1}^{n} x_i.$$

Note that as n increases, the hyperparameters and hence the posterior density are dominated by the data. For large n, the posterior is approximately normal with mean $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and variance $\frac{\widehat{\theta}(1-\widehat{\theta})}{n}$.

3.6.3 Computing Posterior Density

3.6.3.1 Single Parameter

The posterior density is

$$\pi(\theta|x_1, \cdots, x_n) = c(x_1, \cdots, x_n)\pi(\theta)\mathcal{L}(\theta).$$

Typically $\pi(\theta)$ and $\mathcal{L}(\theta)$ are easy to evaluate so we just need to compute

$$c(x_1, \cdots, x_n) = \left[\int_\Theta \pi(\theta)\mathcal{L}(\theta)\mathrm{d}\theta\right]^{-1}.$$

The integral may be simple (for example if we have a conjugate prior) but typically we need to use some sort of numerical integration:

$$\int_\Theta \pi(\theta)\mathcal{L}(\theta)\mathrm{d}\theta \approx \sum_{k=1}^{N} w_k \pi(\theta_k)\mathcal{L}(\theta_k).$$

3.6.3.2 Multiparameter

Numerical integration becomes much more difficult and we can use Monte Carlo methods to estimate the posterior density.

- ***Importance sampling*** (***Monte Carlo integration***):

$$\int_\Theta \pi(\theta)\mathcal{L}(\theta)\mathrm{d}\theta \approx \frac{1}{N}\sum_{k=1}^{N} \frac{\pi(\theta_k)\mathcal{L}(\theta_k)}{g(\theta_k)}$$

  where $\theta_1, \cdots, \theta_N$ are sampled from a joint density g. Ideally, $g(\theta)$ should have a similar shape to $\pi(\theta)\mathcal{L}(\theta)$.

- Sample from the posterior (***Markov chain Monte Carlo***): Draw a sample $\theta_1, \cdots, \theta_N$ from the posterior density. We can then use statistical methods (e.g., density estimation) to estimate marginal posterior densities, etc.

Note that we do not need to know the normalizing constant to use these methods.

3.6.4 Bayesian Interval Estimation

Definition 3.9. Given a posterior density $\pi(\theta|x_1, \cdots, x_n)$, an interval or a set $\mathcal{I} = \mathcal{I}(\mathbf{x})$ is a $100p\%$ ***credible interval*** or ***credible region*** for θ if

$$\int_{\mathcal{I}(\mathbf{x})} \pi(\theta|x_1, \cdots, x_n)\mathrm{d}\theta = p.$$

Definition 3.10. A $100p\%$ credible interval or region $\mathcal{I}$ is called a $100p\%$ ***highest posterior density*** (HPD) ***interval*** or ***region*** for θ if for all $\theta \in \mathcal{I}$ and all $\theta' \notin \mathcal{I}$,

$$\pi(\theta|x_1, \cdots, x_n) > \pi(\theta'|x_1, \cdots, x_n).$$

An HPD interval or region will be the smallest credible interval or region. At the end points, $l(\mathbf{x})$ and $u(\mathbf{x})$ we have

$$\pi(l(\mathbf{x})|\mathbf{x}) = \pi(u(\mathbf{x})|\mathbf{x}).$$

3.6.4.1 Credible versus Confidence Interval

Credible interval and confidence interval attempt to do the same thing, albeit in different ways. Confidence interval is defined in terms of coverage over repeated experiments:

$$P_\theta[\theta \in \mathcal{I}(\mathbf{X})] = P_\theta[l(\mathbf{X}) \leqslant \theta \leqslant u(\mathbf{X})] = p, \forall\theta \in \Theta$$

where P_θ is the probability distribution of $\mathbf{X} = (X_1, \cdots, X_n)$. Credible interval is defined in terms of the posterior distribution of θ which depends on the prior distribution and the data:

$$\int_{\mathcal{I}(\mathbf{x})} \pi(\theta|x_1, \cdots, x_n)\mathrm{d}\theta = p.$$

Example 3.33 (Multiple Mark-Recapture Experiment)**.** Recall that M_i is the the number of marked individuals captured at stage i. The number of marked individuals in the population at stage i is

$$t_i = \sum_{j=0}^{i-1}(n_j - m_j) = t_{i-1} + \underbrace{(n_{i-1} - m_{i-1})}_{\text{Unmarked at stage } (i-1)}$$

where $m_0 = 0$. At stage i, M_i has a hypergeometric distribution depending on N and t_i. The likelihood function for N is a product of hypergeometric distributions

$$\mathcal{L}(N) = \prod_{i=1}^{k}\left[\frac{\binom{t_i}{m_i}\binom{N-t_i}{n_i-m_i}}{\binom{N}{n_i}}\right]$$

where $N \geqslant \sum_{i=0}^{k}(n_i - m_i)$. Setting $\omega = \frac{1}{N}$, we obtain an approximate Poisson log-likelihood function

$$\ln\mathcal{L}(\omega) = \sum_{i=1}^{k}[m_i\ln(n_i t_i\omega) - n_i t_i\omega - \ln(m_i!)].$$

For Bayesian analysis, we can either put a prior on N (discrete parameter) or on ω (pseudo-continuous parameter). For example, we can use the following prior density for ω :

$$\pi(\omega) = \lambda^2\frac{\omega\exp(-\lambda\omega)}{1-(1+\lambda)\exp(-\lambda)} \text{ for } 0 \leqslant \omega \leqslant 1$$

where $\lambda > 0$ is a hyperparameter. The corresponding prior density for $N = \frac{1}{\omega}$ has mean approximately λ and infinite variance.

3.6.5 Point Estimation from Posterior Density

The ***posterior mean*** is

$$\widehat{\theta} = \int_{\Theta} \theta \pi(\theta|x_1, \cdots, x_n) \mathrm{d}\theta.$$

The ***maximum a posteriori*** (MAP) ***estimate*** $\widehat{\theta}$ is the posterior mode

$$\pi(\theta|x_1, \cdots, x_n) \geqslant \pi(\theta|x_1, \cdots, x_n), \forall \theta \in \Theta.$$

Posterior mean and MAP estimate are often used in situations where MLE is unstable or is undefined.

3.6.6 Bayes Estimate and Regularization

The prior density is used to regularize the problem. Essentially, we force the distribution of $\widehat{\theta}$ to stay within a bounded (compact) subset of Θ. By doing so, we reduce the variance of $\widehat{\theta}$, albeit at the expense of possibly increasing the bias. Regularization in statistics almost inevitably involves a trade-off between bias and variance. For example, in kernel density estimation bias and variance depend on the bandwidth parameter h.

Example 3.34 (Non-Regular Location Estimation)**.** Suppose $X_1, \cdots, X_n$ are independent r.v.s. with

$$f(x; \theta) = \frac{|x - \theta|^{-1/2}}{2\sqrt{\pi}} \exp(-|x - \theta|).$$

The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[\frac{|x_i - \theta|^{-1/2}}{2\sqrt{\pi}} \exp(-|x_i - \theta|) \right]$$

and MLE is non-unique/undefined since $\mathcal{L}(\theta) \uparrow \infty$ as $\theta \to x_i$. To regularize the estimation problem, we put a Cauchy prior on θ :

$$\pi(\theta) = \frac{10}{\pi(100 + \theta^2)}.$$

The posterior density is then

$$\pi(\theta|x_1, \cdots, x_n) = c(\mathbf{x}) \frac{\prod_{i=1}^{n} [|x_i - \theta|^{-1/2} \exp(-|x_i - \theta|)]}{100 + \theta^2}$$

where $c(\mathbf{x})$ is the normalizing constant. The mean of the posterior can be determined via numerical integration.

Example 3.35 (LASSO)**.** Suppose $Y_i = \beta_0 + \mathbf{x}_i^T \beta + \varepsilon_i$ where $\{\varepsilon_i\}$ are independent $\mathcal{N}(0, \sigma^2)$ r.v.s.. If p is large relative to n with p possible larger than n, then LS estimators can have very large variances. If $p \geqslant n$, then the LS estimators are not unique. We need to regularize the problem by adding a penalty term to reduce the variability of the estimators, such as ridge regression or LASSO.

LASSO is short for least absolute shrinkage and selection operator. Assume predictors are centered and scaled to have mean 0 and variance 1:

$$\frac{1}{n} \sum_{i=1}^{n} x_{ji} = 0 \text{ and } \frac{1}{n} \sum_{i=1}^{n} x_{ji}^2 = 1.$$

Then define $\widehat{\beta}_0 = \overline{Y}$ and $\widehat{\theta}(\lambda)$ to minimize

$$\sum_{i=1}^{n}(Y_i - \overline{Y} - \mathbf{x}_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

$\widehat{\beta}(\lambda)$ can be viewed as an MAP estimate where the prior density for $(\beta_1, \cdots, \beta_p)$ has form

$$\pi(\beta_1, \cdots, \beta_p) = C\exp\left(-\frac{\lambda}{2\sigma^2}\sum_{j=1}^{p}|\beta_j|\right).$$

In practice, λ acts as a tuning parameter: As we increase λ, components of $\widehat{\beta}$ shrink towards 0 and for λ sufficiently large, $\widehat{\beta}(\lambda) = \mathbf{0}$.

The LASSO plot is the plot of $\widehat{\beta}_j(\lambda)$ against $s(\lambda)$ for each $j = 1, \cdots, p$ where

$$s(\lambda) = \frac{\sum_{j=1}^{p}|\widehat{\beta}_j(\lambda)|}{\sum_{j=1}^{p}|\widehat{\beta}_j(0)|}.$$

The LASSO plot gives an idea of which predictors are most important: $\widehat{\beta}_j(\lambda)$ will shrink to 0 slowest as λ increases or equivalently as $s(\lambda)$ decreases.

3.7 Application: Two State Markov Chain Model

Recall that a Markov chain $\{X_i\}$ is a discrete-time stochastic process s.t. the conditional distribution of X_i given $X_{i-1} = x_{i-1}, X_{i-2} = x_{i-2}, \cdots$ depends only on x_{i-1}. Thus the joint PDF/PMF of $(X_1, \cdots, X_n)$ can be written as

$$f(x_1, \cdots, x_n) = f_1(x_1)\prod_{i=2}^{n} f_{i|i-1}(x_i|x_{i-1}).$$

We now consider the simplest possible Markov chain: X_i takes on only two values (0 and 1). The case is useful for modeling dependence in sequences of binary outcomes.

3.7.1 Transition/Conditional Probability

Recall that the transition probabilities are

$$P(X_{i+1} = 1|X_i = 0) = \alpha, P(X_{i+1} = 0|X_i = 0) = 1 - \alpha$$
$$P(X_{i+1} = 1|X_i = 1) = 1 - \beta, P(X_{i+1} = 0|X_i = 1) = \beta$$

where $0 < \alpha, \beta < 1$. We can define a transition matrix

$$\mathbf{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

$\mathbf{P}$ is useful for determining $P(X_{i+k} = x|X_i = y)$, i.e., $(mathbfP^k$. The two eigenvalues of $\mathbf{P}$ are 1 and $\rho := 1 - \alpha - \beta$ and so the eigenvalues of $\mathbf{P}^k$ are 1 and ρ^k.

3.7.2 Stationary/Invariant Distribution

Recall that the stationary distribution of the Markov chain by looking at the limit $\mathbf{P}^k$ as $k \to \infty$:

$$\mathbf{P}^k \to \mathbf{P}_0 = \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}.$$

Therefore, the stationary distribution of X_i is

$$P(X_i = 1) = f(1; \alpha, \beta) = \frac{\alpha}{\alpha+\beta} = \theta$$
$$P(X_i = 0) = f(0; \alpha, \beta) = \frac{\beta}{\alpha+\beta} = 1 - \theta.$$

It follows that

$$\text{Cor}(X_i, X_{i+1}) = \frac{\text{Cov}(X_i, X_{i+1})}{\text{Var}(X_i)} = 1 - \alpha - \beta = \rho.$$

We have an alternative convenient parametrization: (θ, ρ) rather than (α, β). If $\rho < 0$, then θ cannot be too close to 0 or 1.

3.7.3 Run

Definition 3.11. Suppose we observe a sequence of 0s and 1s: $x_1 x_2 \cdots x_n$. A ***run*** is defined to be a subsequence consisting of all 0s or all 1s.

Example 3.36. The sequence 0001101111001110 has 7 runs of lengths 3, 2, 1, 4, 2, 3, 1.

If $X_1, \cdots, X_n$ come from the two state Markov chain then the number of runs is

$$R = 1 + \sum_{i=1}^{n-1} I(X_i \neq X_{i+1}).$$

Intuitively, R should provide some information about ρ : If ρ is close to 1 then we should see fewer runs; if ρ is close to -1 then we should see more runs. Note that

$$\begin{aligned} \mathbb{E}[R] &= 1 + \sum_{i=1}^{n-1} P(X_i \neq X_{i+1}) = 1 + \frac{2(n-1)\alpha\beta}{\alpha+\beta} \\ &= 1 + 2(n-1) \underbrace{\theta(1-\theta)}_{\text{Variance}} \underbrace{(1-\rho)}_{\in[0,2]} \end{aligned}$$

which confirms our intuition.

3.7.4 MoM Estimation of θ and ρ

We can use the proportion of 1s and the number of runs to obtain MoM estimators of θ and ρ :

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} I(X_i = 1)$$
$$\widehat{\rho} = 1 - \frac{R-1}{2(n-1)\widehat{\theta}(1-\widehat{\theta})}$$

3.7.5 MLE

Given $X_1 = x_1, \cdots, X_n = x_n$, the likelihood function of (α, β) is

$$\mathcal{L}(\alpha, \beta) = f(x_1; \alpha, \beta) \prod_{i=1}^{n-1} f(x_{i+1}|x_i; \alpha, \beta)$$

where

$$f(x_1; \alpha, \beta) = \left(\frac{\alpha}{\alpha+\beta}\right)^{x_1} \left(\frac{\beta}{\alpha+\beta}\right)^{1-x_1}$$
$$f(x_{i+1}|x_i; \alpha, \beta) = [\alpha^{(1-x_i)x_{i+1}}(1-\alpha)^{(1-x_i)(1-x_{i+1})}][\beta^{x_i(1-x_{i+1})}(1-\beta)^{x_i x_{i+1}}]$$

.

We can also define a conditional likelihood function

$$\mathcal{L}_{\text{cond}}(\alpha, \beta) = \prod_{i=1}^{n-1} f(x_{i+1}|x_i; \alpha, \beta)$$

where we condition on the first observation.

The MLE of α and β cannot be written in closed-form and we need to determine these numerically for given data. On the other hand, maximizing the conditional likelihood is straightforward. The conditional MCLE is

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n-1}(1-X_i)X_{i+1}}{\sum_{i=1}^{n-1}(1-X_i)}$$
$$\widehat{\beta} = \frac{\sum_{i=1}^{n-1}(1-X_{i+1})X_i}{\sum_{i=1}^{n-1}X_i}$$

When n is not too small, the MLE and MCLE are close.

3.7.6 Bayesian Analysis

We put a prior density on (α, β) which implies a prior on (θ, ρ) where $\theta = \frac{\alpha}{\alpha+\beta}$ and $\rho = 1 - \alpha - \beta$. If $\pi(\alpha, \beta)$ is a prior density for (α, β) then the prior density for (θ, ρ) is

$$\pi(\theta(1-\rho), (1-\theta)(1-\rho)) \underbrace{(1-\rho)}_{\text{Jacobian}}$$

on the set

$$\left\{(\theta, \rho) : -1 < \rho < 1, \frac{-\min(\rho, 0)}{1 - \min(\rho, 0)} < \theta < \frac{1}{1 - \min(\rho, 0)}\right\}.$$

We know that

$$\pi(\alpha, \beta|x_1, \cdots, x_n) = c(x_1, \cdots, x_n)\pi(\alpha, \beta)\mathcal{L}(\alpha, \beta)$$

where the normalizing constant $c(x_1, \cdots, x_n)$ may or may not be easy to compute. If we replace the likelihood function by the conditional likelihood $\mathcal{L}_{\text{cond}}$ then a conjugate family of priors are independent Beta distributions for α and β. However, even if we know the posterior density of (α, β), determining the posterior density of $\rho = 1 - \alpha - \beta$ is non-trivial.

Markov chain Monte Carlo (MCMC) method is very useful: Draw a sample $(\alpha_1, \beta_1), \cdots, (\alpha_N, \beta_N)$ from the posterior density that gives $\rho_1, \cdots, \rho_N$ from posterior for ρ. We can then estimate the posterior density of ρ using (for example) kernel density estimation.

3.7.6.1 Independence MCMC Sampler

We want to generate independent proposals $(A_1, B_1). \cdots$ from some joint density $g(a, b)$ and define $(\alpha_1, \beta_1), \cdots$ using these proposals. In particular, we define $(\alpha_1, \beta_1), \cdots$ as follows:

$$(\alpha_i, \beta_i) = \begin{cases} (A_i, B_i), & \text{w.p. } P_i \\ (\alpha_{i-1}, \beta_{i-1}), & \text{w.p. } 1 - P_i \end{cases}$$

where

$$P_i = \min \left\{ 1, \frac{\pi(A_i, B_i)\mathcal{L}(A_i, B_i)g(\alpha_{i-1}, \beta_{i-1})}{\pi(\alpha_{i-1}, \beta_{i-1})\mathcal{L}(\alpha_{i-1}, \beta_{i-1})g(A_i, B_i)} \right\}$$

3.8 Bayesian Inference and Updating

The posterior density for θ given $X_1 = x_1, \cdots, X_n = x_n, X_{n+1} = x_{n+1}, \cdots, X_{n+k} = x_{n+k}$ is

$$\begin{aligned} \pi(\theta|x_1, \cdots, x_n, x_{n+1}, \cdots, x_{n+k}) \propto \pi(\theta)\mathcal{L}_{n+k}(\theta) &= \underbrace{\pi(\theta)\mathcal{L}_n(\theta)}_{\propto \pi(\theta|x_1, \cdots, x_n)} \frac{\mathcal{L}_{n+k}(\theta)}{\mathcal{L}_n(\theta)} \\ &\propto \pi(\theta|x_1, \cdots, x_n)\mathcal{L}_{\text{cond}}(\theta) \end{aligned}$$

where

$$\mathcal{L}_{\text{cond}}(\theta) = \frac{f(x_1, \cdots, x_n, x_{n+1}, \cdots, x_{n+k}; \theta)}{f(x_1, \cdots, x_n; \theta)}$$

is the conditional likelihood function of θ given $X_1 = x_1, \cdots, X_n = x_n$. If X_i are independent with PDF/PMF $f(x; \theta)$, then

$$\pi(\theta|x_1, \cdots, x_n, x_{n+1}, \cdots, x_{n+k}) \propto \pi(\theta|x_1, \cdots, x_n) \prod_{i=1}^{k} f(x_{n+i}; \theta).$$

3.9 Predictive Distribution

Example 3.37 (Normal Distribution)**.** Suppose $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. We want to predict X_{n+1} based on $X_1, \cdots, X_n$ and find a prediction interval $\mathcal{I}$ s.t. $P(X_{n+1} \in \mathcal{I}) = p$. We estimate μ by $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Since $\overline{X}_n \perp X_{n+1}$, then

$$X_{n+1} - \overline{X}_n \sim \mathcal{N}\left(0, \sigma^2 + \frac{\sigma^2}{n}\right).$$

Therefore a 95% prediction interval for X_{n+1} is

$$\left[\overline{X}_n \mp 1.96\sigma\sqrt{1 + \frac{1}{n}}\right].$$

If σ^2 is unknown, we replace σ by S and 1.96 by a Student's t quantile.

We can use Bayesian method: We have the conditional PDF/PMF

$$f(x_{n+1}|x_1, \cdots, x_n; \theta) = \frac{f_{n+1}(x_1, \cdots, x_{n+1}; \theta)}{f_n(x_1, \cdots, x_n; \theta)}.$$

Thus, $f(x_{n+1}|x_1,\cdots,x_n;\theta)\pi(\theta|x_1,\cdots,x_n)$ is the PDF/PMF of (X_{n+1},θ) given $X_1=x_1,\cdots,X_n=x_n$. To obtain the predictive PDF/PMF of X_{n+1}, we integrate w.r.t. θ :

$$f(x_{n+1}|x_1,\cdots,x_n)=\int_\Theta f(x_{n+1}|x_1,\cdots,x_n;\theta)\pi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta.$$

Example 3.38 (Poisson Distribution)**.** Suppose X_i are independent Poisson r.v.s. with

$$f(x;\lambda)=\frac{\exp(-\lambda)\lambda^x}{x!},x=0,1,\cdots$$

where $\lambda>0$ is unknown. A conjugate family of priors is the Gamma distribution

$$\pi(\lambda)=\frac{\lambda^{\alpha-1}\beta^\alpha\exp(-\beta\lambda)}{\Gamma(\alpha)},\lambda>0$$

where $\alpha,\beta>0$ are hyperparameters. Th posterior density given $X_1=x_1,\cdots,X_n=x_n$ is

$$\pi(\lambda|x_1,\cdots,x_n)=\frac{\lambda^{\alpha+t-1}(\beta+n)^{\alpha+t}\exp[-(\beta+n)\lambda]}{\Gamma(\alpha+t)},\lambda>0$$

where $t=x_1+\cdots+x_n$. Since we assume X_{n+1} is independent of $X_1,\cdots,X_n$, the predictive PMF of X_{n+1} is

$$f(x|x_1,\cdots,x_n)=\int_0^\infty\frac{\exp(-\lambda)\lambda^x}{x!}\pi(\lambda|x_1,\cdots,x_n)\mathrm{d}\lambda=\frac{\Gamma(\alpha+t+x)(\beta+n)^{\alpha+t}}{x!\Gamma(\alpha+t)(\beta+n+1)^{\alpha+t+x}}$$

for $x=0,1,\cdots$ and $t=x_1+\cdots+x_n$, which is called a ***negative binomial distribution***.

Example 3.39 (Bernoulli Distribution)**.** Suppose X_i are independent Bernoulli r.v.s. with

$$f(x;\theta)=\theta^x(1-\theta)^{1-x},x=0,1$$

where $0<\theta<1$ is unknown. A conjugate family of priors is the Beta distribution

$$\pi(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},0<\theta<1$$

where $\alpha,\beta>0$ are hyperparameters. The posterior density given $X_1=x_1,\cdots,X_n=x_n$ is

$$\pi(\theta|x_1,\cdots,x_n)=\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+t)\Gamma(\beta+n-t)}\theta^{\alpha+t-1}(1-\theta)^{\beta+n-t-1}$$

for $0<\theta<1$ where $t=x_1+\cdots+x_n$. The predictive distribution of $S=X_{n+1}+\cdots+X_{n+1}\sim$ Binomial(k,θ). The predictive distribution of S given $X_1+\cdots+X_n=t$ is called a ***Beta-Binomial distribution***:

$$\begin{aligned}f(s|x_1,\cdots,x_n)&=\int_0^1\binom{k}{s}\theta^s(1-\theta)^{k-s}\pi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta\\&=\binom{k}{s}\frac{\Gamma(\alpha+\beta+n)\Gamma(\alpha+t+s)\Gamma(\beta+n+k-t-s)}{\Gamma(\alpha+t)\Gamma(\beta+n-t)\Gamma(\alpha+\beta+n+k)}\end{aligned}$$

for $s=0,1,\cdots,k$. For large k, we can approximate the Beta-Binomial distribution by a normal distribution with mean and variance

$$\begin{aligned}\mu&=\frac{k(\alpha+t)}{\alpha+\beta+n}\\\sigma^2&=\frac{k(\alpha+t)(\beta+n-t)(\alpha+\beta+n+k)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}\end{aligned}$$

Example 3.40 (Real-Time Election Forecasting)**.** Suppose tow candidates: A and B. It is known that A is more popular in urban areas; B is more popular in non-urban areas. A total of n votes are cast, n_1 from urban areas, and n_0 form non-urban areas with $n = n_0 + n_1$. Assume n_0, n_1 are known and a majority of votes is needed to win, i.e., $\left\lfloor \frac{n}{2} \right\rfloor + 1$.

We can build a simple model to forecast the result based on partial returns. Assume we have data of the form $(A_i, t_i, u_i), i = 1, \cdots, m$ from each precinct where A_i is votes for A, t_i is total number of votes, and $u_i = 1$ if urban precinct; $u_i = 0$ if non-urban.

The model is $A_i | u_i = j \sim \text{Binomial}(t_i, \theta_j)$. The prior densities on θ_0 and θ_1 should reflect the knowledge of the voting dynamics: Prior for θ_1 should have most of its mass on values greater than 0.5; prior for θ_0 should have most of its mass on values less than 0.5. However, the choice of prior is not too important since the data will quickly dominate.

Define cumulative vote totals:

$$V_1(k) = \sum_{i=1}^{k} A_i u_i = \text{Urban votes for } A \text{ in 1st } k \text{ precincts}$$
$$V_0(k) = \sum_{i=1}^{k} A_i (1 - u_i) = \text{Non-Urban votes for } A \text{ in 1st } k \text{ precincts}$$

We can use the observed values of $V_0(k)$ and $V_1(k)$ with the prior distributions on θ_0 and θ_1 to compute

$$P\left(V_0(m) + V_1(m) \geqslant \left\lfloor \frac{n}{2} \right\rfloor + 1 | V_0(k) = v_0(k), V_1(k) = v_1(k) \right).$$

The predictive distributions of $V_0(m) - v_0(k)$ and $V_1(m) - v_1(k)$ are independent Beta-Binomials.

3.10 Bias-Variance Tradeoff

$\widehat{\theta}$ is an estimator of θ (in a parametric model) or of $\theta(F)$ (in a non-parametric model). The mean square error (MSE) of $\widehat{\theta}$ is defined by

$$\text{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = \text{Var}_\theta[\widehat{\theta}] + (\mathbb{E}_\theta[\widehat{\theta}] - \theta)^2 = \text{Var}_\theta[\widehat{\theta}] + \text{Bias}^2_\theta(\widehat{\theta}).$$

Example 3.41 (Exponential Distribution)**.** Suppose $X_1, \cdots, X_n$ are independent exponential r.v.s. with

$$f(x; \lambda) = \lambda \exp(-\lambda x), x \geqslant 0$$

where $\lambda > 0$ is unknown. The MLE of λ is

$$\widehat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^{-1} = \frac{1}{\overline{X}}.$$

We can determine the mean and variance of $\widehat{\lambda}$ since

$$\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda).$$

Hence for $r > -n$,

$$\mathbb{E}_\lambda \left[\left(\sum_{i=1}^{n} X_i \right)^r \right] = \frac{\Gamma(n + r)}{\Gamma(n)} \lambda^{-r}.$$

It follows that

$$\mathbb{E}_\lambda[\widehat{\lambda}] = \frac{n\Gamma(n-1)}{\Gamma(n)}\lambda = \frac{n}{n-1}\lambda$$
$$\mathbb{E}_\lambda[\widehat{\lambda}^2] = \frac{n^2\Gamma(n-2)}{\Gamma(n)}\lambda^2 = \frac{n^2}{(n-1)(n-2)}\lambda^2$$

Therefore,

$$\text{Bias}_\lambda(\widehat{\lambda}) = \mathbb{E}_\lambda[\widehat{\lambda}] - \lambda = \frac{\lambda}{n-1} \approx \frac{\lambda}{n}$$
$$\text{Var}_\lambda[\widehat{\lambda}] = \mathbb{E}_\lambda[\widehat{\lambda}^2] - \mathbb{E}^2_\lambda[\widehat{\lambda}] = \frac{n^2}{(n-1)^2(n-2)}\lambda^2 \approx \frac{\lambda^2}{n}$$

Thus,

$$\text{MSE}_\lambda(\widehat{\lambda}) \approx \frac{\lambda^2}{n} + \frac{\lambda^2}{n^2}$$

and for large $n, \text{MSE}_\lambda(\widehat{\lambda}) \approx \text{Var}_\lambda[\widehat{\lambda}]$. Note that we can make $\widehat{\lambda}$ unbiased:

$$\widetilde{\lambda} = \frac{n-1}{n}\widehat{\lambda}.$$

Note also that $\widetilde{\lambda}$ has smaller variance:

$$\text{Var}_\lambda[\widetilde{\lambda}] = \underbrace{\left(\frac{n-1}{n}\right)^2}_{<1} \text{Var}_\lambda(\widehat{\lambda}) < \text{Var}_\lambda[\widehat{\lambda}].$$

If we want to estimate $\mu = \mathbb{E}_\lambda(X_i) = \frac{1}{\lambda}$, we have

$$\mathbb{E}_\lambda\left[\frac{1}{\widetilde{\lambda}}\right] = \mathbb{E}_\lambda\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{\lambda}$$

which is unbiased, and

$$\text{Var}_\lambda\left[\frac{1}{\widetilde{\lambda}}\right] = \text{MSE}_\lambda\left(\frac{1}{\widetilde{\lambda}}\right) = \frac{1}{n\lambda^2}.$$

On the other hand,

$$\text{MSE}_\lambda\left(\frac{1}{\widetilde{\lambda}}\right) = \text{Var}_\lambda\left[\frac{1}{\widetilde{\lambda}}\right] + \text{Bias}^2_\lambda\left(\frac{1}{\widetilde{\lambda}}\right) = \frac{n}{(n-1)^2\lambda^2} + \frac{1}{(n-1)^2\lambda^2} = \frac{n+1}{(n-1)^2\lambda^2} > \frac{1}{n\lambda^2}.$$

3.10.1 Regularization and Shrinkage

We should not worry obsessively about bias if $\text{Bias}^2_\theta(\widehat{\theta}) \ll \text{Var}_\theta[\widehat{\theta}]$. The MSE is almost always dominated by the variance in nice estimation problems where for example estimators are functions of sample means (or can be approximated thus). Typically in these cases

$$\text{Bias}_\theta(\widehat{\theta}) = \frac{a_1(\theta)}{n}, \text{Var}_\theta(\widehat{\theta}) = \frac{a_2(\theta)}{n}$$

so that

$$\text{MSE}_\theta(\widehat{\theta}) = \frac{a_2(\theta)}{n} + \frac{a_1^2(\theta)}{n^2} \approx \frac{a_2(\theta)}{n}.$$

If $\widetilde{\theta} = a\widehat{\theta}$ for $|a| < 1$ then

$$\mathrm{Var}_\theta(\widetilde{\theta}) < \mathrm{Var}_\theta(\widehat{\theta})$$
$$\mathrm{MSE}_\theta(\widetilde{\theta}) < \mathrm{MSE}_\theta(\widehat{\theta}) \text{ if } \widetilde{\theta} \text{ is unbiased.}$$

Suppose $Y_i = \theta_i + \varepsilon_i$ for $i = 1, \cdots, n$. For simplicity, assume $\{\varepsilon_i\}$ are independent $\mathcal{N}(0, \sigma^2)$ r.v.s. with σ^2 known. To reduce variance, we can assume that $\{\theta_i\}$ have some sort of smoothness. For example, $|\theta_i - \theta_{i-1}|$ is small or $\theta_i = g\left(\frac{i}{n}\right)$ where g is continuous. To estimate $\{\theta_i\}$, we exploit the smoothness assumption. For example, Hodrick-Prescott filter, estimating $\{\theta_i\}$ by minimizing

$$\sum_{i=1}^{n}(Y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n-1}(\theta_{i+1} - 2\theta_i + \theta_{i-1})^2.$$

Increasing λ shrinks the estimates thereby reducing their variance.

In practice, we should be aware of sources of potential bias in our data:

- Misspecified model:

  * $X_1, \cdots, X_n$ are independent and we assume a common PDF/PMF $f(x;\theta)$ but true PDF/PMF is $g(x)$ with $g(x) \neq f(x;\theta).\forall\theta$.

  * Omitted predictors in linear regression: Assumed model is $Y_i = \mathbf{x}_i^T\beta + \varepsilon_i$ but true model is $Y_i = \mathbf{x}_i^T\beta + g(z_i) + \varepsilon_i$ for some omitted predictor $\{z_i\}$.

- Contaminated data: A small fraction of observations are outliers.

Example 3.42 (A Model for Contaminated Data)**.** Suppose X_i are independent with PDF/PMF $f(x;\theta)$ for some $\theta \in \Theta$, which is the nominal model. Suppose X_i are independent with PDF/PMF $(1-\varepsilon)f(x;\theta) + \varepsilon g(x)$ for some $\varepsilon > 0$ and unknown PDF/PMF $g(x)$, which is the true model. ε represents the fraction of bad observations and $g(x)$ represents the population generating the bad data.

We to find an estimator of θ that has minimal absolute bias across ε and contaminating $g(x)$ and such an estimator is called ***bias robust***. Consider contaminated normal model: X_i are independent with PDF

$$(1-\varepsilon)\underbrace{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}_{\mathcal{N}(\mu,\sigma^2)} + \varepsilon g(x).$$

The sample median is a bias robust estimator of μ in this model.

3.10.2 Bias and Reduce Bias

With messy data (for example ,contaminated data or very noisy data), it is often expedient to use methods that are able to extract structure from the bulk of the data. Structure means observations following a simple parametric model. It is easier to identify sub-groups, anomalous observations (e.g., outliers) within the data:

$$\underbrace{\{x_1, \cdots, x_n\}}_{\text{Data}} = \underbrace{\{x_1, \cdots, x_m\}}_{\text{Structure}} \cup \underbrace{\{x_{m+1}, \cdots, x_n\}}_{\text{Outliers}}.$$

The estimators underlying such methods necessarily have a smaller bias in the presence of messy data. The price that we have to pay is a larger variance.

Example 3.43 (Least Median of Squares Estimation)**.** Suppose a simple linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

The nominal assumption is $\{\varepsilon_i\}$ are independent $\mathcal{N}(0, \sigma^2)$ r.v.s. and under the assumption the least squares estimators of β_0 and β_1 are MLEs.

Suppose that the data $\{(x_i, y_i)\}$ contains observations that do not conform to the model, then LS estimates can be driven to $\pm\infty$ by a single observation. A robust alternative is least median of squares (LMS): $\widehat{\beta}_0, \widehat{\beta}_1$ minimize

$$\text{Median}\{(Y_i - \beta_0 - \beta_1 x_i)^2 : i = 1, \cdots, n\}.$$

LMS is effectively a mode estimation method. The variance of LMS estimator tend to 0 with n at a slower rate than that for least squares estimators:

$$\text{Var}[\widehat{\beta}_1^{(\text{LS})}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = O(n^{-1})$$
$$\text{Var}[\widehat{\beta}_1^{(LMS)}] = O(n^{-2/3})$$

We simulate data from model $Y_i = x_i + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}\left(0, \frac{1}{4}\right)$ for $i = 1, \cdots, 100$ with (x_i, Y_i) uniform on $[-3, -2] \times [2, 3]$ for $i = 101, \cdots, 125$. The R code is:

```
library(MASS)
x = rnorm(100)
y = x + rnorm(100, 0, 0.5)
x = c(x, runif(25, -3, -2))
y = c(y, runif(25, 2, 3))
plot(x, y, pch=20, cex=2)
r1 = lm(y ~ x)
r2 = lmsreg(y ~ x, nsamp="exact")
```

3.10.3 Bias-Variance Balance

Example 3.44 (Divide and Conquer)**.** Recall four V's of big data: volume, velocity, variety, veracity. If the volume of data is large then computation of estimates may become difficult. Computational time often scales poorly as the amount of data increases. For example, in linear regression with n observations and p predictors, the computational time for least squares is $O(np^2)$. In other cases, the amount of memory needed to compute estimates is prohibitive.

There is one possible solution, divide and conquer: Divide the data into disjoint subsets and compute estimates on each subset; combined the subset estimates by some sort of averaging.

We have an estimator $\widehat{\theta}$ of some parameter θ based on n observations, and assume for simplicity that

$$\text{Bias}_\theta(\widehat{\theta}) = \frac{c_1(\theta)}{n^\alpha}$$
$$\text{Var}_\theta[\widehat{\theta}] = \frac{c_2(\theta)}{n^\beta}$$
$$\text{MSE}_\theta(\widehat{\theta}) = \frac{c_2(\theta)}{n^\beta} + \frac{c_1^2(\theta)}{n^{2\alpha}}$$

for $\alpha, \beta > 0$. In nice parametric models, we typically have $\alpha = \beta = 1$: variance dominates squared bias. For some estimators, we try to balance the variance and squared bias terms so that $2\alpha = \beta < 1$. Divide the data into k_n subsets of size m_n so that $n = k_n m_n$, and define subset estimators $\widehat{\theta}_j$ for $j = 1, \cdots, k_n$.

Each $\widehat{\theta}_j$ is based on m_n observations so that

$$\begin{aligned}
\text{Bias}_\theta(\widehat{\theta}_j) &= \frac{c_1(\theta)}{m_n^\alpha} \\
\text{Var}_\theta(\widehat{\theta}_j) &= \frac{c_2(\theta)}{m_n^\beta} \\
\text{MSE}_\theta(\widehat{\theta}_j) &= \frac{c_2(\theta)}{m_n^\beta} + \frac{c_1^2(\theta)}{m_n^{2\alpha}}
\end{aligned}$$

Define the averaged estimator

$$\overline{\theta} = \frac{1}{k_n} \sum_{j=1}^{k_n} \widehat{\theta}_j.$$

Assume that $\{\widehat{\theta}_j\}$ are independent, then we have

$$\begin{aligned}
\text{Bias}_\theta(\overline{\theta}) &= \frac{c_1(\theta)}{m_n^\alpha} \\
\text{Var}_\theta(\overline{\theta}) &= \frac{c_2(\theta)}{k_n m_n^\beta} \\
\text{MSE}_\theta(\overline{\theta}) &= \frac{c_2(\theta)}{k_n m_n^\beta} + \frac{c_1^2(\theta)}{m_n^{2\alpha}}
\end{aligned}$$

Unless $\widehat{\theta}$ is unbiased, i.e., $c_1(\theta) = 0$, then $|\text{Bias}(\overline{\theta})| > |\text{Bias}_\theta(\widehat{\theta})|$. Thus if bias is a concern, we should take larger subsets, i.e., bigger m_n and smaller k_n.

The variance of $\overline{\theta}$ relative to that of $\widehat{\theta}$ depends on β :

$$\text{Var}_\theta(\overline{\theta}) \begin{cases} = \text{Var}_\theta(\widehat{\theta}), & \beta = 1 \\ < \text{Var}_\theta(\widehat{\theta}), & \beta < 1 \\ > \text{Var}_\theta(\widehat{\theta}), & \beta > 1 \end{cases}$$

The variance reduction phenomenon for $\beta < 1$ is similar to the variance reduction for ensemble methods in machine learning, i.e., combine estimators (for example, classifiers) from different methods to improve prediction. In these cases, taking increasing k_n reduces the variance at the expense of increasing bias.

Example 3.45 (Non-Parametric Regression)**.** Suppose $Y_i = g(x_i) + \varepsilon_i$ where g is an unknown but smooth function and $\{\varepsilon_i\}$ are independent with mean 0 and variance σ^2. We want to estimate the unknown function g :

- Parametric approach: Write g as a sum of basis functions

$$g(x) = \beta_0 + \sum_{j=1}^{p} \beta_j \phi_j(x)$$

  and estimate $\beta_0, \cdots, \beta_p$ using least squares.

- Non-Parametric approach: Estimate $g(x)$ by a weighted average of $\{Y_i : |x_i - x| \leqslant h\}$.

We will consider non-parametric estimators of the form

$$\widehat{g}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

for some weights $w_1(x), \cdots, w_n(x)$. $w_i(x)$ is largest for x_i close to x with $w_i(x)$ decreasing as $|x_i - x|$ increases. Note that the weights $\{w_i(x)\}$ satisfy $\sum_{i=1}^{n} w_i(x) = 1$ and some of the weights can be negative. The form of $\widehat{g}(x)$ makes it very easy to analyze the mean and variance of $\widehat{g}(x)$:

$$\mathbb{E}[\widehat{g}(x)] = \sum_{i=1}^{n} w_i(x)\mathbb{E}[Y_i] = \sum_{i=1}^{n} w_i(x) g(x_i)$$
$$\text{Var}[\widehat{g}(x)] = \sum_{i=1}^{n} w_i^2(x)\text{Var}[Y_i] = \sigma^2 \sum_{i=1}^{n} w_i^2(x)$$

For example, consider a simple local linear smoother (in a neighborhood of each x, g will be approximately linear). For each x, define the set

$$\mathcal{S}_h(x) = \{i : |x_i - x| \leqslant h\}$$

where $h > 0$ is a bandwidth parameter. Define $m(x)$ to be the size of $\mathcal{S}_h(x)$ and $\overline{x}_x$ to be the average of $\{x_i : i \in \mathcal{S}_h(x)\}$. We estimate $g(x)$ by $\widehat{g}(x) = \widehat{\beta}_0(x) + \widehat{\beta}_1(x)x$ where $\widehat{\beta}_0(x)$ and $\widehat{\beta}_1(x)$ minimize

$$\sum_{i \in \mathcal{S}_h(x)} (Y_i - \beta_0 - \beta_1 x)^2.$$

We then get

$$\widehat{g}(x) = \sum_{i \in \mathcal{S}_h(x)} \underbrace{\left[\frac{1}{m(x)} + \frac{(x_i - \overline{x}_x)(x - \overline{x}_x)}{\sum_{i \in \mathcal{S}_h(x)} (x_i - \overline{x}_x)^2} \right]}_{w_i(x)} Y_i.$$

Also

$$\text{Bias}[\widehat{g}(x)] = \sum_{i=1}^{n} w_i(x) g(x_i) - g(x) = \sum_{i=1}^{n} w_i(x)[g(x_i) - g(x)].$$

Note that $\text{Var}[\widehat{g}(x)]$ depends only on $\{w_i(x)\}$ and not the function g. The bias depends on the relationship between $w_i(x)$ and $g(x_i) - g(x)$: To make the bias small in absolute terms, we should set $w_i(x)$ to be close to 0 when $|g(x_i) - g(x)|$ is large and $w_i(x)$ to be larger when $|(g_i) - g(x)|$ is small. However we do not know g but under smoothness assumption, $|g(x_i) - g(x)|$ will be small if $|x_i - x|$ is small, which agrees with the prescription for $w_i(x)$: it is largest for x_i close to x with $w_i(x)$ decreasing as $|x_i - x|$ increases.

To analyze bias further, we need to make some additional assumptions about g and $w_i(x)$:

- g is (at least) twice differentiable:

$$g(x_i) - g(x) \approx g'(x)(x_i - x) + \frac{1}{2} g''(x)(x_i - x)^2.$$

- For some $h > 0, w_i(x) = 0$ when $|x_i - x| > h$ and for $|x_i - x| \leqslant h$,

$$w_i(x) \propto w\left(\frac{x_i - x}{h}\right)$$

where the kernel $w(t)$ is symmetric around 0, i.e., $w(t) = w(-t)$, and

$$\int_{-1}^{1} w(t)\mathrm{d}t = 1.$$

- $\{x_i\}$ are approximately uniformly distributed over $[a, b]$ for some $a < b$.

For simplicity and w.l.o.g., take $[a, b] = [0, 1]$, in which case

$$w_i(x) = \frac{1}{nh} w\left(\frac{x_i - x}{h}\right).$$

Therefore,

$$\begin{aligned}
\mathrm{Bias}(\widehat{g}(x)) &= \sum_{i=1}^{n} w_i(x)[g(x_i) - g(x)] \\
&\approx \sum_{i=1}^{n} w_i(x)\left[g'(x)(x_i - x) + \frac{1}{2}g''(x)(x_i - x)^2\right] \\
&\approx \frac{h^2 g''(x)}{2nh} \sum_{|x_i - x| \leqslant h} w\left(\frac{x_i - x}{h}\right)\left(\frac{x_i - x}{h}\right)^2 \\
&\approx \frac{h^2 g''(x)}{2} \int_{-1}^{1} t^2 w(t)\mathrm{d}t \\
\mathrm{Var}[\widehat{g}(x)] &= \sigma^2 \sum_{i=1}^{n} w_i^2(x) \\
&\approx \frac{\sigma^2}{n^2 h^2} \sum_{|x_i - x| \leqslant h} w^2\left(\frac{x_i - x}{h}\right) \\
&\approx \frac{\sigma^2}{nh} \int_{-1}^{1} w^2(t)\mathrm{d}t \\
\mathrm{MSE}(\widehat{g}(x)) &\approx \left[\frac{g''(x)}{2} \int_{-1}^{1} t^2 w(t)\mathrm{d}t\right]^2 h^4 + \frac{\sigma^2}{nh} \int_{-1}^{1} w^2(t)\mathrm{d}t
\end{aligned}$$

The true function g effects the bias of $\widehat{g}(x)$ via $g''(x)$ which is the curvature of the function: If g is a nearly linear function then the bias will be quite small, regardless of h. Note that the parameter h has a different effect on the bias and variance terms: As $h \to 0$, the bias tends to 0; on the other hand, the variance blows up as $h \to 0$.