

STA355 Notes

Ian Zhang

July 29, 2024

Contents

| | | |
|----------|---|----------|
| 1 | Probability | 3 |
| 1.1 | Convergence | 3 |
| 2 | Statistical Models | 4 |
| 2.1 | Statistical Functionals | 4 |
| 2.1.1 | Substitution Principle | 4 |
| 2.2 | Order Statistics | 4 |
| 2.2.1 | Distribution of $X_{(k)}$ | 5 |
| 2.2.2 | Central Order Statistics | 5 |
| 2.3 | Spacings Density Estimation | 5 |
| 2.3.1 | Spacings From a Continuous F | 5 |
| 2.3.2 | Hazard Functions | 5 |
| 2.4 | Kernel Density Estimation | 6 |
| 3 | Point and Interval Estimation | 6 |
| 3.1 | Point Estimation | 6 |
| 3.1.1 | Consistency | 7 |
| 3.2 | Standard Error | 7 |
| 3.2.1 | Delta Method Standard Error Estimator | 7 |
| 3.2.2 | Jackknife Standard Error Estimator | 7 |
| 3.3 | Method of Moments Estimation | 8 |
| 3.3.1 | One Unknown | 8 |
| 3.3.2 | k unknowns | 8 |
| 3.4 | Interval Estimation | 9 |
| 3.4.1 | Confidence Intervals | 9 |
| 3.4.2 | Pivotal Method | 9 |
| 3.5 | Maximum Likelihood Estimation | 10 |
| 3.5.1 | Sufficiency | 10 |

| | | |
|----------|---|-----------|
| 3.5.2 | Computing MLE | 11 |
| 3.5.3 | One parameter models | 11 |
| 3.5.4 | Theoretical justification of MLE | 11 |
| 3.5.5 | Consistency of MLE | 12 |
| 3.6 | Asymptotic Normality of MLEs | 12 |
| 3.6.1 | Bartlett Identities | 13 |
| 3.6.2 | Invariance of the Likelihood to injective transformations | 14 |
| 3.6.3 | More general MLE | 14 |
| 3.7 | Misspecified Models | 15 |
| 3.7.1 | Asymptotic Normality of the MLE in a misspecified model | 16 |
| 3.8 | MoM vs MLE | 16 |
| 3.9 | Bayesian Inference | 18 |
| 3.9.1 | Choice of prior | 19 |
| 3.9.2 | Bayesian Interval Estimation | 19 |
| 3.10 | Bias and Variance Tradeoffs | 19 |
| 3.11 | Non-parametric Regression | 20 |
| 3.11.1 | Smoothing Matrices | 20 |
| 4 | Hypothesis Testing | 21 |
| 4.1 | Power of a Test | 21 |
| 4.1.1 | Formal Setup of a Hypothesis Test | 21 |
| 4.2 | Neyman-Pearson Lemma | 22 |
| 4.2.1 | Likelihood Ratio Test | 24 |
| 4.3 | p values | 25 |
| 4.3.1 | Stochastic Ordering | 25 |
| 4.3.2 | Combining p values | 26 |
| 4.4 | Multiple Hypothesis Testing | 26 |
| 4.4.1 | Family-Wise Error Rate | 26 |
| 4.4.2 | False Discovery Rate | 27 |

1 Probability

1.1 Convergence

Definition. A sequence of random variables $(X_n)_n$ converges in probability to a random variable X if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0$$

If so, then $X_n \xrightarrow{p} X$.

Theorem (Weak Law of Large Numbers). If X_1, X_2, \dots are independent random variables with finite mean μ , then

$$\bar{X}_n \xrightarrow{p} \mu$$

Definition. A sequence of random variables $(X_n)_n$ converges in distribution to a random variable X if

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x)$$

If so, then $X_n \xrightarrow{d} X$.

Theorem (Central Limit Theorem). If X_1, X_2, \dots are independent random variables with common cdf F , mean $\mu < \infty$ and finite variance $\sigma^2 < \infty$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

This theorem implies that the distribution of \bar{X}_n is approximately Normal with mean μ and variance $\frac{\sigma^2}{n}$ when n is sufficiently large.

Theorem (Slutsky's Theorem). Suppose $X_n \xrightarrow{d} X \sim G$ and $Y_n \xrightarrow{p} \theta$ for some constant $\theta \in \mathbb{R}$. Then

1. $X_n + Y_n \xrightarrow{d} X + \theta$
2. $X_n Y_n \xrightarrow{d} \theta X$

Theorem (Delta Method). Suppose $a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n \uparrow \infty$. If g is differentiable at θ with derivative $g'(\theta)$, then

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

Definition. Let X_1, X_2, \dots be independent with cdf F , mean μ and variance $\sigma^2 = \varphi(\mu)$. A variance stabilizing transformation is a function g such that

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

- g satisfies

$$[g'(\mu)]^2 \varphi(\mu) = 1$$

2 Statistical Models

2.1 Statistical Functionals

Let X_1, \dots, X_n be independent with unknown cdf F .

- Often, we're interested in estimating characteristics $\theta(F)$ of F
 - A statistical functional is a function $\theta(\cdot) : \mathcal{F} \rightarrow \mathbb{R}$ where \mathcal{F} is a family of distributions with $F \in \mathcal{F}$

2.1.1 Substitution Principle

Given $X_1, \dots, X_n \stackrel{iid}{\sim} F$, we estimate $\theta(F)$ by first estimating $F \rightarrow \hat{F}$ and substituting \hat{F} for F into $\theta(F)$:

$$\hat{\theta}(F) = \theta(\hat{F})$$

If $\theta(\cdot)$ is continuous and $\hat{F} \approx F$, then $\theta(\hat{F}) \approx \theta(F)$.

Definition. A simple estimator of F is the Empirical Distribution Function (edf) defined by

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$$

Properties of the edf:

- $E[\hat{F}(x)] = F(x)$
- $\text{Var}[\hat{F}(x)] = \frac{F(x)(1-F(x))}{n}$
- Since the edf is effectively a sample mean of independent Bernoulli trials, the CLT and WLLN both hold
 - $\hat{F}(x) = \hat{F}_n(x) \xrightarrow{p} F(x)$ for each x
 - $\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1 - F(x)))$

2.2 Order Statistics

Let X_1, \dots, X_n be independent with continuous cdf F and pdf f .

Definition. The order statistics of X_1, \dots, X_n are the ordered values of the X_i :

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

- If $k \approx \tau n$ for some $\tau \in (0, 1)$, then $X_{(k)}$ should tend to $F^{-1}(\tau)$ as n increases \rightarrow
 $X_{(k)} \approx F^{-1}(\tau)$ for large n

2.2.1 Distribution of $X_{(k)}$

The cdf of $X_{(k)}$ is

$$P(X_{(k)} \leq x) = \sum_{i=k}^n \binom{n}{i} F(x)^i [1 - F(x)]^{n-i}$$

2.2.2 Central Order Statistics

Definition. Suppose $k = k_n \approx \tau n$ for some $\tau \in (0, 1)$ but not too close to 0 nor 1. We say $X_{(k)}$ is a central order statistic.

- $X_{(k)}$ is an estimator of the τ -quantile $F^{-1}(\tau)$ so expect $X_{(k)} \rightarrow F^{-1}(\tau)$ as $n \rightarrow \infty$ if $\frac{k}{n} \rightarrow \tau$

Theorem (Convergence in Distribution of Central Order Statistics). If $(k_n)_n$ is a sequence of integers with $\sqrt{n} \left(\frac{k_n}{n} - \tau \right) \rightarrow 0$ for some $\tau \in (0, 1)$ and $f(F^{-1}(\tau)) > 0$, then

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

2.3 Spacings Density Estimation

Given order statistics $X_{(1)} \leq \dots \leq X_{(n)}$, define $n - 1$ spacings as

$$D_k = X_{(k+1)} - X_{(k)}$$

for $k \in \{1, \dots, n - 1\}$.

If $\tau \approx \frac{k+1}{n} \approx \frac{k}{n}$, then $X_{(k+1)} \approx X_{(k)} \approx F^{-1}(\tau)$. Since the number of observations around $F^{-1}(\tau)$ should increase as $f(F^{-1}(\tau))$ increases, D_k should be smaller if $f(F^{-1}(\tau))$ is large and conversely larger if $f(F^{-1}(\tau))$ is small.

2.3.1 Spacings From a Continuous F

Suppose F is continuous. Then if $\frac{k_n}{n} \rightarrow \tau$ for some $\tau \in (0, 1)$ and $f(F^{-1}(\tau)) > 0$, then

$$nD_{k_n} \xrightarrow{d} \text{Exponential with mean } \frac{1}{f(F^{-1}(\tau))}$$

2.3.2 Hazard Functions

Definition. If X is a positive continuous random variable with cdf F and pdf f , its hazard function is

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

- $F(x) = 1 - \exp\left(-\int_0^x h(t) dt\right)$

- $f(x) = h(x) \exp\left(-\int_0^x h(t) dt\right)$

Suppose X_1, \dots, X_n are independent positive continuous random variables with hazard function h . Define normalized spacings $nX_{(1)}, (n-1)D_1, \dots, D_{n-1}$. If $\frac{k}{n} \rightarrow \tau$ for some $\tau(0, 1)$, then


$$(n-k)D_k \xrightarrow{d} \text{Exponential}(h(F^{-1}(\tau)))$$

Proof. Note that

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

and since $D_k \xrightarrow{d} \text{Exponential}(f(F^{-1}(\tau)))$

$$\begin{aligned} (n-k)D_k &= \left(1 - \frac{k}{n}\right) nD_k \\ &\approx (1 - \tau)nD_k \\ &\xrightarrow{d} \text{Exponential}\left(\frac{1 - \tau}{f(F^{-1}(\tau))}\right) \end{aligned}$$

by linearity of expectation. 

2.4 Kernel Density Estimation

Let $w(x)$ be a density called a kernel. Given a kernel w , a bandwidth parameter h , define the kernel density estimator to be

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

3 Point and Interval Estimation

3.1 Point Estimation

Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be an estimator of θ .

Definition. The sampling distribution of $\hat{\theta}$ is its probability distribution

- This will depend on θ

Definition. The Mean Squared Error (MSE) of $\hat{\theta}$ is

$$\text{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] = \text{Var}_\theta(\hat{\theta}) + [\text{Bias}_\theta(\hat{\theta})]^2$$

- If $\hat{\theta}$ is an unbiased estimator of θ , then

$$\text{MSE}(\hat{\theta}) = \text{Var}_\theta(\hat{\theta})$$

3.1.1 Consistency

Definition. The sequence of estimators $(\hat{\theta}_n)_n$ is consistent for θ if for all $\varepsilon > 0$ and $\theta \in \Theta$,

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

- $(\hat{\theta}_n)_n$ is consistent for θ if $\hat{\theta}_n \xrightarrow{P} \theta$

3.2 Standard Error

Assume $\hat{\theta}$ is an estimator of θ .

Definition. The standard error of $\hat{\theta}$ is the standard deviation of the sampling distribution of $\hat{\theta}$ and is denoted $\text{se}(\hat{\theta})$.

Often, $\text{se}(\hat{\theta})$ includes unknown parameters (since the sampling distribution is dependent on θ), so we have to resort to estimating the standard error. There are 2 methods for estimating: the Delta Method and Jackknife.

3.2.1 Delta Method Standard Error Estimator

Let X_1, \dots, X_n be independent with some unknown cdf F . Suppose $\hat{\theta} = g(\bar{X})$. If g is differentiable, by the Delta Method

$$\hat{\theta} = g(\bar{X}) \sim \mathcal{N}\left(0, [g'(\mu)]^2 \frac{\sigma^2}{n}\right)$$

where $\sigma^2 = \text{Var}(X_i)$. Then since we can estimate σ^2 with S^2 , by the substitution principle, we can estimate standard error with

$$\widehat{\text{se}}(\hat{\theta}) = \frac{|g'(\bar{X})|S}{\sqrt{n}}$$

where $S := \sqrt{S^2}$.

- Idea is to use the Delta Method to obtain a sampling distribution of $\hat{\theta}$, and then estimate the standard error using the substitution principle

3.2.2 Jackknife Standard Error Estimator

Suppose $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ is an estimator of θ . Define the leave-one-out estimator $\hat{\theta}_{-i}$ as

$$\hat{\theta}_{-i} = \hat{\theta}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$$

Suppose we can approximate $\hat{\theta}$ as a sample mean

$$\hat{\theta} \approx \frac{1}{n} \sum_{i=1}^n \varphi(X_i)$$

where φ is unknown. Then the leave-one-out estimators satisfy

$$\hat{\theta}_{-i} \approx \frac{1}{n-1} \sum_{j \neq i} \varphi(X_j)$$

Since φ is unknown, define pseudo-values

$$\phi_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i} \approx \varphi(X_i)$$

which are used to recover $\varphi(X_i)$.

Given pseudo-values ϕ_1, \dots, ϕ_n , the Jackknife Standard Error estimator is defined as

$$\widehat{\text{se}}(\hat{\theta}) = \left(\frac{1}{n(n-1)} \sum_{i=1}^n (\phi_i - \bar{\phi})^2 \right)^{\frac{1}{2}} = \left(\frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{-i} - \hat{\theta}_{\bullet})^2 \right)^{\frac{1}{2}}$$

where

$$\hat{\theta}_{\bullet} := \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{-i}$$

3.3 Method of Moments Estimation

Consider X_1, \dots, X_n with joint pdf/pmf

$$f(x_1, \dots, x_n; \theta_1, \dots, \theta_k)$$

for an unknown finite number unknowns $\theta_1, \dots, \theta_k$.

3.3.1 One Unknown

Find a statistic $T(X_1, \dots, X_n)$ such that

$$E_{\theta}(T(X_1, \dots, X_n)) = h(\theta)$$

where h has a well-defined inverse. Then propose an estimator

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

3.3.2 k unknowns

If we have k unknown parameters, then we need k moment conditions

- Ensure that there is an injective mapping between the θ_i and the moments

We can also use quantiles as moment conditions: If X_1, \dots, X_n are independent with cdf F_{θ} and $F_{\theta}^{-1}(\tau) = h(\theta)$, then we can define $\hat{\theta} = h^{-1}(X_{(k)})$ where $k \approx n\tau$.

Example:

Suppose X_1, \dots, X_n are independent with pdf

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1} \exp(-\lambda x)}{\Gamma(\alpha)} \text{ for } x \geq 0$$

where $\lambda, \alpha > 0$ are unknown. Note that

$$E(X_i) = \frac{\alpha}{\lambda} \quad \text{Var}(X_i) = \frac{\alpha}{\lambda^2}$$

Set

$$\bar{X} = \frac{\hat{\alpha}}{\hat{\lambda}} \quad S^2 = \frac{\hat{\alpha}}{\hat{\lambda}^2}$$

Then

$$\hat{\alpha} = \frac{\bar{X}^2}{S^2} \quad \hat{\lambda} = \frac{\bar{X}}{S^2}$$

3.4 Interval Estimation

Let (X_1, \dots, X_n) have a distribution with unknown parameter θ . Interval estimation is when we define an interval

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

that we believe will contain θ with probability 1.

3.4.1 Confidence Intervals

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

$$P_\theta(\theta \in \mathcal{I}) = p \quad \forall \theta \in \Theta$$

3.4.2 Pivotal Method

Let X_1, \dots, X_n be independent with unknown cdf F . Estimate $\theta = \theta(F)$ by $\hat{\theta}$ with

$$\hat{\theta} \sim \mathcal{N}(\theta, [\text{se}(\hat{\theta})]^2)$$

If $\widehat{\text{se}}(\hat{\theta})$ is a good estimator of $\text{se}(\hat{\theta})$, then

$$\frac{\hat{\theta} - \theta}{\widehat{\text{se}}(\hat{\theta})} \sim \mathcal{N}(0, 1)$$

The idea of the pivotal method is to find a random variable $g(X_1, \dots, X_n, \theta)$ with a distribution independent of θ and any other unknowns

- $g(X_1, \dots, X_n, \theta)$ is called a **pivot**

Given a pivot, choose a, b such that

$$p = P_\theta(a \leq g(X_1, \dots, X_n, \theta) \leq b) = G(b) - G(-a)$$

where G is the cdf of $g(X_1, \dots, X_n, \theta)$ and is completely known. From this, we can manipulate to

$$p = P_\theta(l(X_1, \dots, X_n) \leq \theta \leq u(X_1, \dots, X_n))$$

Choosing $g(X_1, \dots, X_n, \theta)$: If we have a point estimator $\hat{\theta}$, select $g(\hat{\theta}, \theta)$ such that its distribution is independent of θ .

Choosing a, b : If G is cdf of the pivot, default is defining a, b such that

$$G(a) = \frac{1-p}{2} \quad G(b) = \frac{1+p}{2}$$

We want the CI to be as short as possible, so we have to minimize the distance $b - a$.

3.5 Maximum Likelihood Estimation

Model: Let (X_1, \dots, X_n) be random variables with joint pdf/pmf

$$f(x_1, \dots, x_n; \theta_1, \dots, \theta_k)$$

where $\theta_1, \dots, \theta_k$ are unknown parameters.

Define the likelihood function for **fixed** data x_1, \dots, x_n as

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

Definition. Suppose for each $\mathbf{x} = (x_1, \dots, x_n)$, $(T_1(\mathbf{x}), \dots, T_k(\mathbf{x}))$ maximizes $\mathcal{L}(\theta_1, \dots, \theta_k)$.

The MLEs of $\theta_1, \dots, \theta_k$ are

$$\hat{\theta}_j = T_j(\mathbf{x})$$

for all $j \in \{1, \dots, k\}$.

- The MLE need not be unique nor exist

3.5.1 Sufficiency

Definition. A statistic $T = (T_1(X), \dots, T_m(X))$ is a sufficient statistic for θ if the conditional distribution of X given $T = t$ depends solely on t and not θ .

Theorem (Neyman-Factorization Theorem). Suppose the joint pdf/pmf of $X = (X_1, \dots, X_m)$ is $f(x; \theta)$. The statistic $T = (T_1(X), \dots, T_m(X))$ is a sufficient statistic of θ iff

$$f(x; \theta) = g(T(X); \theta)h(x)$$

where h doesn't depend on θ .

3.5.2 Computing MLE

There are 2 scenarios depending on the parameter set Θ .

1. $\mathcal{L}(\theta)$ is differentiable and Θ is open. Then $\hat{\theta}$ is the MLE if it satisfies the likelihood equation

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

2. $\hat{\theta}$ occurs on a boundary:

- $\hat{\theta} \in \partial\Theta$
- $\hat{\theta}$ is an extreme of the data (ie: $\hat{\theta} = X_{(n)}$)

3.5.3 One parameter models

Let X_1, \dots, X_n be random variables with joint pdf/pmf $f(x_1, \dots, x_n; \theta)$ where $\theta \in \Theta$ is unknown. Consider the likelihood function

$$\mathcal{L}(\theta) = f(x_1, \dots, x_n; \theta)$$

which is maximized over Θ by the MLE. If Θ is open and \mathcal{L} differentiable, then we can determine $\hat{\theta}$ by the likelihood equation

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

Given the MLE $\hat{\theta}$, defined the observed Fisher information as

$$-\frac{d^2}{d\theta^2} \ln(\mathcal{L}(\hat{\theta}))$$

The Fisher information is used to estimate standard error of the MLE:

$$\widehat{\text{se}}(\hat{\theta}) = \left(-\frac{d^2}{d\theta^2} \ln(\mathcal{L}(\hat{\theta})) \right)^{-\frac{1}{2}}$$

The Fisher information tells us about the absolute curvature of the log-likelihood function at its maximum; the greater the curve, the more well-defined the MLE is, thus as Fisher information increases, uncertainty in the estimate decreases.

3.5.4 Theoretical justification of MLE

The MLE $\hat{\theta} = \hat{\theta}_n$ maximizes

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

over Θ . It also maximizes

$$\frac{1}{n}(\ln(\mathcal{L}(\theta)) - \ln(\mathcal{L}(\theta_0))) = \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) =: \varphi_n(\theta)$$

By the WLLN, for each $\theta \in \Theta$,

$$\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) \xrightarrow{p} E_{\theta_0} \left[\ln \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) \right] =: \varphi(\theta)$$

Proposition (Jensen's Inequality). If g is strictly concave (ie: $g'' < 0$), then

$$E[g(Y)] \leq g[E(Y)]$$

with equality iff Y is constant.

Claim. $0 = \varphi(\theta_0) > \varphi(\theta)$ for all $\theta \neq \theta_0$.

Proof. Suppose f is a pdf without loss of generality. By Jensen's Inequality, since $\ln(x)$ is strictly concave, for $\theta \neq \theta_0$,

$$\begin{aligned} \varphi(\theta) &= E_{\theta_0} \left[\ln \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) \right] \\ &< \ln \left(E_{\theta_0} \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) \right) \\ &= \ln \left(\int_{\mathbb{R}} \frac{f(x; \theta)}{f(x; \theta_0)} f(x; \theta_0) dx \right) \\ &= \ln \left(\int_{\mathbb{R}} f(x; \theta) dx \right) \\ &= 0 \end{aligned}$$

as required. 

Since, $\hat{\theta}_n$ maximizes $\varphi_n(\theta)$, which converges in P for each θ to $\varphi(\theta)$, which is maximized at $\theta = \theta_0$, this justifies $\hat{\theta}$ as an estimator of θ_0 .

3.5.5 Consistency of MLE

If the MLE is a function of a sample mean, then $\hat{\theta}_n = g(\hat{X}_n) \xrightarrow{p} g(\mu_0) = \theta_0$ if g is continuous.

- Since conditions on $f(x; \theta)$ and Θ facilitate consistency, assume consistency of $(\hat{\theta}_n)_n$.

3.6 Asymptotic Normality of MLEs

Let X_1, \dots, X_n be independent with pdf/pmf $f(x; \theta)$ for some $\theta \in \Theta$. Assume

- Θ is open
- $A = \{x : f(x; \theta) > 0\}$ is independent of θ
- $l(x; \theta) = \ln(f(x; \theta))$ is 3 times differentiable WRT θ for all $x \in A$ with derivatives $l'(x; \theta), l''(x; \theta), l'''(x; \theta)$.

If θ_0 is the true parameter, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$$

where $I(\theta_0) = \text{Var}[l'(x; \theta)] = -E[l''(x; \theta)]$ Notice that this implies

$$\widehat{\text{se}}(\hat{\theta}_n) = \left(-\sum_{i=1}^n l''(x_i; \hat{\theta}_n)\right)^{-\frac{1}{2}} = \left(-\frac{d^2}{d\theta^2} \ln(\mathcal{L}(\hat{\theta}_n))\right)^{-\frac{1}{2}}$$

by log identities. Since the MLE is approximately Normally distributed, the CIs are based on the pivot

$$\frac{\hat{\theta} - \theta}{\widehat{\text{se}}(\hat{\theta})} \sim \mathcal{N}(0, 1)$$

We can also use the approximate pivot

$$2[\ln(\mathcal{L}(\hat{\theta})) - \ln(\mathcal{L}(\theta))] \approx nI(\theta)(\hat{\theta} - \theta)^2 \sim \chi_1^2$$

thus the approximate 100p% CI is

$$\{\theta : 2[\ln(\mathcal{L}(\hat{\theta})) - \ln(\mathcal{L}(\theta))] \leq q_p\}$$

where q_p is the p th quantile of the χ_1^2 distribution.

3.6.1 Bartlett Identities

Suppose $f(x; \theta)$ is a pdf and let $A = \{x : f(x; \theta) > 0\}$ be independent of θ , thus for all $\theta \in \Theta$,

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

thus taking any k th order derivative of $\int_A f(x; \theta) dx$ would equal 0. If we suppose we can differentiate under the \int sign, the Bartlett identities hold:

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

The $k = 1$ and $k = 2$ identities imply that $\text{Var}[l'(x_i; \theta)] = -E[l''(x_i; \theta)]$, which is what we would expect.

3.6.2 Invariance of the Likelihood to injective transformations

We know that if X_1, \dots, X_n are independent, then the MLE $\hat{\theta}$, under regularity conditions on $f(x; \theta)$, is asymptotically Normally distributed with

$$\hat{\theta} \sim \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

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

What if X_1, \dots, X_n aren't independent?

Suppose that the X_1, \dots, X_n have joint pdf/pmf $f(x_1, \dots, x_n; \theta)$. Define $\mathbf{Y} = (Y_1, \dots, Y_n) = g(X_1, \dots, X_n)$ where

$$g(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix}$$

is an injective function.

Claim. The likelihoods for θ based on \mathbf{x} and $\mathbf{y} = g(\mathbf{x})$ are the same up to a multiplicative constant:

$$\mathcal{L}(\theta|\mathbf{x}) = \mathcal{L}(\theta|\mathbf{y})K(\mathbf{y})$$

where K is independent of θ .

- In the discrete case, take $K(\mathbf{y}) = 1$
- In the continuous case, take $K(\mathbf{y}) = \text{Jacobian of } g^{-1}(\mathbf{y})$
- The MLE based on x_1, \dots, x_n is the same as the MLE based on y_1, \dots, y_n

If the X_1, \dots, X_n are dependent but the Y_1, \dots, Y_n are independent, then we can still apply iid theory to the MLE, such as using Fisher information to estimate standard errors.

3.6.3 More general MLE

Suppose (X_1, \dots, X_n) have joint pdf/pmf $f(x_1, \dots, x_n; \theta)$. We can rewrite $f(x_1, \dots, x_n; \theta)$ as a product of conditional pdfs/pmfs.:

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

Thus, the log-likelihood function is

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

If θ_0 is the true parameter value, then $l'(X_1; \theta_0)$ and $l''(X_i|X_1, \dots, X_{i-1}; \theta)$ for $i \geq 2$ each have mean 0 and are uncorrelated. Additionally,

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

- The numerator is the sum of uncorrelated random variables with expected value 0, so should be approximately Normally distributed
- The variance of the numerator is equal to the negative expected value of the denominator

Thus

$$\text{Var}(\hat{\theta}) \approx \left(-\sum_{i=1}^n l''(X_i|X_1, \dots, X_{i-1}; \theta_0) \right)^{-1}$$

which implies

$$\widehat{\text{se}}(\hat{\theta}) = \left(-\sum_{i=1}^n l''(X_i|X_1, \dots, X_{i-1}; \hat{\theta}) \right)^{-\frac{1}{2}} = \left(-\frac{d^2}{d\theta^2} \ln(\mathcal{L}(\theta)) \right)^{-\frac{1}{2}}$$

3.7 Misspecified Models

Suppose X_1, \dots, X_n are independent with pdf/pmf $f(x; \theta)$ for some $\theta \in \Theta$. Suppose the true pdf/pmf is some $g(x) \neq f(x; \theta)$ for all $\theta \in \Theta$.

- This model is a misspecified model
- Ideally, we want $g(x) \approx f(x; \theta_0)$

Define $\hat{\theta}$ as the solution to

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

Since $\hat{\theta}$ is the MLE, then $\hat{\theta} = \hat{\theta}_n$ also maximizes

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

By the WLLN, for all $\theta \in \Theta$,

$$\varphi_n(\theta) \xrightarrow{p} \varphi(\theta) = E_g \left[\ln \left(\frac{f(X_i; \theta)}{g(X_i)} \right) \right]$$

Define θ_0 to be the value of θ maximizing φ over Θ .

- This is consistent with the theoretical justification of the MLE

This implies

$$E_g(l'(X_i; \theta_0)) = 0$$

3.7.1 Asymptotic Normality of the MLE in a misspecified model

For a misspecified model, the asymptotic normality of the MLE is a bit different. Suppose the regularity conditions of asymptotic normality of $\hat{\theta}$ for a specified model hold. Then

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

where

$$I_g(\theta_0) = -E_g[l''(X_i; \theta_0)] \quad J_g(\theta_0) = \text{Var}[l'(X_i; \theta_0)]$$

so

$$\hat{\theta} \rightsquigarrow \mathcal{N}\left(\theta_0, \frac{J_g(\theta_0)}{nI_g^2(\theta_0)}\right)$$

We can estimate J_g and I_g using

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

- Since $\hat{\theta} \xrightarrow{p} \theta_0$ and $E_g[l'(X_i; \theta_0)] = 0$

and

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

So, we estimate standard error using

$$\widehat{\text{se}}(\hat{\theta}) = \left(\frac{\hat{J}_g}{n\hat{I}_g^2}\right)^{\frac{1}{2}}$$

3.8 MoM vs MLE

Suppose the regularity conditions for asymptotic normality of MLEs holds and that $E_\theta[g(X_i)] = \psi(\theta)$ where ψ is injective.

MoM:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n g(X_i) &= \psi(\tilde{\theta}_n) \\ \tilde{\theta}_n &= \psi^{-1}\left(\frac{1}{n} \sum_{i=1}^n g(X_i)\right) \end{aligned}$$

From the CLT, we know that

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

From the Delta Method,

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

This implies that

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

For the MLE, since

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

we have

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

where \mathcal{N}_2 denotes the bivariate normal distribution and $C(\theta)$ is a 2×2 covariance matrix defined as

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

where $\eta(\theta)$ is further defined

$$\eta(\theta) = \frac{1}{I(\theta)\psi'(\theta)} \text{Cov}_\theta[g(X_i), l'(X_i; \theta)]$$

By the Cauchy-Schwarz,

$$(\text{Cov}_\theta[g(X_i), l'(X_i; \theta)])^2 \leq \text{Var}[g(X_i)]\text{Var}[l'(X_i; \theta)] = \sigma^2(\theta)[\psi'(\theta)]^2 I(\theta)$$

so

$$\eta^2(\theta) \leq \frac{\sigma^2(\theta)}{I(\theta)}$$

Furthermore

$$\begin{aligned} \psi'(\theta) &= \frac{d}{d\theta} \int_A g(x) f(x; \theta) dx \\ &= \int_A g(x) \frac{\partial}{\partial \theta} f(x; \theta) dx \\ &= \int_A g(x) l'(x; \theta) f(x; \theta) dx \\ &= E_\theta[g(X_i) l'(X_i; \theta)] \\ &= \text{Cov}[g(X_i), l'(X_i; \theta)] \quad \text{since } E[l'(X_i; \theta)] = 0 \end{aligned}$$

Thus,

$$\eta(\theta) = \frac{1}{I(\theta)\psi'(\theta)} \text{Cov}[g(X_i), l'(X_i; \theta)] = \frac{1}{I(\theta)}$$

But since we also showed that

$$\eta^2(\theta) \leq \frac{\sigma^2(\theta)}{I(\theta)}$$

this implies that

$$\frac{1}{I^2(\theta)} \leq \frac{\sigma^2(\theta)}{I(\theta)} \implies \text{Var}_\theta(\hat{\theta}) = \frac{1}{nI(\theta)} \leq \frac{\sigma^2(\theta)}{n} = \text{Var}_\theta(\tilde{\theta})$$

which shows the MLE has a lower standard error than MoM

- This means the MLE is a better estimator of than the MoM

Definition. Estimators $(\hat{\theta}_n)_n$ are efficient estimators if

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

3.9 Bayesian Inference

Bayesian approach: Quantify a priori information about θ via probability distributions

- Think of $f(x_1, \dots, x_n; \theta)$ as representing conditional pmf/pdf of (X_1, \dots, X_n) given a value of θ where θ has some probability distribution on Θ

$$f(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n | \theta)$$

The information about θ is given through a **priority density function** $\pi(\theta)$.

Theorem (Bayes'). Suppose θ is discrete-valued with prior pmf $\pi(\theta)$ on $\Theta = \{\theta_1, \dots\}$. Then the posterior density function is

$$\begin{aligned} \pi(\theta_j | x_1, \dots, x_n) &= \frac{\pi(\theta_j)f(x_1, \dots, x_n; \theta_j)}{\sum_k \pi(\theta_k)f(x_1, \dots, x_n; \theta_k)} \\ &= c(x_1, \dots, x_n)\pi(\theta_j)\mathcal{L}(\theta_j) \\ &\propto \pi(\theta_j)\mathcal{L}(\theta_j) \end{aligned}$$

where

$$c(x_1, \dots, x_n) = \left[\sum_k \pi(\theta_k)f(x_1, \dots, x_n; \theta_k) \right]^{-1}$$

For a continuous parameter space, define the posterior density by

$$\begin{aligned} \pi(\theta | x_1, \dots, x_n) &= \frac{\pi(\theta)f(x_1, \dots, x_n; \theta)}{\int_{\Theta} \pi(s)f(x_1, \dots, x_n; s) ds} \\ &= c(x_1, \dots, x_n)\pi(\theta)\mathcal{L}(\theta) \\ &\propto \pi(\theta)\mathcal{L}(\theta) \end{aligned}$$

where

$$c(x_1, \dots, x_n) = \left[\int_{\Theta} \pi(s) f(x_1, \dots, x_n; s) ds \right]^{-1}$$

The reason we have a function $c(x)$ is to normalize $\pi(\theta)\mathcal{L}(\theta)$ such that $\pi(\theta | x)$ is a pdf/pmf.

The general framework of Bayesian inference is to use the prior information on the parameter θ in the form of the prior distribution and then observing the information on θ after collecting data (through the posterior distribution).

If we have multiple parameters $(\theta_1, \dots, \theta_k)$, the general framework is the same.

$$\pi(\theta_1, \dots, \theta_k | x_1, \dots, x_n) = c(x_1, \dots, x_n) \pi(\theta_1, \dots, \theta_k) \mathcal{L}(\theta_1, \dots, \theta_k)$$

3.9.1 Choice of prior

- If Θ is bounded, we can set $\pi(\theta) = d$ for some constant d for $\theta \in \Theta$

Conjugate priors

Given a model, choose a prior density such that the posterior density has the same form as the prior

$$\pi_{\alpha}(\theta) \xrightarrow{\text{data}} \pi_{\alpha'}(\theta | x_1, \dots, x_n)$$

where α' is dependent on (x_1, \dots, x_k) .

3.9.2 Bayesian Interval Estimation

Definition. Given a posterior density $\pi(\theta | x_1, \dots, x_n)$, an interval $\mathcal{I} = \mathcal{I}(x)$ is a $100p\%$ credible interval for θ if

$$\int_{\mathcal{I}(X)} \pi(\theta | x_1, \dots, x_n) d\theta = p$$

Definition. A $100p\%$ credible interval \mathcal{I} is a 100% highest posterior density for θ if for all $\theta \in \mathcal{I}$ and $\theta' \notin \mathcal{I}$,

$$\pi(\theta | x_1, \dots, x_n) > \pi(\theta' | x_1, \dots, x_n)$$

Note that the coverage of a confidence interval is defined in terms of the distribution of X while the coverage of a credible interval is defined in terms of the posterior density.

3.10 Bias and Variance Tradeoffs

Recall that the mean squared error is defined as

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

which can be rewritten as

$$\text{Var}_{\theta}(\hat{\theta}) + \text{Bias}_{\theta}(\hat{\theta})^2$$

In a parametric model, the variance component of the MSE is typically much larger than the bias squared component. However, if the model is a poor approximation or contains outliers, then the bias may pose an issue. A bias-variance tradeoff usually occurs in a non-parametric estimation where tuning parameters are being used. To combat this, we can use a “divide and conquer” method: Divide the data into k subsets of size m and define

$$\bar{\theta} = \frac{1}{k} \sum_{i=1}^k \hat{\theta}_i$$

where $\hat{\theta}_i$ is based on the i th subset.

3.11 Non-parametric Regression

Model: $Y_i = g(X_i) + \varepsilon_i$ for $i = 1, \dots, n$ where g is unknown but smooth, the ε_i are independent with mean 0 and variance σ^2 . We estimate $g(x)$ by

$$\hat{g}(x) = \sum_{i=1}^n w_i(x) Y_i$$

where $\sum_{i=1}^n w_i(x) = 1$. This implies

$$\begin{aligned} E[\hat{g}(x)] &= \sum_{i=1}^n w_i(x) E(Y_i) = \sum_{i=1}^n w_i(x) g(x_i) \\ \text{Var}[\hat{g}(x)] &= \sum_{i=1}^n w_i^2(x) \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^n w_i^2(x) \end{aligned}$$

Thus,

$$\text{Bias}[\hat{g}(x)] = \sum_{i=1}^n w_i(x) [g(x_i) - g(x)]$$

To make bias small, we set $w_i(x)$ close to 0 when $|g(x_i) - g(x)|$ is large and $w_i(x)$ larger when the distance is small. Since we assume g is smooth, the size of $|g(x_i) - g(x)|$ directly relates to $|x_i - x|$, which minimizes error in our regression.

3.11.1 Smoothing Matrices

In matrix form,

$$\hat{\mathbf{g}} = \begin{bmatrix} \hat{g}(x_1) \\ \vdots \\ \hat{g}(x_n) \end{bmatrix} = \begin{bmatrix} w_1(x_1) & \cdots & w_n(x_1) \\ \vdots & \ddots & \vdots \\ w_1(x_n) & \cdots & w_n(x_n) \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \mathbf{S}\mathbf{Y}$$

where \mathbf{S} is the smoothing matrix. Then

$$\hat{\mathbf{g}} - \mathbf{g} = \mathbf{S}(\mathbf{g} + \boldsymbol{\varepsilon}) - \mathbf{g} = (\mathbf{S} - \mathbf{I})\mathbf{g} + \mathbf{S}\boldsymbol{\varepsilon}$$

which represents the bias-variance trade-off (the $(\mathbf{S} - \mathbf{I})\mathbf{g}$ term represents bias, the $\mathbf{S}\boldsymbol{\varepsilon}$ represents variance).

4 Hypothesis Testing

A hypothesis test involves comparing a simple model to a more complicated one

- Comparing the null hypothesis, H_0 , to the alternate hypothesis, H_1

There are 2 types of error:

Type I Rejecting H_0 in favour of H_1 when H_0 is true

Type II Accepting H_0 when H_1 is true

The goal of a classical hypothesis test is to minimize the probability of Type II error subject to a bound on the probability of a Type I error.

4.1 Power of a Test

Definition. The power of a test is the probability of rejecting H_0 .

If H_0 is true, then

$$\text{power} = P(\text{Type I error})$$

If $P(\text{type I error}) \leq \alpha$, then α is the level/size of the test.

- If H_1 is true, then $\text{power} = 1 - P(\text{Type II error})$

4.1.1 Formal Setup of a Hypothesis Test

Let $X = (X_1, \dots, X_n)$ have a joint pdf/pmf $f(x; \theta)$ for some $\theta \in \Theta$. Split $\Theta = \Theta_0 \cup \Theta_1$ where Θ_0 and Θ_1 are disjoint. We want to test

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$$

at level α .

Define the power function $\text{power}(\theta) \leq P_\theta[\text{reject } H_0]$.

- If $\theta \in \Theta_0$, then $\text{power}(\theta) \leq \alpha$
- If $\theta \in \Theta_1$, then $\text{power}(\theta) = 1 - P_\theta[\text{type II error}]$

Define $\varphi(X)$ to take values 0 and 1 such that

$$\varphi(X) = 0 \implies \text{accept } H_0$$

$$\varphi(X) = 1 \implies \text{reject } H_0$$

thus the power(θ) = $E_\theta[\varphi(X)] = P_\theta(\varphi(X) = 1)$. So, we want φ such that $P_\theta(\varphi(X) = 1) \leq \alpha$ for all $\theta \in \Theta_0$.

The goal is the maximize power $E_\theta[\varphi(X)]$ for $\theta \in \Theta_1$ over all α -level test functions.

Example:

Consider the following simple model: Suppose Θ takes only 2 possible values $\{\theta, \theta_1\}$. Define $f_0(x) = f(x; \theta_0)$ and $f_1(x) = f(x; \theta_1)$ and consider the test

$$H_0 : X \sim f_0(x) \quad H_1 : X \sim f_1(x)$$

If $\varphi(X)$ is an α -level test function, then φ satisfies

$$E_0[\varphi(X)] = \int \varphi(x) f_0(x) dx \leq \alpha$$

Thus, we want to find φ that maximizes

$$E_1[\varphi(X)] = \int \varphi(x) f_1(x) dx$$

subject to the above constraint.

4.2 Neyman-Pearson Lemma

Suppose Θ takes only 2 possible values $\{\theta, \theta_1\}$. Define $f_0(x) = f(x; \theta_0)$ and $f_1(x) = f(x; \theta_1)$ and consider the test

$$H_0 : X \sim f_0(x) \quad H_1 : X \sim f_1(x)$$

For some finite $k > 0$, define the test function

$$\varphi^*(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} \geq k \\ 0 & \text{if } \frac{f_1(x)}{f_0(x)} < k \end{cases}$$

and set $\alpha = E_0[\varphi^*(X)]$. φ^* maximizes $E_1[\varphi(X)]$ over all test functions φ with $E_0[\varphi(X)] \leq \alpha$.

- In other words, φ^* is the most powerful (MP) α -level test of H_0 vs. H_1
- The **test statistic** is

$$T = \frac{f_1(x)}{f_0(x)} = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)}$$

In practice, we want k such that $\alpha = P\left(\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \geq k\right)$ to define the MP α -level test

Consider the following one-sided testing problem: Let X_1, \dots, X_n be independent Exponential random variables with parameter λ . We want to test

$$H_0 : \lambda \leq \lambda_0 \quad H_1 : \lambda > \lambda_0$$

First, we look at

$$H'_0 : \lambda \leq \lambda_0 \quad H'_1 : \lambda = \lambda_1 > \lambda_0$$

The N-P lemma gives us the MP α -level test of H'_0 vs H'_1 . For this test, the test statistic is

$$T(X) = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left[(\lambda_0 - \lambda_1) \sum_{i=1}^n X_i\right]$$

Note that $T(X)$ is a decreasing function of $\sum_{i=1}^n X_i$ since $\lambda_0 < \lambda_1$, thus the MP test of H'_0 vs H'_1 rejects H'_0 at level α if

$$\sum_{i=1}^n X_i \leq k \text{ where } P_{\lambda_0} \left[\sum_{i=1}^n X_i \leq k \right] = \alpha$$

Since

1. The MP α -level test of H'_0 vs H'_1 is the same for all $\lambda_1 > \lambda_0$
2. For $\lambda \leq \lambda_0$,

$$P_\lambda \left[\sum_{i=1}^n X_i \leq k \right] \leq \alpha$$

the MP α -level test of H'_0 vs H'_1 is (from 2. above) an α -level test of H_0 vs. H_1 , and furthermore a uniformly most powerful (UMP) α -level test of H_0 vs H_1 .

Now suppose the scenario is a two-sided test:

$$H_0 : \lambda = \lambda_0 \quad H_1 : \lambda \neq \lambda_0$$

In this case, the MP α -level test of

$$H'_0 : \lambda = \lambda_0 \quad H'_1 : \lambda = \lambda_1$$

depends on if $\lambda_1 > \lambda_0$ or $\lambda_1 < \lambda_0$.

- If $\lambda_0 < \lambda_1$, we reject if $\sum_{i=1}^n X_i \leq k_1$
- If $\lambda_0 > \lambda_1$, we reject if $\sum_{i=1}^n X_i \geq k_2$

This shows there is no UMP test of H_0 vs. H_1 .

4.2.1 Likelihood Ratio Test

Let $X = (X_1, \dots, X_n)$ have pdf/pmf $f(x; \theta)$ where $\theta = (\theta_1, \dots, \theta_n) \in \Theta$. Suppose Θ_0 is a lower-dimensional subset of Θ . Define the Likelihood Ratio Statistic

$$\Lambda = \frac{\sup_{\theta \in \Theta} f(x; \theta)}{\sup_{\theta \in \Theta_0} f(x; \theta)} = \frac{f(x; \hat{\theta})}{f(x; \hat{\theta}_0)}$$

where $\hat{\theta}$ is the MLE of θ and $\hat{\theta}_0$ is the MLE under H_0 .

Reject H_0 at level α for $\Lambda \geq k_\alpha$ where

$$\sup_{\theta \in \Theta_0} P_\theta(\Lambda \geq k_\alpha) \leq \alpha$$

- $2 \ln(\Lambda) \sim \chi_{p-r}^2$ where $p = \dim(\Theta)$ and $r = \dim(\Theta_0)$

Example:

Suppose that $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and we want to test

$$H_0 : \mu = 0 \quad H_1 : \mu \neq 0$$

at level α . The likelihood function is given by

$$\mathcal{L}(\mu, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

The unrestricted MLEs are

$$\hat{\mu} = \bar{X} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

while under H_0 , the MLEs are

$$\hat{\mu}_0 = 0 \quad \hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

thus the likelihood ratio test statistic is

$$\Lambda = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}} = 1 + \frac{n\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = 1 + \frac{1}{n-1} \left(\frac{\bar{X}}{S/\sqrt{n}} \right)^2$$

which is an increasing function. Note that the quantity

$$T = \frac{\bar{X}}{S/\sqrt{n}}$$

has Student's t_{n-1} distribution, thus T^2 has Snedcor's $F_{1, n-1}$ distribution. By the LRT, we reject H_0 for large values of H_0 , thus we observe the quantiles of $F_{1, n-1}$ distribution.

4.3 p values

Define tests for all $\alpha \in (0, 1)$.

Definition. The p value is the smallest α for which we reject H_0 (any α -level test for α less than the p value will always accept H_0)

For given data x ,

$$p \text{ value} = P_{H_0}(X \text{ is more extreme than } x) = P_{H_0}(T(X) \geq T(x))$$

where $T = T(X)$ is a test statistic.

Suppose we test $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$. For each $\alpha \in (0, 1)$, define test functions

$$\varphi_\alpha(X) = \begin{cases} 1 & \text{if } X \in \mathcal{R}_\alpha \\ 0 & \text{if } X \notin \mathcal{R}_\alpha \end{cases}$$

where $E_\theta[\varphi_\alpha(X)] = P_\theta(X \in \mathcal{R}_\alpha) \leq \alpha$.

- \mathcal{R}_α is the rejection region of the α -level test

The rejection regions $\{\mathcal{R}_\alpha : \alpha \in (0, 1)\}$ are nested such that $\mathcal{R}_{\alpha_1} \subseteq \mathcal{R}_{\alpha_2}$ if $\alpha_1 < \alpha_2$.

Define the p value as

$$L(x) = \inf\{\alpha : x \in \mathcal{R}_\alpha\} = \inf\{\alpha : \mathcal{R}_\alpha \text{ is nonempty}\}$$

- We often use the p value in favour of H_1 over H_0

sec

4.3.1 Stochastic Ordering

Definition. Suppose F and G are cdfs. F is stochastically greater than G , denoted $F \succeq_s G$ if $F(x) \leq G(x)$ for all x .

- If $F \succeq_s G$ and $G \succeq_s F$, then $F = G$
- In terms of quantiles, $F \succeq_s G$ if $F^{-1}(t) \geq G^{-1}(t)$ for all t

Proposition. Define $L(X)$ to be a p value based on a family of rejection regions $\{\mathcal{R}_\alpha : \alpha \in (0, 1)\}$. Then for any $\theta \in \Theta_0$,

$$L(X) \succeq_s \text{Unif}(0, 1)$$

Corollary. If there exists some $\theta \in \Theta_0$ such that $P_\theta[X \in \mathcal{R}_\alpha] = \alpha$ for all $\alpha \in (0, 1)$, then

$$P_{\theta_0}[L(X) \leq \alpha] = \alpha$$

- This implies that under parameter θ_0 , the p value is $\text{Unif}(0,1)$ distributed
- For good tests, $L(X) \prec_s \text{Unif}(0, 1)$ when H_1 is true

4.3.2 Combining p values

Suppose now that the same H_0 vs H_1 testing problem is considered in m independent studies. Let the p values be L_1, \dots, L_m , which are independent random variables on $(0, 1)$ with distribution G .

- Under H_0 , $G = \text{Unif}(0, 1)$ or $G \succ_s \text{Unif}(0, 1)$
- Under H_1 , $G \prec \text{Unif}(0, 1)$

Thus, we can rewrite the testing problem as

$$H_0 : G = \text{Unif}(0, 1) \quad H_1 : G \prec_s \text{Unif}(0, 1)$$

so we need a test statistic based off L_1, \dots, L_m . Suppose $T = T(L_1, \dots, L_m)$ be such a test statistic. If we reject H_0 when T is large, then $T(l_1, \dots, l_m)$ should be a decreasing function in each l_i .

4.4 Multiple Hypothesis Testing

Consider testing $H_0^{(k)}$ vs $H_1^{(k)}$ for $k = 1, \dots, m$. Let $\varphi_1(x), \dots, \varphi_k(x)$ be the test functions of each test:

$$\varphi_k(X) = \begin{cases} 1 & \text{reject } H_0^{(k)} \\ 0 & \text{accept } H_0^{(k)} \end{cases}$$

and define $\alpha_k = E_0[\varphi_k(x)]$. How can we choose α_k such that false rejections of $H_0^{(k)}$ can be avoided?

4.4.1 Family-Wise Error Rate

Suppose all the null hypotheses are true (i.e.: the global null hypothesis holds). Define the family-wise error rate (FWER) as

$$\begin{aligned} \text{FWER}(\varphi_1, \dots, \varphi_m) &= P_0(\exists k \text{ s.t. } \varphi_k(X) = 1) \\ &= 1 - P_0(\varphi_1(X) = 0, \varphi_2(X) = 0, \dots, \varphi_m(X) = 0) \\ &= 1 - E_0 \left[\prod_{k=1}^m (1 - \varphi_k(X)) \right] \end{aligned}$$

where P_0 and E_0 assume the global null hypothesis holds. We want $\text{FWER}(\varphi_1, \dots, \varphi_m) \leq \alpha$. If the $\varphi_k(X)$ are independent, then

$$\begin{aligned} \text{FWER}(\varphi_1, \dots, \varphi_m) &= 1 - \prod_{k=1}^m E_0[1 - \varphi_k(X)] \\ &= 1 - P_0(\varphi_1(X) = 0, \dots, \varphi_m(X) = 0) \end{aligned}$$

thus we can get $\text{FWER}(\varphi_1, \dots, \varphi_m) = \alpha$ by taking

$$\alpha_k = 1 - (1 - \alpha)^{\frac{1}{m}}$$

Since $P_0(\varphi_k(X) = 1) = \alpha_k$, then $P_0(\varphi_k(X) = 0) = 1 - \alpha_k$, so

$$1 - P_0(\varphi_1(X) = 0, \dots, \varphi_m(X) = 0) = 1 - \prod_{k=1}^m (1 - \alpha_k) = 1 - (1 - \alpha)^{\frac{m}{m}} = \alpha$$

as required. On the other hand, if the $\varphi_k(X)$ cannot be simply assumed to be independent, then we can use **Bonferroni correction**:

$$\begin{aligned} \text{FWER}(\varphi_1, \dots, \varphi_m) &= P_0(\text{at least 1 } \varphi_k(X) = 1) \\ &= P\left(\bigcup_{k=1}^m \{\varphi_k(X) = 1\}\right) \leq \sum_{k=1}^m P(\varphi_k(X) = 1) = \sum_{k=1}^m \alpha_k \end{aligned}$$

Thus, taking $E_0[\varphi_k(X)] = \frac{\alpha}{m}$ guarantees $\text{FWER}(\varphi_1, \dots, \varphi_m) \leq \alpha$.

The issue with controlling the FWER, however, lies in the fact that the assumption is that all null hypotheses hold true. An alternative approach is to control the false discovery rate.

4.4.2 False Discovery Rate

Now we no longer make assumptions about the number of true/false null hypotheses. Define the set

$$\mathcal{S} = \{k : H_0^{(k)} \text{ is true}\}$$

where $\mathcal{S} = \emptyset$ is a possibility. For test functions $\varphi_1, \dots, \varphi_m$, define

$$\begin{aligned} R &= \sum_{k=1}^m \varphi_k(X) = \text{number of rejected null hypotheses} \\ V &= \sum_{k \in \mathcal{S}} \varphi_k(X) = \text{number of falsely rejected null hypotheses} \end{aligned}$$

We want to minimize $\frac{V}{R}$, aka the proportion of rejected null hypotheses that are falsely rejected. For the FWER, we know that $V = R$ since $\mathcal{S} = \{\text{all } H_0^{(k)}\}$, thus

$$\text{FWER}(\varphi_1, \dots, \varphi_m) = P_0(V \geq 1)$$

Define the false discovery rate (FDR) as

$$\text{FDR}(\varphi_1, \dots, \varphi_m) = E\left(\frac{V}{R} \mid R > 0\right) P(R > 0) = E\left(\frac{V}{\max(R, 1)}\right)$$

Once again, we want $\text{FDR}(\varphi_1, \dots, \varphi_m) \leq \alpha$. Order the p values

$$L_{(1)} \leq \dots \leq L_{(m)}$$

Define

$$\hat{k} = \max\left\{k : L_{(k)} \leq \alpha \frac{k}{m}\right\}$$

then reject all null hypotheses corresponding to $L_{(1)}, \dots, L_{(\hat{k})}$.