STAT-4320 Notes

November 9, 2025

Contents

1 Class 1

1.1 The sample mean

Definition 1.1. (Sample mean): Given $X_1, \dots X_n$ iid from F, and $\mathbb{E}[X_i] = \mu$, the sample mean is the random variable defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Remark. The sample mean is random, so it has an expectation

$$\begin{aligned}
\mathbb{E}\left[\bar{X}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \qquad \text{by linearity} \\
&= \mu
\end{aligned}$$

The expectation of the sample mean is the population mean.
The sample mean is an unbiased estimator of the population mean.

Remark.

$$\begin{aligned}
Var\left(\bar{X}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad \text{since } \mathbb{E}[cX_i] = c\mathbb{E}[X_i], Var(cX_i) = c^2 Var(X_i), X_i \text{ indepedent} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Remark. We say that $\bar{X}$ follows a sampling distribution. Think of this as a thought experiment. If a sample of size n is taken many times, we expect to see the sample mean exhibit the above expectation and variability.

1.2 Central Limit Theorem

Theorem 1.2. Theorem (Central Limit Theorem): Let $X_1, \dots X_n$ be iid from arbitrary distribution F wit mean μ and variance σ^2, then as $n \to \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1),$$

OR

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

2 Chapter 2

2.1 Breakdown Point and Efficiency

Recall that sample mean is defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

An issues with the sample mean is that corruption in 1 or few data points can make sample mean unstable. The sample median is more robust alternative.

2.1.1 Sample and population median

Definition 2.1. (Sample median): The sample median is the *middle value* when a list of numbers are sorted in non-decreasing order.

$$X_{med} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ odd} \\ X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)} & \text{if } n \text{even} \end{cases}$$

Where $X_{(i)}$ denotes the i-th smallest value in $X_1, \ldots X_n$

Definition 2.2. (population median): The population median of distribution F with density function f is the point m such that

$$\int_{\infty}^{m} f(x)dx = \int_{m}^{\infty} f(x)dx = \frac{1}{2}$$

2.1.2 Breakdown point

Definition 2.3. (Breakdown Point): The breakdown point of an estimate $\hat{\theta}_n$ based on data $X_1 \ldots X_n$ is the fraction of data points that have to be moved to infinity for the esimate to also move to infinity.

Example. The breakdown point

- For sample mean $= \frac{1}{n}$
- For sample median $\approx \frac{1}{2}$

Remark. Note that this is **not** a direct consequence of CLT.

For example, $F = N\left(\mu, \sigma^2\right)$.

The sample mean follows **exactly** a normal distribution

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The sample median approximately follows

$$X_{med} \approx N\left(\mu, \frac{1}{4f(\mu)^2 n}\right)$$

Recall that

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Hence

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$$

$$X_{med} \approx N\left(\mu, \frac{\pi\sigma^2}{2n}\right)$$

2.1.3 Efficiency

Definition 2.4. (Efficiency): The efficiency of two estimates is the ratio of their variances.

$$\text{Efficiency}\left(\widetilde{X}_{med}, \bar{X}\right) = \frac{Var(\bar{X})}{Var\left(\widetilde{X}_{med}\right)}$$

Example. For sample mean and sample median, the efficnecy is $\frac{2}{\pi}$.

3 Class 3

3.1 Convergence of random variables

There are two kinds of convergence

- convergence in probability
- convergence in distribution

Definition 3.1. (Convegence in probability): We say a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in probability to X if for any $\epsilon > 0$,

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

Denoted

$$X_n \xrightarrow{p} X$$

Definition 3.2. (Convegence in distribution): We say a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in distribution to X if

$$\mathbb{P}\left(X_n \leq x\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(X \leq x\right)$$

Which is equivalent to

$$F_n(x) \underset{n\to\infty}{\longrightarrow} F(x)$$

Denoted

$$X_n \xrightarrow{d} X$$

3.2 Slutsky's Lemma and Continuous Mapping Theorem

Lemma 3.3. Lemma (Slutsky's Lemma): If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for constant c, then the following hold

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n Y_n \xrightarrow{d} cX$
- $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ if $c > 0$

Theorem 3.4. (Continuous mapping): If g is a continuous function, then

$$X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$$

$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$$

3.3 Weak Law of Large Numbers and Central Limit Theorem

The weak law of large numbers is an example of convergence in probability. The central limit theorem is an example of convergence in distribution.

Theorem 3.5. Weak law of large numbers: Suppose $X_1 \dots X_n$ iid from F with $\mathbb{E}\left[X_1\right] = \mu$ and $Var(X) = \sigma^2$, then

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

Proof.

$$\begin{aligned}
\mathbb{P}\left(|\bar{X} - \mu| > \epsilon\right) &= \frac{Var\left(\bar{X}\right)}{\epsilon^2} \quad \text{by Chebyshev's Inequality} \\
&= \frac{\sigma^2}{n\epsilon^2}
\end{aligned}$$

$$\frac{\sigma^2}{n\epsilon^2} \underset{n\to\infty}{\longrightarrow} 0$$

✓

Theorem 3.6. Markov's Inequality: For any random variable X, and non-negative constant a,

$$P(|X| \geq a) \leq \frac{\mathbb{E}\left[|X|\right]}{a}$$

Alternatively, for any non-negative random variable X,

$$P(X \geq a) \leq \frac{\mathbb{E}\left[X\right]}{a}$$

Proof. We prove the general case, for any random variable X, let $Y = |X|$

$$\begin{aligned}
\mathbb{E}\left[Y\right] &= \mathbb{E}\left[Y|Y \geq a\right] P(Y \geq a) + \mathbb{E}\left[Y|Y < a\right] P(Y < a) \text{ by Law of Total Expectation} \\
&\geq \mathbb{E}\left[Y|Y \geq a\right] P(Y \geq a) \\
&\geq aP(Y \geq a) \\
\implies P(Y \geq a) &\leq \frac{\mathbb{E}\left[Y\right]}{a}
\end{aligned}$$

✓

Theorem 3.7. Chebyshev's Inequality: Let X be a random variable with mean μ and variance σ^2, then for any $a > 0$

$$P\left(|X - \mu| \geq a\right) \leq \frac{\sigma^2}{a^2}$$

Proof.

$$\begin{aligned}
P\left(|X - \mu| \geq a\right) &= P\left((X-\mu)^2 \geq a^2\right) \\
&\leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{a^2} \quad \text{by Markov's Inequality} \\
&= \frac{Var(X)}{a^2}
\end{aligned}$$

✓

Theorem 3.8. Central limit theorem: Suppose $X_1 \dots X_n$ iid from F with $\mathbb{E}\left[X_1\right] = \mu$ and $Var(X) = \sigma^2$, then

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

Alternatively, let

$$Z_n = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma}$$

$$\mathbb{P}\left(Z_n \leq t\right) = P(N(0,1) \leq t) \text{ for all } t \in \mathbb{R}$$

Remark.

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Although this statement has no mathematical content. We are taking the limit of the distribution of $\bar{X}$ as n gets large. $N\left(\mu, \frac{\sigma^2}{n}\right)$ cannot be a limit.

3.4 Delta method

CLT gives asymptotic distribution of $\bar{X}$. We want to get the asymptotic distribution of functions of $\bar{X}$.
By Continuous Mapping Theorem, we get this for free

$$\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \xrightarrow{d} N(0,1) \implies g\left(\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma}\right) \xrightarrow{d} g(N(0,1))$$

However, we don't just want statements about $g(Z_n)$, we want statements about $g(\bar{X})$

Theorem 3.9. Delta Method: Suppose $X_1, \dots X_n$ iid F, with $\mathbb{E}\left[X_1\right] = \mu$ and $Var(X) = \sigma^2$ and g is a function such that the derivative of $g'(\mu) \neq 0$. Then

$$\sqrt{n}\left(g(\bar{X}) - g(\mu)\right) \xrightarrow{d} N\left(0, \sigma^2 \left(g'(u)\right)^2\right)$$

Remark. Note: We know by Continuous Mapping Theorem that the in-probability limit of $g(\bar{X})$ is $g(\mu)$

$$g(\bar{X}) \xrightarrow{p} g(\mu)$$

Subtracting away the in-probability limit and taking the Z-score, delta method tells us that the z-score follows a normal distribution.

Proof.
Recall Taylor's Expansion

$$f(x) = f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a) + \dots$$

$$g(\bar{X}) - g(\mu) = (\bar{X} - \mu)g'(\mu) + \text{ error terms}$$
$$\sqrt{n}\left(g(\bar{X}) - g(\mu)\right) = \sqrt{n}(\bar{X} - \mu)g'(\mu) + \text{ error terms}$$

By CLT, we know that

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

By Slutsky's,

$$\sqrt{n}(\bar{X} - \mu)g'(\mu) \xrightarrow{d} N(0, \sigma^2)g'(\mu) = N(0, \sigma^2(g'(\mu))^2)$$

Note: if $g'(\mu) = 0$, by taking higher orders in the Taylor expansion, we get

$$g(\bar{X}) - g(\mu) \approx \frac{1}{2}\left(\bar{X} - \mu\right)^2 g''(\mu)$$

Since

$$n\left(\bar{X} - \mu\right)^2 = \left(\sqrt{n}\left(\bar{X} - \mu\right)\right)^2 \xrightarrow{d} \left(\sigma N(0,1)\right)^2 = \sigma^2 \chi_1^2$$

We get

$$n\left(g(\bar{X}) - g(\mu)\right) \xrightarrow{d} \frac{1}{2}\sigma^2 \chi_1^2 g''(\mu)$$

✓

3.5 Multivariate Data

For each unit of study, the number of measurements is greater than 1. For example, for n data points and p observed variables

$$\mathbf{X}_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{pmatrix}, \mathbf{X}_2 = \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2p} \end{pmatrix}, \mathbf{X}_3 = \begin{pmatrix} X_{31} \\ X_{32} \\ \vdots \\ X_{3p} \end{pmatrix}$$

Where $\mathbf{X}_i$ are iid p-dimentional observations with distribution F.
The mean vector is

$$\boldsymbol{\mu} = \mathbb{E}\left[\mathbf{X}_1\right] = \begin{pmatrix} \mathbb{E}\left[X_{11}\right] \\ \mathbb{E}\left[X_{12}\right] \\ \vdots \\ \mathbb{E}\left[X_{1p}\right] \end{pmatrix} \in \mathbb{R}^p$$

The covariance matrix is denoted

$$\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$$

Where the $i-j$-th element is

$$\boldsymbol{\Sigma}_{i,j} = Cov(X_{1i}, X_{1j}) = \mathbb{E}\left[X_{1i}X_{1j}\right] - \mathbb{E}\left[X_{1i}\right]\mathbb{E}\left[X_{1j}\right]$$

Hence, we can epxress Σ as the difference of two matrices

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{X}_1 \boldsymbol{X}_1^T\right] - \mathbb{E}\left[\boldsymbol{X}\right]\mathbb{E}\left[\boldsymbol{X}\right]^T$$

3.6 Multivariate Normal

Definition 3.10. (Multivariate Normal): A p-dimensional random vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots X_p \end{pmatrix}$ is said to follow the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and positive definite (pd) covariance matrix $\boldsymbol{\Sigma}$ if it has a density function $f : \mathbb{R}^p \to \mathbb{R}$ of the form

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Remark. This definition only works with a *pd* covariance matrix.

Remark. A multivariate normal can be defined without the *pd* matrix, using the linear combination definition.

Definition 3.11. (Multivariate CLT): Suppose $\boldsymbol{X}_1 \dots \boldsymbol{X}_n$ are iid p-dimensional random vectors with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma}$
The sample mean

$$\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

has the following distribution

$$\sqrt{n} \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu} \right) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$$

4 Class 4

4.1 Linear Algebra Review

4.1.1 Positive definite matrices

Definition 4.1. Definition (positive definite): A symmetric $p \times p$ matrix is said to be positive definite (*pd*) if for all $\boldsymbol{x} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$,

$$\boldsymbol{x}^T A \boldsymbol{x} > 0$$

Remark. All eigenvalues of a *pd* matrix are positive

Remark. By spectral decomposition, any *pd* matrix can be written as

$$A = P \Lambda P^T,$$

Where Λ is a diagonal matrix with the eigenvalues of A on the diagonals

$$\Lambda = \begin{bmatrix} \lambda_1 & \dots & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ & \vdots & & \\ 0 & \dots & \dots & \lambda_p \end{bmatrix}$$

and P is an orthogonal matrix

$$P^T P = P P^T = I_{p \times p}$$

Remark. A^{-1} exists and is given by

$$A^{-1} = P \Lambda^{-1} P^T$$

Proof.

$$A^{-1} A = P \Lambda^{-1} P^T P \Lambda P^T = P \Lambda^{-1} \Lambda P^T = P P^T = I$$

✓

Remark. A has a square root. Given a *pd* matrix A, we say that B is the square root of A if $BB = A$

$$B = P \begin{bmatrix} \sqrt{\lambda_1} & \dots & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ & \vdots & & \\ 0 & \dots & \dots & \sqrt{\lambda_p} \end{bmatrix} P^T$$

Remark. Sum of eigenvalues is the trace of A

$$\sum_{i=1}^{n} \lambda_i = tr(A)$$

Remark. What does it mean to assume that the covariance matrix is *pd*?

Consider any $\boldsymbol{a} \in \mathbb{R}^p \setminus \{0\}$. Recall that

$$\begin{aligned} \Sigma &= \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right] - \mathbb{E}\left[\boldsymbol{X}\right]\mathbb{E}\left[\boldsymbol{X}\right]^T \\ &= \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{x}\right])(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right])^T\right] \end{aligned}$$

Consider $\boldsymbol{a}^T \Sigma \boldsymbol{a}$,

$$\begin{aligned} \boldsymbol{a}^T \Sigma \boldsymbol{a} &= \mathbb{E}\left[(\boldsymbol{a}^T(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right]))(\boldsymbol{a}^T(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right]))^T\right] \text{ since } (AB)^T = B^T A^T \\ &= \mathbb{E}\left[(a^T(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right]))^2\right] \\ &= Var(\boldsymbol{a}^T X) \\ &> 0 \end{aligned}$$

i.e. *projected onto any direction $\boldsymbol{a}$, the variance of $\boldsymbol{X}$ is nonzero.*

i.e. The RV is non-degenerate along every direction $\boldsymbol{a} \in \mathbb{R}^p \setminus \{0\}$.

4.2 Moment generating functions

Definition 4.2. (Moment generating function) For a random variable X, the *mgf* is a function $\mathbb{R} \to \mathbb{R}_{\geq 0}$,

$$\phi_X(t) = \mathbb{E}\left[e^{tX}\right]$$

Example. For $X \sim N(\mu, \sigma^2)$

$$\phi_X(t) = \exp\left(t\mu + \frac{1}{2}\sigma^2 t^2\right)$$

If X is a p-dimensional random variable, the *mgf* is a function $\mathbb{R}^p \to \mathbb{R}_{\geq 0}$

$$\phi_X(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}\right)$$

4.3 Properties of the multivariate normal

Proposition 4.3. If $\boldsymbol{X}$ is a p-dimensional normal, $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, and $\boldsymbol{A}$ is a $k \times p$ matrix such that $rank(\boldsymbol{A}) = k \leq p$ (i.e. full row rank) and $\boldsymbol{b} \in \mathbb{R}^k$ is a fixed vector,

$$\boldsymbol{AX} + \boldsymbol{b} \sim N_k\left(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T\right)$$

Example. An important case of this is when $\boldsymbol{b} = \boldsymbol{0}$ and $k = 1$, then $\boldsymbol{A}$ is a row vector $\boldsymbol{A} = [a_1, a_2, \ldots, a_p]$. Take any $\boldsymbol{a} \in \mathbb{R}^p$, then

$$\boldsymbol{a}^T\boldsymbol{X} \sim N_1(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{\Sigma a})$$

This can also be expressed as

$$\boldsymbol{a}^T\boldsymbol{X} = \sum_{i=1}^p a_i X_i \sim N_1\left(\sum_{i=1}^p a_i\mu_i, \boldsymbol{a}^T\boldsymbol{\Sigma a}\right)$$

Proof. Exercise for enthusiasts: Prove 1 using *mgfs*. ✓

Proposition 4.4. Suppose $\boldsymbol{X}$ is a p_1-dimensional RV and $\boldsymbol{Y}$ is a p_2-dimensional RV, such that

$$\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix} \in \mathbb{R}^{p_1+p_2}, \begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix} \sim N_{p_1+p_2}\left(\begin{pmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}}\\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}}\end{pmatrix}\right)$$

where $\boldsymbol{\mu_1} \in \mathbb{R}^{p_1}, \boldsymbol{\mu_2} \in \mathbb{R}^{p_2}$, and

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}}\\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}}\end{pmatrix} \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}$$

then we say that

$$\boldsymbol{X} \perp \boldsymbol{Y} \Leftrightarrow \boldsymbol{\Sigma_{11}} = \boldsymbol{\Sigma_{21}^T} = \boldsymbol{0}$$

Proof. The forward direction is obvious.

The converse is not true in general for a 1-dimensional normal, but it is true in multivariate normal. ✓

Example. Important case: $p_1 = p_2 = 1$, Suppose

$$\begin{pmatrix}X\\ Y\end{pmatrix} \sim N\left((\mu_1, \mu_2), \begin{bmatrix}\sigma_1^2 & \sigma_{21}\\ \sigma_{12}^2 & \sigma_2^2\end{bmatrix}\right)$$

Then $X \perp Y$ if and only if $\sigma_{12} = 0$.

Remark. Relationship

- If $X \perp Y$, then $Cov(X, Y) = 0$. This is always true

$$X \perp Y \implies Cov(X, Y) = 0$$

- The converse is not true in general
- However, if (X, Y) are jointly normal or jointly bernoulli, then

$$Cov(X, Y) = 0 \implies X \perp Y$$

4.4 Sample Variance

The sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Has a sampling distribution

$$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

5 Class 5

5.1 Chi-squared distribution

Definition 5.1. The χ^2_d is the sum of d iid standard normals squared.

$$Z_1, \dots Z_d \sim N(0,1), \quad \sum_{i=1}^{d} Z_i \sim \chi^2_d$$

Result 5.2. The *chi-squared* distribution has expectation

$$\begin{aligned}\mathbb{E}\left[\chi^2_d\right] &= d\mathbb{E}\left[Z_1^2\right] \\ &= d\end{aligned}$$

Proof.

$$\mathbb{E}\left[Z_1^2\right] = Var(Z_1) - \mathbb{E}\left[Z_1\right]^2$$

✓

Result 5.3. The *chi-squared* distribution has variance

$$\begin{aligned}Var(\chi^2_d) &= Var\left(\sum_{i=1}^{d} Z_i^2\right) \\ &= \sum_{i=1}^{d} Var(Z_i^2) \\ &= 2d\end{aligned}$$

Proof.

$$\begin{aligned}Var(Z_1^2) &= \mathbb{E}\left[Z_1^4\right] - \mathbb{E}\left[Z_1^2\right]^2 \\ &= \mathbb{E}\left[Z_1^4\right] - 1 \\ &= 3 - 1 \\ &= 2\end{aligned}$$

✓

5.2 Sample Variance, cont'd

Result 5.4. The sample variance follows a *chi-squared* distribution.

$$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

Proof. Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}$$

We express $\boldsymbol{Y} = \boldsymbol{AX}$ for some matrix $\boldsymbol{A}$.

$$\boldsymbol{Y} = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} \end{pmatrix} \boldsymbol{X}$$

Since

$$\boldsymbol{X} \sim N_n\left(\mu\mathbf{1}, \sigma^2\boldsymbol{I}\right)$$

Hence

$$\begin{aligned}\boldsymbol{Y} &\sim N_n\left(\boldsymbol{A}\boldsymbol{\mu}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T\right)\\ &= N_n(\mu\boldsymbol{A}\boldsymbol{1}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T)\\ &= N_n(\boldsymbol{0}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T) \text{ since } \boldsymbol{A}\boldsymbol{1} = \boldsymbol{0}\\ &= N_n(\boldsymbol{0}, \sigma^2\boldsymbol{A}^2) \text{ since } \boldsymbol{A} \text{ symmetric}\end{aligned}$$

Note that

$$\boldsymbol{A} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$$

Where $\boldsymbol{1}\boldsymbol{1}^T$ is a rank 1 symmetric matrix. Hence $\boldsymbol{1}\boldsymbol{1}^T$ has at most 1 non-zero eigenvalue. Since the sum of eigenvalues is the trace, and $tr(\boldsymbol{1}\boldsymbol{1}^T) = n$, the non-zero eigenvalue is n. $\boldsymbol{1}\boldsymbol{1}^T$ has eigenvalues $(n, 0, 0, \ldots, 0)$. $\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$ has eigenvalues $(1, 0, 0, \ldots)$.

Therefore, $\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^n$ has eigenvalues $(1, 1, \ldots, 1, 0)$ (This only works because of $\boldsymbol{I}$).

We can express s^2 in terms of Y.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\boldsymbol{Y}^T\boldsymbol{Y}$$

It suffices to show that

$$\boldsymbol{Y}^T\boldsymbol{Y} \sim \sigma^2\chi^2_{n-1}$$

Note: Y has elements that are normal, i.e. $Y_i = X_i - \bar{X}$ is a difference of normals. However, Y_i is not independent due to $\bar{X}$.

Fact: Denote $\boldsymbol{\Sigma} = \boldsymbol{A}^2$, and we define $\boldsymbol{Z} = N_n(\boldsymbol{0}, \boldsymbol{I})$. Then,

$$Y \overset{d}{=} \sigma\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{A})$$

This is true because

$$\begin{aligned}\sigma\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} &\sim N_n\left(\boldsymbol{0}, \sigma^2\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^T\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\\ &= N_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{A}^T\boldsymbol{A}\right)\\ &= N_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{A}^2\right)\end{aligned}$$

Hence

$$\begin{aligned}\boldsymbol{Y}^T\boldsymbol{Y} &= \sigma^2\boldsymbol{Z}^T\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^T\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}\\ &= \sigma^2\boldsymbol{Z}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{Z}\\ &= \sigma^2\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}\end{aligned}$$

Hence it suffices to show that

$$\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_{n-1}$$

Note that if $\boldsymbol{A} = \boldsymbol{I}$, then $\boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2_n$.

By spectral decomposition,

$$\boldsymbol{Z}\boldsymbol{A}^T\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T\boldsymbol{Z}$$

Denote $\boldsymbol{P}^T\boldsymbol{Z} = \boldsymbol{W} \in \mathbb{R}^n$.

$$\boldsymbol{Z}\boldsymbol{A}^T\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T\boldsymbol{Z} = \boldsymbol{W}^T\boldsymbol{\Lambda}\boldsymbol{W}$$

Because of spectral decomposition, we know that $\boldsymbol{P} and \boldsymbol{P}^T$ are orthogonal matrices whose product is the identity. Applying an orthogonal matrix to a multivariate standard normal, i.g. $\boldsymbol{A}\boldsymbol{Z}$, does not change the multivariate standard normal distribution.

$$\boldsymbol{W} \sim N_n(\boldsymbol{0}, \boldsymbol{P}\boldsymbol{P}^T) = N_n(\boldsymbol{0}, \boldsymbol{I})$$

Hence

$$\begin{aligned}\boldsymbol{W}^T\boldsymbol{\Lambda}\boldsymbol{W} &= \boldsymbol{W}^T\begin{pmatrix}1 & 0 & \ldots & 0 & 0\\ 0 & 1 & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & 0\\ 0 & 0 & \ldots & 1 & 0\\ 0 & 0 & \ldots & 0 & 0\end{pmatrix}\boldsymbol{W}\\ &= \sum_{i=1}^{n-1}W_i^2\\ &\sim \chi^2_{n-1} \quad \checkmark\end{aligned}$$

✓

5.3 Joint distribution of sample mean and sample variance

We know the marginal distributions for the sample mean and variance

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

Result 5.5. If $X_1, X_2, \dots X_n$ iid normal, $\bar{X}, s^2$ independent.

Proof. Define

$$\boldsymbol{Z} = \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \\ \bar{X} \end{pmatrix}$$

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{X}, \boldsymbol{B} \in \mathbb{R}^{(n+1)\times n}, \boldsymbol{B} = \begin{pmatrix} & & \boldsymbol{A} & \\ & & & \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$$

Hence $\boldsymbol{Z}$ is $(n+1)$-dimensional normal.

If we show that $X_1 - \bar{X} \perp \bar{X}$, $X_2 - \bar{X} \perp \bar{X} \dots X_n - \bar{X} \perp \bar{X}$, then

$$\begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots X_n - \bar{X} \end{pmatrix} \perp \bar{X} \implies s^2 \perp \bar{X}$$

Since $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{X}$ is multivariate normal, it suffices to check that the covariance is 0.

To show $Cov(\bar{X}, X_i - \bar{X}) = 0$ for all $1 \le i \le n$.
Note that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} a_i X_i, \quad a_1 = a_2 = \dots a_n = \frac{1}{n}$$

$$X_1 - \bar{X} = \sum_{i=1}^{n} b_i X_i, \quad b_1 = 1 - \frac{1}{n}, b_2 = b_3 = \dots b_n = -\frac{1}{n}$$

$$\begin{aligned} Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j Cov(X_i X_j) \\ &= \sum_{i=1}^{n} a_i b_i Var(X_i) \\ &= \sigma^2 \sum_{i=1}^{n} a_i b_i \\ &= 0 \end{aligned}$$

✓

Remark. As a general strategy, to show independence, we can show in two steps

1. show jointly normal
2. show 0 covariance

6 Class 6

6.1 Basic Framework of Statistical Estimation

Given iid samples $X_1, X_2, \dots X_n$, how do we infer / estimate parameters of F?

Example. (Bernoulli): Given $X_1, X_2, \dots, X_n$ iid $Ber(p)$,

$$X_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p \end{cases}$$

An estimate of p is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{T}{n}$$

Where T is the number of heads and $T \sim Binom(n, p)$.

Example. (Normal): $X_1, X_2, \dots X_n$ iid $N(\mu, \sigma^2)$. Estimates for parameters are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

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

6.2 Properties of Estimators

Definition 6.1. (Unbiasedness): An estimate $\hat{\theta}$ is said to be unbiased for a parameter θ if

$$\mathbb{E}[\hat{\theta}] = \theta$$

for all values of the population.

Example. (Bernoulli):

$$\mathbb{E}\left[\hat{p}\right] = \frac{\mathbb{E}\left[T\right]}{n} = \frac{np}{n} = p$$

Hence $\hat{p}$ is an unbiased estimate of p.

Note that $\hat{p}^2$ s not an unbiased estimate of p^2.

$$\begin{aligned}
\mathbb{E}\left[\hat{p}^2\right] &= \mathbb{E}\left[T^2\right]\left(\frac{1}{n^2}\right) \\
&= \left(np(1-p) + n^2p^2\right)\left(\frac{1}{n^2}\right) \\
&= p^2 + \frac{p(1-p)}{n} \\
&\neq p^2
\end{aligned}$$

Note that we can rearrange terms to get

$$\begin{aligned}
\mathbb{E}\left[T^2\right] &= n^2p^2 + np(1-p) \\
&= (n^2 - n)p^2 + np \\
\frac{\mathbb{E}\left[T^2\right]}{n(n-1)} &= p^2 + \frac{p}{n-1}
\end{aligned}$$

We try estimating $\frac{p}{n-1}$ by $\frac{T}{(n-1)n}$.

Consider the estimate

$$\tilde{p} = \frac{T^2}{n(n-1)} - \frac{T}{n(n-1)} = \frac{T(T-1)}{n(n-1)}$$

The expectation is

$$\mathbb{E}\left[\tilde{p}\right] = p^2 + \frac{p}{n-1} - \frac{p}{n-1} = p^2$$

Note that (Proof left as exercise)

$$\mathbb{E}\left[\frac{T(T-1)}{n(n-1)}\right] = \sum_{r=2}^{n} \frac{r(r-1)}{n(n-1)} \binom{n}{r} p^r \left(1-p\right)^{n-r} = p^2$$

In general, an unbiased estimate for p^k is

$$\frac{T(T-1)(T-2)\dots(T-k+1)}{n(n-1)(n-2)\dots(n-k+1)}$$

An unbiased estimate of $2p^2 + 5p^3$ is

$$2\frac{T(T-1)}{n(n-1)} + 5\frac{T(T-1)(T-2)}{n(n-1)(n-2)}$$

Example. (Normal): Estimate σ^2 with

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sim \chi^2_{n-1}\left(\frac{\sigma^2}{n-1}\right)$$

The expectation of sample variance is

$$\begin{aligned}\mathbb{E}\left[s^2\right] &= \frac{\sigma^2}{n-1}\mathbb{E}\left[\chi^2_{n-1}\right] \\ &= \frac{\sigma^2}{n-1}(n-1) \\ &= \sigma^2\end{aligned}$$

Example. (Network sampling)
A *network* refers to a graph of vertices that are connected if they have an interaction.

Network sampling helps with understanding *how networks look* by studying a small section of the network, and understanding *features of a large unobserved network* from a sample subgraph.

Motifs refer to patterns of small subgraphs, such as an edge, or a triangle.

Motif estimation refers to estimating the number of a particular type of motif, based on an observed sample subgraph.

Let G_n be a population graph on n-vertices. Our subgraph sampling model involves sampling each vertex of G_n with probability $p \in (0,1)$ independently, and then observing the subgraph on the set of sampled vertices.

Goal: estimate the number of edges in G_n based on the observed graph.

- initial guess: count the number of edges in the observed graph
- Denote $\hat{E}(G_n)$ as the number of edges in observed graph
- Denote $E(G_n)$ as number of edges in population graph

Result:

$$\mathbb{E}\left[\hat{E}(G_n)\right] = p^2$$

Hence,

$$\frac{\hat{E}(G_n)}{p^2} = \frac{\text{\# edges in observed graph}}{p^2}$$

is an unbiased estimate of the number of edges in the population.

proof: Note that

$$E(G_n) = \sum_{1\le i\le j\le n} a_{ij}$$

Where

$$a_{ij} = \begin{cases} 1 \text{ if } (i,j) \text{ edge in } G_n \\ 0 \text{ otherwise} \end{cases}$$

Denote S the set of sampled vertices.

$$\begin{aligned}
\hat{E}(G_n) &= \sum_{1\le i\le j\le n, i\in S, j\in S} a_{ij} \\
&= \sum_{1\le i\le j\le n} a_{ij}\mathbf{1}[i\in S]\mathbf{1}[j\in S] \\
\mathbb{E}\left[\hat{E}(G_n)\right] &= a_{ij}\sum_{1\le i\le j\le n} \mathbb{E}\left[\mathbf{1}[i\in S]\mathbf{1}[j\in S]\right] \\
&= \sum_{1\le i\le j\le n} a_{ij}\mathbb{P}(i\in S)\mathbb{P}(j\in S) \\
&= p^2\sum_{1\le i\le j\le n} a_{ij} \\
&= p^2 E(G_n)
\end{aligned}$$

Definition 6.2. (Variance of an estimate): The variance of an estimate $\hat{\theta}$ is

$$Var(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right]$$

Example. (Bernoulli):

$$Var(\hat{p}) = Var\left(\frac{Binom(n,p)}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Example. (Normal):

$$\begin{aligned}
Var\left(\hat{\mu}\right) &= Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \\
Var\left(\hat{\sigma}^2\right) &= Var(s^2) \\
&= Var\left(\frac{\sigma^2}{n-1}\chi^2_{n-1}\right) \\
&= \left(\frac{\sigma^2}{n-1}\right)^2 Var\left(\chi^2_{n-1}\right) \\
&= \frac{\sigma^4}{(n-1)^2}2(n-1) \\
&= \frac{2\sigma^4}{n-1}
\end{aligned}$$

Definition 6.3. Definition (Mean Squared Error): The mean squared error of an estimate $\hat{\theta}$ is

$$MSE(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Result 6.4.

$$MSE(\hat{\theta}) = (Bias(\hat{\theta}))^2 + Var\left(\hat{\theta}\right)$$

Proof.

Proof:

$$\begin{aligned}
MSE(\hat{\theta}) &= \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)^2+2\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)\left(\mathbb{E}[\hat{\theta}]-\theta\right)+\left(\mathbb{E}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)^2\right]+2\left(\mathbb{E}[\hat{\theta}]-\theta\right)\mathbb{E}\left[\hat{\theta}-\mathbb{E}[\hat{\theta}]\right]+\left(\mathbb{E}[\hat{\theta}]-\theta\right)^2 \\
&= \mathrm{Var}(\hat{\theta})+0+\left(\mathbb{E}[\hat{\theta}]-\theta\right)^2 \\
&= \mathrm{Var}(\hat{\theta})+\left(\mathrm{Bias}(\hat{\theta},\theta)\right)^2
\end{aligned}$$

✓

Corollary 6.5. If $\hat{\theta}$ unbiased, then $MSE(\hat{\theta}) = Var(\hat{\theta})$

7 Class 7

7.1 Consistency

Definition 7.1. (Consistency): Suppose $\hat{\theta}_n$ is an estimate of θ based on n iid samples. Then, $\hat{\theta}_n$ is said to be consistent for θ if

$$\hat{\theta}_n \xrightarrow{p} \theta$$

as $n \to \infty$

Result 7.2. MSE converging to 0 in the limit implies consistency

$$MSE(\hat{\theta}_n) \longrightarrow 0 \implies \hat{\theta}_n \xrightarrow{p} \theta$$

Proof. By Markov's Inequality, for any $\epsilon > 0$

$$P\left(\left|\hat{\theta}_n - \theta\right| > \epsilon\right) \leq \frac{\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]}{\epsilon^2}$$

✓

Remark. To prove consistency, we can

- Show MSE vanishes
- Use Continuous Mapping
- Use Slutsky's Lemma

Example. (Bernoulli):

$$Bias(\hat{p}) = 0, Var(\hat{p}) = \frac{p(1-p)}{n}$$

As $n \to \infty$

$$MSE(\hat{p}) = Var(\hat{p}) = \frac{p(1-p)}{n} \longrightarrow 0$$

Example. (Normal):

$$\hat{\mu} = \bar{X} \implies bias(\hat{\mu}) = 0, Var(\hat{\mu}) = Var(\bar{X}) = \frac{\sigma^2}{n}$$

As $n \to \infty$

$$MSE(\bar{X}) = \frac{\sigma^2}{n} \longrightarrow 0$$

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

$$bias(\hat{\sigma}_2) = 0, Var(\hat{\sigma^2}) = Var\left(\sigma^2 \frac{\chi^2_{n-1}}{n-1}\right) = \frac{2\sigma^4}{n-1}$$

As $n \to \infty$

$$MSE(\hat{\sigma}^2) = Var(\hat{\sigma}^2) \longrightarrow 0$$

7.2 Method of moments estimation

Definition 7.3. (Method of Moments): Suppose $X_1, X_2, \ldots X_n$ iid from distribution F which as k unknown parameters $(\theta_1, \theta_2, \ldots \theta_k)$.

Denote $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$.

For each $1 \leq j \leq k$, compute the j-th moment of F

$$\alpha_j(\boldsymbol{\theta}) = \mathbb{E}\left[X^j\right] = \int X_j f(x)dx$$

Define the j-th sample moment as

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Set up k equations

$$\begin{cases} \alpha_1 = & \hat{\alpha}_1 \\ \alpha_2 = & \hat{\alpha}_2 \\ & \vdots \\ \alpha_k = & \hat{\alpha}_k \end{cases}$$

The method of moments estimate $\hat{\theta}_n$ is the solution to the system of equations.

Remark. Note that

- The above system is a system of k equations in k unknows
- The above system may have $0, 1$ or multiple solutions. If solution is unique, then the solution is the method of moments estimate of $\hat{\theta}$

Example. (Bernoulli): $X_1, \ldots X_n \sim Ber(p)$

$$\alpha_1 = \alpha_1(p) = \mathbb{E}\left[X\right] = p$$
$$\hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Solving for p:

$$\hat{\alpha}_1 = \alpha_1$$
$$\implies \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example. (Normal) $X_1, \ldots X_n \sim N(\mu, \sigma^2)$
The population moments are

$$\alpha_1 = \mu$$
$$\alpha_2 = \mathbb{E}\left[X^2\right] = \sigma^2 + \mu^2$$

Equating to sample moments

$$\mu = \alpha_1 = \hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\mu^2 + \sigma^2 = \alpha_2 = \hat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Therefore,

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned}$$

The last equality follows because

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\\
=&\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{2}{n}\bar{X}\sum_{i=1}^{n}X_i+\bar{X}^2\\
=&\frac{1}{n}\sum_{i=1}^{n}X_i^2-\bar{X}^2
\end{aligned}$$

Note: $\hat{\sigma}^2_{MLE}$ is a biased estimator of σ^2.

$$\mathbb{E}\left[\hat{\sigma}^2\right]=\frac{n-1}{n}\mathbb{E}\left[s^2\right]=\frac{n-1}{n}\sigma^2$$

As $n\to\infty$

$$\begin{aligned}
Bias(\hat{\theta}^2)&=\frac{n-1}{n}\sigma^2-\sigma^2\to 0\\
Var(\hat{\theta}^2)&=Var\left(\frac{n-1}{n}s^2\right)\\
&=\left(\frac{n-1}{n}\right)^2\left(\frac{2\sigma^4}{n-1}\right)\\
&=\frac{n-1}{n^2}2\sigma^4\\
&\to 0
\end{aligned}$$

Hence $MSE(\hat{\sigma})\to 0$, and $\hat{\sigma}^2_{MLE}$ is consistent despite being biased.

Example. Ex (Pearson, 1984).
Contributions to the mathematical theory of evolution (1894). Weldon (1893) crab data: body length.
The histogram of the data did not look like the bell-shaped Gaussian curve.

Conjecture: maybe two species, each Gaussian? (Gaussian Mixture Model.)
Pearson conjectured that the data consists of two different kinds of crabs, each with its own Gaussian distribution.

Gaussian Mixture Model (GMM)
Let W be a random variable such that

$$W\sim\begin{cases}N(\mu_1,\sigma_1^2), & \text{with prob. } p,\\ N(\mu_2,\sigma_2^2), & \text{with prob. } (1-p).\end{cases}$$

Distribution Function

$$P(W\le t)=P(W\le t,Z=1)+P(W\le t,Z=2),$$

where $Z=\begin{cases}1 & \text{if group 1 is sampled,}\\ 2 & \text{if group 2 is sampled.}\end{cases}$
So,

$$P(W\le t)=p\,P(N(\mu_1,\sigma_1^2)\le t)+(1-p)P(N(\mu_2,\sigma_2^2)\le t).$$

Note: GMM is *not* adding two normals. Adding two normals gets another normal, but GMM is not a sum. It is a mixture.

Density Function

$$f(t)=\frac{d}{dt}P(W\le t)=p\,\phi\left(\frac{t-\mu_1}{\sigma_1}\right)+(1-p)\,\phi\left(\frac{t-\mu_2}{\sigma_2}\right),$$

where $\phi(t)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)$.
Estimation of Parameters of GMM

1. Compute sample moments:

$$\hat{m}_r=\frac{1}{n}\sum_{i=1}^{n}X_i^r,\quad 1\le r\le 5.$$

2. Compute population moments:
$$m_r = \mathbb{E}[X^r], \quad 1 \leq r \leq 5.$$
3. Solve for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p$ from
$$\hat{m}_r = m_r, \quad r = 1, \ldots, 5.$$

Pearson's Sixth-Moment Test

In general, the system of equations can have multiple roots. To choose a root, Pearson's approach: choose the root $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p})$ that is closest to the sixth sample moment.

That is, the feasible root with the smallest value of
$$|\hat{m}_6 - m_6| = \left| \frac{1}{n} \sum_{i=1}^{n} X_i^6 - \mathbb{E}[X^6] \right|.$$

Theorem (Kalai, Moitra, Valiant, 2010): the sixth-moment method gives consistent estimates of the parameters of GMM and can be computed efficiently.

8 Class 8

8.1 Maximum Likelihood Estimation

Definition 8.1. (Maximum Likelihood Estimation): Let $X_1, \dots X_n$ iid from F with *pdf* or *pmf* f_θ. Consider the joint *pdf* or *pmf*

$$L(\theta|x_1, x_2, \dots x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

The maximum likelihood estimator is given by

$$\hat{\theta}_n = \arg\max_{\Theta} L(\theta|x_1, x_2, \dots x_n)$$

Equivalently, define

$$\begin{aligned} l(\theta|x_1, \dots x_n) &:= \log L(\theta|x_1, \dots x_n) \\ &= \sum_{i=1}^{n} \log f_\theta(x_i) \end{aligned}$$

$$\hat{\theta}_n = \arg\max_{\Theta} l(\theta|x_1, x_2, \dots x_n)$$

Example. (Bernoulli) $X_1, \dots X_n \sim Ber(p)$.

$$\begin{aligned} f_p(x) &= P(X = x) \\ &= \begin{cases} p \text{ if } x = 1 \\ 1 - p \text{ if } x = 0 \end{cases} \\ &= p^x(1-p)^{1-x} \end{aligned}$$

The likelihood function is

$$\begin{aligned} L(p|X_1, \dots X_n) &= \prod_{i=1}^{n} f_p(X_i) \\ &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \\ &= p^{\sum_{i=1}^{n} X_i}(1-p)^{n - \sum_{i=1}^{n} X_i} \\ l(p|X_1, X_2, \dots X_n) &= \left(\sum_{i=1}^{n} X_i\right)\log p + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-p) \end{aligned}$$

Define $T = \sum_{i=1}^{n} X_i$,

$$\begin{aligned} l(p|X_1, \dots X_n) &= T \log p + (n - T)\log(1-p) \\ \frac{d}{dp} l(p|X_1, \dots X_n) &= \frac{T}{p} - \frac{n-T}{1-p} \\ &= 0 \end{aligned}$$

Hence

$$\hat{p}_{MLE} = \frac{T}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

To show that this is indeed a maximum,

$$\frac{d^2}{d^2p} l(p|X_1, \dots X_n) = -\frac{T}{p^2} - \frac{T}{(1-p)^2} < 0$$

Hence $l(p|X_1, \dots X_n)$ is concave and $\hat{p} = \frac{T}{n}$ is the unique maximum.

By LLN,

$$\hat{p} \xrightarrow{p} p$$

By CLT,

$$\sqrt{n}\,(\hat{p} - p) \xrightarrow{d} N(0, p(1-p))$$

Example. (Normal): $X_1, \ldots X_n$ iid $N(\mu, \sigma^2)$

$$\begin{aligned}
L(\mu, \sigma^2 | X_1, \ldots X_n) &= \prod_{i=1}^{n} f_{\mu,\sigma^2}(X_i) \\
&= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2 \right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right)
\end{aligned}$$

The log likelihood is

$$\begin{aligned}
l(\mu, \sigma^2 | X_1, \ldots X_n) &= \log\big(L(\mu, \sigma^2 | X_1, \ldots X_n)\big) \\
&= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \\
&= -\frac{n}{2} \log 2\pi\sigma^2 - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2
\end{aligned}$$

Define

$$g(\mu, \sigma^2) := -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

Setting first order partials to zero, we get

$$\begin{aligned}
\frac{\partial g(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0 \\
\frac{\partial g(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2} \left(\frac{1}{\sigma^2} \right) + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2 = 0
\end{aligned}$$

Hence

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$

To show that this is the global max, we can either

- Find the Hessian and show that it is negative definite, i.e. the function is concave
- Argue that the derivative is 0 at only one point, and show that for extreme values of μ and σ^2 (i.e. $\mu \to -\infty$, $\mu \to \infty$, $\sigma^2 \to 0$, $\sigma^2 \to \infty$)

$$l(\mu, \sigma^2) \to -\infty$$

Note

1. $\hat{\mu} \to \mu$ by LLN
2. $\sqrt{n}\,(\hat{\mu} - \mu) \stackrel{d}{=} N(0, \sigma^2)$
3. $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ by computing MSE.
4. $\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4)$

Proof of 4:

$$\begin{aligned}
\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) &= \sqrt{n} \left(\frac{\sigma^2}{n-1} \chi^2_{n-1} - \sigma^2 \right) \\
&= \sigma^2 \sqrt{n} \left(\frac{\chi^2_{n-1}}{n-1} - 1 \right)
\end{aligned}$$

Note that

$$\sqrt{n}\left(\frac{\chi_n^2}{n} - 1\right) \xrightarrow{d} N(0, 2)$$

Hence

$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4)$$

REVISIT: check with the professor regarding this proof on asymptotic normality of $\hat{\sigma}^2_{MLE}$ without using Fisher Information.

9 Class 9

9.1 Fisher Information

Definition 9.1. (Score Function): The score function is defined as

$$s(\theta) = \frac{d}{d\theta} \log f_\theta(X)$$

Definition 9.2. Definition (Fisher Information): The Fisher Information is defined as

$$I(\theta) = \mathbb{E}\left[\left(s(\theta)^2\right)\right] = \mathbb{E}\left[\left(\frac{d}{d\theta} \log f_\theta(X)\right)^2\right]$$

Lemma 9.3. Lemma (Properties of the score function):

1. Zero expectation
$$\mathbb{E}\left[s(\theta)\right] = \mathbb{E}\left[\frac{d}{d\theta} \log f_\theta(X)\right] = 0$$
2. The variance is Fisher Information
$$Var\left(\frac{d}{d\theta} \log f_\theta(x)\right) = I(\theta)$$
3. The expectation of the second derivative is negative Fisher information
$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2} \log f_\theta(X)\right]$$

Proof. (1):

$$\begin{aligned}
\mathbb{E}\left[\frac{d}{d\theta} \log f_\theta(X)\right] &= \int \frac{d}{d\theta} \log f_\theta(x) f_\theta(x) dx \\
&= \int \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\
&= \int \frac{d}{d\theta} f_\theta(x) dx \\
&= \frac{d}{d\theta} \int f_\theta(x) dx \\
&= \frac{d}{d\theta} 1 \\
&= 0
\end{aligned}$$

(2): Proof is immediate

$$\begin{aligned}
Var\left(s(\theta)\right) &= \mathbb{E}\left[(s(\theta))^2\right] - \mathbb{E}\left[s(\theta)\right]^2 \\
&= \mathbb{E}\left[(s(\theta))^2\right] \\
&= I(\theta)
\end{aligned}$$

✓

9.2 Asymptotic properties of the MLE estimator

Result 9.4. (Asymptotic properties of MLE): Suppose $X_1, \ldots X_n$ iid with *pdf* or *pmf* $f_{\theta_0}(x), \theta_0 \in \Omega$ where $\theta_0$is the true parameter and Ω is the parameter space.

Denote by $\hat{\theta}$ the MLE estimate based on $X_1, \ldots X_n$.

Suppose the following assumptions

- The density / *pmf* f_θ has the same support for all $\theta \in \Omega$, i.e.
$$\{x : f_\theta(x) > 0\}$$

does not depend on θ

- θ_0 is an interor point of Ω
- The log-likelihood function $l_n(\theta) = l(\theta|X_1, \dots X_n)$ is differentiable in θ
- $\hat{\theta}$ is the unique value of $\theta \in \Omega$ such that

$$l_n'(\theta) = 0$$

Then the following hold

1. $\hat{\theta}$ consistent for θ_0, i.e. as $n \to \infty$

$$\hat{\theta} \to \theta_0$$

2.

and

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

Proof. Normality implies consistency: if we know that the asymptotic distribution is normal with variance being the inverse of Fisher Information, then

$$\left(\hat{\theta} - \theta_0\right) = \frac{\sqrt{n}\left(\hat{\theta} - \theta\right)}{\sqrt{n}} \xrightarrow{p} 0 \text{ by Slutsky's}$$

Asymptotic normality:

$$l_n'\left(\hat{\theta}\right) = 0$$

By Taylor Expansion around θ_0,

$$\begin{aligned}
l_n'(\hat{\theta}) &\approx l_n'(\theta_0) + \left(\hat{\theta} - \theta_0\right) l_n''(\theta_0) \\
\implies \left(\hat{\theta} - \theta_0\right) &\approx \frac{-l_n'(\theta_0)}{l_n''(\theta_0)} \\
\implies \sqrt{n}\left(\hat{\theta} - \theta_0\right) &\approx \frac{-\frac{1}{\sqrt{n}} l_n'(\theta_0)}{\frac{1}{n} l_n''(\theta_0)}
\end{aligned}$$

Consider the denominator

$$\begin{aligned}
&\frac{1}{n} l_n''(\theta_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i)\bigg|_{\theta=\theta_0} \\
&\xrightarrow{p} \mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i)\bigg|_{\theta=\theta_0}\right] \text{ by law of large numbers} \\
&= -I(\theta_0)
\end{aligned}$$

Consider the numerator

$$\begin{aligned}
&\frac{1}{\sqrt{n}} l_n'(\theta_0) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i)\bigg|_{\theta=\theta_0} \\
&= \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i)\bigg|_{\theta=\theta_0}\right) \\
&\xrightarrow{d} N\left(0, I(\theta_0)\right) \text{ by CLT}
\end{aligned}$$

Hence

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}-\theta_0\right)\\
\approx&-\sqrt{n}\left(\frac{l_n'(\theta_0)}{l_n''(\theta_0)}\right)\\
\approx&\frac{-\frac{1}{\sqrt{n}}l_n'(\theta_0)}{\frac{1}{n}l_n''(\theta_0)}\\
\xrightarrow{d}&\frac{N\left(0, I(\theta_0)\right)}{I(\theta_0)} \text{ by Slutsky's}\\
=&N\left(0, \frac{1}{I(\theta_0)}\right)
\end{aligned}$$

✓

Example. (Bernoulli): $X_1, x_2, \ldots X_n \sim Ber(P)$. Then,

$$\begin{aligned}
f_p(x) &= p^x(1-p)^x\\
\log f_p(x) &= x\log p + (1-x)\log(1-p)\\
\frac{\partial}{\partial p}\log f_p(x) &= \frac{x}{p} - \frac{1-x}{1-p}\\
-\frac{\partial^2}{\partial p^2}\log f_p(x) &= \frac{x}{p^2} + \frac{1-x}{(1-p)^2}
\end{aligned}$$

Therefore

$$\begin{aligned}
I(p) &= -\mathbb{E}\left[\left(\frac{\partial^2}{\partial p^2}\log f_p(x)\right)\right]\\
&= \frac{p}{p^2} + \frac{1-p}{(1-p)^2}\\
&= \frac{1}{p} + \frac{1}{1-p}\\
&= \frac{1}{p(1-p)}
\end{aligned}$$

Hence

$$\begin{aligned}
\sqrt{n}\left(\hat{p}_{MLE} - p\right) \xrightarrow{p} &N\left(0, \frac{1}{I(p)}\right)\\
=&N\left(0, p(1-p)\right)
\end{aligned}$$

Example. (Normal): $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$. The *pdf* of X is

$$\begin{aligned}
f_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\\
\log f_{\mu,\sigma^2}(x) &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2
\end{aligned}$$

Fisher information for μ:

$$\begin{aligned}
\frac{\partial}{\partial\mu}\log f(x) &= \frac{1}{\sigma^2}(x-\mu)\\
-\frac{\partial^2}{\partial\mu^2}\log f(x) &= \frac{1}{\sigma^2}
\end{aligned}$$

Therefore,

$$I(\mu) = \frac{1}{\sigma^2}$$

Hence

$$\sqrt{n}(\hat{\mu}_{MLE} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Fisher information for σ^2:

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}\log f(x) &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2\\
-\frac{\partial^2}{(\partial\sigma^2)^2} &= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}(x-\mu)^2
\end{aligned}$$

Therefore,

$$I(\sigma^2) = -\mathbb{E}\left[\frac{\partial^2}{(\partial\sigma^2)^2}\right] = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^4} = \frac{1}{2\sigma^4}$$

And

$$\begin{aligned}\sqrt{n}\left(\hat{\sigma}_{MLE} - \sigma^2\right) &\xrightarrow{d} N\left(0, \frac{1}{I(\sigma^2)}\right) \\ &= N(0, 2\sigma^4)\end{aligned}$$

Remark. As a heuristic

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

9.3 Asymptotic properties of the MLE estimator for multiple parameters

Let $\{f_{\boldsymbol{\theta}}(x) : \boldsymbol{\theta} \in \Omega\}$ be a family of distributions where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots \theta_k) \in \mathbb{R}^k$ is a k-dimensional parameter.

Theorem 9.5. Suppose $X_1, \dots X_n$ iid from distribution with density / *pmf* f_θ. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ be the MLE based on $X_1, X_2, \dots X_n$.
Let the $k \times k$ Fisher Information matrix be defined as

$$\left(I(\boldsymbol{\theta})\right)_{ij} = Cov\left(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(x), \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(x)\right)$$

Under the same conditions as the univariate normal,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_k(\mathbf{0}, I(\boldsymbol{\theta})^{-1})$$

Remark. As an heuristic,

$$\hat{\boldsymbol{\theta}_{\boldsymbol{n}}} \approx N\left(\boldsymbol{\theta}, \frac{1}{n}I(\theta)^{-1}\right)$$

Proof. We WTS that

$$Cov(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(x), \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(x),) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\right]\log f_\theta(x)$$

✓

Example. (Normal):

$$-\frac{\partial^2}{\partial\partial\mu\partial\sigma^2}l_n(\mu, \sigma^2) = \frac{1}{\sigma^4}(x-\mu)$$

$$\mathbb{E}\left[-\frac{\partial^2}{\partial\partial\mu\partial\sigma^2}l_n(\mu, \sigma^2)\right] = 0$$

The Fisher Information matrix is

$$I(\mu, \sigma^2) = \begin{pmatrix}\frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4}\end{pmatrix}$$

Hence

$$I(\mu, \sigma^2)^{-1} = \begin{pmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{pmatrix}$$

and

$$\sqrt{n}\begin{pmatrix}\hat{\mu}_{MLE} - \mu \\ \hat{\sigma^2}_{MLE} - \sigma^2\end{pmatrix} = N_2\left(\mathbf{0}, \begin{pmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{pmatrix}\right)$$

Example. (Gamma distribution)

Let $X_1, \dots X_n$ iid $\Gamma(\alpha, \beta)$ with density

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha\Gamma(\alpha)}$$

The likelihood function is

$$\begin{aligned}L_n(\alpha, \beta) &= \prod_{i=1}^n f_{\alpha,\beta}(X_i) \\ &=\end{aligned}$$

The likelihood function is

$$L_n(\alpha,\beta) = \prod_{i=1}^{n} f_{\alpha,\beta}(X_i) = \frac{1}{(\beta^\alpha \Gamma(\alpha))^n} \prod_{i=1}^{n} X_i^{\alpha-1} e^{-\frac{1}{\beta}\sum_{i=1}^{n} X_i}.$$

The log-likelihood function is

$$\ell_n(\alpha,\beta) = \log L_n(\alpha,\beta) = -n\alpha\log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\log X_i - \frac{1}{\beta}\sum_{i=1}^{n} X_i.$$

Taking derivatives with respect to α and β and setting them to zero:

$$\frac{\partial \ell_n(\alpha,\beta)}{\partial\alpha} = -n\log\beta - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\log X_i = 0, \tag{1}$$

$$\frac{\partial \ell_n(\alpha,\beta)}{\partial\beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} X_i = 0. \tag{2}$$

From (2), the MLE satisfies

$$\hat\beta = \frac{\bar X}{\hat\alpha}, \qquad \text{where } \bar X = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Substituting $\hat\beta$ into (1) and dividing by n, we obtain

$$\log\hat\alpha - \frac{\Gamma'(\hat\alpha)}{\Gamma(\hat\alpha)} - \log\bar X + \frac{1}{n}\sum_{i=1}^{n}\log X_i = 0.$$

Consider the function

$$f(\alpha) = \log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}.$$

The function $f(\alpha)$ decreases from ∞ to 0 as α increases from 0 to ∞.
Moreover,

$$\log\bar X - \frac{1}{n}\sum_{i=1}^{n}\log X_i > 0 \quad \text{(by Jensen's inequality)}.$$

Hence,

$$\log\hat\alpha - \frac{\Gamma'(\hat\alpha)}{\Gamma(\hat\alpha)} = \log\bar X - \frac{1}{n}\sum_{i=1}^{n}\log X_i$$

has a unique solution $\hat\alpha$, which is the MLE of α.

10 Class 10

10.1 Cramer-Rao Lower Bound

Theorem 10.1. (Cramer-Rao Lower Bound) Consider a parametric model of distributions $\{f_\theta(x), \theta \in \Omega\}$ satisfying certain *mild regularity conditions*, and T is any unbiased estimator θ based on $X_1, X_2, \ldots X_n$ iid. Then

$$Var(T) \geq \frac{1}{nI(\theta)}$$

Proof. Denote the score function

$$s(\theta, x) = \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)}$$

Let

$$S = s(X_1, S_2, \ldots X_n) = \sum_{i=1}^{n} s(\theta, X_i)$$

For any unbiased estimator T,

$$\begin{aligned} &(Corr(S,T))^2 \leq 1 \\ \implies &Cov(S,T)^2 \leq Var(S) \cdot Var(T) \\ \implies &Var(T) \geq \frac{(Cov(S,T))^2}{Var(S)} \end{aligned}$$

Where

$$\begin{aligned} Var(S) &= nVar(s) \\ &= nVar\left(\frac{\partial}{\partial\theta} \log f_\theta(X_1)\right) \\ &= nI(\theta) \end{aligned}$$

Hence

$$Var(T) \geq \frac{(Cov(S,T))^2}{nI(\theta)}$$

To show that $Cov(S,T) = 1$, by unbiasedness of T

$$\begin{aligned} \theta &= \mathbb{E}\left[(T)\right] \\ &= \int_{\mathbb{R}^n} T(X_1, X_2, \ldots X_n) f_\theta(X_1) f_\theta(X_2) \ldots f_\theta(X_n) dX_1 dX_2 \ldots dX_n \end{aligned}$$

Taking the derivative wrt θ on both sides,

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} T(X_1, X_2, \ldots X_n) \left[\sum_{i=1}^{n} \left(\frac{\partial}{\partial\theta} f_\theta(x_i) \prod_{j\in[1,n], j\neq i} f_\theta(x_j)\right)\right] dX_1, dX_2 \ldots dX_n \\ &= \int_{\mathbb{R}^n} T(X_1 \ldots X_n) S(X_1, \ldots X_n) f_\theta(X_1) f_\theta(X_2) \ldots f_\theta(X_n) dX_1 dX_2 \ldots dX_n \\ &= \mathbb{E}_\theta\left[TS\right] \end{aligned}$$

Since the score function has zero expectation,

$$1 = \mathbb{E}\left[TS\right] - \mathbb{E}\left[T\right]\mathbb{E}\left[S\right] = Cov(S,T)$$

✓

Remark. An unbiased estimator is said to be efficient if its variance is

$$\frac{1}{nI(\theta)}$$

Remark. Since the MLE is asymptotically unbiased and the variance of the MLE attains the Cramer-Rao Lower Bound asymptotically, the MLE is said to be **asymptotically efficient**.

10.2 Simple Linear Regression

10.2.1 Estimating β_0, β_1

Definition 10.2. (Simple Linear Regression) Suppose we have data $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ related by model

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

The **least squares estimates** are obtained via

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The **fitted values** are defined as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The **residuals** are defined as

$$E_i = Y_i - \hat{Y}_i$$

The **sum of squared errors, SSE** are

$$SSE := \sum_{i=1}^{n} E_i = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Definition 10.3. Define

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = (n-1)s_{xy}$$
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s_x^2$$
$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2$$

Remark. Define

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The LS estimates $(\hat{\beta}_0, \hat{\beta}_1)$ are solved via

$$\begin{cases} \frac{\partial f}{\partial \beta_0} = 0 \\ \frac{\partial f}{\partial \beta_1} = 0 \end{cases}$$

Result 10.4. The LS line always passes through the point $(\bar{X}, \bar{Y})$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Remark. (Derivation of regression parameters)

Let

$$f(a, b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

Setting the partial derivatives to 0

$$\frac{\partial f}{\partial a} = -2\sum_{i=1}^{n}(y_i - (a + bx_i)) = 0$$
$$\frac{\partial f}{\partial b} = -2\sum_{i=1}^{n}(y_i - (a + bx_i))x_i = 0$$

This is a system of linear equations in a, b.

$$(2n)a + \left(2\sum_{i=1}^{n} x_i\right) b = 2\sum_{i=1}^{n} y_i$$
$$\left(2\sum_{i=1}^{n} x_i\right) a + \left(2\sum_{i=1}^{n} x_i^2\right) b = 2\sum_{i=1}^{n} x_i y_i$$

Solving for a, b

$$\hat{\beta}_0 = a = \frac{\left(\sum x_i^2\right)\left(\sum y_i\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$
$$\hat{\beta}_1 = b = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Also

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Remark. Note that the results thus far assumes nothign about the ϵ_i's.

10.2.2 Estimating variance

Now assume that

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

By definition

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$
$$\|\epsilon\|^2 = \sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

If β_0, β_1 known, then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\right] = \sigma^2$$

and an unbiased estimator of σ^2 is

$$\frac{\|\epsilon\|^2}{n}$$

Result 10.5. $\hat{\sigma}^2$ defined below is an unbiased estimate of σ^2

$$SSE \sim \sigma^2\chi^2_{n-2} \implies \hat{\sigma}^2 = \frac{SSE}{n-2}$$

10.2.3 Facts about Least Square Estimates

Result 10.6. The LS estimates are unbiased.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$$
$$\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0$$

Result 10.7. The LS estimates are normally distributed.

$$\hat{\beta}_0 \sim N\left(\beta_0 \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n s_{xx}}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

where

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{X})^2$$

Result 10.8. The LS estimates find the vector $\hat{\boldsymbol{v}}$in the plane spanned by the vectors $\mathbf{1}, \boldsymbol{X}$ that is the closest to $\boldsymbol{Y}$, where

$$\hat{\boldsymbol{v}} = \hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \boldsymbol{X}$$

To see this, denote $\boldsymbol{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots v_n \end{pmatrix}$ with $v_i = \beta_0 + \beta_1 x_i$. This can be written as

$$\boldsymbol{v} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}$$

where

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

Therefore, $\boldsymbol{v}$ is a linear combination of $\mathbf{1}$ and $\boldsymbol{x}$.

The LS estimate minimizes

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = |\boldsymbol{Y} - \boldsymbol{v}|^2$$

Result 10.9. $\hat{\beta}_0, \hat{\beta}_1$ are independent of $\hat{\sigma}^2$.

Hence

$$\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{\sum X_i^2}{nS_{xx}}}} \sim t_{n-2}$$

This can be used for constructing confidence intervals and hypothesis testing for β_0 and β_1.

Result 10.10. $\hat{\beta}_0, \hat{\beta}_1$ coincides with the MLE estimates.

Note that $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. The likelihood function is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta x_i)^2\right\}$$

The log likelihood function is

$$-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta x_i)^2 - \frac{n}{2} \log(2\pi\sigma^2)$$

The MLE is obtained by

$$\arg \min_{\beta_0, \beta_1, \sigma^2} l\left(\beta_0, \beta_1, \sigma^2\right)$$

11 Class 11

11.1 Least Absolute Deviation Line

Definition 11.1. (Least Absolute Deviation) Suppose we have data $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$ related by model

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

The **least absolute deviation line** is obtained by minimizing

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 X_i|$$

Result 11.2.

$$\arg\min_{m} \sum_{i=1}^{n} |X_i - m| = median\{X_1, \dots X_n\}$$

Result 11.3. The LAD line passes through a pair of points $(X_i, Y_i), (X_j, Y_j)$.

Proof. For simplicity, assume n odd.

Begin by fixing β_1 and defining

$$Z_i = Y_i - \beta_1 X_i$$

Then

$$\begin{aligned}
\arg\min_{\beta_0} \sum_{i=1}^{n} |Z_i - \beta_0| &= \arg\min_{\beta_0} \sum_{i=1}^{n} |Z_i - \beta_0| \\
&= median\{Z_1, \dots, Z_n\} \\
&= Z_{i_0} \text{ for some } i_0 \\
&= Y_{i_0} - \beta_1 X_{i_0}
\end{aligned}$$

This implies the LAD line for fixed β_1 passes through some point (X_{i_0}, Y_{i_0}).

by shifting the coordinate system, we can assume that the LAD line passes through the origin

$$Y - Y_{i_0} = \beta_1 (X - X_{i_0})$$

We solve for β_1 with

$$\begin{aligned}
&\min_{\beta_1} \sum_{i=1}^{n} |(Y_i - Y_{i_0}) - \beta_1 (X_i - X_{i_0})| \\
=&\min_{\beta_1} \sum_{i=1}^{n} |Y_{\tilde{i}} - \beta_1 X_{\tilde{i}}|
\end{aligned}$$

For simplicity, we just write Y_i, X_i

$$\begin{aligned}
&\min_{\beta_1} \sum_{i=1}^{n} |Y_i - \beta_1 X_i| \\
=&\min_{\beta_1} \sum_{i=1}^{n} |X_i| \left| \frac{Y_i}{X_i} - \beta_1 \right|
\end{aligned}$$

This is a weighted median problem, the sum of absolute deviations is piecewise linear between each data point and convex. Hence, the minimum is attained at some i_*, i.e.

$$\hat{\beta}_1 = \frac{Y_{i_*}}{X_{i_*}}$$

The LAD line is

$$Y = Y_{i_0} + \frac{Y_{i_*} - Y_{i_0}}{X_{i_*} - X_{i_0}} (X - X_{i_0})$$

✓

Remark. The LAD line can be computed by checking over all the $\binom{n}{2}$ pairwise lines determined by the data points. These lines are called the **elemental lines**.

Remarks. Note that

1. The LAD line is one of the elemental lines
2. The LAD estimates are the MLEs where the errors have the Laplace / Double Exponential distribution

$$Ae^{\frac{-|x|}{B}}$$

3. The slope of the LS line is a weighted average of the slopes of the elemental lines.

12 Class 12

12.1 Confidence Intervals

Definition 12.1. For an unknown parameter θ and a sample $X_1, X_2, \dots X_n$, A $100(1-\alpha)\%$ confidence interval for θ is a random interval

$$[L(X_1, X_2, \dots X_n), U(X_1, X_2, \dots X_n)]$$

such that

$$P(L \leq \theta \leq U) = 1 - \alpha$$

Remark. Interpretation of CI: If the experiment is repeated, $100(1-\alpha)\%$ of intervals will contain true θ.

12.1.1 Exact Confidence Intervals

Definition 12.2. (t-distribution). If $U \sim N(0,1)$ and $V \sim \chi^2_d$ and U, V independent, then

$$\frac{U}{\sqrt{\frac{V}{d}}} \sim t_d$$

i.e. the t-distribution with d degrees of freedom.

Example. (CI for μ, σ^2, normal data) Suppose $X_1, X-2, \dots X_n \sim$ iid $N(\mu, \sigma^2)$ where μ, σ^2 unknown.

Confidence interval for μ:

1. Estimate μ with $\hat{\mu} = \bar{X}$
2. Find distribution of estimate

$$\begin{aligned}
&\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \\
\implies &Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \\
\implies &P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right) = 1 - \alpha \\
\implies &P\left(\left|\bar{X} - \mu\right| \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha \\
\implies &P\left(-\frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha \\
\implies &P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha
\end{aligned}$$

3. Estimate σ^2:

$$s^2 = \frac{1}{n-1} \sum_{i-1}^{n} (X_i - \bar{X})^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}$$

   Then, define

$$\begin{aligned}
\hat{Z} &:= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\
&= \frac{\sqrt{n}\left(\bar{X} - \mu\right)/\sigma}{\sqrt{s^2/\sigma^2}} \\
&= \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \\
&\sim t_{n-1}
\end{aligned}$$

Hence

$$
\begin{aligned}
&P(\left|\hat{Z}\right| \leq t_{n-1,\alpha/2}) = 1-\alpha \\
\implies &P\left(-t_{n-1,\alpha/2} \leq \left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}\right) = 1-\alpha \\
\implies &P\left(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right)
\end{aligned}
$$

Confidence interval for σ^2

1. Estimate σ^2 with
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
2. Find the distribution of the estimate
$$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1} \implies \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$$
3. Since the distribution is free of unknown parameters, we can immediately find a confidence interval. Define $\chi_{n-1,\alpha}$ be the $(1-\alpha)$-th percentile of the χ^2_{n-1} distribution.
$$
\begin{aligned}
&P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}\right) = 1-\alpha \\
\implies &P\left(\frac{(n-1)s^2}{\chi_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}}\right) = 1-\alpha
\end{aligned}
$$

12.1.2 Asymptotic Confidence Intervals

Example. (Confidence interval for μ, arbitrary distribution): Suppose $X_1, X_2, \ldots X_n \sim iid\ F$ with $\mathbb{E}[X_1] = \mu, Var(X_1) = \sigma^2$.

Confidence interval for μ:

1. By CLT
$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1)$$
2. By LLN,
$$s^2 \xrightarrow{p} \sigma^2 \implies s \xrightarrow{p} \sigma$$
3. By Slutsky's
$$\frac{\sqrt{n}(\bar{X}-\mu)}{s} \xrightarrow{d} N(0,1)$$
4. Find the CI
$$
\begin{aligned}
&P\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}},\right]\right) \xrightarrow{n\to\infty} 1-\alpha \\
\implies &\left[\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right] \text{ is an asymptotically } 100(1-\alpha)\ \% \text{ confidence interval for } \mu
\end{aligned}
$$

Remark. CLT of estimate and consistent estimation of variance gives asymptotic confidence interals.

Example. (Poisson distribution) Let $X_1, \ldots X_n \sim Pois(\lambda)$. To find CI for λ,

1. Estimate λ by $\hat{\lambda} = \bar{X}$
2. By CLT
$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) \xrightarrow{d} N(0,\lambda)$$
Hence
$$Z = \frac{\sqrt{n}\left(\hat{\lambda}-\lambda\right)}{\sqrt{\lambda}} \xrightarrow{d} N(0,1)$$
3. Hence, by Slutsky's
$$\hat{Z} = \frac{\sqrt{n}\left(\hat{\lambda}-\lambda\right)}{\sqrt{\hat{\lambda}}} \xrightarrow{d} N(0,1)$$

Hence,

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\hat{\lambda}}} \leq z_{\alpha/2} \overset{n\to\infty}{\longrightarrow} 1-\lambda\right)$$

Definition 12.3. (Wald Interval) Suppse $\{f_\theta(x) : \theta \in \Omega\}$ is a parametric family of distributions, and $X_1, \dots X_n$ iid from f_θ, the **Wald Interval** for θ is the asymptotically $100(1-\alpha)\%$ CI for θ given by

$$\left[\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}, \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}},\right]$$

where $\hat{\theta}$ is the MLE estimate.

Proof. Since

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

Estimate $I(\theta)$ by $I(\hat{\theta})$. If $I(\theta)$ continuous in θ< by continuous mapping,

$$I\left(\hat{\theta}\right) \xrightarrow{p} I(\theta)$$

By Slutsky's

$$\sqrt{nI(\hat{\theta})}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N(0,1)$$

Hence

$$P\left(-z_{\alpha/2} \leq \sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta) \leq z_{\alpha/2}\right) \overset{n\to\infty}{\longrightarrow} 1-\alpha$$

✓

13 Class 13

13.1 Proof of Weak of Law of Large Numbers without finite variance

Theorem 13.1. (Weak law of large numbers)

Proof. ✓

14 Class 14

14.1 Hypothesis testing

Remark. Suppose $X_1, \ldots, X_n$ iid from $f_\theta(x)$, where $\theta \in \Omega$. We want to test the following hypotheses

1. Simple vs simple
$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta = \theta_1$$
2. Simple vs composite
$$\begin{aligned} H_0 : \theta = \theta_0 \text{ vs } H_1 &: \theta \neq \theta_1 \text{ OR} \\ H_1 &: \theta < \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$
3. Composite vs composite
$$\begin{aligned} H_0 : \theta < \theta_0 \text{ vs } H_1 : \theta > \theta_1 \text{ OR} \\ H_0 : \theta > \theta_0 \text{ vs } H_1 : \theta < \theta_1 \text{ OR} \end{aligned}$$

Definition 14.1. (Test function) A test function ϕ is a function $\phi : (X_1, X_2, \ldots, X_n) \to \{0, 1\}$ such that

$$\begin{aligned} \phi = 0 &\implies H_0 \text{ is accepted} \\ \phi = 1 &\implies H_0 \text{ is rejected} \end{aligned}$$

Definition 14.2. (Type I, Type II error) The two types of error in testing problems are

- Type I error: rejecting H_0 when H_0 true (false positive)
- Type II error: not rejecting H_0 when H_1 true (false negative)

Definition 14.3. (Level of a test): The significance level of a test is the probability of Type I error

$$\alpha = P_{H_0}(\text{ reject } H_0) = P_{H_0}(\phi = 1) = \text{ Type I error rate} = \text{ False Positive Rate}$$

Definition 14.4. (p-value): The p-value is the smallest significance level at which the test rejects H_0.

Definition 14.5. (Power): Let β be the probability of Type II error

$$\beta = P_{H_1}(\text{ accept } H_0) = P_{H_1}(\phi = 0) = \text{ Type II error rate} = \text{ False Negatve Rate}$$

The power of a test is
$$1 - \beta = P_{H_1}(\text{ reject } H_0) = P_{H_1}(\phi = 1)$$

14.1.1 Likelihood Ratio Test and Neyman-Pearson Lemma

Definition 14.6. (MP test): The **most powerful (MP) test** is a test which maximizes power $(1 - \beta)$ at a given level (α), i.e.

$$\begin{aligned} &\max_\phi \mathbb{E}_{H_1}[\phi] \\ &\text{subject to } \mathbb{E}_{H_0}[\phi] \leq \alpha \end{aligned}$$

Definition 14.7. (Likelihood ratio test): Reject H_0 for small values of the likelihood ratio

$$L(X_1, \ldots X_n) = \frac{f_{\theta_0(X_1, \ldots X_n)}}{f_{\theta_1}(X_1, \ldots X_n)}$$

Remark. "The points which give the strongest evidence in favor of H_1 over H_0". check what this means

Theorem 14.8. (Neyman-Pearson Lemma) Under a simple vs simple hypothesis test, for a constant $c > 0$, define the test function

$$\phi_0\left(X_1, X_2, \ldots X_n\right) = L(X_1, \ldots X_n) < c$$

where c is chosen such that

$$P(\text{Type I error}) = P_{H_0}(L(X_1, \ldots X_n)) = \alpha$$

Then ϕ_0 is a MP test for the testing problem

Proof. Let ϕ be any other test with significance level at most α. WTS that

$$\mathbb{E}_{H_1}\left[\phi_0\right] \geq \mathbb{E}_{H_1}\left[\phi\right]$$

Let

$$g(\boldsymbol{x}) = \left(\phi_0(\boldsymbol{x}) - \phi(\boldsymbol{x})\right)\left(c f_{\theta_1}(\boldsymbol{x}) - f_{\theta_0}(\boldsymbol{x})\right)$$

Note that

$$\begin{aligned} \phi_0(\boldsymbol{x}) > \phi(\boldsymbol{x}) &\implies f_{\theta_0}(\boldsymbol{x}) < c f_{\theta_1}(\boldsymbol{x}) \\ \phi_0(\boldsymbol{x}) < \phi(\boldsymbol{x}) &\implies f_{\theta_0}(\boldsymbol{x}) \geq c f_{\theta_1}(\boldsymbol{x}) \end{aligned}$$

This means

$$\begin{aligned} & g(\boldsymbol{x}) \geq 0 \\ \implies & \int g(\boldsymbol{x}) d\boldsymbol{x} \geq 0 \\ \implies & \int \left(\phi_0(\boldsymbol{x}) - \phi(\boldsymbol{x})\right)\left(c f_{\theta_1}(\boldsymbol{x}) - f_{\theta_0}(\boldsymbol{x})\right) d\boldsymbol{x} \geq 0 \\ \implies & c \int \left(\phi_0(\boldsymbol{x}) - \phi(\boldsymbol{x})\right) f_{\theta_1}(\boldsymbol{x}) d\boldsymbol{x} \geq \int \left(\phi_0(\boldsymbol{x}) - \phi(\boldsymbol{x})\right) f_{\theta_0}(\boldsymbol{x}) d\boldsymbol{x} \\ \implies & c\left(\mathbb{E}_{H_1}\left[\phi_0\right] - \mathbb{E}_{H_1}\left[\phi\right]\right) \geq 0 \end{aligned}$$

Since

$$\begin{aligned} \int \phi_0(\boldsymbol{x}) f_{\theta_0}(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{H_0}\left[\phi_0\right] = \alpha \\ \int \phi(\boldsymbol{x}) f_{\theta_0}(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{H_0}\left[\phi\right] \leq \alpha \end{aligned}$$

Hence,

$$\mathbb{E}_{H_1}\left[\phi_0\right] \geq \mathbb{E}_{H_1}\left[\phi\right]$$

✓

Example. (Testing normal means) $X_1, \ldots X_n \sim N(\mu, \sigma^2)$, σ^2 known. Want to test

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu = \mu_1 \neq \mu_0$$

The likelihood ratio is

$$\begin{aligned} L\left(X_1, \ldots X_n\right) &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_1)^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \mu_1)^2\right)\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(2\left(\mu_1 - \mu_0\right)\sum_{i=1}^{n} X_i + \mu_0^2 - \mu_1^2\right)\right\} \\ &= \exp\left\{-\frac{\left(\mu_1 - \mu_0\right)}{\left(\sigma^2\right)}\sum_{i=1}^{n} X_i\right\} \exp\left\{-\frac{1}{2\sigma^2}\left(\mu_0^2 - \mu_1^2\right)\right\} \end{aligned}$$

Case 1: $\mu_1 > \mu_0$, then we reject for large values of $\sum_{i=1}^{n} X_i$.

Choose c such that

$$P_{H_0}\left(\sum_{i=1}^{n} X_i > c\right) = \alpha$$

Since under H_0

$$\sum_{i=1}^{n} X_i \sim N(n\mu_0, n\sigma^2)$$

therefore

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu_0}{\sqrt{n}\sigma} \sim N(0,1)$$

Therefore we need to choose c such that

$$P_{H_0}\left(\frac{\sum_{i=1}^{n} X_i - n\mu_0}{\sqrt{n}\sigma} > \frac{c - n\mu_0}{\sqrt{n}\sigma}\right) = \alpha$$

$$\implies P_{H_0}\left(N(0,1) > \frac{c - n\mu_0}{\sqrt{n}\sigma}\right) = \alpha$$

$$\implies \frac{c - n\mu_0}{\sqrt{n}\sigma} = z_\alpha \implies c = n\mu_0 + z_\alpha \sigma \sqrt{n}$$

Hence the MP test for when $\mu_1 > \mu_0$ rejects H_0 when

$$\sum_{i=1}^{n} X_i > n\mu_0 + z_\alpha \sigma \sqrt{n} \Leftrightarrow \frac{\sqrt{n}\bar{X} - \mu_0}{\sigma} > z_\alpha$$

Case 2: when $\mu_1 < \mu_0$, we reject for small values of $\bar{X}$. Choose c such that

$$P_{H_0}(\bar{X} < c) = \alpha$$

Therefore

$$P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < \frac{\sqrt{n}(c - \mu_0)}{\sigma}\right) = \alpha$$

Hence

$$\frac{\sqrt{n}(c - \mu_0)}{\sigma} = -z_\alpha$$

Hence the MP teste for when $\mu_1 < \mu_0$ rejects H_0 when

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < -z_\alpha$$

14.1.2 One sided composite hypothesis and UMP tests

Definition 14.9. (UMP test) Conider testing the following one-sided composite hypothesis

$$H_0 : \mu = \mu_) \text{ vs } H_1 : \mu > \mu_0$$

The **uniformly most powerful test** is a test such that the power function is uniformly maximized for $\mu \neq \mu_0$ over all level α tests.
i.e. The UMP test is ϕ_0 such that

$$\mathbb{E}_{H_0}[\phi_0] = \alpha$$

and

$$\mathbb{E}_\mu[\phi_0] \geq \mathbb{E}_\mu[\phi]$$

for all $\mu \neq \mu_0$ and any ϕ such that $\mathbb{E}_{H_0}[\phi] = \alpha$

Remark. There exists no UMP test for two sided hypothesis because the UMP tests for 1 sided hypotheses have disjoint rejection regions.

15 Class 15

15.1 Generalized Likelihood Ratio Test

Definition 15.1. (GLRT, simple null) Let $\{f_\theta(x) : \theta \in Omega\}$ be a parametric model and let $\theta_0 \in \Omega$ be a particular parameter value. Consider the testing problem

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

Then, the GLRT rejects H_0 for small values of

$$\wedge = \frac{L(\theta_0|X_1, \ldots X_n)}{\max_{\theta\in\Omega} L(\theta|X_1, \ldots X_n)}$$

Where

$$L(\theta|X_1, \ldots X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

is the likelihood ratio function.

Example. (Testing normal means for two sided hypothesis)

Let $X_1, \ldots X_n \sim N(\mu, \sigma^2), \sigma^2$ known, with test

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

To compute the GLRT, note that

$$\max_{\mu} L(\mu|X_1, \ldots X_n) = L(\hat{\mu}|X_1, \ldots X_n)$$

where $\hat{\mu} = \bar{X}$ is the MLE for μ.

Then

$$\begin{aligned}\wedge &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2} n \left(\bar{X} - \mu_0\right)^2\right\}\end{aligned}$$

We reject for small $\wedge$, which is equivalent to large values of

$$-2\log\wedge = \frac{n\left(\bar{X} - \mu_0\right)^2}{\sigma^2}$$

Under H_0,

$$n\frac{(\bar{X} - \mu_0)^2}{\sigma^2} \sim \chi_1^2$$

Hence we reject H_0 when

$$\frac{\left(\bar{X} - \mu_0\right)^2}{\sigma^2} > \chi^2_{1,\alpha} \Leftrightarrow \left|\frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{\sigma}\right| \leq z_{\alpha/2}$$

Example. (MVG) $X_1, \ldots X_n \sim N_p(\boldsymbol{\mu}, \Sigma)$ where Σ known, and test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} \text{ vs } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$$

GLRT rejects when

$$-2\log\wedge = n\left(\bar{X} - \boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\bar{X} - \mu\right) > \chi^2_{p,\alpha}$$

Example. (One sided hypothesis)
Recall $X_1, \ldots X_n \sim N(\mu, \sigma^2)$ where σ^2 known,

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

Reject H_0 when

$$\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} > z_\alpha$$

We claim that this is the GLRT.

To compute GLRT, we need to first maximize

$$\max_{\mu>\mu_0} L(\mu|X_1,\dots X_n) = \max_{\mu>\mu_0}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}{2\sigma^2}\right\}$$

This equivalently minimizes

$$\min_{\mu>\mu_0}\sum(X_i-\mu)^2 = \min_{\mu>\mu>0}\left(\sum_{i=1}^n X_i^2 - 2n\bar{X} + n\mu_0^2\right)$$

Case 1: $\bar{X} > \mu_0$:

$$\min_{\mu>\mu_0}\sum_{i=1}^n(X_i-\mu)^2 = \sum_{i=1}^n(X_i-\bar{X})^2$$

Case 2: $\bar{X} \leq \mu_0$:

$$\min_{\mu>\mu_0}\sum_{i=1}^n(X_i-\mu)^2 = \sum_{i=1}^n(X_i-\mu_0)^2$$

Hence

$$\begin{aligned}\wedge &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\bar{X})^2\right\}\mathbf{1}\left[\bar{X}>\mu_0\right] + \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\bar{\mu_0})^2\right\}\mathbf{1}\left[\bar{X}\leq\mu_0\right]} \\ &= \begin{cases}\exp\left\{-\frac{n}{2\sigma^2}\left(\bar{X}-\mu_0\right)^2\right\} & \text{if } \bar{X}>\mu_0 \\ 1 \text{ otherwise}\end{cases}\end{aligned}$$

Reject H_0 for small values of $\wedge$, i.e. large values of $\bar{X}-\mu_0$. Hence the level α GLRT rejects H_0 when

$$\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} > z_\alpha$$

15.2 Asymptotic distribution of the GLRT

In general, the exact distribution of $-2\log\wedge$ under H_0 may not have a simple form, but it can be approximated by chi-squared when $n\to\infty$.

Theorem 15.2. Suppose $f_{\boldsymbol{\theta}(x):\boldsymbol{\theta}\in\mathbb{R}^k}$ is a parametric model indexed by k-dim parameter vector.

Let $X_1,\dots X_n$ iid $f(x|\theta)$ and test

$$H_0:\theta=\theta_0 \text{ vs } H_1:\theta\neq\theta_0$$

Under regularity conditions

$$-2\log\wedge \xrightarrow{d} \chi_k^2$$

Proof. Proof for case $k=1$ ✓

15.3 GLRT for submodels

Suppose $\Omega_0\subset\Omega$ is a lower dimension subspace of parameter space $\Omega=\mathbb{R}^k$. We want to test

$$H_0:\theta\in\Omega_0 \text{ vs } H_1:\theta\notin\Omega_0$$

The GLRT is

$$\wedge = \frac{\max_{\theta\in\Omega_0} L(\theta|X_1,\dots X_n)}{\max_{\theta\in\Omega} L(\theta|X_1,\dots X_n)}$$

Theorem 15.3. Let $\{f_\theta:\theta\in\Omega\}$ be a parametric model, and let $X_1,\dots,X_n$ be iid $f_{\theta_0}(x)$. Suppose θ_0 is an interior point of both Ω_0 and Ω, then under regulartiy conditions,

$$-2\log\wedge \xrightarrow{d} \chi^2_d$$

Where

$$d = \dim\Omega - \dim\Omega_0$$

Example. (Testing normal means, unknown variance)

Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, where σ^2 unknown. Test

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

Then

$$\wedge = \frac{\max_{\sigma^2}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}\right)}{\max_{\mu,\sigma^2}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}\right)}$$

The denominator is maximized at

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

The numerator is maximized by

$$\hat{\sigma}^2_{H_0} = \sum_{i=1}^{n}(X_i - \mu_0)^2$$

Therefore

$$\begin{aligned}
\wedge &= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_{H_0}}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\mu_0)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2}\right)^{\frac{n}{2}}
\end{aligned}$$

Since

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$\wedge$ small is equivalent to large values of

$$\left|\frac{\sqrt{n}(\bar{X}-\mu_0)}{S}\right|$$

Hence $GLRT$ rejects H_0 when

$$-2\log\wedge = n\log\left(1 + \frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right) \approx \frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Since $\log(1+x) \approx x$.

Since

$$n(\bar{X}-\mu_0)^2 \xrightarrow{d} \sigma^2\chi^2_1$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \xrightarrow{p} \sigma^2$$

Therefore

$$-2\log\wedge \xrightarrow{d} \chi^2_1$$

16 Lecture 16

16.1 GLRT for submodels, cont'd

Example. (Hardy-Weinberg Equilibrium) An individuals has genotype AA, Aa, aa. Draw n random observations and denote

$$N_{AA}, N_{Aa}, N_{aa}$$

such that

$$N_{AA} + N_{Aa} + N_{aa} = n$$

We model this with multinomial distribution with parameters p_{AA}, p_{Aa}, p_{aa} where

$$p_{AA} + p_{Aa} + p_{aa} = 1$$

Then

$$(N_{AA}, N_{Aa}, N_{aa}) \sim Multi(n, p_{AA}, p_{Aa}, p_{aa})$$

This has PMF

$$P(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n! p_1^{n_1} p_2^{n_2} p_3^{n_3}}{n_1! n_2! n_3!}$$

We want to test

$$H_0 : p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$$

vs

$$H_1 : H_0 \text{ not true}$$

Here, the reduced model is

$$\Omega_0 = \{\theta \in (0,1), P_1 = \theta^2, 2\theta(1-\theta), p_3 = (1-\theta)^2\}$$

and

$$\Omega = \{(p_1, p_2, p_3) : p_1 + p_2 + p_3 = 1\}$$

Note

$$\dim \Omega_0 = 1, \dim \Omega = 2$$

Under full model (H_1), denote

$$\hat{p}_1, \hat{p}_2, \hat{p}_3$$

Under submodel, (H_0), denote

$$\hat{p}_{1,H_0}, \hat{p}_{2,H_0}, \hat{p_{3,H_0}}$$

1. Solve MLE for full model

$$\hat{p}_1 = \frac{n_1}{n}, \hat{p}_2 = \frac{n_2}{n}, \hat{p}_3 = \frac{n_3}{n}$$

2. Solve MLE for submodel

$$\begin{aligned} l(\theta) &\propto \left(\theta^2\right)^{n_1} \left(2\theta(1-\theta)\right)^{n_2} (1-\theta)^{2n_3} \\ &\propto \theta^{2n_1+n_2}(1-\theta)^{n_2+2n_3} \end{aligned}$$

Setting $\frac{\partial}{\partial\theta} \log l(\theta) = 0$,

$$\begin{aligned} \hat{\theta} &= \frac{2n_1 + n_2}{2n_1 + 2n_2 + 2n_3} \\ &= \frac{2n_1 + n_2}{2n} \end{aligned}$$

Then,

$$\wedge = \left(\frac{\hat{p}_{1,H_0}}{\hat{p}_1}\right)^{n_1} \left(\frac{\hat{p}_{2,H_0}}{\hat{p}_2}\right)^{n_2} \left(\frac{\hat{p}_{3,H_0}}{\hat{p}_3}\right)^{n_3}$$

and under H_0

$$-2\log\wedge = 2n_1 \log\left(\frac{\hat{p}_1}{\hat{p}_{1,H_0}}\right) + 2n_2 \log\left(\frac{\hat{p}_2}{\hat{p}_{2,H_0}}\right) + 2n_3 \log\left(\frac{\hat{p}_3}{\hat{p}_{3,H_0}}\right) \xrightarrow{d} \chi_1^2$$

Example. (Two sample t-test, equal variances) Suppose $X_1, \dots X_m \sim N(\mu_1, \sigma^2), Y_1, \dots Y_n \sim N(\mu_2, \sigma^2)$, σ^2 unknown. Test

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

The likelihood function is

$$L(\mu_1, \mu_2, \sigma^2 | X_1, \dots X_m, Y_1, \dots Y_n) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^m \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(X_i - \mu_1)^2\right\} \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(Y_i - \mu_2)^2\right\}$$

The GLRT is based on

$$\wedge = \frac{\max_{\mu,\sigma^2} L(\mu, \mu, \sigma^2 | \boldsymbol{X}, \boldsymbol{Y})}{\max_{\mu_1,\mu_2,\sigma^2} L(\mu_1, \mu_2, \sigma^2 | \boldsymbol{X}, \boldsymbol{Y})}$$

Rejecting H_0 for small values of $\wedge$ is equivalent to rejecting for large values of

$$\left|\frac{\bar{X} - \bar{Y}}{s_P\sqrt{\frac{1}{m} + \frac{1}{n}}}\right|$$

Where s_p^2 is the pooled sample variance

$$s_p^2 = \frac{1}{n+m-2}\{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(X_j - \bar{Y})^2\}$$

Note that

$$\sum_{i=1}^{m}(X_i - \bar{X})^2 \sim \sigma\chi^2_{m-1}$$

$$\sum_{j=1}^{n}(Y_i - \bar{Y})^2 \sim \sigma\chi^2_{n-1}$$

And the following are independent.

$$\bar{X}, \bar{Y}, \sum_{i=1}^{m}(X_i - \bar{X})^2, \sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

Under H_0,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} \sim N(0, 1)$$

and

$$\frac{s_p^2}{\sigma^2} \sim \frac{\chi^2_{m+n-2}}{m+n-2}$$

Hence

$$T = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{n+m-2}$$

The GLRT level α test rejects when

$$|T| \geq t_{m+n-2,\alpha/2}$$

Example. (Two sample t-test, unequal variances)

17 Class 17

18 Class 18

19 Class 19

19.1 Comparison of confidence intervals

Remark. (Normal data): $X_1, \dots X_n \sim N(\mu, \sigma^2)$

Parameter	σ^2 known	σ^2 unknown	Type
μ	$\left[\bar{X} \pm z_{1-\alpha/2} \dfrac{\sigma}{\sqrt{n}}\right]$	$\left[\bar{X} \pm t_{n-1,\, 1-\alpha/2} \dfrac{S}{\sqrt{n}}\right]$	Exact
σ^2	NA	$\left[\dfrac{(n-1)S^2}{\chi^2_{n-1,\, 1-\alpha/2}}, \dfrac{(n-1)S^2}{\chi^2_{n-1,\, \alpha/2}}\right]$	Exact

Remark. (Arbitrary data): $X_1, \dots X_n \sim F$ with $\mathbb{E}[X_1] = \mu, Var(X) = \sigma^2$ unknown

interval	$\left[\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right]$	$\left[\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}\right]$
type of interval	Exact	Asymptotic
requires noramlity of data?	Yes	No

19.2 Confidence set

Definition 19.1. (Confidence set): A confidence set is a set A which is a function of $X_1, \dots X_n$, such that

$$P(\boldsymbol{\mu} \in A(X_1, \dots X_n)) \geq 1 - \alpha$$

Result 19.2. The multivariate sample mean is multivariate normal

$$\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \bar{X}_i \sim N_p\left(\boldsymbol{\mu}, \frac{\Sigma}{n}\right)$$

Example. When $p = 1$,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

When: $p \geq 2$
Define

$$\boldsymbol{Z} = \sqrt{n}\,\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p).$$

Then

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{i=1}^{p} Z_i^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Hence,

$$P\left(\boldsymbol{Z}^\top \boldsymbol{Z} \leq \chi^2_{p,\,1-\alpha}\right) = 1 - \alpha,$$

or equivalently,

$$P\left(n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \chi^2_{p,\,1-\alpha}\right) = 1 - \alpha.$$

Confidence set for $\boldsymbol{\mu}$
If Σ known

$$\{\boldsymbol{\mu} \in \mathbb{R}^p : n\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \leq \chi^2_{p,\alpha}\}$$

is a $100(1-\alpha)\%$ confidence set for μ.
For the case where $p = 2, \Sigma = I$, the confidence set is

$$\{\boldsymbol{\mu} \in \mathbb{R}^2 : \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)^T \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \leq \frac{\chi_{p,\alpha}}{n}\}$$

which describes a circle centered at $\bar{\boldsymbol{X}}$ and radius

$$\sqrt{\frac{\chi^2_{p,\alpha}}{n}}$$

For $p = 2, \sigma \neq I$

$$\{\boldsymbol{\mu} \in \mathbb{R}^2 : \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \leq \frac{\chi_{p,\alpha}}{n}\}$$

If Σ unknown, By multivariate CLT

$$\sqrt{n}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \xrightarrow{d} N_p\left(0, \Sigma\right)$$

Therefore,

$$Z = \sqrt{n}\Sigma^{-1}\left(\bar{\boldsymbol{X}} - \mu\right) \xrightarrow{d} N_p(0,1)$$

To estimate Σ, we use the sample covariance matrix

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^2$$
$$S \xrightarrow{p} \Sigma$$

By multivariate Slutsky's

$$\tilde{Z} = \sqrt{n}S^{-\frac{1}{2}}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \xrightarrow{d} N_p(\mathbf{0}, I)$$

$$\begin{aligned}\tilde{Z}^T\tilde{Z} &= n\left(\bar{X} - \mu\right)S^{-1}(\bar{X} - \mu)\\ &\xrightarrow{d} \chi_p^2\end{aligned}$$

Hence,

$$\{\boldsymbol{\mu} \in \mathbb{R}^p : n\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)S^{-1}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \leq \chi^2_{p,\alpha}\}$$

is an asymptotically $100(1-\alpha)$ % confidence set of μ.

19.3 Hypothesis tests

19.3.1 Simple vs simple

Recall that for simple vs simple hypothesis, the most powerful test is given by the Neyman-Pearson lemma, where we reject for large values of

$$\wedge(X_1, \ldots X_n) = \frac{f_{\theta_0}(X_1, \ldots X_n)}{f_{\theta_1}(X_1, \ldots X_n)}$$

19.3.2 Simple vs composite

Example. Recall that for testing normal means, for $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known, the following are UMP tests

H_0	H_1	UMP Test
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	Reject $\bar{X} > \mu_0 + \frac{Z_\alpha}{\sqrt{n}}$
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	Reject $\bar{X} < \mu_0 - \frac{Z_\alpha}{\sqrt{n}}$
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	No UMP Test

For the last case, we can use the GLRT. The GLRT says to reject if

$$\left|\sqrt{n}\frac{(\bar{X} - \mu_0)}{\sigma}\right| > z_{\alpha/2}$$

Remark. Recall that for GLRT, testing

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

$$\wedge = \frac{L(\theta_0|X_1, \ldots X_n)}{\max_{\theta\in\Omega} L(\theta|X_1, \ldots X_n)}$$

Under H_0:

$$-2\log\wedge \xrightarrow{d} \chi_k^2$$

Remark. Recall GLRT for submodels, testing

$$H_0 : \theta \in \Omega_0 \subset \Omega \text{ vs } H_1 : \theta \notin \Omega_0$$

$$\wedge = \frac{\max_{\theta\in\Omega_0} L(\theta|X_1, \ldots X_n)}{\max_{\theta\in\Omega} L(\theta|X_1, \ldots X_n)}$$

Under H_0:

$$-2\log\wedge \xrightarrow{d} \chi_d^2$$

where

$$d = \dim\Omega - \dim\Omega_0$$

19.3.3 Composite vs composite

Remark. For composite vs composite tests, the GLRT test of level α rejects when

$$\left|\frac{\sqrt{n}\left(\bar{X}-\mu_0\right)}{S}\right| > t_{n-1,\frac{\alpha}{2}}$$

19.3.4 Asymptotic power of tests

Definition 19.3. (Consistent test) A level α test with power converging to 1 is known as a **consistent** test.

Example. $X_1, \ldots X_n \sim N(\theta, 1)$, test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta = \theta_1 > 0$$

The z-test rejects H_0 when

$$\sqrt{n}\bar{X} > z_\alpha$$

Under H_1:

$$X_1, \ldots X_n \sim N(\theta_1, 1) \implies \sqrt{n}\bar{X} \sim N\left(\sqrt{n}\theta_1, 1\right)$$

The power of this test is

$$\begin{aligned}
Power(\theta_1) &= P_{H_1}\left(\sqrt{n}\bar{X} > z_\alpha\right) \\
&= P\left(\sqrt{n}\bar{X} - \sqrt{n}\theta_1 > z_\alpha - \sqrt{n}\theta_1\right) \\
&= P\left(N(0,1) > z_\alpha - \sqrt{n}\theta_1\right) \\
&= 1 - \Phi\left(z_\alpha - \sqrt{n}\theta_1\right) \\
&\overset{n\to\infty}{\longrightarrow} 1
\end{aligned}$$

If instead, we take $\theta_1 = \frac{h}{\sqrt{n}}$ where h fixed.

$$Power(h) = 1 - \Phi(z_\alpha - h)$$

This is the **asymptotic local power**.

Example. GLRT for $X_i \sim \text{Uniform}(0, \theta)$
Let $X_1, X_2, \ldots, X_n \sim \text{Uniform}(0, \theta)$ i.i.d. We want to test

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta < 1.$$

The likelihood under the uniform model is

$$L(\theta \mid X_1, \ldots, X_n) = \begin{cases} \theta^{-n}, & \text{if } X_{(n)} \le \theta, \\ 0, & \text{otherwise,} \end{cases}$$

where $X_{(n)} = \max\{X_1, \ldots, X_n\}$.
The MLE of θ is therefore

$$\hat{\theta} = X_{(n)}.$$

The generalized likelihood ratio is

$$\Lambda = \frac{L(H_0)}{\max_\theta L(\theta)} = \frac{L(\theta = 1)}{L(\theta = \hat{\theta})} = \frac{1^{-n}\mathbf{1}_{\{X_{(n)} \le 1\}}}{\hat{\theta}^{-n}\mathbf{1}_{\{X_{(n)} \le \hat{\theta}\}}} = X_{(n)}^n.$$

We reject H_0 for small values of Λ, that is,

$$\Lambda < c_1 \quad \Leftrightarrow \quad X_{(n)} < c_1^{1/n}.$$

Determining c_1 (level α test) We choose c_1 so that the test has size α:

$$P_{\theta=1}(X_{(n)} < c) = \alpha.$$

Under H_0, since $X_i \sim \text{Uniform}(0, 1)$,

$$P_{\theta=1}(X_{(n)} < c) = P(X_1 < c, \ldots, X_n < c) = c^n.$$

Hence,

$$c = \alpha^{1/n}.$$

Therefore, the GLRT of level α rejects H_0 when

$$X_{(n)} < \alpha^{1/n}$$

Power function For $\theta = \theta_0 < 1$,

$$\text{Power}(\theta_0) = P_{\theta_0}(X_{(n)} < \alpha^{1/n}) = P_{\theta_0}(X_1 < \alpha^{1/n}, \dots, X_n < \alpha^{1/n}) = \left(\frac{\alpha^{1/n}}{\theta_0}\right)^n,$$

if $\alpha^{1/n} < \theta_0$, and 1 otherwise.
That is,

$$\text{Power}(\theta_0) = \begin{cases} \left(\frac{\alpha^{1/n}}{\theta_0}\right)^n, & \alpha^{1/n} < \theta_0, \\ 1, & \alpha^{1/n} \geq \theta_0. \end{cases}$$

Local power For local alternatives $\theta_0 = 1 - \dfrac{h}{n}$, we have

$$\text{Power}(h) = \begin{cases} \dfrac{\alpha}{(1-\frac{h}{n})^n}, & \alpha^{1/n} < 1 - \frac{h}{n}, \\ 1, & \alpha^{1/n} \geq 1 - \frac{h}{n}. \end{cases}$$

As $n \to \infty$,

$$(1 - \tfrac{h}{n})^{-n} \to e^h,$$

so that

$$\text{Power}(h) = \begin{cases} \alpha e^h, & \text{if } \alpha e^h < 1, \\ 1, & \text{if } \alpha e^h \geq 1. \end{cases}$$