STAT-4320 - Class Notes

October 7, 2025

Contents

Chapter 1: Class 1

1.1 The sample mean

Definition (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

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

Note

1. 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. OR

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

2. The variance of $\bar{X}$ is given by

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

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 (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

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

Chapter 2: Chapter 2

2.1 Breakdown Point and Efficiency

Recall that sample mean is defined as

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

Issues with the sample mean: corruption in 1 or few data points can make sample mean unstable. Alternatively, the sample median is more robust.

2.1.1 Sample and population median

Definition (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 (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 (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.

Ex.

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

2.1.3 Sampling distribution of the sample median

Let $X_1, \dots X_n$ iid F and F has population median m. As $n \to \infty$

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

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

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

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

2.1.4 Efficiency

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

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

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

Chapter 3: Class 3

3.1 Convergence of random variables

There are two kinds of convergence

- convergence in probability
- convergence in distribution

Definition (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 (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 (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 (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.

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: Fix $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(|\bar{X} - \mu| > \epsilon\right) &= \frac{Var\left(\bar{X}\right)}{\epsilon^2} \quad \text{by Chebyshev's Inequality}\\ &= \frac{\sigma^2}{n\epsilon^2}\\ \frac{\sigma^2}{n\epsilon^2} &\underset{n\to\infty}{\longrightarrow} 0\end{aligned}$$

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 $Y = |X|$

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

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: We prove using Markov's Inequality

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

Central limit theorem: Suppose $X_1 \ldots X_n$ 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

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

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

Remarks: informally, we can say

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

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 $g(\bar{X})$

Delta Method: Suppose $X_1, \ldots X_n$ 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

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

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

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

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

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

The covariance matrix is denoted

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

Where the $i-j$-th element is

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

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

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

3.6 Multivariate Normal

Definition (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

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

Note:

- this definition only works with a *pd* covariance matrix.
- a multivariate normal can be defined without the *pd* matrix, using the linear combination definition.

Definition (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

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

Chapter 4: Class 4

4.1 Linear Algebra Review

4.1.1 Positive definite matrices

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

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

Note:

- All eigenvalues of a *pd* matrix are positive
- 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

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

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

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

  Proof:

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

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

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

- Sum of eigenvalues is the trace of A

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

Note: 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 $\boldsymbol{a}^T\Sigma\boldsymbol{a}$,

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

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

For a random variable X, the *mgf* is a function $\mathbb{R} \to \mathbb{R}_{\geq 0}$,

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

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

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

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

4.3.1 Property 1

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,

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

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

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

Exercise for enthusiasts: Prove 1 using *mgfs*.

4.3.2 Property 2

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 $\boldsymbol{X} \perp \boldsymbol{Y}$ if and only if $\boldsymbol{\Sigma_{11}} = \boldsymbol{\Sigma_{21}^T} = \boldsymbol{0}$.
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.
Important case: $p_1 = p_2 = 1$, Suppose

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

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

Remarks

- 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, 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

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

Chapter 5: Class 5

5.1 Chi-squared distribution

Note that χ^2_d is the sum of d iid standard normals squared.

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

The *chi-squared* distribution has expectation

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

Since

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

The *chi-squared* distribution has variance

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

Since

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

5.2 Sample Variance, cont'd

Result: the sample variance follows a *chi-squared* distribution.

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

Proof:

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

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

Note that

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

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

Therefore, $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^n$ has eigenvalues $(1, 1, \dots, 1, 0)$ (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 $\boldsymbol{Z} = N_n(\mathbf{0}, \boldsymbol{I})$. Then,

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

This is true because

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

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

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

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: 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.

Note: As a general strategy, to show independence, we can show in two steps

1. show jointly normal
2. show 0 covariance

It suffices to show $Cov(\bar{X}, X_i - \bar{X}) = 0$ for all $1 \leq i \leq n$.
Note that

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

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

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

Chapter 6: Class 6

6.1 Basic Framework of Statistical Estimation

Given iid samples $X_1, X_2, \ldots X_n$, how do we infer / estimate parameters of F?
Ex 1 (Bernoulli): Given $X_1, X_2, \ldots, X_n$ 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 $T \sim Binom(n,p)$.
Ex 2 (Normal): $X_1, X_2, \ldots X_n$ iid $N(\mu, \sigma^2)$. Estimates for parameters are

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

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

6.1.1 Properties of Estimators

Definition (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.

Ex (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

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

Ex (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

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

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

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

Ex (Bernoulli):

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

Ex (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 (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:

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

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: If $\hat{\theta}$ unbiased, then $MSE(\hat{\theta}) = Var(\hat{\theta})$

Chapter 7: Class 7

7.1 Consistency

Definition (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: 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 $\epsilon > 0$

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

Note: To prove consistency, we can

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

Ex (Bernoulli):

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

As $n \to \infty$

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

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

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

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.

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 $\hat{\theta}$

Ex (Bernoulli): $X_1, \dots X_n \sim Ber(p)$

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

Solving for p:

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

Ex (Normal) $X_1, \dots X_n \sim N(\mu, \sigma^2)$
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.
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:
$$m_r = \mathbb{E}[X^r], \quad 1 \le r \le 5.$$
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)

Pearson's method works: the sixth-moment method gives consistent estimates of the parameters of GMM and can be computed efficiently.

Chapter 8: Class 8

8.1 Maximum Likelihood Estimation

Definition (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

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

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

Ex (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 functino is

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

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

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

Hence

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

To show that this is indeed a maximum,

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

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

By LLN,

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

By CLT,

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

Ex (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. $\sqrt{n}\left(\hat{\mu} - \mu\right) \overset{d}{=} N(0, \sigma^2)$
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

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

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.

Chapter 9: Class 9

9.1 Fisher Information and asymptotic properties of the MLE estimator

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

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

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 (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

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

Definition (Asymptotic properties of MLE): Suppose $X_1, \dots X_n$ 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 $X_1, \dots X_n$.

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, as $n \to \infty$

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

and

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

Note: as a heuristic,

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

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,

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

Consider the denominator

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

Consider the numerator

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

Hence

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