Unit 3 - part 3
Fisher Information, Asymptotic Normality of MLE; Method of Moments

2. Review: Covariance Matrices and the Log-Likelihood Function

Exercises due Jul 14, 2020 20:59 JST

Bookmark this page

Let $\mathbf{X}$ be a random vector of dimension $d \times 1$ with expectation $\mu_{\mathbf{X}}$. Recall from Lecture 10 that the covariance matrix Σ is defined as the following matrix outer product:

$$\Sigma = \mathbb{E}\left[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T\right].$$

It can be shown (similar to the covariance of random variables X, Y in Lecture 10) that

$$\begin{aligned}\Sigma &= \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] - \mathbb{E}\left[\mathbf{X}\right]\mathbb{E}[\mathbf{X}]^T \\ &= \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^T.\end{aligned}$$

Review of Covariance Matrices

1/3 points (graded)

Consider the following random vector of dimension $d \times 1$: $\mathbf{X} = \left[X^{(1)}, X^{(2)}, \ldots, X^{(d)}\right]^T$ is equally likely to be one of $[1, 0, \ldots, 0]^T$, $[0, 1, \ldots, 0]^T$, ..., $[0, 0, \ldots, 1]^T$. That is, $\mathbf{X}$ is equal to any of the unit vectors along the coordinate axes with probability $\frac{1}{d}$.

Let us compute the entries of the covariance matrix $\Sigma_{ij} = \text{Cov}\left(X^{(i)}, X^{(j)}\right)$.

$\text{Cov}\left(X^{(i)}, X^{(i)}\right) =$ 1/d ✖ **Answer:** 1/d - 1/d^2

$\frac{1}{d}$

With $i \neq j$, $\text{Cov}\left(X^{(i)}, X^{(j)}\right) =$ 2/d ✖ **Answer:** -1/d^2

$\frac{2}{d}$

Is Σ a singular (i.e. not invertible) covariance matrix? **Note:** A matrix Σ is singular if $\det(\Sigma) = 0$.

- (●) Yes
- ( ) No

✔

Solution:

For any $i \in \{1, 2, \ldots, d\}$,

$$\begin{aligned} \mathsf{Cov}\left(X^{(i)}, X^{(i)}\right) &= \mathsf{Var}\left(X^{(i)}\right) \\ &= \frac{1}{d} - \frac{1}{d^2}, \end{aligned}$$

as each $X^{(i)}$ is equal to 1 with probability $\frac{1}{d}$ and equal to 0 with probability $1 - \frac{1}{d}$.

For any $i \neq j$, $\mathbb{E}\left[X^{(i)}X^{(j)}\right] = 0$ as $X^{(i)}$ and $X^{(j)}$ are never both equal to 1 at the same time. Therefore,

$$\begin{aligned} \mathsf{Cov}\left(X^{(i)}, X^{(j)}\right) &= \mathbb{E}\left[X^{(i)}X^{(j)}\right] - \mathbb{E}\left[X^{(i)}\right]\mathbb{E}\left[X^{(j)}\right] \\ &= -\frac{1}{d^2}. \end{aligned}$$

The covariance matrix looks as follows:

$$\Sigma = \begin{bmatrix} \frac{1}{d} - \frac{1}{d^2} & -\frac{1}{d^2} & \cdots & -\frac{1}{d^2} \\ -\frac{1}{d^2} & \frac{1}{d} - \frac{1}{d^2} & \cdots & -\frac{1}{d^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{d^2} & -\frac{1}{d^2} & \cdots & \frac{1}{d} - \frac{1}{d^2} \end{bmatrix}.$$

Adding all the rows and replacing row 1 with the result yields

$$\widehat{\Sigma} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -\frac{1}{d^2} & \frac{1}{d} - \frac{1}{d^2} & \cdots & -\frac{1}{d^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{d^2} & -\frac{1}{d^2} & \cdots & \frac{1}{d} - \frac{1}{d^2} \end{bmatrix}.$$

From the above, we can see that the determinant of $\widehat{\Sigma}$ is equal to 0. This means that Σ, which is row-equivalent to $\widehat{\Sigma}$, is a singular covariance matrix.

Dimensions of Gradient of Log-Likelihood Function

1/2 points (graded)

Let $\left(E, (\mathbf{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a random vector $\mathbf{X}$ of dimension $k \times 1$. Let $f_\theta(\mathbf{x})$ be the joint pdf of $\mathbf{X}$ and let $\theta \in \mathbb{R}^d$.

Let the log-likelihood function associated with one observation of $\mathbf{X}$ be denoted $\ell_1(\mathbf{x}, \theta)$. For simplicity, let $\ell_1(\mathbf{x}, \theta)$ be denoted $\ell(\theta)$, where it is assumed that $\mathbf{x}$ is fixed.

Assuming that $\ell(\theta)$ is differentiable with respect to θ for almost all $\mathbf{x}$, what are the dimensions of the gradient $\nabla\ell(\theta)$?

Number of rows in $\nabla\ell(\theta)$: k ✖ **Answer:** d + 0*k

k

Number of columns in $\nabla\ell(\theta)$: 1 ✔ **Answer:** 1 + 0*d + 0*k

1

Solution:

$\ell(\theta)$, at any given $\mathbf{x}$, is a real-valued function of d variables in the parameter $\theta \in \mathbb{R}^d$.

Therefore, the gradient vector $\nabla\ell(\theta)$ is of size $d \times 1$.

Log-Likelihood Function of a Bernoulli-like Random Variable

0/1 point (graded)

Consider the following experiment: You take a coin that lands a head (H) with probability $0 < p < 1$ and you toss it twice. Define X as the following random variable:

$$X = \begin{cases} 1 & \text{if outcome is HH} \\ 0 & \text{otherwise} \end{cases}$$

Let $\ell(p)$ be the log-likelihood function of X when written as a random function, i.e. all of the x in the function written as X. What is $\ell(p)$?

Hint: Write the pmf of X as a one-line formula.

(Enter **X** for X, and **ln(y)** for $\ln(y)$. Do not enter "log".)

$\ell(p) =$ ln(p*(1-p)) ✖ **Answer:** 2*X*ln(p) + (1-X)*ln(1-p^2)

$\ln(p \cdot (1-p))$

Solution:

First, note that X takes on the value 1 with probability p^2 and the value 0 with probability $1 - p^2$.

Finding the log-likelihood function involves writing down the pmf of X as a one-line equation:

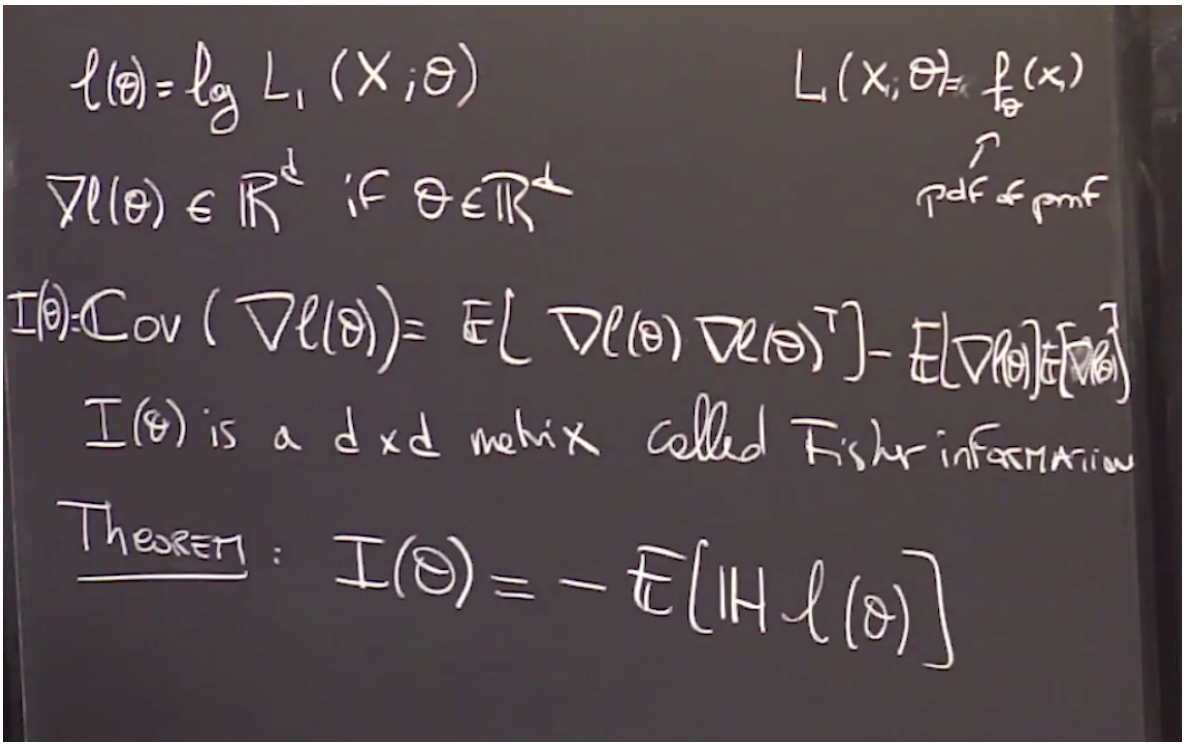
$$f(x; p) = \left(p^2\right)^x \cdot \left(1 - p^2\right)^{1-x}, \quad \text{where} \quad x \in \{0, 1\}.$$

Taking logarithm and replacing all x with X yields the desired log-likelihood function written as a random function.

Note: because X takes only two values $\{0, 1\}$, there is more than one way to write down the pmf, and consequently the log-likelihood function.

slide 36

just showing the definition here



Let $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ denote a continuous statistical model. Let $f_\theta(x)$ denote the pdf (probability density function) of the continuous distribution $\mathbf{P}_\theta$. Assume that $f_\theta(x)$ is twice-differentiable as a function of the parameter θ.

In the next few problems, you will derive the formula

$$\mathcal{I}(\theta) = \int_{-\infty}^{\infty} \frac{\left(\frac{\partial f_\theta(x)}{\partial\theta}\right)^2}{f_\theta(x)}\, dx$$

using the definition $\mathcal{I}(\theta) = \mathsf{Var}(\ell'(\theta))$ and the basic formula $\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ for any random variable X.

For computations, it is sometimes convenient to use the above formula for the Fisher information.

Note: The derivation in the next set of problems is presented as a proof in the video that follows, but we encourage you to attempt these problems before watching the video.

Deriving a Useful Formula for the Fisher Information I

2/2 points (graded)

Let $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta \in \mathbb{R}})$ denote a statistical model for a continuous distribution $\mathbf{P}_\theta$. Let f_θ denote the pdf (probability density function) of the continuous distribution $\mathbf{P}_\theta$. Recall that

$$\int_{-\infty}^{\infty} f_\theta(x)\ dx = 1$$

for all $\theta \in \mathbb{R}$.

For the next two questions, assume that you are allowed to interchange derivatives and integrals.

What is

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_\theta(x)\ dx \quad ?$$

0 ✔ **Answer:** 0.0

What is

$$\int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta^2} f_\theta(x)\ dx \quad ?$$

0 ✔ **Answer:** 0.0

Solution:

Since we know $\int_{-\infty}^{\infty} f_\theta(x)\ dx = 1$, this implies that

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f_\theta(x)\ dx = \frac{\partial}{\partial \theta} 1 = 0.$$

Since we are allowed to interchange the integral and derivative, this implies that

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_\theta(x)\ dx = 0.$$

Similarly for the second derivative,

$$\int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta^2} f_\theta(x)\ dx = \frac{\partial^2}{\partial \theta^2} \int_{-\infty}^{\infty} f_\theta(x)\ dx = \frac{\partial^2}{\partial \theta^2} 1 = 0.$$

Remark: If f is "nice enough," analytically speaking, then we can rigorously justify interchanging the integral and derivative.

Deriving a Useful Formula for the Fisher Information II

1/1 point (graded)

As before, let f_θ denote the pdf (probability density function) of the continuous distribution $\mathbf{P}_\theta$. By definition,

$$\ell(\theta) = \ln L_1(X, \theta) = \ln f_\theta(X)$$

where $X \sim \mathbf{P}_\theta$. Differentiating, we see

$$\ell'(\theta) = \frac{\partial}{\partial \theta} \ln f_\theta(X) = \frac{\frac{\partial}{\partial \theta} f_\theta(X)}{f_\theta(X)}.$$

What is

$$\mathbb{E}[\ell'(\theta)] = \mathbb{E}\left[\frac{\frac{\partial}{\partial \theta} f_\theta(X)}{f_\theta(X)}\right]?$$

0 ✔ **Answer:** 0.0

(Note that $X \sim \mathbf{P}_\theta$.)

Solution:

Observe that

$$\mathbb{E}\left[\frac{\frac{\partial}{\partial \theta} f_\theta(X)}{f_\theta(X)}\right] = \int_{-\infty}^{\infty} \left(\frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)}\right) f_\theta(x)\, dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_\theta(x)\, dx = 0,$$

by the computation in the previous question.

Deriving a Useful Formula for the Fisher Information III

0/1 point (graded)

As before, let f_θ denote the pdf (probability density function) of the continuous distribution $\mathbf{P}_\theta$. By definition,

$$\ell(\theta) = \ln L_1(X, \theta) = \ln f_\theta(X)$$

where $X \sim \mathbf{P}_\theta$.

Using the previous question, which of the following are equal to $\mathsf{Var}(\ell'(\theta)) = \mathsf{Var}\left(\frac{\partial}{\partial\theta} \ln f_\theta(X)\right)$? (Choose all that apply.)

- [x] $\mathcal{I}(\theta)$ ✔
- [ ] $\mathbb{E}[\ell'(\theta)]$
- [ ] $\mathbb{E}[(\ell'(\theta))^2]$ ✔
- [x] $\int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2}{f_\theta(x)}\, dx$ ✔

✖

Solution:

We consider the choices in order.

- By definition, $\mathcal{I}(\theta) = \mathsf{Var}(\ell'(\theta))$, so the first answer choice $\mathcal{I}(\theta)$ is correct.
- By the previous question, $\mathbb{E}[\ell'(\theta)] = 0$, so this answer choice is incorrect.
- By definition of variance,

  $$\mathsf{Var}(\ell'(\theta)) = \mathbb{E}[\ell'(\theta)^2] - \mathbb{E}[\ell'(\theta)]^2,$$

  and $\mathbb{E}[\ell'(\theta)] = 0$, by the previous question. Hence, $\mathbb{E}[(\ell'(\theta))^2] = \mathsf{Var}(\ell'(\theta))$, and so the answer choice $\mathbb{E}[(\ell'(\theta))^2]$ is correct.
- The last choice $\int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2}{f_\theta(x)}\, dx$ is correct because, using the previous bullet,

  $$\begin{aligned}\mathsf{Var}\left(\ell'(\theta)\right) &= \mathbb{E}\left[(\ell'(\theta))^2\right] \\ &= \mathbb{E}\left[\left(\frac{\frac{\partial}{\partial\theta} f_\theta(X)}{f_\theta(X)}\right)^2\right] \\ &= \int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2}{f_\theta(x)}\, dx.\end{aligned}$$

Remark: A convenient way to compute the Fisher information is to use the fourth answer choice, which gives the useful formula

$$\mathcal{I}(\theta) = \int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2}{f_\theta(x)}\, dx.$$

lecture version of above
assuming that we can take the derivative of the integral
f(x) is a pdf so integrates to 1

X is continuous with pdf $f_\theta(x) = L_1(x,\theta)$

$$\int f_\theta(x)\,dx = 1 \qquad \frac{\partial}{\partial\theta}\int f_\theta(x)\,dx = 0$$

also

$$\int \frac{\partial}{\partial\theta} L_1(x;\theta)\,dx = 0 \quad (1)$$

$$\int \frac{\partial^2}{\partial\theta^2} L_1(x;\theta)\,dx = 0 \quad (2)$$

$$l'(\theta) = \frac{\partial}{\partial\theta}\log L_1(x;\theta)$$

$$= \frac{\frac{\partial}{\partial\theta} L_1(x,\theta)}{L_1(x,\theta)}$$

$$E[\ell'(\theta)] = \int \frac{\frac{\partial}{\partial\theta} L_1(x,\theta)}{L_1(x;\theta)} L_1(x;\theta)\, dx = 0 \quad (\text{by } (1))$$

$$\Rightarrow \text{Var}[\ell'(\theta)] = E[(\ell'(\theta))^2]$$

$$= \int \left(\frac{\frac{\partial}{\partial\theta} L_1(x;\theta)}{L_1(x,\theta)}\right)^2 L_1(x,\theta)\, dx$$

$$= \int \frac{\left(\frac{\partial}{\partial\theta} L_1(x;\theta)\right)^2}{L_1(x;\theta)}\, dx.$$

$$\ell''(\theta) = \frac{\frac{\partial^2}{\partial\theta^2} L_1(x,\theta) L_1(x,\theta) - \left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{(L_1(x,\theta))^2}$$

$$-E[\ell''(\theta)] = -\int \frac{\frac{\partial^2}{\partial\theta^2} L_1(x,\theta)\cdot L_1(x,\theta) - \left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{L_1(x;\theta)^2} L_1(x;\theta)\, dx$$

$$= -\int \frac{\partial^2}{\partial\theta^2} L_1(x,\theta)\, dx + \text{Var}(\ell'(\theta))$$

$$= 0 \quad (\text{by } (2))$$

$$\Rightarrow -E[\ell''(\theta)] = \text{Var}(\ell'(\theta))$$

Definition of Fisher Information

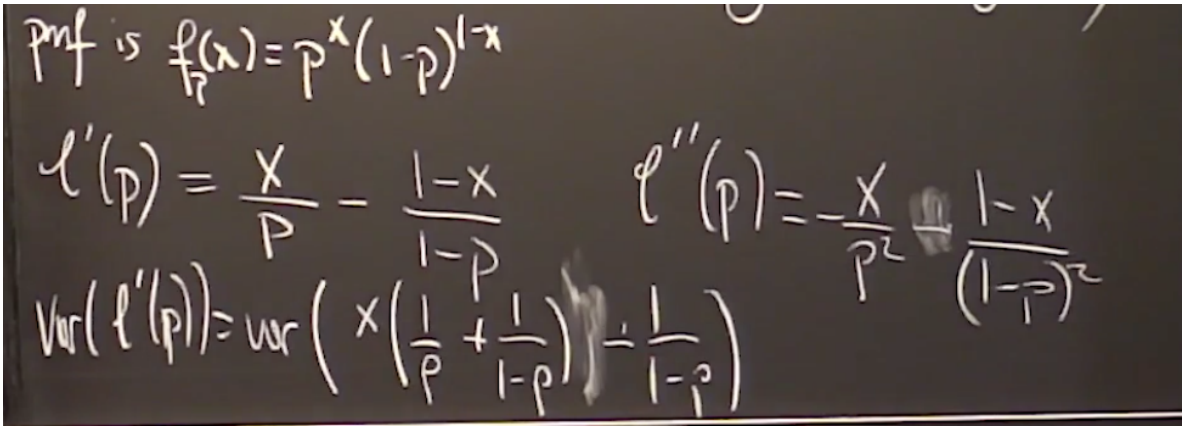
Let $\theta \in \Theta \subset \mathbb{R}^d$ and let $\left(E, \{\mathbf{P}_\theta\}_{\theta \in \Theta}\right)$ be a statistical model. Let $f_\theta(\mathbf{x})$ be the pdf of the distribution $\mathbf{P}_\theta$. Then, the Fisher information of the statistical model is

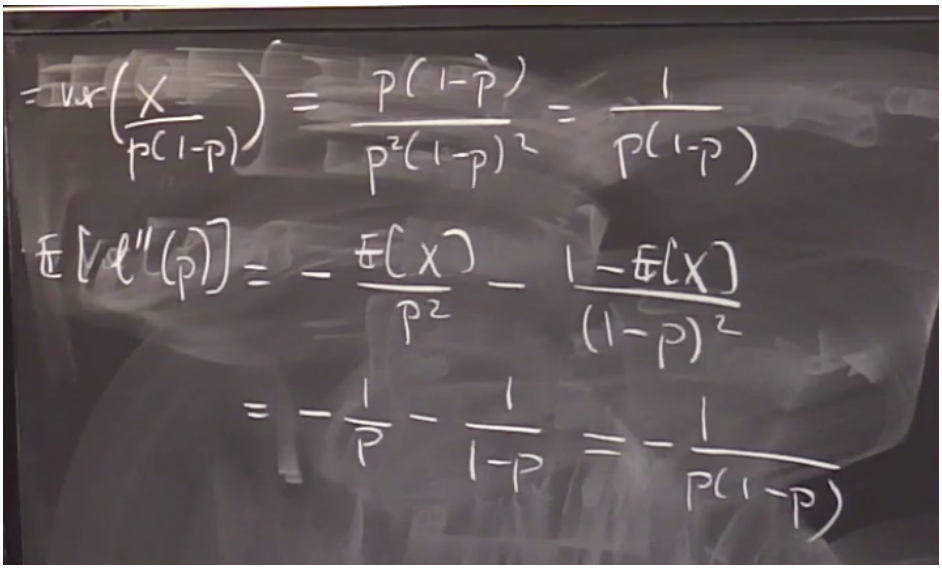
$$\mathcal{I}(\theta) = \mathrm{Cov}\left(\nabla \ell(\theta)\right) = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right],$$

where $\ell(\theta) = \ln f_\theta(\mathbf{X})$.

The definition when the distribution has a pmf $p_\theta(\mathbf{x})$ is also the same, with the expectation taken with respect to the pmf.

slide 37
this way of writing the PMF for Bernoulli is a clever way to say that when x is 1 the pmf is x and when x is 0 it is 1-p (instead of writing the 1(x=1) odd function)





Fisher Information of the Binomial Random Variable

1/1 point (graded)

Let X be distributed according to the binomial distribution of n trials and parameter $p \in (0, 1)$. Compute the Fisher information $\mathcal{I}(p)$.

Hint: Follow the methodology presented for the Bernoulli random variable in the above video.

$\mathcal{I}(p)$: n/(p*(1-p))

✔ **Answer:** n/(p*(1-p))

$\frac{n}{p \cdot (1-p)}$

pmf	$\binom{n}{k} p^k q^{n-k}$

Solution:

The logarithm of the pmf of a binomial random variable X, treated as a random function, can be written as

$$\ell(p) \triangleq \ln \binom{n}{X} + X \ln p + (n - X) \ln (1 - p), \quad X \in \{0, 1, \ldots, n\}.$$

The derivative of $\ell(p)$ with respect to p is

$$\ell'(p) = \frac{X}{p} - \frac{n - X}{1 - p},$$

which means the second derivative is

$$\ell''(p) = -\frac{X}{p^2} - \frac{n - X}{(1 - p)^2}.$$

The Fisher information $\mathcal{I}(p)$, therefore, is

$$\begin{aligned} \mathcal{I}(p) = -\mathbb{E}[\ell''(p)] &= \mathbb{E}\left[\frac{X}{p^2} + \frac{n - X}{(1 - p)^2}\right] \\ &= \frac{np}{p^2} + \frac{n - np}{(1 - p)^2} \\ &= \frac{n}{p(1 - p)}. \end{aligned}$$

Fisher Information of a Bernoulli-Like Random Variable

0/1 point (graded)

Consider the following experiment: You take a coin that lands a head (H) with probability $0 < p < 1$ and you toss it twice. Define X as the following random variable:

$$X = \begin{cases} 1 & \text{if outcome is HH} \\ 0 & \text{otherwise} \end{cases}$$

Compute the Fisher information $\mathcal{I}(p)$.

$\mathcal{I}(p)$: 1/(p*(1-p^2))

✖ **Answer:** 4/(1-p^2)

$\frac{1}{p \cdot (1-p^2)}$

Solution:

Following the Bernoulli and binomial examples,

$$\ell(p) \triangleq 2X \ln p + (1-X) \ln\left(1-p^2\right), \quad X \in \{0,1\}.$$

The derivative of $\ell(p)$ with respect to p is

$$\ell'(p) = \frac{2X}{p} - 2p \cdot \frac{1-X}{1-p^2},$$

which means the second derivative is

$$\ell''(p) = -\frac{2X}{p^2} - 2 \cdot \frac{(1-X)}{1-p^2} - 4p^2 \cdot \frac{1-X}{\left(1-p^2\right)^2}.$$

The Fisher information $\mathcal{I}(p)$, therefore, is

$$\begin{aligned}
\mathcal{I}(p) = -\mathbb{E}\left[\ell''(p)\right] &= \mathbb{E}\left[\frac{2X}{p^2} + 2 \cdot \frac{(1-X)}{1-p^2} + 4p^2 \cdot \frac{1-X}{\left(1-p^2\right)^2}\right] \\
&= \frac{2p^2}{p^2} + \frac{2\left(1-p^2\right)}{\left(1-p^2\right)} + 4p^2 \cdot \frac{1-p^2}{\left(1-p^2\right)^2} \\
&= 4 + \frac{4p^2}{1-p^2} \\
&= \frac{4}{1-p^2}
\end{aligned}$$

Fisher Information of a Modified Gaussian Random Vector

0/4 points (graded)

Let $\mathbf{X}$ be a Gaussian random vector with **independent** components $X^{(i)} \sim \mathcal{N}(\alpha + \beta t_i, 1)$ for $i = 1, \ldots, d$, where t_i are known constants and α and β are unknown parameters.

Compute the Fisher information matrix $\mathcal{I}(\theta)$ using the formula $\mathcal{I}(\theta) = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right]$.

Use **S_1** for $\sum_{i=1}^{d} t_i$ and **S_2** for $\sum_{i=1}^{d} t_i^2$.

$\mathcal{I}(\theta)_{1,1} =$ 0 ✖ **Answer:** d + 0*S_1 + 0*S_2

0

$\mathcal{I}(\theta)_{1,2} =$ 0 ✖ **Answer:** 0*d + S_1 + 0*S_2

0

$\mathcal{I}(\theta)_{2,1} =$ 0 ✖ **Answer:** 0*d + S_1 + 0*S_2

0

$\mathcal{I}(\theta)_{2,2} =$ 0 ✖ **Answer:** 0*d + 0*S_1 + S_2

0

Hint: Let $\theta = [\alpha \ \ \beta]^T$ denote the paramaters of the statistical model. $\ell(\theta)$ is a real-valued function of θ as given by the joint pdf at any fixed $\mathbf{x}$.

Solution:

Let $\theta = [\alpha \ \ \beta]^T$ denote the paramaters of the statistical model. The joint distribution of the random vector $\mathbf{X}$ is equal to the product of the marginal distributions of its components, $(2\pi^{-1/2}e^{-(x^(i)-\alpha-\beta t_i)/2}$, by independence. Therefore, the random vector $\mathbf{X}$ has the density

$$f_\theta(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\sum_{i=1}^{d}\left(x^{(i)}-\alpha-\beta t_i\right)^2}, \quad \text{where} \quad \mathbf{x} = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(d)} \end{bmatrix}^T \in \mathbb{R}^d.$$

Taking the log of the pdf yields (written as a random function)

$$\ell(\theta) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\left[\sum_{i=1}^{d}\left(\left(X^{(i)}-\beta t_i\right)^2 - 2\alpha\left(X^{(i)}-\beta t_i\right) + \alpha^2\right)\right]$$

Therefore,

$$\nabla\ell(\theta) = \begin{bmatrix} \sum_{i=1}^{d}\left(X^{(i)}-\beta t_i-\alpha\right) \\ \sum_{i=1}^{d}\left(t_i X^{(i)}-\beta t_i^2-\alpha t_i\right) \end{bmatrix},$$

from which we can obtain the hessian

$$\mathbf{H}\ell(\theta) = \begin{bmatrix} \sum_{i=1}^{d}(-1) & \sum_{i=1}^{d}(-t_i) \\ \sum_{i=1}^{d}(-t_i) & \sum_{i=1}^{d}(-t_i^2) \end{bmatrix}.$$

Therefore,

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right] = \begin{bmatrix} d & \sum_{i=1}^{d} t_i \\ \sum_{i=1}^{d} t_i & \sum_{i=1}^{d} t_i^2 \end{bmatrix},$$

where the expectation is taken with respect to the pdf of the random vector $\mathbf{X}$. Since none of the entries of the hessian contained any $X^{(i)}$, the expectation was simply the hessian matrix itself.

A Geometric View on the Fisher Information

1/1 point (graded)

Let $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ denote a statistical model. Recall that the MLE (maximum likelihood estimator) for one observation maximizes the log-likelihood for one observation, which is the random variable $\ell(\theta) = \ln L_1(X, \theta)$ where $X \sim \mathbf{P}_\theta$. Suppose we observe $X_1 = x_1$, and now consider the graph of the function $\theta \mapsto \ln L_1(x_1, \theta)$.

What does the Fisher information $\mathcal{I}(\theta)$ represent?

Hint: Use the definition $\mathcal{I}(\theta) = -\mathbb{E}\left[\ell''(\theta)\right]$.

- ◯ It gives you an approximation for the true parameter θ^*.
- ◯ It tells you the average slope of the function $\theta \mapsto \ln L(x_1, \theta)$ is.
- ◉ It tells you, on average, how curved the function $\theta \mapsto \ln L(x_1, \theta)$ is.

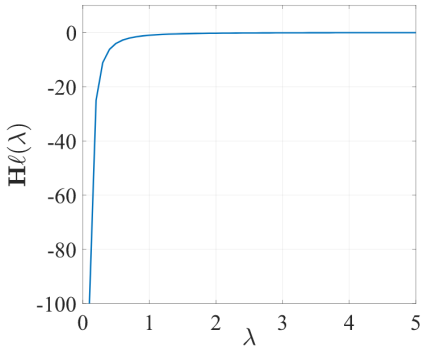
✔

Solution:

The third choice "It tells you, on average, how curved the function $\theta \mapsto \ln L(x_1, \theta)$ is." is correct. Recall that the Fisher information is also equal to the expected second-derivative of the log-likelihood: $\mathcal{I}(\theta) = -\mathbb{E}[\ell''(\theta)]$. Since the second-derivative measures concavity/convexity (how curved a function is at a particular point), $\mathcal{I}(\theta)$ measures the *average* curvature of the function $\theta \mapsto \ell(\theta) = \ln L_1(x_1, \theta)$.

Remark: It turns out that the Fisher information tells how curved (on average) the log-likehood $\ln L_n(x_1, \ldots, x_n, \theta)$ for several samples $X_1 = x_1, \ldots, X_n = x_n$ is. In particular, $\mathcal{I}(\theta^*)$ tells how curved (on average) the log-likelihood is near the true parameter. As a rule of thumb, if the Fisher information $\mathcal{I}(\theta^*)$ is large, then we expect the MLE to give a good estimate for θ^*.

Consider the exponential statistical model with $f_\lambda(x) = \lambda e^{-\lambda x}$, $x \geq 0$ and $\lambda \in (0, \infty)$. The second derivative of $\ell(\lambda)$, which is $\mathbf{H}\ell(\lambda) = \ell''(\lambda)$ (you will compute later in a homework exercise), does not depend upon x and is shown in the following figure.



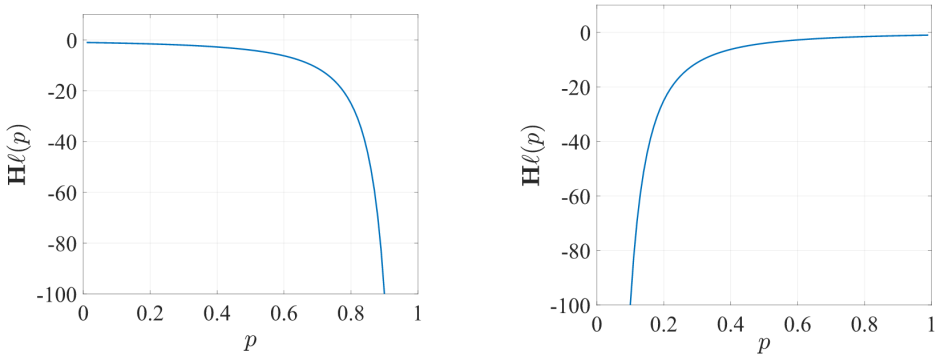
$\ell''(\lambda)$ for the exponential statiscal model with parameter λ for all X.

The Fisher information, $\mathcal{I}(\theta) = -\mathbb{E}[\mathbf{H}\ell(\theta)]$, captures the negative of the expected curvature of $\ell(\theta)$. For example, for the exponential statistical model, the expected curvature of $\ell(\lambda)$ is $\ell''(\lambda)$ itself and this is shown in the figure above. The Fisher information in this case is always positive. The fact that $\ell''(\lambda)$ is negative for all x also means that the log-likelihood function $\ell(\lambda)$ is a concave function of λ for all x.

If we consider the Bernoulli statistical model with parameter $p \in (0, 1)$, we derived in a lecture video (also can be seen in the slides) that

$$\ell''(p) = -\frac{X}{p^2} - \frac{1 - X}{(1 - p)^2}, \quad X \in \{0, 1\}.$$

Here, we see that $\ell''(p)$ is a random function that depends upon X. The following two figures show $\ell''(p)$ for $X = 0$ and $X = 1$.



$\ell''(p)$ for the Bernoulli statiscal model with parameter $p \in (0, 1)$ for $X = 0$ (left) and $X = 1$ (right).

The **asymptotic normality of the ML estimator** , which will be discussed in the upcoming video, depends upon the Fisher information. For a one-parameter model (like the exponential and Bernoulli), the asymptotic normality result will say something along the lines of following: that the asymptotic variance of the ML estimator is inversely proportional to the value of Fisher information at the true parameter θ^* of the statistical model. This means that if the value of Fisher information at θ^* is high, then the asymptotic variance of the ML estimator for the statistical model will be low.

Weak Consistency of the MLE

1/1 point (graded)

Let $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ denote a statistical model associated to a statistical experiment $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ for some true parameter θ^* that we would like to estimate. You construct the maximum likelihood estimator $\widehat{\theta}_n^{\text{MLE}}$ for θ^*. Which of the following conditions is **not** necessary for the MLE $\widehat{\theta}_n^{\text{MLE}}$ to converge to θ^* in probability ?

- ○ The model $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ is identified. (Recall that the parameter θ is identified if the map $\theta \mapsto \mathbf{P}_\theta$ is injective.)
- ○ For all $\theta \in \Theta$, the support of $\mathbf{P}_\theta$ does not depend on θ.
- ◉ The MLE $\widehat{\theta}_n^{\text{MLE}}$ is given by the sample average.
- ○ The Fisher information $\mathcal{I}(\theta)$ is non-zero in an interval containing true parameter θ^*. (Note that this is what it means for a 1×1 matrix, a scalar, to be invertible.)

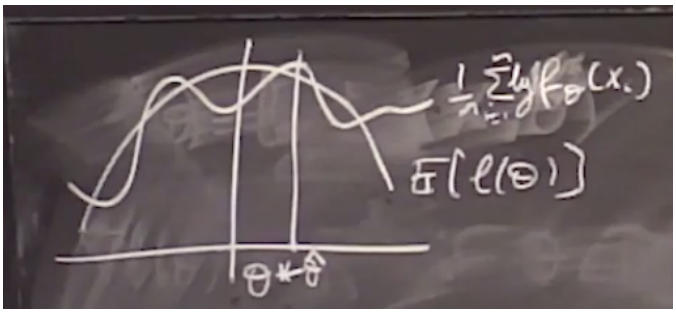
✔

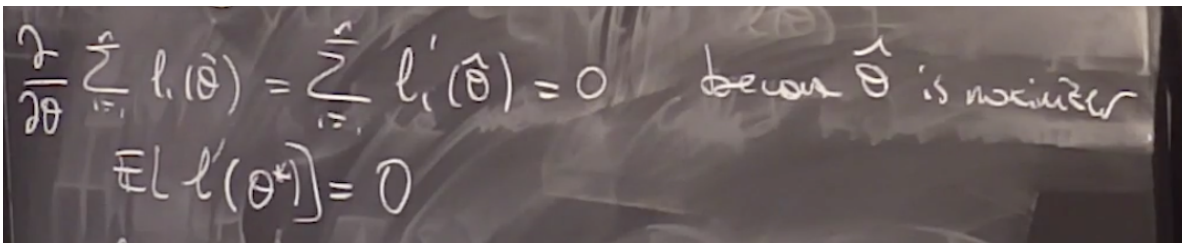
Solution:

The power of the theorem for the convergence of the MLE is that it applies even in situations where the MLE is *not* the sample average. Hence, the third choice, "The MLE $\widehat{\theta}_n^{\text{MLE}}$ is given by the sample average.", is correct, as it is not an assumption needed for the theorem statement. On the other hand, the first, second, and fourth choices are all hypotheses in the theorem statement regarding the convergence of the MLE. Hence, "The model $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ is identified.", "For all $\theta \in \Theta$, the support of $\mathbf{P}_\theta$ does not depend on θ.", and "The Fisher information $I(\theta)$ is invertible in an interval containing the true parameter θ^*" are all incorrect responses.

continuation of slide 38
what is the Fisher Information for

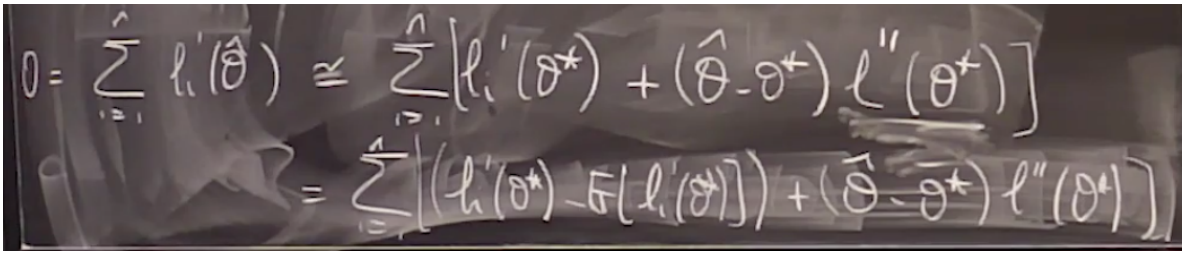
using an estimator (squiggle line)
want closeness in y axis to translate to closeness in x axis



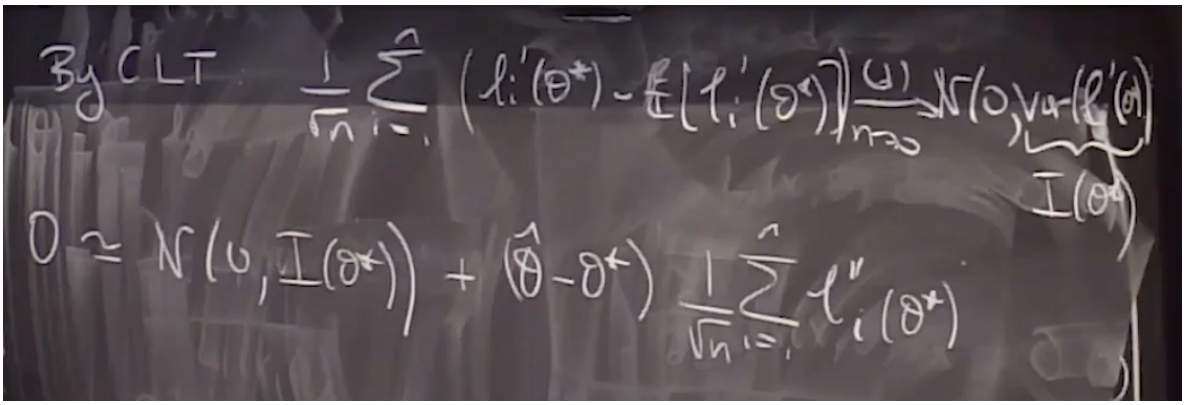


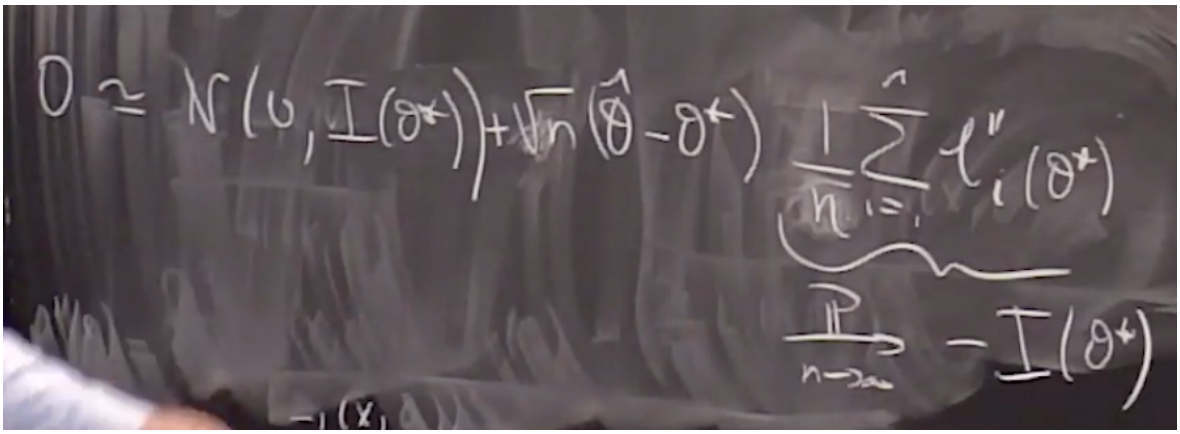
equal to 0 because theta* is the maximum of the expected log likelihood

1st order taylor expansion of l(theta_hat) around l(theta*)

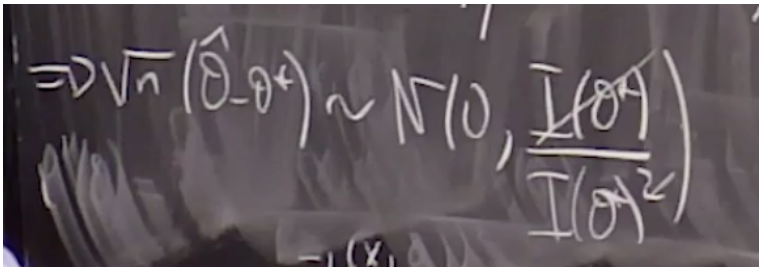


now have sums of random variables minus their expectations
so the central limit theorem comes in





swapped the n around so now we know what the term on the right converges in probability to (Fisher information)



which is the formula at the bottom of slide 38 (1/Fisher information)

Asymptotic = approaches but never touches the line

Fisher Information and Asymptotic Normality of the MLE

1/1 point (graded)

Consider the statistical model $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ associated to the statistical experiment $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$, where θ^* is the true parameter. Assume that the conditions of the theorem for the convergence of the MLE hold. Which of the following statements about the Fisher information $\mathcal{I}(\theta)$ is true?

- ◯ The Fisher information $\mathcal{I}(\theta^*)$ at the true parameter gives a good approximation for θ^*.
- ◯ The Fisher information $\mathcal{I}(\theta^*)$ at the true parameter determines the asymptotic mean of the random variable $\widehat{\theta}_n^{\mathrm{MLE}}$.
- ◉ The Fisher information $\mathcal{I}(\theta^*)$ at the true parameter determines the asymptotic variance of the random variable $\widehat{\theta}_n^{\mathrm{MLE}}$.

✔

Solution:

As stated in the theorem,

$$\sqrt{n}(\widehat{\theta}_n^{\mathrm{MLE}} - \theta^*)$$

converges to a normal random variable $\mathcal{N}(0, \mathcal{I}(\theta^*)^{-1})$. Hence, the Fisher information determines the asymptotic variance, and so the third choice is correct.

Asymptotic Normality of the MLE

0/1 point (graded)

Consider the statistical model $(\{0,1\}, \{\mathrm{Ber}(\theta)\}_{\theta\in(0,1)})$. Let $\ell(\theta)$ denote the **log-likelihood of one observation** of this model. You observe samples $X_1, \ldots, X_n \sim \mathrm{Ber}(\theta^*)$ and construct the MLE $\widehat{\theta}_n^{\mathrm{MLE}}$ for θ^*. By the theorem for the convergence of the MLE (you are allowed to assume that all necessary conditions for this theorem hold), this implies that

$$\sqrt{n}(\widehat{\theta}_n^{\mathrm{MLE}} - \theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$

for some constant σ^2 that depends on θ^*. The quantity σ^2 is referred to as the **asymptotic variance** . Use the theorem for the convergence of the MLE to find the expression for σ^2.

What is σ^2? Express your answer in terms of $T := \theta^*$.

Type **T** for T, using the variable T to stand for θ^*.

$\sigma^2 =$ 1/T ✖ **Answer:** T*(1- T)

$\frac{1}{T}$

Solution:

We have that for this model the Fisher information is $\mathcal{I}(\theta) = \frac{1}{\theta} + \frac{1}{(1-\theta)} = \frac{1}{\theta(1-\theta)}$. Applying the theorem for the convergence of the MLE,

$$\sqrt{n}(\widehat{\theta}_n^{\mathrm{MLE}} - \theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \mathcal{I}(\theta^*)^{-1}\right)$$

Hence,

$$\sigma^2 = \mathcal{I}(\theta^*)^{-1} = \theta^*(1-\theta^*).$$

Remark: Alternatively, the asymptotic variance can be computed directly from the MLE, which is given, in the Bernoulli case, by the sample mean $\overline{X}_n$.

Estimating Moments

0/1 point (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ be integer-valued random variables and let $(\mathbb{Z}, \{\mathbf{P}_\theta\}_{\theta\in\Theta})$ be the associated statistical model. Let p_θ denote the pmf of $\mathbf{P}_\theta$. Assume that for all $\theta \in \Theta$, the ***k*-th moment**

$$m_k(\theta) := \mathbb{E}[X^k] = \sum_{x\in\mathbb{Z}} x^k p_\theta(x)$$

exists for all $k \geq 1$. Use the law of large numbers to fill in the formula so that $\widehat{m}_k(\theta)$ is a consistent estimator for $m_k(\theta)$.

(Type **X_i** for X_i.)

$\widehat{m}_K(\theta) = \frac{1}{n}\sum_{i=1}^{n}$ k*X_i ✖ **Answer:** X_i^k

$k \cdot X_i$

STANDARD NOTATION

Solution:

The weak law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \to \mathbb{E}[X^k],$$

where the convergence is in probability.

Mapping Parameters to Moments I

2/2 points (graded)

Let $(\mathbb{R}, \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma > 0})$ be the statistical model of a normal random variable X. Let

$$m_k(\mu, \sigma) = \mathbb{E}[X^k]$$

denote the k-th moment of X. Let $\psi : \mathbb{R} \times (0, \infty) \to \mathbb{R}^2$ be defined by $\psi(\mu, \sigma) = (m_1(\mu, \sigma), m_2(\mu, \sigma))$. (Since we have two parameters of interest, μ and σ, it makes sense to work with the first two moments. The hope is that the two moments will uniquely determine the paramters of interest μ and σ.)

Express $m_1(\mu, \sigma)$ and $m_2(\mu, \sigma)$ in terms of μ and σ.

$m_1(\mu, \sigma) =$ mu ✔ **Answer:** mu

μ

$m_2(\mu, \sigma) =$ mu^2 + sigma^2 ✔ **Answer:** mu^2 +sigma^2

$\mu^2 + \sigma^2$

STANDARD NOTATION

Solution:

Note that

$$\begin{aligned} m_1(\mu, \sigma) &= \mathbb{E}[X] = \mu \\ m_2(\mu, \sigma) &= \mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \left(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\right) = \mu^2 + \sigma^2. \end{aligned}$$

Hence, $\psi(\mu, \sigma) = (\mu, \mu^2 + \sigma^2)$.

Mapping Parameters to Moments II

2/3 points (graded)

Let

$$\begin{aligned} \psi : \mathbb{R} \times (0, \infty) &\to \mathbb{R}^2 \\ (\mu, \sigma) &\mapsto (m_1(\mu, \sigma), m_2(\mu, \sigma)). \end{aligned}$$

denote the moments map considered in the previous problem, where $m_k(\mu, \sigma)$ denotes the k-th moment of the distribution $N(\mu, \sigma^2)$.

Is ψ one-to-one on the domain $\mathbb{R} \times (0, \infty)$? (Equivalently, given the outputs m_1 and m_2, can we use them to uniquely reconstruct $\mu \in \mathbb{R}$ and $\sigma > 0$?)

- (●) Yes
- ( ) No

✔

If ψ is one-to-one on the given domain and $\psi(\mu, \sigma) = (m_1, m_2)$, what is μ expressed in terms of m_1 and m_2? (If ψ is not one-to-one, enter 0.)

$\mu =$ m_1 ✔ **Answer:** m_1

m_1

If ψ is one-to-one on the given domain and $\psi(\mu, \sigma) = (m_1, m_2)$, what is σ expressed in terms of m_1 and m_2? (If ψ is not one-to-one, enter 0.)

$\sigma =$ m_2 - (m_1)^2 ✖ **Answer:** sqrt(m_2 - m_1^2)

$m_2 - (m_1)^2$

Solution:

Note that

$$\begin{aligned} m_1(\mu, \sigma) &= \mathbb{E}[X] = \mu \\ m_2(\mu, \sigma) &= \mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \left(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\right) = \mu^2 + \sigma^2. \end{aligned}$$

Hence, $\psi(\mu, \sigma) = (\mu, \mu^2 + \sigma^2)$. This function is one-to-one on the domain $\mathbb{R} \times (0, \infty)$. Since $m_1(\mu, \sigma) = \mu$, we can reconstruct the first parameter directly from the first moment: $\mu = m_1$.

Next, since we know $m_2(\mu, \sigma) = \sigma^2 + \mu^2$, we can back-solve for σ:

$$\sigma = \sqrt{m_2 - \mu^2} = \sqrt{m_2 - m_1^2}.$$

Above we took the positive square-root because we have insisted a priori that $\sigma > 0$.

Remark: Assuming that m_1 and m_2 are one of the outputs of ψ, we have essentially shown how to construct $\psi^{-1}(m_1, m_2)$. In general, computing inverses is a computationally difficult problem, but in this particular example, the function ψ is simple enough that it is possible to invert by hand.

Method of Moments Concept Question I

2/2 points (graded)

Let $(E, \{\mathbf{P}_\theta\}_{\theta\in\Theta})$ denote a statistical model associated to a statistical experiment $X_1, \dots, X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ where $\theta^* \in \Theta$ is the true parameter. Assume that $\Theta \subset \mathbb{R}^d$ for some $d \geq 1$. Let $m_k(\theta) := \mathbb{E}[X^k]$ where $X \sim \mathbf{P}_\theta$. $m_k(\theta)$ is referred to as the ***k*-th moment of $\mathbf{P}_\theta$** . Also define the moments map:

$$\begin{aligned} \psi : \Theta &\to \mathbb{R}^d \\ \theta &\mapsto (m_1(\theta), m_2(\theta), \dots, m_d(\theta)). \end{aligned}$$

Assume that ψ is one-to-one (and hence, invertible).

Which of the following is equal to θ^*?

- $(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$
- $\psi(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$
- **$\psi^{-1}(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$**
- $\psi^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \dots, \frac{1}{n}\sum_{i=1}^n X_i^d\right)$

✔

Which of the following is the method of moments estimator for θ^*?

- $(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$
- $\psi(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$
- $\psi^{-1}(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$
- **$\psi^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \dots, \frac{1}{n}\sum_{i=1}^n X_i^d\right)$**

✔

Solution:

Observe that $\psi(\theta^*) = (m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$ by definition of ψ. Since ψ is invertible, then we know that $\psi^{-1}(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*)) = \theta^*$. Hence, $\psi^{-1}(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*)$ is the correct response to the first question.

The remaining choices are incorrect.

- $(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$ is the list of moments of $\mathbf{P}_{\theta^*}$, not the parameter θ^* itself.
- $\psi(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$ is incorrect, as this is really $\psi^2(\theta^*)$, which is not necessarily θ^*.
- This is the method of moments estimator, not the true parameter θ^*.

The method of moments estimator is given by

$$\widehat{\theta}_n^{\text{MM}} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \dots, \frac{1}{n}\sum_{i=1}^n X_i^d\right),$$

Remark: The above expression is consistent with the procedure we followed in the previous problems that we used to construct the method of moments estimator for a Gaussian statistical model with unknown mean and variance. Namely, we expressed the true parameters in terms of the true moments, and then plugged in the sample means into that expression. Informally, since we expect the sample means to give a good approximation for the true moments, plugging in the sample moments into the expression for the true parameters (in terms of the moments) should also give a good approximation for the true parameters. This is the strategy of the method of moments, and in general, the strategy of replacing expectations with averages is a recurring theme in statistics and in this course.

Applying the Method of Moments to a Gaussian Statistical Model

1/2 points (graded)

We let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu^*, (\sigma^*)^2)$ and consider the associated statistical model $(\mathbb{R}, \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma > 0})$. Let

$$\begin{aligned} \psi : \mathbb{R} \times (0, \infty) &\to \mathbb{R}^2 \\ (\mu, \sigma) &\mapsto (m_1(\mu, \sigma), m_2(\mu, \sigma)). \end{aligned}$$

denote the moments map considered in the previous problem, where $m_k(\mu, \sigma)$ denotes the k-th moment of the distribution $N(\mu, \sigma^2)$.

To answer the next question, you should recall:

1. your result on writing μ and σ in terms of m_1 and m_2 obtained in the previous problem, and
2. the estimators $\widehat{m_1}$ and $\widehat{m_2}$ (the sample moments) for the true moments m_1 and m_2.

Suppose we observe the data-set $X_1 = 0.5, X_2 = 1.8, X_3 = -2.3, X_4 = 0.9$.

What is the method of moments estimator $\widehat{\mu}^{\mathrm{MM}}$ for μ^* evaluated on this data-set? (You are encouraged to use whatever computational tools may be helpful.)

$\widehat{\mu}^{\mathrm{MM}}(0.5, 1.8, -2.3, 0.9) =$ 0.225 ✔ **Answer:** 0.225

What is the method of moments $\widehat{\sigma}^{\mathrm{MM}}$ estimator for σ^* evaluated on this data-set? (You are encouraged to use whatever computational tools may be helpful.)

$\widehat{\sigma}^{\mathrm{MM}}(0.5, 1.8, -2.3, 0.9) =$ 0 ✖ **Answer:** 1.532

Solution:

Since we computed ψ^{-1} explicitly in the previous problem, we can apply the method of moments to estimate the true parameters μ^* and σ^*. Let

$$\widehat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

denote the first and second sample moments, respectively. Then the **method of moments estimator** is defined to be

$$(\widehat{\mu}_n^{\mathrm{MM}}, \widehat{\sigma}_n^{\mathrm{MM}}) := \psi^{-1}\,(\widehat{m}_1, \widehat{m}_2) = (\widehat{m}_1, \sqrt{\widehat{m}_2 - \widehat{m}_1^2})\,.$$

Using a calculator (or other computational software) we can compute,

$$\widehat{m}_1\,(0.5, 1.8, -2.3, 0.9) = \frac{0.5 + 1.8 - 2.3 + 0.9}{4} \approx 0.225$$

and

$$\widehat{m}_2\,(0.5, 1.8, -2.3, 0.9) = \frac{(0.5)^2 + (1.8)^2 + (-2.3)^2 + (0.9)^2}{4} \approx 2.3975.$$

Applying the method of moments,

$$\widehat{\mu}_n^{\mathrm{MM}}\,(0.5, 1.8, -2.3, 0.9) = \widehat{m}_1\,(0.5, 1.8, -2.3, 0.9) \approx 0.225.$$

and

$$\widehat{\sigma}_n^{\mathrm{MM}}\,(0.5, 1.8, -2.3, 0.9) = \sqrt{\widehat{m_2} - (\widehat{m_1})^2} \approx \sqrt{2.3975 - (0.225)^2} \approx 1.532.$$