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 $1 - p^2$.

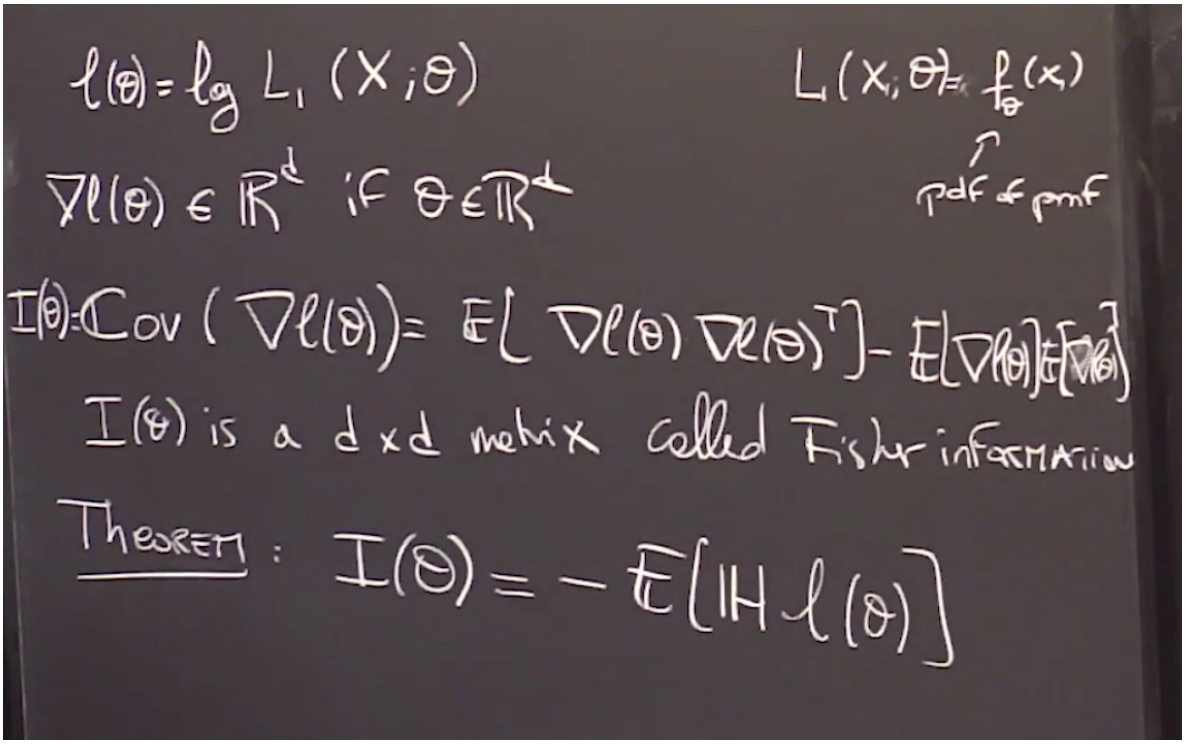
Finding the log-likelihood function involves writing down the pmf of X as a one-line equation:

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

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.

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] $\int_{-\infty}^{\infty} \frac{(\frac{\partial}{\partial\theta} f_\theta(x))^2}{f_\theta(x)}\, dx$ ✔

✖

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 $\int_{-\infty}^{\infty} \frac{(\frac{\partial}{\partial\theta} f_\theta(x))^2}{f_\theta(x)}\, dx$ is correct because, using the previous bullet,

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

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

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

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

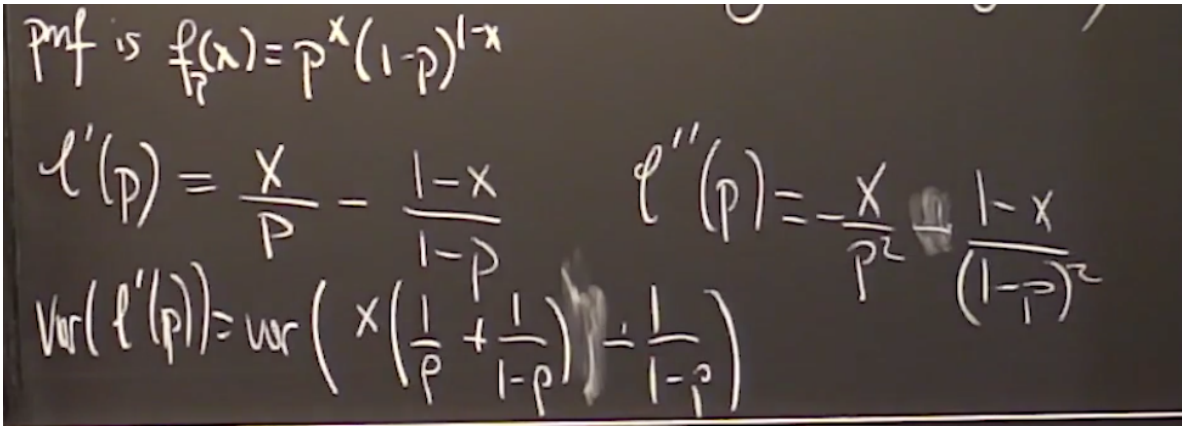
Let $\theta \in \Theta \subset \mathbb{R}^d$ and let $\left(E, \{\mathbf{P}_\theta\}_{\theta \in \Theta}\right)$ be a statistical model. Let $f_\theta(\mathbf{x})$ be the pdf of the distribution $\mathbf{P}_\theta$. Then, the Fisher information of the statistical model is

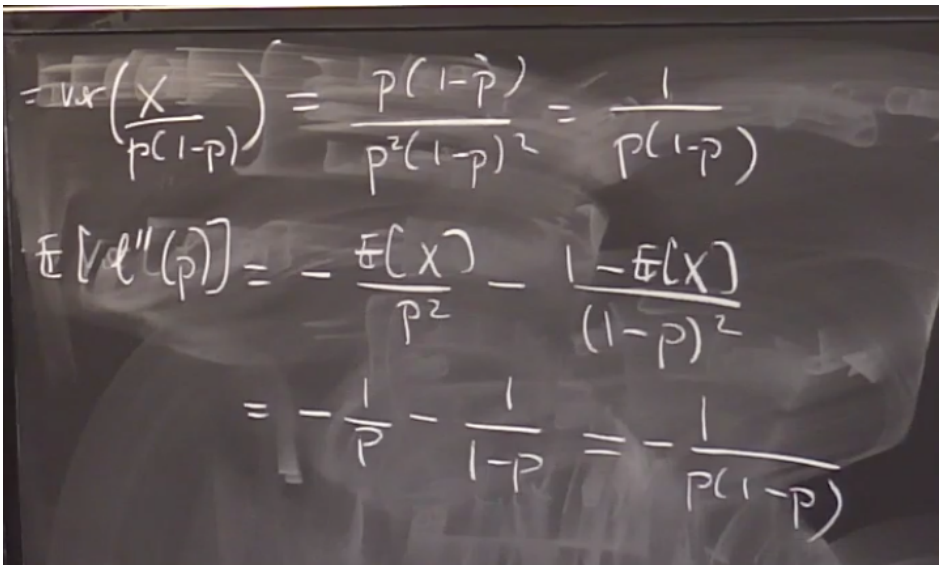
$$\mathcal{I}(\theta) = \mathrm{Cov}\left(\nabla \ell(\theta)\right) = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right],$$

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,

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

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

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

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

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