The Cramer-Rao Lower Bound – Derivation and Examples

I will build up the Cramer-Rao lower bound in a series of steps, showing the general formulas for each concept and then examples using the normal distribution and the binomial distribution.

The Score

When we do maximum likelihood expectation, we are treating the likelihood as a function of the population parameter, and we are considering the data fixed. However, we can also treat the likelihood as a function of the data points. These data points are random quantities, and functions of random quantities are other random quantities. Viewed in this way, the likelihood is a random variable – i.e. a quantity that depends on the experiment – the particular random quantities drawn. To derive some of the properties of the likelihood, we need to think of it both as a function of the population parameter(s) and as a function of the sample, i.e. as a random variable. Similarly, the log-likelihood can be thought of as a function of the population parameter(s) and as a function of the sample.

When we view the log-likelihood as a function of the population parameter, we can differentiate it with respect to the parameter. This derivative is called the score. Because it is also a function of the particular data values, we write this derivative as a partial derivative:

$$\frac{\partial}{\partial\theta} l\left(\mathbf{x}\,|\,\theta\right) = \frac{\partial}{\partial\theta} \log f\left(\mathbf{x}\,|\,\theta\right)$$

Example - Normal Distribution

For the estimator of the mean of the normal distribution, the parameter θ is the mean, μ.

The likelihood (which is the same as the normal probability density) is:

$$L\left(\mu\,|\,x_1,...,x_n,\sigma\right) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right)$$

The log-likelihood is:

$$l\left(\mu\,|\,x_1,...,x_n,\sigma\right) = -n\log\sigma - \frac{n}{2}\log\left(2\pi\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$$

The score is the derivative. I will use the letter *g* for the score.

$$g\left(\mu\right) = \frac{\partial l}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right) = \frac{n}{\sigma^2}\left(\overline{x}-\mu\right)$$

Notice that the score is a linear function of the sample mean, and is directly proportional to the difference between the sample mean and the population mean. Let's look at what the score means for some set of data. If the data points are on average below the true population mean, then the score is negative. The score is a derivative with respect to the mean, so the negative score implies that the log-likelihood decreases as the mean increases. This makes sense – we are moving the mean farther away from the data, making the probability lower. On the other hand, if the data points are on average above the population mean, the score is positive. The log-likelihood increases as the population mean increases, which brings it closer to the data.

Example - Binomial Distribution

The binomial probability for sample size n and population probability π is:

$$\text{prob}\left(X = k\right) = \frac{n!}{k!\left(n-k\right)!}\pi^k\left(1-\pi\right)^{n-k}.$$

The log-likelihood is:

$$l\left(\pi;k\right) = \log\left(n!\right) - \log\left(k!\right) - \log\left(\left(n-k\right)!\right) + k\log\pi + \left(n-k\right)\log\left(1-\pi\right)$$

The score is the derivative with respect to π, which is

$$\frac{\partial}{\partial\pi}l\left(\pi;k\right) = \frac{k}{\pi} - \frac{\left(n-k\right)}{1-\pi}.$$

You should confirm that if k / n is lower than π, the score will be negative. If k / n is higher than π, the score will be positive.

Here's a table of scores for a binomial where $n = 10$, $\pi = 0.35$.

k	score
0	-15.3846
1	-10.989
2	-6.59341
3	-2.1978
4	2.197802
5	6.593407
6	10.98901
7	15.38462
8	19.78022
9	24.17582
10	28.57143

You will notice that the sign is as predicted, and also that the values of the score are farther away from 0 when the data is far away from the population parameter. You might imagine the score as a measure of "force" that the data is exerting on the population parameter, trying to bring it more in line with the data. A negative score is trying to lower the population parameter, and a positive score is trying to raise it. (With a negative score, lowering the population parameter will raise the probability.)

The Expectation of the Score

You can see that the score takes on both positive and negative values for different values of the data.

We can show that the expectation of the score is 0:

$$E\left(\frac{\partial}{\partial\theta}\log f\left(\mathbf{x}\,|\,\theta\right)\right)=0$$

Proof:

Recall that the derivative of a logarithm is of the form $\frac{d}{dx}\log\left(g\left(x\right)\right)=\frac{1}{g\left(x\right)}\frac{d}{dx}g\left(x\right)$.

So $\frac{\partial}{\partial\theta}\log f\left(\mathbf{x}\,|\,\theta\right)=\frac{1}{f\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)$.

Taking expectations:

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}\log f\left(\mathbf{x}\,|\,\theta\right)\right)=\mathrm{E}\left(\frac{1}{f\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)$$

Note here that **x** can be a vector of data points.

To take an expectation, we integrate or sum over all possible values for all the observations, weighted by the probability or the probability density.

$$\mathrm{E}\left(\frac{1}{f\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)=\int\left(\frac{1}{f\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)f\left(\mathbf{x}\,|\,\theta\right)d\mathbf{x}=\int\left(\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)d\mathbf{x}$$

or

$$\mathrm{E}\left(\frac{1}{f\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)=\sum_{\text{all }\mathbf{x}\text{'s}}\left(\frac{1}{f\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)f\left(\mathbf{x}\,|\,\theta\right)=\sum_{\text{all }\mathbf{x}\text{'s}}\left(\frac{\partial}{\partial\theta}f\left(\mathbf{x}\,|\,\theta\right)\right)$$

Notice how using the log-likelihood, and then taking the derivative, allows us to "cancel out" the probability or probability density. This is a reason why the log-likelihood is of theoretical as well as practical interest. Also notice that even if *f* is a discrete probability, and so the **x**'s only take on discrete values, typically the parameter θ will be continuous, so we can take the derivative.

If the probability density or probability function is well-behaved (the buzzword is that it has to meet certain "regularity conditions"), we can reverse the order of the integral/summation and the derivative, to get

$$\int\left(\frac{\partial}{\partial\theta} f\left(\mathbf{x}\mid\theta\right)\right)d\mathbf{x} = \frac{\partial}{\partial\theta}\int\left(f\left(\mathbf{x}\mid\theta\right)\right)d\mathbf{x} = \frac{\partial}{\partial\theta}(1) = 0$$

or QED.

$$\sum_{\text{all x's}}\left(\frac{\partial}{\partial\theta} f\left(\mathbf{x}\mid\theta\right)\right) = \frac{\partial}{\partial\theta}\sum_{\text{all x's}}\left(f\left(\mathbf{x}\mid\theta\right)\right) = \frac{\partial}{\partial\theta}(1) = 0$$

This probably looks like pulling a rabbit out of a hat, so I'll give a couple of examples.

Example - Normal Distribution

From above, the score is:

$$g\left(\mu\right) = \frac{\partial l}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right).$$

$$E\left(g\left(\mu\right)\right) = E\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)\right] = 0, \text{ since each } x_i \text{ has expectation } \mu.$$

Intuitively, each x_i has a symmetric distribution around μ, so the positive and negative contributions to the score have to exactly average out to 0.

Example - Binomial Distribution

This can be seen most easily by using a numeric table. To within the precision of computer arithmetic, the expectation (the sum of $score_k$ * $prob_k$) is 0.

Given n=10, p = 0.35			
k	likelihood	score	likelihood*score
0	0.013462743	-15.385	-0.2071
1	0.072491695	-10.989	-0.7966
2	0.175652953	-6.5934	-1.1582

3	0.252219625	-2.1978	-0.5543
4	0.237668493	2.1978	0.52235
5	0.153570411	6.59341	1.01255
6	0.0689098	10.989	0.75725
7	0.021203015	15.3846	0.3262
8	0.004281378	19.7802	0.08469
9	0.000512302	24.1758	0.01239
10	2.75855E-05	28.5714	0.00079
		Sum	-8E-16

The Fisher Information

I described the score as a measure of "force" that the data is exerting on the population parameter, trying to bring it more in line with the data. Alternatively, it could be seen as a measure of force that the population parameter is exerting on the data, or more symmetrically, as the force driving them to be close to each other. To switch metaphors, we can also think of it as a measure of "information" that the data is providing about the parameter. The stronger the force, the closer the data is likely to be to the parameter, and the more information about the parameter we are getting from the data.

Let's take a look at the score function for the normal distribution again:

$$g\left(\mu;\mathbf{x}\right)=\frac{\partial l}{\partial \mu}=\frac{n}{\sigma^2}\left(\overline{x}-\mu\right)$$

As described above, the farther apart the sample mean and the population mean, the greater the absolute value of the score. Also, the score is proportional to n/σ^2. That says that the score goes up as the number of data points goes up, and goes up as the variance (of the x values) goes down.

The values of the score will get more extreme as the observed data is farther away from the population parameter. We would like some measure of that "force" that doesn't depend on the specific data points, but is still in some sense a measure of how strongly the data is pulled toward the population parameter. We can't just take an expectation (which would reflect the information from all possible outcomes), because we have seen the expectation is 0. Instead, we take the expectation of the square. This expectation is called the Fisher Information.

$$I\left(\theta\right)=\mathrm{E}\left(\left(g\left(\theta;\mathbf{x}\right)\right)^2\right)=E\left(\left(\frac{\partial}{\partial\theta}\log f\left(\mathbf{x}\,|\,\theta\right)\right)^2\right)=\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f\left(\mathbf{x}\,|\,\theta\right)\right)$$

Note that the expectation of the square is equal to the variance, because the expectation of the score is 0. Also note that the Fisher information is an expectation taken over possible data values, so it is not a function of the x values, but it may be a function of θ.

Example - Normal Distribution

$$I = E\left(\left(g\left(\mu;\mathbf{x}\right)\right)^2\right) = \frac{1}{\left(\sigma^2\right)^2}\sum_{i=1}^{n} E\left(x_i-\mu\right)^2 = \frac{n\sigma^2}{\left(\sigma^2\right)^2} = \frac{n}{\sigma^2}$$

Notice that for the normal distribution, the information is not a function of the mean, but it is a function of the variance. Also notice that the information is directly proportional to the number of data points, and indirectly proportional to the variance. Hopefully, this makes sense for something called "information." If we have more data points, the data provides more information about the population mean. If the variance of the values is higher, each data point provides less information about the population mean.

Example - Binomial Distribution

For the binomial, it turns out that the information is $\frac{n}{\pi\left(1-\pi\right)}$. This is derived later. For now, you can confirm it for the specific case we've been working with,

$$\frac{n}{\pi\left(1-\pi\right)} = \frac{10}{.35*.65} = 43.956:$$

Given n=10, p = 0.35			
k	likelihood	score^2	likelihood*score^2
0	0.013462743	236.686	3.18645
1	0.072491695	120.758	8.75398
2	0.175652953	43.473	7.63616
3	0.252219625	4.83033	1.21831
4	0.237668493	4.83033	1.14802
5	0.153570411	43.473	6.67617
6	0.0689098	120.758	8.32143
7	0.021203015	236.686	5.01847
8	0.004281378	391.257	1.67512
9	0.000512302	584.47	0.29943
10	2.75855E-05	816.327	0.02252
		Sum	43.956

Again, the information is proportional to the sample size *n*, and is inversely proportional to the variance (recall that the variance of a Bernoulli variable with probability π is $\pi(1-\pi)$).

The Cramer-Rao Lower Bound

The Cramer-Rao bound says that for any unbiased estimator of a population parameter, the lowest possible variance is $1/I$, where I is the Fisher information. There are some technical regularity conditions, but these hold for the usual probability distributions and estimators that we use in statistics.[1] Before proving this, I will use the same two examples.

Example - Normal Distribution

From above:

$$I = \frac{n}{\sigma^2}, \quad \frac{1}{I} = \frac{\sigma^2}{n}$$

Notice that $1/I$ is exactly the variance of the sample mean. This tells us that the sample mean is the "best" unbiased estimator, i.e. there can be no unbiased estimator with a lower variance.

Example - Binomial Distribution

From above, the information is $\frac{n}{\pi(1-\pi)}$. The Cramer-Rao lower bound is $\frac{\pi(1-\pi)}{n}$. This is the variance of a binomial proportion (the binomial count k divided by n). So if we take the sample proportion k / n as the natural estimator of π, its variance matches the Cramer-Rao lower bound.

Proof

The proof is very clever, but it's another one of those "pulling a rabbit out of a hat" derivations. Recall that the score, being a function of the data as well as the parameter, is a random variable. Also, the estimator, being a function of the data, is a random variable. I pointed out that for our two examples, the score is positive when the "data" is higher than the parameter. To be more precise, if we use $\bar{x}$ as an estimator for μ, then the score is positive when $\bar{x} > \mu$, and if we use k / n as an estimator for π, then the score is positive when $k / n > \pi$. That implies that the score is positively correlated with these

[1] There are also versions of the Cramer-Rao lower bound for estimators that are not unbiased. I won't cover those here.

estimators. That's the best I can come up with as a motivation for the trick used in the proof.

The trick is, we look at the covariance between the score and the estimator. Consider any estimator of the parameter θ. It is a function of the data, and I will call it W(**x**). I will refer to the score as $g(\mathbf{x};\theta)$ in this derivation, to emphasize the fact that it is a function of the data. Recall that $\mathrm{E}(g(\mathbf{x};\theta)) = 0$.

We need some basic facts about random variables:

1) for any two random variables *X* and *Y*, if <u>either</u> random variable has 0 expectation, then $\mathrm{Cov}(X,Y) = \mathrm{E}(XY)$
2) the correlation between two random variables $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$ and $|\rho(X,Y)| \le 1$

Putting these together, if *X* or *Y* has 0 expectation,

$$\mathrm{Var}(X) = \frac{(\mathrm{Cov}(X,Y))^2}{(\rho(X,Y))^2\,\mathrm{Var}(Y)} \ge \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(Y)} = \frac{(\mathrm{E}(XY))^2}{\mathrm{Var}(Y)}$$

Letting *X* be *W*(**x**) and *Y* be $g(\mathbf{x};\theta)$, $\mathrm{Var}(W(\mathbf{x})) \ge \dfrac{\mathrm{E}(W(\mathbf{x})g(\mathbf{x};\theta))}{\mathrm{Var}(g(\mathbf{x};\theta))}$.

To evaluate the expectation in the numerator, we use a technique similar to the one used to show that the expectation of the score was 0. I will write this out using an integral, but a similar derivation can be done with discrete random variables using summations.

$$\mathrm{E}(W(\mathbf{x})g(\mathbf{x};\theta)) = \mathrm{E}\left(W(\mathbf{x})\frac{\partial}{\partial\theta}\log f(\mathbf{x}\mid\theta)\right) \qquad \text{(Using definition of } g\text{)}$$

$$= \mathrm{E}\left(\frac{W(\mathbf{x})}{f(\mathbf{x}\mid\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x}\mid\theta)\right) \qquad \text{(Derivative of a log)}$$

$$= \int\left(\frac{W(\mathbf{x})}{f(\mathbf{x}\mid\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x}\mid\theta)\right)f(\mathbf{x}\mid\theta)\,d\mathbf{x} \qquad \text{(Definition of expectation as integral)}$$

$$= \int\left(W(\mathbf{x})\frac{\partial}{\partial\theta}f(\mathbf{x}\mid\theta)\right)d\mathbf{x} \qquad \text{(Cancelling } f(\mathbf{x}\mid\theta)\text{)}$$

Now $W(\mathbf{x})$ is not (directly) a function of θ, so we can move it inside the derivative. Under "regularity conditions", we can also take the derivative outside the integral giving

$$\int\left(W(\mathbf{x})\frac{\partial}{\partial\theta}f(\mathbf{x}\mid\theta)\right)d\mathbf{x}=\frac{\partial}{\partial\theta}\int\left(W(\mathbf{x})f(\mathbf{x}\mid\theta)\right)d\mathbf{x}=\frac{\partial}{\partial\theta}\mathrm{E}\left(W(\mathbf{x})\right).$$

Since $W(\mathbf{x})$ is an unbiased estimator of θ, $\frac{\partial}{\partial\theta}\mathrm{E}\left(W(\mathbf{x})\right)=\frac{\partial}{\partial\theta}\theta=1.$

Putting it all together, $\mathrm{Var}\left(W(\mathbf{x})\right)\geq\frac{\mathrm{E}\left(W(\mathbf{x})g(\mathbf{x};\theta)\right)}{\mathrm{Var}\left(g(\mathbf{x};\theta)\right)}=\frac{1}{\mathrm{Var}\left(g(\mathbf{x};\theta)\right)}.$

<u>An Alternative Expression for the Fisher Information</u>

Under some additional regularity conditions, it turns out that

$$\mathrm{Var}\left(g(\mathbf{x};\theta)\right)=-\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\log\left(f(\mathbf{x};\theta)\right)\right).$$

This is a remarkable result. It says that the variance of the score is the <u>curvature</u> of the log-likelihood function. I will put the proof of this at the end. For now, I will demonstrate with our two examples.

Example - Normal Distribution

$$g(\mu;\mathbf{x})=\left(\frac{\partial}{\partial\mu}\log\left(f(\mathbf{x};\mu)\right)\right)=\frac{n}{\sigma^2}(\bar{x}-\mu)$$

The second derivative is $\left(\frac{\partial^2}{\partial\mu^2}\log\left(f(\mathbf{x};\mu)\right)\right)=-\frac{n}{\sigma^2}.$

In this case, the second derivative is <u>not</u> a function of the data, so its expectation is simply $-\frac{n}{\sigma^2}$, and $-\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\log\left(f(\mathbf{x};\theta)\right)\right)=\frac{n}{\sigma^2}.$

Example -Binomial Distribution

$$\frac{\partial^2}{\partial\pi^2}\log f(\pi;k)=-\frac{k}{\pi^2}-\frac{(n-k)}{(1-\pi)^2}$$

$$-\mathrm{E}\left(\frac{\partial^2}{\partial\pi^2}\log f\left(\pi;k\right)\right)=\mathrm{E}\left(\frac{k}{\pi^2}\right)+\mathrm{E}\left(\frac{\left(n-k\right)}{\left(1-\pi\right)^2}\right)$$

$$=\frac{n\pi}{\pi^2}+\frac{n\left(1-\pi\right)}{\left(1-\pi\right)^2}=n\left(\frac{1}{\pi}+\frac{1}{1-\pi}\right)=\frac{n}{\pi\left(1-\pi\right)}$$

This is the formula that was stated but not derived earlier.

The Observed Information

Notice that for the normal distribution, we didn't need to calculate an expectation when we used the second derivative formula for the Fisher information, although if we don't know σ^2, we will need to estimate it. Also notice that, although we calculated an expectation for the binomial distribution, it turns out that if we plugged in $k = n\,\pi$ in the second derivative, we would get the same formula. This suggests that we might calculate $\frac{n}{\hat{\sigma}^2}$ as an estimate for the information in the normal distribution, and $\frac{n}{\hat{p}\left(1-\hat{p}\right)}$ for the binomial. This calculation is called the observed information. The formula based on taking expectations is called the expected information. Given the observed information, we can calculate an estimated variance without taking an expectation and without knowing the true population parameter. For the normal, the estimated variance is $\frac{\hat{\sigma}^2}{n}$; for the binomial, the estimated variance is $\frac{\hat{p}\left(1-\hat{p}\right)}{n}$.

The Information Matrix

When there is more than one population parameter (for example, the β coefficients in multiple regression), there is a form of the Cramer-Rao bound that applies to all the parameters. In this case, we need to look at the variance-covariance matrix of all the estimates, and the Cramer-Rao bound is a matrix. There is a way in which the magnitude of matrices can be compared,[2] and the Cramer-Rao bound is the smallest possible variance-covariance matrix. In this case, the Fisher information is called the information matrix. As described in the previous paragraph, we can have both an expected information matrix and an observed information matrix.

Maximum Likelihood Estimation

In addition to its theoretical importance, the Cramer-Rao bound plays a role in maximum likelihood estimation. It turns out that the maximum likelihood estimator is consistent even in cases where it is not unbiased. It is also asymptotically efficient, meaning that as the sample size gets larger and larger, its variance is as low as possible, i.e. it meets the

[2] A matrix A is "greater than" a matrix B if A-B is a positive definite matrix, and is "greater than or equal to" B if A-B is non-negative definite.

Cramer-Rao lower bound. Although this is only an asymptotic result, in practice the variance of the maximum likelihood estimator is calculated by using the Cramer-Rao formula (the observed information matrix, in particular), so when standard errors are given in a maximum likelihood estimator, they are based on the Cramer-Rao formula.

Proof of the Alternative Expression for the Fisher Information

We need to prove: $\mathrm{Var}\left(g\left(\mathbf{x};\theta\right)\right) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\log\left(f\left(\mathbf{x};\theta\right)\right)\right)$

(To reduce the amount of notation, I will use *l* for the log-likelihood, *f* for the probability density, and *g* for the score, without writing in the dependence on **x** and θ. Looking at the second derivative of the log-likelihood:

$$\frac{\partial^2 l}{\partial\theta^2} = \frac{\partial}{\partial\theta}\frac{\partial l}{\partial\theta} = \frac{\partial}{\partial\theta}\left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right) = \left[\frac{\partial}{\partial\theta}\left(\frac{1}{f}\right)\right]\frac{\partial f}{\partial\theta} + \frac{1}{f}\left[\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial\theta}\right)\right] \qquad \text{(Product rule)}$$

$$= \left[\left(\frac{-1}{f^2}\right)\frac{\partial f}{\partial\theta}\right]\frac{\partial f}{\partial\theta} + \frac{1}{f}\left(\frac{\partial^2 f}{\partial\theta^2}\right) \qquad \text{(Derivative formulas)}$$

$$= -1\left[\left(\frac{1}{f}\right)\frac{\partial f}{\partial\theta}\right]^2 + \frac{1}{f}\left(\frac{\partial^2 f}{\partial\theta^2}\right) \qquad \text{(Algebra)}$$

$$= -g^2 + \frac{1}{f}\left(\frac{\partial^2 f}{\partial\theta^2}\right) \qquad \text{(Definition of score)}$$

Taking expectations,

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial\theta^2}\right) = \mathrm{E}\left(g^2\right) + \mathrm{E}\left[\frac{1}{f}\left(\frac{\partial^2 f}{\partial\theta^2}\right)\right]$$

We can prove the equivalence of alternative expression if we can prove that the last expectation is 0. Using the same trick as before

$$\mathrm{E}\left[\frac{1}{f}\left(\frac{\partial^2 f}{\partial\theta^2}\right)\right] = \int\left[\frac{1}{f}\left(\frac{\partial^2 f}{\partial\theta^2}\right)\right] f\, d\mathbf{x} = \int\left(\frac{\partial^2 f}{\partial\theta^2}\right) d\mathbf{x} = \frac{\partial^2}{\partial\theta^2}\int f\, d\mathbf{x} = \frac{\partial^2}{\partial\theta^2}(1) = 0$$