Frequentist Statistics

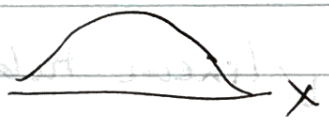
Estimation and Hypothesis testing:

Sampling distribution of an estimator

$\theta \rightarrow$ parameter $\quad \hat{\theta}(D) \rightarrow$ estimator

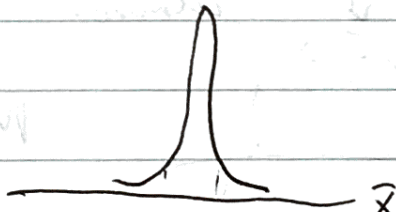
(D: data)

Example:



μ = mean, $\quad \bar{X} = \frac{X_1 + \cdots + X_N}{N}$

Parameter $\qquad$ Estimator.



$\bar{X}$ has its own distribution

In general, one can think of datsets $\mathcal{D}^{(s)}$ from some true model $p(\cdot|\theta^*)$. $\mathcal{D}^{(s)} = \{x_i^{(s)}\}_{i=1}^{N}$

Estimator $\{\hat{\theta}(\mathcal{D}^{(s)})\}_{s=1}^{S}$ with $S \rightarrow \infty$ gives us the sampling distribution

Some desirable properties of estimators

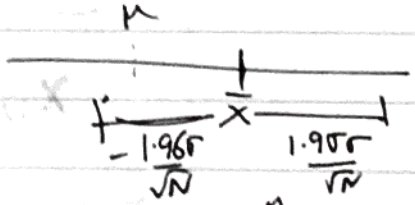
In practice, how do we know what the sampling distribution is?

Why is knowing the sampling distribution important?

Consider large sample limit of the distribution. If X is a random variable with finite mean μ and variance σ^2, $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{N}$.

$$P\left(\mu - \frac{1.96\sigma}{\sqrt{N}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{N}}\right) = 95\%.$$

We can rewrite it as



$$P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{N}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{N}}\right) = 95\%.$$

↑ confidence interval

That tells us something about μ. If we don't know σ we could replace it by a sample estimate too.

What if we do not have such large sample theory guarantees?

Bootstrap:

Parametric bootstrap

$x_i^{(s)} \sim p(\cdot \mid \theta^*)$

$\hat{\theta}^s = f(x_{1:N}^s)$ estimated parameter

Use $x_i^{(s)} \sim p(\cdot \mid \hat{\theta}(D))$ to generate $\hat{\theta}^{(s)}$ and its distrn.

Non-parametric bootstrap

$x_1, \ldots, x_N$

Generate random samples of size N with replacement.

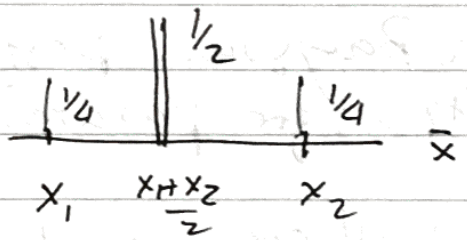
Each x_i could now be repeated

How could we get something almost out of nothing?

Imagine we have only two values

x_1, x_2

$\frac{1}{4}$ prob	x_1, x_1	$\rightarrow$	$\bar{x} = x_1$
$\frac{1}{2}$ prob	x_1, x_2	$\rightarrow$	$\bar{x} = \frac{x_1+x_2}{2}$
$\frac{1}{4}$ prob	x_2, x_2	$\rightarrow$	$\bar{x} = x_2$



Sampling distrn Var $\sim \frac{1}{4}\left(x_2 - \frac{x_1+x_2}{2}\right)^2 + \frac{1}{2}\,0 + \frac{1}{4}\left(x_1 - \frac{x_1+x_2}{2}\right)^2$

$= \frac{1}{2}\left(\frac{x_2-x_1}{2}\right)^2 = \frac{1}{8}(x_2-x_1)^2$

$S^2_{N-1} = \frac{\left(x_2 - \frac{x_1+x_2}{2}\right)^2 + \left(x_1 - \frac{x_1+x_2}{2}\right)^2}{2-1}$

$= \frac{1}{4}(x_1-x_2)^2$

Essentially we are sampling from the

empirical distribution

$$\hat{\theta}^s = \hat{\theta}(x^s_{1:N})$$

↑ Bootstrap samples

Contrast with Bayesian posterior probabilities for parameters.

As we will see local sampling of θ by MC is more efficient than having to find many samples (Sim#) and extract $\hat{\theta}$ for each sample.

Large sample theory for the Maximum Likelihood Estimator (MLE)

Log likelihood function

$\log P(D|\theta)$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log P(D|\theta)$$

Example $x_i \sim N(\mu, \sigma^2)$

$$P(D|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\sum_{i=1}^{N}(x_i-\mu)^2/2\sigma^2}$$

$$\log P = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

Vary μ

$$\sum_{i=1}^{N}(x_i-\hat{\mu}) = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}$$

Vary σ^2

$$-\frac{N}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x_i-\hat{\mu})^2}{N} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}$$

$$= S_N^2$$

Exponential distribution

$\frac{1}{b} e^{-\frac{x}{b}} = p(x)$

$P(D|b) = \frac{1}{b^N} e^{-\frac{\sum x_i}{b}}$

$\log P = -N \ln b - \frac{\sum_{i=1}^{N} x_i}{b}$

$\frac{\partial}{\partial b} \log P \Big|_{b=\hat{b}} = 0 \qquad -\frac{N}{\hat{b}} + \frac{\sum_{i=1}^{N} x_i}{\hat{b}^2} = 0$

$\hat{b} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}$

In general

$S(\theta, D) = \nabla_\theta \log P(D|\theta)$

Score function

For MLE $S(\hat{\theta}(D), D) = 0$

In observed information matrix

$J(\hat{\theta}(D)) = -\nabla S(\theta)|_{\theta=\hat{\theta}}$

$= -\nabla_\theta \nabla_\theta \log P(D|\theta)$

Hessian of log-likelihood.

The J matrix tell us how peaked things are in the theta space.

Let us go back to easy example of normal distribution and mean. Assume σ is fixed.

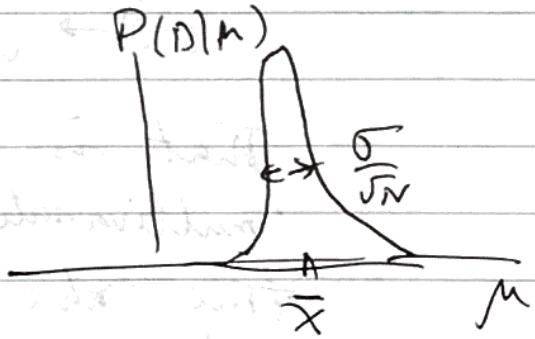
$$\log P(D|\mu) \sim \frac{\sum (x_i - \mu)^2}{2\sigma^2}$$

$$S(\mu) = \frac{N}{\sigma^2}(\bar{x} - \mu)$$

$$J(\mu) = -\frac{d}{d\mu} S(\mu) = \frac{N}{\sigma^2}$$

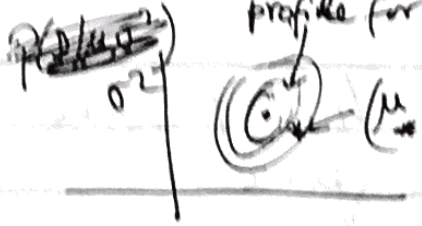
$$P(D|\mu) \sim e^{-\frac{\sum (x_i - \mu)^2}{2\sigma^2}} \sim e^{-\frac{\sum (x_i - \bar{x})^2}{2\sigma^2}} e^{-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}}$$

$$\propto e^{-\frac{N(\mu - \bar{x})^2}{2\sigma^2}}$$



Things are a bit more complicated if we looke at the joint variation μ, σ^2

Likelihood profile for one sample $(x_1, \ldots, x_N)$



Average of different samples for true parameters of the observed information matrix $\rightarrow$ Fisher information

$$I_N(\hat{\theta}|\theta^*) \triangleq E_{\theta^*}[J(\hat{\theta}, D)]$$

We are still keeping $\hat{\theta}$ and D independent, denoting some estimator function of the sample x's.
Sometimes written as $I_N(\theta)$.

Since $\hat{\theta}$ is built from additive contribution from independent samples after averaging $I_N(\hat{\theta}) = N I_1(\hat{\theta}) \triangleq N I(\hat{\theta})$

$J = -\nabla_\theta \nabla_\theta \log p(x_i|\theta)|_{\theta=\hat{\theta}}$, $E_{\theta^*}(J) = -N \int dx \, \nabla_\theta \nabla_\theta \log p(x|\theta)|_{\theta} \, p(x|\theta^*)$

For $\hat{\theta}(D) = \hat{\theta}_{MLE}(D)$

$$\hat{\theta}(D) \rightarrow \mathcal{N}(\theta^*, I_N(\theta^*)^{-1}) = \mathcal{N}\left(\theta^*, \frac{I(\theta^*)^{-1}}{N}\right)$$

That is $\hat{\theta}(D)$ is asymptotically (multivariate) normal with mean at true value and the covariance matrix given by inverse of Fisher information

Note that that the covariance matrix $\rightarrow 0$ as $N \rightarrow \infty$

Let us see how that comes out to be. Do it for one parameter first, with an example, the exponential distribution

$$\log P(D|b) = \log \frac{1}{b^N} e^{-\frac{\sum x_i}{b}} = -N \log b - \frac{\sum x_i}{b}$$

$$\frac{\partial \log P}{\partial b} = S(b, D) = -\frac{N}{b} + \frac{\sum x_i}{b^2} = \sum_{i=1}^{N} \frac{x_i - b}{b^2}$$

↑ Sum of N independent random variables

$$J = -\frac{\partial}{\partial b} S = -\frac{N}{b^2} + \frac{2\sum x_i}{b^3}$$

$$= \sum_i \frac{2x_i - b}{b^3}$$

← Sum of N independent random variables.

If x_i are generated from the distribution $\frac{1}{b^*} e^{-x_i/b^*}$, $x_i \geq 0$ $\quad E_{b^*}[X_i] = b^*$

$$\text{So } I_N(b|b^*) = \sum_i \frac{2E_{b^*}[X_i] - b}{b^3} = \frac{N(2b^* - b)}{b^3}$$

$I(b^*) = \frac{1}{b^{*2}}$

Now let us look at $\hat{b}(D)$ distribution when $\hat{b}(D)$ is the MLE.

$\hat{b}$ is decided by the equation

$$\partial_b \sum_i \log p(x_i|b) = 0 \quad \text{or} \quad \sum_i \frac{x_i - b}{b^2} = 0$$

Of course, we could write $\hat{b} = \frac{1}{n}\sum x_i$ and analyze it from there. However, here is general procedure

$$\sum_i \partial_b \log p(x_i|b) = 0 \longrightarrow \text{Taylor expand:}$$

$$0 \approx \sum_i \partial_b \log p(x_i|b^*)\Big|_{b^*} + (b-b^*)\left[\sum_i \partial_b^2 \log p(x_i|b)\Big|_{b^*}\right] + o((b-b^*)^2)$$

In particular,

$$0 = \sum_i \frac{x_i - b^*}{b^{*2}} + (b-b^*)\left[-\sum_{i=1}^{N} \frac{(2x_i - b^*)}{b^{*3}}\right] + o((b-b^*)^2)$$

Consider the random variable
$S(b,D) = \sum \partial_b \log p(x_i|b)$. It is a sum of indept variables. We need to know its mean and variance.
What is its expectation?

$$E_{b^*}[S(b,D)] = N \partial_b \int \log(p(x|b))\, p(x|b^*)\, dx$$
$$= N \int \partial_b p(x|b)\, \frac{p(x|b^*)}{p(x|b)}\, dx$$

Note that $E_{b^*}[S(b^*,D)]$
$$= N \int \partial_b p(x|b)\Big|_{b^*} dx$$

But $\int p(x|b)\, dx = 1$ for every b,
so that $\int \partial_b p(x|b)\, dx = 0$

So $E_{b^*}[S(b^*,\theta)] = 0$

What is $\mathrm{Var}_{b^*}[S(b^*,\theta)]$

$$\text{Var}_{b^*}[S(b^*, D)] = \left[\int \left(\partial_b \log p(x|b) \right)^2 p(x|b)\, dx \right]_{b=b^*}$$

$$= \int \frac{\left(\partial_b p(x|b) \right)^2}{p(x|b)} dx \Bigg|_{b=b^*}$$

Remember for two distributions p_i and $p_i + \delta p_i$ KL divergence

$$= \frac{1}{2} \sum_i \frac{\delta p_i^2}{p_i}$$

Also

$$- \int \partial_b^2 \left[\log p(x|b) \right] p(x|b)\, dx$$

$$= - \int \partial_b \left(\frac{\partial_b p(x|b)}{p(x|b)} \right) p(x|b)\, dx$$

$$= - \int \left[\frac{\partial_b^2 p(x|b)}{p(x|b)} - \frac{\partial_b p(x|b)}{p(x|b)^2} \right] p(x|b)\, dx$$

$$= -\partial_b^2 \int p(x|b)\,dx + \int \frac{(\partial_b p(x|b))^2}{p(x|b)}\,dx$$

zero ↙

First derive square ↓ 2 ↓ av

$$\mathrm{Var}_{b^*}[S(b^*, D)] = N \int (\partial_b \log p(x|b))\, p(x|b)\,dx \Big|_{b^*}$$

$$= -N \int \partial_b^2 [\log p(x|b)]\, p(x|b)\,dx \Big|_{b^*}$$

↑ Second deriv ~~version~~ av.

Any way, going back to the Taylor expanded equation.

$$0 = S(b^*, D) + (b - b^*) \left[\sum_i \partial_b^2 \log(x_i|b) \right]_{b^*} + \cdots$$

Sum of many variables. Usually nonzero average. Replace by

$$N \int dx\, p(x|b^*) \left[\partial_b^2 \log p(x|b)\right]_{b=b^*}$$

That is just $N I_1(b^*)$

So $S(b^*, D) = N(b - b^*) I_1(b^*) + \cdots$

$$\hat{b}(D) - b^* = \frac{1}{N} I_1(b^*)^{-1} S(b^*, D)$$

$S(b^*, D)$: mean zero

Variance $\rightarrow \frac{1}{N^2} I_1(b^*)^{-2} N I_1(b^*)$

$= I_1(b^*)^{-1} / N$

Thus $\hat{b}(D) \sim \mathcal{N}\left(b^*, \frac{I_1(b^*)^{-1}}{N}\right)$

In detail

$$\sum_i \frac{x_i - b^*}{b^{*2}} \approx \sum_i \frac{2x_i - b}{b^{*3}} (\hat{b} - b^*)$$

$$\simeq \frac{N}{b^{*2}} (\hat{b} - b^*)$$

$$E\left[\frac{X_i - b^*}{b^{*2}}\right] = 0$$

$$\mathrm{Var}\left[\sum_i \frac{X_i - b^*}{b^{*2}}\right] = N \mathrm{Var}\left[\frac{X - b^*}{b^{*2}}\right]$$

$$= N \frac{\mathrm{Var}[X]}{b^{*4}} = N \frac{b^{*2}}{b^{*4}} = \frac{N}{b^{*2}}$$

$$\hat{b} = b^* + \frac{b^{*2}}{N} \sum \frac{X_i - b^*}{b^{*2}}$$

$$= b^* + \frac{1}{N} \sum_i (X_i - b^*)$$

mean zero

variance $\frac{b^{*2}}{N}$

$$\hat{b} \sim \mathcal{N}\left(b^*, \frac{b^{*2}}{N}\right)$$

In the multiparameter case

$$S(\theta, D) = \sum_i \nabla_\theta \log p(x_i|\theta)$$

Sum of independent random vectors.

$$E_{\theta^*}[S(\theta^*, D)] = 0$$

$$\text{Cov}_{\theta^*}[S(\theta^*, D)] \quad \cancel{N \int \nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)}$$

$$= N \int dx\, p(x|\theta^*) \cdot \left[\nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)\right]_{\theta=\theta}$$

$$= -N \int dx\, p(x|\theta^*) \left[\nabla_\theta \nabla_\theta \log p(x|\theta)\right]_{\theta=\theta}$$

$$= N I_{\theta}(\theta^*)$$

Taylor expanded equation

$$S(\theta, D) = 0$$

$$\Rightarrow S(\theta^*, D) + \underbrace{\sum_i \nabla_\theta \nabla_\theta \log p(x_i|\theta)\Big|_{\theta=\theta^*}}_{\text{Replace by } -N I_\theta(\theta^*)} (\theta - \theta^*) = 0$$

So $\underbrace{\theta - \theta^*}_{\text{vector}} = \frac{1}{N} \underbrace{I(\theta^*)^{-1} S(\theta^*; D)}_{\text{Mean zero}}$

$$\text{Cov}_{\theta^*} = \frac{1}{N} I^{-1}(\theta^*) I(\theta^*) I^{-1}(\theta^*)$$

$$= \frac{1}{N} I(\theta^*)^{-1}$$

means $\theta - \hat{\theta} \sim \frac{1}{\sqrt{N}}$.

We will also see that the log likelihood at the MLE ~~estimate~~ point goes as ~~log P(D|θ)~~ a constant ~~plus~~ minus half a χ^2 variable.

To see that expand in Taylor Series log likelihood directly around $\theta = \theta^*$

$$\log P(D|\theta) = \log P(D|\theta^*) + \left[\nabla_\theta \log P(D|\theta)\right]_{\theta=\theta^*}^T (\theta - \theta^*)$$

$$+ \frac{1}{2} (\theta - \theta^*)^T \left[\nabla_\theta \nabla_\theta \log P(D|\theta)\Big|_{\theta=\theta^*}\right] (\theta - \theta^*)$$

$$+ \cdots$$

$$\log P(D|\theta)$$

$$\approx \log P(D|\theta^*) + S(\theta^*|D)^T (\theta - \theta^*) - \frac{1}{2} N (\theta - \theta^*)^T I(\theta^*)(\theta - \theta^*)$$

If $\theta = \hat{\theta}$, the max. likelihood estimator, $S = N I^{-1} I^{-1} (\hat{\theta} - \theta^*)$

$$\log P(D|\theta) = \log P(D|\theta^*) + \frac{N}{2} (\hat{\theta} - \theta^*)^T I(\theta^*)(\hat{\theta} - \theta^*)$$

So $2 \log \frac{P(D|\theta)}{P(D|\theta^*)} = N (\hat{\theta} - \theta^*)^T I(\theta^*)(\hat{\theta} - \theta^*)$

This will important ~~for testing~~ later.

Note that $P(\hat{\theta}) \sim e^{-\frac{1}{2} N (\hat{\theta}-\theta^*)^T I(\theta^*)(\hat{\theta}-\theta^*)}$

If I choose coordinate in the θ space, s.t.

$\theta^* \to 0$ & $y_\alpha \sim O(\theta - \theta^*)$

↑ orthogonal matrix

so that $O I O^{-1} = A = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_k \end{pmatrix}$ $a_i \geq 0$

$$P(\hat{\theta}) \sim e^{-\frac{1}{2} N y^T A y}$$

$$= e^{-\frac{N}{2}\sum_\alpha a_\alpha y_\alpha^2}$$

Define $z_\alpha = \sqrt{N a_\alpha}\, y_\alpha$

$$P \sim e^{-\frac{1}{2}\sum_\alpha z_\alpha^2}$$

$N(\hat{\theta}-\theta^*)^T I (\hat{\theta}-\theta^*) = \sum_\alpha z_\alpha^2$

This is a sum of k independent standard normal variables.

Why does max likelihood choose the right distribution asymptotically?

Consider a discrete distribution

$$P(D|\theta) = \prod_a p_a(\theta)^{n_a}$$

↑ events, not samples

$$\sum n_a = N$$

$$\log P(D|\theta) = N \sum_a \frac{n_a}{N} \log p_a(\theta)$$

If events are from $p_a(\theta^*)$ and we have a lot of data

$$\approx N \sum_a p_a(\theta^*) \log p_a(\theta)$$

$$= -N\, H(p_a(\theta^*), p_a(\theta))$$

↑ cross entropy

$$KL(P\|Q) = \sum P \log \frac{P}{Q}$$

$$= H(P,Q) - H(P)$$

$$-\frac{1}{N} \log P(D|\theta) - H(P(\theta^*))$$

$$\approx KL(P(\theta) \| P(\theta^*))$$

$$\delta\theta^T I(\theta)\, \delta\theta = \int dx \frac{1}{P(x,\theta)} \delta\theta^T \nabla_\theta \log P(x) \nabla_\theta \log P(x) \cdot \delta\theta$$

$$= \int \frac{\delta P^2(x,\theta)}{P(x,\theta)} dx$$

Explicit Example: Normal Distribution

$$\log P(\mathcal{D}|\mu,\sigma^2) = \log\left[\frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

I will treat σ^2 as a variable, say θ

$$\log P = -\frac{N}{2}\log(2\pi\theta) - \frac{1}{2\theta}\sum_i (x_i-\mu)^2$$

$$S(\mu,\theta) = \left(\frac{\partial}{\partial\mu}\log P, \frac{\partial}{\partial\theta}\log P\right)$$

$$= \left(\frac{1}{\theta}\sum_{i=1}^{N}(x_i-\mu), -\frac{N}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$

$$= (S_\mu, S_\theta)$$

$$J(\mu,\theta) = -\begin{pmatrix} \frac{\partial S_\mu}{\partial\mu} & \frac{\partial S_\mu}{\partial\theta} \\ \frac{\partial S_\theta}{\partial\mu} & \frac{\partial S_\theta}{\partial\theta} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{N}{\theta} & \frac{1}{\theta^2}\sum_{i=1}^{N}(x_i-\mu) \\ \frac{1}{\theta^2}\sum_{i=1}^{N}(x_i-\mu) & -\frac{N}{2\theta^2} + \frac{1}{\theta^3}\sum_{i=1}^{N}(x_i-\mu)^2 \end{pmatrix}$$

$$\underset{\mu^*,\theta^*}{E}[J(\mu,\theta)] = N\begin{pmatrix} \frac{1}{\theta} & \frac{\mu^*-\mu}{\theta^2} \\ \frac{\mu^*-\mu}{\theta^2} & \frac{2\theta^*\blacksquare-\theta+2(\mu^*-\mu)^2}{2\theta^3} \end{pmatrix}$$

Since $\sum_{i=1}^{N}(X_i-\mu)^2 = \sum_i (X_i-\mu^*)^2 + 2(\mu^*-\mu)\sum_i (X_i-\mu^*) + N(\mu^*-\mu)^2$

At $\mu=\mu^*, \theta=\theta^*$ $\quad J(\mu^*,\theta^*) = NI(\mu^*,\theta^*)$

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

[redacted] the max likelihood estimators,

$S_\mu = 0 \Rightarrow \frac{1}{n}\sum x_i = \hat{\mu}$

$S_\theta = 0 \Rightarrow \frac{1}{n}\sum_i (x_i-\hat{\mu})^2 = \hat{\theta} = \hat{\sigma}^2$

$\frac{1}{N}\sum x_i = \bar{x}$ [redacted] has variance $\frac{\sigma^2}{N}$

$\frac{1}{N}\sum_i (x_i-\bar{x})^2$ can be rewritten as $\frac{1}{n}\sum_{j=1}^{N-1} y_j^2$

where $y_j \overset{i.i.d}{\sim} N(0,\sigma^2)$.

$$\text{Var}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right]$$

$$=\text{Var}\left[\frac{1}{N}\sum_{j=1}^{N-1}Y_j^2\right]$$

$$=\sum_{j=1}^{N-1}\frac{1}{N^2}\,\text{Var}[Y_j^2]$$

$$=\sum_{j=1}^{N-1}\frac{1}{N^2}\left(E[Y_j^4]-E[Y_j^2]^2\right)$$

$$=\frac{N-1}{N^2}(3\sigma^4-\sigma^4)$$

$$=\frac{2(N-1)\sigma^4}{N^2}$$

Asymptotically, this goes as

$$\frac{2\sigma^4}{N}$$

It turns out that $\bar{x}$ and $\sum_i (x_i-\bar{x})^2$ are independent

So the covariance matrix of (μ,σ^2) is,

for large N, given by

$$\begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{pmatrix}$$

This is just the inverse of the Fisher information matrix upto the factor N.

$$\begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{pmatrix}$$