This time, multi-parameter problems

next time:

Last time we worked out that the likelihood function in this model is

\[ L(\theta | y, B) = L(\mu, \sigma | y, B) \]

\[ = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \]

Hence, the log-likelihood function is therefore

\[ \ell(\theta | y, B) = \ell(\mu, \sigma | y, B) \]

\[ = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \]

Expand out sum:

\[ \sum_{i=1}^{n} (y_i - \mu)^2 = \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + \mu^2 \]

So immediately \((\bar{y}, \frac{1}{n} \sum_{i=1}^{n} y_i^2)\) is sufficient.
But a more familiar set of sufficient statistics can be found with a small trick:

\[
\sum (x_i - \mu)^2 = \sum \left[ (x_i - \bar{y}) + (\bar{y} - \mu) \right]^2 \\
= \sum (x_i - \bar{y})^2 + 2(\bar{y} - \mu) \sum (x_i - \bar{y}) \\
+ n(\bar{y} - \mu)^2
\]

So \( \sum (x_i - \bar{y}) \) is also sufficient for \( \theta = (\mu, \sigma) \) by the fact:

In repeated sampling from this sampling model, \( E_R \left[ \frac{1}{n-1} \sum (x_i - \bar{x})^2 \right] = \sigma^2 \)

Definition: If an estimator \( \hat{\theta}_n \) of a (1-dimensional) parameter \( \theta \) is such that

\( E_R (\hat{\theta}_n) = \theta \), \( \hat{\theta}_n \) is said to be an unbiased estimator of \( \theta \)
Restatement of math fact: $\bar{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{X}_i - \bar{X})^2$ is unbiased for $\sigma^2$ in the $\mathcal{N}(\mu, \sigma^2)$ sampling model.

"Official" set of sufficient statistics in this model: $(\bar{y}, \hat{S}_n)$.

Using track:

$$f(\mu, \sigma | \bar{y}, n) \sim \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (n-1)\hat{S}_n^2 + n(\bar{y} - \mu)^2 \right] \right\}$$

$$\mathcal{L}(\mu, \sigma | \bar{y}, n) \sim -n \log \sigma - \frac{(n-1)\hat{S}_n^2}{2\sigma^2} - \frac{(\bar{y} - \mu)^2}{2\sigma^2}$$

Let's set the MLEs:

$$\frac{\partial \mathcal{L}(\mu, \sigma | \bar{y}, n)}{\partial \mu} = n(\bar{y} - \mu) \frac{\sigma^2}{\sigma^2} = 0 \quad \text{(seems ok)}$$

Set $\sigma^2 = 0$, so solve for $\mu$:

$$\mu = \bar{y}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma | \bar{y}, n)}{\partial \sigma^2} = -\frac{n}{\sigma} - \frac{(\bar{y} - \mu)^2}{2\sigma^2}$$
The MLE \( \hat{\theta}_{\text{MLE}} = (\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}}) \) is the solution of 2 equations in 2 unknowns:

\[
\begin{align*}
\frac{\partial \mathcal{L}(\mu, \sigma^2 | Y_B)}{\partial \mu} &= 0 \\
\frac{\partial \mathcal{L}(\mu, \sigma^2 | Y_B)}{\partial \sigma^2} &= 0
\end{align*}
\]

It happens in this problem that the first equation, for \( \hat{\mu}_{\text{MLE}} \), can be solved without reference to the second equation; this will in general not be true.

\[
\frac{\partial \mathcal{L}(\mu, \sigma^2 | Y_B)}{\partial \sigma^2} = 0 \quad \Rightarrow \quad -n \sigma^2 + (n-1) s^2 + n(\bar{Y}_B)^2 = 0
\]

If \( \hat{\sigma}^2_{\text{MLE}} = \frac{n-1}{n} s^2 + \left( \frac{\hat{\mu}_{\text{MLE}}}{\hat{\mu}_{\text{MLE}}} \right)^2 \), then \( \hat{\sigma}^2_{\text{MLE}} = \frac{
}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \).
Important property of maximum likelihood

Suppose we parameterize this model in terms of

\[ (\mu, \sigma^2) \leftrightarrow (\mu, \eta) \]

instead of \((\mu, \sigma)\)

\[ \theta' = \theta \]

\[ \theta = \theta \]

\[ \mathbb{E}(\theta' | y, B) = \mathbb{E}(\mu, \eta | y, B) = \frac{1}{\eta} \sum_{i=1}^{n} \left( x_i - \frac{1}{\eta} \right)^2 \]

\[ \mathbb{E}(\sigma^2 | y, B) = \text{same steps for } \eta \]

\[ \sigma^2 \sum_{i=1}^{n} \left( x_i - \mu \right)^2 \]

\[ \mathbb{E}(\eta | y, B) = \frac{1}{\eta} \sum_{i=1}^{n} \left( x_i - \frac{1}{\eta} \right)^2 \]

\[ \sigma^2 \sum_{i=1}^{n} \left( x_i - \mu \right)^2 \]
\[ \text{MLE}(\theta', \gamma ') = -\frac{1}{2} \log 2 - \frac{1}{2} \sum_{i=1}^{n} (\gamma_i - \mu)^2 \]

\[ \frac{d}{d\mu} \text{MLE}(\theta', \gamma ') = -\frac{1}{2} \sum_{i=1}^{n} (\gamma_i - \mu) (-1) \]

*0: if \( \sum_{i=1}^{n} \gamma_i = n \mu \) and \( \mu_{\text{MLE}} = \bar{Y} \) define

\[ \frac{d}{d\mu} \text{MLE}(\theta', \gamma ') = -\frac{1}{2} \sum_{i=1}^{n} (\gamma_i - \mu) \]

Together with above

\[ -\eta \mu + \frac{1}{2} \sum_{i=1}^{n} (\gamma_i - \mu)^2 \]

iff \( \eta = \eta_{\text{MLE}} = \frac{1}{2} \sum_{i=1}^{n} (\gamma_i - \mu_{\text{MLE}})^2 \]

So:

\[ \hat{\sigma}_{\text{MLE}}^2 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\gamma_i - \bar{Y})^2} = \sqrt{\eta_{\text{MLE}}} \]

\[ \eta = \sigma^2 + \frac{\eta_{\text{MLE}}}{\sigma^2} g(\eta) + \hat{\sigma}_{\text{MLE}}^2 g(\hat{\eta}_{\text{MLE}}) \]
Ex. If \( f(-2) = \frac{1}{3} \) and \( f(2) = \frac{2}{3} \), then \( f(0) \) must be 1.

If \( f(x) \) is defined as \( f(x) = x^2 \), then the value of \( f(-2) \) is \( 4 \).

In the N(\mu, \sigma^2) sampling model, if \( \mu = 0 \) and \( \sigma^2 = 1 \), then this model is not invertible for \( \mu \neq 0 \).

If \( f(x) \) is an invertible function, and \( g(y) = \mu \), then invertible for \( \phi \) is a fixed parameter.
What about \( \sigma^2 \)?

If your sample model is

\[
\begin{align*}
(x_i | \mu, \sigma^2) & \sim N(\mu, \sigma^2) \\
(\mu | \sigma^2) & \sim \text{known}
\end{align*}
\]

s. that \( \Theta = \mu \) (1-dimensional),

then the Fisher information is

\[
\hat{I}(\hat{\mu}_{\text{MLE}}) = \left[ -\frac{\partial^2 \log p(x | \Theta)}{\partial \mu^2} \right]_{\mu = \hat{\mu}_{\text{MLE}}}.
\]

When \( \Theta \) has dimension greater than 1, the analogue of this is a matrix of second partial derivatives of \( \log p(x | \Theta) \).
**Definition** θ dimension p ≥ 1: the Hessian of \( p(\theta | y, B) \) is the \((p \times p)\) matrix of second partial derivatives:

\[
\begin{bmatrix}
\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\theta | y, B) & \ldots & \frac{\partial^2}{\partial \theta_i \partial \theta_p} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2}{\partial \theta_p \partial \theta_i} & \ldots & \frac{\partial^2}{\partial \theta_p \partial \theta_p}
\end{bmatrix}
\]

The \((i, j)\) element of \( H \) is

\[
\left(H_{ij}\right) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\theta | y, B)
\]

**Definition** Fisher information matrix \( I(\hat{\theta}_{\text{MLE}}) \) is

\[
I(\hat{\theta}_{\text{MLE}}) = [-H]
\]
In the $N(\mu, \sigma^2)$ sampling model, $\theta = (\mu, \sigma)$

we already worked out that

\[
E(\theta | y) = -n \log \sigma - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2
\]

\[
\frac{\partial}{\partial \mu} E(\theta | y) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)
\]

\[
\frac{\partial^2}{\partial \mu^2} E(\theta | y) = \frac{n}{\sigma^4}
\]

\[
\frac{\partial}{\partial \sigma} E(\theta | y) = -\frac{n}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)
\]

\[
\frac{\partial^2}{\partial \sigma^2} E(\theta | y) = \frac{1}{\sigma^2} - \frac{n}{\sigma^4} 
\]

\[
\frac{\partial}{\partial \sigma} E(\theta | y) = -\frac{1}{\sigma} \sum_{i=1}^{n} (y_i - \mu)
\]

\[
\frac{\partial^2}{\partial \sigma^2} E(\theta | y) = \frac{n}{\sigma^4} - \frac{3 n}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2
\]
As usual the mixed partial derivatives \( \frac{\partial^2}{\partial \mu_1 \partial \mu_2} \) and \( \frac{\partial^2}{\partial \mu_2 \partial \mu_1} \) are equal, so you only need to compute one of them and you might as well compute the easier one:

\[
\frac{\partial}{\partial \mu} \mathcal{L}(\theta | y, \mathbf{B}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \sum_{i=1}^{m} (\gamma_i - \mu)
\]

So

\[
\frac{\partial}{\partial \mu} \mathcal{L}(\theta | y, \mathbf{B}) = -2 \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \sum_{i=1}^{m} (\gamma_i - \mu)
\]

So

\[
\mathbf{H} = \begin{bmatrix}
\frac{-n}{\sigma^2} & -2 \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \\
-2 \frac{1}{\sigma^2} \sum_{i=1}^{m} (\gamma_i - \mu) & \frac{n \sigma^2 - 3 \sum_{i=1}^{n} (y_i - \mu) \sum_{i=1}^{m} (\gamma_i - \mu)}{\sigma^4}
\end{bmatrix}
\]
and $\hat{I} (\hat{\theta}_{\text{MLE}}) = \begin{bmatrix} -H \end{bmatrix} \theta = \hat{\theta}_{\text{MLE}}$

\[
\begin{bmatrix}
\frac{\mu}{\sigma^2} & \frac{2 \sum (x_i - \mu)}{\sigma^3} \\
\frac{2 \sum (x_i - \mu) - n \sigma^2 + 3 \sum (x_i - \mu)^2}{\sigma^4}
\end{bmatrix}
\]

For $\theta = \bar{x}$,

\[\sigma = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}\]

\[
\begin{bmatrix}
\frac{\mu}{\sigma_{\text{MLE}}^2} & 0 \\
0 & \frac{2 n \sigma_{\text{MLE}}^2}{\sigma_{\text{MLE}}^2}
\end{bmatrix}
\]

\[\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2\]

So

\[n \sigma_{\text{MLE}}^2 = \sum (x_i - \bar{x})^2\]
If $X_1, \ldots, X_n$ form an IID random sample from a population density such that $E(X_i^4)$ exists, then

$$V_{E_{X}}(\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Cautionary note:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$$ is unbiased for $\sigma^2$.

Crypt: If $S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2}$ unbiased for $\sigma^2$, A: No; if $\theta_n$ is unbiased for $\theta$, $g(\theta_n)$ is not unbiased for $g(\theta)$ unless $g(\cdot)$ is linear.
\[
\begin{bmatrix}
1 \\
2 \\
9
\end{bmatrix}
\cdot \begin{bmatrix}
7, -\bar{y} \\
7, -\bar{y} \\
7, -\bar{y}
\end{bmatrix}
\cdot \frac{1}{n} \cdot \begin{bmatrix}
-3 \\
-2 \\
x + 5
\end{bmatrix}
\rightarrow
\begin{bmatrix}
(-3)^2 \\
(-2)^2 \\
(x + 5)^2
\end{bmatrix}
\]

mean \( y = \bar{y} \)

\[
\sqrt{X^2} \begin{cases}
\text{free} & n = 3 \\
\text{not free} & n \neq 3
\end{cases}
\]

mean \( \bar{y} = 4 \)

\[\text{Fisher: degrees of freedom (df)}\]

\[\text{A dataset } \bar{y} = (7_1, \ldots, 7_n) \text{ only has (n-1) df for measuring spread (scale), if center (mean) must also be estimated}\]
Def. \( \text{Bias}(\theta) = E[\hat{\theta}(x)] - \theta \)

\( r = 1 \)

\( x_0 = 0 \)

\( v = 0 \)

\( v = 0 \)

\( r = 0 \)

\( r = 0 \)

\( v = 0 \)

\( r = 0 \)

\( r = 0 \)

Fact:

\( \theta \) is about \( \mu \) in

\( \theta \) is necessary

about \( \mu \)
Fact: To get $\sqrt{\text{SE}(\hat{\theta}_i)_{\text{MLE}}}$, invert the Fisher information matrix $\mathcal{I}(\hat{\theta}_{\text{MLE}})$ and take the square root of diagonal elements $i$ in $\mathcal{I}^{-1}$ (e.g., $N(\mu, \sigma^2)$ model).

$$\mathcal{I}^{-1}(\hat{\theta}_{\text{MLE}}) = \begin{pmatrix} \frac{\hat{\sigma}^2}{\sigma_{\text{MLE}}} & 0 \\ 0 & \frac{\hat{\sigma}^2}{2\sigma_{\text{MLE}}} \end{pmatrix}$$

So $\text{SE}(\hat{\mu}_{\text{MLE}}) = \sqrt{\frac{\hat{\sigma}^2}{\sigma_{\text{MLE}}} / n}$, $\text{SE}(\hat{\sigma}) = \frac{\hat{\sigma}^2}{\sigma_{\text{MLE}}} / 2n$

Similar tovariance inference problem: ex. $(Y_i | \theta) \overset{iid}{\sim} \text{uniform}(0, \theta)$
Finally, we consider the components of $\hat{\sigma}^2_{\text{MLE}}$ when $n$ is large, and the components of $\hat{\sigma}^2_{\text{MLE}}$ follow normal distributions:

$\begin{align*}
(\hat{\theta}_i)_{\text{MLE}} & \sim N(\theta_i, I'(\hat{\theta}^2_{\text{MLE}})) \\
\text{var} & \sim 1.96 \frac{\hat{\sigma}^2_{\text{MLE}}}{\sqrt{n}} \quad \text{(familiar CI)}
\end{align*}$

Thus, to get an approximate (large-sample) 100(1 - 2\alpha/2)% CI for $\mu$,

$\mu_{\text{MLE}} \pm \hat{\theta}_i(1 - \frac{\alpha}{2}) \text{SE}(\hat{\mu}_{\text{MLE}})$
ex. 95\% CI for $\sigma$:

large-sample approx.

$$\hat{\sigma}_{MLE} \pm 1.96 \frac{\hat{\sigma}_{MLE}}{\sqrt{n}}$$

Not so good when $n$ is small:
correct interval for $\sigma$ should be skewed.