MA 413 - Statistics for Data Science

Solutions to Exercise 10

1. (a) The loglikelihood is

$$\begin{aligned}\ell(\lambda) &= \log\left(\prod_{i=1}^n f_{Y_i}(Y_i;\lambda)\right) = \sum_{i=1}^n \log f_{Y_i}(Y_i;\lambda) = \sum_{i=1}^n \log\left(\frac{\lambda^3}{\Gamma(3)} Y_i^2 e^{-\lambda Y_i}\right)\\ &= \sum_{i=1}^n [3\log\lambda - \lambda Y_i + 2\log Y_i - \log\Gamma(3)]\\ &= 3n\log\lambda - \lambda\sum_{i=1}^n Y_i + 2\sum_{i=1}^n \log Y_i - n\log\Gamma(3)\end{aligned}$$

Setting $\ell'(\lambda) = 0$ we get

$$\frac{3n}{\lambda} - \sum_{i=1}^n Y_i = 0$$

and the unique root is

$$\hat{\lambda} = \frac{3n}{\sum_{i=1}^n Y_i}$$

Note that

$$\ell''(\lambda) = -\frac{3n}{\lambda^2} < 0$$

Therefore, $\hat{\lambda}$ is the MLE of λ.

(b) Note that $Y_i \sim \text{Gamma}(3,\lambda)$ implies that $S_n := \sum_{i=1}^n Y_i \sim \text{Gamma}(3n,\lambda)$. Therefore, $E\left(S_n^{-1}\right)$ can be calculated as

$$E\left(S_n^{-1}\right) = \int_0^\infty s^{-1}\frac{\lambda^{3n}}{\Gamma(3n)} s^{3n-1} e^{-\lambda s} ds$$

$= \int_0^\infty \frac{\lambda}{3n-1}\frac{\lambda^{3n-1}}{\Gamma(3n-1)} s^{(3n-1)-1} e^{-\lambda s} ds$ (we have used $\Gamma(x+1) = x\Gamma(x)$)

$$\begin{aligned}&= \tfrac{\lambda}{3n-1}\int_0^\infty \tfrac{\lambda^{3n-1}}{\Gamma(3n-1)} s^{(3n-1)-1} e^{-\lambda s} ds \quad (*)\\ &= \tfrac{\lambda}{3n-1}\end{aligned}$$

where the integrand in $(*)$ is the density function of Gamma $(3n-1,\lambda)$, and thus integrating it over the whole domain yields 1. The expectation of $\hat{\lambda}$ is then computed as

$$E(\hat{\lambda}) = E\left(\frac{3n}{\sum_{i=1}^n Y_i}\right) = 3nE\left(S_n^{-1}\right) = \lambda\frac{3n}{3n-1}$$

(c) Given that $\tilde{\lambda}_c$ is an unbiased estimator of λ, we note that

$$\lambda = E\left(\bar{\lambda}_c\right) = E(c\hat{\lambda}) = cE(\hat{\lambda}) = c\lambda\frac{3n}{3n-1}$$

so we get

$$c = \frac{3n-1}{3n}$$

(d) The key is to compute the variance of S_n^{-1}. Similarly to (b), we note that

$$\begin{aligned}E\left(S_n^{-2}\right) &= \int_0^\infty s^{-2}\frac{\lambda^{3n}}{\Gamma(3n)} s^{3n-1} e^{-\lambda s} ds\\ &= \int_0^\infty \frac{\lambda^2}{(3n-1)(3n-2)}\frac{\lambda^{3n-2}}{\Gamma(3n-2)} s^{(3n-2)-1} e^{-\lambda s} ds\\ &= \frac{\lambda^2}{(3n-1)(3n-2)}\int_0^\infty \frac{\lambda^{3n-2}}{\Gamma(3n-2)} s^{(3n-2)-1} e^{-\lambda s} ds\\ &= \frac{\lambda^2}{(3n-1)(3n-2)}\end{aligned}$$

and

$$\operatorname{Var}\left(S_n^{-1}\right)=E\left(S_n^{-2}\right)-\left[E\left(S_n^{-1}\right)\right]^2=\frac{\lambda^2}{(3n-1)(3n-2)}-\frac{\lambda^2}{(3n-1)^2}=\frac{\lambda^2}{(3n-1)^2(3n-2)}$$

Then the MSE of $\hat{\lambda}$ and $\bar{\lambda}_c$ is computed as follows,

$$\begin{gathered}\operatorname{Var}(\hat{\lambda})=\operatorname{Var}\left(3nS_n^{-1}\right)=9n^2\operatorname{Var}\left(S_n^{-1}\right)=\lambda^2\tfrac{9n^2}{(3n-1)^2(3n-2)}\\ \operatorname{MSE}(\hat{\lambda},\lambda)=[E(\hat{\lambda})-\lambda]^2+\operatorname{Var}(\hat{\lambda})=\left[\lambda\tfrac{3n}{3n-1}-\lambda\right]^2+\operatorname{Var}(\hat{\lambda})=\lambda^2\tfrac{3n+2}{(3n-1)(3n-2)}\end{gathered}$$

and as $\bar{\lambda}_c$ is unbiased,

$$\operatorname{MSE}\left(\tilde{\lambda}_c,\lambda\right)=\operatorname{Var}\left(\tilde{\lambda}_c\right)=\operatorname{Var}(c\hat{\lambda})=c^2\operatorname{Var}(\hat{\lambda})=\left(\frac{3n-1}{3n}\right)^2\lambda^2\frac{9n^2}{(3n-1)^2(3n-2)}=\frac{\lambda^2}{3n-2}$$

We further note that

$$\frac{\operatorname{MSE}(\hat{\lambda},\lambda)}{\operatorname{MSE}\left(\bar{\lambda}_c,\lambda\right)}=\frac{3n+2}{3n-1}>1$$

so $\tilde{\lambda}_c$ has a smaller MSE and is preferred.

(e) Yes. With MSE, the quadratic function $(\hat{\lambda}-\lambda)^2$ is used as the loss function. As λ can take values in $(0,\infty)$, such a loss function penalizes the over estimation (when $\hat{\lambda}>\lambda$) more heavily than the under estimation (when $0<\lambda<\lambda$). Moreover, while the under estimation is bounded, the over estimation is not. Therefore, a loss function that can "evenly" penalize the over estimation and the under estimation would be more appropriate; see Problem 3 for example.

2. (a) The loglikelihood is

$$\begin{aligned}\ell(\theta)&=\log\left(\prod_{i=1}^n f_{Y_i}\left(Y_i;\theta\right)\right)=\sum_{i=1}^n\log f_{Y_i}\left(Y_i;\theta\right)=\sum_{i=1}^n\log\left(\theta^{Y_i}(1-\theta)^{1-Y_i}\right)\\&=\log\theta\sum_{i=1}^n Y_i+\log(1-\theta)\left(n-\sum_{i=1}^n Y_i\right)\end{aligned}$$

Setting $\ell'(\theta)=0$ we get the unique root

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^n Y_i$$

We verify that

$$\ell''(\theta)=-\theta^{-2}\sum_{i=1}^n Y_i-(1-\theta)^{-2}\left(n-\sum_{i=1}^n Y_i\right)<0$$

holds as $0\leq\sum_{i=1}^n Y_i\leq n$. Therefore, $\hat{\theta}$ is the MLE of θ We compute the MSE of $\hat{\theta}$ as follows,

$$\begin{gathered}E(\hat{\theta})=E\left(\tfrac{1}{n}\sum_{i=1}^n Y_i\right)=\tfrac{1}{n}\sum_{i=1}^n E\left(Y_i\right)=\theta\\ \operatorname{Var}(\hat{\theta})=\operatorname{Var}\left(\tfrac{1}{n}\sum_{i=1}^n Y_i\right)=\tfrac{1}{n^2}\sum_{i=1}^n\operatorname{Var}\left(Y_i\right)=\tfrac{\theta(1-\theta)}{n}\end{gathered}$$

and

$$\operatorname{MSE}(\hat{\theta},\theta)=[E(\hat{\theta})-\theta]^2+\operatorname{Var}(\hat{\theta})=\frac{\theta(1-\theta)}{n}$$

(b) We compute the MSE of $\tilde{\theta}_c$ as follows,

$$\begin{gathered}E\left(\bar{\theta}_c\right)=E(c\hat{\theta})=cE(\hat{\theta})=c\theta\\ \operatorname{Var}\left(\bar{\theta}_c\right)=\operatorname{Var}(c\theta)=c^2\operatorname{Var}(\hat{\theta})=c^2\tfrac{\theta(1-\theta)}{n}\end{gathered}$$

and

$$\operatorname{MSE}\left(\bar{\theta}_c,\theta\right)=\left[E\left(\bar{\theta}_c\right)-\theta\right]^2+\operatorname{Var}\left(\bar{\theta}_c\right)=(c-1)^2\theta^2+c^2\frac{\theta(1-\theta)}{n}$$

(c) Define $g(c) = (c-1)^2\theta^2 + c^2\theta(1-\theta)/n$. Setting $g'(c) = 0$ we get

$$\hat{c} = \frac{n\theta}{(n-1)\theta + 1}$$

and as $g''(c) = 2\theta^2 + 2\theta(1-\theta)/n > 0, \hat{c}$ is the unique minimum. In other words, $\hat{c}$ is the optimal value such that the MSE of $\bar{\theta}_c$ attains the minimum for fixed n and θ.

3. Recall that the MLE is

$$\hat{\lambda} = \frac{3n}{\sum_{i=1}^n Y_i} = \frac{3n}{S_n}$$

and the bias corrected estimator is

$$\tilde{\lambda}_c = c\hat{\lambda} = \frac{3n-1}{3n}\frac{3n}{\sum_{i=1}^n Y_i} = \frac{3n-1}{\sum_{i=1}^n Y_i} = \frac{3n-1}{S_n}$$

where $S_n = \sum_{i=1}^n Y_i \sim \text{Gamma}(3n, \lambda)$. The expected loss of $\hat{\lambda}$ is

$$\begin{aligned} E[\mathcal{L}(\lambda, \hat{\lambda})] &= E[\lambda/\hat{\lambda} - 1 - \log(\lambda/\hat{\lambda})] = E\left[\frac{\lambda}{3n}S_n - 1 - \log\left(\frac{\lambda}{3n}\right) + \log S_n\right] \\ &= \frac{\lambda}{3n}E\left(\sum_{i=1}^n Y_i\right) - 1 - \log\left(\frac{\lambda}{3n}\right) + E\left(\log S_n\right) = \frac{\lambda}{3n}n\frac{3}{\lambda} - 1 - \log\left(\frac{\lambda}{3n}\right) + E\left(\log S_n\right) \\ &= -\log\left(\frac{\lambda}{3n}\right) + E\left(\log S_n\right) \end{aligned}$$

and the expected loss of $\bar{\lambda}_c$ is

$$\begin{aligned} E\left[\mathcal{L}\left(\lambda, \bar{\lambda}_c\right)\right] &= E\left[\lambda/\bar{\lambda}_c - 1 - \log\left(\lambda/\bar{\lambda}_c\right)\right] = E\left[\frac{\lambda}{3n-1}S_n - 1 - \log\left(\frac{\lambda}{3n-1}\right) + \log S_n\right] \\ &= \frac{\lambda}{3n-1}n\frac{3}{\lambda} - 1 - \log\left(\frac{\lambda}{3n-1}\right) + E\left(\log S_n\right) \\ &= \frac{1}{3n-1} - \log\left(\frac{\lambda}{3n-1}\right) + E\left(\log S_n\right) \end{aligned}$$

Note that one may further compute $E\left(\log S_n\right) = \psi(3n) - \log\lambda$, where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function (check the wikipedia page on Gamma Distribution (section on "Logarithmic expectation and variance") for more details). We also note that

$$E\left[\mathcal{L}\left(\lambda, \tilde{\lambda}_c\right)\right] - E[\mathcal{L}(\lambda, \hat{\lambda})] = \frac{1}{3n-1} - \log\left(\frac{3n}{3n-1}\right) = \left(\frac{3n}{3n-1} - 1\right) - \log\left(\frac{3n}{3n-1}\right) > 0$$

where we have used $(x-1) - \log x > 0$ for $x > 1$.

4. We would like to see if there is evidence in the data that the mean is larger than μ_0, which mathematically is to say: reject H_0 if $T > c$ where $T = \sum_{i=1}^n X_i$ is a test statistic and c is is a critical value. Note that $X_i \sim$ Poisson (μ_0) iid implies $T \sim$ Poisson $(n\mu_0)$ Now we fix the significance level α and set $P(T > c) = \alpha$ to find c. It follows that

$$\alpha = P(T > c) = 1 - P(T \leq c) = 1 - F_T(c)$$

or equivalently,

$$F_T(c) = 1 - \alpha$$

where F_T is the CDF of Poisson $(n\mu_0)$. We note that for Equation (1) to have a solution $c, (1-\alpha)$ can only take certain discrete values as Poisson is a discrete distribution, that is,

$$1 - \alpha \in \left\{e^{-n\mu_0}, e^{-n\mu_0}\left(1 + n\mu_0\right), e^{-n\mu_0}\left(1 + n\mu_0 + \frac{\left(n\mu_0\right)^2}{2}\right), \ldots\right\}$$

In conclusion, we reject H_0 if $\sum_{i=1}^n X_i > c$, or in other words, the data shows certain level of evidence (depending on α) to suggest the mean is greater than μ_0.

5. Similar to Problem 4, we would like to see if there is evidence in the data that the mean is larger than μ_0, which states: reject H_0 if $T > c$ where $T = \frac{1}{n}\sum_{i=1}^n X_i$ is a test statistic and c is is a critical value to be determined. Note that $T \sim N\left(\mu_0, \sigma_0^2/n\right)$, or equivalently, $\frac{\sqrt{n}(T-\mu_0)}{\sigma_0} \sim N(0,1)$ Now we fix the significance level α and set $P(T > c) = \alpha$ to find c. It follows that

$$\alpha = P(T > c) = 1 - P(T \le c) = 1 - P\left(\frac{\sqrt{n}\,(T-\mu_0)}{\sigma_0} \le \frac{\sqrt{n}\,(c-\mu_0)}{\sigma_0}\right) = 1 - \Phi\left(\frac{\sqrt{n}\,(c-\mu_0)}{\sigma_0}\right)$$

Solving for c we get

$$c = \frac{\sigma_0 \Phi^{-1}(1-\alpha)}{\sqrt{n}} + \mu_0$$

In conclusion, we reject H_0 if $\frac{1}{n}\sum_{i=1}^n X_i > c$, or in other words, the data shows certain level of evidence (depending on α) to suggest the mean is greater than μ_0.

6. (a) Differentiating $\ell(\boldsymbol{\mu})$ with respect to μ_i, we get

$$\frac{\partial \ell(\boldsymbol{\mu})}{\partial \mu_i} = Y_i - \mu_i$$

Solving for $\partial \ell(\boldsymbol{\mu})/\partial \mu_i = 0$ we get $\hat{\mu}_i = Y_i$, or equivalently,

$$\hat{\boldsymbol{\mu}} = (Y_1, \ldots, Y_n)^T$$

(b) We verify that the matrix $\left[-\frac{\partial^1 \ell(\mu)}{\partial \mu_i \partial \mu_j}\right]$ is an identity matrix, so it is positive definite and $\hat{\boldsymbol{\mu}}$ is the maximum.

(c) The expected square loss is the MSE of $\hat{\boldsymbol{\mu}}$, and it is computed as follows,

$$\mathrm{MSE}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = E\left[\sum_{i=1}^n (Y_i - \mu_i)^2\right] = \sum_{i=1}^n E\left(Y_i - \mu_i\right)^2 = \sum_{i=1}^n \mathrm{Var}\left(Y_i\right) = n$$

(d) Show that the expected loss of $\widetilde{\mu}_i$ is

$$\begin{aligned}
\mathcal{L}(\widetilde{\mu}_i, \mu_i) &= \mathrm{E}\left\{(\widetilde{\mu}_i - \mu_i)^2\right\} \\
&= \mathrm{E}\left\{(Y_i - \mu_i)^2(|Y_i| > \theta)\right\} + \mathrm{E}\left\{(0 - \mu_i)^2(|Y_i| < \theta)\right\} \\
&= 2\int_\theta^\infty (y-\mu_i)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2} dy + 2\mu_i^2 \int_0^\theta \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2}\, dy \\
&= 2\int_\theta^\infty (y-\mu_i)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2} dy + 2\mu_i^2 \left\{\Phi(\theta - \mu_i) - (\Phi(-\mu_i)\right\}.
\end{aligned}$$

We also define

$$\begin{aligned}
I(\mu_i, \theta) &\equiv 2\int_\theta^\infty (y-\mu_i)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2} dy \\
&= \int_\theta^\infty 2(y-\mu_i)\cdot(y-\mu_i) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2} dy \qquad (1) \\
&= \left[-2(y-\mu_i)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2}\right]_\theta^\infty + 2\int_\theta^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2} dy \qquad (2) \\
&= \left[-2(y-\mu_i)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_i)^2}\right]_\theta^\infty + 2\{1 - \Phi(\theta - \mu)\} \\
&= 2(\theta - \mu_i)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-\mu_i)^2} + 2\{1 - \Phi(\theta - \mu)\}. \qquad (3)
\end{aligned}$$

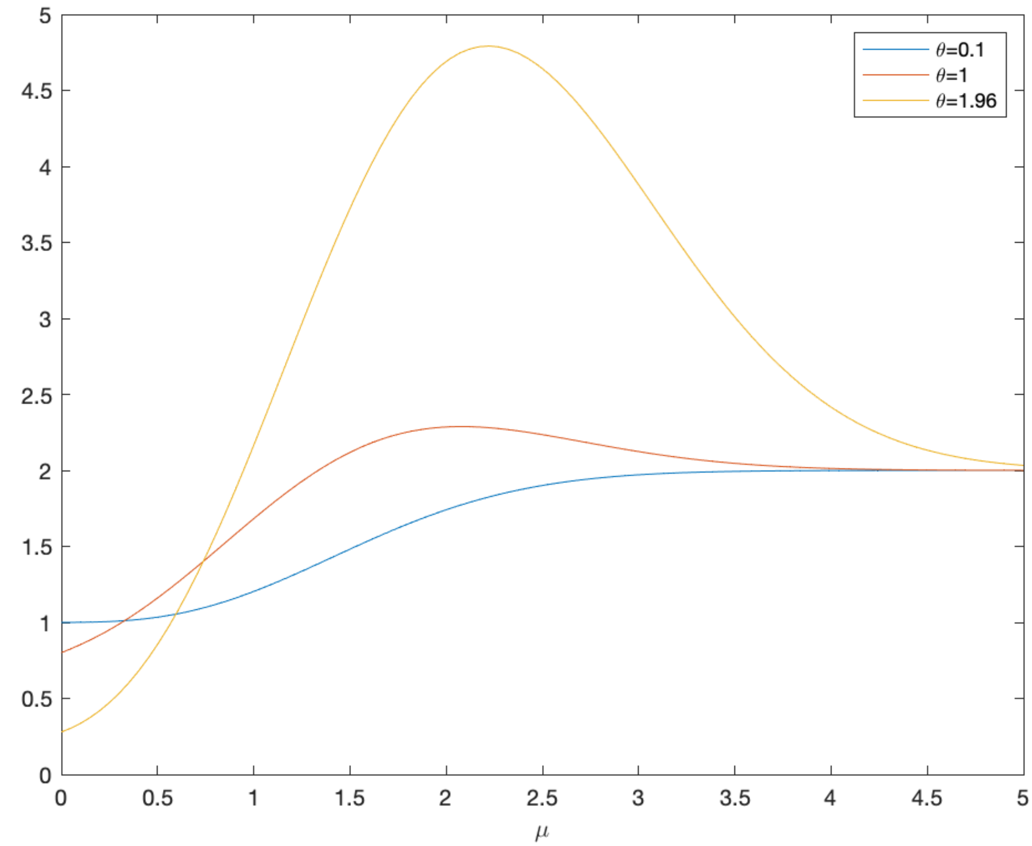
This is the first integral in $\mathcal{L}(\widetilde{\mu}_i, \mu_i)$, integrated by parts. Combining this simplifications we arrive at

$$\mathcal{L}(\widetilde{\mu}_i, \mu_i) = 2(\theta - \mu_i)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-\mu_i)^2} + 2\{1 - \Phi(\theta - \mu)\} + 2\mu_i^2 \left\{\Phi(\theta - \mu_i) - \Phi(-\mu_i)\right\}.$$

(e) Combining these expressions we get another for the estimation risk of $\widetilde{\mu}_i$. We have already obtained

$$\mathcal{L}(\widetilde{\mu}_i, \mu_i) = 2(\theta - \mu_i)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\theta-\mu_i)^2} + 2\{1 - \Phi(\theta - \mu)\} + 2\mu_i^2\left\{\Phi(\theta - \mu_i) - (\Phi(-\mu_i)\right\}.$$

Plot this using matlab as a function of μ_i taking i) $\theta = 0.1$, ii) $\theta = 1$ and iii) $\theta = 2\log(n)$ (I cheated in the last one). How do these cases vary? Basically if μ is small, then we are better off by taking a larger threshold.
What does this tell us about the estimation problem? Surprisingly, as μ/σ decreases it is better to threshold more.
It is reasonable to use unscaled numbers (in σ) as the variance is fixed to unity. No as if we redo the calculations above we will add a division by σ. The threshold should behave linearly in σ.



7. (a) Based on the rejection region $\{M_n > r\}$, the power of the test is

$$\beta(\phi) = P_\phi(M_n > r) = 1 - P\left(\max_i Y_i \le r\right) = 1 - \prod_{i=1}^{n} P_\phi\left(Y_i \le r\right) = 1 - (r/\phi)^n.$$

(b) Let

$$0.1 = P_{\phi_0}(M_n > r) = 1 - (r/\phi_0)^n = 1 - (2r)^n,$$

so $r = 0.9^{\frac{1}{n}}/2$.

(c) With the observed value $M_n = 0.47$, the p-value is

$$p = P_{\phi_0}\left(\max_i Y_i > 0.47\right) = 1 - (0.47/0.5)^{22} \approx 0.74,$$

so we do not reject H_0.

8. (a) Following the method introduced in slide 4, lecture 11,

$$\alpha = \sup_{\mu\in\Theta_0} \beta(\mu)$$

where

$$\beta(\mu) = \Pr_{\mu}\{T(\underline{Y}) > c\}.$$

Given the null hypothesis $H_0 : \mu = 0$, the unique value of parameter μ in Θ_0 is $\mu = 0$. Since

$$\beta(0) = \Pr_{\mu=0}\{T(\underline{Y}) > c\} = \Pr\{\mathcal{N}\left(0, n^{-1}\right) > c\} = 1 - \Phi(c\sqrt{n})$$

So the value of c is determined by

$$c = \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha).$$

(b)

$$\beta(1) = \Pr_{\mu=1}\{T(\underline{Y}) > c\} = \Pr\{\mathcal{N}\left(1, n^{-1}\right) > c\} = 1 - \Phi((c-1)\sqrt{n}).$$

(c)

$$\beta(1) = 1 - \Phi((c-1)\sqrt{n}) = 1 - \Phi(\Phi^{-1}(1-\alpha) - \sqrt{n}) \longrightarrow 1, \quad n \to \infty.$$

9. The ML estimator of the Poisson mean is

$$\hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

and the estimated variance of the estimator is

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n^2} n\mu = \frac{\mu}{n},$$

so

$$\hat{\mathrm{se}} = \sqrt{\frac{\mu}{n}}.$$

Which gives the Wald statistics to be:

$$W = n\frac{\hat{\mu} - \mu}{\mu}.$$

10. With μ known and given, the MLE of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

The variance of the estimator is

$$\mathrm{Var}(\widehat{\sigma}^2) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2\Big).$$