Example Sheet 1 Solutions
Example Class: 19 Feb 2019, 1pm, MR5

Part III Astrostatistics

I have included solutions for all problems (1-4). Where relevant, the accompanying code and data can be found on the website.

1 Combining Multiple Distance Estimates

Suppose N Cepheid variable stars are observed in the same external galaxy, and analyses of their brightness time series yields unbiased, independent estimates of their distance moduli, $\hat{\mu}_i$, $i = 1 \dots N$. The distance modulus is a logarithmic measure of their distance d from us on Earth.

$$\mu = 25 + 5 \log_{10}[d \text{ Mpc}^{-1}], \tag{1}$$

where Mpc is a mega-parsec, a unit of distance. Since the distances between galaxies are very much larger than the distances within galaxies, the Cepheid stars must all be effectively at the same true distance modulus μ. However, the estimates $\hat{\mu}_i$ have different sampling variances σ_i^2 around the true μ, because of observational heteroskedastic measurement error. We wish to combine the N independent estimates from the N individual stars to determine the "best" single estimate of the distance modulus μ to the galaxy.

1. Consider $N = 2$ stars. Consider all estimators that are linear combinations of the data $\hat{\mu}_1, \hat{\mu}_2$: $\hat{\mu} = \alpha_1 \hat{\mu}_1 + \alpha_2 \hat{\mu}_2$. What restriction is required of all *unbiased* linear estimators of μ?

   Solution:

   $$\mathbb{E}[\hat{\mu}] = \alpha_1 \mathbb{E}(\hat{\mu}_2) + \alpha_2 \mathbb{E}(\hat{\mu}_2) = (\alpha_1 + \alpha_2)\mu$$

   Therefore to be unbiased, $\alpha_2 = 1 - \alpha_1$ and the estimators are

   $$\hat{\mu} = \alpha_1 \hat{\mu}_1 + (1 - \alpha_1)\hat{\mu}_2$$

2. For $N = 2$, what is the sampling variance $\text{Var}[\hat{\mu}]$ of the unbiased linear estimators in part 1? Find the *minimum variance* unbiased linear estimator by solving for the appropriate coefficients. Show that they can be expressed as $\alpha_i = K\sigma_i^\gamma$, and determine K and γ. What is the variance of the minimum variance unbiased linear estimator?

   Solution: Because of independence,

   $$V(\alpha_1) \equiv \mathbf{Var}[\hat{\mu}] = \alpha_1^2 \mathbf{Var}[\hat{\mu}_1] + (1 - \alpha_1)^2 \mathbf{Var}[\hat{\mu}_2]$$

   Taking the first derivative wrt α_1, and setting it equal to zero, yields:

   $$\alpha_1 = \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} = K\sigma_1^\gamma$$

and

$$\alpha_2 = 1 - \alpha_1 = \frac{\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} = K\sigma_2^{\gamma}$$

therefore $\alpha_i = K\sigma_i^{\gamma}$ with $\gamma = -2$ and $K = (\sigma_1^{-2} + \sigma_2^{-2})^{-1}$. The second derivative,

$$\frac{d^2V}{d\alpha_1^2} = K > 0$$

indicating this is a minimum.

3. Now generalise to $N > 2$. Consider all linear estimators of the form $\hat{\mu} = \sum_{i=1}^{N} \alpha_i \hat{\mu}_i$. What are the coefficients of the minimum variance unbiased linear estimator? Verify that they satisfy the first- and second-derivative conditions for a local minimum.

   Solution: Similarly, by computing the expectation and requiring unbiasedness, we find $\sum_{i=1}^{N} \alpha_i = 1$. The variance is

$$V(\boldsymbol{\alpha}) = \sum_{i=1}^{N-1} \alpha_i^2 \sigma_i^2 + \left(1 - \sum_{i=1}^{N-1} \alpha_i\right)^2 \sigma_N^2$$

   We then have $N - 1$ partial derivatives to set to zero.

$$\frac{\partial V}{\partial \alpha_i} = 2\alpha_i \sigma_i^2 - 2\left(1 - \sum_{i=1}^{N-1} \alpha_i\right)\sigma_N^2 = 0$$

   Generalising from Part 2, we can use the ansatz $\alpha_i = K\sigma_i^{-2}$, where $K^{-1} = \sum i = 1^N \sigma_i^{-2}$. It is easy to show this solution satisfies the first-order conditions. The second order conditions are that the Hessian $\boldsymbol{H}$ is a positive definite matrix.

$$H_{ij} = \frac{\partial^2 V}{\partial \alpha_i \, \partial \alpha_j}$$

   For $i = 1, \ldots N - 1$,

$$\frac{\partial^2 V}{\partial \alpha_i^2} = 2\sigma_i^2 + 2\sigma_N^2$$

   and for off-diagonal entries $i = j$

$$\frac{\partial^2 V}{\partial \alpha_i \, \partial \alpha_j} = 2\sigma_n^2$$

   Therefore $\boldsymbol{H} = \boldsymbol{D} + 2\boldsymbol{C}$, where $D_{ij} = 2\sigma_i^2 \delta_{ij}$ and $C_{ij} = \sigma_N^2$ and δ_{ij} is the Kronecker delta. Because for any vector $\boldsymbol{z} \in \mathbb{R}^{N-1}$, we can show that

$$\boldsymbol{z}^T \boldsymbol{H} \boldsymbol{z} = \boldsymbol{z}^T \boldsymbol{D} \boldsymbol{z} + 2\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z} > 0$$

   since $\boldsymbol{z}^T \boldsymbol{D} \boldsymbol{z} = \sum_{i=1}^{N-1} z_i^2 \sigma_i^2 > 0$ and $\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z} = 2\sigma_N^2 (\sum_{i=1}^{N-1} z_i)^2 > 0$.

4. For $N > 2$, suppose all the uncertainties of the individual estimates are the same, $\sigma_i = \sigma$ for $i = 1, \ldots, N$. What is the variance of the minimum variance unbiased linear estimator, and how does it scale with the number of stars N?

Solution: Since the variance is:

$$V(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i^2 \sigma_i^2 = K^2 \sum_{i=1}^{N} \sigma_i^{-2}$$

if $\sigma_i = \sigma \ \forall i$, then $V = \sigma^2/N$. The variance scales inversely with N, and the standard deviation decreases by $\sqrt{N}$.

5. Now suppose, because of systematic uncertainties, the distance errors for $N > 2$ stars are jointly Gaussian and correlated between stars, with known pairwise covariances $\mathrm{Cov}[\hat{\mu}_i, \hat{\mu}_j] \equiv C_{ij} = \sigma_i \, \sigma_j \, \rho_{ij}$, and correlation coefficients $|\rho_{ij}| < 1$. What is required for the matrix $\boldsymbol{C}$ to be a valid covariance matrix? Assuming $\boldsymbol{C}$ is a valid covariance matrix, derive the maximum likelihood estimator $\hat{\mu}_{\mathrm{MLE}}$. Compute the bias and variance of the MLE. Compare the variance to the Cramér-Rao bound. You may leave your answers in terms of elements Λ_{ij} of the inverse of the covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{C}^{-1}$.

Solution: The likelihood function is

$$P(\hat{\boldsymbol{\mu}}|\mu) = N(\hat{\boldsymbol{\mu}}|\, \mathbf{1}\mu, \boldsymbol{C}) = |2\pi \boldsymbol{C}|^{-1/2} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{\mu}} - \mathbf{1}\mu)^T \boldsymbol{C}^{-1} (\hat{\boldsymbol{\mu}} - \mathbf{1}\mu)\right)$$

where 1 is a vector of ones. The matrix $\boldsymbol{C}$ must be symmetric and positive definite. The inverse $\boldsymbol{C}^{-1}$ exists and the determinant $|\boldsymbol{C}| > 0$. The log likelihood function is

$$l(\mu) = \mathbf{const} - \frac{1}{2} \sum_{i,j} (\hat{\mu}_i - \mu) \Lambda_{ij} (\hat{\mu}_j - \mu)$$

Setting the first derivative with respect to μ to zero, we find

$$\hat{\mu}_{\mathbf{mle}} = \sum_j w_j \hat{\mu}_j$$

where $w_j = (\sum_i \Lambda_{ij})/(\sum_{i,m} \Lambda_{im})$. The second derivative is

$$\frac{d^2 l}{d\mu^2} = -\left(\sum_{i,j} \Lambda_{ij}\right) < 0$$

We see that this is negative because $\sum_{ij} \Lambda_{ij} = \mathbf{1}^T \boldsymbol{C}^{-1} \mathbf{1}$. Suppose $\boldsymbol{x} = \boldsymbol{C}^{-1}\mathbf{1}$, then $\mathbf{1}^T \boldsymbol{C}^{-1} \mathbf{1} = (\boldsymbol{C}x)^T \boldsymbol{C}^{-1} \boldsymbol{C}\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{C}\boldsymbol{x} > 0$, since $\boldsymbol{C}$ is positive definite. Therefore the second order conditions are satisfied and this is a maximum. We can straightforwardly compute the expectation and variance

$$\mathbb{E}[\hat{\mu}_{\mathbf{mle}}] = \mu$$

$$\mathbf{Var}[\hat{\mu}_{\mathbf{mle}}] = \mathbf{Cov}[\hat{\mu}_{\mathbf{mle}}, \hat{\mu}_{\mathbf{mle}}]$$

after some matrix algebra, we find $\mathbf{Var}[\hat{\mu}_{\mathbf{mle}}] = (\sum_{ij} \Lambda_{ij})^{-1}$. The Fisher information is

$$I(\mu) = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \mu^2}\right] = \left(\sum_{i,j} \Lambda_{ij}\right)$$

The Cramér-Rao bound states that

$$\mathbf{Var}[\hat{\mu}_{\mathbf{mle}}] \geq I(\mu)^{-1}$$

However, we find that $\mathbf{Var}[\hat{\mu}_{\mathbf{mle}}] = I(\mu)^{-1}$ so the bound is said to be *saturated*.

2 Estimating the Hubble Constant on the Local Distance Ladder

Type Ia supernovae (SNe Ia) are thermonuclear explosions of white dwarf stars. They are used as "standard candles," objects with a narrow range of peak absolute magnitude (log luminosity), so their distances can be judged from their apparent magnitudes (log apparent brightness or flux). Suppose the peak absolute magnitudes of SNe Ia come from an intrinsic Gaussian distribution with population mean M_0 and variance σ^2_{int}:

$$M_s \sim N(M_0, \sigma^2_{\text{int}}) \tag{2}$$

for every supernova s. The true absolute magnitude is related to the true apparent magnitude m_s via the true distance modulus μ_s, which is a logarithmic measure of the true distance d_s.

$$m_s = M_s + \mu_s. \tag{3}$$

The definition of the distance modulus is $\mu = 25 + 5\log_{10}[d \text{ Mpc}^{-1}]$, where Mpc is a megaparsec, a unit of astronomical distance. For every supernova s, we obtain an estimate of its peak apparent magnitude $\hat{m}_s$ with known error variance $\sigma^2_{m,s}$.

$$\hat{m}_s | \, m_s \sim N(m_s, \sigma^2_{m,s}) \tag{4}$$

Type Ia supernovae can be observed at great distances, and are used to estimate the current expansion rate of the Universe, the Hubble constant H_0. However, to do this, their luminosities (absolute magnitudes) must be calibrated. *Calibration* is the statistical task of estimating M_0 and σ^2_{int}. (In most of this problem we will assume known values of σ^2_{int} for simplicity). Ideally, calibration is done simultaneously with the estimation of H_0 to ensure proper propagation of uncertainties.

Suppose we have a calibrator set of $k = 1, \ldots, K$ supernovae located in nearby galaxies in which we can observe Cepheid variable stars. We can use the observations of the apparent brightness and periodicity of the variable stars, along with the Cepheid period-luminosity relation, to estimate their distances. For each calibrator supernova, we have an unbiased distance modulus estimate $\hat{\mu}_{C,k}$ with a Gaussian error with variance $\sigma^2_{C,k}$, obtained from the Cepheid stars in the same galaxy.

$$\hat{\mu}_{C,k} | \, \mu_k \sim N(\mu_k, \sigma^2_{C,k}) \tag{5}$$

We also have a much larger, "Hubble Flow" set of $i = 1, \ldots, N$ supernovae which are much further away, so the Cepheids stars cannot be observed in their galaxies. However, they are far enough away that they participate in the smooth, overall expansion of the Universe. Thus, they follow the Hubble Law, the linear relation between their recession velocities $v_i = c\, z_i$ and their distances d_i:

$$d_i = \frac{c}{H_0} z_i \tag{6}$$

where c is the speed of light and z_i is the redshift. Assume the redshift is measured exactly for each supernova i. In terms of the distance modulus,

$$\mu_i = 25 + 5\log_{10}\left[\frac{c\, z_i}{H_0} \text{ Mpc}^{-1}\right]. \tag{7}$$

The units of the Hubble Constant are km s^{-1} Mpc^{-1}. Let

$$\theta = 5\log_{10}[H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})] \tag{8}$$

1. Write down the likelihood function of (M_0, θ) in terms of the data of the calibrator set $\{\hat{m}_k, \hat{\mu}_{C,k}\}$ and the Hubble flow sample $\{\hat{m}_i, z_i\}$, and the relevant variances.

   Solution: First we derive the likelihood for the Cepheid sample. We could begin by writing down the product of sampling distributions of the data given the latent variables and parameters, and then integrating out the latent (true) variables.

$$\int N(\hat{m}_k|\, m_k, \sigma^2_{m,k})\, N(\hat{\mu}_{C,k}|\mu_k, \sigma^2_{C,k})\, N(M_k = m_k - \mu_k|\, M_0, \sigma^2_{\mathbf{int}})\, dm_k\, d\mu_k$$

   After doing some Gaussian integrals and algebraic simplifications, we would find:

$$= N(\hat{m_k} - \hat{\mu}_k|\, M_0, \sigma^2_{\mathbf{int}} + \sigma^2_{m,k} + \sigma^2_{C,k})$$

   Another way to derive this is to separate the means from the random variables:

$$\hat{m}_k = m_k + \epsilon_{m,k} = M_k + \mu_k + \epsilon_{m,k} = M_0 + \epsilon_{\mathbf{int},k} + \hat{\mu}_{C,k} + \epsilon_{C,k} + \epsilon_{m,k}$$

   or

$$\hat{m}_k - \hat{\mu}_{C,k} = M_0 + \epsilon_{\mathbf{int},k} + \epsilon_{C,k} + \epsilon_{m,k}$$

   where the ϵ's are independent mean zero Gaussian random variables with variances $\sigma^2_{\mathbf{int},k}, \sigma^2_{C,k}, \sigma^2_{m,k}$, respectively. Because the sum of independent Gaussian random variables is also Gaussian, with total mean equal to the sum of the means, and the total variance equal to the sum of the variances, we find:

$$P(\hat{m}_k, \hat{\mu}_{c,k}|M_0, \sigma^2_{\mathbf{int}}) = N(\hat{m_k} - \hat{\mu}_k|\, M_0, \sigma^2_{\mathbf{int}} + \sigma^2_{m,k} + \sigma^2_{C,k})$$

   as before. Multiplying the independent supernovae, the likelihood function is:

$$L_{\mathbf{cal}}(M_0, \sigma^2_{\mathbf{int}}) = \prod_{k=1}^{K} N(\hat{m_k} - \hat{\mu}_k|\, M_0, \sigma^2_{\mathbf{int}} + \sigma^2_{m,k} + \sigma^2_{C,k})$$

   We see that the likelihood function only depends on the data through $\hat{M}_k \equiv \hat{m_k} - \hat{\mu}_k$. Using similar logic for the Hubble Flow sample,

$$\hat{m}_i = m_i + \epsilon_{m,i} = M_i + \mu_i + \epsilon_{m,i} = M_0 + \epsilon_{\mathbf{int}} + \mu_i + \epsilon_{m,i}.$$

   We can rewrite the distance modulus to factor out the Hubble constant:

$$\mu_i = 25 + 5\log_{10}(cz_i/H_0) = 25 - \theta + 5\log_{10}(cz_i/100).$$

   If we define $\tilde{M}_i = \hat{m}_i - 25 - 5\log_{10}(cz_i/100)$, then we can write

$$P(\hat{m}_i|\, z_i) = N(\tilde{M}_i|\, M_0 - \theta, \sigma^2_{\mathbf{int}} + \sigma^2_{m,i})$$

   So the likelihood function for the Hubble flow sample is:

$$L_{\mathbf{HF}}(M_0, \theta, \sigma^2_{\mathbf{int}}) = \prod_{i=1}^{N} N(\tilde{M}_i|\, M_0 - \theta, \sigma^2_{\mathbf{int}} + \sigma^2_{m,i}).$$

   So the total likelihood function of both independent datasets is

$$L(M_0, \theta, \sigma^2_{\mathbf{int}}) = \left[\prod_{k=1}^{K} N(\hat{m_k} - \hat{\mu}_k|\, M_0, \sigma^2_{\mathbf{int}} + \sigma^2_{m,k} + \sigma^2_{C,k})\right] \times \left[\prod_{i=1}^{N} N(\tilde{M}_i|\, M_0 - \theta, \sigma^2_{\mathbf{int}} + \sigma^2_{m,i})\right]$$

For convenience, I will define $\tau_k^2 = \sigma_{\mathbf{int}}^2 + \sigma_{C,k}^2 + \sigma_{m,k}^2$ as the total variance of each calibrator supernova, and $\sigma_i^2 = \sigma_{\mathbf{int}}^2 + \sigma_{m,i}^2)$ as the total variance for each Hubble flow supernova.

2. Assume that all error variances are known, as well as the intrinsic dispersion $\sigma_{\rm int}$. Derive the maximum likelihood estimators for (M_0, θ).

   Solution: The log likelihood function is

$$l(M_0, \theta, \sigma_{\mathbf{int}}^2) = \sum_k -\frac{1}{2}\frac{(\hat{M}_k - M_0)^2}{\tau_k^2} - \frac{1}{2}\log(2\pi\tau_k^2) + \sum_i -\frac{1}{2}\frac{(\tilde{M}_i - M_0 + \theta)^2}{\sigma_i^2} - \frac{1}{2}\log(2\pi\sigma_i^2)$$

   We find the maximum likelihood solutions for M_0 and θ by finding where the first derivatives are zero (and the second derivatives are negative, see below).

$$\frac{\partial l}{\partial M_0} = \sum_k \frac{(\hat{M}_k - M_0)}{\tau_k^2} + \sum_i \frac{(\tilde{M}_i - M_0 + \theta)}{\sigma_i^2} = 0$$

$$\frac{\partial l}{\partial \theta} = \sum_i -\frac{(\tilde{M}_i - M_0 + \theta)}{\sigma_i^2} = 0$$

   Therefore,

$$\hat{M}_0 - \hat{\theta} = \frac{\sum_i \sigma_i^{-2}\tilde{M}_i}{\sum_i \sigma_i^{-2}}$$

   and

$$\hat{M}_0 = \frac{\sum_k \tau_k^{-2}\hat{M}_k}{\sum_k \tau_k^{-2}}$$

   so

$$\hat{\theta} = \frac{\sum_k \tau_k^{-2}\hat{M}_k}{\sum_k \tau_k^{-2}} - \frac{\sum_i \sigma_i^{-2}\tilde{M}_i}{\sum_i \sigma_i^{-2}} \tag{9}$$

3. Evaluate the bias and variance your estimators $(\hat{M}_0, \hat{\theta})$, and compare against the Cramér-Rao bound.

   Solution: The expectations of these estimators show that they are unbiased:

$$\mathbb{E}[\hat{M}_0] = \frac{\sum_k \tau_k^{-2}\mathbb{E}[\hat{M}_k]}{\sum_k \tau_k^{-2}} = M_0$$

$$\mathbb{E}[\hat{\theta}] = M_0 - \frac{\sum_i \sigma_i^{-2}\mathbb{E}[\tilde{M}_i]}{\sum_i \sigma_i^{-2}} = M_0 - (M_0 - \theta) = \theta$$

   The variances are

$$\mathbf{Var}[\hat{M}_0] = \frac{\sum_k \tau_k^{-4}\tau_k^2}{(\sum_k \tau_k^{-2})^2} = \frac{1}{\sum_k \tau_k^{-2}}$$

   and similarly,

$$\mathbf{Var}[\hat{\theta}] = \frac{1}{\sum_k \tau_k^{-2}} + \frac{1}{\sum_i \sigma_i^{-2}}$$

We can compute the Fisher matrix using the second derivatives of the negative log likelihood function

$$-\frac{\partial^2 l}{\partial M_0^2} = \sum_k \tau_k^{-2} + \sum_i \sigma_i^{-2}$$

$$-\frac{\partial^2 l}{\partial M_0 \partial \theta} = -\sum_i \sigma_i^{-2}$$

$$-\frac{\partial^2 l}{\partial \theta^2} = \sum_i \sigma_i^{-2}$$

Therefore the Fisher information matrix is:

$$\boldsymbol{I} = \begin{pmatrix} \sum_k \tau_k^{-2} + \sum_i \sigma_i^{-2} & -\sum_i \sigma_i^{-2} \\ -\sum_i \sigma_i^{-2} & \sum_i \sigma_i^{-2} \end{pmatrix}$$

and the inverse is

$$\boldsymbol{I}^{-1} = \det(\boldsymbol{I})^{-1} \begin{pmatrix} \sum_i \sigma_i^{-2} & \sum_i \sigma_i^{-2} \\ \sum_i \sigma_i^{-2} & \sum_k \tau_k^{-2} + \sum_i \sigma_i^{-2} \end{pmatrix}$$

where the determinant is

$$\begin{aligned} \det(\boldsymbol{I}) &= \left(\sum_k \tau_k^{-2} + \sum_i \sigma_i^{-2}\right)\left(\sum_i \sigma_i^{-2}\right) - \left(\sum_i \sigma_i^{-2}\right)^2 \\ &= \left(\sum_k \tau_k^{-2}\right)\left(\sum_i \sigma_i^{-2}\right). \end{aligned}$$

Therefore a lower bound on the variance is

$$\mathbf{Var}[\hat{M}_0] \geq (\boldsymbol{I}^{-1})_{11} = \left(\sum_k \tau_k^{-2}\right)^{-1}$$

$$\mathbf{Var}[\hat{\theta}] \geq (\boldsymbol{I}^{-1})_{22} = \frac{1}{\sum_k \tau_k^{-2}} + \frac{1}{\sum_i \sigma_i^{-2}}$$

We can see from our previous calculation of the variances, that they saturate their Cramér-Rao bounds.

4. Simplify for the case where each source of error is homoskedastic, i.e. $\sigma_{C,k} = \sigma_C$ for all calibrators, and $\sigma_{m,s} = \sigma_m$ for all supernovae. Derive an expression for the variance of $\hat{\theta}$, and propagate the error to derive the standard deviation of $\hat{H}_0$.

   Solution: with $\tau^2 = \sigma_{\mathbf{int}}^2 + \sigma_C^2 + \sigma_m^2$, and $\sigma^2 = \sigma_{\mathbf{int}}^2 + \sigma_m^2$

$$\mathbf{Var}[\hat{\theta}] = \tau^2/K + \sigma^2/N$$

Using propagation of errors, the approximate variance of $\hat{h} = \exp(\hat{\theta}\log(10)/5)$ is

$$\mathbf{Var}[\hat{h}] \approx \left|\frac{\partial \hat{h}}{\partial \hat{\theta}}\right|^2 \mathbf{Var}[\hat{\theta}]$$

Therefore

$$\mathbf{STD}[\hat{h}] \approx 0.46\,\hat{h}\,\sqrt{\tau^2/K + \sigma^2/N}$$

In particular the fractional uncertainty is:

$$\frac{\mathbf{STD}[\hat{h}]}{\hat{h}} = \frac{\mathbf{STD}[\hat{H}_0]}{\hat{H}_0} = 0.46\,\sqrt{\tau^2/K + \sigma^2/N}$$

5. Suppose $K = 19, N = 300$ and $\sigma_C = 0.1$, $\sigma_m = 0.05$, and $\sigma_{\rm int} = 0.1$. What uncertainty (standard deviation of error) would you expect for $\hat{H}_0$? What would decrease the uncertainty the most: obtaining one more calibrator supernova, or one more Hubble flow supernova?

   For these values, we find

$$\mathbf{STD}[\hat{\theta}] = 0.0350$$

   resulting in a fractional uncertainty:

$$\frac{\mathbf{STD}[\hat{h}]}{h} = 0.46\,\mathbf{STD}[\hat{\theta}] = 0.0161$$

   Increasing $K = 20$, yields a 2.5% reduction of $\mathbf{STD}[\hat{\theta}] = 0.0342$ and increasing $N = 301$ yields a negligible improvement to $\mathbf{STD}[\hat{\theta}] = 0.0350$. Therefore observing one more calibrator supernova is more valuable.

6. Return to the heteroskedastic errors. Look up the paper "Measuring the Hubble constant with Type Ia supernovae as near-infrared standard candles" by Dhawan, Jha & Leibundgut (2018, Astronomy & Astrophysics 609, A72). For the calibrator sample use the "m_J","$\sigma_{\rm fit}$", "$\mu_{\rm Ceph}$" and "$\sigma_{\rm Ceph}$" columns in Table 1 for $\{\hat{m}_k, \sigma_{m,k}, \hat{\mu}_{C,k}, \sigma_{C,k}\}$, respectively. For the Hubble flow sample, use the "$z_{\rm CMB}$ (flow-corrected)", "m_J" and "$\sigma_{\rm fit}$" columns in Table 2 for $\{z_i, \hat{m}_i, \sigma_{m,i}\}$. Assuming a value of $\sigma_{\rm int} = 0.10$, what estimate do you obtain for the Hubble constant and its standard error? What if $\sigma_{\rm int} = 0.16$?

   Solution: Plugging in the data into our estimators, for $\sigma_{\rm int} = 0.11$, we find the MLEs and standard deviations, $\hat{M}_0 = -18.52 \pm 0.04$, $\hat{\theta} = -0.724 \pm 0.05$, and $\hat{h} = 0.72$ with a fraction uncertainty of 2.3%.

   Plugging in the data into our estimators, for $\sigma_{\rm int} = 0.16$, we find the MLEs and standard deviations, $\hat{M}_0 = -18.53 \pm 0.06$, $\hat{\theta} = -0.724 \pm 0.068$, and $\hat{h} = 0.72$ with a fraction uncertainty of 3.1%.

7. (Bonus) Now allowing $\sigma_{\rm int}$ to be a free parameter, compute the joint maximum likelihood estimates $(\hat{M}_0, \hat{\theta}, \hat{\sigma}_{\rm int})$, and estimate the standard errors from the inverse of the observed Fisher matrix at the MLE.

3 Likelihood with Observational Selection Effects

Suppose the masses of N stars $\{y_i\}$ observed in a star-forming region follow a Pareto distribution:

$$P(y|\gamma) = \begin{cases} 0, & y < t_0 \\ A\,y^{-\gamma}, & y \geq t_0 \end{cases}, \tag{10}$$

where $t_0 = 1$ is a lower limit, A is a normalisation constant, and γ is the exponent of astrophysical interest.

1. Solve for A. What conditions on γ must hold for y to have finite expectation and variance?

   Solution: We will repeatedly use the following integral.

$$\int_a^b A\, y^{-\gamma+n}\, dy = \frac{A}{\gamma - n - 1}\left[a^{-(\gamma-n-1)} - b^{-(\gamma-n-1)}\right]$$

   For $n = 0$, and $\int_{t_0}^{\infty} P(y|\,\gamma)\, dy = 1$, we obtain $A = (\gamma-1)t_0^{(\gamma-1)}$, which is only finite when $\gamma > 1$. The expected value is obtained with $n = 1$,

$$\mathbb{E}[y] = \int y\, P(y|\,\gamma)\, dy,$$

   and the integral is only finite when $\gamma > 2$. The variance is

$$\mathbf{Var}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2.$$

   The first term is obtained with $n = 2$, and is only finite when $\gamma > 3$.

2. Derive the maximum likelihood estimator for γ. Using the dataset of $\{y_i\}$ provided, compute the maximum likelihood estimate. Bootstrap the dataset 100 times to compute its standard deviation.

   Solution: The likelihood for N stars is

$$P(\boldsymbol{y}|\,\gamma) = \prod_{i=1}^{N}(\gamma-1)\, t_0^{\gamma-1} y^{-\gamma}$$

   and the log likelihood is

$$l(\gamma) = \log P(\boldsymbol{y}|\,\gamma) = N\log(\gamma-1) + N(\gamma-1)\log t_0 - \gamma\sum_{i=1}^{N}\log(y_i).$$

   The first derivative is

$$\frac{dl}{d\gamma} = \frac{N}{\gamma-1} - \sum_{i=1}^{N}\log(y_i/t_0)$$

   Setting this equal to zero, we obtain the MLE:

$$\hat{\gamma} = \frac{N}{\sum_{i=1}^{N}\log(y_i/t_0)} + 1$$

   We also check that the second derivative is negative,

$$\frac{d^2l}{d\gamma^2} = -\frac{N}{(\gamma-1)^2} - \sum_{i=1}^{N}\log(y_i/t_0) < 0$$

   so this is a maximum. Using the dataset, we find $\hat{\gamma} = 3.03$. Using 100 bootstraps, we find that its standard deviation is 0.14 (the exact number will vary slightly due to randomisation).

3. After you've done all that work, the astronomer who obtained the data now tells you that she would have been unable to see any stars with masses $y > t_1$ due to observational selection effects. This can be modeled by introducing an *inclusion* vector $\boldsymbol{I}$ with binary components

$$I_i = \begin{cases} 0, & y_i \text{ is not observed,} \\ 1, & y_i \text{ is observed.} \end{cases} \tag{11}$$

The inclusion vector has a probability representing the data collection process:

$$P(I_i = 1 |\, y_i, t_1) = \begin{cases} 1, & y_i \le t_1 \\ 0, & y_i > t_1 \end{cases} \tag{12}$$

where $t_1 = 5$ is the known observational limit. Let $y_i^{\rm obs}$ indicate the the mass y_i was observed, $I_i = 1$. Derive the likelihood function $P(y_i^{\rm obs}|I_i = 1, \gamma)$ for one star, and then the likelihood for the full sample.

Solution: For a single *observed* object ($I_i = 1$), the likelihood for $y_i^{\rm obs}$ is, by the law of conditional probability

$$\begin{aligned} P(y_i^{\rm obs} |\, I_i = 1, \gamma) &= \frac{P(I_i = 1, y_i |\, \gamma)}{\int dy_i P(I_i = 1, y_i |\, \gamma)} \\ &= \frac{P(I_i = 1 |\, y_i) P(y_i |\, \gamma)}{\int dy_i P(I_i = 1 |\, y_i) P(y_i|\gamma)}. \end{aligned}$$

Therefore,

$$P(y_i^{\rm obs} |\, I_i = 1, \gamma) = \begin{cases} (y_i^{\rm obs})^{-\gamma} \left[\frac{(\gamma-1)}{t_0^{-(\gamma-1)} - t_1^{-(\gamma-1)}} \right], & t_0 < y_i^{\rm obs} < t_1 \\ 0, & \text{otherwise.} \end{cases}$$

Then the observed data likelihood is just the product of N terms for each observed data point:

$$P(\boldsymbol{y}_{\rm obs} |\, \boldsymbol{I}, \gamma) = \prod_{i=1}^{N} P(y_i^{\rm obs} |\, I_i = 1, \gamma, t_1)$$

and a maximum likelihood estimate comes from optimizing this over possible values of γ.

4. Compute the corrected MLE on the dataset provided for $\boldsymbol{y}^{\rm obs}$. Bootstrap the dataset 100 times to compute its standard deviation.

Solution: By plotting the corrected likelihood function on a find grid, we find $\hat{\gamma} = 2.6$. Using bootstrap we find its standard deviation to be 0.20. Here is a plot of the likelihood functions:

5. Using the corrected MLE value $\hat{\gamma}$ you found, generate 100 new samples of the same size by drawing random numbers from the observationally-truncated distribution $P(y_i^{\rm obs}|I_i = 1, \hat{\gamma})$. Compute the corrected MLE on each simulated dataset and compute its standard deviation over simulations.

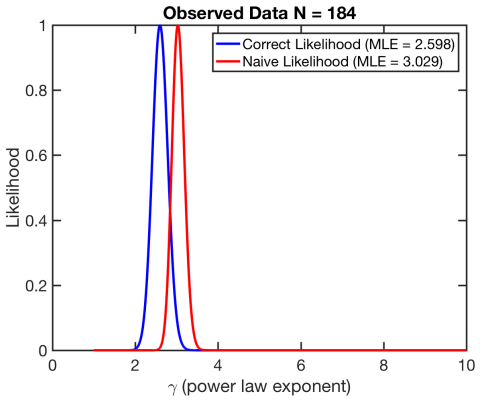

Figure 1: **The naive and corrected likelihoods.**

Solution: We can generate random samples from the truncated Pareto distribution using the inverse CDF (inverse transform) method. Using the corrected sampling distribution, we can compute the cumulative distribution function

$$F(y) = \int_{-\infty}^{y} P(y'|\, I_i = 1, \gamma)\, dy'$$
$$= \begin{cases} 0 & y < t_0 \\ \frac{t_0^{-(\gamma-1)} - y^{-(\gamma-1)}}{t_0^{-(\gamma-1)} - t_1^{-(\gamma-1)}} & t_0 \le y \le t_1 \\ 1 & y > t_1 \end{cases}$$

Therefore the inverse CDF is

$$F^{-1}(u) = \left[u\left(t_1^{-(\gamma-1)} - t_0^{-(\gamma-1)}\right) + t_0^{-(\gamma-1)}\right]^{1/(\gamma-1)}$$

We can generate a random number from the truncated Pareto distribution by first sampling $u \sim U(0,1)$, and then computing $y = F^{-1}(u)$. In the included code (ex1p3_soln.m), we generate 100 samples of $N = 184$ stars each, and then compute the corrected MLE on each simulated sample. We find the mean MLE is 2.56 with standard deviation 0.17.

4 Bayesian Inference for Gaussian data with unknown mean and variance

Suppose instead that log stellar masses $\{y_i\}$ are from Gaussian distribution with unknown population mean μ and variance σ^2:

$$y_i \sim N(\mu, \sigma^2) \tag{13}$$

for $i = 1, \ldots, N$.

1. Derive the likelihood function $P(\boldsymbol{y}|\, \mu, \sigma^2)$, expressed in terms of the sufficient statistics: the sample mean,

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{14}$$

and sample variance:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2. \tag{15}$$

Solution:

$$\begin{aligned}
P(\boldsymbol{y}|\,\mu,\sigma^2) &= \prod_{i=1}^{N} N(y_i|\,\mu,\sigma^2) = \prod_{i=1}^{N}(2\pi\sigma^2)^{-1/2}\, e^{-(y_i-\mu)^2/2\sigma^2} \\
&= (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right) \\
&= (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y}+\bar{y}-\mu)^2\right) \\
&= (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[(y_i-\bar{y})^2 + 2(\bar{y}-\mu)(y_i-\bar{y}) + (\bar{y}-\mu)^2\right]\right) \\
&= (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{N}(y_i-\bar{y})^2\right) + N(\bar{y}-\mu)^2\right]\right) \\
&= (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{(N-1)\,s^2}{2\sigma^2}\right)\exp\left(-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2\right).
\end{aligned}$$

where $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2$ is the sample variance.

2. Adopt a "non-informative" improper prior density $P(\mu,\sigma^2) \propto \sigma^{-2}$ for $\sigma^2 > 0$. Derive the posterior density $P(\mu,\sigma^2|\,\boldsymbol{y})$. Show that:

$$P(\mu|\,\sigma^2,\boldsymbol{y}) = N(\mu|\,\bar{y},\sigma^2/n) \tag{16}$$

and

$$P(\sigma^2|\,\boldsymbol{y}) = \text{Inv-}\chi^2(\sigma^2|n-1,s^2) \tag{17}$$

where the scaled inverse χ^2 distribution has an unnormalised density:

$$\text{Inv-}\chi^2(\theta\,|\,\nu,s^2) \propto \theta^{(-\nu/2+1)}\exp(-\nu s^2/(2\theta)). \tag{18}$$

Solution: For $\sigma^2 > 0$, we have the posterior

$$\begin{aligned}
P(\mu,\sigma^2|\,\boldsymbol{y}) &\propto P(\boldsymbol{y}|\,\mu,\sigma^2)\times\sigma^{-2} \\
&\propto (\sigma^2)^{-N/2-1}\exp\left(-\frac{(N-1)\,s^2}{2\sigma^2}\right)\exp\left(-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2\right).
\end{aligned}$$

By inspection, we see that for any fixed σ^2, the conditional posterior depends on μ only through the factor

$$P(\mu|\,\sigma^2,\boldsymbol{y}) \propto \exp\left(-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2\right).$$

The normalisation constant can be inferred by requiring a Gaussian integral to integrate to unity. Finally,

$$P(\mu|\,\sigma^2,\boldsymbol{y}) = N(\mu|\,\bar{y},\sigma^2/N)$$

The marginal posterior of σ^2 can be computed by integrating out μ from the joint posterior.

$$\begin{aligned} P(\sigma^2|\,\boldsymbol{y}) &\propto (\sigma^2)^{-N/2-1} \exp\left(-\frac{(N-1)\,s^2}{2\sigma^2}\right) \int \exp\left(-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2\right) d\mu \\ &\propto (\sigma^2)^{-N/2-1} \exp\left(-\frac{(N-1)\,s^2}{2\sigma^2}\right) (2\pi\sigma^2/N)^{1/2} \\ &\propto (\sigma^2)^{-N/2-1/2} \exp\left(-\frac{(N-1)\,s^2}{2\sigma^2}\right) \\ &= \textbf{Inv-}\chi^2(\sigma^2\,|\,n-1, s^2) \end{aligned}$$

where the normalisation constant can be inferred to be the same as that normalising the inverse χ^2 distribution by the fact that they expression have the same functional form with respect to $\theta = \sigma^2$.

3. Show that the marginal $P(\mu|\,\boldsymbol{y})$ is a t-distribution and derive its parameters. A t-random variable has unnormalised density:

$$t_\nu(\theta|\,\mu, \sigma^2) \propto \left[1 + \frac{1}{\nu}\Big(\frac{\theta-\mu}{\sigma}\Big)^2\right]^{-(\nu+1)/2}. \tag{19}$$

Solution: The marginal posterior of μ can be obtained by integration:

$$P(\mu|\,\boldsymbol{y}) \propto \int_0^\infty (\sigma^2)^{-N/2-1} \exp\left(-\frac{(N-1)\,s^2}{2\sigma^2}\right) \exp\left(-\frac{N}{2\sigma^2}(\bar{y}-\mu)^2\right) d\sigma^2.$$

By changing variables to $z = A/2\sigma^2$, where $A = (n-1)s^2 + n(\mu-\bar{y})^2$, we find

$$\begin{aligned} P(\mu|\,\boldsymbol{y}) &\propto \int_0^\infty \left(\frac{A}{2z}\right)^{-n/2-1} \exp(-z)\frac{A}{2z^2}\,dz \\ &\propto A^{-n/2} \int_0^\infty z^{(n-2)/2} \exp(-z)\,dz \end{aligned}$$

The integral is a dimensionless gamma function, which we do not need to compute as it is a constant. Now, we have

$$\begin{aligned} P(\mu|\,\boldsymbol{y}) &\propto A^{-n/2} \\ &\propto \left[(n-1)s^2 + n(\mu-\bar{y})^2\right]^{-n/2} \propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2} \\ &= t_{n-1}(\mu|\,\bar{y}, s^2/n). \end{aligned}$$

Therefore the marginal is a $t-$distribution with those parameters.

4. Now adopt the informative conjugate prior from class:

$$P(\mu, \sigma^2) = P(\mu|\,\sigma^2)P(\sigma^2) = N(\mu|\,\mu_0, \sigma^2/\kappa_0) \times \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \tag{20}$$

where μ_0 is a prior mean, σ_0^2 is the prior scale of the variance, and κ_0 and ν_0 quantify the strength of the prior information in terms of the number of "prior observations". Derive the

posterior $P(\mu, \sigma^2 | \, \boldsymbol{y})$. Show that is of the same form as the prior, and derive its parameters $\mu_n, \kappa_n, \nu_n, \sigma_n^2$.

Solution: The prior density on $\sigma^2 > 0$ can be written as

$$P(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]\right).$$

Then the posterior is the likelihood times the prior:

$$\begin{aligned} P(\mu, \sigma^2 | \, \boldsymbol{y}) &\propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]\right) \\ &\times (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right) \\ &\propto \sigma^{-1}(\sigma^2)^{(\nu_n/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + (n+1)s^2 + \kappa_0(\mu_0 - \mu)^2 + n(\bar{y} - \mu)^2]\right) \end{aligned}$$

where $\nu_n \equiv \nu_0 + n$ is the update of the ν parameter. We expect that the posterior mean in μ will be a weighted average of the prior mean and the sample mean:

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}$$

Thus we examine the terms in the brackets that depend on μ:

$$\begin{aligned} n(\bar{y} - \mu)^2 + \kappa_0(\mu_0 - \mu)^2 &= n(\bar{y} - \mu_n + \mu_n - \mu)^2 + \kappa_0(\mu - \mu_n + \mu_n - \mu_0)^2 \\ &= (n + \kappa_0)(\mu - \mu_n)^2 + n(\bar{y} - \mu_n)^2 + \kappa_0(\mu_n - \mu_0)^2 \\ &+ 2n(\bar{y} - \mu_n)(\mu_n - \mu) + 2\kappa_0(\mu - \mu_n)(\mu_n - \mu_0) \end{aligned}$$

You can show that the terms proportional to $(\mu - \mu_n)$ cancel out, by expanding the definition of μ_n. The remaining terms $n(\bar{y} - \mu_0)^2 + \kappa_0(\mu_n - \mu_0)^2$ can similarly be simplified down to $\kappa_0 n(\bar{y} - \mu_0)^2/(\kappa_0 + n)$. Therefore

$$n(\bar{y} - \mu)^2 + \kappa_0(\mu_0 - \mu)^2 = (n + \kappa_0)(\mu - \mu_n)^2 + \kappa_0 n(\bar{y} - \mu_0)^2/(\kappa_0 + n).$$

Thus the posterior is

$$P(\mu, \sigma^2 | \, \boldsymbol{y}) \propto \sigma^{-1}(\sigma^2)^{(\nu_n/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2]\right).$$

where $\kappa_n \equiv \kappa_0 + n$ and

$$\nu_n\sigma_n^2 \equiv \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2$$

The posterior is of the same form as the conjugate prior, but with updated parameters.

5. Derive the marginal density $P(\mu|\boldsymbol{y})$. What is the limiting form of this density as $n \to \infty$? Use the fact that the t-distribution tends to a Gaussian as its degrees of freedom tends toward infinity. Show that the limiting posterior is independent of the prior.

Solution: We can perform the integral over σ^2 in the same way as above.

$$\begin{aligned} P(\mu|\,\boldsymbol{y}) &\propto \left(1 + \frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2} \\ &= t_{\nu_n}(\mu|\,\mu_n, \sigma_n^2/\kappa_n) \end{aligned}$$

As $n \to \infty$, we have $\nu_n \to n$, $\mu_n \to \bar{y}$, $\sigma_n^2 \to s^2$, $\kappa_n \to n$, so

$$P(\mu|\,\boldsymbol{y}) \to t_n(\mu|\,\bar{y}, s^2/n) \to N(\mu|\,\bar{y}, s^2/n)$$

which is independent of the prior, and dominated by the (sufficient statistics of) the likelihood.