Coursework 4: STAT 570

Philip Pham

November 5, 2018

1. Consider the so-called Neyman-Scott problem in which $Y_{ij} \mid \mu_i, \sigma_2 \sim_{\text{ind}} \mathcal{N}\left(\mu_i, \sigma^2\right), i = 1, \ldots, n, j = 1, 2$.

   (a) Obtain the MLE of σ^2 and show that it is inconsistent. Why does this inconsistency arise in this example?

   Solution: The likelihood is

   $$\begin{aligned} L\left(\mu, \sigma\right) &= \prod_{i=1}^{n} \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y_{ij} - \mu_i\right)^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left[\left(Y_{i1} - \mu_i\right)^2 + \left(Y_{i2} - \mu_i\right)^2\right]\right), \end{aligned} \tag{1}$$

   so the log-likelikehood is

   $$l\left(\mu, \sigma\right) = -n\log\left(2\pi\right) - n\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\left(Y_{i1} - \mu_i\right)^2 + \left(Y_{i2} - \mu_i\right)^2\right]. \tag{2}$$

   Taking the derivative with respect to σ^2, we have

   $$\frac{\partial}{\partial\sigma^2} l\left(\mu, \sigma^2\right) = -\frac{n}{\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left[\left(Y_{i1} - \mu_i\right)^2 + \left(Y_{i2} - \mu_i\right)^2\right]. \tag{3}$$

   Solving Equation 3, where $\frac{\partial}{\partial\sigma^2} l\left(\hat{\mu}, \hat{\sigma}^2\right) = 0$, we have

   $$\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}\left[\left(Y_{i1} - \hat{\mu}_i\right)^2 + \left(Y_{i2} - \hat{\mu}_i\right)^2\right]. \tag{4}$$

   Taking the derivative of Equation 2 with respect to μ_i, we have

   $$\frac{\partial}{\partial\mu_i} l\left(\mu, \sigma^2\right) = \frac{1}{\sigma^2}\left(Y_{i1} + Y_{i2} - 2\mu_i\right). \tag{5}$$

   Solving Equation 5, where $\frac{\partial}{\partial\mu_i} l\left(\hat{\mu}, \hat{\sigma}^2\right) = 0$, we have

   $$\hat{\mu}_i = \frac{Y_{i1} + Y_{i2}}{2}. \tag{6}$$

   Substituting Equation 6 into Equation 4, we have

   $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_{i1} - Y_{i2}}{2}\right)^2. \tag{7}$$

Taking the expected value of Equation 7, we have

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{4n}\sum_{i=1}^{n}\left(\mathbb{E}\left[Y_{i1}^2\right] + \mathbb{E}\left[Y_{i2}^2\right] - 2\mathbb{E}\left[Y_{i1}Y_{i2}\right]\right) \\ &= \frac{1}{4n}\sum_{i=1}^{n}\left(\left(\sigma^2+\mu_i^2\right) + \left(\sigma^2+\mu_i^2\right) - 2\mu_i^2\right) \\ &= \frac{\sigma^2}{2}. \end{aligned} \tag{8}$$

Clearly, $\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2/2 \nrightarrow \sigma^2$, so the estimator is not consistent.
This is because the MLE estimate of σ^2 depends on $\mu_1, \ldots, \mu_n$, so the number of parameters being estimated increases with n. Thus, the model is not well-defined.

(b) Derive the posterior distribution corresponding to the prior

$$\pi\left(\mu_1, \ldots, \mu_n, \sigma^2\right) \propto \sigma^{-n-2} \tag{9}$$

and show that

$$\mathbb{E}\left[\sigma^2 \mid Y\right] = \frac{1}{2(n-1)}\sum_{i=1}^{n}\frac{\left(Y_{i1}-Y_{i2}\right)^2}{2}. \tag{10}$$

Solution: Using the likelihood in Equation 1 and the prior in Equation 9. We have that

$$p\left(\mu, \sigma^2 \mid Y\right) \propto L\left(\mu, \sigma^2\right)\pi\left(\mu_1, \ldots, \mu_n, \sigma^2\right). \tag{11}$$

We have that

$$\begin{aligned} p\left(Y\right) &= \int_0^\infty \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty L\left(\mu, \sigma^2\right)\pi\left(\mu_1, \ldots, \mu_n, \sigma^2\right)\,\mathrm{d}\mu_1 \cdots \mathrm{d}\mu_n\,\mathrm{d}\sigma^2 \\ &= \int_0^\infty \frac{1}{2^n\pi^n\left(\sigma^2\right)^{(3n+2)/2}}\left(\pi\sigma^2\right)^{n/2}\prod_{i=1}^{n}\exp\left(-\frac{1}{4\sigma^2}\left(Y_{i1}-Y_{i2}\right)^2\right)\,\mathrm{d}\sigma^2 \\ &= \int_0^\infty \frac{1}{2^n\pi^{n/2}\left(\sigma^2\right)^{n+1}}\exp\left(-\frac{1}{4\sigma^2}\sum_{i=1}^{n}\left(Y_{i1}-Y_{i2}\right)^2\right)\,\mathrm{d}\sigma^2 \\ &= -\frac{2^n}{\pi^{n/2}}\left(\sum_{i=1}^{n}\left(Y_{i1}-Y_{i2}\right)^2\right)^{-n}\int_\infty^0 u^{n-1}\exp\left(-u\right)\,\mathrm{d}u \\ &= \frac{1}{\pi^{n/2}}\left(\sum_{i=1}^{n}\frac{\left(Y_{i1}-Y_{i2}\right)^2}{2}\right)^{-n}\Gamma(n). \end{aligned} \tag{12}$$

Normalizing Equation 11 with the evidence Equation 12, we have the posterior

$$p\left(\mu, \sigma^2 \mid Y\right) = \frac{\left(\sigma^2\right)^{-(3n+2)/2}}{2^n\pi^{n/2}\Gamma(n)}\left(\sum_{i=1}^{n}\frac{\left(Y_{i1}-Y_{i2}\right)^2}{2}\right)^n\prod_{i=1}^{n}\prod_{j=1}^{2}\exp\left(-\frac{1}{2\sigma^2}\left(Y_{ij}-\mu_i\right)^2\right). \tag{13}$$

Marginalizing μ in Equation 13, we get

$$\begin{aligned} p\left(\sigma^2 \mid Y\right) &= \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty p\left(\mu, \sigma^2 \mid Y\right)\,\mathrm{d}\mu_1 \cdots \mathrm{d}\mu_n \\ &= \frac{\left(\sigma^2\right)^{-n-1}}{2^n\Gamma(n)}\left(\sum_{i=1}^{n}\frac{\left(Y_{i1}-Y_{i2}\right)^2}{2}\right)^n\exp\left(-\frac{1}{4\sigma^2}\sum_{i=1}^{n}\left(Y_{i1}-Y_{i2}\right)^2\right). \end{aligned} \tag{14}$$

Taking the expectation over the distribution in Equation 14, we have that

$$\begin{aligned}
\mathbb{E}\left[\sigma^2 \mid Y\right] &= \int_0^\infty \sigma^2 p\left(\sigma^2 \mid Y\right) \mathrm{d}\sigma^2 \\
&= \frac{1}{\Gamma(n)} \int_0^\infty \left(\frac{1}{4\sigma^2} \sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2\right)^n \exp\left(-\frac{1}{4\sigma^2} \sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2\right) \\
&= \frac{\sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2}{4\Gamma(n)} \int_0^\infty u^{n-1-1} \exp\left(u\right) \mathrm{d}u \\
&= \frac{\Gamma(n-1)}{2\Gamma(n)} \sum_{i=1}^n \frac{\left(Y_{i1} - Y_{i2}\right)^2}{2} \\
&= \frac{1}{2(n-1)} \sum_{i=1}^n \frac{\left(Y_{i1} - Y_{i2}\right)^2}{2}, \qquad (15)
\end{aligned}$$

which is the desired result.

(c) Hence, using Equation 15, show that $\mathbb{E}\left[\sigma^2 \mid Y\right] \to \sigma^2/2$ as $n \to \infty$, so that the posterior mean is inconsistent.

Solution: From Equation 15, we have that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\sigma^2 \mid Y\right] &= \lim_{n\to\infty} \frac{1}{2(n-1)} \sum_{i=1}^n \frac{\mathbb{E}\left[\left(Y_{i1} - Y_{i2}\right)^2\right]}{2} \\
&= \lim_{n\to\infty} \frac{1}{2(n-1)} \sum_{i=1}^n \frac{\mathrm{Var}\left(Y_{i1} - Y_{i2}\right)}{2} \\
&= \lim_{n\to\infty} \frac{n\sigma^2}{2(n-1)} \\
&= \frac{\sigma^2}{2} \neq \sigma^2, \qquad (16)
\end{aligned}$$

so the posterior mean is inconsistent.

(d) Examine the posterior distribution corresponding to the prior

$$\pi\left(\mu_1, \ldots, \mu_n \sigma^2\right) \propto \sigma^{-2}. \qquad (17)$$

Solution: If we use Equation 17, Equation 12 becomes

$$\begin{aligned}
p\left(Y\right) &= \int_0^\infty \frac{1}{2^n \pi^{n/2} \left(\sigma^2\right)^{n/2+1}} \exp\left(-\frac{1}{4\sigma^2} \sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2\right) \mathrm{d}\sigma^2 \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\pi^{n/2}} \left(\sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2\right)^{-n/2}. \qquad (18)
\end{aligned}$$

With Equation 18, the posterior becomes

$$p\left(\mu, \sigma^2 \mid Y\right) = \frac{\left(\sigma^2\right)^{-n-1}}{2^n \pi^{n/2} \Gamma(n/2)} \left(\sum_{i=1}^n \frac{\left(Y_{i1} - Y_{i2}\right)^2}{2}\right)^n \prod_{i=1}^n \prod_{j=1}^2 \exp\left(-\frac{1}{2\sigma^2} \left(Y_{ij} - \mu_i\right)^2\right). \qquad (19)$$

Marginalizing Equation 19 over μ, we have

$$p\left(\sigma^2 \mid Y\right) = \frac{1}{\sigma^2 \Gamma\left(n/2\right)} \left(\frac{\sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2}{4\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n \left(Y_{i1} - Y_{i2}\right)^2}{4\sigma^2}\right). \qquad (20)$$

Length (mm)	0	1	2	3	4	5	6	7	8	9	10	11	12
1	2.247	2.640	2.842	2.908	3.099	3.126	3.245	3.328	3.355	3.383	3.572	3.581	3.681
10	1.901	2.132	2.203	2.228	2.257	2.350	2.361	2.396	2.397	2.445	2.454	2.454	2.474
20	1.312	1.314	1.479	1.552	1.700	1.803	1.861	1.865	1.944	1.958	1.966	1.997	2.006
50	1.339	1.434	1.549	1.574	1.589	1.613	1.746	1.753	1.764	1.807	1.812	1.840	1.852

Table 1: Failure stress data for four groups of fibers.

Equation 20 is quite similar to Equation 14, but with n replaced by $n/2$ in the gamma function and the exponent of the sum of squares.

(e) Is the posterior mean for σ^2 consistent in this case?

Solution: Yes. Taking the expectation with Equation 20, we have

$$\begin{aligned}
\mathbb{E}\left[\sigma^2 \mid Y\right] &= \int_0^\infty p\left(\sigma^2 \mid Y\right) \, \mathrm{d}\sigma^2 \\
&= \frac{1}{4\Gamma\left(n/2\right)} \sum_{i=1}^{n} \left(Y_{i1} - Y_{i2}\right)^2 \int_0^\infty u^{n/2-1-1} \exp\left(-u\right) \, \mathrm{d}u \\
&= \frac{\Gamma\left(n/2-1\right)}{4\Gamma\left(n/2\right)} \sum_{i=1}^{n} \left(Y_{i1} - Y_{i2}\right)^2 \\
&= \frac{1}{2\left(n/2-1\right)} \sum_{i=1}^{n} \frac{\left(Y_{i1} - Y_{i2}\right)^2}{2} \\
&= \frac{1}{n-2} \sum_{i=1}^{n} \frac{\left(Y_{i1} - Y_{i2}\right)^2}{2}. \qquad (21)
\end{aligned}$$

Taking the limit of Equation 21, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\sigma^2 \mid Y\right] &= \lim_{n\to\infty} \frac{1}{n-2} \sum_{i=1}^{n} \frac{\mathbb{E}\left[\left(Y_{i1} - Y_{i2}\right)^2\right]}{2} \\
&= \lim_{n\to\infty} \frac{n}{(n-2)} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{Var}\left(Y_{i1} - Y_{i2}\right)}{2} \\
&= \lim_{n\to\infty} \frac{n}{(n-2)} \frac{1}{n} \sum_{i=1}^{n} \frac{2\sigma^2}{2} \\
&= \lim_{n\to\infty} \frac{n}{(n-2)} \sigma^2 \\
&= \sigma^2, \qquad (22)
\end{aligned}$$

so the posterior mean is consistent when the prior doesn't depend on n.

2. The data in Table 1 contain data on a typical reliability experiment and give the failure stresses (in GPa) of four samples of carbon fibers of lengths 1, 10, 20 and 50mm.

(a) Consider a Bayesian analysis with an exponential likelihood and a gamma prior, $\lambda \sim \mathrm{Gamma}(a, b)$. Derive the form of the posterior distribution for λ.

Solution: Suppose that we observe independent and identically distributed $Y_i \sim \mathrm{Exponential}(\lambda)$ for $i = 1, \ldots, n$. Let $Y = (Y_1, \ldots, Y_n)$. Then, we have

Length (mm)	a'	b'	Mean	Standard error
1	16.634063	48.275125	0.344568	0.084484
10	16.634063	37.320125	0.445713	0.109284
20	16.634063	30.025125	0.554005	0.135836
50	16.634063	28.940125	0.574775	0.140928

Table 2: The results of updating the prior belief in Part 2b with the data.

the likelihood function

$$L(\lambda) = p\left(Y \mid \lambda\right) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} Y_i\right). \tag{23}$$

From Equation 23, we have the posterior

$$\begin{aligned} p\left(\lambda \mid Y\right) &\propto p\left(Y \mid \lambda\right) p\left(\lambda\right) \\ &\propto \left(\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} Y_i\right)\right) \left(\frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp\left(-b\lambda\right)\right) \\ &\propto \lambda^{a+n-1} \exp\left(-\left(b + \sum_{i=1}^{n} Y_i\right)\lambda\right), \end{aligned} \tag{24}$$

which equal to the Gamma probability density function up to a constant factor independent of λ. Thus, we have that $\lambda \mid Y \sim \Gamma\left(a+n, b+\sum_{i=1}^{n} Y_i\right)$, and

$$p\left(\lambda \mid Y\right) = \frac{\left(b + \sum_{i=1}^{n} Y_i\right)^{a+n}}{\Gamma\left(a+n\right)} \lambda^{a+n-1} \exp\left(-\left(b + \sum_{i=1}^{n} Y_i\right)\lambda\right). \tag{25}$$

(b) Choose a and b so that the prior probability that λ lies between 0.05 and 1 is 0.95.

Solution: For a specified mean μ, we specify our prior as the gamma distribution $\text{Gamma}\left(k\mu, k\right)$ for some k. Let F be the cumulative distribution function. We choose k such that

$$F\left(1\right) - F\left(0.05\right) = 0.95. \tag{26}$$

Choosing $a \approx 3.634$ and $b \approx 7.268$ satisfies Equation 26. These values were obtained numerically in `failure_stresses.ipynb`.

(c) Obtain the posterior means and posterior standard deviations for λ and give histogram representations of $p\left(\lambda \mid y\right)$ for each of the groups in Table 1. Compare inference with the frequentist analyses. Also give histogram representations of the posterior for λ^{-1}, again for each of the four groups.

Solution: The posterior parameters, means, and standard deviations can be seen in Table 2. The first two columns are the shape and rate parameters of the gamma posterior, respectively.

The means are similar to the estimates from frequentist analyses in the previous homework but are drawn towards the prior mean of 0.5. The

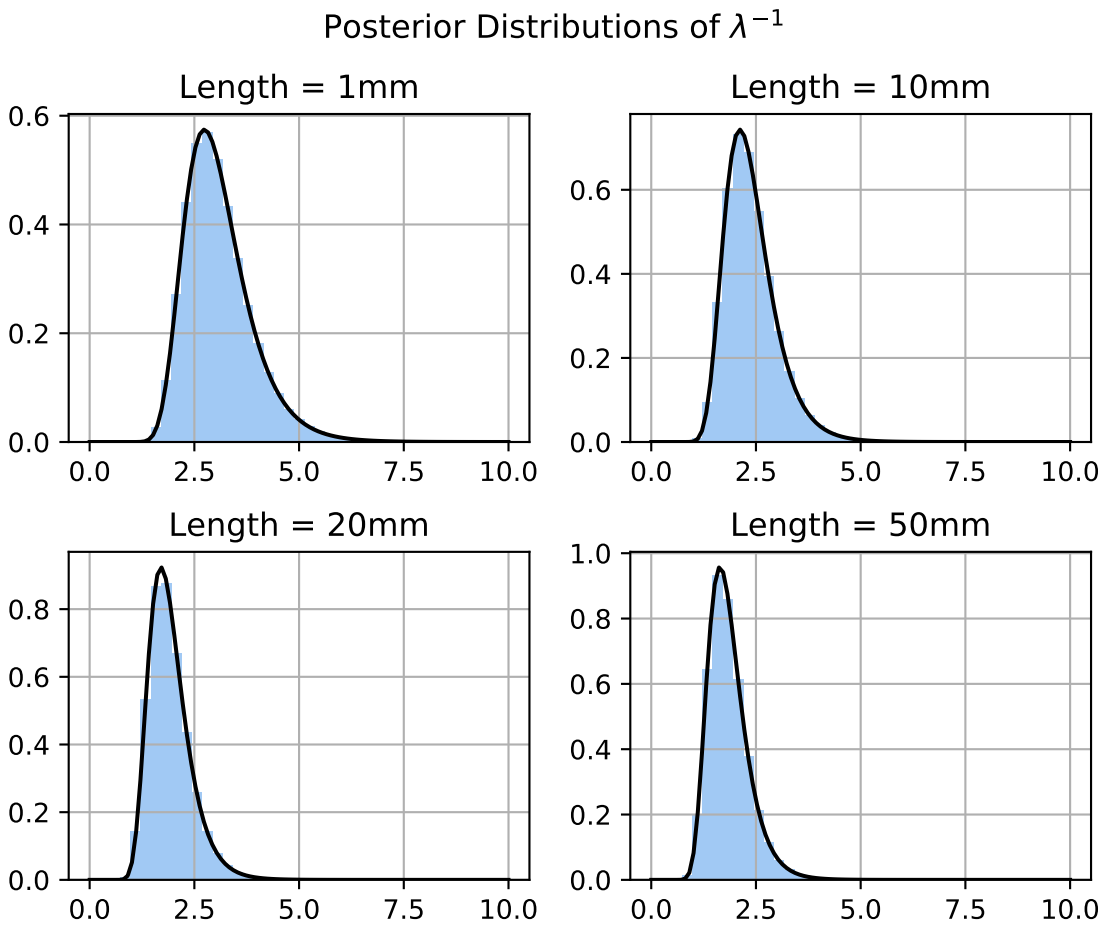

Figure 1: Histograms and probability density for samples drawn from the posteriors for λ^{-1}.

standard errors are slightly smaller tahn those obtained from fitting the exponential model but are much larger than either the quasi-likelihood model or sandwich estimates.

For the posterior distribution of λ^{-1}, we have that

$$p\left(\lambda^{-1} \mid Y\right) = p\left(\lambda \mid Y\right) \left| \frac{\mathrm{d}}{\mathrm{d}\lambda^{-1}} \left(\lambda^{-1}\right)^{-1} \right| = p\left(\lambda \mid Y\right) \lambda^2$$
$$= \frac{\left(b + \sum_{i=1}^{n} Y_i\right)^{a+n}}{\Gamma\left(a+n\right)} \lambda^{a+n+1} \exp\left(- \left(b + \sum_{i=1}^{n} Y_i \right) \lambda \right), \qquad (27)$$

where we used Equation 25 and transformation of random variables, so $\lambda^{-1} \mid Y \sim \text{InverseGamma}\left(a+n, b+\sum_{i=1}^{n} Y_i\right)$.

Theoretical histograms along with their density are plotted in Figure 1.

Unlike the MLE, the posterior mean is not invariant under reparameterization:

$$\mathbb{E}\left[\lambda^{-1} \mid Y\right] = \frac{b'}{a'-1} \gtrless \frac{b'}{a'} = \frac{1}{\mathbb{E}\left[\lambda \mid Y\right]}.$$

See `failure_stresses.ipynb` for code to reproduce plots.

3. In this question we will consider inference when the sampling model is multivariate hypergeometric. Suppose a population contains objects of K different types, with $X_1, \ldots, X_K$ being the number of each type, $\sum_{k=1}^{K} X_k = N$. A simple random sample of size n is taken and the number of each type, $Y_1, \ldots, Y_K$, is recorded (so that $\sum_{k=1}^{K} y_k = n$).

   An obvious model for $Y_1, \ldots, Y_K$, is the multivariate hypergeometric distribution:

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_K = y_K \mid x_1, \ldots, x_K\right) = \frac{\prod_{k=1}^{K} \binom{x_k}{y_k}}{\binom{N}{n}}, \tag{28}$$

   with means and variances:

$$\mathbb{E}\left[Y_k \mid x_k\right] = n\frac{x_k}{N} \tag{29}$$

$$\mathrm{Var}\left(Y_k \mid x_k\right) = n\frac{x_k}{N}\left(1 - \frac{x_k}{N}\right)\frac{N-n}{N-1}. \tag{30}$$

   Suppose we take a sample from a population of K distinct objects, and record $y_1, \ldots, y_K$, but the numbers $X_1, \ldots, X_K$ are unknown (but N is known).

   (a) Using Equation 29, write down an estimator for X_k, $k = 1, \ldots, K$. We will refer to this a method of moments estimator. Using Equation 30, give a form for the variance of this estimator, along with the estimator of this variance.

   Solution: A simple estimator for X_k can be obtained from Equation 29 by substituting $\mathbb{E}\left[Y_k \mid x_k\right]$ with the observed y_k:

$$y_k = n\frac{\hat{x}_k}{N} \implies \hat{x}_k = \frac{N}{n} y_k. \tag{31}$$

   If we treat $\hat{x}_k$ like a random variable, then the variance of our estimator is

$$\begin{aligned}
\mathrm{Var}\left(\hat{x}_k\right) &= \mathrm{Var}\left(\frac{N}{n} Y_k\right) = \left(\frac{N}{n}\right)^2 \mathrm{Var}\left(Y_k\right) \\
&= \left(\frac{N}{n}\right)^2 \left(n\frac{x_k}{N}\left(1 - \frac{x_k}{N}\right)\frac{N-n}{N-1}\right) \\
&= x_k \left(\frac{N}{n}\right)\left(1 - \frac{x_k}{N}\right)\frac{N-n}{N-1}
\end{aligned} \tag{32}$$

   Since we don't know x_k, we substitute x_k with $\hat{x}_k$ to get the estimated variance:

$$\begin{aligned}
\hat{\mathrm{Var}}\left(\hat{x}_k\right) &= \hat{x}_k \left(\frac{N}{n}\right)\left(1 - \frac{\hat{x}_k}{N}\right)\frac{N-n}{N-1} \\
&= y_k \left(\frac{N}{n}\right)^2 \left(1 - \frac{y_k}{n}\right)\frac{N-n}{N-1} \\
&= (N-n)\frac{y_k}{n}\left(1 - \frac{y_k}{n}\right)\left(\frac{N^2}{n\left(N-1\right)}\right)
\end{aligned} \tag{33}$$

   (b) We now consider a Bayesian approach to inference. Consider a multinomial distribution for counts $X_1, \ldots, X_K$,

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_K = x_K \mid p_1, \ldots, p_K\right) = \frac{N!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k}, \tag{34}$$

with $p_k > 0$ and $\sum_{k=1}^K p_k = 1$. Show that the Dirichlet:

$$\pi\left(p_1, \ldots, p_K\right) = \frac{\Gamma\left(\alpha_+\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k\right)} \prod_{k=1}^K p_k^{\alpha_k - 1}, \tag{35}$$

where $\alpha_k > 0$, $k = 1, \ldots, K$ and $\alpha_+ = \sum_{k=1}^K \alpha_k$ is the conjugate distribution to the multinomial sampling model.

Solution: We use the definition of conjugate distributions and show that $(p_1, \ldots, p_K) \mid (X_1 = x_1, \ldots, X_K = x_k)$ is of the same distribution family as $(p_1, \ldots, p_K)$.
We have that

$$\begin{aligned} p\left(p_1, \ldots, p_K \mid x_1, \ldots, x_k\right) &\propto \mathbb{P}\left(x_1, \ldots, x_K \mid p_1, \ldots, p_K\right) \pi\left(p_1, \ldots, p_K\right) \\ &\propto \prod_{k=1}^K p_k^{x_k + \alpha_k - 1}, \end{aligned} \tag{36}$$

which we recognize as the form of the Dirichlet family.
Thus, we have that

$$p\left(p_1, \ldots, p_K \mid x_1, \ldots, x_k\right) = \frac{\Gamma\left(\alpha_+ + N\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k + x_k\right)} \prod_{k=1}^K p_k^{\alpha_k + x_k - 1}, \tag{37}$$

so the Dirichlet distribution is the conjugate distribution to the multinomial sampling model.

(c) The compound multinomial distribution, $\mathrm{CMult}\left(N, \alpha\right)$, is defined as

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_K = x_K\right) = \frac{N! \Gamma\left(\alpha_+\right)}{\Gamma\left(N + \alpha_+\right)} \prod_{k=1}^K \frac{\Gamma\left(x_k + \alpha_k\right)}{x_k! \Gamma\left(\alpha_k\right)} \tag{38}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)$. Show that the prior predictive distribution, obtained as the marginal distribution when the likelihood is Equation 34 and the prior is Equation 35, is of the compound multinomial form with parameters that you should identify.

Solution: Let $\mathbf{p} = (p_1, \ldots, p_K)$. We simply integrate:

$$\begin{aligned} \mathbb{P}\left(X_1 = x_1, \ldots, X_K = x_K\right) &= \int \mathbb{P}\left(x_1, \ldots, x_K \mid p_1, \ldots, p_K\right) \pi\left(p_1, \ldots, p_K\right) \, \mathrm{d}\mathbf{p} \\ &= \frac{N!}{\prod_{k=1}^K x_k!} \frac{\Gamma\left(\alpha_+\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k\right)} \int \prod_{k=1}^K p_k^{x_k + \alpha_k - 1} \, \mathrm{d}\mathbf{p} \\ &= \frac{N! \Gamma\left(\alpha_+\right)}{\prod_{k=1}^K x_k! \Gamma\left(\alpha_k\right)} \left(\frac{\prod_{k=1}^K \Gamma\left(x_k + \alpha_k\right)}{\Gamma\left(N + \alpha_+\right)} \right) \\ &= \frac{N! \Gamma\left(\alpha_+\right)}{\Gamma\left(N + \alpha_+\right)} \prod_{k=1}^K \frac{\Gamma\left(x_k + \alpha_k\right)}{x_k! \Gamma\left(\alpha_k\right)}, \end{aligned}$$

where we have used that the Dirichlet probability density function must integrate to 1 to compute the integral.

(d) Find the mean $\mathbb{E}[X_k]$ and variance $\mathrm{Var}(X_k)$, $k = 1, \ldots, K$, of a compound multinomial distribution.

Solution: We can use the law of total expectation for the mean:

$$\mathbb{E}[X_k] = \mathbb{E}_\pi[\mathbb{E}[X_k \mid p_k]] = \mathbb{E}_\pi[Np_k] = N\frac{\alpha_k}{\alpha_+}, \tag{39}$$

where we have taken advantage of the known means of the multinomial and Dirichlet distributions.
For the variance we can use the law of total variance:

$$\begin{aligned}
\mathrm{Var}(X_k) &= \mathbb{E}_\pi[\mathrm{Var}(X_k \mid p_k)] + \mathrm{Var}(\mathbb{E}[X_k \mid p_k]) \\
&= \mathbb{E}_\pi[Np_k(1-p_k)] + \mathrm{Var}(Np_k) \\
&= N\left(\mathbb{E}_\pi[p_k] - \mathbb{E}_\pi\left[p_k^2\right]\right) + N^2\mathrm{Var}(p_k) \\
&= N\left(\frac{\alpha_k}{\alpha_+} - \frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)} - \left(\frac{\alpha_k}{\alpha_+}\right)^2\right) + N^2\frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)} \\
&= N\frac{\alpha_k}{\alpha_+}\left(1 - \frac{1+\alpha_k}{1+\alpha_+}\right) + N\frac{\alpha_k}{\alpha_+}\left(N\frac{\alpha_+ - \alpha_k}{\alpha_+(\alpha_+ + 1)}\right) \\
&= N\frac{\alpha_k}{\alpha_+}\left(\frac{\alpha_+(\alpha_+ - \alpha_k)}{\alpha_+(1+\alpha_+)} + \frac{N(\alpha_+ - \alpha_k)}{\alpha_+(1+\alpha_+)}\right) \\
&= N\frac{\alpha_k}{\alpha_+}\left(\frac{(\alpha_+ - \alpha_k)(N+\alpha_+)}{\alpha_+(1+\alpha_+)}\right) \\
&= N\frac{\alpha_k}{\alpha_+}\left(1 - \frac{\alpha_k}{\alpha_+}\right)\left(\frac{N+\alpha_+}{1+\alpha_+}\right).
\end{aligned} \tag{40}$$

(e) Let $W_k = X_k - y_k$ represent the unobserved counts, $k = 1, \ldots, K$. Show that the posterior distribution $\mathbb{P}(W_1, \ldots, W_K | y_1, \ldots, y_K)$ is compound multinomial $\mathrm{CMult}(N - n, \alpha + y)$, where $y = (y_1, \ldots, y_K)$.

Solution: Given y, $W = (W_1, \ldots, W_K)$ is just a reparameterization of $X = (X_1, \ldots, X_k)$. However, the domain is different since $W_k \geq 0$, so we need a normalization constant.
Using Equation 38 as our prior and Equation 28 as our likelihood, we can compute the posterior

$$\begin{aligned}
&\mathbb{P}(W_1 = w_1, \ldots, W_K = w_K \mid y_1, \ldots, y_K) \\
&\quad\propto \mathbb{P}(y_1, \ldots, y_K \mid X)\,\mathbb{P}(X_1 = w_1 + y_1, \ldots, X_K = w_K + y_K) \\
&\quad\propto \left(\frac{\prod_{k=1}^K \binom{w_k + y_k}{y_k}}{\binom{N}{n}}\right)\left(\frac{N!\Gamma(\alpha_+)}{\Gamma(N+\alpha_+)}\prod_{k=1}^K \frac{\Gamma(w_k + y_k + \alpha_k)}{(w_k + y_k)!\Gamma(\alpha_k)}\right) \\
&\quad\propto \frac{(N-n)!n!\Gamma(\alpha_+)}{\Gamma(N+\alpha_+)}\prod_{k=1}^K \frac{\Gamma(w_k + y_k + \alpha_k)}{w_k! y_k!\Gamma(\alpha_k)} \\
&\quad\propto \prod_{k=1}^K \frac{\Gamma(w_k + y_k + \alpha_k)}{w_k!},
\end{aligned} \tag{41}$$

nn which we recognize as the from the compound multionomial distribution. Since Equation 41 must sum to 1 to be a proper probability distribution

over the support of the prior $\sum_{k=1}^{K} W_k = N - n$, where W_k are nonnegative integers, we get that

$$\begin{aligned}
&\mathbb{P}\left(W_1 = w_1, \ldots, W_K = w_K \mid y_1, \ldots, y_K\right) \\
&\quad = \frac{(N-n)!\Gamma\left(n + \alpha_+\right)}{\Gamma\left((N-n) + (n+\alpha_+)\right)} \prod_{k=1}^{K} \frac{\Gamma\left(w_k + (y_k + \alpha_k)\right)}{w_k!\Gamma\left(y_k + \alpha_k\right)},
\end{aligned} \tag{42}$$

which is the $\text{CMult}\left(N - n, y + \alpha\right)$ distribution.

(f) Write down the posterior mean and posterior variance of X_k, $k = 1, \ldots, K$. Comment on the case when $\alpha_k = 0$, $k = 1, \ldots, K$.

Solution: Using the results from Equation 42, 39, and 40, we have

$$\begin{aligned}
\mathbb{E}\left[X_k \mid y_1, \ldots, y_K\right] &= \mathbb{E}\left[W_k \mid y_1, \ldots, y_K\right] + y_k \\
&= y_k + (N-n)\frac{y_k + \alpha_k}{n + \alpha_+}. \qquad (43) \\
\text{Var}\left(X_k \mid y_1, \ldots, y_K\right) &= \text{Var}\left(W_k \mid y_1, \ldots, y_K\right) \qquad (44) \\
&= (N-n)\frac{y_k + \alpha_k}{n + \alpha_+}\left(1 - \frac{y_k + \alpha_k}{n + \alpha_+}\right)\left(\frac{N + \alpha_+}{1 + n + \alpha_+}\right).
\end{aligned}$$

When $\alpha = \mathbf{0}$, the expectation and variance are similar to those of the multinomial distribution with $p_k = y_k/n$, but the variance has the additional $\frac{N}{1+n}$ factor. These are similar to the method of the moments estimators in Equations 31 and 33. The mean estimate is same as posterior mean, and the variance has a $\frac{N^2}{n(N-1)}$ has is quite close to $\frac{N}{1+n}$, especially for large values of N and n.

(g) A certain infectious disease can be caused by one of three different pathogens, A, B, or C. Over a 1 year period population surveillance is carried out, and 750 individuals are observed to be infected. A random sample of 60 cases is selected for labtesting, i.e., to determine the pathogen responsible. Of these 60 selected cases,the numbers who were infected by pathogens A, B, C, were 42, 18, 0, respectively. We wish to estimate the numbers of the total population of cases that were infected by each of the pathogens.

i. Calculate the method of moments estimators and the associated standard errors.

Solution: Let X_1, X_2, and X_3 correspond to the total population with pathogens A, B, and C, respectively. Let Y_1, Y_2, and Y_3 be the corresponding samples.

We substitute $N = 750$, $n = 60$, $y_1 = 42$, $y_2 = 18$, and $y_3 = 0$ into Equations 31 and 32.

For the estimated means, we have

$$\begin{aligned}
\hat{x}_1 &= 750\frac{42}{60} = 525 \\
\hat{x}_2 &= 750\frac{18}{60} = 225 \\
\hat{x}_3 &= 750\frac{0}{60} = 0. \qquad (45)
\end{aligned}$$

For estimated standard errors, we have

$$\begin{aligned}
\sqrt{\hat{\text{Var}}\left(\hat{x}_1\right)} &\approx 42.587 \\
\sqrt{\hat{\text{Var}}\left(\hat{x}_2\right)} &\approx 42.587 \\
\sqrt{\hat{\text{Var}}\left(\hat{x}_3\right)} &= 0.
\end{aligned} \tag{46}$$

ii. Calculate the Bayesian posterior mean and posterior standard deviation, with prior specification, $\alpha_k = 1$, $k = 1, \ldots, K$. Which estimates are the most reasonable?

Solution: Here, we substitute into Equations 43 and 44.

We have that

$$\begin{aligned}
\mathbb{E}\left[X_1 \mid y\right] &= 42 + (690)\,\frac{42+1}{60+3} = 42 + \frac{9890}{21} \approx 512.952 \\
\mathbb{E}\left[X_2 \mid y\right] &= 18 + (690)\,\frac{18+1}{60+3} = 18 + \frac{4370}{21} \approx 226.095 \\
\mathbb{E}\left[X_3 \mid y\right] &= 0 + (690)\,\frac{1}{63} = \frac{230}{21} \approx 10.952,
\end{aligned} \tag{47}$$

and

$$\begin{aligned}
\sqrt{\text{Var}\left[X_1 \mid y\right]} &\approx 41.941 \\
\sqrt{\text{Var}\left[X_2 \mid y\right]} &\approx 41.352 \\
\sqrt{\text{Var}\left[X_3 \mid y\right]} &\approx 11.261.
\end{aligned} \tag{48}$$

The estimates in Equations 47 and 48 are similar to those in Equations 45 and 46 for X_1 and X_2 but are pulled slightly closer together since the prior specifies that the pathogens occur in about equal proportions.

The estimate for X_3 is much different, however. In Equations 45 and 46, the estimate and standard error is 0 because there were no observations. If we know that at least one person carries pathogen C, this estimate is deeply disturbing. It could simply be that no one in our sample carried the pathogen, so the Bayesian estimates are most reasonable.