(1) **[45 points]** Suppose we have a sample of employment x_i, $i = 1, \ldots N$ at N firms in a particular county and industry and we wish to model them as i.i.d. draws from a common distribution. Suppose further that we assume the distribution is exponential, meaning it has probability density function (pdf) ae^{-ax_i}, with a an unknown parameter.

(a) Recall that the pdf of a Gamma(n, α) random variable z is $\alpha^n z^{n-1} e^{-\alpha z}/\Gamma(n)$ and that the sum of independent Gamma$(n, 1)$ and Gamma$(m, 1)$ variates is distributed as Gamma$(n + m, 1)$. Also that the Γ function satisfies $n\Gamma(n) = \Gamma(n + 1)$.
What is the posterior pdf for a in our sample of N firm employment levels, under a flat prior on a?

(b) What is the posterior expectation of the mean $1/a$ of the x_i distribution?

(c) Show that the sample mean $\bar{x} = \sum x_i/N$ is a (frequentist) unbiased estimate of $1/a$.

(d) Show that $a \sum_{i=1}^{n} x_i$ is a pivotal statistic and explain how to use it to generate a 90% confidence interval for $1/a$.

(e) How could you generate a 90% posterior probability posterior credibility interval for $1/a$ from this sample?

(f) Suppose that contrary to what was assumed above, the data set from which this sample is drawn, covering many counties and industries besides the industry and county in this sample, top-codes large firms. That is, even though the true distribution of employment levels across firms in the sample for any county and industry is i.i.d. exponential as we have assumed, the data would record "500" for any observed true employment level of 500 or over. However in the particular county and industry sample at hand there are no $x_i = 500$ observations. Does recognizing the top-coding in the sampling scheme affect the Bayesian posterior distribution you derived above? Does it affect the claim of unbiasedness for the sample mean or the calculation of the frequentist confidence interval? Explain your answers.

(2) **[15 points]** Suppose y is distributed as $N(1,1)$. Derive the pdf of $z = 1/y$ and sketch its shape. What are the mean and variance of z?

(3) **[30 points]** Consider the linear regression equation

$$y_i = a + bx_i + \varepsilon_i \,. \qquad (*)$$

A common practice is to estimate a and b in an equation like this by ordinary least squares and use the "heteroskedasticity-consistent" estimate of the standard errors of the estimated coefficients to produce confidence intervals.

(a) What is the formula for the heteroskedasticity-consistent coefficient covariance matrix estimate? What assumptions, beyond i.i.d. data, are needed to justify it?

(b) Suppose that in fact the data are i.i.d. across i with

$$E[y_i \mid x_i] = x_i/(1+x_i) \,.$$

In other words, there is a nonlinear regression relation between y_i and x_i. Does this violate the assumptions that justify fitting a linear relation by OLS and using the heteroskedasticty-consistent covariance matrix for the estimates? Why or why not?

(c) Suppose instead that in fact the data are i.i.d. across i and ε_i in $(*)$ satisfies $\varepsilon_i \mid x_i \sim N(0,1)$. Also assume that x_i is drawn from a distribution whose pdf declines at the rate $x^{-2.5}$ as $x \to \infty$. Does this violate the assumptions that justify fitting a linear relation by OLS and using the heteroskedasticty-consistent covariance matrix for the estimates? Why or why not? Would use of the usual $\hat{\sigma}^2(X'X)^{-1}$ covariance matrix be justified under these assumptions? Why or why not?