Problem Set #2

MACS 40200, Dr. Evans
Due Monday, Jan. 22 at 1:30pm

1. **Health claim amounts and the GB family of distributions (10 points).** For this problem, you will use 10,619 health claims amounts from a fictitious sample of households. These data are in a single column of the text file `clms.txt` in the PS2 folder. Health claim amounts are reported in U.S. dollars. For this exercise, you will need to use the generalized beta family of distributions shown in the figure in Section 7 of your MLE Jupyter notebook.

   (a) (1 points) Calculate and report the mean, median, maximum, minimum, and standard deviation of monthly health expenditures for these data. Plot two histograms of the data in which the y-axis gives the percent of observations in the particular bin of health expenditures and the x-axis gives the value of monthly health expenditures. Use percentage histograms in which the height of each bar is the percent of observations in that bin (see instructions in Jupyter notebook `PythonVisualize.ipynb` in Section 1.2). In the first histogram, use 1,000 bins to plot the frequency of all the data. In the second histogram, use 100 bins to plot the frequency of only monthly health expenditures less-than-or-equal-to \$800 ($x_i \leq 800$). Adjust the frequencies of this second histogram to account for the observations that you have not displayed ($x_i > 800$). That is, the heights of the histogram bars in the second histogram should not sum to 1 because you are only displaying a fraction of the data. Comparing the two histograms, why might you prefer the second one?

   (b) (2 points) Using MLE, fit the gamma $GA(x; \alpha, \beta)$ distribution to the individual observation data. Use $\beta_0 = Var(x)/E(x)$ and $\alpha_0 = E(x)/\beta_0$ as your initial guess.[1] Report your estimated values for $\hat{\alpha}$ and $\hat{\beta}$, as well as the value of the maximized log likelihood function $\ln \mathcal{L}(\hat{\theta})$. Plot the second histogram from part (a) overlayed with a line representing the implied histogram from your estimated gamma (GA) distribution.

   (c) (2 points) Using MLE, fit the generalized gamma $GG(x; \alpha, \beta, m)$ distribution to the individual observation data. Use your estimates for α and β from part(b), as well as $m = 1$, as your initial guess. Report your estimated values for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{m}$, as well as the value of the maximized log likelihood function $\ln \mathcal{L}$. Plot the second histogram from part (a) overlayed with a line representing the implied histogram from your estimated generalized gamma (GG) distribution.

[1] These initial guesses come from the property of the gamma (GA) distribution that $E(x) = \alpha\beta$ and $Var(x) = \alpha\beta^2$.

(d) (2 points) Using MLE, fit the generalized beta 2 $GB2(x; a, b, p, q)$ distribution to the individual observation data. Use your estimates for α, β, and m from part (c), as well as $q = 10,000$, as your initial guess. Report your estimated values for $\hat{a}$, $\hat{b}$, $\hat{p}$, and $\hat{q}$, as well as the value of the maximized log likelihood function $\ln \mathcal{L}$. Plot the second histogram from part(a) overlayed with a line representing the implied histogram from your estimated generalized beta 2 (GB2) distribution.

(e) (2 points) Perform a likelihood ratio test for each of the estimated in parts (b) and (c), respectively, against the GB2 specification in part (d). This is feasible because each distribution is a nested version of the GB2. The degrees of freedom in the $\chi^2(p)$ is 4, consistent with the GB2. Report the $\chi^2(4)$ values from the likelihood ratio test for the estimated GA and the estimate GG.

(f) (1 point) Using the estimated GB2 distribution from part (d), how likely am I to have a monthly health care claim of more than $1,000? How does this amount change if I use the estimated GA distribution from part (b)?