Homework 1: (Log)-Normal Families

36-313, Fall 2022

Due by 6 pm on Thursday, 8 September 2022 on Gradescope

> **Agenda**: Visualizing distributions; working with quantiles and distribution functions; getting a feel for both how much money typical households make, and the range of the distribution above and below that; fitting a model and comparing it to the data; quantifying the uncertainty in the fit.

The "Current Population Survey" is an on-going survey conducted by the US Census Bureau which is one the main sources of information about social and economic conditions in this country, including information about inequality. Part of the survey involves asking about household income; this is defined as the total income of everyone living in the household, but only counting income in money, not in-kind benefits. (So it doesn't count, for instance, the value of employer-provided health insurance, if someone has that, or the value of food which farmers grow for their own use.) Table HINC-01 gives the results for total money income, in the form of showing how many households have incomes within particular ranges, sometimes called "brackets". This table shows this information for various ways of breaking up the country, but in this assignment we are only going to look at the top line of the table, showing totals for the entire country.

Before proceeding, download the data for 2019[1]. Be sure to download the table for "All Races"; the Excel version is currently at [https://www2.census.gov/programs-surveys/cps/tables/hinc-01/2020/hinc01_1.xlsx]. You may need to do some work to get the data into R.

1. *Getting a feel for the distribution*
   a. (6) Create a histogram showing the number of households in each income bracket, i.e., from 0 to \$4,999, from \$5,000 to \$9,999, etc., all the way up to \$200,00 and above. *Hint*: The `hist()` function in R would actually *not* be a very convenient way to do this (why?).
   b. (5) Create a plot showing the cumulative distribution function (CDF), i.e., for each dollar value x, it shows what *fraction* of households make $\leq x$ dollars per year. You will need to somehow interpolate between the boundaries of income brackets; there are lots of sensible choices here, but explain your choice and why you made it. *Hint*: This figure will be helpful in many of the later parts of this problem.
   c. (6) Where would a household with an income of \$10,000 fall in the distribution? That is, what is the probability that a randomly-chosen household makes \$10,000 or less? What about \$20,000? \$50,000? \$100,000? \$199,999?
   d. (4) CMU's "full fare" or "sticker price" tuition (without financial aid) for the 2022–2023 academic year is \$59,864 [https://www.cmu.edu/sfs/tuition/undergraduate/index.html]. Where would that level of household income fall in the distribution? The total "full fare" cost for students living on campus is \$80,540; where would that fall? The Federal poverty guideline value for a family of four is \$27,500 [https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines/prior-hhs-poverty-guidelines-federal-register-references]. Where would that fall in the distribution?
   e. (6) What annual income would put a household at the 10th percentile of the distribution? At the 25th percentile? The 50th? The 80th? The 90th? The 99th? If you can't answer these questions exactly from the table, approximate, and explain how you made your approximations.
   f. (5) Use the table to calculate an *approximate* mean household income. Explain why using the

[1] The most recent year of data available is 2020, but the pandemic made lots of things very weird (including survey responses!).

table can only give an approximation. The data file also gives the actual mean household income; your value should be noticeably smaller than this — why?

2. *Fitting a distribution (crudely)* Recall from lecture that a random variable X follows a **log-normal distribution** when $\log X$ follows a Gaussian distribution. We write $X \sim LN(\mu, \sigma^2)$, but be warned that the mean of X is not μ and the variance of X is not σ^2! In fact, the mean of a lognormal distribution is $e^{\mu+\sigma^2/2}$. (You do not have to prove that last fact, but you will use it in a moment.)
   a. (5) The median of a lognormal distribution is e^{μ}; explain why.
   b. (6) Write d for the median and a for the mean. Starting from the expressions for mean and median above, find equations for μ and σ in terms of d and a.
   c. (6) Estimate μ and σ for this data by plugging in the observed median and mean. Call your estimates $\hat{\mu}$ and $\hat{\sigma}$; report them to reasonable precision. (R's default precision is unreasonable for these purposes[2].)
3. *Checking the fit of the log-normal distribution*
   a. (5) What median income is implied by your estimated parameters? Does it match the median from the data? Is this evidence that the log-normal distribution is a good fit?
   b. (6) What values are implied by your estimated parameters for the other percentiles you found above in Q1e? Do these predictions match the data? Is this evidence for a good fit? *Hint*: `qlnorm()`.
   c. (5) Similarly, where would the levels of income from problem Q1c] fall in the distribution implied by your parameter estimates? Does this match what you found in the data? Is this evidence for a good fit?
   d. (6) Using the `plnorm()` function, plot the CDF of your estimating log-normal distribution, and compare it to the CDF you estimated from the table, over the range of incomes from \$0 to \$200,000. (Why those limits?) Do the two curves match over this range? Is this evidence that the model is a good fit?
4. *Summary measures of income concentration*
   a. (5) Using the data, construct a Lorenz curve. Recall from lecture that this is a 2D plot, where the horizontal axis p shows different percentiles of household income (from 0 to 1), and the vertical axis shows what fraction of the total income goes to households at or below percentile p. If you need to make some approximations, describe them. *Hint*: Use the mean household income (given in the table) and the total number of households to figure out the total income.
   b. (4) The Lorenz curve for a log-normal distribution is given by $\Phi(\Phi^{-1}(p) - \sigma)$, where Φ is the CDF of the standard Gaussian distribution. (You don't need to show this.) Add this curve to your plot from the previous figure. Does it seem like a good match?
   c. (3) Recall from lecture that the Gini coefficient of inequality (or concentration) is the area between the 45 degree diagonal and the Lorenz curve, divided by the area under the 45 degree diagonal. Calculate this from your data-based Lorenz curve. (There are a couple of ways to do this; whatever you want to do is fine so long as you can explain your approach.)
   d. (3) Compare your Gini coefficient to the one given in the table. Why might the two not match?
   e. (3) The Gini coefficient of a log-normal distribution is $2\Phi(\sigma/\sqrt{2}) - 1$. (You don't need to show this.) What is that from your estimated parameters, and how well does it match the values from the table and from your Lorenz curve?
5. *Timing* (1) About how long did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

[2] R's default precision for printing numbers is perfectly reasonable if you are trying to debug numerical code, but that's a very different goal than trying to communicate statistical results to human beings.

Extra credit: Read the handout on the class website about the "propagation of error" technique for calculating standard errors[3].

a. (2) Referring to Q2b, find the four partial derivatives $\partial\mu/\partial d$, $\partial\mu/\partial a$, $\partial\sigma/\partial d$ and $\partial\sigma/\partial a$.

b. (2) Find formulas for the standard errors (not variances) in $\hat{\mu}$ and $\hat{\sigma}$ as functions of d, a, and standard errors in d and a, say ϵ_d and ϵ_a. You can assume, for the purposes of this problem, that measurement errors in d and a are uncorrelated. *Hint*: you'll need ECa.

c. (3) Use your formulas from ECb, and the standard errors for the mean and median given in the table, to find standard errors for $\hat{\mu}$ and $\hat{\sigma}$ for this data set.

d. (3) Using the log-normal model, and propagation of error, find a standard error for the Gini coefficient. *Hint*: You'll need the derivative of Φ; that function has a name we know (what is it?), and an R command to calculate it (what command?).

[3]You remember from intro. stats. that every estimator of a parameter has its own "standard error", which is just its standard deviation. (An estimator is a random variable, because it's a function of the data, and the data are random.) We use the name "standard error" as a reminder that, in this case, the standard deviation is telling us about the random, merely-statistical errors that arise from using this estimator. (The estimator may also make *systematic* errors.) An important special case of this is the "standard error of the mean", when we use the sample mean to estimate the population expected value. There is a simple formula for the standard error of the mean, $\sigma/\sqrt{n}$. But, again, *every* estimator has its own standard error, which will have to be calculated in a different way. We are *not* using the standard error of the mean at any point in this problem set, and will probably never use it in this course.