Homework 3: A First Look at Group Comparisons

36-313, Fall 2021

Due at 6 pm on Thursday, 23 September 2021

In HW 1, I introduced you to the Current Population Survey (CPS), which is an random-sampling survey of the whole American population conducted on an on-going basis by the US Census Bureau. An important part of that is the Annual Social and Economic Supplement (ASEC), which surveys people about things like household finances. A random sub-sample, with identifying information removed, is available as a "public use microdata sample" ("PUMS"), via the IPUMS center at the University of Minnesota [https://cps.ipums.org/]. I have used this to create a datafile[1] for this assignment, using the 2020 ASEC; for various reasons I can't make that file public, but you can find it on Canvas under Homework 3.

The data file contains a lot of variables; we won't be using most of them in this time, but we'll come back to this data in later weeks. Many of the variables are really categorical, but are represented through numerical values or "codes"; the "codebook" file, also on Canvas, explains the codes for each variable. In R, you might find it convenient to convert these variables to factors or to ordered variables, but that's not necessary to do the assignment.

1. *Summary statistics and sub-populations* The `HHINCOME` variable records annual household income, in dollars[2].
   a. (4) Find the over-all median and mean income; the 10th, 25th, 75th, 90th, 95th and 99th percentiles; and the standard deviation.
   b. (4) Using the `RACE` variable, calculate the same summary statistics for those who self-identify as white (only, not white-and-something). *Hint*: Use the codebook file to see which value of `RACE` corresponds to "white (only)".
   c. (4) Using the `RACE` variable, calculate these summary statistics for those who self-identify as black (only).
   d. (4) As Q1b and Q1c, but for Asians.
   e. (4) Create a single plot which shows all these percentiles, and the means (but not the standard deviations), distinguishing the different racial categories by color or plotting symbol (or some other visual clue). You may find it helpful to use a log scale on the axis showing dollars. (As in the Lecture 2 examples of income trends over time, I suggest putting dollars on the vertical axis, but do whatever's convenient and clear to read.)
   f. (5) Try to describe, in words, the differences in income between these three groups.
2. *Summary statistics and different sub-populations*
   a. (4) Using the `EDUC` variable, find the mean income, the standard deviation of income, and the same percentiles as in Q1a, for everyone whose education stopped with less than a bachelor's degree. (For short, we will call these "non-college-educated" going forward, even though some of them have some college education.) *Hint*: The numerical codes for education levels are arranged so that everyone who hasn't completed at least a bachelor's degree has a code below some threshold. (See the codebook.)
   b. (4) As Q2a, but for those who have at least finished a bachelor's degree. (For short, we'll call these people "college-educated".)

[1] The original data contains a row for every member of every surveyed household, but we're only interested in household income for this assignment, so I have picked one member of each household as a representative.

[2] The original data contained 12 households reporting negative incomes. I have replaced these with 0s, to simplify your analysis. This increased the mean income by 1.2 dollars/yr.

c. (4) As Q1e, but comparing the college-educated to the non-college-educated.
d. (4) As Q1f, but comparing the college-educated to the non-college-educated.

3. *Q-Q plots* One common visual device for comparing two distributions is a **quantile-quantile** plot, or **Q-Q plot**. (You may have encountered these in other courses already.) We make these by plotting the quantiles of group A on one axis against the same quantiles of group B on the other axis. In symbols: say $Q_A(p)$ is the p quantile of group A, $\mathbb{P}(X \leq Q_A(p)) = p$, and likewise for $Q_B(p)$, $\mathbb{P}(Y \leq Q_B(p)) = p$. We make up a sequence of probabilities between 0 and 1, say $0 < p_1 < p_2 < \ldots < p_m < 1$, and for each p_i, we plot the point $(Q_A(p_i), Q_B(p_i))$ until we have m points for our m quantiles in the plot. In R, the function `qqplot()` conveniently does this, including making up a sensible sequence of p_is.
   a. (3) Explain why, if the two groups have the same distribution, the Q-Q plot should be on, or very close to, the 45 degree diagonal.
   b. (4) Going the other way, if the Q-Q plot follows the 45 degree diagonal, does that imply the two groups have the same distribution?
   c. (4) Suppose all the points from the Q-Q plot are on or below the diagonal. Which group, A or B, has higher values? Explain.
   d. (5) Create a Q-Q plot comparing black and white incomes. Explain, in words, what the shape of the plot tells you about the two income distributions.
   e. (5) As Q3d, but comparing college-educated to non-college-educated.
   f. (5) ("Intersectional analysis") As Q3d, but now compare college-educated blacks to non-college-educated whites.
4. *Uncertainty in comparisons*
   a. (2) What is the difference between the median income of blacks to the median income of whites?
   b. (5) Using bootstrapping, give a 95% confidence interval for this difference. *Hint*: Lecture 6.
   c. (2) Why should we *not* use a permutation test in Q4b?
   d. (3) Does this interval contain 0? What can we conclude from this?
   e. (2) What is the ratio of mean income of the college-educated to the non-college educated?
   f. (5) Using bootstrapping, give a 99% confidence interval for this ratio.
   g. (3) Does this interval contain 1? What can we conclude from this? Why do we not care if this interval contains 0?
5. *Timing* (1) How long, roughly, did you spend on this assignment?

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.

Extra credit (5): Use the bootstrap to find 95% confidence envelopes for the Q-Q plots you made in Q3. Do the confidence envelopes change any of your conclusions? *Hint*: Read section 9.8 from Zieffler, Harring and Long's *Comparing Groups* (link on the course webpage) on "Bootstrapping the Confidence Envelope for a Q-Q Plot".

Note on sampling weights

Because ASEC is a random-sampling survey, each household it contacts is standing in for, or representing, some number of other, similar households in the total population. In a simple random sample, where there are N households in the population and we survey n of them, each data point is representing N/n households. We'd say that each surveyed household has a **sampling weight** of N/n. The people who run ASEC know that they're more likely to survey some types of households than others, so they estimate a *different* weight for each household. These are contained in the `ASECWTH` variable. To capture what's going on the population, we should really use these weights. We will explore how to do this later in the course; for now, I just want to you to be aware that the calculations I'm asking you to do are not quite accurate.