0. Title and author(s):

Title: **The Categorization of Multilingualism in Washington State**
Author: Tully McCombs

1. Summary of Research Questions:

1) I**s there a difference in the age distribution of those who speak a non-English language at home for different language groups? How does this compare between a *linguistic* categorization of languages and the standard *demographic* categorization of languages commonly used in surveys like the Census?**
   Yes, there is a difference. For example, people who speak Spanish at home in Washington State are more likely to be children than is the default, whereas speakers of North American languages such as Aleut or Cherokee tend to be in their 50s. In the more common demographic categorization, many of these differences seem to disappear due to the broadness of the categories.

2) **Are there other language groups than Spanish which are more likely to be linguistically isolated than Spanish, and thus be more likely to need help accessing essential services?**
   Yes. The results of the Chi-Square test were X^2=348.279, p=1.12e-63, so we can reject the null hypothesis that there is no difference in the proportion of households which are linguistically isolated depending on the linguistic group. Specifically, more than one-fifth of respondents in households which primarily speak Sino-Tibetan languages, Afroasiatic languages, Austroasiatic languages, and European languages which belong to no known or only very small language groups are linguistically isolated—over 20% of those who said they speak that language at home live in a household in which no-one speaks English very well. When making the equivalent graph for the broader categorization system, while they both had a very similar mean proportion between all the categories, the differences tended to be several percentage points more extreme for the more precise categories than they were for the broader system, with a range of about 0.08 for the broad categories vs about 0.22 for the precise categories.

3) **Are there significant demographic differences between non-English speakers at a more granular degree than the broad demographic categorization allows for?**
   Yes. The results of the Chi-Square test were X^2 = 134.980, p = 2.09e-27 for linguistic isolation among 'Asian languages' and X^2 =16.606, p = 0.005 for the proportion receiving public assistance income among people who primarily speak an 'Asian language' at home. Both values are below 0.01, which means that they reach significance, and we can *reject* the null hypothesis that there is no difference in the proportion of people

living in linguistically isolated households or receiving public assistance income depending on which Asian language they speak at home. In fact, on the graphs, proportions of linguistic isolation among Asian languages alone varied by a little over 0.20, instead of the 0.08 range for the same proportion among all languages with the broader categories. For public assistance income, the values were generally much smaller, but among Asian languages they ranged from 0.0 (likely due to small samples) to 0.037. Given that these are the first two statistics I tried, I think it's safe to say that there are generally differing demographic patterns between these linguistic groups, all of which fit in the 'Asian and Pacific Island languages' bin in the broader categorization.

2. Motivation and Background:

Many important services in the US are only offered or described in a limited selection of languages, sometimes only English, or English and Spanish. In a country with speakers of many different languages (such as the US), this can pose a serious issue to families without fluent English speakers. Understanding not only what languages people speak, but also their degree of fluency and when that fluency tends to be obtained is important to determine how to prioritize different languages for different services. Understanding the difference in English ability between language groups and how that impacts the speaker's ability to receive government aid is useful for the same reason: it allows people to make directed attempts at outreach in languages whose speakers are more likely to need the help.

Unfortunately, the most common grouping of languages in surveys and so on in the US have only four broader options that are not English:

- Spanish (which is only one language, even if it is the most common in the US besides English)
- Other Indo-European languages (which is one of the world's primary language families)
- Asian and Pacific Island languages (which includes at least *six* of the world's primary language families, all of which are largely unrelated to each other)
- All other languages (which includes *all languages* coming out of both American continents as well as the entire continent of Africa, parts of Europe, and debatably the Middle East)

This grouping has at least two very big problems:

First, each of these categories is of a completely different scale. Most languages in the world fall in the "Other" category, and only a tiny fraction under "Spanish" or "Other Indo-European languages".

Second, the members of "Asian and Pacific Island languages" and "All other languages" have very little in common with the other members of their own category. The *only* thing they have in common linguistically, for the most part, is that they are all languages.

Therefore, my research questions center around testing where differences occur within what would usually be considered a single language category. As there are well over a hundred individual languages among the respondents to the survey, I instead divided them up mostly by family—comparing, for example, Indo-European languages to Dravidian languages, or to Sino-Tibetan languages, or to Niger-Congo languages. Because Spanish is by far the most common non-English language spoken in the US, I still chose to separate it out from the other Indo-European languages. However, the benefit of categorizing by language *family* is that it reflects itself in the cultures and demographics of the people who speak them. Languages are related because they were once the *same* language, spoken by the same group of people. Languages are tied to people, not places, and so when people migrate somewhere, their languages follow them—meaning that even without the other issues, trying to categorize languages by their geographic location is confusing and vague at best. Understanding which linguistic families are, for example, most in need of translations to access basic services is a lot more helpful than knowing that that help is needed by speakers of some unknown language that is not Indo-European or from Asia.

3. Dataset:

I will be using data from the American Community Survey for 2018, downloaded here: https://www2.census.gov/programs-surveys/acs/data/pums/2018/1-Year/

The key for converting the categorical data from numbers to descriptive labels can be found at https://data.census.gov/mdat/ . The precise language recodes, specifically, I transcribed and categorized in the files language_recode.txt and shortened_fam_labels.txt, as they were not in a format I can easily read into Python.

The two main datasets are listed on the first link as csv_hwa and csv_pwa, and should unpack with the names psam_h53.csv and psam_p53.csv .

4. Methodology:

Step 1: Load in and **merge datasets**, trimming all columns except SERIALNO, LNGI, HHLANP, and HHL from the household dataset and trimming all except SERIALNO, SPORDER, AGEP, LANP, LANX, and PAP from the individual dataset.

Step 2: Recode all categorical variables from their existing numeric interpretations to descriptive labels for ease of reading, and add a column recoding both specific language columns (HHLANP and LANP) by linguistic group. Also recode PAP as a Boolean variable, with 0 as False and any non-zero amount of public assistance income as True.

Step 3: Question 1—Graph relationship between linguistic family spoken at home and age of respondent, as well as broad category of language spoken at home an age of respondent, in

both cases as a violin plot where the width of the pyramid corresponds to the amount of respondents of that age for that language group.

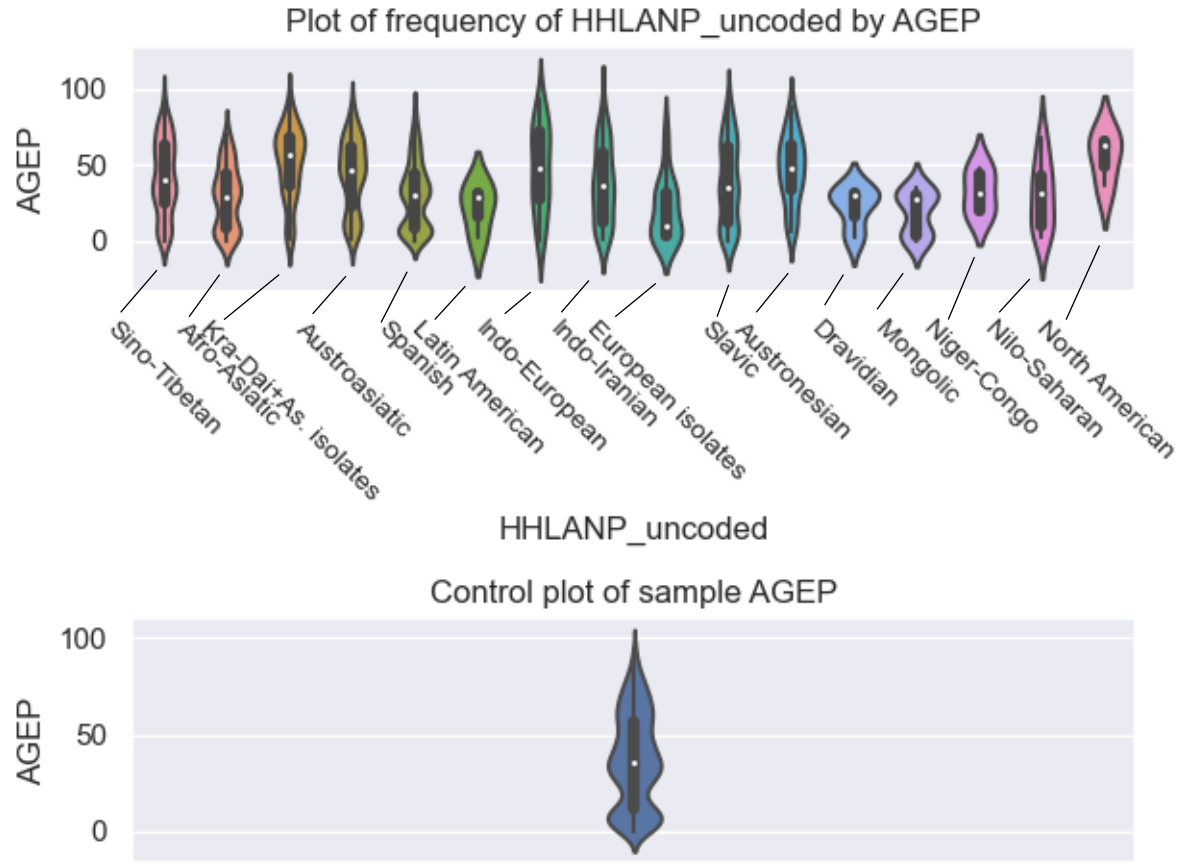
Step 4: Question 2—Graph proportion of speakers of each language group which are in linguistically isolated households as a bar graph for both categorization methods. Conduct a **Chi-Square goodness of fit test** to test whether the differences present could reasonably be due to chance.

Step 5: Question 3—Graph as a bar plot the proportion of speakers of each 'Asian' language group which are in linguistically isolated, and conduct a **Chi-Square goodness of fit test** to check whether the differences present could reasonably be due to chance. **Repeat** for the proportion of respondents receiving public assistance income.

5. Results:

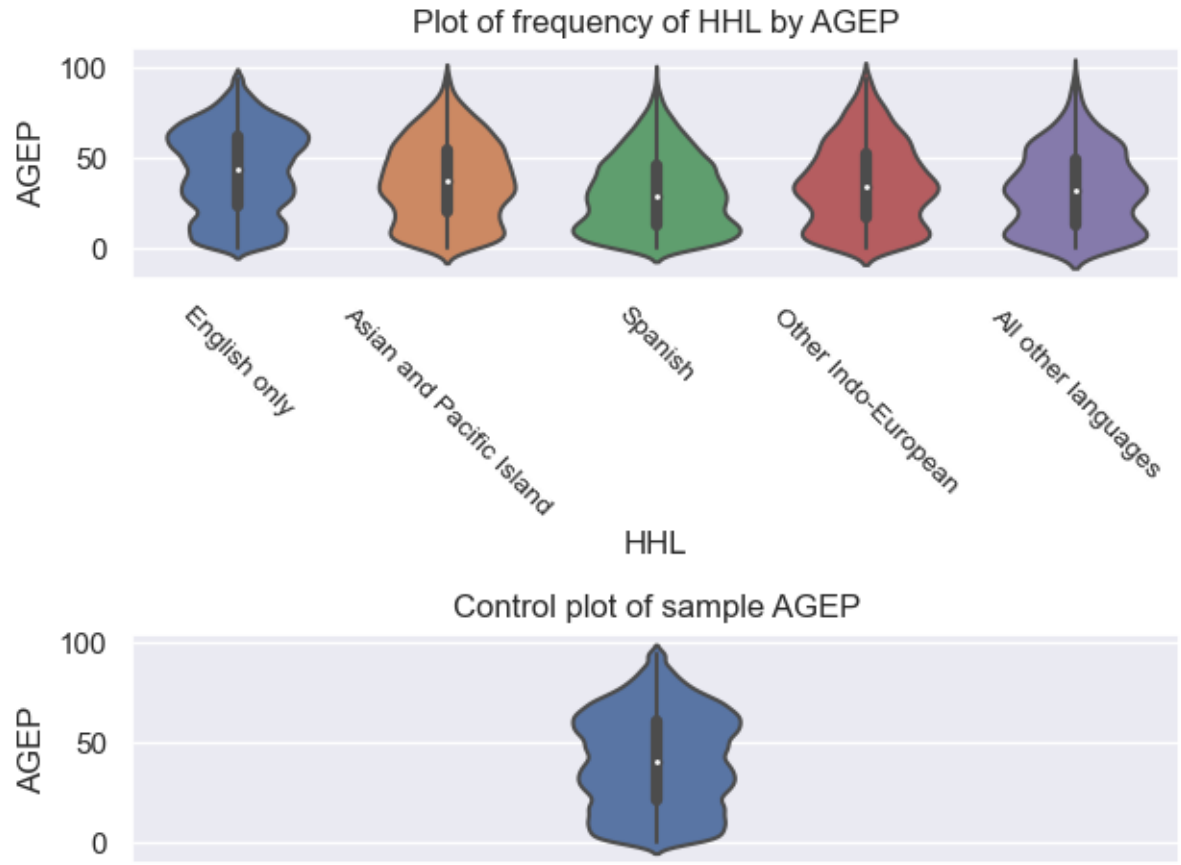
Is there a difference in the age distribution of those who speak a non-English language at home for different language groups? How does this compare between a *linguistic* categorization of languages and the standard *demographic* categorization of languages commonly used in surveys like the Census?

Yes, there is a difference.
As can be seen in the plot below, each language group has a different distribution of ages, both compared to each other and to the control plot, in this case the age distribution for all respondents with a household language other than English.



For example, respondents with Spanish as their household language tend to primarily be either children or in their 30s. Speakers of languages from the Kra-Dai family or Asiatic language isolates, on the other hand, seem to be primarily over age 60. Several language groups, notably Afro-Asiatic, Spanish, Slavic, and Sino-Tibetan seem to more-or-less follow the same age pattern as the control, though still with some variation. Austronesian languages, non-Spanish Indo-European languages, and Austroasiatic languages all seem to have a peak around age 50.

Speakers of Indo-Iranian languages, small European language families, and Nilo-Saharan languages all seem to have a greater proportion of children than is standard; while Central/South American, Dravidian, Mongolic, Niger-Congo, and North American languages likely don't have enough respondents in Washington State to be very reliable on this graph, as seen in the way that entire age ranges appear here to be truncated entirely.



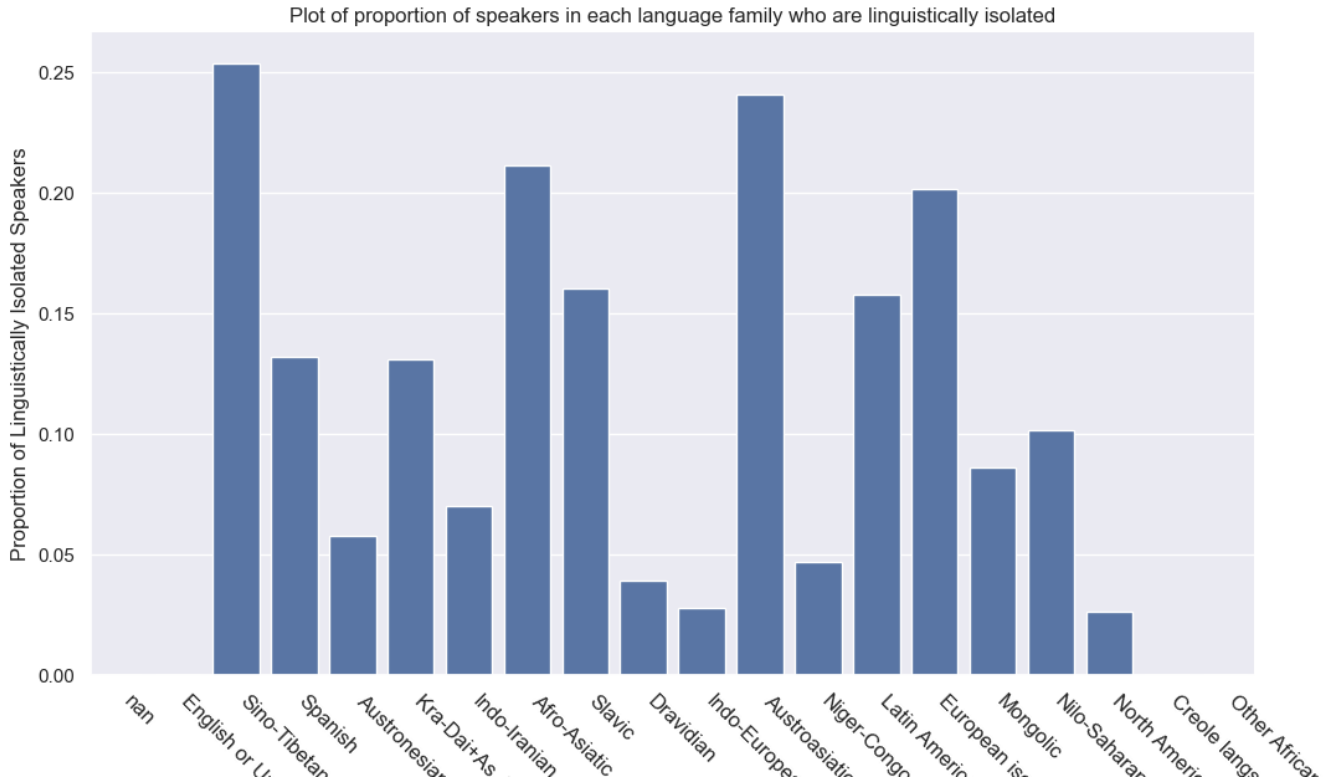
With the broader categorization system, most of these distinctions disappear.
Again, Spanish-speaking households seem to have more children than they do any other age group, and all four non-English options appear to have far fewer speakers above age 50 than either English-speakers or the control. But that's pretty much all you can see.

Are there other language groups than Spanish which are more likely to be linguistically isolated than Spanish, and thus be more likely to need help accessing essential services?

Yes.
The results of the Chi-Square test were X^2=348.279, p=1.12e-63, so we can *reject* the null hypothesis that there is no difference in the proportion of households which are linguistically isolated depending on the linguistic group.

More than one-fifth of respondents in households which primarily speak Sino-Tibetan languages (such as most Chinese languages), Afroasiatic languages (such as Hausa, Hebrew, and Arabic), Austroasiatic languages (such as Vietnamese), and European languages which belong to no known or only very small language groups (such as Finnish, Turkish, and Basque) are linguistically isolated—in other words, over 20% of those who said they speak that language at home live in a household in which no-one speaks English "very well" or better.

Without counting language groups with zero linguistically isolated speakers, when categorizing by language family, proportions range from about 0.03 to just over 0.25, showing clear differences depending on the language family. Again, the Chi-Square test agrees with p=1.12e-63 that it is extremely unlikely these differences are all due to chance.

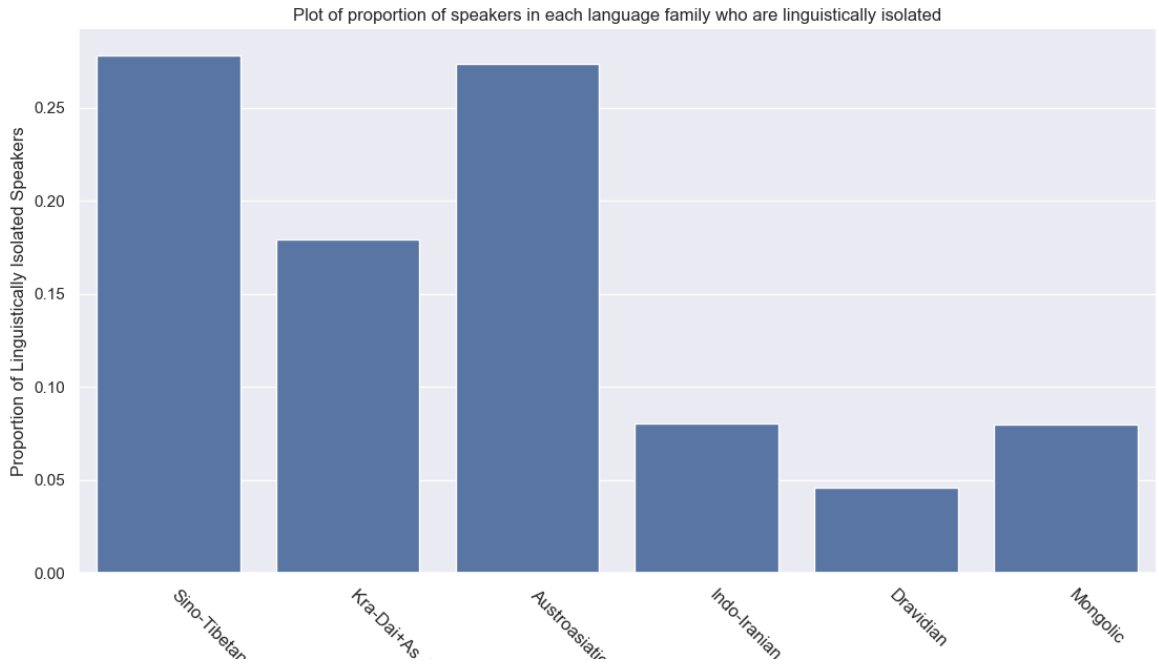


As far as having no English-speakers in the household goes, Spanish is more-or-less in the middle. It therefore does makes a lot of sense to prioritize Spanish-speaking households for translated materials, as they are also the most numerous category with 4483 entries. But speakers of Sino-Tibetan languages are the most likely of all the language families to have no competent English speakers in the household with *over a quarter* of their speakers being linguistically isolated, and have a solid 1277 entries in this survey, out of 11,766 who entered a language other than English as their language spoken at home. The 761 speakers of Austroasiatic languages, as well as the smaller groups that make up Afro-Asiatic language speakers and speakers of European language isolates/small language family members are also significantly more likely than Spanish-speakers to need the help.

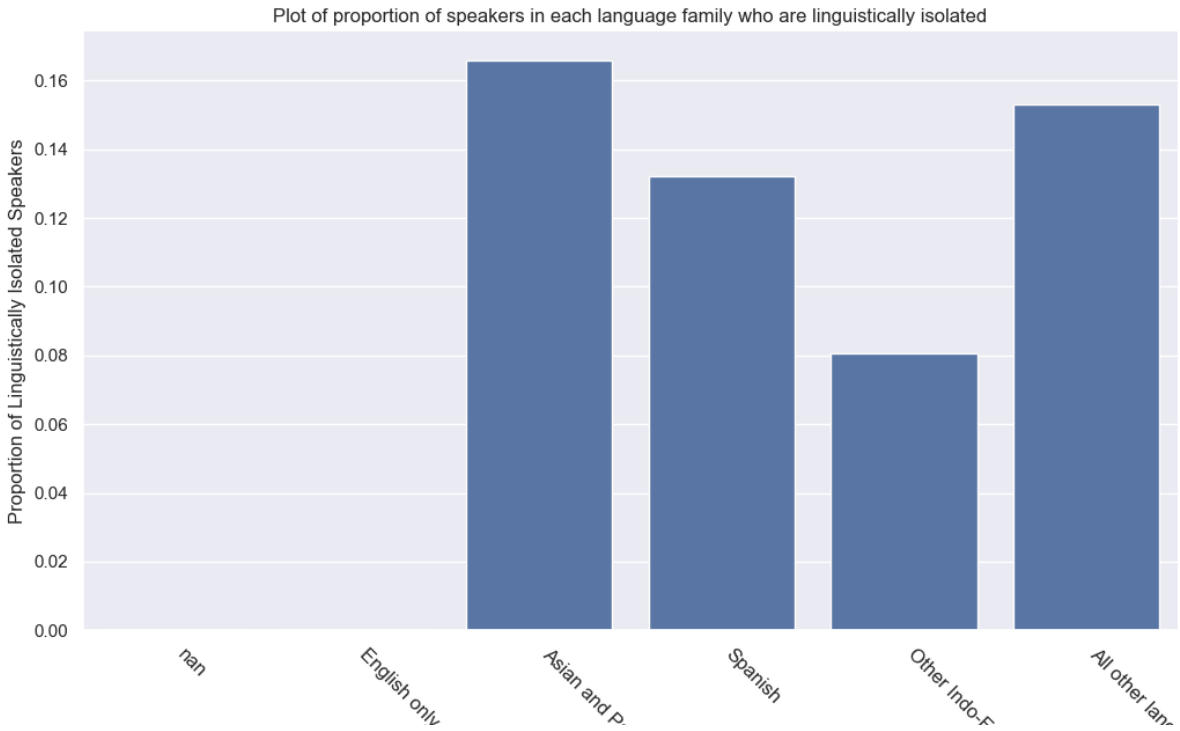
Are there significant demographic differences between non-English speakers at a more granular degree than the broad demographic categorization allows for?

Yes. To test this, I graphed and carried out the same statistical test for a couple of different variables among Asian languages (one of the groups in the broad categorization system), and between Asian languages and the other broad groups.

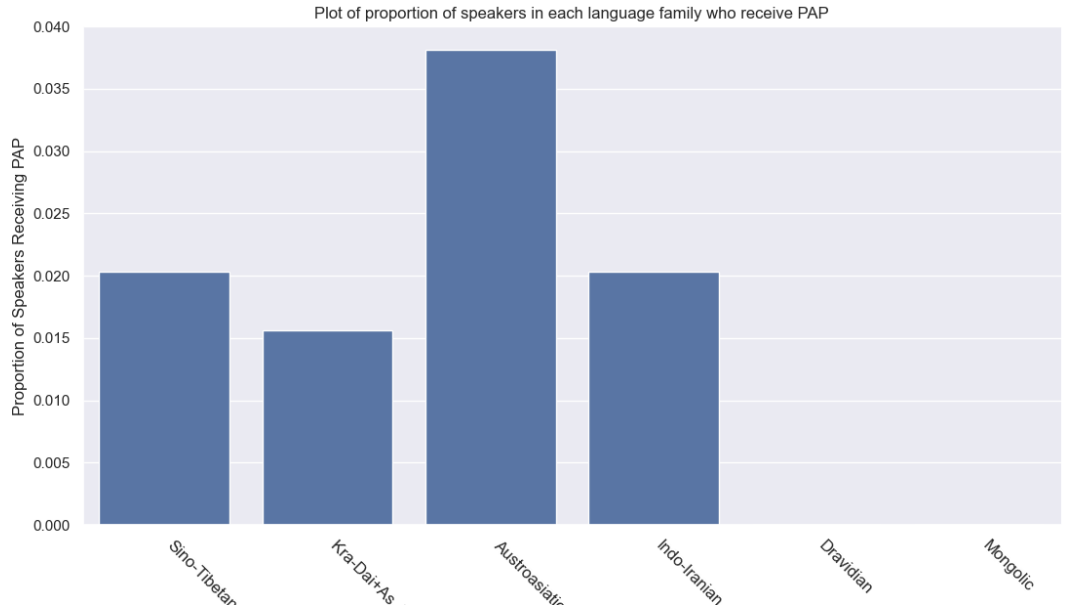
The first test was for proportion of linguistically isolated speakers again. Here the range was *slightly* smaller than for all language families, with Dravidian languages just below 0.05 and Sino-Tibetan languages a little bit over 0.25. If I were using the national dataset instead of just the data for Washington State, I probably could have included the 15 other major language families primarily spoken in Asia separately, but as it is, they're all lumped in together with Kra-Dai and the language isolates, especially since the categorization for many of them is still debated. With only a p = 2.09e-27 chance of these differences being coincidence according to the Chi-Square test, there are pretty clearly major differences in between different 'Asian' language groups.



By contrast, the equivalent plot using the much more common broad categorization has far less variety. All values range between about 0.08 and 0.16, a much smaller range than before, which makes it far harder to pick out what patterns *might* exist, much less find them big enough to be statistically significant—which is entirely justified, because for five out of six Asian language families with a notable presence in Washington State, its 0.16 estimate is flat-out wrong, by at least 0.08 in either direction. That is a massive loss of information and patterns that are otherwise pretty clear, even in language groups with dozens of languages.

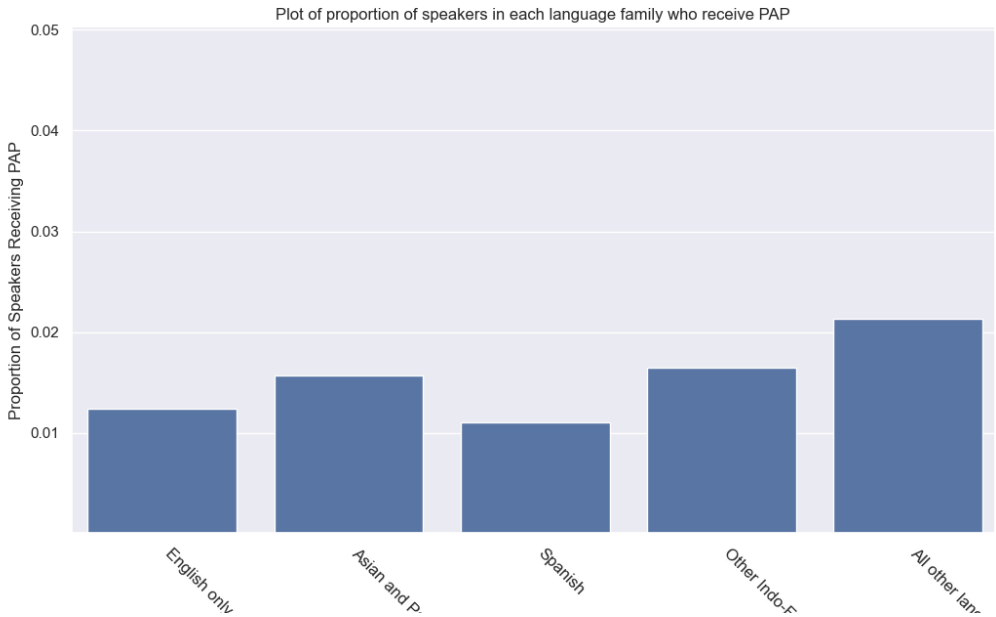


Next, I looked at the proportion of speakers of each Asian language group receiving public assistance income (PAP).



All of the non-zero proportions ranged between 0.015 and 0.04, but as seen with the broader categories below, that actually goes unusually high for this statistic. With a p-value of p = 0.005 among Asian language families, these differences from family to family are still unlikely to

appear by chance. We can, again, *reject* the null hypothesis that there is no significant difference between Asian language families in the proportion of people who receive public assistance income—if with somewhat less confidence than we could before.



In summary, at least for the 'Asian and Pacific Island languages' category, there is clearly more information to be found if you divide them up by linguistic family instead of grouping all these extremely disparate languages and cultures together. Grouping them by linguistic family allows for much stronger patterns in demographic data as well as purely linguistic data, making it far easier to find statistically significant differences.

This is why category distinctions are so important: if all members of the category *have things in common,* then measuring them together as a category doesn't obscure too much of their properties. On the other hand, if all members of a category are chosen pretty much at random, their characteristics as a group will always trend toward the overall average value for that characteristic regardless of the properties of each individual member.

6. Challenge Goals:

Multiple datasets: I completed the multiple datasets challenge goal by using both household and per-person datasets from the ACS, and merging them on household ID.

Result validity: For the majority of my questions, I used a Chi-Square Goodness of Fit test to verify that the test results were unlikely to happen by chance. A Chi-Square test is the best to use for situations where what is being measured is the relative frequency of samples from different groups meeting a shared condition—such the frequency that people who speak a certain language at home will be in a linguistically isolated household. I was planning to also implement the same test for age bins in my first question, and possibly also do a more specific test for each language group on some questions, but I ran out of time before I could implement them.

7. Work Plan Evaluation:

1) **Join, clean, and recode data (3 hrs)**– This step took much longer than planned! Specifically the recoding data part, as I had not quite realized that there were over a hundred languages that I would need to recode, and absolutely did not want to do that by hand the way I dealt with those variables that only had a handful of options.

2) **Graph questions and check requirements (2 hrs)–** This step took somewhat longer than planned, about 4 hours as I forgot I would also have to sometimes do math on the data in order to get a number that could be interesting to graph. It often was also tricky trying to format the graphs themselves in a way where they continued to be legible, given the large number of rather long labels that tend to overlap.

3) **Calculate test models (3 hrs)**—This part took about as much time as anticipated, because as a largely categorical survey, it doesn't really have uncertainty measurements. This greatly reduced my options for statistical tests, into mostly just Chi-Square tests for most of my questions unless I *really* wanted to do an ANOVA, so the only thing I really had to do here was produce observed and expected frequency tables for whatever variable I was looking at.

4) **Conduct solo tests (3 hrs)**— This took much less time than expected! I started to implement the Chi-Square test by myself, went "hey wait a minute I'm sure someone's already made a function for this", and learned to do it through SciPy instead of writing my own.

5) **Conduct ANOVA tests (2 hrs)—** Don't know, I did not have time for this.

8. Testing:

Most of my testing was done through print statements that I don't have much of a record of. Wherever there was math involved, I checked each stage of whatever value I was calculating against what I computed for it based on the csv files; such as in calculating the observed and expected values for the statistical tests. To do so, I would usually check my language group ranges and filter the Excel copy of the relevant csv to show only those language ranges, and use the total at the bottom as what the total for that group of items should be, and calculate the rest from there. As an example, there were 62 LANP entries for North American languages according to my program, so when I filter to only LANP entries in the range 6800-7124 (inclusive), there should be exactly 62 entries. This did get especially tricky when working with HHLANP, for example, which is only in the household-level spreadsheet and not the individual-level spreadsheet, because I couldn't just count it when comparing it to an individual-level variable like PAP or LNGI. In those cases, I stuck to very sparse variables in my testing, so I could copy and filter to all the relevant household ids from the household spreadsheet on the individual spreadsheet without it taking hours to set up the filter. I did not have time to write assert_equals tests.

9. Collaboration:

I didn't collaborate with anyone. My main online sources were the Pandas and Seaborn documentation, but I did also occasionally refer to questions on StackExchange with similar issues to my own (though I did not ask about my specific code issues) and got some help formatting my Seaborn graphs by reading through a blog post: https://www.drawingfromdata.com/how-to-rotate-axis-labels-in-seaborn-and-matplotlib