[redacted]

Class: ST_511

Data Analysis 1

Date: 11/07/2014

INTRODUCTION

The data set used in this analysis was obtained from the American Community Survey (ACS). Every year the U.S. Census Bureau contacts households in order to receive information on social, economic, demographic, and other characteristics of U.S. population. The data subset represents a sample of households (randomly selected) in Oregon in 2013 that consist of opposite gender married couples. The information obtained by the Bureau includes a unique ID number assigned to each household; age of the husband and the wife; total annual income of the husband and the wife; number of bedrooms in the home; monthly cost of electricity and gas; the number of children in the home; whether the home has internet access; the way the household took the survey; whether the residents own the house with or without a mortgage or rent; the primary language spoken in the home; and the decade the home was built.

The main questions of interest in this study can be summarized as follows:

1. Do husbands tend to be older than their wives?
2. Do households with children spend more on electricity?
3. Do households that own their houses without mortgage or loan earn more money than those that rent houses?

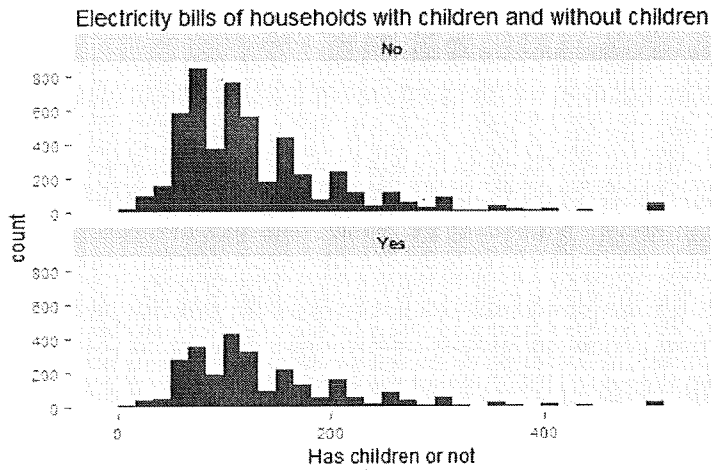
To answer these questions, first, it would be useful to obtain summary statistics of the relevant variables. The data set provides information on 7811 households. The ages of husbands vary from 17 to 95 years with the mean value close to 54 years; the ages of wives vary from 19 to 95 years with the mean value of 52 years. The lower bound of total income is the same for husbands and wives and equals - 6300. Therefore, certain individuals suffered losses during 2013. The upper bound of total income for husbands equals $756 000 and for wives $421 000; the mean values are $59 831 and $28 985 respectively. The monthly cost of electricity is between $4 and $500; on average households pay $131.9 per month. 5151 households don't have children, and the rest have up to 12 children. Also, there are 76 households that occupied homes without payment of rent; 1896 that owned free and clear; 4505 households that owned homes with mortgage or loan, and 1334 that rented homes.

Second, we should examine the graphical representation of data involved. The distribution of the differences in ages of husbands and wives is centered around 2 years and varies from -22 to 40 years.

Figure 1

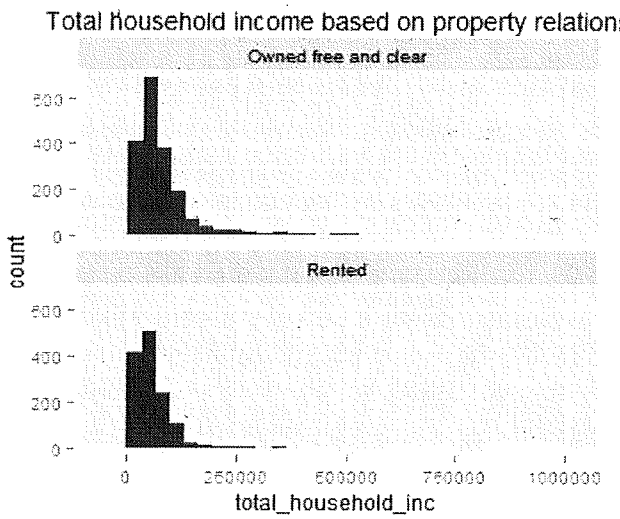
When comparing electricity bills that households with children and without children pay, one can notice that the spread of values is approximately the same. The data is concentrated on the left of the histograms, the right tails are longer.

Figure 2



The histogram of total household income based on different types of property relations shows that data for both groups is significantly skewed to the right.

Figure 3



METHODS

The methods chosen to answer the questions are paired t-test and two sample t-test since they allow to make inferences about the mean difference and the difference in means where appropriate. The following section discusses if the t-tests are suitable for the questions and data, the assumptions of the tests, and whether those assumptions are met or not.

1. **Do husbands tend to be older than their wives?**

The data is this case is pared since we look at the differences in ages between two spouses who belong to the same household. Therefore, we have only one random sample of sample differences and conduct a one sample t-test on the differences in two responses from our paired data.

Figure 4

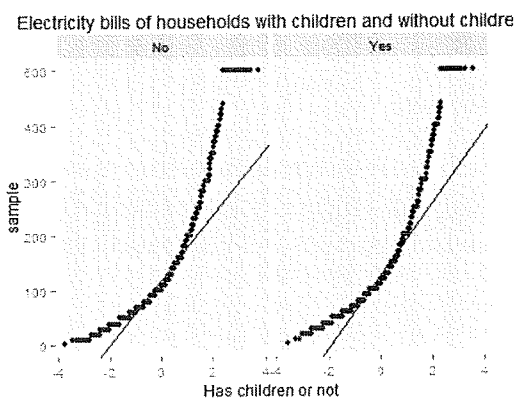
To test for *the normality assumption* (sample was drawn from a normally distributed population) we generate a normal probability plot. Systematic deviations from a straight line provide evidence that the sample of differences is not from normal population. Histogram confirms long-tailedness of the

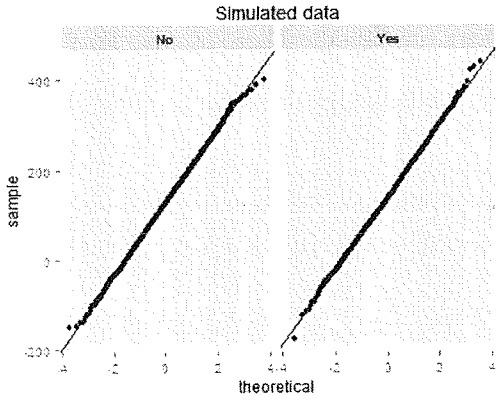
data. However, the sample size is very large (7811 observations). We, thus, claim robustness of the paired t-test to this violation recalling the Central Limit Theorem.

2. **Do households with children spend more on electricity?**

Two sample t-test is appropriate in this case since we have two populations and two samples (households with children and households without children). The assumption of independence is satisfied by relying on random samples from general populations. As mentioned above, the spread of values of payments for electricity is approximately the same for two groups (Figure 2). Moreover, the sample sizes are quite large. Therefore, we should not worry about *the assumption of equal variances* even though the sample sizes differ (5151 households don't have children and 2660 do). To test for *the normality assumption* we again generate a normal probability plot. The data is probably not distributed normally since it deviates from the straight line in both groups. The histogram (Figure 2) also shows that it's positively skewed. We also look at the plot that is based on simulated data with similar samples sizes and averages as our data but drawn from normal populations with equal standard deviations. The simulated data looks very different from our data. Therefore, there is evidence that the samples are not from normal populations. However, since the spreads are the same, population shapes are the same, and the sample sizes are large, we can still use t-tools to answer our questions.

Figures 5 and 6

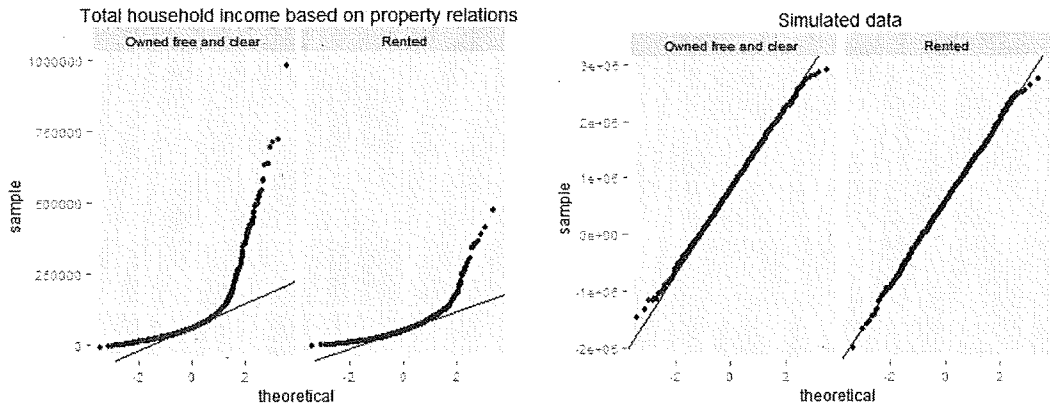




3. **Do households that own their houses without mortgage or loan earn more money than those that rent houses?**

We have similar to the aforementioned concerns when we try to answer the third question. We chose to use two sample t-test in this situation since we have two populations and two samples (households that own their homes free and clear and households that rent). Again, there are possible issues with *the normality assumption*. However, we claim the robustness of the t-test referring to 1) the Central Limit Theorem (our sample sizes are quite large: 1896 households that owned free and clear and 1334 households that rented homes); 2) the fact that population shapes are roughly the same (the data in both groups is skewed to the right) (Figure 3); and 3) the spreads of values in the two samples are not very different (it may be due to sampling variability).

Figures 7 and 8



SUMMARY

1. There is convincing evidence the mean difference in the ages of husbands and wives is not zero (two sided p-value < 0.01, paired t-test). We estimate that the age of husbands is 2.2 years greater on average than the age of wives. A 95% confidence interval for the difference is 2.1 to 2.3. Therefore, with 95% confidence the husband's age is on average between 2.1 and 2.3 years greater than the wife's age.
2. There is convincing evidence that the mean monthly cost of electricity in households that have children is not equal to the mean monthly cost of electricity in households that don't have children (two sample t-test, two-sided p-value < 0.01). The mean monthly cost of electricity in households that have children is estimated to be $9.1 bigger than the mean monthly cost of electricity in households that don't have children. With 95% confidence the mean monthly cost of electricity in households that have children is between 5.3 and 12.9 dollars bigger than the mean monthly cost of electricity in households that don't have children.
3. There is convincing evidence that the mean total income in households that owned their homes free and clear is not equal to the mean total income in households that rented their homes (two sample t-test, two-sided p-value < 0.01). The mean total income in households that owned their homes free and clear is estimated to be $23039 bigger than the mean total income in households that rented their homes. With 95% confidence the mean total income in households that owned their homes free and clear is between 17974 and 28103 dollars bigger than the mean total income in households that rented their homes.

We cannot make causal inference since we sampled randomly from populations and did not do any experiments. We can generalize to households in Oregon in 2013 that consist of opposite gender married couples because of randomization procedure. However, we are not able to make more general conclusions (for instance, about all households in the U.S.).