Data Analysis #1

Dustin Leatherman

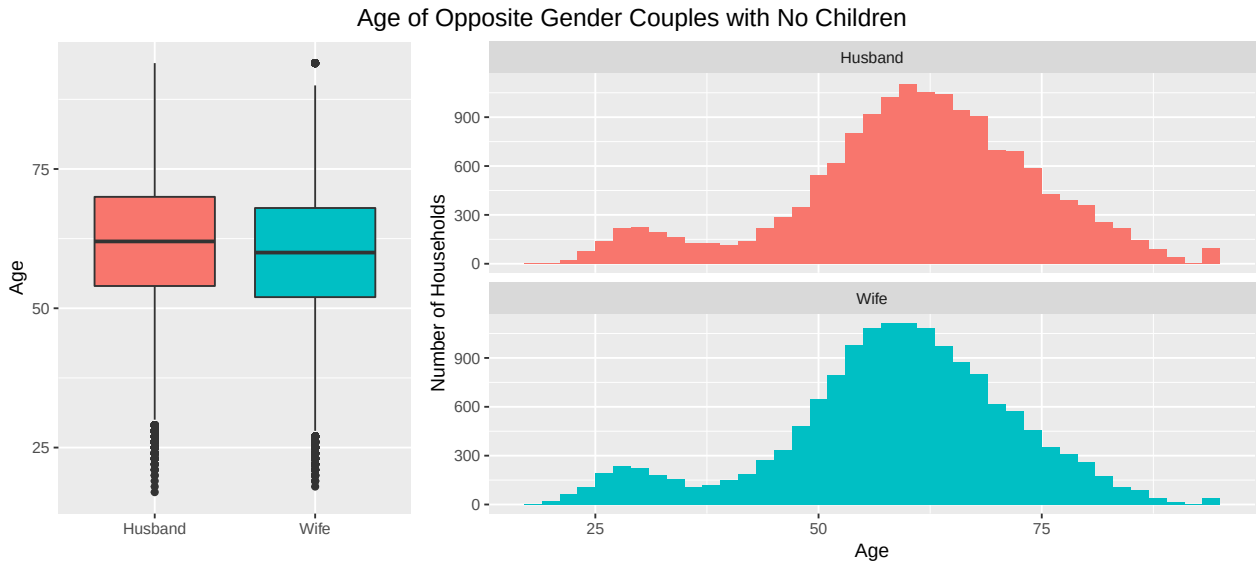
October 13, 2018

Introduction

The American Community Survey (ACS) is a large survey taken by the U.S Census Bureau in the years between the Census. The data of interest is a subset of public data from this survey in 2016. Particularly, this dataset contains households with opposite gender married couples. There are a few items of interest that can be gleaned from this dataset.

1. In households with no children, do husbands tend to be older than their wives? If so, by how much?
2. Do household in houses built in the 1960s or earlier spend more on electricity than those built in the 1970s or later? By house much?
3. Are households with higher electric and gas bills more likely to have a higher household income? If so, by how much?

```
noChildren <- data %>% filter(number_children == 0)  %>% select(age_husband, age_wife)
bucket <- list(Husband=noChildren$age_husband, Wife=noChildren$age_wife) %>% melt
hist <- ggplot(bucket, aes(x = value, fill = L1)) +
  geom_histogram(position = "stack", binwidth=2, show.legend = FALSE) +
  xlab("Age") + ylab("Number of Households") +
  guides(fill = guide_legend(title = "Age"))  +
  facet_wrap(~L1, ncol = 1)

bplot <- ggplot(bucket, aes(x = L1, y = value, fill = L1)) +
  geom_boxplot() +
  xlab("") +
  ylab("Age") +
  guides(fill = FALSE)

grid.arrange(bplot, hist,
             widths = c(1, 2),
             top = textGrob("Age of Opposite Gender Couples with No Children",
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```

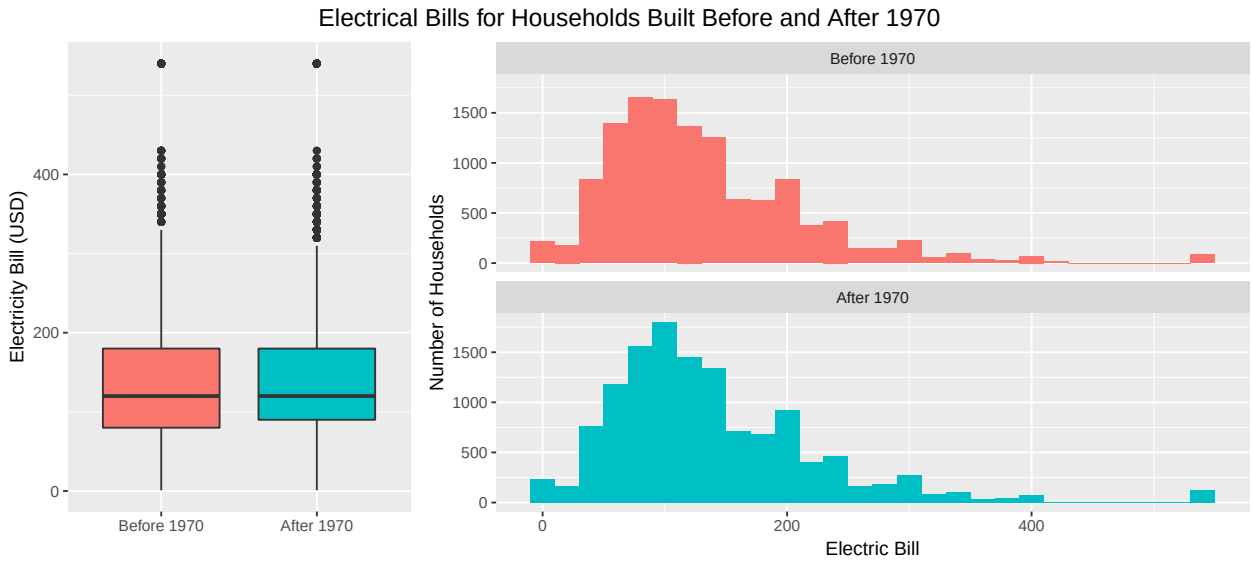



```
knitr::kable(bucket %>%
               group_by(L1) %>%
               summarize_all(funs(
                 mean,
                 sd,
                 min,
                 "25%"=quantile(value, probs = 0.25),
                 median,
                 "75%"=quantile(value, probs = 0.75),
                 max,
                 length)),
             col.names = c("", "Mean", "Std. Dev", "Min", "1st Quartile", "Median", "3rd Quartile", "Ma
             align = c('l'))
```

	Mean	Std. Dev	Min	1st Quartile	Median	3rd Quartile	Max	Count
Husband	61.17114	13.61846	17	54	62	70	94	15344
Wife	58.95633	13.23001	18	52	60	68	94	15344

```
hist <- ggplot(data,
               aes(x = electricity,
                   fill = factor(gte1970, labels = c("Before 1970", "After 1970")))
               ) +
  geom_histogram(position = "stack", binwidth = 20, show.legend = FALSE) +
  xlab("Electric Bill") + ylab("Number of Households") +
  guides(fill = guide_legend(title = "Decade Built")) +
  facet_wrap(~factor(gte1970, labels = c("Before 1970", "After 1970")), ncol = 1)

bplot <- ggplot(data,
                aes(x = factor(gte1970, labels = c("Before 1970", "After 1970")),
                    y = electricity,
                    fill = factor(gte1970, labels = c("Before 1970", "After 1970")))) +
  geom_boxplot() +
  xlab("") +
  ylab("Electricity Bill (USD)") +
```

```
  guides(fill = FALSE)

grid.arrange(bplot, hist,
             ncol = 2,
             widths = c(1, 2),
             top = textGrob("Electrical Bills for Households Built Before and After 1970",
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```

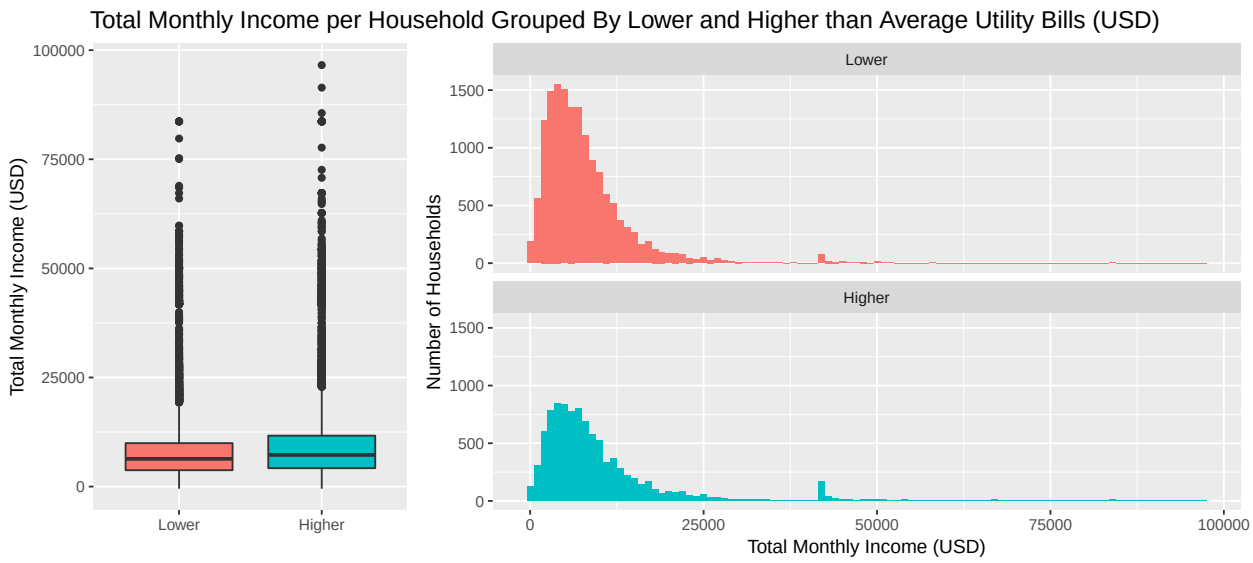



```
knitr::kable(data %>%
               group_by(if_else(gte1970, "After 1970", "Before 1970")) %>%
               summarize_at(c("electricity"),
                            funs(
                              mean,
                              sd,
                              min,
                              "25%"=quantile(electricity, probs = 0.25),
                              median,
                              "75%"=quantile(electricity, probs = 0.75),
                              max,
                              length)),
             col.names = c("", "Mean", "Std. Dev", "Min", "1st Quartile", "Median", "3rd Quartile", "Ma
             align = 'l')
```

	Mean	Std. Dev	Min	1st Quartile	Median	3rd Quartile	Max	Count
After 1970	142.2760	84.76693	1	90	120	180	540	12777
Before 1970	136.3112	81.84608	1	80	120	180	540	12331

```
data$total_monthly_income = (data$income_husband + data$income_wife) / 12
data$total_monthly_bills = data$gas + data$electricity

data$gte_mean_total_monthly_bills <-
  data$total_monthly_bills >= mean(data$total_monthly_bills)

hist <- ggplot(data,
```

```
                aes(x = total_monthly_income,
                    fill = factor(gte_mean_total_monthly_bills,
                                  labels = c("Lower", "Higher")))
                ) +
  geom_histogram(position = "stack", binwidth = 1000, show.legend = FALSE) +
  xlab("Total Monthly Income (USD)") + ylab("Number of Households") +
  guides(fill = guide_legend(title = "Higher or Lower than Average")) +
  facet_wrap(~factor(gte_mean_total_monthly_bills, labels = c("Lower", "Higher")), ncol = 1)


bplot <- ggplot(data,
                aes(x = factor(gte_mean_total_monthly_bills,
                               labels = c("Lower", "Higher")),
                    y = total_monthly_income,
                    fill = factor(gte_mean_total_monthly_bills, labels = c("Lower", "Higher")))
                ) +
  geom_boxplot() +
  xlab("") + ylab("Total Monthly Income (USD)") +
  guides(fill = FALSE)

grid.arrange(bplot, hist,
             ncol = 2,
             widths = c(1, 2),
             top = textGrob("Total Monthly Income per Household Grouped By Lower and Higher than Average
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```




```
knitr::kable(data %>%
               group_by(if_else(gte_mean_total_monthly_bills, "Higher", "Lower")) %>%
               summarise_at(c("total_monthly_income"),
                            funs(
                              mean,
                              sd,
                              min,
                              "25%"=quantile(total_monthly_income, probs = 0.25),
                              median,
                              "75%"=quantile(total_monthly_income, probs = 0.75),
```

```
                            max,
                            length)),
             col.names = c("", "Mean", "Std. Dev", "Min", "1st Quartile", "Median", "3rd Quartile", "Ma
            align = 'l')
```

	Mean	Std. Dev	Min	1st Quartile	Median	3rd Quartile	Max	Count
Higher	9984.261	9961.289	-491.6667	4216.667	7250	11666.67	96583.33	9715
Lower	7926.211	6926.901	-491.6667	3750.000	6350	9950.00	83666.67	15393

Methods

In households with no children, do husbands tend to be older than their wives? If so, by how much?

Given that the ACS dataset has information by household and each household contains both the husband's and wife's age, a paired t-test can be done. A paired t-test can be performed under the following assumptions: the sample sizes and Std. Devs are the same, and the households were randomly sampled. Per the summary table for the first set of graphs, both sample sizes and Std. Devs are approximately the same. The latter assumption is true since this dataset is a random sample of a different, larger dataset which is assumed to be sampled in such a way to make it statistically relevant.

Do household in houses built in the 1960s or earlier spend more on electricity than those built in the 1970s or later? By house much?

The appropriate test is a Welch's two-sampled t-test. This test can be performed under the assumption that the samples are random and without the assumption that the Std. Devs are equivalent. There is moderate evidence that the Std. Devs for monthly electricity bill between houses constructed before and after 1970 are not equivalent (Levene's Test. P-value = 0.01665). Since this is the case, Welch's two-sample t-test must be used instead of the standard two-sample t-test. The reasoning applied above for using independence is still valid for this case.

```
results <- levene.test(data$electricity, data$gte1970)
knitr::kable(
  tidy(results) %>%
    select(p.value),
  digits = 4,
  col.names = c("P-Value"),
  align = 'l',
  caption = "Levene's Test for Electricity and Construction Date Before & After 1970")
```

Table 4: Levene's Test for Electricity and Construction Date Before & After 1970

P-Value
0.0166

Are households with higher electric and gas bills more likely to have a higher household income? If so, by how much?

The appropriate test for this question would be a regression test. Unfortunately, that has not yet been covered in class so the Welch's two-sample t-test will be applied instead. The concept of 'higher utility bills' is calculated by whether or not the Total Monthly Income for a household is greater than or equal to the mean of all Total Monthly Incomes.

In order to determine whether or not the underlying populations are normal, a Shapiro-Wilk test is performed. Unfortunately, the Shapiro-Wilk test in R can only contain a maximum of 5000 subjects. There is convincing evidence that both populations are not normally distributed (Shapiro-Wilk, P-value < 2.2e-16). However, Welch's t-test can still be applied as it is robust to the non-normality assumption under large sample sizes.

```
resultsGreaterThan <- sample(
  subset(data, data$gte_mean_total_monthly_bills)$total_monthly_income, 5000
  ) %>% shapiro.test
resultsLessThan <- sample(
  subset(data, data$gte_mean_total_monthly_bills)$total_monthly_income, 5000
  ) %>% shapiro.test

knitr::kable(
  tidy(resultsGreaterThan) %>%
    select(p.value),
  digits = 4,
  col.names = c("P-Value"),
  align = 'l',
  caption = "Shapiro-Wilk Test for Total Monthly Income Higher than Average Utility Bills")
```

Table 5: Shapiro-Wilk Test for Total Monthly Income Higher than Average Utility Bills

P-Value
0

```
knitr::kable(
  tidy(resultsLessThan) %>%
    select(p.value),
  digits = 4,
  col.names = c("P-Value"),
  align = 'l',
  caption = "Shapiro-Wilk Test for Total Monthly Income Lower than Average Utility Bills")
```

Table 6: Shapiro-Wilk Test for Total Monthly Income Lower than Average Utility Bills

P-Value
0

Summary

In households with no children, do husbands tend to be older than their wives? If so, by how much?

For households with no children, there is convincing evidence that the difference between the age of husbands and the age of their wives is greater than zero (Two Sample t-test. P-value < 2.2e-16). It is estimated that husbands tend to be 2.215 years older than their wives. With 95% confidence, Husbands are on average between 1.914 and 2.515 years older than their wives.

Given the size of this sample, one could say that these results are applicable to all opposite gender married households in Illinois.

```
results <- t.test(noChildren$age_husband, noChildren$age_wife, var.equal = T)

knitr::kable(
  tidy(results) %>%
    select(estimate1, estimate2, p.value, conf.low, conf.high) %>%
    mutate(estimate=estimate1 - estimate2),
  digits = 3,
  col.names = c("Husband Mean Age", "Wife Mean Age", "P-Value", "C.I - Lower", "C.I - Upper", "Estimate
  align = 'l')
```

Husband Mean Age	Wife Mean Age	P-Value	C.I - Lower	C.I - Upper	Estimate
61.171	58.956	0	1.914	2.515	2.215

Do household in houses built in the 1960s or earlier spend more on electricity than those built in the 1970s or later? By house much?

There is convincing evidence that the difference for electricity bills between houses built in the 1960's or earlier are different than houses build in the 1970's or after (two-sample t-test. P-value = 1.45e-08). It is estimated that households built in the 1960's or earlier pay $5.97 less per month than households built in the 1970's or after. With 95% confidence, households built in the 1960's or before on average pay between $3.90 and $8.03 less per month than households built in the 1970's or after.

Given the size of this sample, one could say that these results are applicable to all opposite gender married households in Illinois.

```
results <- t.test(electricity ~ factor(gte1970, labels = c("Before", "After")), data = data)

knitr::kable(
  tidy(results) %>%
    select(estimate1, estimate2, p.value, conf.low, conf.high) %>%
    mutate(estimate=estimate1 - estimate2),
  digits = 3,
  col.names = c("Before 1970 Mean Cost", "After 1970 Mean Cost", "P-Value", "C.I - Lower", "C.I - Upper
  align = 'l')
```

Before 1970 Mean Cost	After 1970 Mean Cost	P-Value	C.I - Lower	C.I - Upper	Estimate
136.311	142.276	0	-8.026	-3.904	-5.965

Are households with higher electric and gas bills more likely to have a higher household income? If so, by how much?

There is convincing evidence that the difference between average monthly incomes from housesholds that pay higher utilities than average monthly incomes from households who pay lower utilities (Welch's two-sample t-test. P-value < 2.2e-16). It is estimated that households who pay higher than average utilities earn $2058.41 per month more than households who pay lower than average utilities. With 95% confidence, households who pay higher than average utilities earn between $1831.74 and $2284.36 more per month than households who pay lower than average utilities.

Given the size of this sample, one could say that these results are applicable to all opposite gender married households in Illinois.

```
results <- t.test(data$total_monthly_income ~ factor(data$gte_mean_total_monthly_bills, labels = c("Low
                  data = data)
```

```r
knitr::kable(
  tidy(results) %>%
    select(estimate1, estimate2, p.value, conf.low, conf.high) %>%
    mutate(estimate=estimate1 - estimate2),
  digits = 3,
  col.names = c("Lower than Avg Mean Income", "Higher than Avg Mean Income", "P-Value", "C.I - Lower",
  align = 'l')
```

Lower than Avg Mean Income	Higher than Avg Mean Income	P-Value	C.I - Lower	C.I - Upper	Estimate
7926.211	9984.261	0	-2284.364	-1831.735	-2058.049

* *Note that p-values appear as 0 due to rounding*