STATS1: Second Group Project

Aliza, Sat, Paloma

Due Friday, April 19

By completing this group project, you will be able to:

- Generate a question that can be answered with provided data.
- Perform EDA for two (binary) categorical variables.
- Perform inference for one proportion and two proportions
- Write accurate inferential conclusions and interpretations for each of these procedures

Useful resources for this homework:

- Book chapters 16 and 17.
- In class activities from Unit 4

Introduction

As in Part 1 of the project we will be using one of three datasets provided for you.

- Data Set 1: Cook County Housing Data
- Data Set 2: Canine Assisted Interventions
- Data Set 3: Predicting Classroom Performance

The datasets are available under the folder *Stats 172 S24/Class/Project/Project_data*.

Descriptions of the three datasets are available under the folder

Stats 172 S24/Class/Project/Project_data_sets.Rmd.

```
library(tidyverse)
library(ggplot2)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
#Download cook county housing data
cook_county <- read_csv("~/Stats 172 S24/Class/Project/Project_data/group1_cook_county_data_Project2.csv
```

```
## Rows: 1000 Columns: 27

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (14): pin, construction_quality, garage_attached, basement_type, centra...
## dbl  (12): township_code, num_bedrooms, num_fireplaces, num_full_baths, num_...
## date  (1): sale_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Overview

You will be exploring at least three binary categorical variables. You are encouraged to select variables that can contribute to a broader narrative related to the dataset you chose. You can build on questions your identified in the first project submission. A strong conclusion should reflect connections between the variables you choose for this portion of the project.

0. Clearly identify the dataset your group will be working with. You may continue with the same dataset or choose a different one to explore. Load in that dataset and provide a brief overview of origin of the data.

We will be using Cook County Housing Data dataset for this analysis.

The dataset covers a wide range of properties sold within Cook County, having detailed information about property sales in the Chicago area, including both the square footage, year the properties were built, sale price, and many others. The dataset is publicly available, hosted on GitHub

Section A: One Proportion

1. Select one of your three binary categorical variables and identify a question that can be answered with it. Clearly state this question for a general audience and explain the variable in context of the data collection.

`central_air`: Central A/C, Yes/No Question: What is the proportion of houses that have Central A/C?

2. For your variable provide appropriate summary statistic(s) and then describe those statistic(s) in context.

```
cook_county$central_air<-ifelse((cook_county$central_air=="Central A/C"), "Yes (A/C)", "No (A/C)")

table(cook_county$central_air)
```

```
##
##  No (A/C) Yes (A/C)
##       447       553
```

```
table(cook_county$central_air) %>%
  proportions() %>% round(3)
```

```
##
##  No (A/C) Yes (A/C)
##     0.447     0.553
```

In the data collected from housing in Cook County (Chicago Area) around 45% of the houses does not have central A/C and 55% does have it.

3. State the null and alternative hypothesis in symbols and in words for an appropriate statistical test.

$$H_0 : p = 0.5$$

50% of houses in Cook County have central A/Cs

$$H_A : p \neq 0.5$$

the percentage of houses that have central A/C is different than 50%

4. Run your hypothesis test, report your test statistic and p-value, and state your statistical conclusion in context.

```r
prop.test(x=553, n = 447+553, p=0.5, alternative="two.sided")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  553 out of +553 out of 447553 out of 553
## X-squared = 11.025, df = 1, p-value = 0.0008989
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5215390 0.5840488
## sample estimates:
##     p
## 0.553
```

```r
n<-447+553
p_est<-0.553
null<-0.5

x<-(p_est-null)

SE<-(sqrt((x/n)))

test_stats<-(x/SE)

test_stats
```

```
## [1] 7.28011
```

The p-value is significant smaller than the threshold, therefore we reject the null hypothesis and favor of the alternative hypothesis. The percentage of houses that have central A/C in Cook county is different than 50%.

The test statistic of approximately 7.28 indicates that the observed proportion of houses with central A/C in Cook County is 7.28 standard errors away from the hypothesized proportion of 50%. A test statistic like this is considered to be highly significant, which suggests that the true proportion of houses with central A/C in Cook County is very unlikely to be the 50% predicted by the null hypothesis.

5. Interpret an appropriate 95% confidence interval in context.

We are 95% confident that the percentage of houses at Cook county that have central A/C is between 52% and 58%. Therefore, we can conclude that the percentage of houses that have central A/C in Cook county is not only different than 50% but is grater than 50%.

6. Comment on the role of assumptions/conditions for the test.

Based on the dataset description, we assume that the observations in our dataset are independent of each other, meaning that the presence or absence of central air conditioning in one house does not influence the presence or absence in another. Additionally, we ensure that our sample size is sufficiently large, with at least 10 houses having central air conditioning and 10 houses without it. These conditions being met are essential for the proper application of the mathematical model that we used in this analysis.

Section B: Two Proportions

1. Identify a question that can be answered with two binary categorical variables in the dataset. Clearly state this question for a general audience, identify the explanatory and response variable, and explain the two variables in context of the data collection.

`central_air`: Central A/C, Yes/No - explanatory variable

`porch`: Porch status ("None" or "Finished Porch") - response variable

Question: Is there a relationship between the presence of air conditioning systems and the presence of porches in houses within Cook County?

We are investigating whether there exists a correlation between the presence of central air conditioning and the presence of porches in houses located in Cook County. Our objective is to determine whether the installation of central air conditioning is associated with the presence of porches in the houses.

2. Explore and describe the relationship between the two variables with appropriate summary statistics and visualizations. Make sure to describe the results of your summary and visualization.

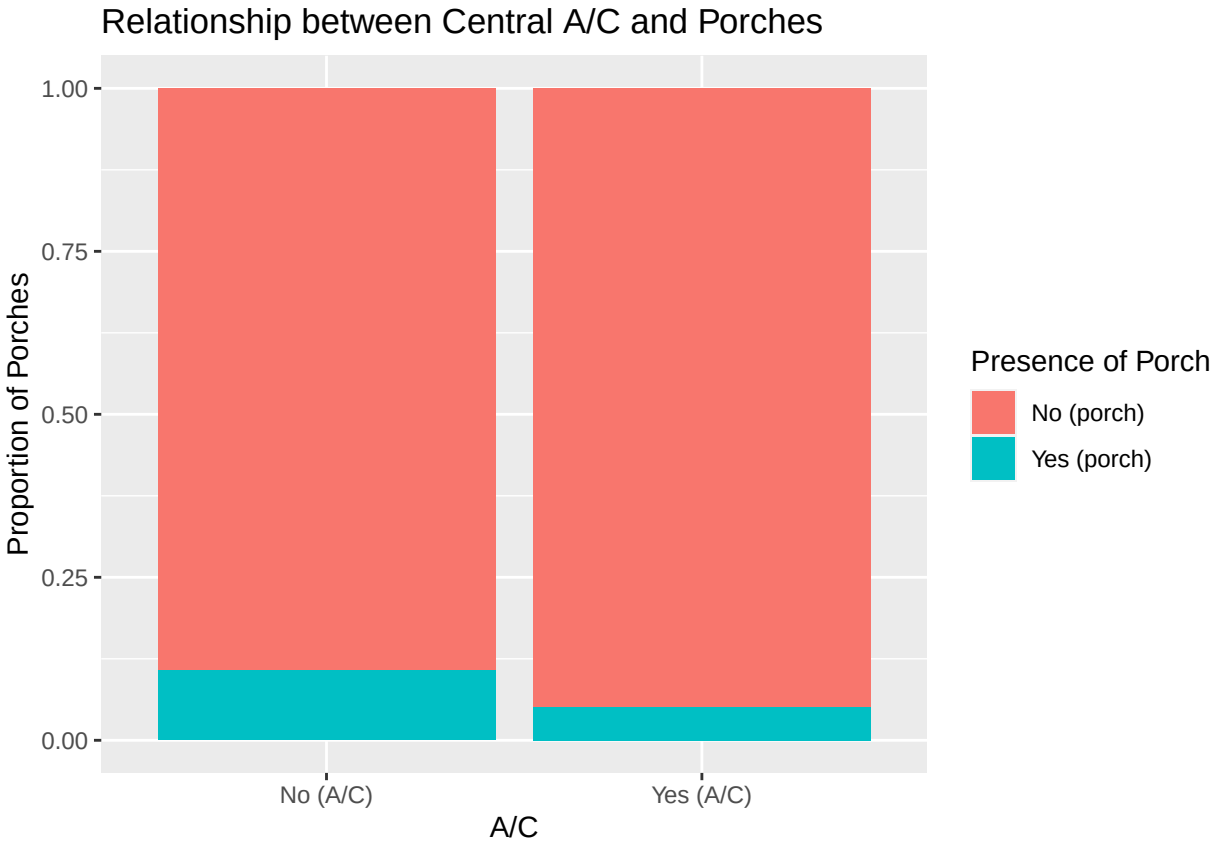
```
cook_county$porch<-ifelse((cook_county$porch=="Finished porch"), "Yes (porch)", "No (porch)")
table(cook_county$central_air, cook_county$porch) |>
  proportions()
```

```
##
##             No (porch) Yes (porch)
##   No (A/C)       0.399       0.048
##   Yes (A/C)      0.525       0.028
```

```
table(cook_county$central_air, cook_county$porch) %>%
  addmargins()
```

```
##
##             No (porch) Yes (porch)  Sum
##   No (A/C)         399          48  447
##   Yes (A/C)        525          28  553
##   Sum              924          76 1000
```

```
ggplot(cook_county, aes(x = central_air, fill = porch)) +
  geom_bar(position = "fill") +
  labs(title = "Relationship between Central A/C and Porches",
       x = "A/C",
       y = "Proportion of Porches",
       fill = "Presence of Porch")
```



In order to have a better understanding of the variables we were analyzing and to explore the relationship between central air conditioning (A/C) and the presence of porches, we first transformed the porch variable into a binary category: "Yes (porch)" and "No (porch)." Then, we examined the proportions of houses with central A/C that have or not porches.

The table of proportions indicates that the proportion of houses with central A/C and a porch is lower than the proportion of houses with central A/C but no porch. The bar chart we used, shows that a smaller proportion of houses with central A/C have porches compared to those without central A/C.

3. State the null and alternative hypothesis in symbols and in words for an appropriate statistical test.

$H_0 : p_{porch} - p_{cool} = 0$ There is no difference in the proportion of houses with A/C that have a porch and the proportion of houses with A/C that does not have a porch at Cook County

$H_A : p_{porch} - p_{cool} \neq 0$ There is a difference in the proportion of houses with A/C that have a porch and the proportion of houses with A/C that does not have a porch at Cook County

4. Run your hypothesis test, report your test statistic and p-value, and state your statistical conclusion in context.

```
#prop.test(x = c(28,48), n = c(553,447), conf.level = 0.95, alternative = "two.sided")
mosaic::prop.test(porch~central_air, data=cook_county, conf.level = 0.95, alternative = "two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
```

```
## data:  tally(porch ~ central_air)
## X-squared = 10.543, df = 1, p-value = 0.001167
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.09279665 -0.02070262
## sample estimates:
##    prop 1    prop 2
## 0.8926174 0.9493671
```

```
sqrt(10.543)
```

```
## [1] 3.246999
```

The p-value is significantly smaller than the threshold, being 0.001167, therefore we to reject the null hypothesis. Thus, there is a significant difference in proportion of houses with with A/C and the houses with porches at Cook County.

The test statistic of approximately 3.25 indicates that the observed difference between the sample proportions of houses with central A/C that have a porch and house with centra A/C that does not have a house in Cook County is 3.25 standard errors away from the expected difference under the null hypothesis. A test statistic like this is considered to be significant, which suggests that the true proportion of houses with central A/C in Cook County is quite unlikely to be the predicted.

5. Interpret an appropriate 95% confidence interval in context.

We are 95% confident that the difference in proportions of houses with A/C and the roof material is between -0.09 and -0.02 at Cook County. Given that the range between the lower and upper bound does not include zero, being negative, we can conclude that there is a slightly lower proportions of houses with A/C that have a porch then house with A/C that do not have a porch at Cook County.

6. Comment on the role of assumptions/conditions for the test.

The dataset does not reveal any evidence of dependence between variables, nor is there documentation indicating paired variables. Consequently, we infer that the presence of air conditioning systems is not directly determinative of the presence of porches. We confirm the independence of data within and between groups. Additionally, both variables we used have a minimum of 10 observations for each of their levels, ensuring the appropriate application of the statistical model.

Section C: Two Proportions

Identify a new explanatory variable (with the same response) or a new response variable (with the same explanatory) and repeat Part B. When choosing variables consider the broader narrative of your questions and how these questions relate to each other.

`central_air`: Central A/C, Yes/No - explanatory variable

`garage_attached`: Garage attached to home? Yes/No - response variable

Question: Is there a relationship between the presence of air conditioning systems and the presence of garages in houses within Cook County?

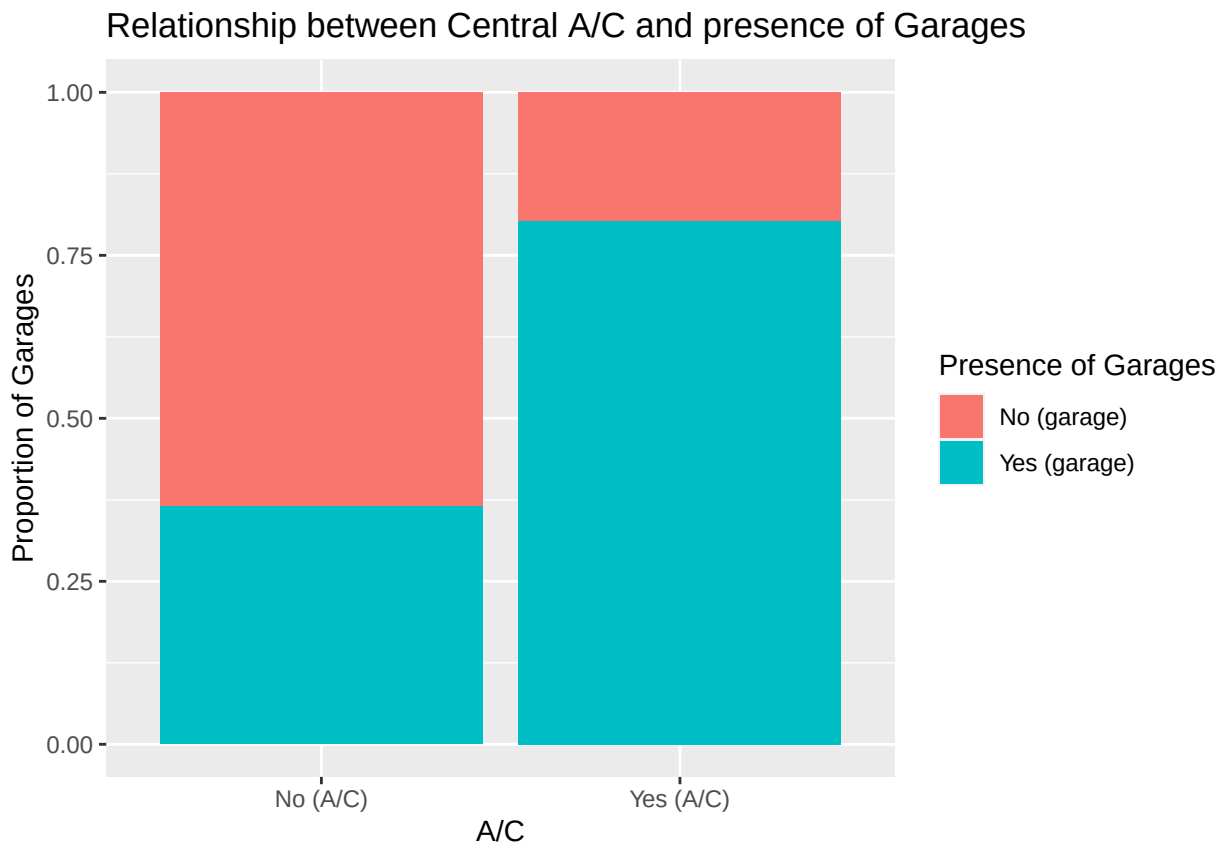
```
cook_county$garage_attached<-ifelse((cook_county$garage_attached=="Yes"), "Yes (garage)", "No (garage)")

table(cook_county$central_air, cook_county$garage_attached) |>
  proportions()
```

```
##
##             No (garage) Yes (garage)
##   No (A/C)        0.284        0.163
##   Yes (A/C)       0.109        0.444
```

```
table(cook_county$central_air, cook_county$garage_attached) %>%
  addmargins()
```

```
##
##             No (garage) Yes (garage)  Sum
##   No (A/C)          284          163  447
##   Yes (A/C)         109          444  553
##   Sum               393          607 1000
```

```
ggplot(cook_county, aes(x = central_air, fill = garage_attached)) +
  geom_bar(position = "fill") +
  labs(title = "Relationship between Central A/C and presence of Garages",
       x = "A/C",
       y = "Proportion of Garages",
       fill = "Presence of Garages")
```



Again we needed to transform the garage_attached variable into a binary category: "Yes (garage)" and "No (garage)" in order to have a better understanding of the variables we were analyzing and to explore the relationship between central air conditioning (A/C) and the presence of garages. Then, we examined the proportions of houses with central A/C that have or not garages attached

The table of proportions indicates that the proportion of houses with central A/C and a garage is higher than the proportion of houses with central A/C but no garage attached. The bar chart we used, shows that a higher proportion of houses with central A/C have garages compared to those without central A/C.

- The null and alternative hypothesis

$H_0 : p_{cool} - p_{garage} = 0$ There is no difference in the proportion of houses with A/C that have garages attached and the proportion of houses with A/C that does not have garages attached at Cook County

$H_A : p_{cool} - p_{garage} \neq 0$ There is a difference in the proportion of houses with A/C that have garages attached and the proportion of houses with A/C that does not have garages attached at Cook County

- Hypothesis test and test statistics

```
#prop.test(x = c(444, 163), n = c(553, 447), conf.level = 0.95, alternative = "two.sided")
prop.test(garage_attached~central_air, data=cook_county)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(garage_attached ~ central_air)
## X-squared = 197.18, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.3806262 0.4958539
## sample estimates:
##    prop 1    prop 2
## 0.6353468 0.1971067
```

```
sqrt(197.18)
```

```
## [1] 14.04208
```

The p-value is extremely small ($2.2e-16$) which leads us to reject the null hypothesis. Thus, we conclude that there is a significant difference in the proportions of houses with attached garages and houses with central A/C at Cook County.

The test statistic of approximately 14.04 indicates that the observed difference between the sample proportions of houses with central A/C that have a garage and house with centra A/C that does not have a garage in Cook County is 14.04 standard errors away from the expected difference under the null hypothesis. A test statistic like this is considered to be highly significant, which suggests that the true proportion is very unlikely to be the predicted.

- Interpret an appropriate 95% confidence interval in context.

We are 95% confident that the different in proportions of houses with central air and garage attached is between 0.38 and 0.49 at Cook County. This interval does not contain zero, further supporting the rejection of the null hypothesis and indicating that the proportion of houses with central A/C that have a garage is slightly higher than the proportion of houses with A/C that don't have a garage at Cook County.

- Comment on the role of assumptions/conditions for the test.

Prior to conducting our analysis, we ensured the independence of data both within and between groups. Moreover, each variable utilized in our analysis contained a minimum of 10 observations for every level, ensuring the ensuring the appropriate application of the statistical model. The dataset exhibited no visible evidence of interdependence between variables, nor did it indicate the presence of paired variables. Consequently, we infer that the presence of air conditioning systems does not directly dictate the presence of garages.

Section D: Conclusion

Write a paragraph summarizing your findings from your 3 analyses to a general audience. In addition to summarizing conclusions, you should consider how broadly the results apply (generalizability), whether or not there may be confounding variables (causation), and if there are any shortcomings or limitations to the study or its conclusions. When presenting findings to a general audience it is common to report confidence intervals and statements of significance (or not) but you should refrain from using words like "null", "alternative", "p-value", "test-statistic", etc.

Our study shows the prevalence of central air conditioning systems in homes across Cook County and their relationship with other housing features. We found that a significant proportion of homes with central air conditioning. Interestingly, our analysis revealed that the presence of a porch does not appear to be correlated with the presence of central air conditioning, suggesting that these two features may be independent of each other. However, we observed an association between central air conditioning and the presence of garages, indicating a potential link between these variables.

While our findings provide some insights into housing trends in Cook County, it's essential to acknowledge some limitations of this analysis. Unaccounted confounding variables may influence observed relationships, and the results might not apply to other regions, as housing trends can vary significantly between different areas. Furthermore, our analysis focused on a limited set of variables, and there may be other factors that could impact the relationships we observed.

Despite these limitations, this analysis uncovers trends in overall house appliances and their relationships.

Knit this file. Submit your pdf to Moodle.