# Experimental_Methodology

*Mark Paluta, Krysten Thompson, Chris Ventura*

*April 19, 2019*

## Contents



# 1 Introduction / Background / Motivation

###–> CHRIS, can you link any of comments below to the articles under "References"? I see there's reference to NYT article but not sure if others referenced below. (You are likely most familiar w article content.)

In a recent study, Zillow found that the share of adults living with a non-spouse, non-partnered roomate increased from 21% in 2005 to 30% at the end of 2017[1]. Whether it is due to changing attitudes about living with roommates, a necessity due to increased rents in major cities, or some other factor, the trend is clear: more Americans are cohabitating with someone other than a spouse or partner. While college and post-college aged urbanites often expect to have roommates, the trend is increasing for those older than 30, as a recent New York Times profile depicts[2]. As employment opportunities continue to cluster in dense metropolitan areas such as New York and San Francisco, an older cohort may be forced to cohabitate in order to pursue economic opportunity.

In spite of this demographic trend, little research has been done into whether the age of someone looking for a place to live with roommates has an impact on their options. As the New York Times profile describes,

until recently the idea of having a roommate was mostly relegated to young professionals in their 20s[2]. Furthermore, in the United States it is not illegal to descriminate on the basis of age for roommates and other housing related decisions[3]. If age presents itself as a barrier to finding a roommate or living situation, this could add an additional obstacle to those beyond "typical" roommate age wishing to move to a dense metropolitan area to pursue employment, impacting economic opportunity for this age cohort. With this implication, our study sought to examine the existence of possible age bias in responses to roommate inquiries.

# 2 Research Question

Through this experiment, we set out to determine if post-college-age professionals face an age penalty when seeking a roommate. We investigate whether age affects the probability of receiving a "favorable reply" when responding to a roommate request posting on Craigslist. We hypothesized that older ages would generally receive fewer favorable replies than those shortly past typical college-age (mid-20's).

We define favorable replies as those that are 1) received within 72 hours of sending our email and 2) asking for more information or requesting next steps. An unfavorable reply is any other reply received within 72 hours. No response received within 72 hours is considered no reply.

# 3 Initial Experiment Idea

Our initial experiment sought to test both gender and age discrimination as well as their interaction. We created two email accounts for a female (age 27 and 43) and two email accounts for a male (age 27 and 43). We chose these ages to avoid round numbers, have the younger persona sufficiently past typical college-age, and have the older person over 40. We wanted the two treatment ages to be sufficiently different as to measure a material difference without pushing the extremes of old or young to still reasonably represent a large portion of roommate seekers.

We also considered running this experiment in several different cities in the U.S. to see if regional differences exist in terms of gender and age.

A pilot was conducted in early March to determine if our initial approach would work and identify potential obstacles. We discovered issues with Craigslist and Gmail fraud mechanisms that led us to modify our approach. See Pilot Study below for further details.

## 3.1 Pilot Study and Outcomes

We conducted a pilot study for three days to identify any problems that may surface during the full study. Major learnings from the pilot:

1) VPNs are difficult to use with Craigslist. We initially wanted to use a VPN to avoid being flagged as fraud by Craigslist. We tried several popular VPNs but Craigslist blocked all the associated IP addresses from sending emails.

2) Google sent one of our accounts a notification that it was shutting down one of our accounts because it detected the account was not being used according to Google's guidelines.

We did not want to send email on behalf of multiple personas from the same IP address for fear of fraud detection. Given this preference and the additional learnings from the pilot study, we concluded that maintaining too many personas would be difficult and decided to limit our full study to 2 personas. We were able to each imitate two separate individuals by using separate devices and/or a different internet connection such as a mobile hotspots. We anticipated that response rates could be low and thought females might have higher response rates, so we elected to use two female personas. In the pilot study, more postings seemed to

be open to females, and we speculated that males would be more willing to live with females than females would be willing to live with males.

Not all posts seemed appropriate for measuring effects on a typical applicant. The following types of posts were excluded:

a) Posts with sexual implications including "friends with benefits" or reduced rent in exchange for sex
b) Posts soliciting other favors in exchange for reduced rent such as babysitting
c) Posts for students only (as we were targeting a conclusion for working professionals)
d) Posts without an email option or posts explicitly stating that email replies will not be answered
e) Posts explicitly stating males only

# 4 Experiment Design & Methodology

The experiment design consisted of sending emails to listings under the Craigslist heading "Housing" > "Rooms/Shared".

In order to be able to draw conclusions with the best chance of generalizing to the US population, Indianapolis was selected as the target market. Indianapolis is a mid-sized city located in the midwest.

We sent emails every 1-3 days across a two-week time period. We were not particularly concerned with the exact times or days we collected data since we were collecting a number of time-based covariates in order to model any temporal effects. We did make a conscious effort to cover a broad range of variation in our temporal covariates by sending emails on every day of the week and including a mix of morning, afternoon, and evening collections.

The following procedure was implemented each time emails were sent:

1) Open all postings since the last data collection (or fewer if subject to time constraints, prioritizing recent postings)
2) Go through each post and throw out any that meet our exclusion criteria
3) Obtain a count of the number of included (remaining) posts
4) Generate an array of this length, randomly assigning each post to one of our treatment conditions with equal probability
5) Beginning with the first persona in the array, log into their respective email account and begin emailing posters in order from most recent to least, recording covariates as we go

Emails were sent only to listings that were new, or listings that had been posted days or weeks ago but were "refreshed" in the postings list. The data was collected manually in a spreadsheet. See the Data section for all variables collected.

Only responses received within 72 hours were flagged as a legitimate reply. We chose 72-hours as an appropriate window of time to account for people on vacation or possibly having something at work prevent them from responding within 24 hours. If a response to our treatment arrived after 72 hours, the response was considered a "no reply".

**Clear statement of the experiment**

```
Experimental materials (e.g. treatment materials)
Measurement of variables
Modeling choices
```

**end**

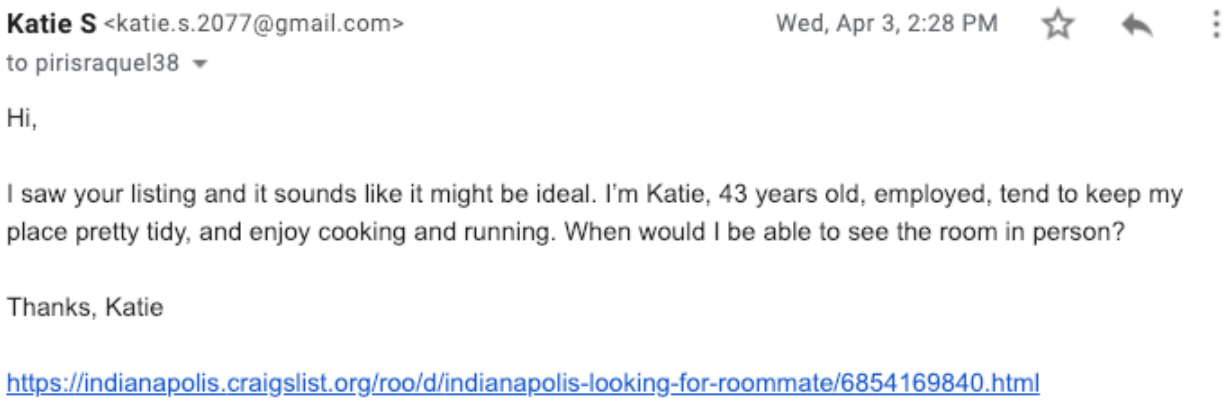
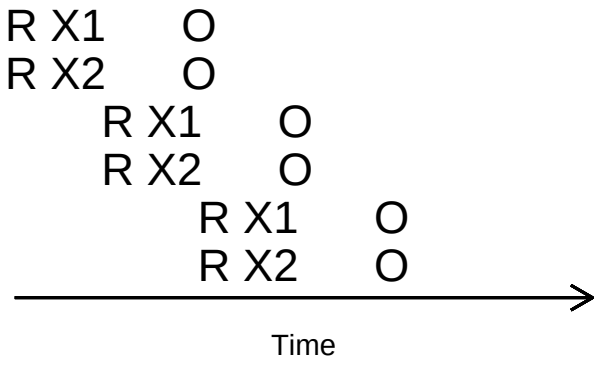



Figure 1: Katie Email

### 4.1 ROXO Grammar

Treatment was applied at various times since our data was collected over several weeks. We did not always have observations on the previous cohort (group of 10-20 emails sent) prior to treating a new cohort due to our impossed 72 hour waiting period. This was not a concern as we believed there to be little chance posters would interact with each other.

We demonstrate these characteristics using ROXO grammar in the following diagram. Each cohort is randomized, treated, and sometime over the next 72 hours observed.




## 5 Data

### 5.1 Data Types

The following variables were collected: spanning posting metadata, contents (for archiving purposes), time information (for temporal effects), covariates, and treatment applied. As we were not sure how clean our data would be and what would prove useful, we erred on the side of collecting more variables than would likely be needed.

| Variable | Description | Example | % Missing |
|---|---|---|---|
| Post Title | Title of the Craigslist post | Roommate/shared house ASAP | 0% |

| Variable | Description | Example | % Missing |
|---|---|---|---|
| Post Body | Body of the Craigslist post | ALL INCLUSIVE 10 minutes from downtown... | 0% |
| Post URL | URL of the Craigslist post | https://indianapolis.craigslist.org... | 0% |
| Poster Age | Age of the resident who posted | 34 | 81% |
| Poster Gender | Gender of the resident who posted | Female | 64% |
| Listing Type | Type of residence | Apartment | 8% |
| Listing Price | Price of the listing | 450 | 0% |
| Posting Timestamp | Timestamp that the post went up | 3/21/2019 22:38 | 0% |
| Email Sent Timestamp | Timestamp that we emailed | 3/22/2019 6:59 | 0% |
| Treatment | Which alias we emailed from | Katie 27 | 0% |
| Reply | Did we receive a reply? | Y | 0% |
| Favorable Reply | Did we receive a favorable reply? | Y | 0% |
| Reply Timestamp | Timestamp of a reply, if received | 3/23/2019 16:53 | 0% |
| Weekend | Was our email sent on a weekend? | Y | 0% |

In total, we sent 113 emails - 59 for "Katie 27"" and 54 for "Katie 43". Several variables were missing a significant portion of data and were note used in our models. In particular, we had hoped the variables 'poster age' and 'gender' could hold predictive value and reduce the standard error on our treatment effect. Our hypothesis was be that all else being equal, people prefer to live with those of the same gender and/or age. However, gender and age were each missing in over 50% of posts, so we excluded them from any analysis.

Since some replies we received did contain further information on the poster, which sometimes included a name, age, or gender, we considered filling in some missing data from replies where possible. We eventually decided against this for a few reasons. First, we still would not get close to 100% completeness on those variables and would need a separate category for "Unknown". Introducing imputations from responses might even need to be modeled as a separate category distinct from the same information as obtained directly from the post. We also believe that getting more complete covariate data from *the types of people who reply to our emails* would distort the relationship between that covariate and probability of a reply. This could bias the effect of the covariate and in turn bias our estimate of the treatment effect. Finally, since this information is not readily available to someone browsing Craigslist, it is of limited use to them because by the time they receive it, they already have their reply.

After examining our logged data, we looked for duplicate URLs to ensure that we hadn't accidentally emailed

the same poster more than once. This could have made the poster suspicious of our behavior, especially if we sent them otherwise identical emails with two different ages, but even in the event that we sent the same email twice. We did find four postings for which we had sent emails from both "Katie 27" and "Katie 43", a total of eight observations in our dataset. While a few of these emails garnered replies – in some cases for both "Katies" – we removed all eight records from our dataset as we could not assess any possible spillover that may have occurred. We briefly considered only throwing out the second email and keeping the first in the event that we had already received a response to our first email before sending our second. We decided against this as we did not want to selectively throwing out data on the types of people who did not respond quickly and favorably. The effect would have on average been balanced between our two treatments, but it would have biased our point estimates of response rates. Given our randomization scheme, we believe we can throw out all instances of duplicated subjects indiscriminately without imposing any bias into our experiment.

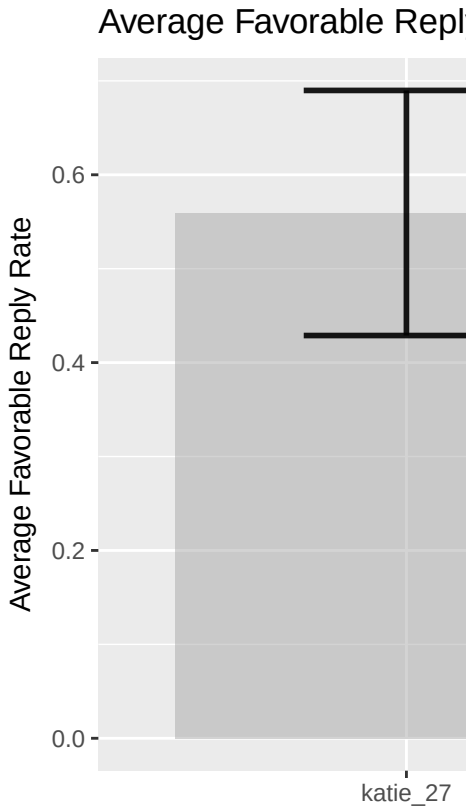
## 5.2 Data Transformations

**- Describe any transformations or cleanup steps that would be interesting to a reader**

We noticed a right-skew tendancy in our continuous variables. This is suggestive that multiplicative differences are more meaningful than additive differences. Given this observation we decided to perform log transformations to facilitate this interpretation. It also may aid our modeling because a lopsided clustering of data points at one end of the range of possible values could reduce the quality of our fit at the other end. This will to some extent even out the distribution of values at either end.

# 6 Results

## 6.1 Summary Results




The figure below shows the average rate of favorable replies by treatment specification:

As you can see, "Katie 27" had a slightly higher rate of favorable replies, however both rates are well within each others' margins of error. This is examined in greater detail in the regression section.

## 6.2 Regression

To estimate our treatment effects, we employed a variety of regression specifications. All of the regression specifications began with the following format:

$$Y = \beta_0 + \beta_1(Treatment) + \epsilon$$

For regressions with additional covariates, the specification included additional beta coefficients for each covariate.

$$Y = \beta_0 + \beta_1(Treatment) + \epsilon + \beta_2(Covariate_1) + ... + \beta_{n+1}(Covariate_n) + \epsilon$$

We used a linear regression for each of our four regression models, regressing the binary variable of a favorable reply on our treatment and, for the final three models, our chosen covariates. While our outcome variable is binary, using a linear regression allows us to directly measure the expected change in probability of getting a favorable reply, with the coefficients of the independent variables corresponding to the probability increase or decrease.

The first regression contains only the treatment variable.

The second regression contains treatment and the log of hours between post time and the email being sent. We will detail this in the next section, but we wanted to ensure the only statistically significant covariate had no impact on the treatment effect.

The third regression contains the treatment variable, the log of hours between post time and the email sent, and the binary indicator of the email being sent on the weekend. This regression specification is meant to help ensure that we randomized effectively and that varibales related to the listing type were not correllated with the treatment.

Finally, the fourth regression regresses favorable reply on the treatment, the log of list price, binary indicators for house and unknown list types (leaving apartments as the base case), the log of hours between the post time and email sent, and a binary value indicating if the treament or control email was sent on a weekend. Given that we have only 113 observations, we were wary of including all of our variables for fear of overfitting our regressions and to preserve degrees of freedom. A number of our variables were sparsely populated, so we selected these variables based on their near-complete observations as well as likelihood of explaining variations in a poster's response.

```
robust.se.1
```

```
##              (Intercept)        treatmentkatie_43          log(list_price)
##               1.10536435               0.09139384                0.17850417
##           list_typehouse         list_typeunknown log(hours_post_to_email)
##               0.10008533               0.18878794                0.02948244
##          sent_is_weekend
##               0.10082994
```

Examining the results of the four regression specifications, we see a fairly consistent treatment effect ranging from -0.04 to -0.02. This suggests that our randomization worked appropriately, and that the measured effect of being a 43 year old female as opposed to a 27 year old females was to reduce the probability of receiving a response to roommate listing inquiries by 0.02 to 0.04. In spite of this practical significance, however, this result lacked statistical signifiance for all four model specifications with p-values ranging from 0.68 to 0.81. Therefore, our effect was statistically indistinguishable from zero. We report robust standard errors to ensure that our reported standard errors are robust to any heteroskedasticity found in the data.

In fact, the only variable to show any statistical significance in our model specifications is the log of hours between the post time and the sending of our treatment or control email. The coefficients of -0.08 to -0.08 suggests that a 100% increase in response time corresponds with a -8% to -8% change in likelihood of a response. However, we should also be careful not to place too much weight in the significance of a covariate. Since hours from post to initial email was not the variable we experimented on, it is possible that it is being confounded by unmeasured variables.

## 6.3 Power Calculations

To better understand if and how our experiment might be lacking in power, we ran a 2-Sample Z-test and Power Tests. The Z-test indicates any differences in the proportions between Katie_27 and Katie_43 receiving a favorable reply. Z-test results:

**Estimated probability of success:**

| | |
|---|---|
| Katie_27 | $0.559$ |
| Katie_43 | $0.537$ |

**P-value:** 0.961

Table 2:

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | fave_reply | | | |
| | (1) | (2) | (3) | (4) |
| treatmentkatie_43 | −0.022 | −0.033 | −0.033 | −0.038 |
| | (0.094) | (0.091) | (0.092) | (0.091) |
| log(list_price) | | | | −0.028 |
| | | | | (0.179) |
| list_typehouse | | | | 0.099 |
| | | | | (0.100) |
| list_typeunknown | | | | 0.010 |
| | | | | (0.189) |
| log(hours_post_to_email) | | −0.077*** | −0.076*** | −0.077*** |
| | | (0.029) | (0.030) | (0.029) |
| sent_is_weekend | | | 0.004 | 0.016 |
| | | | (0.100) | (0.101) |
| Constant | 0.559*** | 0.807*** | 0.805*** | 0.919 |
| | (0.065) | (0.117) | (0.130) | (1.105) |
| Observations | 113 | 113 | 113 | 113 |
| $R^2$ | 0.001 | 0.060 | 0.060 | 0.069 |
| Adjusted $R^2$ | −0.009 | 0.043 | 0.034 | 0.017 |
| Residual Std. Error | 0.502 (df = 111) | 0.489 (df = 110) | 0.491 (df = 109) | 0.496 (df = 106) |
| F Statistic | 0.056 (df = 1; 111) | 3.522** (df = 2; 110) | 2.327* (df = 3; 109) | 1.319 (df = 6; 106) |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

**Confidence Interval at 95%:** $-0.179$ - 0.224

The high p-value of 0.96, combined with a confidence interval wide enough to have both estimates of the probability of success overlap, leads to a conclusion that the proportion of favorable replies between the two Katies' is not significantly different.

In fact, running a simple power test reveals a power of only 0.05, suggesting the probability of finding an effect, given the setup of the esperiment, was only 0.05. To identify what would have been necessary to detect a statistically significant effect, we then ran two power tests, one to calculate the number of observations we would need to obtain for *each* treatment given this effect size, the other to determine the required difference between the outcome variable (favorable replies) in order to observe a statistically significant effect from only 113 observations.

We would need 32434 subjects per treatment condition for our results to be statistically significant. In order to be able to observe a statistically significant effect in age on the outcome variable "favorable replies" using only 113 observations, the treatment effect would need to be 0.37, approximately 10-15 times our observed effect size, depending on the regression specification used. We anticipated a much larger effect size, which contributed to our lack of power in this experiment.

Power increases as sample size or effect size go up, and/or standard deviation is reduced. Because of the requirement of over $32,000$ observations needed for each treatment, in addition to being unsure the treatment effect would be as high as 37% , we conclude that for this particular research project, we do not find any meaningful difference in the rate of favorable replies based on age.

### 6.4 Randomization Inference

In addition, we also used randomization inference to examine the probability of calculating our treatment effect by chance. Randomization inference has the advantage of being a non-parametric method and as a result doesn't rely on large sample sizes or central limit theorem assumptions. Examining the difference in means, we calculate an average treatment effect of -0.0223. Running randomization inference with 1000 iteractions, we find 400 random iterations with at least as significant a treatment effect as our actual. This implies a p-value of 0.4, a non-statistically significant result. This is consistent with our linear regression results showing a non-statistically significant treatment effect and is roughly on the same order of magnitude as our p-values from regression and our z-test.

### 6.5 Causal Mechanism

If an effect existed, it could have been driven by a multitude of factors. Had we been able to capture age and gender of every poster, we may have seen a "birds of a feather flock together" pattern.

In hindsight, there may be confounding factors contributing to the effect such as demographics in Indianapolis, supply and demand of housing, and the unemployment rate (to name a few.)

When it comes to experiments of this nature, there are often unmeasurable factors that come in to play such as the lifestyle of the person looking for a roommate, unconscious bias toward dissimilar age groups, as well unconscious gender bias. Stereotypes such as "men are messier than women" exist as human perception and sometimes partial truths.

## 7 Improvement Opportunities7

In retrospect, had we found a concrete effect in this experiment, it would have been easy (and rewarding) to claim that some level of age bias exists. We also speculated on what would have driven an effect had we seen one.

If we saw an effect of 15%, we speculated that some level of age bias exists. While we would have wanted to conclude that the bias was a result of people generally wanting to reside with others of the same gender and age, the lack of data for these two variables would not have allowed us to draw that conclusion. Even having enough data for type of dwelling, rental price, and sent and reply timestamps, we're not convinced these variables play a determinant role in roommate selection.

If an effect of 30% was observed, we would conclude that something went wrong with our experiment. Seeing a 30% effect from only 113 observations is extreme. Note the Power Test indicated a necessary 37% effect to be statistically significant using only 113 observations. In either case, the results would be difficult to accept as valid.

Moving forward, an experiment executed using a similar methodology would require rigorous preparation. Considerations such the population demographics of the area of study, unemployment rates, renting/roommate trends, supply/demand of rental units would all have an impact on the outcome.

In addition, the manual collection of data is tedious at best, and often leaves room for error. Because of this, a more controlled experiment is recommended.

A controlled experiment could leverage an online platform that starts by providing a description of an apartment or house to participants. Participants will be informed that they have placed this ad on a free website and that they are looking for a roommmate. (The roommate would pay their equal share, etc.) Participants would then be asked to review email inquiries from potential renters and indicate how likely they (the participant) would be to respond to the email. Email inquiries would be virtually identical with the exception of age.

This manner of execution would also allow for blocking by participant gender and clustering by participant age. With a large enough, randomly selected participant group, other variables such as renter gender could also be tested.

Although specifics would need to be flushed out in greater detail, this type of experiment would provide a "contained" environment, potentially reduce the amount of confounding variables, and allow researchers to avoid manual data collection.

# 8 Conclusions

Based on the outputs of the regression analyses, we are unable to draw conclusions from this experiement. Although an impact between-0.04 to -0.02 was observed, the result lacked statistical signficance in all four of the models, as well as in our randomization inference scenario. In addition, the Power Tests indicated the need for either a much larger sample size, or a more pronounced effect. Over the course of approximately two weeks of data collection, we identified 113 roommate-wanted listings that conformed to our specifications (noted under Experiment Design). If we were to gather 32K+ observations for each treatment (or 64,000 observations total) in order to draw definitive conclusions, it would take 22.7 years of replying to roommate listings in Indianapolis (see math in Appendix).

This does not suggest that age bias doesn't exist. In the manner with which our experiment was executed, we agree that it is not possible to reach a conclusion and that another method to test for age, such as the design outlined under Improvement Opportunities be explored.

# 9 Appendix

Math calculation for Conclusion:

- 113 new, non-sexual, no 'male only' listings in a two week period
- 26 weeks per year
- Estimate 2,938 listings per year to receive a "Katie" email

- 64,000 observations needed / 2,938 listings per year = **22.7 years of listings**

# 10 References

[1] Bretz, Lauren, *As Rents Rise, More Renters Turn to Doubling Up*(2017), https://www.zillow.com/research/rising-rents-more-roommates-17618/ [2] Pappu, Sridhar, *Age 31 and Up, With Roommates. You Got a Problem With That?*(2016), https://www.nytimes.com/2016/05/06/fashion/mens-style/adult-men-roommates-new-york.html [3] Department of Justice, *Fair Housing Act*, (https://www.justice.gov/crt/fair-housing-act-1)