Homework Week 3

Steven Simonsen

2024-02-03

1.4: Buteyko method, study components. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.

(a) Identify the main research question of the study.

Does the Buteyko method reduce asthma symptoms and improve quality of life?

(b) Who are the subjects in this study, and how many are included?

The subjects are 600 asthma patients aged 18-69 who relied on medication for asthma treatment.

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

Variable 1 is quality of life, categorical/ordinal. Variable 2 is activity and is categorical/ordinal. Variable 3 is asthma symptoms and is categorical/ordinal. Variable 4 is medication reduction and is categorical/ordinal.Variable 5 is age, and is numerical, discrete. For variable 5, I contemplated whether this was a discrete or continuous variable. Given the details provided, age seems to be provided in whole numbers on a scale from 18-69. However, had age been given to decimal places, or even to a point in time in which a person was born, I would have considered it to have been continuous.

1.12 UN Votes: The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.

(a) List the variables used in creating this visualization.

Year, %Yes, Country, Arms control and disarmament, colonialism, economic development, human rights, nuclear weapons and materials, Palestinian conflict.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

Year: Numerical, discrete. %Yes: Numerical, continuous. Country: Categorical, nominal. Arms control and disarmament: Categorical, nominal. Colonialism: Categorical, nominal. Economic development: Categorical, nominal. Human rights: Categorical, nominal. Nuclear weapons and materials: Categorical, nominal. Palestinian conflict: Categorical, nominal.

1.34 Exercise and mental health: A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41-55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?

Experiment, since treatment is applied (exercise) to the subjects.

(b) What are the treatment and control groups in this study?

Treatment group: Half of the subjects from each age group instructed to exercise twice per week. Control group: The other half of the subjects instructed not to exercise.

(c) Does this study make use of blocking? If so, what is the blocking variable?

No, the subjects are not grouped into blocks based on variables other than the treatment, which in this case is exercise.

(d) Does this study make use of blinding?

No, subjects are aware they are in the control group because they are told not to exercise. There is no placebo.

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

I cannot reasonably conclude that the results of the study can be used to establish a causal relationship between exercise and health. Although the study randomizes the sample fairly well, it does not control for

variables such as how long each subject exercises, the intensity of the exercises, and the type of exercise such as cardiovascular and/or weight training. Additionally, the mental health of participants in each group may be affected due to knowing they are in the control group and treatment group (ex: A participant in the control group may suffer from negative mental health knowing they aren't exercising). For the same reasons, I would not feel comfortable generalizing the results of the study to the population at large.

(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

I would have reservations about funding the study. Better controls and better replication should be introduced into the study. A solution to improve in these areas could be to assign the same exercise intensity, type, and time to each subject. The study could be repeated to ensure sound results, if funding is available to do so. Blocking could be introduced to separate by suspected confounding variables as well, such as underlying diseases or conditions the participants may have.

2.16 Distributions and appropriate statistics, Part II: For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

Examining a histogram where count of houses is on the x-axis and home values in dollars are on the y-axis, I expect the data to be left skewed, as there are a meaningful number of houses far above the the Q3 point of the data. The median would better represent the data as a robust statistic because it is less effected by extreme values, and I would consider a meaningful number of houses costing above $6,000,000 to be extreme. For the same reason, I would expect the IQR as a better representation of the variability as it is less impacted by extreme values being a robust statistic.

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

I expect the data to by distributed relatively symmetrically, as the data seems to be fairly evenly distributed across Q1, the median, and Q3. Additionally, there are few houses costing more than $1,200,000. The mean would better represent the data as a non-robust statistic as there are few extreme values in the dataset, and it will accurately measure the center of the distribution of the data. For this same reason, I would expect the standard deviation to be a better measure of the data as a non-robust statistic, and it will accurately display how far the the typical observation is from the mean.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

Examining a histogram where count of drinks are on the x-axis, and number of students are on the y-axis, I expect the data to be right skewed since most students do not drink and only a few drink excessively. The median would better represent the data as a robust statistic since the data is somewhat extreme considering the majority of the values are populated at 0. For the same reason I would expect the IQR to be a better representation of variability because it is less impacted by many values populated at 0.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

Assuming the salary is on the x-axis of a histogram, and the count of employees is on the y-axis, I would expect this data to be right skewed. Since a few of the employees make a much higher salary, the mean may give an incorrect outlook of the data because it would be inflated by the outliers. Therefore, the median would be a better indicator of the dataset. Additionally, the IQR would better measure the variability for the same reason, as the standard deviation would be much more heavily impacted by the outliers in the dataset.

2.22 Views on immigration: 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.

(a) What percent of these Tampa, FL voters identify themselves as conservatives?

```
library(scales)
percent(372/910)
```

```
## [1] "41%"
```

(b) What percent of these Tampa, FL voters are in favor of the citizenship option?

```
percent(278/910)
```

```
## [1] "31%"
```

(c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

```
percent(57/910)
```

```
## [1] "6%"
```

(d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

```
percent(57/372) #conservative
```

```
## [1] "15%"
```

```
percent(120/363) #moderate
```

```
## [1] "33%"
```

```
percent(101/175) #liberal
```

```
## [1] "58%"
```

(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

```
percent(372/910) #Total number of voters that identify as conservative
```

```
## [1] "41%"
```

```
percent(363/910) #Total number of voters that identify as moderate
```

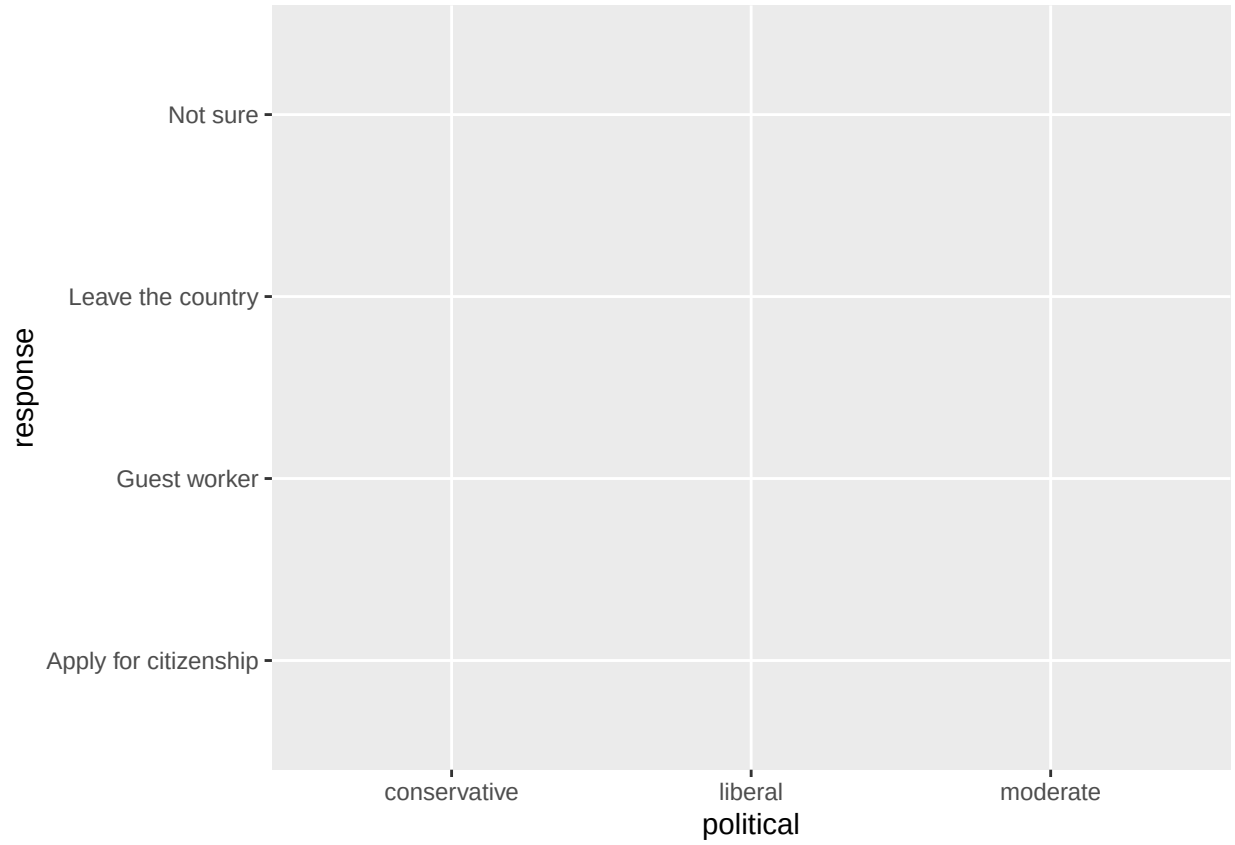
```
## [1] "40%"
```

```
percent(175/910) #Total number of voters that identify as liberal
```

```
## [1] "19%"
```

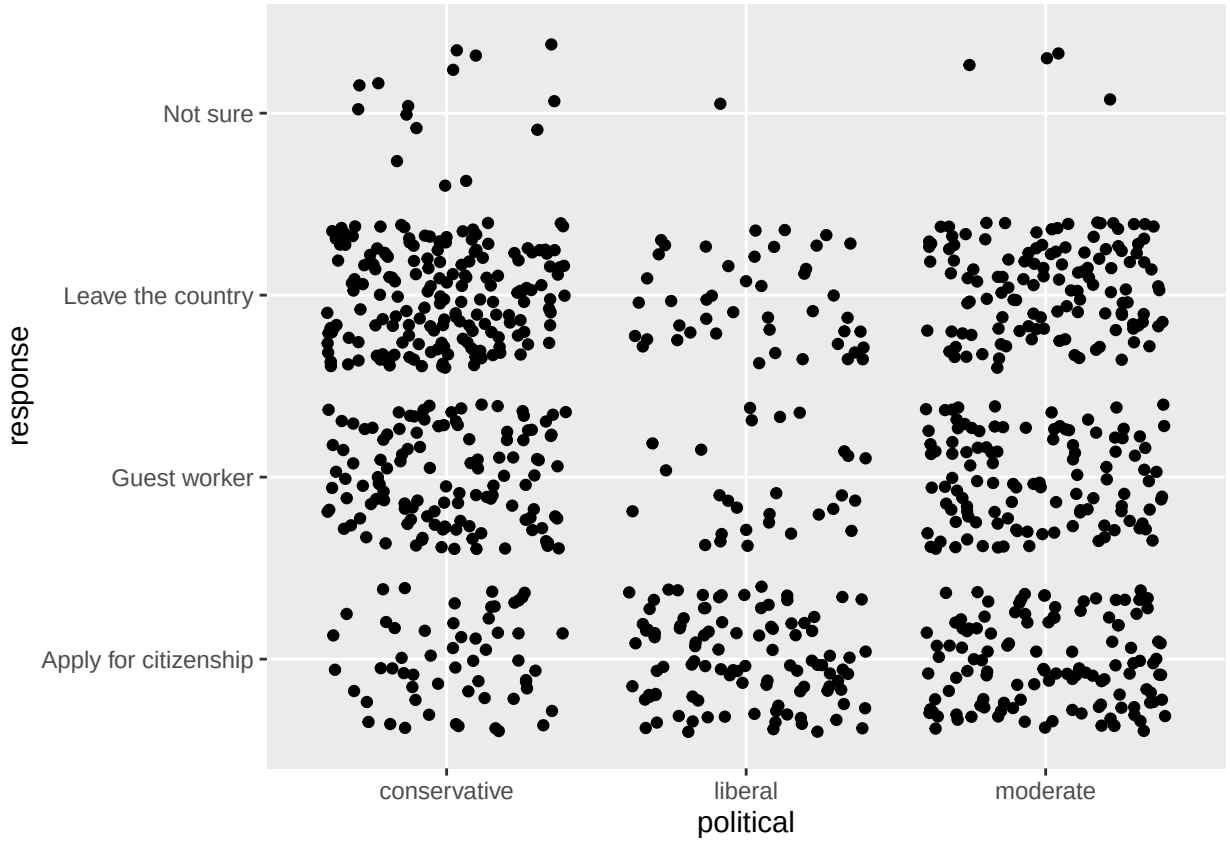
```
library(ggplot2)
library(tidyr)
immigration <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 3/immigration.csv")

#geom-smooth shows no association between the data
immigration %>%
  ggplot() +
  geom_smooth(aes(x=political, y=response))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

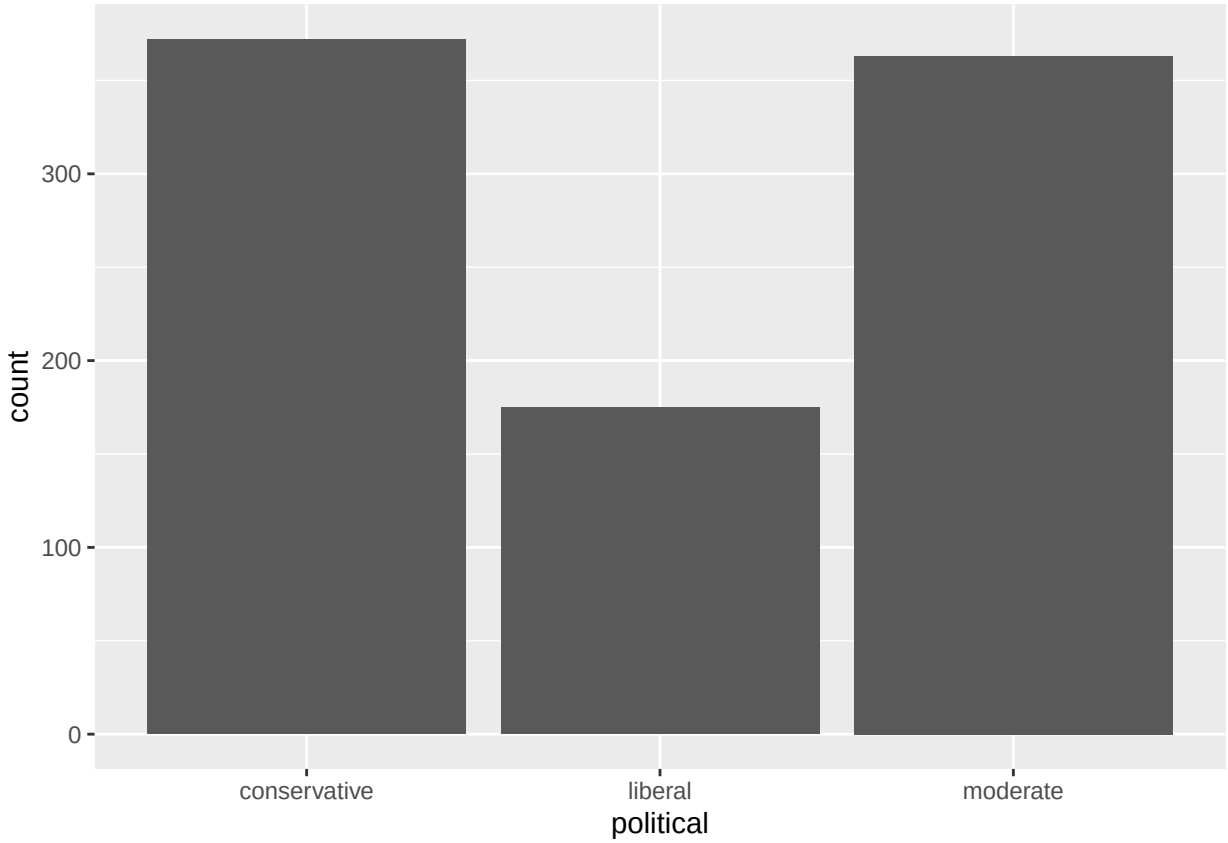



```
#geom-jitter confirms this finding, cannot spot trends/association between the
#data
immigration %>%
```

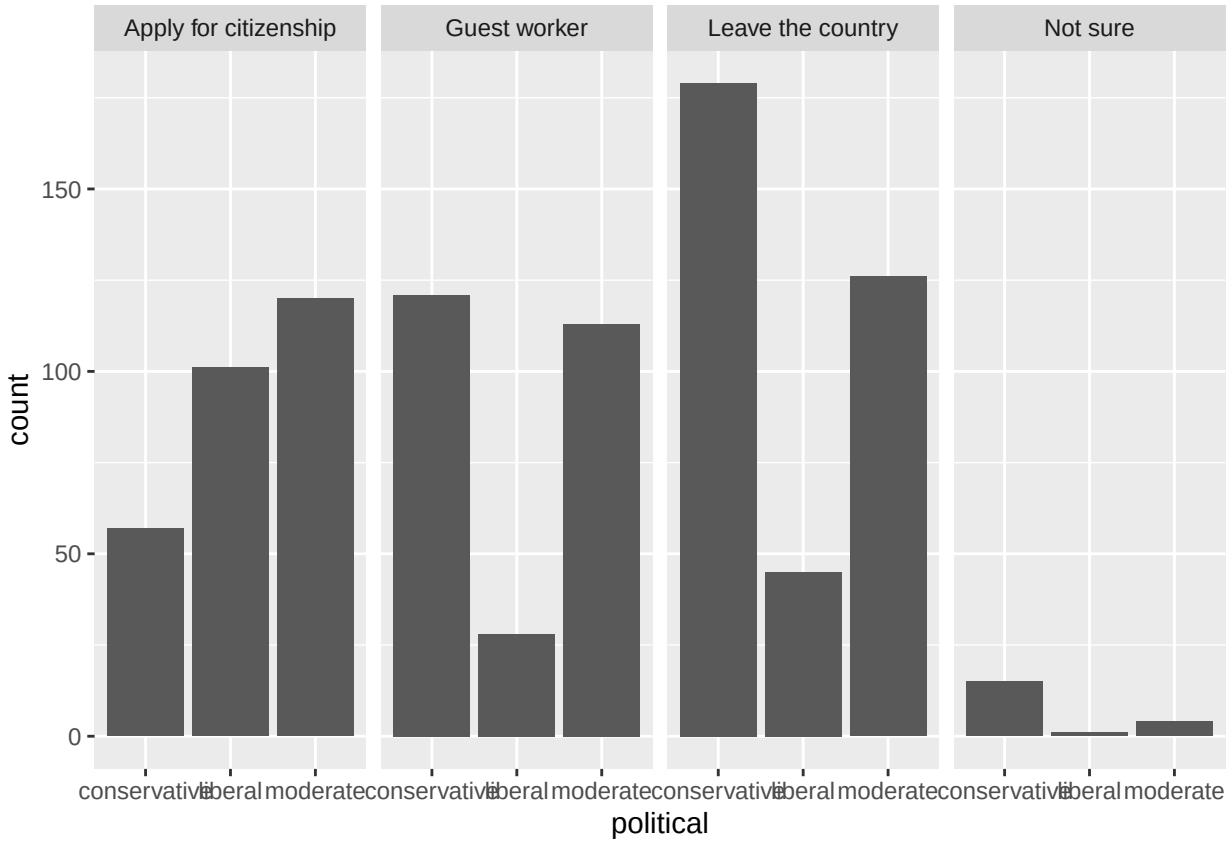
```
  ggplot() +
  geom_jitter(aes(x=political, y=response))
```




```
#bar chart shows distribution of conservative, moderate, and liberal voters
#from sample
immigration %>%
  ggplot() +
  geom_bar(aes(x=political))
```




```
#A Different view of the data. It would appear
#that conservatives feel strongest about leaving the country, but the highest
#number participants from the sample are also conservative.
immigration %>%
  ggplot() +
  geom_bar(aes(x=political)) +
  facet_grid(cols=vars(response))
```



Based on the sample collected, political ideology and views on immigration appear to be independent. Based on data discovery, 41% of the sample Tampa, FL voters identify as conservative, 40% identify as moderate, and 19% identify as liberal. Upon importing the dataset (immigration) and conducting data visualization using the geom_smooth() function, there is no pattern to be mapped. This can also be seen when performing the geom_jitter() function, as all dots appear in a non-patterned formation. The percentage of conservative, moderate, and liberal voters appears again using the geom_bar() function, and provides a visual representation to illustrate the significantly higher number of conservative voters and moderate voters as opposed to liberal, as shown in the bi-modal distribution of the data. Finally, the last faceted bar plot really drives the point of the difference in how the sampled voters identify within the political party. Even in instances where liberals had the highest response (101 in option i), the moderate response to this option showed a higher number (120), even though the highest number of responses from moderates was in support of option iii (126). So, although the variables appear to be independent, I would not feel comfortable scaling this model outside of the Tampa, FL region. Although this may be indicative of Tampa, FL, to apply this model to the larger population, better randomized sampling techniques should be used such as stratified sampling across multiple regions.