‘Lab 1: Question 3’

w203: Statistics for Data Science

Contents

1 Are survey respondents who have had someone in their home infected by COVID-19 more likely to disapprove of the way their governor is handling the pandemic?

1.1 Description of Data

It is often the governors who make a bigger impact on people everyday lives. The coronavirus pandemic has shined a spotlight on the role of these leaders, who have largely been responsible for leading their states' responses to the public health crisis. With little guidance from the federal government, state governors have decided the rules on the shutdown of nonessential businesses and the timelines for reopening. They've been responsible for getting personal protective gear into the hands of medical workers and first responders, as well as expanding coronavirus testing in their states. And they're the ones people look to, to find out exactly how their state is doing and what they can expect during this time of uncertainty.

To espondents who have had someone in their home infected by COVID-19 more likely to disapprove of the way their governor is handling the pandemic? we will address this question using data from the 2018 American National Election Studies (ANES). This is an observational dataset, based on a sample of respondents drawn from the YouGov platform. The first thing to note in this data is that there are differences in the rank ordering of approval and disapproval, ie. strong approval/disapproval, or normal approval/sidapproval. The next figure plots the relative prevalence of ANES respondent's feelings of approval and disapproval. While there are about the same proportion of respondents who report being the lowest category in both fear and anger,

there are more who report being somewhat afraid than report being somewhat angry. This is offset by fewer individuals reporting high levels of fear.

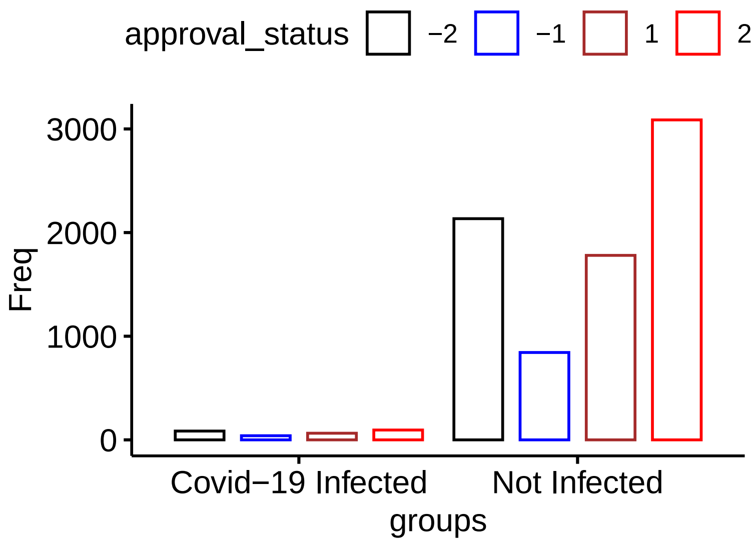
1.2 Importance and Context

COVID-19 has affected every state, but no two governors have handled it in exactly the same way. Some, like Gov. Andrew Cuomo in New York and Gov. Gavin Newsom in California, have instituted long-term stay-at-home orders and mandatory shutdowns of nonessential businesses. Others, like Gov. Kristi Noem of South Dakota and Gov. Asa Hutchinson of Arkansas, have resisted placing sweeping restrictions on travel and business operations. While governors' responses to the public health emergency have varied, one thing is true across all 50 states: Constituents are keeping a careful eye on how authorities are addressing COVID-19, and judging them accordingly.

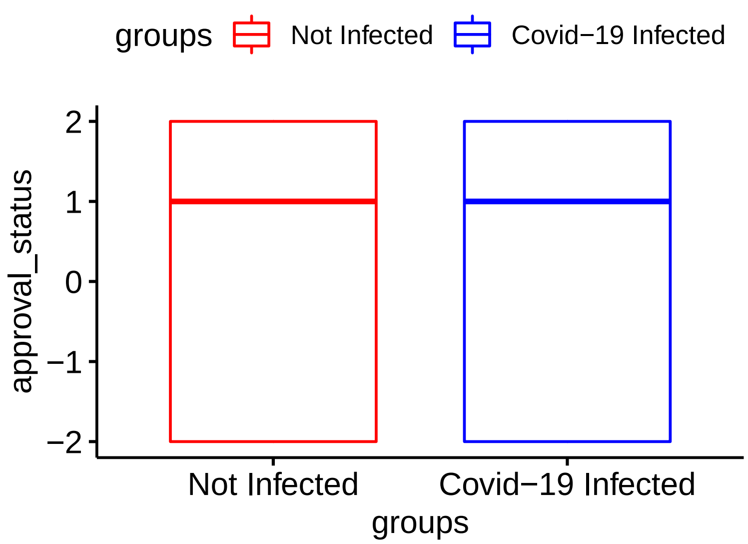
According to a Fox News report on February 2 local time, the approval rating of California Governor Gavin Newsom has plummeted, and this situation may exacerbate the possibility of his removal. A poll conducted by the Berkeley Institute of Government on more than 10,000 registered voters in California found that Newsom's work performance was only approved by 46%, a sharp drop from his 64% approval rate in September last year. In addition, Newsom has been widely criticized for the previous strict stay-at-home order. Up to now, 1.3 million people have signed a petition calling for his removal as governor. The report pointed out that 36% of California voters currently support the removal of Newsom, and most of them support the Republican Party. They are dissatisfied with Newsom's performance in handling the epidemic. In addition, less than one-third of the interviewees said that the governor had "excelled" in responding to the epidemic. In recent weeks, the number of new confirmed cases of new coronary pneumonia in a single day in California has declined. According to statistics from the state, only 12,000 new confirmed cases were reported on the 2nd, while in December last year, the highest number of new confirmed cases in a single day exceeded 50,000.

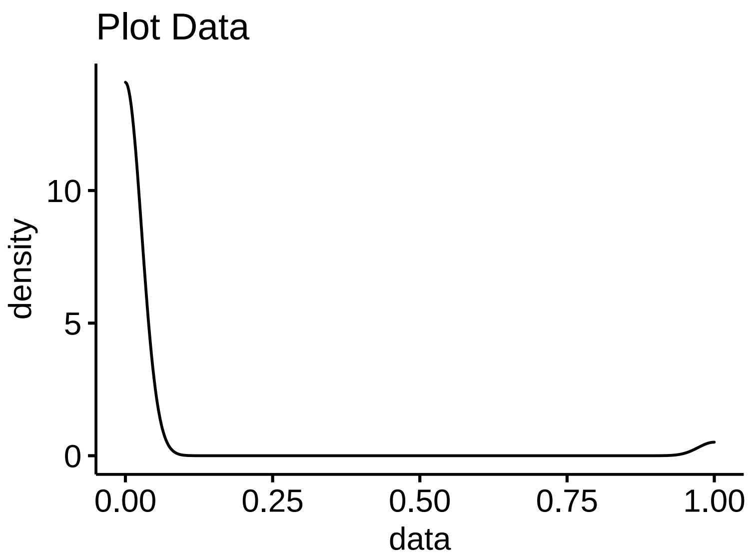
```
##                     approval_status
## groups                 -2   -1    1    2
##   Covid-19 Infected    85   40   64   95
##   Not Infected       2134  843 1780 3087
```

```
##              groups approval_status Freq
## 1 Covid-19 Infected              -2   85
## 2      Not Infected              -2 2134
## 3 Covid-19 Infected              -1   40
## 4      Not Infected              -1  843
## 5 Covid-19 Infected               1   64
## 6      Not Infected               1 1780
```




```
## # A tibble: 2 x 4
##   groups            count  mean    sd
## * <chr>             <int> <dbl> <dbl>
## 1 Covid-19 Infected   284 0.155  1.70
## 2 Not Infected       7844 0.362  1.69
```





I use the R function shapiro.test() to perform the Shapiro-Wilk test of normality for one variable (univariate):

Check if it is normal distribution

```
##
##  Shapiro-Wilk normality test
##
## data:  approval_status[groups == "Covid-19 Infected"]
## W = 0.78207, p-value < 2.2e-16
```

1.3 Most appropriate test

I then test whether this subgroup reports being more angry or more afraid. Because these variables are measured on ordinal scales, a non-parametric test is appropriate. Furthermore, the data is paired, since each individual has a pair of measurements. Some researchers may use a signed-rank test in these circumstances, but that test, while technically non-parametric, requires data on a metric scale. Instead, I will use a sign test, implemented in R using `binom.test`.

The sign test requires the following assumptions to be true:

- i.i.d. data. The ANES 2018 pilot uses a panel of individuals that use YouGov. This is an online system that rewards individuals for filling out surveys. There is a possibility that this introduces dependencies. For example, participants may tell friends or family members about YouGov, resulting in a cluster of individuals that give similar responses. Nevertheless, YouGov claims to have millions of users, which suggests that links between individuals should be rare.
- Ordinal scale. The data levels showing an increase in intensity from "not at all" to "extremely." There is some question, however, as to whether participants use the same scale when they think about anger and fear. Another way of noting this is to ask whether voters hold the same anchoring points on these scales.

An underlying assumption for appropriate use of the tests described was that the continuous outcome was approximately normally distributed or that the samples were sufficiently large (usually n1> 30 and n2> 30) to justify their use based on the Central Limit Theorem. I use the R function shapiro.test() to perform the Shapiro-Wilk test of normality for one variable (univariate): Check if it is normal distribution

- Hypothesis 1: Are the two samples independent?

Yes, because the samples from Covid-19 infected and not infected have nothing to do.

- Hypothesis 2: Does the data of each of the two groups obey a normal distribution?

I use the function of shapiro.test() to calculate the Shapiro-Wilk test for each set of samples. Check if it is normal distribution

```
##
##  Shapiro-Wilk normality test
##
## data:  approval_status[groups == "Covid-19 Infected"]
## W = 0.78207, p-value < 2.2e-16
```

Shapiro-Wilk normality test for p-value < 2.2e-16

In the output result, the two p-values are less than the significance level of 0.05, indicating that the distribution of the two sets of data is significantly different from the normal distribution. The hypothesis test that the data distribution does not conform to the normal distribution is established.

- Hypothesis 3: Do these two populations conform to the homogeneity of variance?

We will use the F test to test the homogeneity of variance. You can use the var.test() function to perform the following operations:

```r
res.ftest <- var.test(approval_status ~ groups, data = df)
res.ftest
```

```
##
##  F test to compare two variances
##
## data:  approval_status by groups
## F = 1.0076, num df = 283, denom df = 7843, p-value = 0.9084
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8574517 1.1998611
## sample estimates:
## ratio of variances
##           1.007593
```

The F test is p = 0.9084. It is greater than the significance level alpha = 0.05. Therefore, there is no significant difference between the variances of the two sets of data. Therefore, we believe that the variance of the two groups of men and women is equal (homogeneity of variance).

Due to the combination of the above three assumptions, and because the data does not conform to a normal distribution, we cannot use the student-t test. A nonparametric test is appropriate, particularly two independent samples Wilcoxon rank sum test.

I then test espondents who have had someone in their home infected by COVID-19 more likely to disapprove of the way their governor is handling the pandemic? The modules on hypothesis testing presented techniques for testing the equality of means in two independent samples, i.e. a group has family member infected, and another group without infected.

A nonparametric test to compare outcomes between two independent groups is the Wilcoxon Rank Sum Test. The parametric test compares the means between independent groups. In contrast, the null and two-sided research hypotheses for the nonparametric test are stated as follows:

H0: The two populations are equal versus H1: The two populations are not equal.

This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality. A one-sided research hypothesis is used if interest lies in detecting a positive or negative shift in one population as compared to the other. The procedure for the test involves pooling the observations from the two samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to n1+n2, respectively.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  approval_status by groups
## W = 1038950, p-value = 0.04289
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  approval_status by groups
## W = 1038950, p-value = 0.02144
## alternative hypothesis: true location shift is less than 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  approval_status by groups
## W = 1038950, p-value = 0.9786
## alternative hypothesis: true location shift is greater than 0
```

```
res <- cor.test(df0$groups, df0$approval_status,
                    method = "pearson")
res
```

```
##
##  Pearson's product-moment correlation
##
## data:  df0$groups and df0$approval_status
## t = -2.0291, df = 8126, p-value = 0.04248
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.044222374 -0.000763539
## sample estimates:
##         cor
## -0.02250359
```

Small size of association.

1.4 Test Limitations

We have conducted this test based on the data available in the ANES. However, even after an appropriate analysis of the data on hand, it would not be possible know why people chose to **not** vote in the 2016 election. Perhaps someone chose not to vote in the 2016 election because they were angry. If this were to be the case, then our study design cannot find an effect of anger-inducing language in the 2018 data. And so, while the data that test on does fails to reject the null hypothesis that there is a difference in anger and fear as a motivation for increased voting in 2018, it is possible that the very nature of this observational data has precluded our ability to produce a satisfactory answer to the question.