Correlation Study Example

5/17/2023

Below I have three statistical methods that can be used to evaluate the relationship between two categorical variables: the chi-square test of independence, Fisher's exact test, and Cramer's V. First we will need to read in the data set `df` and create a contingency table, or a count table `tab`. Using the `table` function, we can create a Punnett-square style count table for two different categorical variables.

It's also important to make sure there are no NA or missing values in your contingency table. For this case, I removed all instances where the value was empty.

```
# Read in data set
df <- read.csv("CA WFC_ Legislative Support - Assemblymembers.csv", header = TRUE, skip = 1)

# Create contingency table
tab <- table(df$SB.114..enacted..COVID.19.SPSL..Chamber.opposed.,
             df$AB.123..vetoed..increase.PFL...SDI.to.90...no.funding.mechanism...CA.WFC.top.priority.C
tab
```

```
##
##                  AYE Not in Office NVR
##   AYE             22             0   0
##   NO               1             0   2
##   Not in Office    0            12   0
```

Chi-Square Test of Independence

The first option to evaluate the relationship between two categorical variables is a chi-square test of independence. This is a statistical hypothesis test that uses a contingency table to determine if two categorical variables are independent or dependent. Using the `chisq` function, we can run the chi-square test of independence on our contingency table. If the calculated p-value is below 0.05, we can conclude that the two variables of interest have are associated with each other. On the other hand, if the p-value is greater than 0.05, we can conclude that the two variables are independent.

```
# Run chi-square test of independence
chisq <- chisq.test(tab)
```

```
## Warning in chisq.test(tab): Chi-squared approximation may be incorrect
```

```
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 60.594, df = 4, p-value = 2.176e-12
```

Fisher Exact Test

The second option to evaluate the relationship between two categorical variables is Fisher's exact test. This is another statistical hypothesis test that uses a contingency table to determine if two categorical variables are independent or dependent. However, this test is used when the assumption for the chi-square test of independence that all the expected counts are sufficiently large (5 or greater) is not met. In the example above, some of the expected counts are well below 5, so Fisher's exact test is the more appropriate choice for a statistical test.

Using the `fisher.test` function, we can run Fisher's exact test on our contingency table. If the calculated p-value is below 0.05, we can conclude that the two variables of interest have are associated with each other. On the other hand, if the p-value is greater than 0.05, we can conclude that the two variables are independent.

```
# Run Fisher Exact Test
fisher <- fisher.test(tab)
fisher
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 5.398e-12
## alternative hypothesis: two.sided
```

Cramer's V Correlation

The third option to evaluate the relationship between two categorical variables is Cramer's V. This operates similar to a correlation coefficient, but is used to estimate the statistical correlation between two categorical variables. Also similar to a correlation coefficient, Cramer's V is a value between 0 and 1 and represents the strength of the correlation, with values 0 to 0.4 representing a weak correlation, 0.4 to 0.7 representing a moderate correlation, and 0.7 to 1 representing a strong correlation.

Using the `cramerV` function from the `rcompanion` package, we can calculate Cramer's V of the two variables in our contingency table. We evaluate the result using the criterion mentioned above.

```
# install.packages("rcompanion")
library(rcompanion)
```

```
## Warning: package 'rcompanion' was built under R version 4.1.2
```

```
cramerV(tab)
```

```
## Cramer V
##   0.9049
```

Stacked Bar Chart

Here's an example of how to visually display the relationship between two categorical variables:

```
# install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
ggplot(df[1:30,], aes(x = Party.Affiliation,
                      fill = SB.114..enacted..COVID.19.SPSL..Chamber.opposed.)) + geom_bar()
```