Categorical Data Analysis

Lucy Harlow

09/11/2020

Chi-square Goodness of Fit

I want to know if certain dog breeds are more likely than others to come into my dog grooming salon. I assume that I'm equally likely to get pomeranians, pugs, poodles, bulldogs, and labradoodles.

$H_0 : p = (.2, .2, .2, .2, .2)$

$H_a : p \neq (.2, .2, .2, .2, .2)$

I observe 22 pomeranians, 28 pugs, 30 poodles, 10 bulldogs, and 10 labradoodles. I perform a chi-square goodness of fit test as follows:

```r
x <- c(22, 28, 30, 10, 10)
chisq.test(x, p = c(.2, .2, .2, .2, .2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 18.4, df = 4, p-value = 0.001031
```

I determine that a χ^2 value of 18.4 is significantly high, and that we can therefore reject the null hypothesis. The probabilities of welcoming pomeranians, pugs, poodles, bulldogs, and labradoodles into my dog grooming salon are not equal.

Chi-square test of independence/association

An example of a research question that could use a chi-square test of independence:

- After watching a group of randomly selected German Shepherds and Dutch Shepherds playing with squeaky balls and squeaky bones, I want to know if there's an association between breed and particular toy preference.

A psychologist is interested in testing whether there is a relationship between statistical anxiety level (low versus high) and first degree (humanities versus sciences). She conducts a random sample survey and the results of the study are shown below:

- High stats anxiety and humanities first degree: 45 people
- High stats anxiety and sciences first degree: 30 people
- Low stats anxiety and humanities first degree: 12 people
- Low stats anxiety and sciences first degree: 76 people

Conduct a test of independence. Test at a 5% level of significance.

- I'll notate high anxiety as 1 and low as 0, and a humanities degree as h and a science degree as s. My null hypothesis is that there is no relationship between stats anxiety and first degree.

$H_0: P_{h1} = P_{h0}, P_{s1} = P_{s0}$

$H_0: P_{h1} \neq P_{h0}, P_{s1} \neq P_{s0}$

```
independence <- matrix(
  c(45, 12, 30, 76), # the data elements
  nrow=2,              # number of rows
  ncol=2,              # number of columns
  byrow = TRUE)        # fill matrix by rows

rownames(independence) = c("Humanities", "Sciences")
colnames(independence) = c("High", "Low")
independence
```

```
##            High Low
## Humanities   45  12
## Sciences     30  76
```

```
chisq.test(independence, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  independence
## X-squared = 38.274, df = 1, p-value = 6.147e-10
```

McNemar Test

In a health psychology experiment, 200 students receive an alcohol prevention intervention. Their past-week drinking behaviour (binge drinker status, either Yes or No) is measured at baseline (pre-intervention) and again at four weeks follow-up (post-intervention).

$H_0:$ the intervention has no effect.

$H_a:$ the intervention has an effect.

62 of the 67 binge drinkers before the intervention were also binge drinkers after the intervention. Interestingly, 21 of the non-binge drinkers before the intervention were binge drinkers after the intervention.

```
test <- matrix(
  c(62, 21, 5, 112), # the data elements
  nrow=2,              # number of rows
  ncol=2,              # number of columns
  byrow = TRUE)        # fill matrix by rows

rownames(test) = c("Before: Yes", "Before: No")
colnames(test) = c("After: Yes", "After: No")
test
```

```
##             After: Yes After: No
## Before: Yes         62        21
## Before: No           5       112
```

```
mcnemar.test(test, correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  test
```

```
## McNemar's chi-squared = 9.8462, df = 1, p-value = 0.001702
```