Chi-Square Part 2: Independence

STAT-S520

Arturo Valdivia

(A particular case of Goodness-of-fit)

04-18-23

Independence: Setting

2 partition

- Let S be the sample space of our experiment and
  - $A_1, \ldots, A_r$ partition S into r cells
  - $B_1, \ldots, B_c$ also partition S into c cells
- Think of As and Bs as two variables with different categories (partition)

Independence: Setting

- We care about the relationship between As and Bs
- We define a third partition by

rows

$$E_{ij} = A_i \cap B_j$$

- Partitions $A_1, \ldots, A_r$ and $B_1, \ldots, B_c$ are mutually independent if and only if

columns

$$P(E_{ij}) = P(A_i) \cdot P(B_j)$$

$$p_{ij} = p_{i\cdot}\, p_{\cdot j}$$

for each ij pair.
- We use the chi-squared methods developed above to check if independence holds

Example 2

Two partitions of criminals, one by type of crime (arson, rape, violence, stealing, coining, fraud) and the other by alcohol consumption (drinker, abstainer). Here is the sample (counts) observed:

Observed counts

```
##              drink abstain
## arson           50      43
## rape            88      62
## violence       155     110
## stealing       379     300
## coining         18      14
## fraud           63     144
```

93 Arson

150 rape

753 673 1426

Expected Counts

drink: $\boxed{\dfrac{753 \times 93}{1426}}$

abstain: $\dfrac{673 \times 93}{1425}$

Example 2 (continued)

The expected counts using the outer product (%o%)

observed counts (contingency table)

```
exp = rowSums(obs)%o%colSums(obs)/sum(obs)
exp
```

outer product

```
##                drink  abstain
## arson       49.10870 43.89130
## rape        79.20757 70.79243
## violence   139.93338 125.06662
## stealing   358.54628 320.45372
## coining     16.89762 15.10238
## fraud      109.30645 97.69355
```

Expected counts under H_0

H_0: The variables are independent

H_1: " " are related

Test statistics

As before, the test statistics are

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

or

$$G^2 = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} \log \left(\frac{o_{ij}}{e_{ij}} \right)$$

same statistics as before

Degrees of Freedom

- Unrestricted set: $rc - 1$
- Restricted set: $(r - 1) + (c - 1)$
- Degrees of freedom:
  $\underbrace{(rc - 1) - [(r - 1) + (c - 1)]} = \underbrace{(r - 1)(c - 1)}$

Example 2 (continued)

```
X2 = sum((obs - exp)^2/exp)
X2
```

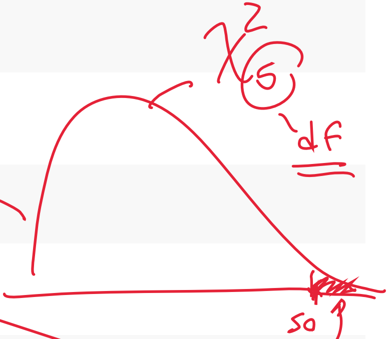
```
## [1] 49.73061
```

```
G2 = sum(2*obs*log(obs/exp))
G2
```

```
## [1] 50.51729
```

```
df = (6 - 1)*(2 - 1)
1 - pchisq(X2, df)
```

```
## [1] 1.573317e-09
```




```
1 - pchisq(G2, df)
```

```
## [1] 1.085962e-09
```

Example 2 Simulation-Based

Contingency table.

Name of first variable (you choose)

Categories of 2nd variables.

Name of 2nd variable

```
library(tidyverse)
df.obs = as.data.frame(obs)
data2 <- df.obs |>
  rownames_to_column("crime") |>
  pivot_longer(cols=c('drink', 'abstain'),
               names_to='alcohol',
               values_to='count') |>
  rowwise() |>
  mutate(count = list(1:count)) |>
  unnest(count) |>
  select(-count)
```

DO NOT CHANGE

Use it only if you need to convert a contingency table into a long-form data frame (one row per individual)

Example 2 Simulation-Based (continued)

```
null_dist <- data2 |>
  specify(alcohol ~ crime) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "Chisq")

null_dist |>
  get_p_value(obs_stat = G2, direction = "greater")
```

$\chi 2$

```
## Warning: Please be cautious in reporting a p-value of 0
## approximation based on the number of `reps` chosen in th
## `?get_p_value()` for more information.

## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Example:

$$
\begin{array}{c|cc}
 & \multicolumn{2}{c}{\text{Alcohol}(A)} \\
\text{Crime}(C) & D & A \\
\hline
a & p_{11} & p_{12} \\
r & p_{21} & p_{22} \\
v & \vdots & \\
s & \vdots & \\
c & & \\
f & p_{61} & p_{62}
\end{array}
$$

$$H_0: \quad p_{11} = p_{1C} \cdot p_{1A}$$
$$p_{21} = p_{2C} \cdot p_{1A}$$
$$\vdots$$
$$\boxed{p_{ij} = p_{iC} \cdot p_{jA}}$$

$i = 1, \dots, 6$

$j = 1, 2$