Exploring Relationships in Categorical Health Data

Contingency Tables for Categorical Variables

For two categorical variables X with I categories and Y with J categories, classification of subjects on both variables have IJ possible combinations. We are interested in their `joint distribution` if both are response variables. Typically one of them, say, Y is a response variable and X is an explanatory variable, we are interested in the `conditional distribution` of Y given the categories of X.

A rectangular table, having I rows for categories of X and J columns for categories of Y, with frequencies in the (IJ) cells is called a contingency table (Karl Pearson, 1904) or a cross-classification table. A contingency table with I rows and J columns is called an $I \times J$ table.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("~/Box/MyDocs/Teaching/Spring/2021/DSCI 610/LectureMaterials/Week 9/Lecture")
df_analysis <-readRDS("analysis.rds")

df_CD<-select(df_analysis, Gender, Race1,  HomeOwn, Diabetes, SmokeNow, HealthGen,Depressed, Marijuana,

df_CD1 <- filter(df_CD, Diabetes !="NA")
df_CD2 <- filter(df_CD, HealthGen !="NA")
```

Example 1: Contingency table for Diabetes and Gender: (2×2) table

```
# Observed frequencies
cross1<- table(df_CD1$Gender,df_CD1$Diabetes)
# Add row and column totals
addmargins(cross1)
```

```
##
##            No  Yes  Sum
##   female 4592  357 4949
##   male   4506  403 4909
##   Sum    9098  760 9858
```

```r
# display column percentages
round(prop.table(cross1,2),digits = 3)
```

```
##
##             No   Yes
##   female 0.505 0.470
##   male   0.495 0.530
```

Comparing two proportions

Often times studies are designed to compare groups on a binary response variable Y. For example, in our NHANES data we may be interested to compare male and female participants' diabetic condition (yes / no). With two groups (male/female), we have a 2×2 contingency table. Proportion of diabetes across male and female participants can be compared with three measures : i) difference of proportions, ii) relative risk, and iii) odds ratio.

Let π_1 denotes the probability that a male participant has diabetes while π_2 denotes the probability that a female participant has diabetes.

Then the `difference of proportions` of these two groups is defined as $\pi_1 - \pi_2$.

The `relative risk` for comparing proportion successes in two groups is defined as the ratio probabilities:

$$\text{relative risk} = \frac{\pi_1}{\pi_2}.$$

A relative risk of 1 indicates that the probability of diabetes does not depend on gender.

The `odds ratio` for comparing the odds of successes in two groups is defined as

$$\text{odds ratio, OR} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

The $OR = 1$ indicates that the odds of diabetes does not depend on gender. When $1 < OR < \infty$, male participants are more likely to have diabetes than female participants. For example, $OR = 2$ indicates that the odds for males to have diabetes is twice the odds for females to have diabetes. When $0 < OR < 1$, male participants are less likely to have diabetes than female participants.

Computing relative risk (risk ratio) using R

```r
library(epitools)
riskratio.wald(cross1)
```

```
## $data
##
##            No Yes Total
##   female 4592 357  4949
```

```
##   male   4506 403  4909
##   Total  9098 760  9858
##
## $measure
##         risk ratio with 95% C.I.
##          estimate     lower   upper
##   female  1.00000        NA      NA
##   male    1.13805 0.9924731 1.30498
##
## $p.value
##         two-sided
##          midp.exact fisher.exact chi.square
##   female         NA           NA         NA
##   male   0.06397691   0.06438682  0.0638329
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

The risk (probability) of having diabetes in male is 1.14 times higher than that in females. However, the relative risk is not statistically significant as the null value 1 is included in the confidence interval of the relative risk and the p-value for the test that the relative risk is 1 is borderline (0.064).

Computing odds ratio using R

```
oddsratio.wald(cross1)
```

```
## $data
##
##           No Yes Total
##   female 4592 357  4949
##   male   4506 403  4909
##   Total  9098 760  9858
##
## $measure
##         odds ratio with 95% C.I.
##          estimate     lower    upper
##   female 1.000000        NA       NA
##   male   1.150396 0.9918778 1.334249
##
## $p.value
##         two-sided
##          midp.exact fisher.exact chi.square
##   female         NA           NA         NA
##   male   0.06397691   0.06438682  0.0638329
##
## $correction
## [1] FALSE
##
## attr(,"method")
```

```
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

The odds of having diabetes in male is 1.15 times higher than that in females. However, the odds ratio is not statistically significant as the null value 1 is included in the confidence interval of the odds ratio and the p-value for the test that the odds ratio is 1 is borderline (0.064).

Note that the relative risk is a valid measure of association for prospective cohort studies but not for retrospective case-control studies. Odds ratio is a valid measure of association for either type of studies.

Testing Independence in two-way contingency tables

Testing independence in two-way contingency tables assumes that the total sample size is fixed with joint probabilities $\{\pi_{ij}\}$ in an $I \times J$ contingency table. The null hypothesis of statistical independence of the row and column variable is :

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j},$$

where, $\pi_{i.}$ is the marginal probability of being in the ith row and $\pi_{.j}$ is the marginal probability of being in the jth column.

Under the assumption that H_0 is true, the expected frequency in the (i, j)th cell, $E(n_{ij}) = \mu_{ij}$. Typically $\pi_{i.}$ and $\pi_{.j}$ are unknown and their MLEs are the sample marginal proportions as follows:

$$\hat{\pi}_{i.} = \frac{n_{i.}}{n}, \quad \text{and} \quad \hat{\pi}_{.j} = \frac{n_{.j}}{n},$$

where $n_{i.}$ and $n_{.j}$ are ith row sum and jth column sum respectively. The estimated expected frequencies under H_0 becomes

$$\hat{\mu}_{ij} = n\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{n_{i.}n_{.j}}{n}.$$

The test statistic for the test of independence has the following expression:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

The statistic χ^2 is called the Pearson statistic (Pearson, 1904, 1922) and has a chi-squared distribution with $(I - 1)(J - 1)$ degrees of freedom.

Example 2: Independence Test for Gender and HealthGen : (2×5) table

```r
# Observed frequencies
cross2<- table(df_CD2$Gender,df_CD2$HealthGen)
# Add row and column totals
addmargins(cross2)
```

```
##
##           Excellent Vgood Good Fair Poor  Sum
##   female        402  1292 1439  491  101 3725
##   male          476  1216 1517  519   86 3814
##   Sum           878  2508 2956 1010  187 7539
```

```r
# display column percentages
round(prop.table(cross2,2),digits = 3)
```

```
##
##           Excellent Vgood  Good  Fair  Poor
##   female      0.458 0.515 0.487 0.486 0.540
##   male        0.542 0.485 0.513 0.514 0.460
```

```r
chisq.test(cross2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cross2
## X-squared = 11.529, df = 4, p-value = 0.02122
```

The observed value of χ^2 is 11.617 which gives a p-value of 0.02044 from a chi-squared distribution of 4 degrees of freedom. Thus, there is evidence against the null hypothesis of independence of general health condition and gender. From the proportion table we see that in general males have better health conditions than females.

Example 3: Independence Test for Race and HealthGen: (5×5) table

```r
cross3<- table(df_CD2$Race1,df_CD2$HealthGen)
addmargins(cross3)
```

```
##
##             Excellent Vgood Good Fair Poor  Sum
##   Black            79   188  370  175   27  839
##   Hispanic         44    98  192   82   14  430
##   Mexican          48   125  286  176   33  668
##   White           643  1918 1860  515   97 5033
##   Other            64   179  248   62   16  569
##   Sum             878  2508 2956 1010  187 7539
```

```r
round(prop.table(cross3,2),digits = 3)
```

```
##
##             Excellent Vgood  Good  Fair  Poor
##   Black         0.090 0.075 0.125 0.173 0.144
##   Hispanic      0.050 0.039 0.065 0.081 0.075
##   Mexican       0.055 0.050 0.097 0.174 0.176
##   White         0.732 0.765 0.629 0.510 0.519
##   Other         0.073 0.071 0.084 0.061 0.086
```

```r
chisq.test(cross3)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  cross3
## X-squared = 358.42, df = 16, p-value < 2.2e-16
```

The observed value of χ^2 is 357.06 which gives essentially a 0 p-value from a chi-squared distribution of 16 degrees of freedom. Thus, there is strong evidence against the null hypothesis of independence of general health condition and race.

References

1. Chapter 2: Describing Contingency Tables. Alan Agresti (2013). Categorical Data Analysis, John Wiley and Sons.