

# Chi Squared

## Instructions

The data comes from the faculty salary example.

There are three variables: sex (sex of professor) 1 = male 2 = female rank (rank of professor) 1 = full professor 2 = associate professor 3 = assistant professor 4 = instructor level (type of program that professor teaches in) 1 = doctoral program 2 = masters program

```
library(pacman) #Package used to load all packages using p_load(); will install missing packages
```

```
## Warning: package 'pacman' was built under R version 3.5.3
```

```
p_load(vcd, MASS, jmv, gmodels)
# jmv and gmodels used for chi-squared
# vcd, MASS used for loglinear
```

Load your data

```
dat <- read.csv("https://www.dropbox.com/s/w2bcd0c2n7qgwzz/Salary-1.csv?dl=1")
head(dat)
```

```
##   sex rank level
## 1   1    1     1
## 2   1    1     1
## 3   1    1     1
## 4   1    1     1
## 5   2    1     1
## 6   2    1     1
```

While this part isn't necessary it will make this entire demo easier to read. You are relabeling the levels of each variable.

```
dat$sex <- factor(dat$sex, levels = c(1,2), labels = c("Male", "Female"))
dat$rank <- factor(dat$rank, levels = c(1,2,3,4), labels = c("Full", "Associate", "Assistant", "Instructor"))
dat$level <- factor(dat$level, levels = c(1,2), labels = c("Doctorate", "Masters"))
head(dat)
```

```
##      sex rank    level
## 1  Male Full Doctorate
## 2  Male Full Doctorate
## 3  Male Full Doctorate
## 4  Male Full Doctorate
## 5 Female Full Doctorate
## 6 Female Full Doctorate
```

## Goodness of Fit

Observed Frequencies for each variable.

```
sex <- table(dat$sex)
sex
```

```
##
##  Male Female
```

```
##      2803      1839
rank <- table(dat$rank)
rank

##
##      Full  Associate  Assistant  Instructor
##      2032      1311      1215        84

level <- table(dat$level)
level

##
## Doctorate  Masters
##      3848      794

## uses descriptives from jmv library - it is mas cute

desc <- descriptives(data = dat,
                     vars = c('sex', 'rank', 'level'),
                     freq = TRUE)
desc
```

```
##
## DESCRIPTIVES
##
## Descriptives
## -----
##           sex      rank      level
## -----
##  N           4642      4642      4642
##  Missing         0         0         0
##  Mean
##  Median
##  Minimum
##  Maximum
## -----
##
##
## FREQUENCIES
##
## Frequencies of sex
## -----
##  Levels      Counts      % of Total      Cumulative %
## -----
##  Male          2803          60.4          60.4
##  Female         1839          39.6          100.0
## -----
##
##
## Frequencies of rank
## -----
##  Levels      Counts      % of Total      Cumulative %
## -----
##  Full          2032          43.8          43.8
##  Associate      1311          28.2          72.0
##  Assistant      1215          26.2          98.2
```

```
##      Instructor      84      1.8      100.0
## -----
##
##
## Frequencies of level
## -----
##      Levels      Counts      % of Total      Cumulative %
## -----
##      Doctorate    3848      82.9      82.9
##      Masters      794      17.1      100.0
## -----
```

Assumptions - 1. Adequate expected cell counts - 5 or more in 2 x 2 or 5 or more in 80% of cells for larger table - Otherwise, Fisher's test - 2. Independence of Observations - otherwise McNemar's test of dependent proportions

Chi Squared Test Goodness of fit (testing if all frequencies are equal)

```
# H0 = equal proportions in each category; Ha = unequal proportions in each category
# Chi-square = Sum[(Observed - Expected)^2/Expected]
# df = # of categories - 1
jmv::propTestN(data = dat,
               var = 'sex',
               expected = TRUE,
               ratio = c(1,1))
```

```
##
## PROPORTION TEST (N OUTCOMES)
##
## Proportions
## -----
##      Level      Count      Proportion
## -----
##      Male      Observed    2803      0.604
##              Expected    2321      0.500
##
##      Female    Observed    1839      0.396
##              Expected    2321      0.500
## -----
##
##
## <U+03C7>2 Goodness of Fit
## -----
##      <U+03C7>2      df      p
## -----
##      200      1      < .001
## -----
```

```
jmv::propTestN(data = dat,
               var = 'rank',
               expected = TRUE,
               ratio = c(1,1,1,1))
```

```
##
## PROPORTION TEST (N OUTCOMES)
##
## Proportions
```

```
## -----
##      Level                Count    Proportion
## -----
##      Full      Observed    2032      0.4377
##                Expected    1160      0.250
##
##      Associate  Observed    1311      0.2824
##                Expected    1160      0.250
##
##      Assistant  Observed    1215      0.2617
##                Expected    1160      0.250
##
##      Instructor Observed     84      0.0181
##                Expected    1160      0.250
## -----
##
##
## <U+03C7>2 Goodness of Fit
## -----
##      <U+03C7>2      df      p
## -----
##      1675      3      < .001
## -----
```

```
jmv::propTestN(data = dat,
               var = 'level',
               expected = TRUE,
               ratio = c(1,1))
```

```
##
## PROPORTION TEST (N OUTCOMES)
##
## Proportions
## -----
##      Level                Count    Proportion
## -----
##      Doctorate  Observed    3848      0.829
##                Expected    2321      0.500
##
##      Masters    Observed     794      0.171
##                Expected    2321      0.500
## -----
##
##
## <U+03C7>2 Goodness of Fit
## -----
##      <U+03C7>2      df      p
## -----
##      2009      1      < .001
## -----
```

However, what if we expected the proportions to be a little different. For example, based on an educated guess:

44% full Professors,  
28% Associate Professors,  
26% Assistant Professors,  
2% Instructors

How does it compare to the Chi-square where all levels were expected to have equal proportions?

```
# H0 = baseline model proportions; Ha = significantly different than baseline model proportions
jmv::propTestN(data = dat,
               var = 'rank',
               expected = TRUE,
               ratio = c(.44, .28, .26, .02))
```

```
##
## PROPORTION TEST (N OUTCOMES)
##
## Proportions
## -----
##      Level                Count    Proportion
## -----
##      Full      Observed    2032      0.4377
##                Expected    2042      0.4400
##
##      Associate  Observed    1311      0.2824
##                Expected    1300      0.2800
##
##      Assistant  Observed    1215      0.2617
##                Expected    1207      0.2600
##
##      Instructor Observed     84      0.0181
##                Expected     93      0.0200
## -----
##
##
## <U+03C7>2 Goodness of Fit
## -----
##      <U+03C7>2    df    p
## -----
##      1.05      3    0.790
## -----
```

## Chi-square Test of Independence

Ha: Is sex dependent upon rank? Is there a relationship between sex and rank?

We have a *new* effect size here (Cramer's V), what does it mean in the context of these results?

```

# Chi-square = Sum[(Observed - Expected)^2/Expected]
# Expected = [(# of row entries for cel)/(# total entries)] * (# of column entries for cel)
# Expected indicates expected values for each category if there is no relationship between two categories
# df = (# rows - 1) * (# columns - 1)
# Cramer's V - small = .1; medium = .3, large = .5; discrepancy between observed and expected scores
jmv::contTables(dat = dat,
                rows = 'sex',
                cols = 'rank',
                exp = TRUE,
                phiCra = TRUE)

```

```

##
## CONTINGENCY TABLES
##
## Contingency Tables
## -----
##      sex                Full    Associate    Assistant    Instructor    Total
## -----
##      Male      Observed    1474          711          583           35      2803
##                Expected    1227          792          734          50.7
##
##      Female      Observed    558          600          632           49      1839
##                Expected    805          519          481          33.3
##
##      Total      Observed    2032          1311          1215           84      4642
##                Expected    2032          1311          1215          84.0
## -----
##
##
## <U+03C7>^2 Tests
## -----
##      Value    df    p
## -----
##      <U+03C7>^2    237    3    < .001
##      N      4642
## -----
##
##
## Nominal
## -----
##      Value
## -----
##      Phi-coefficient    NaN
##      Cramer's V        0.226
## -----

```

*# report APA, magnitude of effect (Cramer's V), direction of effect example (more or less than expected)*

## Chi-square Test of Independence

Is level dependent upon sex? Is there a relationship between level and sex?

```

jmv::contTables(dat = dat,
                rows = 'sex',

```

```
cols = 'level',
exp = TRUE,
phiCra = TRUE)
```

```
##
## CONTINGENCY TABLES
##
## Contingency Tables
## -----
##      sex                Doctorate    Masters    Total
## -----
##      Male      Observed      2332      471      2803
##                Expected      2324      479
##
##      Female    Observed      1516      323      1839
##                Expected      1524      315
##
##      Total     Observed      3848      794      4642
##                Expected      3848      794
## -----
##
##
## <U+03C7>2 Tests
## -----
##      Value    df    p
## -----
##      <U+03C7>2    0.453    1    0.501
##      N      4642
## -----
##
##
## Nominal
## -----
##      Value
## -----
##      Phi-coefficient    0.00988
##      Cramer's V        0.00988
## -----
```

## Chi-square Test of Independence

How about for rank and level?

```
jmv::contTables(dat = dat,
  rows = 'rank',
  cols = 'level',
  exp = TRUE,
  phiCra = TRUE)
```

```
##
## CONTINGENCY TABLES
##
## Contingency Tables
## -----
```

```
##      rank                Doctorate    Masters    Total
## -----
##      Full      Observed      1722      310      2032
##              Expected      1684.4    347.6
##
##      Associate  Observed      1089      222      1311
##              Expected      1086.8    224.2
##
##      Assistant  Observed      971      244      1215
##              Expected      1007.2    207.8
##
##      Instructor Observed      66      18      84
##              Expected      69.6     14.4
##
##      Total      Observed      3848      794      4642
##              Expected      3848.0    794.0
## -----
##
##
## <U+03C7>2 Tests
## -----
##      Value    df    p
## -----
##      <U+03C7>2    13.6    3    0.003
##      N      4642
## -----
##
##
## Nominal
## -----
##      Value
## -----
##      Phi-coefficient    NaN
##      Cramer's V        0.0542
## -----
```

## What happens if we take this a step further...

What if our research question asks: is there a three-way contingency (sex x rank x level)?  $df1 = \# \text{ cells for sex} - 1 = 2 - 1 = 1$   $df2 = \# \text{ cells for rank} - 1 = 4 - 1 = 3$   $df3 = \# \text{ cells for level} - 1 = 2 - 1 = 1$   $N = \text{number of cells in table} (2 \times 4 \times 2) - df1 - df2 - df3$   $df = N - 1 = 10$

Three way contingency test require *log-linear modeling*.

Start with the independence model and end with the saturated model.

**Evidence for model fit: non-significant chi-square value** - no discrepancy between observed and expected values under the null model

Model 1, There are no relationships among the variables.

```
# overall model test
# 2 x 4 x 2 contingency table
# Observed = mytable
# Expected = loglm
```



```

# Expected = Expected frequencies in 2 x 4 x 2 table if there are no relationships

# Null hypothesis means that expected frequencies satisfy our model of expected values
# Alternative Hypothesis means that difference between expected and observed frequencies is significant

mytable<- xtabs(~dat$sex + dat$rank + dat$level) # table of observed values
model1 <- loglm(~dat$sex + dat$rank + dat$level, mytable)
mytable

## , , dat$level = Doctorate
##
##      dat$rank
## dat$sex  Full Associate Assistant Instructor
##   Male   1251      591      464      26
##   Female  471      498      507      40
##
## , , dat$level = Masters
##
##      dat$rank
## dat$sex  Full Associate Assistant Instructor
##   Male    223      120      119      9
##   Female   87      102      125      9

summary(model1)

## Formula:
## ~dat$sex + dat$rank + dat$level
## attr("variables")
## list(dat$sex, dat$rank, dat$level)
## attr("factors")
##      dat$sex dat$rank dat$level
## dat$sex      1      0      0
## dat$rank      0      1      0
## dat$level      0      0      1
## attr("term.labels")
## [1] "dat$sex" "dat$rank" "dat$level"
## attr("order")
## [1] 1 1 1
## attr("intercept")
## [1] 1
## attr("response")
## [1] 0
## attr(".Environment")
## <environment: R_GlobalEnv>
##
## Statistics:
##              X^2 df P(> X^2)
## Likelihood Ratio 254.4448 10      0
## Pearson          251.2707 10      0

Model 2: Rank and Sex are independent but Rank/Level are related and Sex/Level are related.

model2 <- loglm(~(dat$rank+dat$sex)*dat$level, mytable)
summary(model2)

## Formula:

```

```
## ~(dat$rank + dat$sex) * dat$level
## attr("variables")
## list(dat$rank, dat$sex, dat$level)
## attr("factors")
##          dat$rank dat$sex dat$level dat$rank:dat$level dat$sex:dat$level
## dat$rank          1      0      0              1              0
## dat$sex            0      1      0              0              1
## dat$level          0      0      1              1              1
## attr("term.labels")
## [1] "dat$rank"          "dat$sex"          "dat$level"
## [4] "dat$rank:dat$level" "dat$sex:dat$level"
## attr("order")
## [1] 1 1 1 2 2
## attr("intercept")
## [1] 1
## attr("response")
## [1] 0
## attr(".Environment")
## <environment: R_GlobalEnv>
##
## Statistics:
##                X^2 df P(> X^2)
## Likelihood Ratio 240.6000  6      0
## Pearson          237.0755  6      0
```

Model 3: *All two-way relationships*

```
model3 <- loglm(~dat$rank*dat$level + dat$level*dat$sex + dat$rank*dat$sex, mytable)
summary(model3)
```

```
## Formula:
## ~dat$rank * dat$level + dat$level * dat$sex + dat$rank * dat$sex
## attr("variables")
## list(dat$rank, dat$level, dat$sex)
## attr("factors")
##          dat$rank dat$level dat$sex dat$rank:dat$level dat$level:dat$sex
## dat$rank          1      0      0              1              0
## dat$level          0      1      0              1              1
## dat$sex            0      0      1              0              1
##          dat$rank:dat$sex
## dat$rank          1
## dat$level          0
## dat$sex            1
## attr("term.labels")
## [1] "dat$rank"          "dat$level"        "dat$sex"
## [4] "dat$rank:dat$level" "dat$level:dat$sex" "dat$rank:dat$sex"
## attr("order")
## [1] 1 1 1 2 2 2
## attr("intercept")
## [1] 1
## attr("response")
## [1] 0
## attr(".Environment")
## <environment: R_GlobalEnv>
##
```

```
## Statistics:
##               X^2 df  P(> X^2)
## Likelihood Ratio 0.7852340  3 0.8529955
## Pearson          0.7916546  3 0.8514621
```

Model 4: All two-way relationships *and the three-way relationship*

```
#saturated model or "overfit model
# this takes us one step past parsimony
# this means that the three-way relationship does not add to the model
```

```
# i.e. Chi-squared is zero
# e.g., no degrees of freedom
model14 <- loglm(~dat$rank*dat$level*dat$sex, mytable)
summary(model14)
```

```
## Formula:
## ~dat$rank * dat$level * dat$sex
## attr("variables")
## list(dat$rank, dat$level, dat$sex)
## attr("factors")
##          dat$rank dat$level dat$sex dat$rank:dat$level dat$rank:dat$sex
## dat$rank          1          0          0              1              1
## dat$level          0          1          0              1              0
## dat$sex            0          0          1              0              1
##          dat$level:dat$sex dat$rank:dat$level:dat$sex
## dat$rank                  0              1
## dat$level                  1              1
## dat$sex                    1              1
## attr("term.labels")
## [1] "dat$rank"                "dat$level"
## [3] "dat$sex"                   "dat$rank:dat$level"
## [5] "dat$rank:dat$sex"          "dat$level:dat$sex"
## [7] "dat$rank:dat$level:dat$sex"
## attr("order")
## [1] 1 1 1 2 2 2 3
## attr("intercept")
## [1] 1
## attr("response")
## [1] 0
## attr(".Environment")
## <environment: R_GlobalEnv>
##
## Statistics:
##               X^2 df  P(> X^2)
## Likelihood Ratio   0  0          1
## Pearson            0  0          1
```

Compare Models

```
stats::anova(model1,model2,model3, model4)
```

```
## LR tests for hierarchical log-linear models
##
## Model 1:
## ~dat$sex + dat$rank + dat$level
```

```
## Model 2:
## ~(dat$rank + dat$sex) * dat$level
## Model 3:
## ~dat$rank * dat$level + dat$level * dat$sex + dat$rank * dat$sex
## Model 4:
## ~dat$rank * dat$level * dat$sex
##
##           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
## Model 1    254.444762 10
## Model 2    240.600036  6  13.844726         4      0.00781
## Model 3      0.785234  3 239.814802         3      0.00000
## Model 4      0.000000  0  0.785234         3      0.85300
## Saturated   0.000000  0  0.000000         0      1.00000
```

*#Delta(Dev) is a chi-squared difference test between models  
#once difference is no longer significant, the first model is likely parsimonious fit*

The JMV way produces a cleaner output, but there are some drawbacks. Overall, it's good to know multiple ways but see which may be best for your analyses or purpose(s). For now, stick with loglm function.

*# note the similarities between 'Deviance' values and the model comparison stats with the loglm output.  
# the top table output is unknown - so look it up*

```
jmv::logLinear(
  data = dat,
  counts = NULL,
  factors = c('sex', 'rank', 'level'),
  blocks = list(
    list(
      'sex', 'rank', 'level'),
    list(
      c('sex', 'level'),
      c('rank', 'level')),
    list(
      c('sex', 'rank')),
    list(
      c('sex', 'rank', 'level'))),
  refLevels = list(
    list(
      var = 'sex',
      ref = 'Male'),
    list(
      var = 'rank',
      ref = 'Full'),
    list(
      var = 'level',
      ref = 'Doctorate')),
  modelTest = TRUE)
```

```
##
## LOG-LINEAR REGRESSION
##
## Model Fit Measures
## -----
##   Model   Deviance   AIC   R2-McF   <U+03C7>2   df   p
## -----
```

```
##      1      254.445    374    0.948    4656    5    < .001
##      2      240.600    369    0.951    4669    9    < .001
##      3         0.785    135    1.000    4909   12    < .001
##      4      2.80e-13    140    1.000    4910   15    < .001
```

```
## -----
```

```
##
```

```
##
```

```
## Model Comparisons
```

```
## -----
```

##	Model	Model	<U+03C7> <sup>2</sup>	df	p	
##	1	-	2	13.845	4	0.008
##	2	-	3	239.815	3	< .001
##	3	-	4	0.785	3	0.853

```
## -----
```

```
##
```

```
##
```

```
## MODEL SPECIFIC RESULTS
```

```
##
```

```
## MODEL 1
```

```
##
```

```
## Model Coefficients
```

```
## -----
```

##	Predictor	Estimate	SE	Z	p
##	Intercept	6.925	0.0260	266.0	< .001
##	sex:				
##	Female - Male	-0.421	0.0300	-14.0	< .001
##	rank:				
##	Associate - Full	-0.438	0.0354	-12.4	< .001
##	Assistant - Full	-0.514	0.0363	-14.2	< .001
##	Instructor - Full	-3.186	0.1113	-28.6	< .001
##	level:				
##	Masters - Doctorate	-1.578	0.0390	-40.5	< .001

```
## -----
```

```
##
```

```
##
```

```
## MODEL 2
```

```
##
```

```
## Model Coefficients
```

```
## -----
```

##	Predictor	Estimate	SE	Z	p
##	Intercept	6.9504	0.0274	253.851	< .001
##	sex:				
##	Female - Male	-0.4307	0.0330	-13.053	< .001
##	rank:				
##	Associate - Full	-0.4582	0.0387	-11.835	< .001
##	Assistant - Full	-0.5729	0.0401	-14.276	< .001
##	Instructor - Full	-3.2616	0.1254	-26.004	< .001
##	level:				
##	Masters - Doctorate	-1.7361	0.0696	-24.956	< .001
##	sex:level:				
##	(Female - Male):(Masters - Doctorate)	0.0534	0.0794	0.673	0.501

```
## rank:level:
## (Associate - Full):(Masters - Doctorate) 0.1243 0.0961 1.294 0.196
## (Assistant - Full):(Masters - Doctorate) 0.3335 0.0945 3.528 < .001
## (Instructor - Full):(Masters - Doctorate) 0.4154 0.2730 1.522 0.128
```

```
## -----
##
##
## MODEL 3
##
## Model Coefficients
## -----
## Predictor Estimate SE Z p
## -----
## Intercept 7.12988 0.0279 255.4574 < .001
## sex:
## Female - Male -0.97021 0.0512 -18.9376 < .001
## rank:
## Associate - Full -0.74933 0.0484 -15.4948 < .001
## Assistant - Full -0.98667 0.0521 -18.9290 < .001
## Instructor - Full -3.81665 0.1807 -21.1208 < .001
## level:
## Masters - Doctorate -1.71257 0.0656 -26.1009 < .001
## sex:level:
## (Female - Male):(Masters - Doctorate) -0.00766 0.0816 -0.0938 0.925
## rank:level:
## (Associate - Full):(Masters - Doctorate) 0.12573 0.0972 1.2933 0.196
## (Assistant - Full):(Masters - Doctorate) 0.33539 0.0966 3.4711 < .001
## (Instructor - Full):(Masters - Doctorate) 0.41775 0.2741 1.5239 0.128
## sex:rank:
## (Female - Male):(Associate - Full) 0.80176 0.0745 10.7664 < .001
## (Female - Male):(Assistant - Full) 1.05245 0.0761 13.8385 < .001
## (Female - Male):(Instructor - Full) 1.30832 0.2269 5.7665 < .001
## -----
```

```
##
##
## MODEL 4
##
## Model Coefficients
## -----
## Predictor Estimate SE Z p
## -----
## Intercept 7.1317 0.0283 252.244 < .001
## sex:
## Female - Male -0.9768 0.0541 -18.070 < .001
## rank:
## Associate - Full -0.7499 0.0499 -15.023 < .001
## Assistant - Full -0.9918 0.0544 -18.247 < .001
## Instructor - Full -3.8736 0.1981 -19.549 < .001
## level:
## Masters - Doctorate -1.7245 0.0727 -23.725 < .001
## sex:level:
## (Female - Male):(Masters - Doctorate) 0.0356 0.1375 0.259 0.799
## rank:level:
## (Associate - Full):(Masters - Doctorate) 0.1302 0.1237 1.052 0.294
```

##	(Assistant - Full):(Masters - Doctorate)	0.3638	0.1259	2.890	0.0
##	(Instructor - Full):(Masters - Doctorate)	0.6637	0.3935	1.686	0.0
##	sex:rank:				
##	(Female - Male):(Associate - Full)	0.8056	0.0814	9.900	< .0
##	(Female - Male):(Assistant - Full)	1.0655	0.0840	12.689	< .0
##	(Female - Male):(Instructor - Full)	1.4076	0.2577	5.463	< .0
##	sex:rank:level:				
##	(Female - Male):(Associate - Full):(Masters - Doctorate)	-0.0269	0.2018	-0.133	0.9
##	(Female - Male):(Assistant - Full):(Masters - Doctorate)	-0.0750	0.1986	-0.378	0.7
##	(Female - Male):(Instructor - Full):(Masters - Doctorate)	-0.4664	0.5519	-0.845	0.4
##	-----				