カイニ乗検定

竹内理・水本篤(2014). 『外国語教育研究ハンドブック』松柏社 (pp.147-160) より

Q 「ESL コースで学ぶ学習者の母語を調べてみたところ、表1のようになったんだ。中国人が一番多いんじゃないかな」

表 1

	日本語	タイ語	中国語	合計
人数	18	24	48	90

このように、<u>名義尺度</u>で分類された<u>人数や回数の頻度の差</u>を統計的に検証する場合は、カイ二乗検定を使う。

● カイニ乗検定の適用条件

① データが名義尺度(categorical data) であるか

• カイ二乗検定は、名義尺度ごとにカテゴリーで分類されたカテゴリーデータ (categorical data)に対する検定。

② データが累積の頻度であるか

- %などのデータには使用できない。
- 比率データは実際の人数や頻度に戻してから使わなければならない。

③ データが独立しているか

ある1つ(1人)のサンプルから得られるデータは、1度しかカウントされてはいけない

表 2 2x2 の分割表

	質問あり	質問なし
女子	15	7
男子	8	16

表3 2x3 分割表の例

質問回数				
	0~5 回	6~10 回	11 回以上	
女子	15	11	8	
男子	10	9	7	

④ 期待値が5以上であるか

- カイ二乗検定は、比較されるすべてのグループで、同じ度数が観測されると仮定した場合の値(期待値)と、観察された度数(実測値)がどの程度異なっているかを調べる検定。
- このため、いずれかのグループの期待値が5未満になると、結果が不正確になる。

• このような場合は、より多くのデータを収集して期待値が5以上になるようにするか、フィッシャーの正確(直接)確率検定(Fisher's exact test を使う)。

⑤ どのような分割表になるデータか

- 自由度が1だと、カイ二乗検定では、第1種の誤り(実際には差がないのに差があると主張してしまう)が起こる危険性がある。(表2などが例)
- 自由度が1の場合は、「イェーツの補正」と呼ばれる連続性の補正を用いる。
- ただし、フィッシャーの正確確率検定を用いれば、正確なp値が得られる。

● 適合度検定と独立性の検定

カテゴリー変数が1の場合→適合度検定(表1)2以上の場合→独立性の検定(表2)

● 残差分析と多重比較

- カイ二乗検定では、全体として差がある事は示してくれるが、ここのどのカテゴリー (セル) に差があるのかまでは示してくれない。
- どのセルが期待値よりも統計的に優位に大きい(小さい)かを調べるには、残差分析(residual analysis)を行う。
- 残差とは、実際の値(観測値)がモデルの値(期待値)からどのくらいかけ離れているかを示した値。

● 効果量

- カイ二乗検定で効果量を報告する際には、主に相関係数の一種であるファイ(φ) 係数、またはクラマーの V (Cramer's V) という指標を用いる。
- 自由度は2以上の場合はクラマーのVを使用する。自由度が2以上の時にφ係数を使用すると、効果量が0(最低値)であるはずの場合に0より高い値になってしまう事がある。

Chapter 8: Chi-square

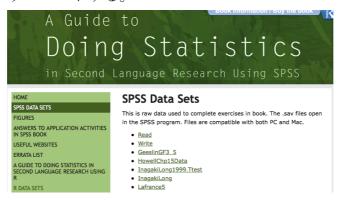
8.1 Summarizing and Visualizing Data8.11 Summary Tables for Goodness-of-Fit Data

- カイ二乗適合度検定
- ▶ スペイン語話者(L1 と L2)
- ▶ 従属変数:動詞の選択(ser, ester, or both)
- ➤ 独立変数: Spanish L1 speaker

Portuguese L1 speaker

Portuguese L1 learner of Spanish L2

- ▶ ここで扱うのは Spanish L1 のデータについてのみ。
- インターネットの SPSS データセットのページに行って、GeeslinGF3_5.sav というファイルをダウンロードする。



◆ 頻度表を作る

- R コマンダーで GeeslinGF3 5.sav ファイルを読み込む。
- 1. Data > import data > from SPSS data set
- 2. geeslin3 という名前で保存。
- 3. Statistics > Summaries > Frequency Distributions
- 4. item3 で ok

結果↓ (頻度とパーセンテージ)

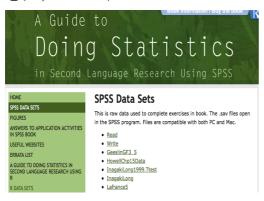
```
counts:
item3
Estar Ser Both
13 4 2

percentages:
item3
Estar Ser Both
68.42 21.05 10.53
```

 R console でも以下のRコードを使うことで、同じ結果を出すことが可能 table(geeslin3\$item3)
 100*table(geeslin3\$item3)/sum(table(geeslin3\$item3))

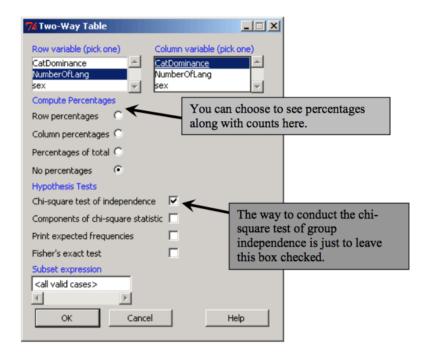
8.1.2 Summaries of Group Comparison Data (Crosstabs)

- ▶ 変数が1つ以上の場合
 - Number of languages
 - Language dominance
- インターネットの SSP データセットのページに行って、BEQ.Dominance というファイルをダウンロード



- Pandey2000
- Torres
- Yates2003
- BeautifulRose
- BEQ.Context
- BEQ.Dominance
- BEQ
- BEQ.Swear
- ClassTime
- DeKeyser2000
- EllisYuan

- ◆ Two-Way-Table を作る
- R コマンダーで BEQ.Dominance ファイルを読み込む。
- 1. Data > import data > from SPSS data set
- 2. beqDom という名前で保存。
- 3. Statistics > Contingency tables > Two-Way Table



4. 結果↓ (Table8.1)

```
Frequency table:
            catdominance
numberoflang YES
                  NO YESPLUS
       Two
              94
                   26
                           17
       Three 159
                   26
                           83
       Four
             148 23
                          110
             157 30
       Five
                          163
        Pearson's Chi-squared test
data:
       . Table
X-squared = 59.581, df = 6, p-value = 5.476e-11
```

● R console でも以下のRコードを使うことで、同じ結果を出すことが可能

.Table <- xtabs(~numberoflang+catdominance, data=beqDom) .Table

```
結果↓
```

```
> .Table <- xtabs(~numberoflang+catdominance, data=beqDom)</pre>
> .Table
            catdominance
numberoflang YES
                   NO YESPLUS
       Two
              94
                   26
                           17
                   26
       Three 159
                           83
       Four 148
                   23
                          110
       Five 157
                   30
                          163
```

● Table8.1 のような結果がすでにあって、そこから 2×2 の表を作りたいとき TM<-matrix(c(12,0,18,16),nrow=2,ncol=2,byrow=T, dimnames=list(c("Relative Clauses", "No +RCs"), c("Method A", "Method B")))

TM

結果↓

- ◆ Multi-Way Table をつくる
 - ▶ 変数が3つ以上の場合
- 1. Statistics > Contingency tables > Multi-Way Table

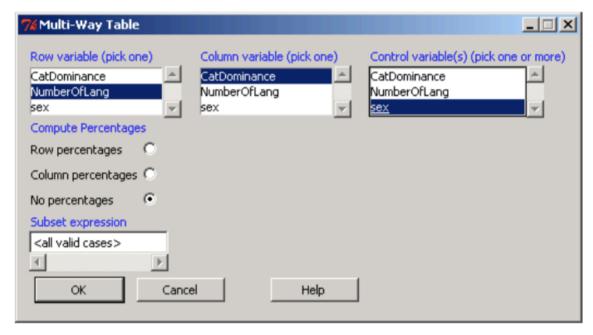


Figure 8.3 How to make a crosstab with three or more categorical variables in R.

◆ 性別ごとの表ができる。

```
Output
Frequency table:
, , sex = F
           catdominance
numberoflang YES NO YESPLUS
      Two
            66 19
      Three 113 18
                        61
      Four 95 18
                        92
      Five
             99 23
                        112
, , sex = M
           catdominance
numberoflang YES NO YESPLUS
      Two
             28
                  7
                         3
      Three 46 8
                         22
      Four 53 5
                         18
      Five
             58 7
                         51
```

◆ Rコンソールでは

Four

Five

53 5

58 7

.Table <- xtabs(~numberoflang+catdominance+sex, data=beqDom)

.Table

```
> .Table <- xtabs(~numberoflang+catdominance+sex, data=beqDom)</p>
> .Table
, sex = F
           catdominance
numberoflang YES NO YESPLUS
      Two
             66 19
      Three 113 18
                         61
      Four
             95 18
                         92
      Five
             99 23
                        112
, , sex = M
           catdominance
numberoflang YES NO YESPLUS
      Two
             28
                7
                         3
      Three 46 8
                         22
```

18

51

◆ 他にもパーセンテージを出せる便利なコードがある

rowPercents(.Table) # Row Percentages

colPercents(.Table) # Column Percentages

totPercents(.Table) # Percentage of Total; this only works with two-way tables

```
> rowPercents(.Table)
, sex = F
          catdominance
numberoflang YES NO YESPLUS Total Count
          66.7 19.2
                      14.1 100.0
      Three 58.9 9.4
                      31.8 100.1
                                   192
      Four 46.3 8.8 44.9 100.0
                                  205
      Five 42.3 9.8 47.9 100.0 234
, , sex = M
          catdominance
numberoflang YES NO YESPLUS Total Count
           73.7 18.4 7.9 100.0
      Two
      Three 60.5 10.5
                       28.9 99.9
                                    76
      Four 69.7 6.6 23.7 100.0
                                   76
      Five 50.0 6.0 44.0 100.0
                                  116
 > colPercents(.Table) # Column Percentages
  , , sex = F
             catdominance
               YES NO YESPLUS
 numberoflana
               17.7 24.4
        Two
                          5.0
        Three 30.3 23.1
                             21.9
               25.5 23.1
        Four
                             33.0
        Five 26.5 29.5
                            40.1
        Total 100.0 100.1
                            100.0
        Count 373.0 78.0
                            279.0
 , , sex = M
             catdominance
               YES
                      NO YESPLUS
 numberoflang
        Two
               15.1 25.9
                            3.2
        Three 24.9 29.6
                            23.4
        Four 28.6 18.5
                           19.1
        Five 31.4 25.9
                           54.3
        Total 100.0 99.9
                           100.0
        Count 185.0 27.0
                           94.0
```

> totPercents(.Table) # Percentage of Total; this only works with two-way tables

```
tab Total
<NA>
        6.4 13.2
      10.9 25.9
<NA>
<NA>
       9.2 27.1
       9.6 33.8
<NA>
<NA>
       1.8 13.2
<NA>
       1.7 25.9
       1.7 27.1
<NA>
<NA>
       2.2 33.8
       1.4 13.2
<NA>
<NA>
       5.9 25.9
<NA>
       8.9 27.1
      10.8 33.8
<NA>
<NA>
       2.7 13.2
       4.4 25.9
<NA>
       5.1 27.1
<NA>
       5.6 33.8
<NA>
       0.7 13.2
0.8 25.9
<NA>
<NA>
       0.5 27.1
<NA>
<NA>
       0.7 33.8
<NA>
       0.3 13.2
       2.1 25.9
<NA>
<NA>
       1.7 27.1
<NA>
       4.9 33.8
Total 100.0 600.0
```

8.1.3 Visualizing Categorical Data

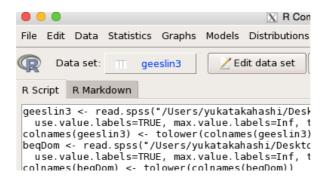
Bar plots の作り方を教えるが、私 (author) は Bar plots はオススメしない。

8.1.4 Bar plots in R

Commander でも作れるが、より sophisticated な Bar plots を作りたければ、R code を 使う必要がある。まずは geeslin3 のデータを使って、commander の操作から。

◆ Bar plot of ONE categorical variable

① Data set を becDom から geeslin3 に切り替える。(青い文字のところ)



② Graphs > Bar graphs > 好きな変数(item3)を選ぶ 結果↓

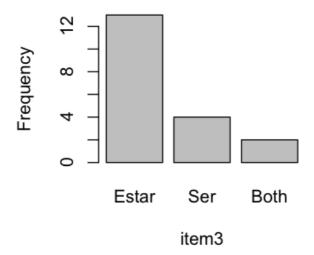
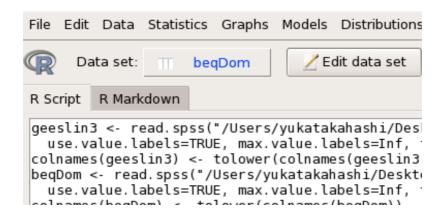


Figure 8.4 Barplot of ONE category variable in R from Geeslin and Guijarro-Fuentes (2006) data

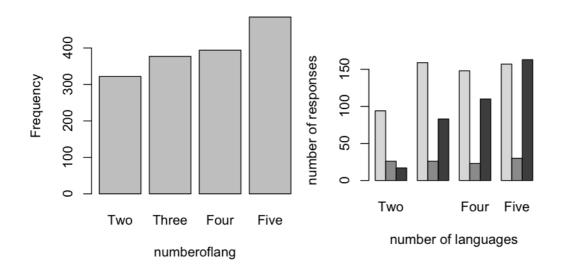
◆ R Console では

barplot(table(geeslin3\$item3), xlab="item3", ylab="frequency")

- Bar plot of TWO categorical variables
- ① Data set を beqDom に切り替える。(青い文字のところ)



②Graphs > Bar graph > numbeoflang > ok



◆ R console では

install.packages("epitools")

library(epitools)
colors.plot(TRUE) #パレットの上で左クリック。右クリックでエスケープ

attach(beqDom)

barplot(tapply(catdominance, list(catdominance, numberoflang), length), col=c("grey83", "grey53", "grey23"), beside=T, ylab="number of responses", xlab="number of languages")

locator() #barplotの上で左クリック。右クリックでエスケープ。

legend(.827,158,legend=c("L1 dominant", "LX dominant", "L1+more dominant"), fill=c("grey83", "grey53", "grey23"))

◆ Raw count dataの代わりにSummary dataがある場合

TM<-matrix(c(12,0,18,16),nrow=2,ncol=2,byrow=T, dimnames=list(c("Relative Clauses", "No +RCs"), c("Method A", "Method B")))

barplot(TM,beside=T, main="Teaching Method and Production of Relative Clauses")

legend(3.24,14.1,c("Rel.clauses", "No RCs"),fill=c("grey22", "grey83"))

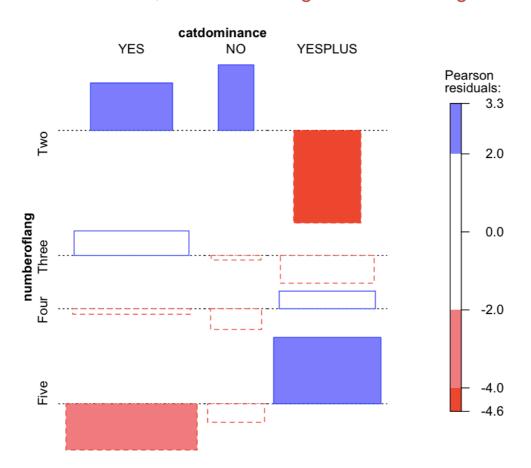
8.1.6 Association Plots

- ① install.packages("vcd") で、パッケージをダウンロードする。
- ② そして、データを contingency table の形にする。
 library(vcd)
 (DOM=structable(catdominance ~ numberoflang,data=beqDom))
 - > (DOM=structable(catdominance ~ numberoflang,data=beqDom))

	catdominance	YES	NO	YESPLUS
numberoflang				
Two		94	26	17
Three		159	26	83
Four		148	23	110
Five		157	30	163
Two Three Four		159 148	26 23	83

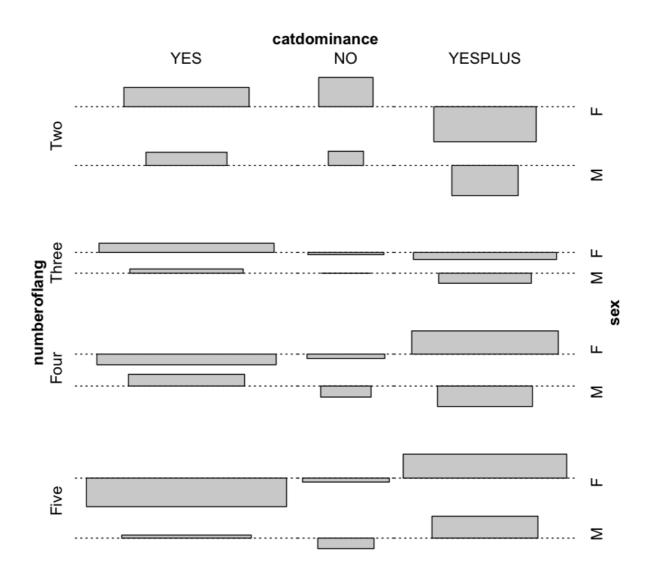
③ Association plot をつくる。

assoc(DOM, gp=shading_Friendly,
labeling_args=list(set_varnames= c(CatDominance ="L1
Dominant", NumberOfLang ="Number of Lges Known")))



◆ 3つめのvariables (sex) を含めてAssociation plot をつくる方法

DOM3=structable(catdominance~numberoflang+sex,data=beqDom) assoc(DOM3)

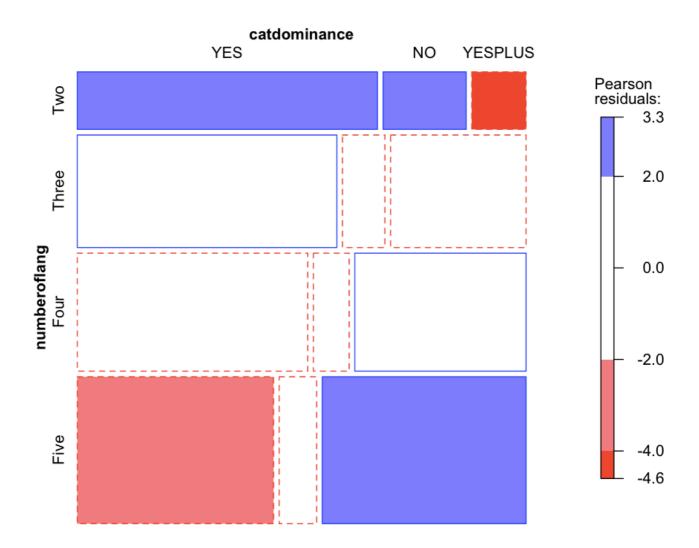


Mosaic plot with TWO variables (Dewaele and Pavlenko data)

①Mozaic plot をつくるコマンド

mosaic(DOM, gp=shading_Friendly, (ここで切る)

labeling_args=list(set_varnames= c(CatDominance="L1 Dominant", NumberOfLangs="Number of Lges Known")))



- ◆ 点線は期待値との有意な差がなかったところ。
- ◆ 青は期待値より有意に頻度が高かったところ
- ◆ 赤は期待値より有意に頻度が低かったところ

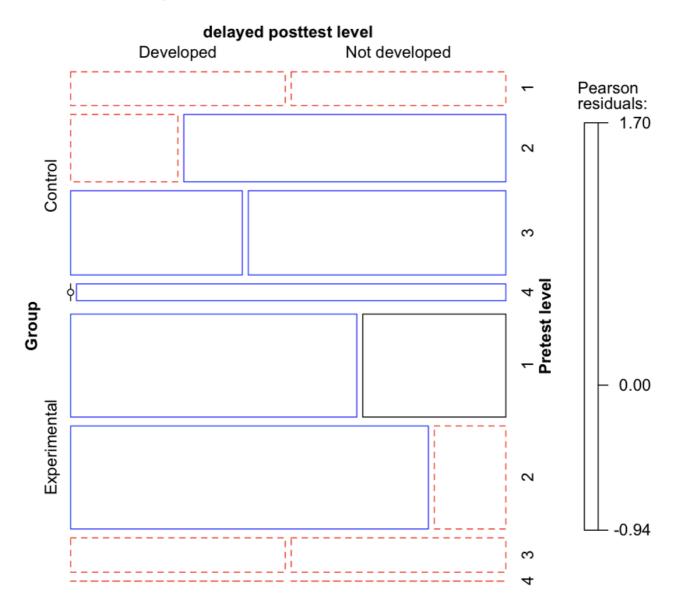
- ◆ Mozaic plot with THREE variables (Mackey and Silver data).
- ① Mackey and Silverのデータをダウンロードする。
 - ▼ Larsonniamrarda
 - LeowMorganShort
 - Lyster.Oral
 - Lyster.Written
 - MackeySilver2005
 - Motivation
 - MunroDerwingMorton
 - Murphy.RepeatedMeasures
 - Obarow.Original
- ②Rコマンダーでデータを読み込み、Mackeyで保存。
- ③Rコンソールで以下のコマンドでテーブルを作り、

(DEV=structable(developdelpost~group+pretest,data=Mackey))

group	pretest		•
Control	1		1 1
	2		1 3
	3		2 3
	4		0 1
Experimental	1		4 2
	2	1	5 1
	3		1 1
	4		0

④以下のコマンドでmosaic plotを作る。

mosaic(DEV, gp=shading_Friendly, labeling_args=list(set_varnames= c(developdelpost="delayed posttest level", group="Group", pretest="Pretest level")))

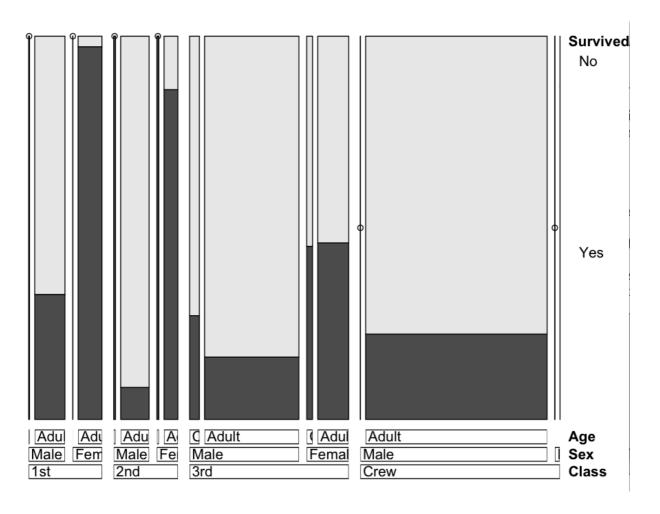


- ◆ 点線は期待値との有意な差がなかったところ。
- ◆ 普通の線は期待値より有意に頻度が異なったところ

◆ タイタニックの生存者/死亡者データ (age, sex, class)

library(vcd)

doubledecker(Survived~Class+Sex+Age, data=Titanic)



8.3 One- Way Goodness-Of-Fit Test

- ◆ 動詞 (ser, ester, both) のpreferenceに違いがあるか。
- ①GeeslinGF3 5.sav fileを読み込み、geeslin3という名前で保存。
- ② Statistics > Summaries > Frequency distributions
- ③ Item3を選択し、Chi-square goodnesss-of-fit-testにチェックを入れる。

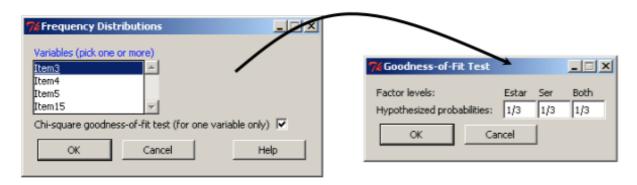


Figure 8.10 Dialogue boxes for goodness-of-fit chi-square test.

```
counts:
item3
Estar
        Ser
             Both
   13
          4
percentages:
item3
Estar
        Ser Both
68.42 21.05 10.53
        Chi-squared test for given
        probabilities
data: .Table
X-squared = 10.842, df = 2,
p-value = 0.004422
```

- ◆ 全て同じ確率 (1/3) という帰無仮説は棄却される (p=0.0044)
- ◆ カイ二乗検定は、どのitemが好まれるかまでは教えてくれない。しかし、このケースでは頻度を見るとestarが好まれている事がわかる。

◆ Rコマンダーでは:

```
chisq.test(table(geeslin3$item3), correct=false)
```

> chisq.test(table(geeslin3\$item3), correct=false)

Chi-squared test for given probabilities

```
data: table(geeslin3$item3)
X-squared = 10.842, df = 2, p-value = 0.004422
```

◆ もし、全て同じ確率ではなく、ser(40%), ester (40%), both(20%)という確率で検定したい場合は、以下のコードを使う。

```
prob=c(.4,.4,.2)
```

chisq.test(table(geeslin3\$item3),correct=false,p=prob)

```
> prob=c(.4,.4,.2)
> chisq.test(table(geeslin3$item3),correct=FALSE,p=prob)
Warning in chisq.test(table(geeslin3$item3), correct = FALSE, p = prob):
    Chi-squared approximation may be incorrect

    Chi-squared test for given probabilities

data: table(geeslin3$item3)
X-squared = 6.3947, df = 2, p-value = 0.04087
```

8.4 Two-Way Group-Independence Test

- ◆ 独立性の検定は、二つの変数の間に何も関連がない事を帰無仮説にします。
- ◆ 話す言語の数によるlanguage dominanceに影響があるかを見る。
- ①コマンダーのData setをbegDomデータに切り替える。
- ③ Statistics > contingency tables > Two-way-tables

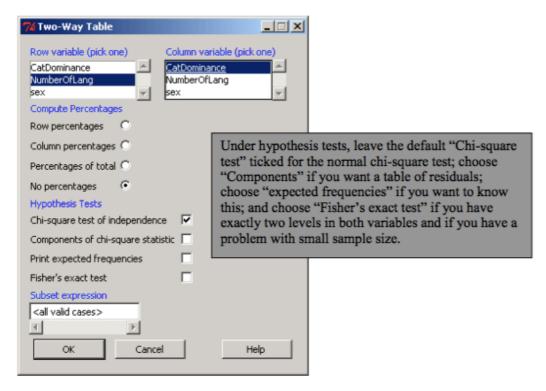


Figure 8.12 How to perform a two-way group-independence chi-square in R.

```
Frequency table:
             catdominance
numberoflang YES
                   NO YESPLUS
               94
                   26
       Two
                            17
       Three 159
                   26
                            83
              148
       Four
                   23
                           110
       Five
              157
                   30
                           163
        Pearson's Chi-squared test
data:
       . Table
X-squared = 59.581, df = 6, p-value = 5.476e-11
```

- ◆ カイ二乗値= (59.6) でとても大きい、p=.00000000055でとても小さい。
- →帰無仮説は棄却される。話す言語の数と、language dominanceには関連がある。

注:今回は大きいデータなので、warningはでてこないが、データの数が少ないと、コマンダーのメッセージボックスに以下のような警告が出る。

```
Messages

WARNING: 3 expected frequencies are less than 1
6 expected frequencies are less than 5
```

Figure 8.13 R Commander's warning about low expected frequencies.

◆ Rコンソールでは:

chisq.test(xtabs(~catdominance+numberoflang, data=beqDom),
correct=FALSE)

```
> chisq.test(xtabs(~catdominance+numberoflang, data=beqDom), correct=FALSE)

Pearson's Chi-squared test

data: xtabs(~catdominance + numberoflang, data = beqDom)
X-squared = 59.581, df = 6, p-value = 5.476e-11
```

- ◆ 期待値が1以下のものがある場合に警告が出る。
 - →1以下のものがいくつあるか見たい場合

.Test= chisq.test(xtabs(~catdominance + numberoflang,data= begDom), correct=FALSE)

.Test\$expected

◆ effect sizeとlikelihood ratio testの結果を求めるとき

library(vcd)

summary(assocstats(xtabs(~catdominance+numberoflang, data=beqDom)))

> summary(assocstats(xtabs(~catdominance+numberoflang, data=beqDom)))

```
Call: xtabs(formula = ~catdominance + numberoflang, data = beqDom)

Number of cases in table: 1036

Number of factors: 2

Test for independence of all factors:

Chisq = 59.58, df = 6, p-value = 5.476e-11

X^2 df P(> X^2)

Likelihood Ratio 63.742 6 7.7904e-12

Pearson 59.581 6 5.4760e-11
```

Phi-Coefficient : NA Contingency Coeff.: 0.233 Cramer's V : 0.17

◆ linear-by-linear association test (線形関連検定)

library(coin)

independence_test(catdominance~numberoflang,data=beqDom,tes tstat="quad")

> independence_test(catdominance~numberoflang,data=beqDom,teststat="quad")

Asymptotic General Independence Test

data: catdominance by
 numberoflang (Two, Three, Four, Five)
chi-squared = 59.523, df = 6, p-value = 5.625e-11