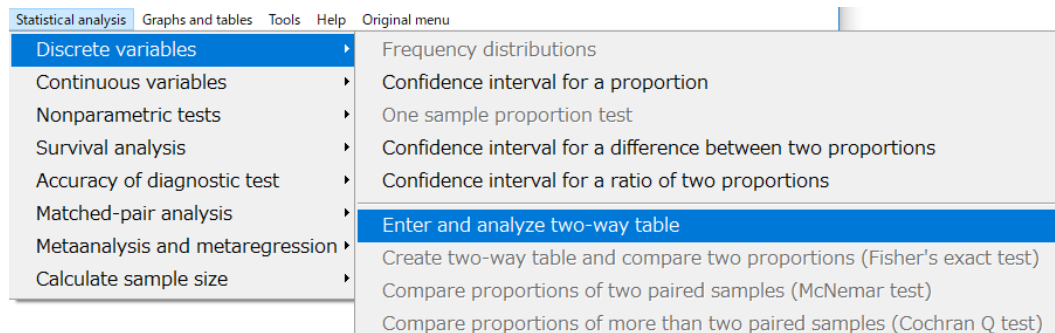


# Cross tabulation

- Comparison of proportions
  - Last week we learned the comparison of means or medians between the 2 quantitative variables
  - In the case of categorical variables, comparison of proportions can be applied (In the case of more than 2 groups, it's similarly applicable).
- Independence of 2 categorical variables
  - Making cross table, then test the null-hypothesis that the 2 variables are independent each other by chi-square test or fisher's exact probability test.
- The analyses above are essentially same.
- (cf.) "How large the effect of the difference of groups on disease occurrence is" can be measured by odds ratios, rate ratios, risk ratios, rate differences or risk differences (already learned)
- (cf.2) Association between 2 categorical variables can be assessed by Pearson's contingency coefficients, phi coefficients, Cramer's V, and polychoric correlation coefficients

# Comparison of 2 proportions

- If we compare the proportions of smoker between patients and controls, 40 among 100 and 20 among 100, respectively, type `prop.test(c(40, 20), c(100, 100))` in the script window, and click [submit], then you can get the result,  $p\text{-value}=0.00337$
- The way to conduct it in EZR menu, see below.



Statistical analysis | Graphs and tables | Tools | Help | Original menu

- Discrete variables
- Continuous variables
- Nonparametric tests
- Survival analysis
- Accuracy of diagnostic test
- Matched-pair analysis
- Metaanalysis and metaregression
- Calculate sample size

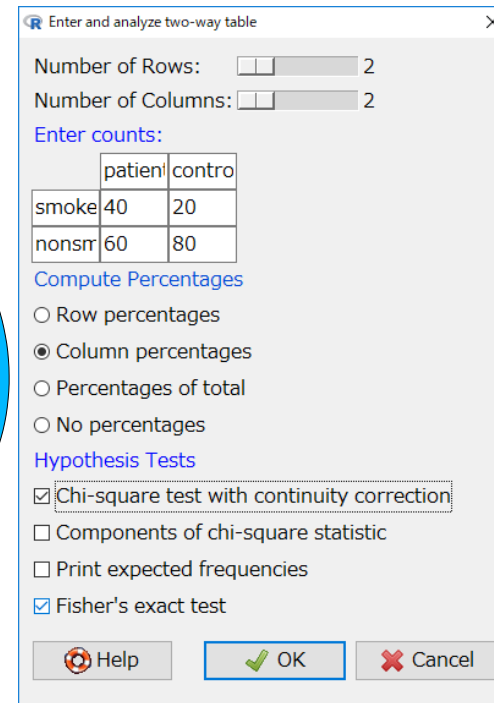
- Frequency distributions
- Confidence interval for a proportion
- One sample proportion test
- Confidence interval for a difference between two proportions
- Confidence interval for a ratio of two proportions
- Enter and analyze two-way table**
- Create two-way table and compare two proportions (Fisher's exact test)
- Compare proportions of two paired samples (McNemar test)
- Compare proportions of more than two paired samples (Cochran Q test)

```
> colPercents(.Table) # Column Percentages
      patient control
smoker      40      20
nonsmoker   60      80
Total      100     100
Count      100     100

> .Test <- chisq.test(.Table, correct=TRUE)
> .Test

      Pearson's Chi-squared test with Yates' continuity correction

data: .Table
X-squared = 8.5952, df = 1, p-value = 0.00337
```



Enter and analyze two-way table

Number of Rows: 2  
Number of Columns: 2

Enter counts:

	patient	contro
smoke	40	20
nonsm	60	80

Compute Percentages

Row percentages  
 Column percentages  
 Percentages of total  
 No percentages

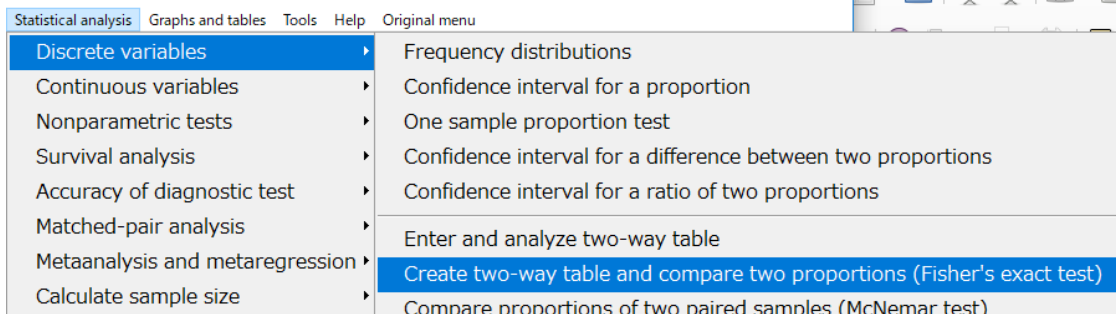
Hypothesis Tests

Chi-square test with continuity correction  
 Components of chi-square statistic  
 Print expected frequencies  
 Fisher's exact test

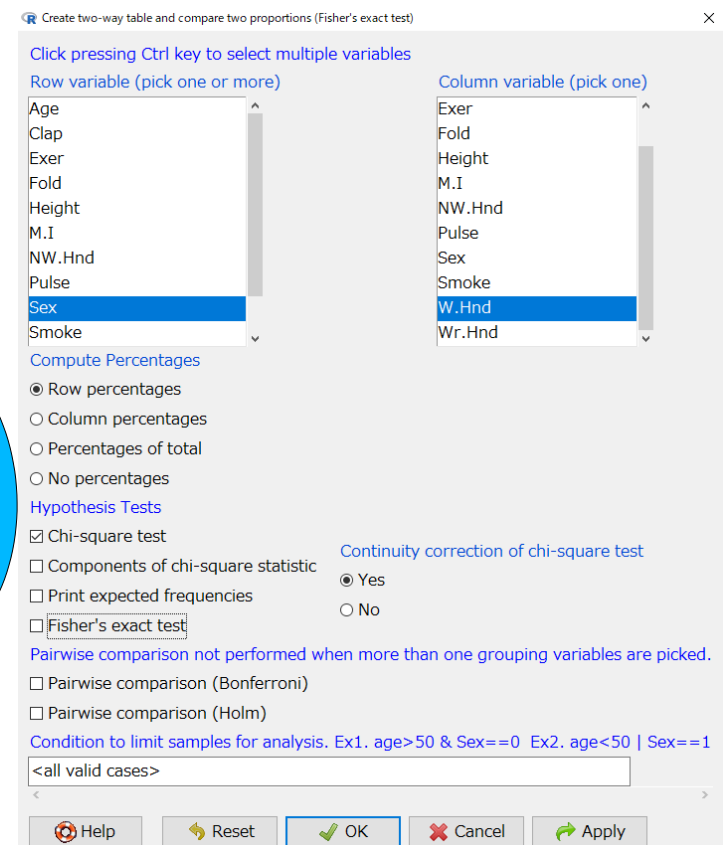
Help OK Cancel

# Comparison of 2 proportions (cont'd)

- From the raw category data, for example, in the case of [survey] data of [MASS] package to compare the proportions of lefties between males and females, type `prop.test(table(survey$Sex, survey$W.Hnd))` in the script window and click [submit].
- The way to conduct it in EZR menu, see below.



```
> Fisher.summary.table
      W.Hnd=Left W.Hnd=Right Chisq.p.value
Sex=Female      7          110          0.627
Sex=Male       10          108
```



# Chi-square test for independence

- In the chi-square test for independence,
  - First, make a cross table (contingency table) by the combination of 2 categorical variables: each number of cells can be denoted as  $n[i, j]$ . If each variable is binary, the table will be a 2x2 cross table.
  - Denote  $N1=n[1, 1]+n[2, 1]$ ,  $N2=n[1, 2]+n[2, 2]$ ,  $M1=n[1, 1] +n[1, 2]$ ,  $M2=n[2, 1]+n[2, 2]$ ,  $N=N1+N2=M1+M2$
  - Calculate the expected number  $x[i, j]$  of each cell if row and column variables are independent.
    - $x[1, 1]=N1*M1/N$ ,  $x[2, 1]=N1*M2/N$ ,  $x[1, 2]=N2*M1/N$ ,  
 $x[2, 2]=N2*M2/N$ .
  - Calculate chi-squared as  $\chi^2=\sum \{(n[i, j] - x[i, j])^2/x[i, j]\}$
  - The value of  $\chi^2$  obeys chi-square distribution of d.f.=1.
- `chisq.test(matrix(c(80, 20, 55, 45), 2, 2))` gives the result for the test for independence between smoking and lung cancer (pp.91-92).
- The way in EZR is similar to the comparison of proportions.

# Fisher's exact probability test for independence

- Chi-square test uses approximation (the combination can take only integer but the chi-square distribution is continuous) and thus is not appropriate for a small-sized sample.
- Fisher invented the direct calculation method to calculate the sum of the probabilities for the given combination and less plausible combinations under the condition of independence, where all marginal numbers are fixed, then the probability obeys hypergeometric distribution.
- This method for the same case is done by `fisher.test(matrix(c(80, 20, 55, 45), 2, 2))`
- The way in EZR is similar to the comparison of proportions, but check the [Fisher's exact test] of the dialogue.

Enter and analyze two-way table

Number of Rows: 2  
Number of Columns: 2

Enter counts:

	Lung.C	Health
Smoke	80	55
No sm	20	45

Compute Percentages

Row percentages  
 Column percentages  
 Percentages of total  
 No percentages

Hypothesis Tests

Chi-square test with continuity correction  
 Components of chi-square statistic  
 Print expected frequencies  
 Fisher's exact test

Help OK Cancel

```
Smoke      Lung.C Healthy Fisher.p.value  
No smoke   20      45      0.000259
```

# Comparison of proportions among 3 or more groups

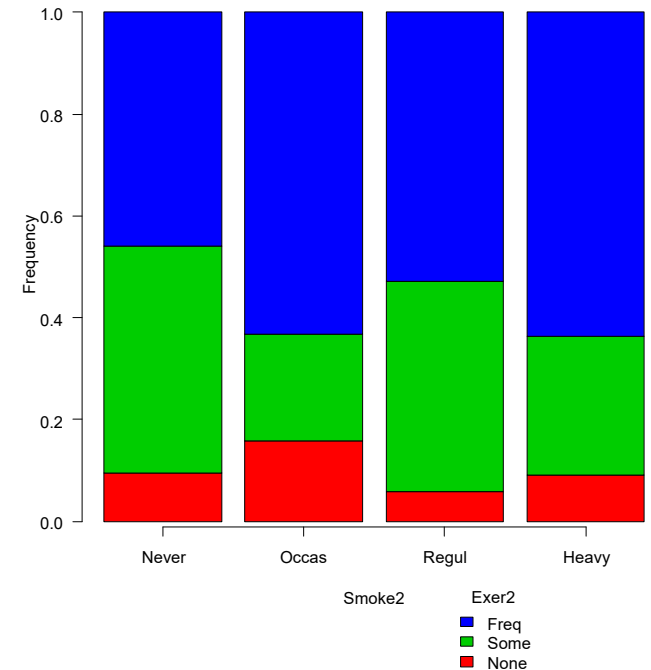
- Comparison of proportions among 3 groups
  - Schistosomiasis positive results of Kato-Katz fecal test were obtained from 60, 30, and 8 in 100 fishermen, 80 farmers, and 30 office workers, respectively, then  
`prop.test(c(60, 30, 8), c(100, 80, 30))`
    - For graph of positive proportions, try  
`barplot(c(60, 30, 8)/c(100, 80, 30))`
    - For pairwise comparisons, try  
`pairwise.prop.test(c(60, 30, 8), c(100, 80, 30), p.adjust.method="fdr")`
  - If we assume the tendency in the risk of subpopulations as 4:2:1, Cochran-Armitage test is applicable  
`prop.trend.test(c(60, 30, 8), c(100, 80, 30), c(4, 2, 1))`
- If the number of groups is more than 3, the way is similar.
- In EZR, directly entering the data of positive and negative numbers instead of positive and total, simple chi-square test is possible.

# Association statistics (1): Pearson's contingency coefficients, and so on

- The extent of association between 2 categorical variables with 2 categories can be measured by the following indicators.
- Let's consider the relationship between smoking habits and lung cancer as already mentioned: The numbers of past smokers were 80 and 55 in Lung cancer patients and healthy controls, respectively.
  - Odds Ratio (rate and risk are unavailable in case-control study) is  $(80/20)/(55/45)$ . It can be obtained by either of `fisher.test(matrix(c(80, 20, 55, 45), 2, 2))`  
`library(fmsb); oddsratio(80, 20, 55, 45)`
  - The results are slightly different because `fisher.test()` uses maximum likelihood estimation, but both significant.
  - Other association statistics can be obtained by `assocstats()` function of `vcd` package, such as `library(vcd); assocstats(matrix(c(80, 20, 55, 45), 2, 2))`
  - Pearson's contingent coefficients and Cramer's V are applicable for the variables with 3 or more levels **without order**.

# Association statistics (2): Polychoric correlation coefficients

- When you like to test the association between non-binary categorical variables **with order**, polychoric correlation coefficient is better than rank correlation or contingency coefficients.
- Please consider the association of smoking habits and exercise in the [survey] data.
  - First, reorder factor levels of [Smoke] and [Exer] as [Smoke2] and [Exer2].
  - `library(polycor)`  
`polychor(survey$Smoke2, survey$Exer2, std.err=TRUE)`
  - It's different from  
`polychor(survey$Smoke, survey$Exer, std.err=TRUE)`



Active data set | Statistical analysis | Graphs and tables | Tools | Help | Original menu

Variables

- Show variables in active data set
- Create new variable
- Bin numeric variable
- Bin numeric variable with specified threshold
- Bin numeric variable to more than 2 groups with specified thresholds
- Logarithmic transformation
- Convert numeric variables to factors
- Convert factors or character variables of numeric data to numeric variables
- Convert all character variables to factors
- Reorder factor levels**
- Recode variables
- Drop unused factor levels

Reorder Factor Levels

Factor (pick one)

- Clap
- Exer
- Fold
- M.I
- Sex
- Smoke**
- W.Hnd

Name for factor

Smoke2

Make ordered factor

Reorder Levels

Old Levels	New order
Heavy	4
Never	1
Occas	2
Regul	3

OK Cancel



# Repeated or Inter-rater agreement of categorical variables (Chap.13)

- When ordered or categorical variables were measured repeatedly or evaluated by multiple raters (observers), the result can be summarized as two-dimensional cross tabulation.
- However, common statistical testing for two-dimensional cross table like chi-square test or fisher's exact test is completely inadequate, because repeated or inter-rater measurements are clearly not independent.
- We have to test (1) the agreement significantly exceeds the expected one by chance, or (2) the agreement significantly worse than the expected one by chance.
  - (1) can be done by Kappa-statistics
  - (2) can be done by McNemar's test

# Kappa-statistics and McNemar's test

- Kappa statistics

- Please assume the clinical test repeated 2 times, summarized as 2 by 2 cross table.
- The agreement probability  $P_o$  is  $(a+d)/(a+b+c+d)$ .
- If the agreement of the 2 test is perfect,  $b=c=0$  ( $P_o=1$ ). When the tests completely disagree,  $a=d=0$  ( $P_o=0$ ).
- If the agreement is completely by chance, expected agreement probability  $P_e$  is  $\{(a+c)(a+d)/(a+b+c+d)+(b+d)(c+d)/(a+b+c+d)\}$
- Kappa statistics can be defined as  $(P_o-P_e)/(1-P_e)$
- `library(fmsb)`  
`Kappa.test(matrix(c(12, 2, 4, 10), 2, 2))`
- In EZR, [Statistical analysis]>[Accuracy of diagnostic test]>[Kappa statistics for agreement of two tests]

Test	Retest	
	Positive	Negative
Positive	a (=12)	b (=4)
Negative	c (=2)	d (=10)

```
> .Table
      Test2 (+) Test2 (-)
Test1 (+)      12         4
Test1 (-)       2        10

> res <- NULL

> res <- epi.kappa(.Table, conf.

> colnames(res$kappa) <- gettext

> res[1]
$kappa
      est      lower      upper
1 0.5714286 0.2674605 0.8753967
```

- McNemar's test

- Evaluate the significant change of binary variable (pos/neg) between before/after intervention
- The result is still 2 by 2 cross table.
- $X^2_0 = (b-c)^2/(b+c)$ , obeys chi-sq dist with d.f.=1
- `mcnemar.test(matrix(c(a, c, b, d), 2, 2))`
- By EZR, from raw data, see right.
- Extended version is Bhapker's test (It's available as `bhapker()` in `irr` package).

Statistical analysis | Graphs and tables | Tools | Help | Original menu

- Discrete variables
  - Frequency distributions
  - Confidence interval for a proportion
  - One sample proportion test
  - Confidence interval for a difference between two proportions
  - Confidence interval for a ratio of two proportions
- Enter and analyze two-way table
  - Create two-way table and compare two proportions (Fisher's exact test)
  - Compare proportions of two paired samples (McNemar test)
  - Compare proportions of more than two paired samples (Cochran Q test)

pos=17)