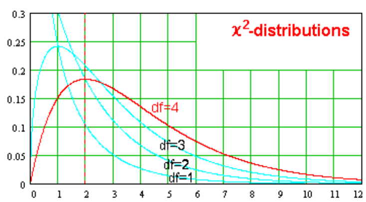
# 2 sample $\chi^2$




## 2 sample $\chi^2$

- This is an extension of the one sample case, we have a somewhat more complicated formula
- We are testing to see if the data we have which is cross classified by 2 variables is consistent with what we would expect if the 2 variables are independent (unrelated to each other)

**Null hypothesis**
Variable A is independent of variable B

**Alternative hypothesis**
Variable A is not independent of variable B

*Tip*
Any numerical variable may be transformed into categories. For example, *Salary* could be divided into 3 groups:
*Under 25K*
*25K to 50K*
*50K and Over*.

*Caution*
Both variables must be categorical. For example, variable A might be a binary variable *Male/Female* with 2 categories and variable B might represent classification *Hourly/Administrative/Executive* with 3 categories.

- For example, we might be interested in how land type is distributed in 2 locations

| Land use classes | | | | | |
|---|---|---|---|---|---|
| | arable | pasture | forest | moorland | total |
| Upland | 5 | 9 | 10 | 15 | 39 |
| Valley | 11 | 11 | 5 | 5 | 32 |
| Total | 16 | 20 | 15 | 20 | 71 |

- we can follow the same steps as before
- 1) $H_0$: no significant differences between upland and valley areas in terms of land use type
- 2) set the level of significance: α=.05
- 3) selection of test statistic: we have frequency counts so chi square is appropriate

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

r rows
k columns in contingency table whose members are $O_{ij}$s

$$E_{ij} = \frac{\sum_{i=1}^{r} O_i \sum_{j=1}^{k} O_j}{\sum_{i=1}^{r}\sum_{j=1}^{k} O_{ij}}$$

- row total times column total / overall total
- $E_{ij}$s
- (39 x 16) /71 = 8.8
- (39 x 20)/71 = 11.0 2 cells with this
- (39 x 15)/71 = 8.2
- (32 x 16)/71= 7.2
- (32 x 20)/71=9.0 2 cells with this
- (32 x 15)/71=6.8

- Expected values

| | | | | |
|---|---|---|---|---|
| upland | 8.8 | 11.0 | 8.2 | 11.0 |
| valley | 7.2 | 9.0 | 6.8 | 9.0 |

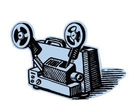
df=(2-1)(4-1)=(1)(3)=3 (r-1)(k-1)

4) compute the test statistic

| | | | | |
|---|---|---|---|---|
| upland | $(5-8.8)^2/8.8=1.64$ | $(9-11)^2/11=.36$ | $(10-8.2)^2/8.2=.40$ | $(15-11)^2/11=1.45$ |
| valley | $(11-7.2)^2/7.2=2.0$ | $(11-9)^2/9=.44$ | $(5-6.8)^2/6.8=.48$ | $(5-9)^2/9=1.78$ |

- $\chi^2$=1.64+.36+.40+1.45+2.0+.44+.48+1.78=8.55
- 5) determine rejection value
- $\chi^2_3$=7.82, α=0.05
- 6) the calculated $\chi^2$ exceeds the critical value
- 7) we reject $H_0$

- 2 way table video
- To see the video online go to
  - http://www.learner.org/resources/series65.html
  - Title: Inference for Two-Way Tables



## special case of $\chi^2$ (2 x 2 table)

- for a 2 x 2 table $\chi^2$ is calculated by

$$\chi^2 = \frac{n(|AD - BC| - n/2)^2}{(A+B)(C+D)(A+C)(B+D)}$$

df=(number of rows-1)(number of columns-1)=1
n = total number of individuals

## Example

| | |
|---|---|
| A | B |
| C | D |

| | |
|---|---|
| 15 A | 11 B |
| 5 C | 12 D |

$$\chi^2 = \frac{43(|180-55| - 43/2)^2}{(26)(17)(20)(23)} = \frac{43(125-21.5)^2}{203320} = 2.26$$

- $\alpha$=0.05, df=1, $\chi^2_c$=3.84
- we accept $H_0$

## Combining cells

- remember the test requires expected cell frequency to be at least 5, if less than 5 the usual practice is to combine cells
- 2 basic options
- combine each end column with the one next to it - only do this if it makes sense, ie. Combining strongly disagree with somewhat disagree
- combine 2 end columns to get a single column - combining the extreme positions on a question

- suspiciously small
  - values close to zero are suspicious
  - so consider whether data has been doctored or small calculation error has occurred
- the additive property of $\chi^2$
  - sometimes an experiment is performed more than once
  - its okay to sum the values of $\chi^2$ obtained from each performance of the experiment, to sum the df and then test the significance

## statistics available in SPSS

- Likelihood ratio chi-square is an alternative to test the hypothesis of no association of columns and rows in nominal-level tabular data.
  - based on maximum likelihood estimation. Though computed differently, likelihood ratio chi-square is interpreted the same way.
- Linear by Linear Association chi-square is a version of chi-square for ordinal data. It assumes equally ordered (interval or near interval) data.

## Mantel-Haenszel chi-square

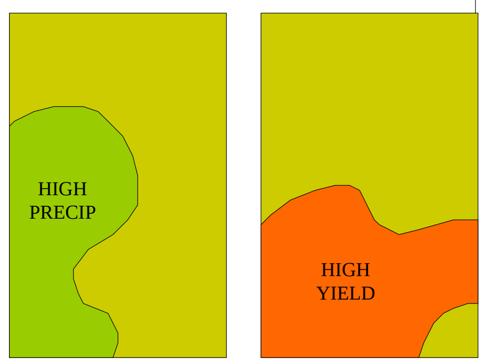
- also called the *Mantel-Haenszel test for linear association*, unlike ordinary and likelihood ratio chi-square, is an ordinal measure of significance. It is preferred when testing the significance of linear relationship between two ordinal variables. If found significant, the interpretation is that increases in one variable are associated with increases (or decreases for negative relationships) in the other greater than would be expected by chance of random sampling.
- Like other chi-square statistics, M-H chi-square should not be used with tables with small cell counts.

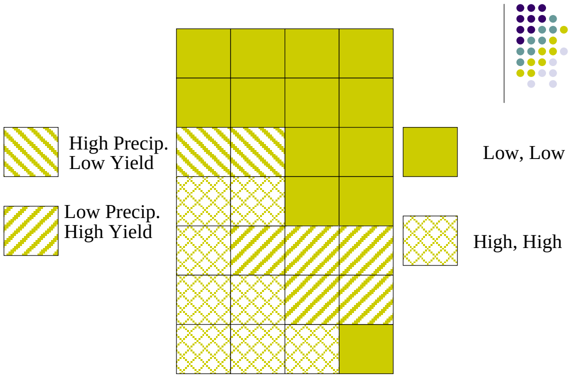
## Fisher Exact Test of Significance

- The Fisher exact test of significance is used in place of the chi-square test in small 2-by-2 tables. It tests the probability of getting a table as strong as the observed or stronger simply due to the chance of sampling, where "strong" is defined by the proportion of cases on the diagonal with the most cases.

## Chi-Square Statistic

- Measures the strength of association between two distributions
- Class Example
  - Relationship between wheat yield and precipitation
  - Two maps showing high and low yields and high and low precipitation




## Chi-Square

- By combining distribution on one map we can better understand the relationship between the two distributions
- In this example we are using a grid
  - The finer the grid, the more precise the measurement
- Four possibilities exist
  - Low rainfall, low yield
  - Low rainfall, high yield
  - High rainfall, low yield
  - High rainfall, high yield

## Chi-Square

- Record the total number of occurrences into a table of observed frequencies

| | | WHEAT | |
|---|---|---|---|
| | | High | Low |
| PRECIP. | High | 8 | 2 |
| | Low | 5 | 13 |

## Compute Chi-Square

- Therefore, in our example we have

Observed

| | High | Low |
|---|---|---|
| High | 8 | 2 |
| Low | 5 | 13 |

$X^2$=6.2

Expected

| | High | Low |
|---|---|---|
| High | 5 | 5 |
| Low | 8 | 10 |

**Contingency Table: Census Division of Respondent Cross-Tabulated with Attitude toward Country Western Music, General Social Survey, 1995**

| | Attitude toward country western music* | | | | | |
|---|---|---|---|---|---|---|
| Census division | Like very much | Like it | Mixed feelings | Dislike it | Dislike very much | Row total |
| New England | 5 (7.8) | 13 (10.6) | 8 (6.8) | 3 (2.7) | 0 (1.1) | 29 (29) |
| Middle Atlantic | 21 (28.1) | 30 (37.9) | 39 (24.3) | 9 (9.9) | 5 (3.8) | 104 (104) |
| E. North Central | 41 (45.2) | 60 (60.8) | 40 (39.1) | 17 (15.8) | 9 (6.1) | 167 (167) |
| W. North Central | 8 (13.0) | 23 (17.5) | 11 (11.2) | 4 (4.5) | 2 (1.8) | 48 (48) |
| South Atlantic | 36 (32.5) | 48 (43.7) | 22 (28.1) | 13 (11.4) | 1 (4.4) | 120 (120) |
| E. South Central | 26 (14.3) | 15 (19.3) | 5 (12.3) | 5 (5.0) | 2 (1.9) | 53 (53) |
| W. South Central | 27 (18.7) | 24 (25.1) | 10 (16.2) | 7 (6.5) | 1 (2.3) | 69 (69) |
| Mountain | 8 (9.5) | 16 (12.7) | 6 (8.2) | 3 (3.3) | 2 (1.3) | 35 (35) |
| Pacific | 28 (30.9) | 40 (41.5) | 32 (26.7) | 9 (10.8) | 5 (4.2) | 114 (114) |
| Column total | 200 (200) | 269 (269) | 173 (173) | 70 (70) | 27 (27) | 739 (739) |

All $E_{ij} = \frac{(R_i)(C_j)}{N}$

For example: $E_{11} = \frac{(R_1)(C_1)}{N} = \frac{(29)(200)}{739} = 7.8$

*Each cell of the table contains the observed frequency count, followed by the expected frequency count, in parentheses

**Worktable for Chi-square Contingency Analysis: Attitude toward Country Western Music by Census Division**

$H_0$: there is no relationship between two variables (variables are statistically independent with only a random association).

$H_A$: there is a relationship between two variables (variables are not statisically independent, but related to one another in some nonrandom fashion.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $O_{ij}$ = the observed frequency count in the $i^{th}$ row and $j^{th}$ column
$E_{ij}$ = the expected frequency count in the $i^{th}$ row and $j^{th}$ column
$r$ = the number of rows in the contingency table
$k$ = the number of columns in the contingency table

The row variable is census division and the column variable is preference of country western music.

$$E_{ij} = \frac{(R_i)(C_j)}{N}$$

$$E_{11} = \frac{(R_1)(C_1)}{N} = \frac{(29)(200)}{739} = 7.8$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(5-7.8)^2}{7.8} + \frac{(13-10.6)^2}{10.6} + \frac{(8-6.8)^2}{6.8} + \dots + \frac{(9-10.8)^2}{10.8} + \frac{(5-4.2)^2}{4.2}$$

$\chi^2 = 52.088$ $p$-value = .014