# Statistical Inference for Frequency Data

Sunthud Pornprasertmanit Chulalongkorn University

When the variables are all categorical, the statistical analysis will be changed. The research question may be correlation or comparison.

In this lecture, you have learned some statistical tests, binomial test, proportion z test, phi coefficient, and McNemar test. However, some statistical tests are new for you, chi-square goodness-of-fit test, chi-square test for proportion, Cochran Q test, Cramer's V, Cohen's Kappa, and Gamma statistics.

## Binomial Test

The binomial test is similar to one-sample *t* test. This test is used for dichotomous variable. The criterion value is not population mean, like one-sample *t* test, but population proportion instead.

| | | |
|---|---|---|
| Null hypothesis | $H_0: p = p_0$ | |
| Alternative hypothesis | $H_1: p \neq p_0$ | (Two-tailed) |
| | $H_1: p > p_0;\ p < p_0$ | (One-tailed) |

The test statistic is based on

1) Exact probability
2) Approximate probability

### Exact Probability

First, specify null hypothetical proportion ($p_0$)

Second, specify sample proportion ($\hat{p}$).

Third, specify sample size (*n*)

Fourth, calculate for number of successes (*r*).

$$r = \hat{p}n$$

Fifth, calculate for *p* value

Formula for calculation

$$p(X = r) = C_r^n p^r q^{n-r}$$

Author Note
This article was written in September 2007 for teaching in Introduction to Statistics in Psychology Class, Faculty of Psychology, Chulalongkorn University
Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com

For example, Flipping the coin 16 times found 2 heads. Is this coin bias if $p(H) = .5$?

$$\begin{aligned} p(X \le 2) &= p(X = 0) + p(X = 1) + p(X = 2) \\ &= C_0^{16} p^0 q^{16-0} + C_1^{16} p^1 q^{16-1} + C_2^{16} p^2 q^{16-2} \\ &= (1)(0.5)^0(0.5)^{16} + (16)(0.5)^1(0.5)^{15} + (120)(0.5)^2(0.5)^{14} \\ &= (1)(0.5)^{16} + (16)(0.5)^{16} + (120)(0.5)^{16} = 137(0.5)^{16} = 0.0021 \end{aligned}$$

MS Excel Function

=BINOMDIST($r$, $n$, $p_0$, cumulative [true/false])

For example, Flipping the coin 16 times found 2 heads. Is this coin bias if $p(H) = .5$?

$p(X \le 2)$ = BINOMDIST(2, 16, 0.5, true) = 0.0021

$p(X = 2)$ = BINOMDIST(2, 16, 0.5, false) = 0.0018

Sixth, specify that the test is one-tailed or two-tailed

Seventh, if the test is two-tailed, the *p* value is multiplied by 2. Otherwise, leave that *p* value.

Eighth, decide whether the p value is less than alpha. Then, decide that the test reject null hypothesis or fail to reject null hypothesis.

## Approximate Probability

The normal distribution can estimate the probability of binomial distribution

$$z = \frac{\hat{p} - E(\hat{p})}{\sigma_p} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The confidence interval based on this null hypothesis testing is

The general form of a two side 100(1 - $\alpha$) % confidence interval for $\mu$ is

$$\text{Prob}\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = .95$$

The general form of a one side 100(1 - $\alpha$) % confidence interval for $\mu$ is

$$\text{Prob}\left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p\right) = .95 \quad \text{or} \quad \text{Prob}\left(p < \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = .95$$

Example

In approximating area by normal distribution in binomial test, it assumes that

1) Random sampling from the population of interest
2) Binomial Population
3) $np_0$ and $n(1-p_0)$ are both greater than 15
4) The population is at least 10 times larger than the sample.

If either third or fourth assumption is violated, the exact probability is preferable.

### Practical Significance

Cohen (1988) presented the effect size measure for binomial test

$$\widehat{w} = \sqrt{\frac{(\hat{p}-p_0)^2}{p_0}}$$

0.1 is a small effect.
0.3 is a medium effect.
0.5 is a large effect.

## Chi-square Goodness-of-fit Test

The chi-square goodness-of-fit test is similar to binomial test. This test is used for polytomous variable.

Null hypothesis $H_0: \hat{p}_1 = p_1'; \hat{p}_2 = p_2'; \ldots; \hat{p}_k = p_k'$

Alternative hypothesis $H_1: \hat{p}_j \neq p_j'$ for one or more categories

The observed value (O) is the observed frequency within each category.

$$O_j = n\hat{p}_j$$

The expected value (E) is the frequency within each category if the null hypothesis is true.

$$E_j = n\hat{p}_j$$

Therefore, the null and alternative hypotheses will change to

Null hypothesis $H_0: O_1 = E_1; O_2 = E_2; \ldots; O_k = E_k$

Alternative hypothesis $H_1: O_j = E_j$ for one or more categories

| | Category 1 | Category 2 | Category 3 | Total |
|---|---|---|---|---|
| **Observed** | $O_1$ | $O_2$ | $O_3$ | ***n*** |
| **Expected** | $\boldsymbol{E_1}$ | $\boldsymbol{E_2}$ | $\boldsymbol{E_3}$ | ***n*** |

For example, the market share of soaps in last year is

Soap 1 = 70 %; Soap 2 = 20 %; Soap 3 = 10 %

In this year, if the market share is not changed and the researcher collected the consumer preference from 1000 people, the expected value will be

Soap 1 = 700; Soap 2 = 200; Soap 3 = 100

However, the real data is

Soap 1 = 720; Soap 2 = 220; Soap 3 = 60

Is the difference between real data and expected data the real difference in population or sampling error?

| | Soap 1 | Soap 2 | Soap 3 | Total |
|---|---|---|---|---|
| **Observed** | 720 | 220 | 60 | **1000** |
| **Expected** | **700** | **200** | **100** | **1000** |

The statistic that measures the deviation between the observed frequency and expected frequency is chi-square.

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

In this example, the chi-square is 18.57.

When there is large discrepancy between observed and expected value, the chi-square statistic is large, and then, it leads us to reject null hypothesis.

How large the chi-square statistic that is unlikely to be sampled from null population?

The chi-square distribution with *df* = *k* -1 will determine the probability of this statistic will be sampled from null population (*p* value).

The MS Excel calculation is

= CHIDIST(*x*, *df*)

For example,

= CHIDIST(18.57, 2) = 0.0000927

If the *p* value is less than desired alpha, reject null hypothesis.

If the *p* value is larger than desired alpha, fail to reject null hypothesis.

In this example, the null hypothesis is rejected, if alpha = .05.

Example

Assumption of this null hypothesis testing

1) Mutually exclusive and exhaustive
2) The observations must be independent.
3) The expected value should be at least 5 in each cell when the degrees of freedom equal 1. However, when the degrees of freedom is larger than 1, the expected value in 80 % of cells should be larger than 5. (Siegel & Castellan, 1988)

### Practical Significance

Cohen (1988) presented the effect size measure for chi-square test

$$\widehat{w} = \sqrt{\frac{\chi^2}{n}}$$

0.1 is a small effect.
0.3 is a medium effect.
0.5 is a large effect.

## Testing Difference for Proportions Using Two Independent Samples

### Dichotomous Variable

The proportion z test is similar to independent t test but the dependent variable is dichotomous. The comparing sample statistic is the proportions of one category rather than sample means.

| | | |
|---|---|---|
| Null hypothesis | $H_0: p_1 = p_2$ | |
| Alternative hypothesis | $H_1: p_1 \neq p_2$ | (Two-tailed) |
| | $H_1: p_1 > p_2; p_1 < p_2$ | (One-tailed) |

The formula for the proportion *z* test is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{Pooled}(1 - \hat{p}_{Pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\hat{p}_{Pooled} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

The confidence interval based on this null hypothesis testing is

A two-sided 100(1-α) % confidence interval for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Lower and upper one-sided 100(1-α) % confidence intervals for $p_1 - p_2$are given by

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2$$

and $$p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example

Assumption of this null hypothesis testing and confidence interval

1) Mutually exclusive and exhaustive
2) The observations must be independent.
3) All the products $n_1\hat{p}_1, n_1(1-\hat{p}_1), n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ are greater than 5 and both populations are at least 10 times larger than their respective samples.

When there is not enough sample size, the Fisher's Exact Test can be used (see from Siegel & Castellan, 1988).

## Practical Significance

Cohen (1988) presented the effect size measure for proportion *z* test

$$\widehat{w} = \sqrt{\frac{z^2}{n}}$$

0.1 is a small effect.
0.3 is a medium effect.
0.5 is a large effect.

## Polytomous Variable

The chi-square testing will be used when the comparing variable is polytomous.

Null hypothesis $$H_0: \begin{bmatrix} P(r_1|c_1) = P(r_1|c_2) \\ P(r_2|c_1) = P(r_2|c_2) \\ \vdots \\ P(r_i|c_1) = P(r_i|c_2) \end{bmatrix}$$

Therefore, the group and categories is statistical independent.

$$P(r_1|c_1) = P(r_1|c_2) = P(r_1)$$

Alternative hypothesis: $H_1: p(r_i|c_1) \neq P(r_i|c_2)$ for one or more category

Therefore, the group and categories is not statistical independent.

$$P(r_1|c_1) \neq P(r_1|c_2) \neq P(r_1)$$

| | Group 1 | Group 2 | Total |
|---|---|---|---|
| Category 1 | $O_{11}$ | $O_{11}$ | $R_1$ |
| Category 2 | $O_{11}$ | $O_{11}$ | $R_2$ |
| Category 3 | $O_{11}$ | $O_{11}$ | $R_3$ |
| Total | $C_1$ | $C_2$ | $n$ |

The observed value of each cell is (assumed that group and categories is not statistical independent)

$$O_{ij} = np(R_i|C_j)p(C_j) \quad or \quad np(C_j|R_i)p(R_i)$$

If null hypothesis is true, the expected value from each cell is

$$E_{ij} = np(R_i)p(C_j) = n\frac{R_i}{n}\frac{C_j}{n} = \frac{R_i C_j}{n}$$

The chi-square statistic testing for independence between group is

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

This statistics is distributed in chi-square distribution with $df = r - 1$

Example

Assumption of this null hypothesis testing

1) Mutually exclusive and exhaustive
2) The observations must be independent.
3) The expected value should be at least 5 in each cell when the degrees of freedom equal 1. However, when the degrees of freedom is larger than 1, the expected value in 80 % of cells should be larger than 5. (Siegel & Castellan, 1988)

### Practical Significance

Cohen (1988) presented the effect size measure for chi-square test

$$\widehat{w} = \sqrt{\frac{\chi^2}{n}}$$

0.1 is a small effect.
0.3 is a medium effect.
0.5 is a large effect.

## Testing Difference for Proportions Two or More Independent Samples

For comparing proportions between two or more groups, the chi-square test will be used when the comparing variable is dichotomous or polytomous.

Null hypothesis $H_0$: $\begin{bmatrix} P(r_1|c_1) = P(r_1|c_2) = \cdots = P(r_1|c_j) \\ P(r_2|c_1) = P(r_2|c_2) = \cdots = P(r_2|c_j) \\ \vdots \\ P(r_i|c_1) = P(r_i|c_2) = \cdots = P(r_i|c_j) \end{bmatrix}$

Alternative hypothesis:

$H_1$: $p(r_i|c_j); P(r_i|c_j); \cdots; P(r_1|c_j)$ differs at least one pair for one or more category

The expected value can be calculated like testing difference between two groups.

The chi-square statistic testing for independence between groups is

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

This statistics is distributed in chi-square distribution with $df = (r-1)(c-1)$

If rejecting null hypothesis, which pairs are statistically different. If the comparing variable is dichotomous, the multiple comparisons can be used. (Marascuilo & McSweeney, 1977)

Example

Assumption of this null hypothesis testing

1) Mutually exclusive and exhaustive
2) The observations must be independent.
3) The expected value should be at least 5 in each cell when the degrees of freedom equal 1. However, when the degrees of freedom is larger than 1, the expected value in 80 % of cells should be larger than 5. (Siegel & Castellan, 1988)

### Practical Significance

Cohen (1988) presented the effect size measure for chi-square test

$$\hat{w} = \sqrt{\frac{\chi^2}{n}}$$

0.1 is a small effect.
0.3 is a medium effect.
0.5 is a large effect.

## Testing Difference for Proportions Using Two Dependent Samples

When researchers want to compare proportion of desired variable between two dependent groups, the McNemar test (adapted binomial test) is suitable.

| | | |
|---|---|---|
| Null hypothesis | $H_0: p_1 = p_2$ | |
| Alternative hypothesis | $H_1: p_1 \neq p_2$ | (Two-tailed) |
| | $H_1: p_1 > p_2; p_1 < p_2$ | (One-tailed) |

To test one of these hypotheses, the data are placed into a 2 × 2 table as follows:

| | | *Sample 2* | | |
|---|---|---|---|---|
| | | Success | Failure | |
| *Sample 1* | Success | $X_{11}$ | $X_{12}$ | $X_{1.}$ |
| | Failure | $X_{21}$ | $X_{22}$ | $X_{2.}$ |
| | | $X_{.1}$ | $X_{.2}$ | $n$ |

You will see that

$$H_0: p_{1.} = p_{.1}$$

$$H_0: X_{12} = X_{21}$$

$X_{12}$ is the change from success in sample 1 to failure in sample 2 and $X_{21}$ is the change from failure in sample 1 to success in sample 2.

If null hypothesis is true, the number of change from success to failure is the same as the number of change from failure to success.

These two frequencies are distributed in binomial distribution and if the sample size is large, the probability can be approximate by z test statistic.

$$z = \frac{X_{12} - X_{21}}{\sqrt{X_{12} + X_{21}}}$$

The confidence interval based on this null hypothesis testing is

A two-sided 100(1-$\alpha$) % confidence interval for $p_1 - p_2$ is given by

$$\frac{X_{12} - X_{21}}{n} - z_{\alpha/2}\sqrt{\frac{(X_{12} + X_{21})(X_{11} + X_{22}) + 4X_{12}X_{21}}{n^3}} < p_1 - p_2 < \frac{X_{12} - X_{21}}{n} + z_{\alpha/2}\sqrt{\frac{(X_{12} + X_{21})(X_{11} + X_{22}) + 4X_{12}X_{21}}{n^3}}$$

Lower and upper one-sided 100(1-$\alpha$) % confidence intervals for $p_1 - p_2$are given by

$$\frac{X_{12} - X_{21}}{n} - z_{\alpha}\sqrt{\frac{(X_{12} + X_{21})(X_{11} + X_{22}) + 4X_{12}X_{21}}{n^3}} < p_1 - p_2$$

and $$p_1 - p_2 < \frac{X_{12} - X_{21}}{n} + z_{\alpha}\sqrt{\frac{(X_{12} + X_{21})(X_{11} + X_{22}) + 4X_{12}X_{21}}{n^3}}$$

Example

Assumption of this null hypothesis testing and confidence interval

1) The samples are dependent.
2) $X_{12} + X_{21} \geq 10$ for two-tailed test and $X_{12} + X_{21} \geq 30$ for one-tailed test

When there is not enough sample size, the exact probability can be used by calculate from binomial distribution.

## Testing Difference for Proportions Using Two or More Dependent Samples

When researchers want to compare proportion of desired dichotomous variable between three or more dependent groups, the Cochran *Q* test is suitable.

Null hypothesis $H_0: p_1 = p_2 = \cdots = p_j$

Alternative hypothesis $H_1: p_i \neq p_j$ for one or more pairs

| | Group 1 | Group 2 | Group 3 | Group 4 | Total |
|---|---|---|---|---|---|
| **Case 1** | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $T_{1.}$ |
| **Case 2** | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{24}$ | $T_{2.}$ |
| **Case 3** | $X_{31}$ | $X_{32}$ | $X_{33}$ | $X_{34}$ | $T_{3.}$ |
| **Case 4** | $X_{41}$ | $X_{42}$ | $X_{43}$ | $X_{44}$ | $T_{4.}$ |
| **Case 5** | $X_{51}$ | $X_{52}$ | $X_{53}$ | $X_{54}$ | $T_{5.}$ |
| **Total** | $T_{.1}$ | $T_{.2}$ | $T_{.3}$ | $T_{.4}$ | $T_{..}$ |

*T* stands for total of each row and each column.

If $H_0: p_{.1} = p_{.2} = \cdots = p_{.j}$ , then $H_0: T_{.1} = T_{.2} = \cdots = T_{.j}$ .

The statistics that test

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} T_{.j}^2 - \left(\sum_{j=1}^{k} T_{.j}\right)^2\right]}{k\sum_{i=1}^{n} T_{i.} - \sum_{i=1}^{n} T_{i.}^2}$$

The *Q* statistic is distributed in chi-square distribution with $df = k - 1$.

If the null hypothesis is rejected, the next question is which pair is statistically different. The multiple comparison technique is shown in Marascuilo & McSweeney (1977).

Example

Assumption for this null hypothesis

1) The samples are dependent.
2) There are 24 or more informative observations (24 cells) distributed over $n \geq 4$ rows.

# Testing for Association between Frequency Data

## Phi Coefficient

The phi coefficient ($r_\phi$) is a measure of the extent of association or relation between two sets of attributes measured on nominal scale, each of which may take on only two values.

Null hypothesis $H_0: \rho_\phi = 0$

Alternative hypothesis $H_1: \rho_\phi \neq 0$

| | | Second Variable | | Total |
|---|---|---|---|---|
| | | Group 1 | Group 2 | |
| First Variable | Group 1 | $X_{11}$ | $X_{12}$ | $R_1$ |
| | Group 2 | $X_{21}$ | $X_{22}$ | $R_2$ |
| Total | | $C_1$ | $C_2$ | $n$ |

The phi coefficient formula is

$$r_\phi = \frac{|X_{11}X_{22} - X_{12}X_{21}|}{\sqrt{R_1 R_2 C_1 C_2}}$$

For testing null hypothesis, the proportion *z* test can test whether both variables are statistical independent or not.

The relationship between phi coefficient and proportion *z* test is

$$r_\phi = \sqrt{\frac{z^2}{n}}$$

Assumption of phi coefficient is the same as proportion z test.

If sample size is small, the significance of phi coefficient can be tested with the Fisher exact test.

## Cramer's V

The Cramer's V statistic is suitable for measuring association between two sets of unordered qualitative variables.

Null hypothesis $H_0: V = 0$

Alternative hypothesis $H_1: V \neq 0$

The Cramer's V formula is

$$\hat{V} = \sqrt{\frac{\chi^2}{n(s-1)}}$$

where *s* is the smaller of the number of rows and columns.

The Cramer's V can range from 0 (indicating complete independence) to 1 (indicating complete dependence)

If *s* = 2, the Cramer's V is equal to phi coefficient.

For testing null hypothesis, the chi-square test can test whether both variables are statistical independent or not.

Assumption of phi coefficient is the same as chi-square test.

Limitations of the Cramer's *V*

1) When *V* is equal to 1 and number of categories is not equal between two variables, there may not be perfect correlation between the variables. (Asymmetric perfect relation)

   For example, the case that *V* equals 1

| | | Attitude toward Abortion | | |
|---|---|---|---|---|
| | | Agree | Neutral | Disagree |
| **Sex** | **Male** | 0 | 25 | 25 |
| | **Female** | 50 | 0 | 0 |

   Attitude toward Abortion predicted Sex (Accuracy = 100 %).

   Sex predicted Attitude toward Abortion (Accuracy < 100 %).

2) The chi-square test of independence assumes that the expected values are large.
3) It is not directly comparable to any other measure of correlation, such as the Pearson r.
4) The Cramer's *V* could not be employed to assess the degree of monotonic association between two ordered variables.
5) It should be cautioned against such an interpretation of *V* or $V^2$.

## Kappa Statistic

When observers rate the objected that may not be ordered but simply assigned into categories, the Kappa (K) statistic measures the degree that observers rating in the same categories.

The kappa coefficient of agreement is the ratio of the proportion of times that rater agree (corrected for chance agreement) to the maximum proportion of times that the raters could agree (corrected for chance agreement).

$$K = \frac{\hat{p}(A) - \hat{p}(E)}{1 - \hat{p}(E)}$$

$\hat{p}(A)$ is the proportion of times that the k raters agree.

$\hat{p}(E)$ is the proportion of times that we would expect the k eaters to agree by chance.

### Two Raters and Two Categories

When there are two raters rating *n* objects into 2 categories,

| | | Rater 2 | | |
|---|---|---|---|---|
| | | Category 1 | Category 2 | |
| Rater 1 | Category 1 | $X_{11}$ | $X_{12}$ | $X_{1.}$ |
| | Category 2 | $X_{21}$ | $X_{22}$ | $X_{2.}$ |
| | | $X_{.1}$ | $X_{.2}$ | $n$ |

$$\hat{p}(A) = \frac{X_{11} + X_{22}}{n}$$

$$\hat{p}(E) = \sum_{j=1}^{2} p_{j.}p_{.j} = p_{1.}p_{.1} + p_{2.}p_{.2} = \left(\frac{X_{1.}}{n}\right)\left(\frac{X_{.1}}{n}\right) + \left(\frac{X_{2.}}{n}\right)\left(\frac{X_{.2}}{n}\right) = \frac{1}{n^2}(X_{1.}X_{.1} + X_{2.}X_{.2})$$

Example

## Two Raters and More than Two Categories

When there are two raters rating *n* objects into *m* categories,

| | | Rater 2 | | | |
|---|---|---|---|---|---|
| | | Category 1 | Category 2 | Category 3 | |
| Rater 1 | Category 1 | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{1.}$ |
| | Category 2 | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{2.}$ |
| | Category 3 | $X_{31}$ | $X_{32}$ | $X_{33}$ | $X_{3.}$ |
| | | $X_{.1}$ | $X_{.2}$ | $X_{.3}$ | $n$ |

$$\hat{p}(A) = \sum_{i=1}^{m} p_{ii} = \frac{1}{n}\sum_{i=1}^{m} X_{ii}$$

$$\hat{p}(E) = \sum_{j=1}^{m} p_{j.}p_{.j} = \sum_{j=1}^{m} \left(\frac{X_{j.}}{n}\right)\left(\frac{X_{.j}}{n}\right) = \frac{1}{n^2}\sum_{j=1}^{m} X_{j.}X_{.j}$$

Example

## More than two raters

When there are *k* raters rating *n* objects into *m* categories,

| Object | Number of rating | | | | | S |
|---|---|---|---|---|---|---|
| | Category 1 | Category 2 | Category 3 | Category 4 | Category 5 | |
| **1** | - | - | - | - | 4 | 12/12 = 1 |
| **2** | 2 | - | 2 | - | - | 4/12 = 0.333 |
| **3** | - | - | - | - | 4 | 12/12 = 1 |
| **4** | 2 | - | 2 | - | - | 4/12 = 0.333 |
| **5** | - | - | - | 1 | 3 | 6/12 = 0.50 |
| **6** | 1 | 1 | 2 | - | - | 2/12 = 0.167 |
| **7** | 3 | - | 1 | - | - | 6/12 = 0.50 |
| **8** | 3 | - | 1 | - | - | 6/12 = 0.50 |
| **9** | - | - | - | - | 4 | 4/12 = 0.333 |
| **10** | 4 | - | - | - | - | 6/12 = 0.50 |
| $C_j$ | 15 | 1 | 8 | 1 | 15 | |
| $p_j$ | 0.375 | 0.025 | 0.20 | 0.025 | 0.375 | |

The sum of each row equals the number of raters.

The proportion of objects assigned to the jth category is

$$p_j = \frac{C_j}{nk}$$

If the raters make their assignment at random, the expected proportion of agreement for each category would be $p_j^2$, and total expected agreement across all categories would be

$$\hat{p}(E) = \sum_{j=1}^{m} p_j^2$$

The extent of agreement among the raters concerning the *i*th subject is the proportion of the number of pairs for which there is agreement to the possible pairs of assignments. For the *i*th subject this is

$$S = \frac{\sum_{j=1}^{m} \binom{n_{ij}}{2}}{\binom{k}{2}} = \frac{1}{k(k-1)} \sum_{j=1}^{m} n_{ij}(n_{ij}-1)$$

To obtain the total proportion of agreement, we find the average of these proportions across all objects rated

$$\hat{p}(A) = \frac{1}{n} \sum_{i=1}^{n} S_i$$

Example

## Statistical Inference

Null hypothesis: $H_0: \kappa = 0$

The degree of rater agreement is less than or equal to chance.

Alternative hypothesis: $H_1: \kappa > 0$

The degree of rater agreement is larger than chance. (Interested only one-tailed test)

The kappa statistic is distributed in normal distribution with variance

$$\hat{\sigma}_K^2 \approx \frac{2}{Nk(k-1)} \frac{P(E) - (2k-3)[P(E)]^2 + 2(k-2)\sum p_j^3}{[1-P(E)]^2}$$

Then, the *z* statistic is

$$z = \frac{K}{\sqrt{\frac{2}{Nk(k-1)} \frac{P(E) - (2k-3)[P(E)]^2 + 2(k-2)\sum p_j^3}{[1-P(E)]^2}}}$$

Confidence interval based on this null hypothesis testing is

Lower and upper one-sided 100(1-α) % confidence intervals for $\kappa$ are given by

$$\kappa > K - z_\alpha \sqrt{\frac{2}{Nk(k-1)} \frac{P(E) - (2k-3)[P(E)]^2 + 2(k-2)\sum p_j^3}{[1-P(E)]^2}}$$

If the lower bound is less than 0, change the lower bound to 0 and use close lower limit. (Do not reject null hypothesis that κ = 0.

Example

Assumption

1) Observed must be independently assigned the objects.
2) The number of objects is large (in Fleiss [1971] use *n* = 30)

Fleiss (1971) provided the formula for calculating agreement of particular category.

## Practical Significance

Guideline for interpretation of kappa statistic

| | |
|---|---|
| *K* < .40 | Poor agreement |
| .40 < *K* < .75 | Good agreement |
| *K* > .75 | Excellent agreement |

## Gamma Statistics

Although there are some statistics (Spearman and Kendall rank order correlation) that measure association between two ordered variables, they are less useful and less appropriate when there are many ties or in any situation in which it is meaningful to cast the data in the form of a contingency table.

The gamma statistic *G* is appropriate for measuring the relation between two ordinally scaled variables.

Null hypothesis $H_0: \gamma = 0$

Alternative hypothesis $H_1: \gamma \neq 0$

The pairs of observations from each unit will fill into ordered contingency table.

| | | Var 2 | | | | Total |
|---|---|---|---|---|---|---|
| | | $B_1$ | $B_2$ | $B_3$ | $B_4$ | |
| Var 1 | $A_1$ | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ | $R_1$ |
| | $A_2$ | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ | $R_2$ |
| | $A_3$ | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{34}$ | $R_3$ |
| | $A_4$ | $n_{41}$ | $n_{42}$ | $n_{43}$ | $n_{44}$ | $R_4$ |
| **Total** | | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $n$ |

The gamma statistic is the difference in the probability that within a pair of observations *A* and *B* are in the same order and the probability that with in a pair of observations the *A* and *B* disagree in their ordering, provided there are no ties in the data.

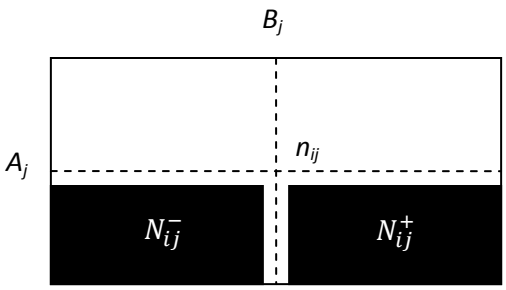
$$G = \frac{P(A \,\&\, B \text{ agree in order}) - P(A \,\&\, B \text{ disagree in order})}{1 - P(A \,\&\, B \text{ are tied})}$$

$$= \frac{P(A \,\&\, B \text{ agree in order}) - P(A \,\&\, B \text{ disagree in order})}{P(A \,\&\, B \text{ agree in order}) + P(A \,\&\, B \text{ disagree in order})}$$

$$= \frac{N(A \,\&\, B \text{ agree in order}) - N(A \,\&\, B \text{ disagree in order})}{N(A \,\&\, B \text{ agree in order}) + N(A \,\&\, B \text{ disagree in order})}$$

$$= \frac{\#(+) - \#(-)}{\#(+) + \#(-)}$$

The #(+) and #(-) can be calculated by




$N_{ij}^+$ is the sum of all of the frequencies below and to the right of the *ij*th cell.

$N_{ij}^+$ is the sum of all of the frequencies below and to the left of the *ij*th cell.

#(+) = The number of agreements in order

$$\#(+) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}N_{ij}^+$$

#(-) = The number of disagreements in order

$$\#(-) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}N_{ij}^-$$

The gamma statistic is equal to 1 if the frequencies in the contingency table are concentrated on the diagonal from the upper left to the lower right of the contingency table and equal to -1 if the frequencies all lie on the diagonal from the upper right corner to the lower left corner of the contingency table.

The G statistic is distributed in normal distribution with variance

$$\sigma_G^2 \leq \frac{N(1-G^2)}{\#(+)-\#(-)}$$

The test statistic is

$$z = (G-\gamma)\sqrt{\frac{\#(+)-\#(-)}{N(1-G^2)}}$$

Then, the test statistic is conservative level. We may infer that the true *p* level is at most the *p* value obtained by this statistic.

The confidence interval based on this null hypothesis testing is

A two-sided 100(1-$\alpha$) % confidence interval for $p_1 - p_2$ is given by

$$G - z_{\alpha/2}\sqrt{\frac{N(1-G^2)}{\#(+)-\#(-)}} < \gamma < G + z_{\alpha/2}\sqrt{\frac{N(1-G^2)}{\#(+)-\#(-)}}$$

Lower and upper one-sided 100(1-$\alpha$) % confidence intervals for $p_1 - p_2$are given by

$$G - z_{\alpha}\sqrt{\frac{N(1-G^2)}{\#(+)-\#(-)}} < \gamma \qquad \text{and} \qquad \gamma < G + z_{\alpha}\sqrt{\frac{N(1-G^2)}{\#(+)-\#(-)}}$$

Assumption of this test statistic and confidence interval

1) Both variables are tied ordered variables.
2) The sample size is large. (In Goodman & Kruskal (1963) use *n* = 300)