Big Data Fundamentals and Applications

Statistical Analysis (VIII)

Correlation Analysis

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

1. Introduction
2. Road Map of Statistical Analysis
3. Hypothesis Testing
4. Type I and Type II Errors
5. Reliability & Validity Analyses
6. Inferential Statistics
7. Test of Normality
8. Differences between Parametric and Nonparametric Statistics
9. Parametric Statistics
10. Nonparametric Statistics
11. Correlation Analysis
12. Question Time

Introduction

Road Map of Statistical Analysis

Hypothesis Testing

Type I and Type II Errors

Reliability & Validity Analyses

Inferential Statistics

Test of Normality

Differences between Parametric and Nonparametric Statistics

Parametric Statistics

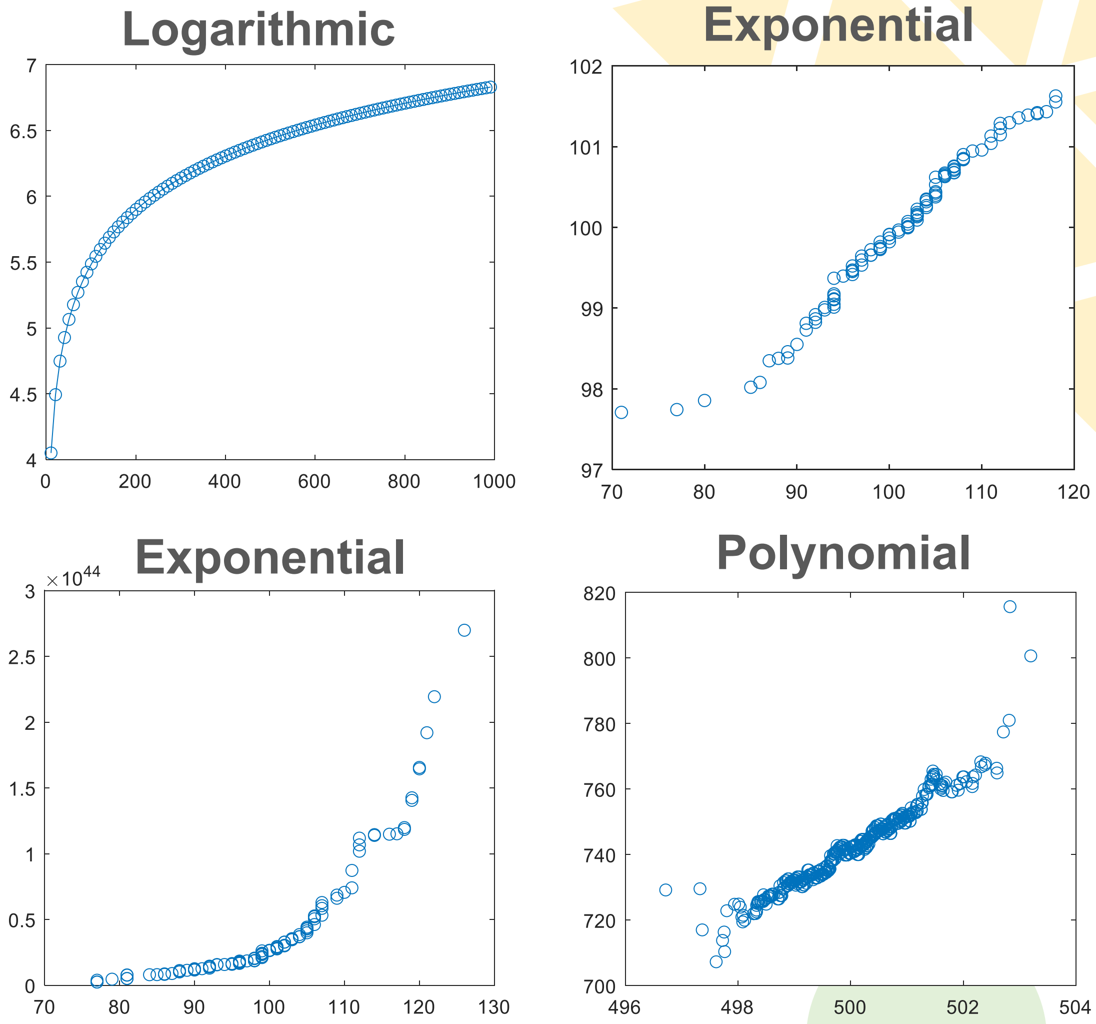
Nonparametric Statistics

Correlation Analysis

Correlation Analysis

Correlation Analysis

- Correlation analysis is an inferential statistics to describe the relationship or association between one variable and other.
- Most formulae for correlation analyses are developed for linear relationship; therefore, other relationships (e.g., logistic, exponential, and cubic) are not suitable. Basically, nonlinear relationship could adopt the performance of curve fitting results.



Logarithmic

Exponential

Exponential

Polynomial

Correlation Analysis

Variable Y/X	Quantitative X	Ordinal X	Nominal X
Quantitative Y	Pearson r	Biserial r_b	Point Biserial r_{pb}
Ordinal Y	Biserial r_b	Spearman ρ/ Tetrachoric r_{tet}	Rank Biserial r_{rb}
Nominal Y	Point Biserial r_{pb}	Rank Biserial r_{rb}	Phi, L, C, V, Lambda

There are two important outcomes from correlation analyses: **significance** and **coefficient**.

1. **Significant** of correlation: the consistency of association between one variable and other.
2. **Coefficient** of correlation: the direction (i.e., positive or negative) and magnitude (i.e., value) of correlation between one variable and other.

Pearson Correlation Coefficient r

- **Pearson correlation coefficient**, also known as Pearson product-moment correlation coefficient (PPMCC), is to measure the linear correlation between two variables or data.
- The definition of Pearson correlation coefficient is calculated by the covariance of the two variables, divided by the product of their standard deviations. Its value ranges from -1 to +1.

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, when\ it\ is\ applied\ for\ population$$

$$r = \frac{cov(X,Y)}{s_X s_Y}, when\ it\ is\ applied\ for\ sample$$

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson Correlation Coefficient r

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, where\ cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

$$then\ \rho = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, and \ldots$$

$$\mu_X = \mathbb{E}[X]; \mu_Y = \mathbb{E}[Y];$$

$$\sigma_X^2 = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2;$$

$$\sigma_Y^2 = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2;$$

$$\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\rho_{XY} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}$$

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson Correlation Coefficient r

- Testing using t-distribution with degrees of freedom $n-2$, where standard error is denoted as,

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}$$

- Therefore, t value is …

$$t = \frac{r}{\sigma_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

The inverse function for determining the critical values for r is …

$$r = \frac{t}{\sqrt{n-2+t^2}}$$

Pearson Correlation Coefficient r

Sleeping/Day, X_i	Relax/Day, Y_i
7.5	1
8	12
9.1	2
6	10
10	5
8.4	6.1
9.1	7
2.4	8.2
6.7	7
6.8	6
9	4.5

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-2.2370}{2.0011 \times 3.0509} = -0.3664$$

$$t = \frac{r}{\sigma_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

$$t = -0.3664 \times \sqrt{\frac{11-2}{1-(-0.3664)^2}} = -1.18143$$

$$t_{-1.18143,9} = 0.133853$$

Biserial r_b

- The **biserial correlation coefficient** is also a correlation coefficient where one of the samples is measured as dichotomous, but where that **sample is really normally distributed**.
- Assuming that we have two sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ where the x_1 are 0 or 1, then the biserial correlation coefficient, denoted r_b, is calculated as follows:

$$r_b = \frac{(m_1 - m_0)p_0p_1}{\sigma_y y}$$

where n_0 is the number of elements in X which are 0, n_1 is the number of elements in X which are 1 (and so $n = n_0 + n_1$), $p_0 = \frac{n_0}{n}, p_1 = \frac{n_1}{n}$, m_0 is the mean of $\{y_i : x_i = 0\}$, m_1 is the mean of $\{y_i : x_i = 1\}$, s is the population standard deviation of Y and y is NORM.S.DIST(NORM.S.INV(p_0),FALSE).

Source: https://www.real-statistics.com/correlation/biserial-correlation/

Biserial r_b

- The biserial correlation coefficient can also be computed from the **point-biserial correlation coefficient** using the following formula.

$$r_b = \frac{r_{pb}\sqrt{p_0 p_1}}{y}$$

y	x
23	1
24	1
15	0
16	0
32	1
32	0
54	1
13	0
12	0
25	0
34	1
36	0
45	1

$\boldsymbol{m_1}$	$mean(y\|x = 1) = 35.3333$	$\boldsymbol{m_0}$	$mean(y\|x = 0) = 35.3333$
$\boldsymbol{n}$	13	$\boldsymbol{n_0}$	6
$\boldsymbol{s}$	$std(y) = 12.21697$	$\boldsymbol{r_{pb}}$	$Corr(y, x) = 0.573219$
$\boldsymbol{p_0}$	$\frac{n_0}{n} = 0.461538$	$\boldsymbol{p_1}$	$1 - p_0 = 0.538462$
$\boldsymbol{z}$	0.096559	$\boldsymbol{y}$	0.397087
$\boldsymbol{r_b}$	0.719642		

Source: https://www.real-statistics.com/correlation/biserial-correlation/

Biserial r_b

The following statistic is standard normally distributed with **Fisher Transformation**. Here, $x' = FISHER(x)$ and the denominator is the standard error. Then, the *P* value can be obtained by standardized normal distribution.

$$z = \frac{\left(\frac{2r_b}{\sqrt{5}}\right)'}{\frac{1}{2}\sqrt{\frac{5}{n}}} = \frac{FISHER(-0.64367)}{0.228218} = \frac{-0.76441}{0.228218} = -3.34947$$

Therefore, *P* value is 0.00081.

Source: https://www.real-statistics.com/correlation/biserial-correlation/

Spearman Rank Correlation ρ

- **The Spearman correlation coefficient** is defined from the Pearson correlation coefficient between the rank variables.
- For a sample of size n, the n raw scores X_i, Y_i are converted to ranks $R(X_i), R(Y_i)$, and r_s is computed as

$$r_s = \rho_{R(X)R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

where ρ denotes the Pearson correlation coefficient with rank variables, $cov(R(X), R(Y))$ is the covariance of the rank variables, $\sigma_{R(X)}\ and\ \sigma_{R(Y)}$ are the standard deviations of the rank variables.

Source: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman Rank Correlation ρ

- Only if all n ranks are distinct integers, it can be computed using the popular formula.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks of each observation, n is the number of observations.

- Significance measurement could be obtained from t distribution, where degree of freedom is $n-2$.

$$t = r_s\sqrt{\frac{n-2}{1-r^2}}$$

Source: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman Rank Correlation ρ

PR, X_i	Reading/Day, Y_i	x_i rank	y_i rank	d_i	d_i^2
60	1	1	2	-1	1
65	1	2	2	0	0
71	2	3	4	-1	1
75	1	4	2	2	4
78	5	5	6	-1	1
81	6	6	7.5	-1.5	2.25
85	7	7	9.5	-2.5	6.25
89	8	8	11	-3	9
91	7	9	9.5	-0.5	0.25
95	6	10	7.5	2.5	6.25
99	4	11	5	6	36

- **Spearman Rank Corr.**

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6\times 67}{11(11^2-1)}$$

$$r_s = 0.695455$$

$$t = r_s\sqrt{\frac{n-2}{1-r^2}} = 2.870262$$

$$t_{2.87,df=9,two} = 0.018469$$

Tetrachoric Correlation r_{tet}

- The **tetrachoric correlation coefficient,** r_{tet}, is used when both variables are dichotomous, like the phi, but we need also to be able to assume both variables really are **continuous** and **normally distributed**.
- Thus it is applied to **ordinal *vs.* ordinal** data which has this characteristic. Ranks are discrete so in this manner it differs from the Spearman.
- The formula involves a trigonometric function called cosine.

Tetrachoric Correlation r_{tet}

X/Y	Y=1	Y=2	Total
X=1	a	b	a+b
X=2	c	d	c+d
Total	a+c	b+d	N

$$r_{tet} = cos\left[\frac{\pi}{1+\sqrt{\frac{ad}{bc}}}\right] = cos\left[\pi\times\frac{\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right]$$

X/Y	Rev>1M	Rev<1M	Total
Doctors	20	30	50
Teachers	15	5	20
Total	35	35	70

$$r_{tet} = cos\left[\pi\times\frac{\sqrt{30\times15}}{\sqrt{20\times5}+\sqrt{30\times15}}\right]$$

$$r_{tet} = 0.266255$$

Point Biserial r_{pb}

- **The point biserial correlation coefficient (r_{pb})** is a correlation coefficient used when one variable (e.g. Y) is dichotomous; Y can either be "naturally" dichotomous, like whether a coin lands heads or tails, or an artificially dichotomized variable.
- In most situations it is not advisable to dichotomize variables artificially. When a new variable is artificially dichotomized the new dichotomous variable may be conceptualized as having an underlying continuity.

Source: https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient

Point Biserial r_{pb}

- **Assumptions**

1. One of your two variables should be measured on a **continuous** scale.
2. Your other variable should be **dichotomous.**
3. There should be no outliers for the continuous variable for each category of the dichotomous variable.
4. Your continuous variable should be **approximately normally distributed** for each category of the dichotomous variable. You can test this using the Shapiro-Wilk test of normality.
5. Your continuous variable should have **equal variances** for each category of the dichotomous variable. You can test this using Levene's test of equality of variances.

Source: https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php
Source: https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient

Point Biserial r_{pb}

- The point-biserial correlation is mathematically equivalent to the Pearson (product moment) correlation; that is, if we have one continuously measured variable X and a dichotomous variable Y, $r_{xy} = r_{pb}$. This can be shown by assigning two distinct numerical values to the dichotomous variable.

$$r_{pb} = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}}, s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where s_{n-1} is the standard deviation used when data are available only for a sample of the population:

Source: https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient

Point Biserial r_{pb}

- The Tate (1954) provides results for the test statistic t calculated as follows, …

$$t = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1-r^2}}, where\ degree\ of\ freedom\ is\ n-2\ (\rho = 0)$$

Point Biserial r_{pb}

Question 1

Given each student's transcript, we want to know if the assignments completed correlate to the final grade with *P* value.

	A	B	C	D	E	F	G	H	I
Assignment	1	0	1	0	1	1	0	0	0
Final grade	90	60	80	20	70	80	60	50	40

Rank Biserial r_{rb}

- A method of reporting the effect size for the Mann–Whitney *U* test is with a measure of rank correlation known as the **rank-biserial correlation**.
- Like other correlational measures, the rank-biserial correlation can range from minus one to plus one, with a value of zero indicating no relationship.
- The correlation is the difference between the proportion of pairs favorable to the hypothesis (f) minus its complement (i.e., the proportion that is unfavorable (u)).

Rank Biserial r_{rb}

- This simple difference formula is just the difference of the common language effect size of each group, and is as follows:

$$r = f - u$$

- 95 of 100 men like the movie of “Star Trek.” The common language effect size is 95%, so the rank-biserial correlation is 95% minus 5%, and the rank-biserial $r = 0.90$.

Phi Coefficient ϕ

$$\phi = \pm\sqrt{\frac{\chi^2}{N}}$$

- **Phi coefficient** is used to measure the relationship between **two binary variables or dichotomous variables**.
- Phi coefficient is calculated by **contingency table** as follows.

X/Y	Male	Female	Total
Doctors	20	30	50
Teachers	15	5	20
Total	35	35	70

X/Y	Y=1	Y=0	Total
X=1	n_{11}	n_{10}	$n_{1\cdot}$
X=0	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	N

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 0}n_{\cdot 1}}} = \frac{20\times 5 - 30\times 15}{\sqrt{50\times 20\times 35\times 35}} = \frac{-350}{1106.7972} = -0.316$$

$$\chi^2 = -0.316^2\times 70 = 6.9999997$$

$$\Rightarrow df = (2-1)(2-1) = 1$$

$$\Rightarrow P\ value = 0.008151$$

Contingency Coefficient (C)

X/Y	Male	Female	Total
Doctors	20	30	50
Teachers	15	5	20
Total	35	35	70

- A technique for determining the correlation between two nominal variables cast in a frequency table larger than 2×2.

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i}.n_{.j}}{N}\right)^2}{\frac{n_{i}.n_{.j}}{N}} = \frac{\left(20 - \frac{50\times35}{70}\right)^2}{\frac{50\times35}{70}} + \frac{\left(5 - \frac{20\times35}{70}\right)^2}{\frac{20\times35}{70}} + \frac{\left(30 - \frac{50\times35}{70}\right)^2}{\frac{50\times35}{70}} + \frac{\left(15 - \frac{20\times35}{70}\right)^2}{\frac{20\times35}{70}}$$

$$\chi^2 = 1 + 2.5 + 1 + 2.5 = 7$$

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{7}{70 + 7}} = 0.311511$$

Cramér's V

X/Y	Y=1	Y=0	Total
X=1	n_{11}	n_{10}	$n_{1\cdot}$
X=0	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	N

- **Cramér's V (**sometimes referred to as **Cramér's phi** and denoted as $\boldsymbol{\varphi_c}$**)** is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive).
- φ_c is the intercorrelation of two discrete variables and may be used with variables having **two or more levels**. φ_c is a symmetrical measure: it does not matter which variable we place in the columns and which in the rows.
- Cramér's V may also be applied to goodness of fit chi-squared models when there is a 1 × *k* table (in this case *r* = 1).
- Cramér's V varies from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when each variable is completely determined by the other.

Source: https://en.wikipedia.org/wiki/Cram%C3%A9r%27s_V

Cramér's V

X/Y	Y=1	Y=0	Total
X=1	n_{11}	n_{10}	$n_{1\cdot}$
X=0	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	N

- Let a sample of size n of the simultaneously distributed variables A and B for $i = 1,2, \dots , r; j = 1,2, \dots , k$, be given by the frequencies.
- $n_{ij} = number\ of\ times\ the\ values\ \left(A_i, B_j\right)\ were\ observed$.
- The chi-squared statistics then is:

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{N}\right)^2}{\frac{n_{i.}n_{.j}}{N}},$$

$$n_{i\cdot} = \sum_j n_{ij}\ \ is\ the\ number\ of\ times\ the\ value\ A_i\ is\ observed\ and$$

$$n_{\cdot j} = \sum_i n_{ij}\ \ is\ the\ nuber\ of\ times\ the\ value\ B_j\ is\ observed.$$

Source: https://en.wikipedia.org/wiki/Cram%C3%A9r%27s_V

Source: https://en.wikipedia.org/wiki/Cram%C3%A9r%27s_V

Cramér's V

- Cramér's V is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1:

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\frac{\chi^2}{N}}{\min(k-1, r-1)}}$$

where:

φ is the phi coefficient.

χ^2 is the derived from Pearson chi-squared test

N is the grand total of observations

k being the number of columns

r being the number of rows

P value for the significance of V is calculated from Pearson chi-squared test.

Cramér's V

X/Y	Male	Female	Total
Doctors	20	30	50
Teachers	15	5	20
Total	35	35	70

- Does the job has gender preference?

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i.n._j}{N}\right)^2}{\frac{n_i.n._j}{N}}$$

$$\chi^2 = \frac{\left(20 - \frac{50\times35}{70}\right)^2}{\frac{50\times35}{70}} + \frac{\left(5 - \frac{20\times35}{70}\right)^2}{\frac{20\times35}{70}} + \frac{\left(30 - \frac{50\times35}{70}\right)^2}{\frac{50\times35}{70}} + \frac{\left(15 - \frac{20\times35}{70}\right)^2}{\frac{20\times35}{70}}$$

$$\chi^2 = 1 + 2.5 + 1 + 2.5 = 7$$

$$V = \sqrt{\frac{\frac{\chi^2}{N}}{\min(k-1, r-1)}} = \sqrt{\frac{\frac{7}{70}}{1}} = \sqrt{0.1} = 0.31622777 \Rightarrow Prob\left(\chi^2_{7,df=1}\right) = 0.008151$$

Goodman and Kruskal’s Lambda

- **Goodman & Kruskal's lambda (λ)** is a measure of proportional reduction in error in cross tabulation analysis.
- For any sample with a nominal independent variable and dependent variable (or ones that can be treated nominally), it indicates the extent to which the modal categories and frequencies for each value of the independent variable differ from the overall modal category and frequency, i.e., for all values of the independent variable together.

Goodman and Kruskal’s Lambda

- λ is defined by the equation.

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} = \frac{F_{iv} - M_{dv}}{N - M_{dv}}, where$$

F_{iv} is sum of the largest cell frequencies within each category of the IV, gender
M_{dv} is the largest marginal total in the categories of the DV, job type
ε_1 is the overall non-modal frequency
ε_2 is the sum of the non-modal frequencies for each value of the independent variable.

- Values for lambda range from **zero** (no association between independent and dependent variables) to one (perfect association).).

Source: https://en.wikipedia.org/wiki/Goodman_and_Kruskal%27s_lambda

Goodman and Kruskal’s Lambda

- Assume that the gender is the independent variable, the job type is the dependent variable, i.e., the question asked is “***can the job type be predicted better if the gender is known?***”

DV/IV	Male	Female	Total
Doctors	**20**	**30**	**50**
Teachers	15	5	20
Total	35	35	70

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} = \frac{F_{iv} - M_{dv}}{N - M_{dv}} = \frac{(20 + 30) - 50}{70 - 50} = 0$$

Goodman and Kruskal’s Lambda

What is the predicted gender based on job type?

DV/IV	Male	Female	Total
Doctors	20	**30**	50
Teachers	**15**	5	20
Total	**35**	35	70

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} = \frac{F_{iv} - M_{dv}}{N - M_{dv}} = \frac{(15 + 30) - 35}{70 - 35} = 0.28571$$

Reading

Nonparametric Correlation Techniques: Techniques for Correlating Nominal & Ordinal Variables
https://staff.blog.ui.ac.id/r-suti/files/2010/05/noparcorelationtechniq.pdf
Further reading for Goodman and Kruskal’s indicators
Goodman and Kruskal’s tau
https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html
Goodman and Kruskal's gamma
https://en.wikipedia.org/wiki/Goodman_and_Kruskal%27s_gamma
More Correlation Coefficients
https://www.andrews.edu/~calkins/math/edrm611/edrm13.htm#PHI

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))