MTMS.01.099 Mathematical Statistics

Lecture 14. Types of variables. Dependence measures

Tõnu Kollo

tonu.kollo@ut.ee

Fall 2016

Three types of data:

- nominal (Favaourite colour, Hobby)
- ordinal (Level of education, Marks at school)
- numerical (Measurements of blood pressure, Number of car accidents)

What can we calculate for these variables?
How to handle them?
We have sevaral characteristics for numerical random variables: mean, standard deviation, moments, central moments, quantiles,mode etc.
What can we find for categorical data?

Categorical data are usually for analysis coded:
variable sex: male - 0; female - 1
When variable has only 2 values, it is considered as *ordinal*
marks at school: 2 ; 3 ; 4 ; 5

Variable hobby: reading - 1, theatre - 2; sports - 3; music - 4 etc.
For nominal data we are not allowed to calculate almost anything, just mode. Which colour is the favourite most often?
And we can find frequencies of possible answers.

Ordinal data can be ordered (or are orded by values)

$$x_1, \ldots, x_n \quad \rightarrow \quad x_{(1)} \leq \ldots \leq x_{(n)}.$$

To ordered sample elements we attribute **ranks**

$$r_1, \ldots, r_n.$$

If $x_{(1)} < \ldots < x_{(n)}$ then rank of $r(x_k) = k$. If some elements in ordered sample are equal, then they all have equal ranks which equals to the average of corresponding places in the ordered sample. For example, if $x_2 < x_{(3)} = x_{(4)} < x_5$, then $x_{(3)}$ and $x_{(4)}$ both have rank 3.5.
So we have also n ranks where the sum of ranks $\sum_{i=1}^{n} r_i = \frac{n(n+1)}{2}$. For ordered sample we can find quantiles and mode. Often sample mean is also calculated but one has to be careful with its interpretation.

Dependence measures

Let us have 2 random variables (numerical) X and Y which are measured on n objects.

$$x_1, \ldots, x_n,$$

$$y_1, \ldots, y_n$$

So we have a sample of size n.
Do tall people have also heavier weight? The first idea to measure this would be using
linear correlation coefficient or
Pearson correlation coefficient

$$r(X, Y) = \frac{Cov(X, Y)}{\sqrt{D(X)D(Y)}}.$$

Linear correlation

Properties:
When X and Y are independent $r(X, Y) = 0$.

$$-1 \leq r(X, Y) \leq 1.$$

When $Y = aX + b$, $r(X, Y) = 1$ when $a > 0$ and $r(X, Y) = -1$ when $a < 0$, $a, b \in R$.
Estimate:

$$\widehat{r(X, Y)} = \frac{1/n \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \overline{x})(y_j - \overline{y})}{s_X s_Y}.$$

Rank correlations

It is not correct to calculate Linear correlation for ordinal data.
Between 2 ordinal variables we can calculate
rank correlations which measure monotone dependence between X and Y.
Variable Y is increasing with respect to X if from $x_i < x_j$ follows $y_i < y_j$, for all values $i, j = 1, \ldots, n$.
Variable Y is decreasing with respect to X if from $x_i < x_j$ follows $y_i > y_j$, for all values $i, j = 1, \ldots, n$.
Assume we have k different values x_i^* of X and l different values y_j^* of Y. Denote different ranks of X by $r_1, \ldots, r_k$ and ranks of Y by $q_1, \ldots, q_l$.
Denote the number of pairs in the sample where $X = x_i^*$ and $Y = y_j^*$ by n_{ij};
$n_{i\bullet}$ is the number of all pairs where X has value x_i^*;
$n_{\bullet j}$ is the number of all pairs where Y has value y_j^*.

Denote

$$D = \sum_{i=1}^{k} \sum_{j=1}^{l} (r_i - q_j)^2 n_{ij}.$$

In the case when all elements x_i and y_j are different
$D = 0$ if X and Y are increasing with respect to each other
When Y depends on X decreasingly D has maximum value

$$S = \frac{1}{3} n(n^2 - 1)$$

Definition Spearman rank correlation coefficient for the sample with non-repeated values is given by

$$\rho(X, Y) = 1 - \frac{2D}{S}.$$

In general case we need 2 more expressions:

$$S_1 = \frac{n(n^2-1)}{3} - \sum_{i=1}^{k} \frac{n_{i\bullet}(n_{i\bullet}^2-1)}{3}$$

$$S_2 = \frac{n(n^2-1)}{3} - \sum_{j=1}^{l} \frac{n_{\bullet j}(n_{\bullet j}^2-1)}{3}.$$

Definition Spearman rank correlation coefficient is given by

$$\rho(X,Y) = 1 - \frac{12D}{S_1+S_2}.$$

Properties:

$$-1 \leq \rho(X, Y) \leq 1$$

When X and Y are strongly monotone increasing, then $\rho(X, Y) = 1$
When X and Y are strongly monotone decreasing, then $\rho(X, Y) = -1$
When $\rho(X, Y) = 0$ we say that there is no monotone dependence between X and Y.
There exist more dependence measures based on ranks (Kendall's Tau, for example)

χ^2-coefficient

How to characterize dependence between 2 nominal variable? These dependence coefficients are based on χ^2-coefficient which is calculated from the frequency table.

$$\chi^2(X, Y) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(nn_{ij} - n_{i\bullet} n_{\bullet j})^2}{n_{i\bullet} n_{\bullet j}}.$$

Properties:
$\chi^2(X, Y) \approx 0$ corresponds to the independence
$\chi^2(X, Y) \leq min[n(k-1), n(l-1)]$
$\chi^2(X, Y) = min[n(k-1), n(l-1)]$ if one variable is the unique function of the other variable.
For instance, Y is a unique function of X if $X = x_i$ determines value of Y.

How to normalize χ^2-coefficient?
There are 2 formulas in literature.
Definition Chuproff correlation coefficient is given by

$$T(X,Y) = \sqrt{\frac{\chi^2(X,Y)}{n\sqrt{(k-1)(l-1)}}}.$$

Properties:

$$0 \leq T(X,Y) \leq 1$$

Value 0 corresponds to the independence of X and Y and $T(X,Y) = 1$ when $l = k$ and there is functional depndence between X and Y.
Definition Cramér's correlation coefficient is given by

$$C(X,Y) = \sqrt{\frac{\chi^2(X,Y)}{n \times min((k-1)(l-1))}}.$$

Correlation ratio

Can we estimate dependence between a numerical random variable and a nominal one?
We are not allowed to find linear correlation or monotone correlation coefficients.
We can find **Correlation ratios**
Let Y be a random variable and X a nominal variable.
We start from the same frequency table where n_{ij} is the number of pairs in the sample where $X = x_i^*$ and $Y = y_j^*$

$$\sum_{j=1}^{l} n_{ij} = n_{i\bullet}; \qquad \sum_{i=1}^{k} n_{ij} = n_{\bullet j}$$

Find values of conditional means:

$$\overline{y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} y_j^* n_{ij} \qquad i = 1, \ldots, k$$

the sample mean of values of Y when $X = x_i^*$.

$$\overline{x}_j = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} x_i^* n_{ij} \qquad j = 1, \ldots, l,$$

the sample mean of values of X when $Y = y_j^*$.

Correlation ratios

Definition The sample correlation ratio of random variable Y and nominal variable X is

$$\hat{\eta}(Y|X) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{k}(\overline{y}_i - \overline{y})^2 n_{i\bullet}}}{s_Y}.$$

The sample correlation ratio of random variable X and nominal variable Y is

$$\hat{\eta}(X|Y) = \frac{\sqrt{\frac{1}{n}\sum_{j=1}^{l}(\overline{x}_j - \overline{x})^2 n_{\bullet j}}}{s_X}.$$

Correlation ratios

When Correlation ratio is not zero we say that between X and Y is regression dependence.
There is no regression dependence if $E(Y|X) = EX$).
Properties:
$0 \leq \hat{\eta}(Y|X) \leq 1$;
When $\hat{\eta}(Y|X) = 0$ there is no regression dependence between Y and X.
When $\hat{\eta}(Y|X) = 1$then Y is functionally dependent on X: $Y = g(X)$.
Correlation ratios $\hat{\eta}(Y|X)$ can be calculated also when X is ordinal or numerical variable.