Chapter 4

Descriptive measures of association for bivariate distributions

Now we come to describe and characterise specific features of bivariate frequency distributions, i.e., intrinsic structures of raw data sets $\{(x_i, y_i)\}_{i=1,\ldots,n}$ obtained from statistical samples S for 2-D variables (X, Y) from some population of study objects. Let us suppose that the spectrum of values resp. categories of X is $a_1, a_2, \ldots, a_k$, and the spectrum of values resp. categories of Y is $b_1, b_2, \ldots, b_l$, where $k, l \in \mathbb{N}$. Hence, for the bivariate joint distribution there exists a total of $k \times l$ possible combinations $\{(a_i, b_j)_{i=1,\ldots,k;j=1,\ldots,l}$ of values resp. categories for (X, Y). In the following we will denote associated absolute (observed) frequencies by $o_{ij} := o_n(a_i, b_j)$, and relative frequencies by $h_{ij} := h_n(a_i, b_j)$.

4.1 $(k \times l)$ contingency tables

Consider a raw data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for a 2-D variable (X, Y) giving rise to $k \times l$ combinations of values resp. categories $\{(a_i, b_j)\}_{i=1,\ldots,k;j=1,\ldots,l}$. The bivariate joint distribution of observed absolute frequencies o_{ij} may be conveniently represented in terms of a $(k \times l)$ contingency table (or cross tabulation) by

$$
\begin{array}{c|ccccccc|c}
o_{ij} & b_1 & b_2 & \ldots & b_j & \ldots & b_l & & \sum_j \\
\hline
a_1 & o_{11} & o_{12} & \ldots & o_{1j} & \ldots & o_{1l} & & o_{1+} \\
a_2 & o_{21} & o_{22} & \ldots & o_{2j} & \ldots & o_{2l} & & o_{2+} \\
\vdots & & & & & & & & \vdots \\
a_i & o_{i1} & o_{i2} & \ldots & o_{ij} & \ldots & o_{il} & & o_{i+} \\
\vdots & & & & & & & & \vdots \\
a_k & o_{k1} & o_{k2} & \ldots & o_{kj} & \ldots & o_{kl} & & o_{k+} \\
\hline
\sum_i & o_{+1} & o_{+2} & \ldots & o_{+j} & \ldots & o_{+l} & & n
\end{array}
\tag{4.1}
$$

where it holds for all $i = 1, \ldots, k$ and $j = 1, \ldots, l$ that

$$0 \leq o_{ij} \leq n \text{ and } \sum_{i=1}^{k} \sum_{j=1}^{l} o_{ij} = n \tag{4.2}$$

The corresponding marginal absolute frequencies of X and of Y are

$$o_{i+} := o_{i1} + o_{i2} + \ldots + o_{ij} + \ldots + o_{il} =: \sum_{j=1}^{l} o_{ij} \tag{4.3}$$

$$o_{+j} := o_{1j} + o_{2j} + \ldots + o_{ij} + \ldots + o_{kj} =: \sum_{i=1}^{k} o_{ij} \tag{4.4}$$

One obtains the related bivariate joint distribution of observed relative frequencies h_{ij} following the systematics of Eq. (**??**) to yield

$$\begin{array}{c|cccccc|c}
h_{ij} & b_1 & b_2 & \dots & b_j & \dots & b_l & \sum_j \\
\hline
a_1 & h_{11} & h_{12} & \dots & h_{1j} & \dots & h_{1l} & o_{1+} \\
a_2 & h_{21} & h_{22} & \dots & h_{2j} & \dots & h_{2l} & o_{2+} \\
\vdots & & & & & & & \vdots \\
a_i & h_{i1} & h_{i2} & \dots & h_{ij} & \dots & h_{il} & h_{i+} \\
\vdots & & & & & & & \vdots \\
a_k & h_{k1} & h_{k2} & \dots & h_{kj} & \dots & h_{kl} & o_{k+} \\
\hline
\sum_i & h_{+1} & h_{+2} & \dots & h_{+j} & \dots & h_{+l} & 1
\end{array} \tag{4.5}$$

Again, it holds for all $i = 1, \dots, k$ and $j = 1, \dots, l$ that

$$0 \leq h_{ij} \leq 1 \text{ and } \sum_{i=1}^{k} \sum_{j=1}^{l} h_{ij} = 1 \tag{4.6}$$

while the marginal relative frequencies of X and of Y are

$$h_{i+} := h_{i1} + h_{i2} + \dots + h_{ij} + \dots + h_{il} =: \sum_{j=1}^{l} h_{ij} \tag{4.7}$$

$$h_{+j} := h_{1j} + h_{2j} + \dots + h_{ij} + \dots + h_{kj} =: \sum_{i=1}^{l} h_{ij} \tag{4.8}$$

On the basis of a $(k \times l)$ contingency table displaying the relative frequencies of the bivariate joint distribution of some 2-D (X, Y), one may define two kinds of related conditional relative frequency distributions, namely

(i) the conditional distribution of X given Y by

$$h(a_i|b_j) := \frac{h_{ij}}{h_{+j}} \tag{4.9}$$

and

(ii) the conditional distribution of Y given X by

$$h(b_j|a_i) := \frac{h_{ij}}{h_{i+}} \tag{4.10}$$

Then, by means of these conditional distributions, a notion of statistical independence of variables X and Y is defined to correspond to the simultaneous properties

$$h(a_i|b_j) = h(a_i) = h_{i+} \text{ and } h(b_j|a_i) = h(b_j) = h_{+j} \tag{4.11}$$

Given these properties hold, it follows from Eqs. (4.9) and (4.10) that

$$h_{ij} = h_{i+}h_{+j}. \tag{4.12}$$

4.2 Measures of association for the metrical scale level

Next, specifically consider a raw data set $\{(x_i, y_i)\}_{i=1,\dots,n}$ from a statistical sample S for some metrically scaled 2-D variable (X, Y). The bivariate joint distribution of (X, Y) in this sample can be conveniently represented graphically in terms of a scatter plot. Let us now introduce two kinds of measures for the description of specific characteristic features of such distributions.

4.2.1 Sample covariance

The first standard measure characterising bivariate joint distributions of metrically scaled 2-D (X, Y) descriptively is the dimensionful sample covariance s_{XY} (metr), defined by

(i) From raw data set:

$$s_{XY} := \frac{1}{n-1}[(x_1 - \overline{x})(y_1 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})] \tag{4.13}$$

$$=: \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \tag{4.14}$$

alternatively:

$$s_{XY} = \frac{1}{n-1}[x_1y_1 + \ldots + x_ny_n - n\overline{x}.\overline{y}] \tag{4.15}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_iy_i - n\overline{x}.\overline{y}). \tag{4.16}$$

(ii) From relative frequency distribution:

$$s_{XY} = \frac{n}{n-1}[(a_1 - \overline{x})(b_1 - \overline{y})h_{11} + \ldots + (a_k - \overline{x})(b_l - \overline{y})h_{kl}] \tag{4.17}$$

$$= \frac{n}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{l}(a_i - \overline{x})(b_j - \overline{y})h_{ij}, \tag{4.18}$$

alternatively:

$$s_{XY} = \frac{n}{n-1}[a_1b_1h_{11} + \ldots + a_kb_lh_{kl} - \overline{x}.\overline{y}] \tag{4.19}$$

$$= \frac{n}{n-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{l}a_ib_jh_{ij} - \overline{x}.\overline{y}\right]. \tag{4.20}$$

Remark 4.2.1 The alternative formulae provided here prove computationally more efficient.

It is worthwhile to point out that in the research literature it is standard to define for bivariate joint distributions of metrically scaled 2-D (X, Y) a dimensionful symmetric (2×2) covariance matrix S according to

$$S := \begin{pmatrix} s_X^2 & s_{XY} \\ s_{YX} & s_Y^2 \end{pmatrix} \tag{4.21}$$

the components of which are defined by Eqs. (**??**) and (4.14). The determinant of S, given by

$$\det(S) = s_X^2 s_Y^2 - s_{XY}^2,$$

is positive as long as $s_X^2 s_Y^2 - s_{XY}^2 > 0$ which applies in most practical cases. Then S is regular, and thus a corresponding inverse S^{-1} exists. The concept of a regular covariance matrix S and its inverse S^{-1} generalises in a straightforward fashion to the case of multivariate joint distributions of metrically scaled $m - D(X, Y, \ldots, Z)$, where $S \in \mathbb{R}^{m \times m}$ is given by

$$S := \begin{pmatrix} s_X^2 & s_{XY} & \ldots & s_{XZ} \\ s_{YX} & s_Y^2 & \ldots & s_{YZ} \\ \vdots & \vdots & \ddots & \vdots \\ s_{ZX} & s_{ZY} & \ldots & s_Z^2 \end{pmatrix} \tag{4.22}$$

Exercise 1 Find the covariance of the eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the two variables.

Exercise 2 Consider the data shown in the next table, on consumption C and income Y for countries A and B (measured in dollars):

	Country A		Country B	
Obs	C	Y	C	Y
1	70	80	75	80
2	65	100	50	100
3	90	120	95	120
4	95	140	120	140
5	110	160	90	160
6	115	180	115	180
7	120	200	110	200
8	140	220	140	220
9	155	240	175	240
10	150	260	140	260

1. For country A, compute the sample means and standard deviations for both variables. Compute the covariance and the correlation coefficient based on the next auxiliar table:

Obs $= i$	C_i	Y_i	$(C_i - \overline{C})^2$	$(Y_i - \overline{Y})^2$	$(C_i - \overline{C})(Y_i - \overline{Y})$

2. Construct a scatter diagram for consumption and income (consumption in the Y axis and income in the X axis).

3. Repeat the exercise for country B. Compare the means and standard deviations.

4. Compare the covariances for both countries. Comment intuitively.

5. Suppose that the data on consumption for country A is altered in the following way:

   (a) Observations on consumption and income are measured in cents instead of dollars.

   (b) Observations on consumption are measured in cents and observations on income in dollars.

   (c) Observations on consumption are measured in dollars, but income is measured in Crowns (1 dollar = 7.85 Crowns).

   (d) To all the observations on income and consumption the number 10 is added (arbitrarily).

   (e) Only to the observations on income the number 10 is added.

   For all the cases, compute the covariance and comment your findings intuitively.

4.2.2 Bravais and Pearson's sample correlation coefficient

The sample covariance s_{XY} constitutes the basis for the second standard measure characterising bivariate joint distributions of metrically scaled 2-D (X, Y) descriptively, the normalised dimensionless sample correlation coefficient r (metr) devised by the French physicist Auguste Bravais (1811-1863) and the English mathematician and statistician Karl Pearson (1857-1936) for the purpose of analysing corresponding raw data $\{(x_i, y_i)\}_{i=1,\dots,n}$ for the existence of linear statistical associations. It is defined in terms of the bivariate sample covariance s_{XY} and the univariate sample standard deviations s_X and s_Y by (cf. Bravais (1846) and Pearson (1901))

$$r := \frac{s_{XY}}{s_X s_Y} \tag{4.23}$$

Due to normalisation, the range of the sample correlation coefficient is $-1 \leq r \leq +1$. The sign of r encodes the direction of a correlation. As to interpreting the strength of a correlation via the magnitude $|r|$, in practice one typically employs the following qualitative.

Rule of thumb:

- $0.0 = |r|$: no correlation
- $0.0 < |r| < 0.2$: very weak correlation
- $0.2 \leq |r| < 0.4$: weak correlation
- $0.4 \leq |r| < 0.6$: moderately strong correlation
- $0.6 \leq |r| \leq 0.8$: strong correlation
- $0.8 \leq |r| < 1.0$: very strong correlation
- $1.0 = |r|$: perfect correlation.

Exercise 3 Student recorded the outside temperature X (in Celsius) and the duration of his way to university Y (in minutes).

x_i	-20	-10	0	10	20
y_i	60	40	35	20	20

How strong is the correlation between these two features?

In addition to Eq. (4.21), it is convenient to define a dimensionless symmetric (2×2) correlation matrix R by

$$R := \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{4.24}$$

which is regular and positive definite as long as $1 - r^2 > 0$. Then its inverse R^{-1} is given by

$$R^{-1} = \frac{1}{1-r^2}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \tag{4.25}$$

Note that for non-correlating metrically scaled variables X and Y, i.e., when $r = 0$, the correlation matrix degenerates to become a unit matrix, $R = 1$.
Again, the concept of a regular and positive definite correlation matrix R, with inverse R^{-1}, generalises to multivariate joint distributions of metrically scaled m-$D(X, Y, \ldots, Z)$, where $R \in \mathbb{R}^{m \times m}$ is given by

$$R := \begin{pmatrix} 1 & r_{XY} & \ldots & r_{XZ} \\ r_{YX} & 1 & \ldots & r_{YZ} \\ \vdots & \vdots & \ddots & \vdots \\ r_{ZX} & r_{ZY} & \ldots & 1 \end{pmatrix} \tag{4.26}$$

Note that R is a dimensionless quantity which, hence, is scale-invariant.

Exercise 4 Statistical packages are able to calculate correlations for multiple pairings of variables, often reporting their findings in a correlation matrix. Correlation matrices report correlation coefficients for all pairing of quantitative variables. We are going to create a correlation matrix for the per capita numbers of cigarettes smoked (sold) in 43 states and the District of Columbia in 1960 and death rates for various forms of cancer. The data, originally from Fraumeni et al.(1968), can be downloaded at

http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html as a text file. Use R to calculate correlation coefficients for each variable pairing. Interpret the correlation coefficients. Which cancers are associated with smoking?

Variable	Description
CIG	cigarettes sold per capita
BLAD	bladder cancer deaths per 100, 000
LUNG	lung cancer deaths per 100, 000
KID	kidney cancer deaths per 100, 000
LEUK	leukemia cancer deaths per 100, 000

4.3 Measures of association for the ordinal scale level

At the ordinal scale level, raw data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for a 2-D variable (X, Y) is not necessarily quantitative in nature. Therefore, in order to be in a position to define a sensible quantitative bivariate measure of statistical association for ordinal variables, one needs to introduce meaningful surrogate data which is numerical. This task is realised by means of defining so-called ranks, which are assigned to the original ordinal data according the procedure described in the following. Begin by establishing amongst the observed values $\{x_i\}_{i=1,\ldots,n}$ resp. $\{y_i\}_{i=1,\ldots,n}$ their natural hierarchical order, i.e.,

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \text{ and } y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)} . \tag{4.27}$$

Then, every individual x_i resp. y_i is assigned a numerical rank which corresponds to its position in the ordered sequences (4.27):

$$x_i \mapsto R(x_i), \;\; y_i \mapsto R(y_i), \;\text{ for all } i = 1, \ldots, n . \tag{4.28}$$

Should there be any "tied ranks" due to equality of some x_i or y_i, one assigns the arithmetical mean of these ranks to all x_i resp. y_i involved in the "tie". Ultimately, by this procedure the entire bivariate raw data undergoes a transformation

$$\{(x_i, y_i)\}_{i=1,\ldots,n} \mapsto \{[R(x_i), R(y_i)]\}_{i=1,\ldots,n} \tag{4.29}$$

yielding n pairs of ranks to numerically represent the original ordinal data. Given surrogate rank data, the means of ranks always amount to

$$\overline{R}(x) := \frac{1}{n}\sum_{i=1}^{n} R(x_i) = \frac{n+1}{2} \tag{4.30}$$

$$\overline{R}(y) := \frac{1}{n}\sum_{i=1}^{n} R(y_i) = \frac{n+1}{2} \tag{4.31}$$

The variances of ranks are defined in accordance with Eqs. (**??**) and (**??**), i.e.,

$$s^2_{R(x)} := \frac{1}{n-1}\left[\sum_{i=1}^{n} R^2(x_i) - n\overline{R}^2(x)\right] = \frac{n}{n-1}\left[\sum_{i=1}^{k} R^2(a_i)h_{i+} - \overline{R}^2(x)\right] \tag{4.32}$$

$$s^2_{R(y)} := \frac{1}{n-1}\left[\sum_{i=1}^{n} R^2(y_i) - n\overline{R}^2(y)\right] = \frac{n}{n-1}\left[\sum_{j=1}^{k} R^2(a_j)h_{+j} - \overline{R}^2(y)\right] \tag{4.33}$$

In addition, to characterise the joint distribution of ranks, a covariance of ranks is defined in line with Eqs. (4.16) and (4.20) by

$$s_{R(x)R(y)} := \frac{1}{n-1}\left[\sum_{i=1}^{n} R(x_i)R(y_i) - n\overline{R}(x)\overline{R}(y)\right] = \frac{n}{n-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{l} R(a_i)R(b_j)h_{ij} - \overline{R}(x)\overline{R}(y)\right] \tag{4.34}$$

On this fairly elaborate technical backdrop, the English psychologist and statistician Charles Edward Spearman FRS (1863-1945) defined a dimensionless rank correlation coefficient r_S (ord), in analogy to Eq. (4.23), by (cf. Spearman (1904))

$$r_S := \frac{s_{R(x)R(y)}}{s_{R(x)}s_{R(y)}} \qquad (4.35)$$

The range of this rank correlation coefficient is $-1 \leq r_S \leq +1$. Again, while the sign of r_S encodes the direction of a rank correlation, in interpreting the strength of a rank correlation via the magnitude $|r_S|$ one usually employs the qualitative.

Rule of thumb:

. $0.0 = |r_S|$: no rank correlation

. $0.0 < |r_S| < 0.2$: very weak rank correlation

. $0.2 \leq |r_S| < 0.4$: weak rank correlation

. $0.4 \leq |r_S| < 0.6$: moderately strong rank correlation

. $0.6 \leq |r_S| \leq 0.8$: strong rank correlation

. $0.8 \leq |r_S| < 1.0$: very strong rank correlation

. $1.0 \leq |r_S|$: perfect rank correlation.

When no tied ranks occur, Eq. (4.35) simplifies to (cf. Hartung et al (2005))

$$r_S = 1 - \frac{6\displaystyle\sum_{i=1}^{n}[R(x_i) - R(y_i)]^2}{n(n^2-1)}. \qquad (4.36)$$

Exercise 5 Two judges at a fete placed the ten entries for the 'best fruit cakes' competition in order as follows (1 denotes first,...)

Entry	A	B	C	D	E	F	G	H	I	J
Judge 1 (x)	2	9	1	3	10	4	6	8	5	7
Judge 2 (y)	6	9	2	1	8	4	3	10	7	5

Is there a linear relationship between the rankings produced by the two judges?

Exercise 6 Find the value of r_S for the following data

Ranks x	1	2	2	5	4	6	7	8
Ranks y	1	3	4	2	5	6	6	6

4.4 Measures of association for the nominal scale level

Lastly, let us turn to consider the case of quantifying descriptively the degree of statistical association in raw data $\{(x_i, y_i)\}_{i=1,\dots,n}$ for a nominally scaled 2-D variable (X, Y) with categories $\{(a_i, b_j)\}_{i=1,\dots,k;j=1,\dots,l}$. The starting point are the observed absolute resp. relative (cell) frequencies o_{ij} and h_{ij} of the bivariate joint distribution of (X, Y), with marginal frequencies o_{i+} resp. h_{i+} for X and o_{+j} resp. h_{+j} for Y. The χ^2-statistic devised by the English mathematical statistician Karl Pearson FRS (1857-1936) rests on the notion of statistical independence of two variables X and Y in that it takes the corresponding formal condition provided by Eq. (4.12) as a reference. A simple algebraic manipulation of this condition obtains

$$h_{ij} = h_{i+}h_{+j} \Rightarrow \frac{o_{ij}}{n} = \frac{o_{i+}}{n}\frac{o_{+j}}{n} \Rightarrow o_{ij} = \frac{o_{i+}o_{+j}}{n} \qquad (4.37)$$

Pearson's descriptive χ^2-statistic (cf. Pearson (1900)) is then defined by

$$\chi^2 := \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(o_{ij} - \frac{o_{i+}o_{+j}}{n}\right)^2}{\frac{o_{i+}o_{+j}}{n}} = n \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(h_{ij} - h_{i+}h_{+j})^2}{h_{i+}h_{+j}} \tag{4.38}$$

whose range of values amounts to $0 \leq \chi^2 \leq \max(\chi^2)$, with $\max(\chi^2) := n[\min(k,l) - 1]$.

Remark 4.4.1 Provided $\frac{o_{i+}o_{+j}}{n} \geq 5$ for all $i = 1, \ldots, k$ and $j = 1, \ldots, l$, Pearson's χ^2-statistic can be employed for the analysis of statistical associations for 2-D variables (X, Y) of almost all combinations of scale levels.

The problem with Pearson's χ^2-statistic is that, due to its variable spectrum of values, it is not clear how to interpret the strength of statistical associations. This shortcoming can, however, be overcome by resorting to the measure of association proposed by the Swedish mathematician, actuary, and statistician Carl Harald Cramér (1893-1985), which basically is the result of a special kind of normalisation of Pearson's measure. Thus, Cramér's V, as it has come to be known, is defined by (cf. Cramér (1946))

$$V := \sqrt{\frac{\chi^2}{\max(\chi^2)}}, \tag{4.39}$$

with range $0 \leq V \leq 1$. For the interpretation of results, one may now employ the qualitative

Rule of thumb:

- $0.0 \leq V < 0.2$: weak association
- $0.2 \leq V < 0.6$: moderately strong association
- $0.6 \leq V \leq 1.0$: strong association.

Exercise 7 The victory of the incumbent, Bill Clinton, in the 1996 presidential election was attributed to improved economic conditions and low unemployment. Suppose a survey of 800 adults taken soon after the election resulted in the following cross-classification of financial condition with education level:

Financial Condition	H.S. Degree or Lower	Some College	College Degree or Higher	Total
Worse off now than before	91	39	18	148
No difference	104	73	31	208
Better off now than before	235	48	161	444
Total	430	160	210	800

At the .05 level of significance, is there evidence of a relationship between financial condition and education level?

Exercise 8 A random sample of 50 women were tested for cholesterol and classified according to age and cholesterol level. The results are given in the following contingency table, where the rows represent age and the columns represent cholesterol level:

	< 210	≥ 210	Row total
< 50	18	7	25
≥ 50	8	17	25
Column total	26	24	50

Test the following null hypothesis at significance level $\alpha = 0.05$ the hypothesis : "age and cholesterol are independent attributes of classification".