Lecture 7: Expectation and Correlation

Dr. Roman V Belavkin

BIS3226

Contents

1 Databases and Random Variables

Databases and Random Variables

Case:	Age	Gender	M. Income (£ K)	M. Expenses (£ K)	Home owner	Credit score
1	21	0	2	1	0	3
2	18	1	1	2	0	1
3	50	1	6	2	1	5
4	23	0	3	1	1	4
5	40	1	3	2	0	2

Variables Age = [1,2,...,100], Gender =[0 (Female), 1 (Male)]

Question 1. *How often does each value appear in the data?*

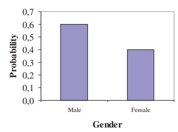
Random variables if each value is associated with probability

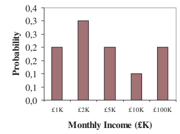
$$P(\text{Male}) = \frac{3}{5}\,, \quad P(\text{Female}) = \frac{2}{5}$$

Note that $P(\text{Male}) + P(\text{Female}) = 1$

Is There Structure in Data?

- Using the concept of random variables, we can analyse the distributions of each variable in the database





- Each case in the database can be seen as a complex (joint) event (e.g. Case 1 is Age=21, Gender=Female, etc).
- Thus, the whole database can be seen as a joint probability

$$P(\text{Case}) = P(\text{Age}, \text{Gender}, \text{Income}, \text{Expenses}, \text{H. owner}, \text{C. score})$$

- Are these variables independent or not?

2 Measures of Location

Measures of Location

- Answer questions such as '*What is the most probable value?*', '*What value should I expect in the long term?*'
- If variable x has n possible values X_1, X_2,...,X_n with probabilities $P(X_1)$, $P(X_2)$,...,$P(X_n)$, then we can compute the *expected value*

$$E\{x\} = X_1 P(X_1) + X_2 P(X_2) + \cdots + X_n P(X_n) = \sum_{i=1}^{n} X_i \, P(X_i)$$

- If all $P(x) = \frac{1}{n}$, then $E\{x\}$ is simply the average (the mean) value.

Example 1. Each value of variable $\texttt{Age} = 21$, 18, 50, 23, 40 occurs once. Therefore $P(\texttt{Age}) = \frac{1}{5}$ and

$$E\{\texttt{Age}\} = \frac{21 + 18 + 50 + 23 + 40}{5} = 30,4$$

Centre of Gravity

- What is the 'average' case?
- The expected value for a joint distribution of m random variables x_1, x_2,...,x_m is a point in an m–dimensional space with coordinates given by expectations of each of the m variables, and is called the *centre of gravity*

$$E\{x\} = (E\{x_1\}, E\{x_2\}, \ldots, E\{x_m\})$$

- For our data, this is the expected case (i.e the average case)

$$E\{\texttt{Case}\} = (E\{\texttt{Age}\}, E\{\texttt{Gender}\}, \ldots, E\{\texttt{C. score}\})$$

3 Measures of Dispersion

Measures of Dispersion

- Answer questions such as '*What is the range of the variable?*', '*What risk is associated with the variable?*'
- We can compute the average deviation from the expected value

$$E\Big\{|x - E\{x\}|\Big\} = \sum_{i=1}^{n} |X_i - E\{x\}| P(X_i)$$

- Or the average squared deviation, called the *variance*

$$Var\{x\} = E\Big\{|x - E\{x\}|^2\Big\} = \sum_{i=1}^{n} |X_i - E\{x\}|^2 P(X_i)$$

- *Standard deviation* is $Sdev\{x\} = \sqrt{Var\{x\}}$

Measures of Dispersion (cont.)

Example 2. Find $Var\{\text{Age}\}$ and $Sdev\{\text{Age}\}$?

1. Earlier we found $E\{\text{Age}\} = 30,4$.
2. We need to find squared deviations from 30,4.

$$(21 - 30,4)^2\ ,\ (18 - 30,4)^2\ ,\ (50 - 30,4)^2\ ,\ (23 - 30,4)^2\ ,\ (40 - 30,4)^2$$

3. Then we multiply each by $P(\text{Age}) = \frac{1}{5}$, and their sum gives the variance

$$Var\{\text{Age}\} = \frac{1}{5}((21 - 30,4)^2 + \cdots + (40 - 30,4)^2) = 154,64$$

4. Standard deviation is a square root of the variance

$$Sdev(Age) = \sqrt{154,64} = 12,44$$

Covariance

- Compares concentration of one variable with respect to another.
- If x and y are two random variables, then their *covariance* is

$$Cov(x, y) = E\Big\{(x - E\{x\})(y - E\{y\})\Big\}$$

- Note that $Cov(x, y) = Cov(y, x)$ and $Cov(x, x) = Var\{x\}$
- If x and y have 'similar' values, then $E\{x\} \approx E\{y\}$ and

$$Cov(x, y) \approx Var\{x\} \approx Var\{y\}$$

- If x and y are not 'similar', then $Cov(x, y) \approx 0$

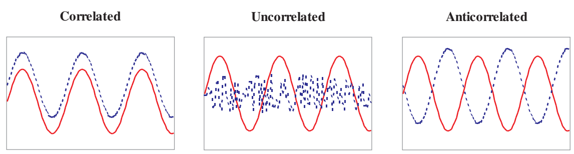
4 Correlation

Correlation

- The ratio of covariance with respect to variances of each variable is called *correlation*

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var\{x\}Var\{y\}}}$$

- If $x = y$, then $Corr(x, y) = 1$ (for $Cov(x, x) = Var\{x\}$)



$Corr(x, y) = 1$ $\quad$ $Corr(x, y) = 0$ $\quad$ $Corr(x, y) = -1$

Correlation Matrix

- Correlations (or covariances) can tell us how ‘similar’ are two random variables.
- Below is the *correlation matrix* showing correlations of each pair of variables in our database

	Age	Gender	Income	Expenses	H. owner	C. score
Age	1,0	0,6	0,9	0,6	0,4	0,5
Gender	0,6	1,0	0,2	1,0	-0,2	-0,3
Income	0,9	0,2	1,0	0,2	0,7	0,9
Expenses	0,6	1,0	0,2	1,0	-0,2	-0,3
H. owner	0,4	-0,2	0,7	-0,2	1,0	0,9
C. score	0,5	-0,3	0,9	-0,3	0,9	1,0

Correlation Is Not Causation

- There is a positive correlation between sales of ice-cream and shark attacks. Does this mean that ice-cream causes shark attacks?
- It is a common fallacy to conclude a causal relation based on correlation
- Often, correlation between x and y can be because they both depend on (or caused by) a third variable z (e.g. both ice-cream sales and shark attacks increase in the summer season)