Тайны корреляционных связей в статистике (Анализ корреляций)

Введение

Структура лекции

1. Кейс «Эффективность работы подготовительных курсов»
2. Связи (зависимости) между переменными
3. Понятие корреляции. Вычисление линейного коэффициента корреляции Пирсона. Условия применимости
4. Частная корреляция. Величина и надежность зависимости
5. Функции распределения. Нормальное распределение
6. Ложные корреляции
7. Некоррелированность и независимость
8. Ранговые коэффициенты корреляции
9. Если распределения ненормальны
10. Закон больших чисел и коэффициент корреляции
11. Закон Гаусса в мире случайного
12. Доверительные границы

Литература 3

Благовещенский Ю.Н. Тайны корреляционных связей в статистике. – М.: Научная книга: ИНФРА-М, 2009

Определение корреляции (двумерные методы исследования)

1. Понятие зависимости (связи двух переменных) не тождественно понятию причинности (каузальной связи);

Связь между переменными означает согласованное изменение двух переменных;

2. Зависимость (связь) носит вероятностный характер;
3. Методы и алгоритмы определения взаимосвязи переменных зависят от типов переменных.

Переменные любых типов связаны (зависимы) между собой, если наблюдаемые значения этих переменных изменяются (распределены) согласованным образом (если зная значение одной переменной, мы можем предсказать значение другой).

Наиболее распространенное понятие для обозначения связи двух переменных – **корреляция.**

(Ф. Гальтон (1822-1911), К. Пирсон (1857-1936) – основоположники корреляционного анализа))

Вычисление коэффициента корреляции К.Пирсона (двумерные методы исследования)

Коэффициент корреляции предполагает:

1. две переменные измеряются по крайней мере в интервальной шкале;

2. определяет степень, с какой значения двух переменных пропорциональны друг другу;

3. является безразмерной величиной, изменяется от -1 до +1;

4. корреляция может быть положительной и отрицательной;

5. рассчитывается по формуле:

$$r^*_{XY} = \frac{\sum_{i=1}^{n}(y_i - \bar{Y}) \cdot (x_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{Y})^2}} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{S_X \cdot S_Y}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- выборочные средние;

$$S_Y^2 = \frac{1}{n}\sum (y_i - \bar{Y})^2 = \overline{Y^2} - (\bar{Y})^2$$

$$S_X^2 = \frac{1}{n}\sum (x_i - \bar{X})^2 = \overline{X^2} - (\bar{X})^2$$

- выборочные дисперсии

Условия применимости коэффициента корреляции

Коэффициент определен только для линейных зависимостей. Это значит, что возможно его искажение по следующим причинам:

1. Наличие выбросов, т. е. нетипичных, резко выделяющихся наблюдений;
2. Отсутствие однородности в имеющихся данных. В таком случае необходимо вычислять корреляцию для каждой отдельной группы данных.
3. Наличие нелинейной зависимости между переменными.

Во всех случаях нужна визуализация данных для проверки всех вышеперечисленных условий (диаграмма рассеяния).

Частные коэффициенты корреляции

$$r_*(Y, X_1; X_2) = \frac{r^*(Y, X_1) - r^*(Y, X_2) \cdot r^*(X_1, X_2)}{\sqrt{(1 - r^{*2}(Y, X_2)) \cdot (1 - r^{*2}(X_1, X_2))}}$$

Пример ложной корреляции, проясняемый частной корреляцией

Параметры	Корреляция и значимость	Возраст	Отношение к приезжим	Посещение церкви
Возраст	Коэффициент Пирсона	1,000	0,468	0,779
	Значимость	-	0,005	0,000
Отношение к приезжим	Коэффициент Пирсона	0,468	1,000	0,432
	Значимость	0,005	-	0,010
Посещение церкви	Коэффициент Пирсона	0,779	0,432	1
	Значимость	0,000	0,010	-

Корреляция характеризуется:

1)Величиной зависимости;

2)Надежностью (истинностью) зависимости (насколько можно распространить полученную на выборке величину зависимости на генеральную совокупность).

Надежность показывает, насколько вероятно, что зависимость будет вновь обнаружена (подтвердится) на данных другой выборки, извлеченной из той же популяции.

Если исследование удовлетворяет некоторым специальным критериям, то надежность найденных зависимостей между переменными выборки можно количественно оценить и представить с помощью стандартной статистической меры, p-уровень, или статистический уровень значимости.

Надежность зависимости 10

Если исследование удовлетворяет некоторым специальным критериям, то надежность найденных зависимостей между переменными выборки можно количественно оценить и представить с помощью стандартной статистической меры, p-уровень, или статистический уровень значимости.

P- уровень – это показатель, находящийся в убывающей зависимости от надежности результата. P-уровень представляет вероятность ошибки, связанной с распространением наблюдаемого результата на всю генеральную совокупность.