## Тема 7. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ.

Любой закон природы или общественного развития может быть представлен описанием совокупности взаимосвязей. Если эти зависимости стохастичны, а анализ осуществляется по выборке из генеральной совокупности, то данная область исследований относится к задачам статистического исследования зависимостей, которые включают в себя *корреляционный и регрессионный анализ*.

**Корреляция** (от лат. *Correlatio* – соотношение, взаимосвязь), ***корреляционная зависимость*** – статистическая взаимосвязь двух или нескольких случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Математической мерой корреляции двух случайных величин служит ***корреляционное отношение*** $\eta$, либо ***коэффициент корреляции*** ***r***.

Впервые в научный оборот термин «корреляция» ввёл французский палеонтолог *Жорж Кювье* в XVIII веке. Он разработал «закон корреляции» частей и органов живых существ, с помощью которого можно восстановить облик ископаемого животного, имея в распоряжении лишь часть его останков. В статистике слово «корреляция» первым стал использовать английский биолог и статистик *Фрэнсис Гальтон* в конце XIX века.

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами X и Y. В качестве меры такой связи используется ***коэффициент корреляции***. Коэффициент корреляции оценивается по выборке объема *n* связанных пар наблюдений ($x_i$, $y_i$) из совместной генеральной совокупности X и Y.

*Коэффициент корреляции характеризует силу связи между изучаемыми признаками и дает представление о ее направленности. Величина коэффициента корреляции изменяется от -1 (строгая обратная зависимость) до 1 (строгая прямая зависимость). При значении 0 зависимости между двумя выборками нет.*

***Чем ближе модуль коэффициента корреляции к единице, тем сильнее или глубже корреляционная взаимосвязь между двумя переменными. Модульное значение выше 0,8 характеризуют сильную взаимосвязь, в интервале 0,8-0,5 – выраженную взаимосвязь, 0,5-0,2 – слабую взаимосвязь, менее 0,2 (0,2 – 0) – отсутствие взаимосвязи*** (рис. 7.1)**.**

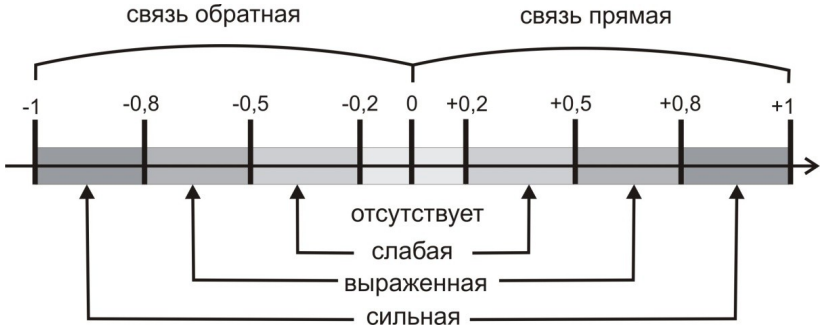


Рис. 7.1. Схема оценки силы корреляционной связи по величине коэффициента корреляции.

Таким образом, ***задача корреляционного анализа*** сводится к *установлению направления (прямая или обратная) и формы (линейная, нелинейная) связи* между варьирующими признаками, *измерению ее тесноты*, и, наконец, к *проверке уровня значимости полученных коэффициентов корреляции*.

В случае несгруппированной совокупности может быть получено наглядное представление о наличии или отсутствии корреляции путем построения диаграммы рассеяния (рис. 7.2).

***Диаграмма рассеяния*** визуализирует зависимость между двумя переменными X и Y. Данные изображаются точками в двумерном пространстве, где оси соответствуют переменным (*X* - горизонтальной, а *Y* - вертикальной оси).

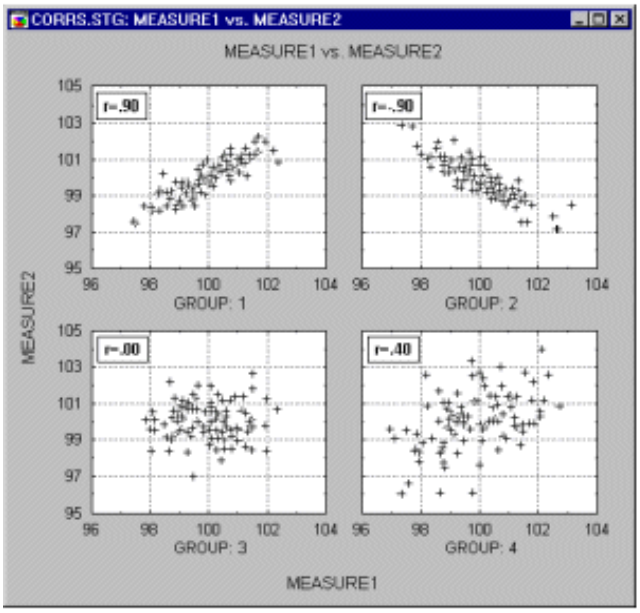


Рис. 7.2. Диаграммы рассеяния для различных значений коэффициентов корреляции

Вытянутость корреляционного поля в диагональном направлении свидетельствует о наличии корреляции между обоими признаками. Если число вариант велико, то корреляционное поле часто имеет вид более или менее правильного эллипса со сгущением точек в центре и сравнительно редким их расположением на периферии; отклонение осей эллипса от координатных направлений указывает на наличие корреляции.

Существует несколько типов коэффициентов корреляции, применение которых зависит от используемой шкалы измерения величин X и Y (табл. 7.1).

Таблица 7.1.

| **Типы шкал** | | **Мера связи** |
|---|---|---|
| ***Переменная X*** | ***Переменная Y*** | |
| Интервальная (или отношений) | Интервальная (или отношений) | Коэффициент Пирсона |
| Ранговая, интервальная (или отношений) | Ранговая, интервальная (или отношений) | Коэффициент Спирмена |
| Ранговая | Ранговая | Коэффициент Кендалла |
| Дихотомическая | Дихотомическая | Коэффициент φ (фи), четырёхполевая корреляция |
| Дихотомическая | Ранговая | Рангово-бисериальный коэффициент |
| Дихотомическая | Интервальная или отношений | Бисериальный коэффициент |
| Интервальная | Ранговая | Не разработан |

Для оценки степени взаимосвязи переменных, измеренных в количественных шкалах, используется коэффициент линейной корреляции (***коэффициент Пирсона***), предполагающий, что выборки X и Y распределены по нормальному закону.

***Коэффициент корреляции Пирсона*** — параметр, характеризующий *степень линейной взаимосвязи* между двумя выборками, рассчитывается по формуле:

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}.$$

В случае работы с данными, распределение которых отлично от нормального, необходимо пользоваться ранговыми методами – вычислять ***коэффициент корреляции Спирмена*** (непараметрический аналог коэффициента Пирсона для интервальных и порядковых переменных), ***коэффициент корреляции Кендалла*** (для порядковых переменных).

***Коэффициент ранговой корреляции Спирмена***

Каждому показателю X и Y присваивается ранг. На основе полученных рангов рассчитываются их разности *d* и вычисляется коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Величина коэффициента корреляции Спирмена также лежит в интервале +1 и -1. Он, как и коэффициент Пирсона, может быть положительным и отрицательным, характеризуя направленность связи между двумя признаками, измеренными в ранговой шкале.

***Коэффициент ранговой корреляции Кендалла***

Применяется для выявления взаимосвязи между количественными или качественными показателями, если их можно ранжировать. Значения показателя X выставляют в порядке возрастания и присваивают им ранги. Ранжируют значения показателя Y и рассчитывают коэффициент корреляции Кендалла:

$$\tau = \frac{2S}{n(n - 1)},$$

где $S = P - Q$, P – суммарное число наблюдений, следующих за текущими наблюдениями с большим значением рангов Y,

Q – суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y (равные ранги не учитываются).

*Важно отметить, что установление корреляции между признаками само по себе еще не дает оснований делать какие-либо заключения о причинно-следственных связях между ними.*

**Вычисление ошибки коэффициента корреляции.**

1. Ошибка коэффициента корреляции, вычисленного методом квадратов (Пирсона):

$$m_{rxy} = \sqrt{\frac{1 - r_{xy}^2}{n - 2}},$$

где $r_{xy}$ – коэффициент корреляции, *n* – объем выборки.

2. Ошибка коэффициента корреляции, вычисленного ранговым методом (Спирмена):

$$m_{pxy} = \sqrt{\frac{1 - p_{xy}^2}{n - 2}},$$

где $p_{xy}$ – коэффициент корреляции, *n* – объем выборки.

**Оценка статистической значимости коэффициента корреляции.**

Значимость коэффициента корреляции определяется с помощью статистики:

$$t = \frac{r_{xy}}{mr_{xy}}$$

или

$$t = \frac{p_{xy}}{mp_{xy}}$$

Критерий *t* оценивается по таблице значений *t* с учетом числа степеней свободы *(n–2)*, где *n* – число парных вариант. Критерий *t* должен быть равен или больше табличного, соответствующего точности оценки данных ≥95%.

**Регрессионный анализ –** статистический метод исследования влияния одной или нескольких независимых переменных $X_1, X_2, ..., X_p$ на зависимую переменную $Y$. Независимые переменные иначе называют *регрессорами* или *предикторами*, а зависимые переменные – *критериальными*. Терминология *зависимых* и *независимых* переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

***Цели регрессионного анализа:***

1. Определение степени детерминированности вариации зависимой переменной предикторами (независимыми переменными)
2. Предсказание значения зависимой переменной с помощью независимой(-ых).
3. Определение вклада отдельных независимых переменных в вариацию зависимой переменной.

Суть регрессионного анализа сводится к установлению уравнения регрессии, т.е. вида кривой между случайными величинами X и Y, оценке тесноты связей между ними, достоверности и адекватности результатов измерений.

Наиболее часто для описания статистической связи признаков используется *линейная форма*. Внимание к линейной связи объясняется четкой интерпретацией ее параметров, ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют (путем логарифмирования или замены переменных) в линейную форму.

В случае линейной парной связи ***линия регрессии –*** прямая линия, строится с использованием уравнения регрессии, имеющего вид. (рис 7.3):

$\boldsymbol{Y = aX + b}$, где

$\boldsymbol{X}$ – факторный признак (независимая переменная),

$\boldsymbol{Y}$ – критериальный признак (зависимая переменная),

$\boldsymbol{a}$ и $\boldsymbol{b}$ – числовые параметры уравнения.

Параметры данного уравнения *a* и *b* чаще всего оцениваются с помощью ***метода наименьших квадратов***, суть которого заключается в том, что сумма квадратов расстояний от точек на диаграмме рассеяния до линии регрессии минимальна.

Прямая линия дает наилучшее приближенное описание линейной зависимости между двумя переменными.

Точность оценки регрессии – ***коэффициент детерминации*** $R^2$ (оценивает % вариации данных вокруг среднего значения, который может быть объяснен с помощью выбранного уравнения регрессии).

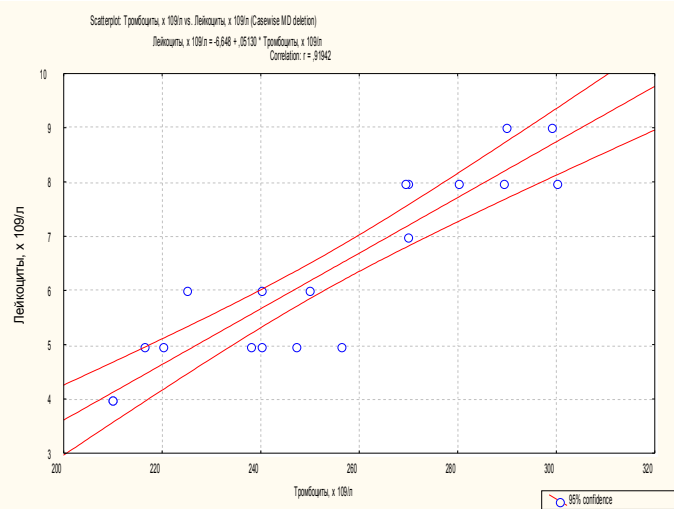


Рис. 7.3. Линия регрессии

Коэффициент детерминации принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50% (в этом случае коэффициент корреляции превышает по модулю 70%). Модели с коэффициентом детерминации выше 80% можно признать достаточно хорошими (коэффициент корреляции превышает 90%). Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.

## Контрольные вопросы

1. В чем суть корреляционного анализа?
2. В чем суть регрессионного анализа?
3. Что характеризует коэффициент корреляции? В каких пределах он находится?
4. Что такое корреляционное поле?
5. Как рассчитываются ошибка коэффициента корреляции?
6. Что такое уравнение регрессии?
7. Что такое коэффициент детерминации?

## Список литературы


1. Гланц С. Медико-биологическая статистика. Пер. с англ. – М.: Практика, 1998. – 459 с.
2. Лях Ю.Е., Гурьянов В.Г., Хоменко В.Н., Панченко О.А. Основы компьютерной биостатистики: анализ информации в биологии, медицине и фармации статистическим пакетом Medstat. – Донецк: Папакица Е.К., 2006. – 214 с.
3. Островок здоровья. – Режим доступа: www.bono-esse.ru
4. Петри А., Сэбин К. Наглядная статистика в медицине. – М.: Издательский дом ГЭОТАР-МЕД, 2003. – 139 с.
5. Платонов А.Е. Статистический анализ в медицине и биологии: задача, терминология, логика, компьютерные методы. – М.: Издательство РАМН, 2000. – 52 с.
6. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: МедиаСфера, 2002. – 312 с.

7. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. - М.: ГЭОТАР-МЕД, 2001. – 256 с.