ОГЛАВЛЕНИЕ

ПРИЛОЖЕНИЯ

Глава V. АНАЛИЗ ВЗАИМОСВЯЗЕЙ

5.1. Функциональная, корреляционная и стохастическая зависимость

Одной из важнейших проблем, с которой нередко встречаются исследователи биологии и медицины, является выявление связи между явлениями. Большинство взаимосвязанных явлений, которые изучает физика и математика, имеют функциональную связь, когда одна величина детерминированно зависима от других значений. Например, расстояние в 20 км из дома на работу вами преодолено в течение 30 мин, что соответствует средней скорости 40 км/ч. Последний показатель детерминирован формулой для вычисления средней скорости движения v = S/t, которая является типичным примером функциональной связи между показателями. И если пробки на дорогах обусловят более медленное передвижение и вы доберетесь до рабочего места за 50 мин, то показатель средней скорости окажется меньше и его также можно легко вычислить, используя представленную выше математическую формулу (24 км/ч). Таким образом, связь между переменными X и Y называют функциональной, если одному значению переменной X соответствует единственное значение переменной Y.

В явлениях, которые изучает биология и медицина, как правило, нет таких математически детерминированных связей, как показано в представленном выше примере, поскольку большинство величин случайные. Однако между многими случайными величинами, находящимися в сфере интересов биологов и врачей, без особого труда прослеживается более или менее оформленная связь. Возможно, например, предположить наличие связи между ростом человека и его весом, и такое предположение не противоречит здравому смыслу. Однако это предположение лишь тенденция, а не функциональная связь, поскольку при реализации подобного исследования можно обнаружить немало невысоких толстяков и высоких худых людей. Подобную зависимость между двумя случайными величинами, с помощью которой с большей или меньшей точностью удается выявить лишь тенденцию к увеличению или уменьшению, называют корреляционной. Такая связь определяет, насколько значения двух переменных соответствуют друг другу.

О стохастической зависимости принято говорить, когда каждому значению переменной X, вследствие влияния большого числа неконтролируемых факторов, может соответствовать множество значений переменной Y.

5.2. Коэффициент линейной корреляции (Correlation analysis)

Британский математик Пирсон в начале прошлого века рассчитал корреляционную зависимость и представил коэффициент, определяющий величину корреляционной связи двух показателей. Разработанный этим математиком показатель называют коэффициентом корреляции Пирсона, а также коэффициентом линейной корреляции. Формула его вычисления в общем виде представлена ниже:

$$r_{pearson} = \frac{\sum_{i=1}^{n}(X1i - \overline{X1}) \times (X2i - \overline{X2})}{\sqrt{\sum_{i=1}^{n}(X1i - \overline{X1})^2 \times \sum_{i=1}^{n}(X1i - \overline{X1})^2}}.$$

Коэффициент корреляции может принимать значения от −1 до 1. Корреляционную связь называют положительной, если при увеличении одного показателя имеется более или менее выраженная тенденция к увеличению другого. И наоборот, коэффициент корреляции имеет отрицательный знак, если при увеличении одного показателя есть тенденция к уменьшению другого.

На практике коэффициент корреляции, равный 1,0 или −1,0, получить практически невозможно. При таком коэффициенте связь становится функциональной. Чем ближе коэффициент корреляции к нулю, тем менее выражена зависимость между изучаемыми показателями. Например, корреляционная зависимость между весом и ростом (за счет наличия большого числа людей среднего роста с большим весом и немалого числа высоких худых лиц) равна лишь 0,3.

Выраженность корреляционной зависимости следует различать по величине коэффициента корреляции.

0–0,19 — корреляционная зависимость выражена очень слабо. В подавляющем большинстве ситуаций такой слабой корреляционной зависимостью следует пренебречь.

0,20–0,39 — слабая корреляционная зависимость. Демонстрировать такие значения корреляционной связи в научной работе возможно, но делать серьезные выводы не следует.

0,40–0,59 — умеренная корреляционная зависимость.

0,60–0,79 — хорошая корреляционная зависимость.

0,80–0,99 — сильная корреляционная связь. Такой выраженной связи следует уделить особое внимание в исследовании.

1,0 — функциональная (математически детерминированная) связь.

Коэффициент корреляции является лишь приближенным значением истинного параметра. Поэтому, вычисляя коэффициент корреляции, следует учитывать лишь те значения, вероятность которых меньше 0,05.

Коэффициент корреляции Пирсона является показателем, эффективно работающим при нормальном распределении выборок, между которыми определяют связь. Поэтому, прежде чем его применить, нужно проверить гипотезу о нормальности распределения анализируемых данных в изучаемых выборках.

5.3. Коэффициенты ранговой корреляции

Для выборок, в которых распределение отличается от нормального, следует применять коэффициенты корреляции, вычисляющие силу связи после преобразования количественных значений в ранги. Такие показатели являются непараметрическими аналогами линейного коэффициента корреляции. Кроме того, для анализа связи качественных значений, представленных в баллах, также необходимо использовать коэффициенты ранговой корреляции.

Существуют несколько подобных коэффициентов. Наиболее часто для вычисления коэффициентов ранговой корреляции применяют методы Спирмена (*Spearman*) и Кендалла (*Kendall*).

Наиболее близким непараметрическим аналогом коэффициента r Пирсона является коэффициент корреляции R Спирмена. При его вычислении учитывается возрастание или убывание парных ранговых переменных. Таким образом, коэффициент ранговой корреляции представляет собой меру совпадения рангов двух изучаемых признаков. Для вычисления коэффициента корреляции Спирмена подменяют фактические значения ранговыми. Равным значениям присваивают одинаковый ранг — среднее значение этих двух рангов.

Для вычисления коэффициента корреляции Спирмена используют формулу:

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

где d — разность между рангами сопряженных значений двух признаков.

При независимом варьировании признаков относительно друг друга коэффициент ранговой корреляции близок нулю. Если будет просматриваться тенденция к увеличению одного значения при возрастании парного, выявляется положительный коэффициент корреляции Спирмена, а при тенденции к уменьшению — отрицательный коэффициент.

В отличие от подхода Спирмена, коэффициент корреляции Кендалла имеет другой принцип вычисления, основанный на вероятности того, что наблюдаемые данные имеют определенный порядок.

5.4. Регрессионный анализ

Этот вид анализа следует применить, если необходимо определить связь между двумя явлениями в виде уравнения, т. е. регрессионный анализ будет эффективно дополнять выявленную корреляционную зависимость. Такая необходимость также возникает при прогнозировании величины показателя или математическом моделировании какого-либо процесса.

Для того чтобы понять суть регрессионного анализа, следует вспомнить из курса математики средней школы уравнение прямой линии. Любое уравнение прямой линии можно выразить в виде формулы $y = ax + b$, где a — это тангенс угла наклона прямой линии, b — точка, в которой прямая пересекает ось ординат (рис. 7). Регрессионный анализ

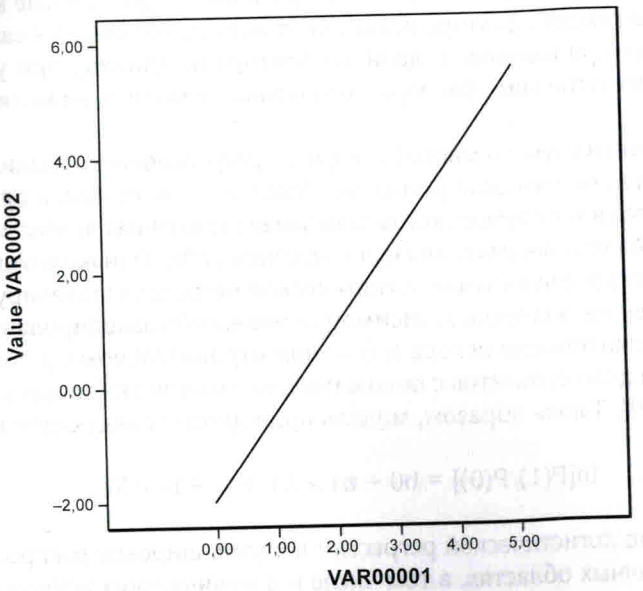

Рис. 7. Функциональная зависимость показателей X и Y на примере уравнения прямой линии Y = 1,5X – 2 (график построен программой SPSS)

позволяет построить прямую и описать ее в виде формулы прямой линии $y = ax + b$. При этом значение a называют коэффициентом наклона, а значение b — коэффициентом сдвига.

Как следует из представленного рисунка, угол наклона прямой равен 48°, прямая пересекает ось ординат в значении –2. Зная величину одного показателя, по формуле можно вычислить значение другого, поскольку любое значение Y детерминировано значением X ($x = 0$, $y = -2$; $x = 1$, $y = -0{,}5$; $x = 2$, $y = 1$ и т. д.). Регрессионный анализ позволяет сделать подобное, однако в этом случае возможно лишь прогнозировать, а не точно определять зависимый показатель. Так, после выполнения регрессионного анализа, имея показатели, предоставляется возможность прогнозировать связанные. Точность прогноза будет тем выше, чем выше коэффициент корреляции, поэтому желательно применять этот вид статистической обработки данных вместе с вычислением коэффициента корреляционной зависимости. Коэффициенты регрессии (a, b) имеют свои ошибки, поскольку они вычисляются на основе выборочных данных.

Если в регрессии анализируется взаимосвязь зависимой переменной (y) с несколькими факторами (x), такой анализ называется многофакторным регрессионным анализом. Полученные регрессионные коэффициенты для каждого фактора показывают, насколько меняется зависимая переменная при изменении данного фактора на единицу при условии, при этом все остальные факторы, включенные в модель, не меняют свои значения.

Частным случаем многофакторного регрессионного анализа является метод логистической регрессии. Модель логистической регрессии используется в тех случаях, когда зависимая переменная является бинарной, т. е. может принимать значения «да/нет» (1/0). Однако в отличие от обычной регрессии в случае логистической регрессии моделируется не само численное значение зависимой переменной (закодированной как 1 при положительном исходе и 0 — при отрицательном), а логарифм отношения доли субъектов с положительным и отрицательным исходом ln[P(1)/P(0)]. Таким образом, модель приобретает следующий вид:

$$\ln[P(1)/P(0)] = b0 + b1 \times X1 + \ldots + bi \times Xi".$$

Метод логистической регрессии получил широкое распространение в различных областях, в том числе и в медицинских исследованиях при изучении бинарных исходов (например, пациент жив/умер, инфаркт миокарда наблюдался / не наблюдался за время исследования и т. д.). Несмотря на более сложную реализацию этого метода по сравнению

с другими методами, применяющимся к категориальным данным (например, хи-квадрат), он позволяет дать количественную оценку взаимосвязи каждого изученного фактора с зависимой переменной с учетом поправки на другие факторы,другими словами, получить более чистую оценку эффекта (например, различия между двумя группами лечения) с учетом нивелировки различий по другим факторам.

Глава VI. АНАЛИЗ ВЫЖИВАЕМОСТИ

Статистические методы, известные сегодня как анализ выживаемости, получили свое название вследствие их изначально широкого применения в медицинских исследованиях для оценки продолжительности жизни при изучении эффективности методов лечения. Сегодня эти методы используются в социальной и страховой сферах и относятся к анализу любых данных типа «время до наступления события». Главная цель анализа выживаемости — оценить время до интересующего события и количественно объяснить, как оно зависит от параметров лечения, индивидуальных особенностей пациентов и других независимых переменных. Анализ выживаемости изучает процесс наступления определенных (терминальных) событий и распределение времен до наступления этих событий среди всех субъектов, находящихся под наблюдением. При этом под термином «событие» может пониматься любой исследуемый исход, как отрицательный, так и положительный. Это может быть длительность нахождения в стационаре, продолжительность заболевания, период действия иммунной защиты и др. В медицине анализ является практическим инструментом, который помогает отвечать на вопросы такого характера, как воздействие различных факторов на выживание, прогнозирование шансов на выживание и др. Для ответа на подобные вопросы необходимо иметь возможность четко определить «время жизни» элемента (период пребывания субъекта в совокупности до наступления исследуемого события). Отличительной особенностью данных типа «время до наступления события» является присутствие неполных наблюдений, в частности, когда терминальное событие для некоторых субъектов не произошло за время наблюдения за данным субъектом. Вместо этого известно лишь, что терминальное событие если даже и произошло, то позже точки окончания наблюдения, т. е. говорят, что данные по выживаемости для этого субъекта цензурированы справа моментом окончания наблюдения за этим субъектом. Например, данные о пациентах с онкологией, которые остались живы на момент окончания пятилетнего периода наблюдения если нас интересует пятилетняя выживаемость), являются цензурированными на момент окончания наблюдения. У них изучаемое событие (исход или в данном случае смерть) не произошло, и неизвестно, когда оно произойдет, поэтому у нас нет точной информации о периоде, прошедшем от постановки диагноза до смерти этих пациентов. Второй вариант цензурированных случаев — это пациенты, которые выпадают из-под

наблюдения (в случае переезда в другой город отказа от наблюдения по иных причин), узнать информацию у этих пациентов об изучаемом событии также не представляется возможным.

6.1. Графическое представление данных об изучаемом событии

На рис. 8 показан ход исследования. Время наблюдения за пациентом представлено горизонтальным отрезком. Левый конец отрезка — это начало наблюдения. На правом конце — черный или белый кружок. Черный кружок означает, что пациент умер (или произошло иное терминальное событие, изучаемое в данном исследовании) и, таким образом, время от начала наблюдения до наступления терминального события нам точно известно. Белый кружок означает, что исследование закончилось до его смерти либо он выбыл из-под наблюдения. Относительно выбывших нам известно только, что они прожили не меньше определенного срока (от начала до момента окончания наблюдения), далее их исход неизвестен.

Все исследования выживаемости должны удовлетворять требованиям для всех субъектов исследования: известно время начала и окончания наблюдения, а также следующие сведения: умер, остался жив или выбыл из исследования по другой причине на момент окончания наблюдения.

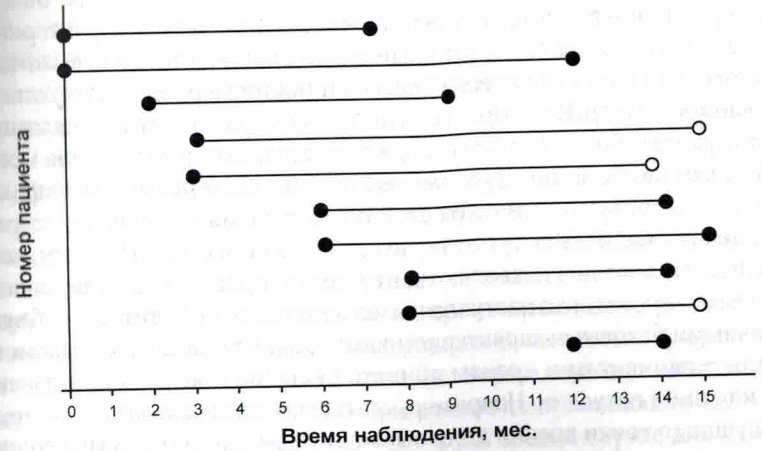

РИС. 8. Общее представление данных наблюдения за пациентами при анализе времени до наступления события

Таким образом, данные по выживаемости состоят из двух компонентов: 1) времени до наступления события, т. е. числа месяцев, дней, часов, минут или иных единиц, прошедших от начала наблюдения за субъектом до момента наступления события либо окончания наблюдения за субъектом, если событие не произошло; 2) цензурирования, который принимает значение 1 в том случае, если событие произошло, и 0 — если не произошло. При этом под «событием» может пониматься не только отрицательный (смерть, инфаркт, инсульт), но и положительный исход (выздоровление, выписка из стационара, полный или частичный ответ на терапию), а также любая комбинация нескольких исходов (например, смерть от сердечно-сосудистого заболевания + инсульт + инфаркт + операция).

6.2. Методы анализа данных типа «время до наступления события»

Кривая, описывающая полученные таким образом данные, называется кривой выживаемости, а соответствующая ей функция — функцией выживания, которая с математической точки зрения представляет собой вероятность того, что время жизни субъекта будет больше заданного времени *t*. Существует несколько статистических методов, которые позволяют провести оценку функции выживания. Наиболее популярными из них являются построение таблиц времен жизни (*life-table*), метод Каплана — Мейера (*Kaplan — Meier analysis*), а также регрессия Кокса (*Cox regression*). С помощью этих методов можно, во-первых, описать цензурированные данные, получив важные статистические характеристики кривых выживания, такие как медиана выживаемости, доля выживших пациентов за заданный интервал времени (например, одно-, двухлетняя выживаемость) и др. Во-вторых, они позволяют сравнивать выживаемость в двух группах и более. В-третьих, при помощи некоторых методов можно оценить взаимосвязь между временем выживания и различными факторами, характеризующими субъектов исследования (например, возраст, пол пациента, наличие сопутствующих заболеваний и др.). Наконец, некоторые из этих методов позволяют дать оценку (провести моделирование) ожидаемого времени до наступления исследуемого события для субъекта с заданными базовыми характеристиками. Знание кривых выживаемости позволяет пациентам и врачам принять решение о назначении лечения в той или иной ситуации. Например, краткосрочная выживаемость может быть лучше (с точки зрения доли выживших) на одном режиме терапии, тогда как долгосрочная выживаемость имеет положительный результат на другом режиме. Кривые выживаемости дают важную информацию о том,

что следует ожидать в каждом конкретном случае. Естественно, фактические времена жизни могут отличаться для отдельных индивидуумов от ожидаемых значений, но эти различия будут не слишком велики, если кривые выживаемости были построены на валидных и надежных данных.

6.3. Построение таблиц времен жизни

Наиболее естественным способом описания выживаемости в выборке является построение таблиц времен жизни. Такую таблицу можно рассматривать как «расширенную» таблицу частот. Весь диапазон возможных времен наступления исследуемого события разбивается на некоторое число равных интервалов, например на временные отрезки по 3 мес. Для каждого интервала вычисляется число и доля пациентов: живых в начале рассматриваемого интервала , умерших в данном интервале и выбывших из-под наблюдения, т. е. цензурированных в данном интервале. При этом предполагается, что пациенты выбыли в середине интервала данного наблюдения, и это является наиболее консервативной из возможных оценок. Таким образом, для каждого интервала наблюдения доля выживших пациентов (p) может быть рассчитана как

$$p = \frac{n - d - c/2}{n - c/2} = 1 - \frac{d}{n - c/2},$$

где n — число пациентов, которые были живы в начале данного интервала; d — число пациентов, которые умерли в данном интервале времени; c — число пациентов, которые выбыли из-под наблюдения в данном интервале (c берется с коэффициентом 0,5, так как мы предполагаем, что цензурирование произошло в середине интервала). В этой формуле числитель представляет собой число выживших пациентов, а знаменатель — число пациентов, находящихся под риском в течение данного интервала. Кумулятивная доля выживших (функция выживания) — это количество выживших к началу соответствующего временного интервала с учетом всех предыдущих временных интервалов. Поскольку вероятности выживания на разных интервалах считаются независимыми друг от друга, то кумулятивная доля равна произведению долей выживших субъектов по всем предыдущим интервалам. В частности, кумулятивная доля выживших (функция выживания) S в течение двух последовательных интервалов рассчитывается как произведение доли выживших пациентов на первом и втором интервале: $S = p_1 \times p_2$. В общем случае для k интервалов $S_k = p_1 \times p_2 \times \ldots \times p_k$. График этой функции от времени называется кривой выживаемости. Поскольку для каждого интервала вероятность не