Анализ зависимостей

Цели

1. Освоить понятия зависимых и независимых переменных.
2. Научиться строить уравнение линейной регрессии.
3. Рассчитать коэффициент корреляции.
4. Проверить статистическую значимость связи.
5. Различать доверительные интервалы и интервалы прогноза.

Клинические примеры зависимостей

- Рост и вес пациента
- Доза лекарства и эффект
- Возраст и давление
- Объем опухоли и уровень онкомаркера

Опасность иллюзорных связей

Не каждая видимая связь реальна.

Пример с диуретиком показывает, что визуальные выводы могут быть ошибочными.

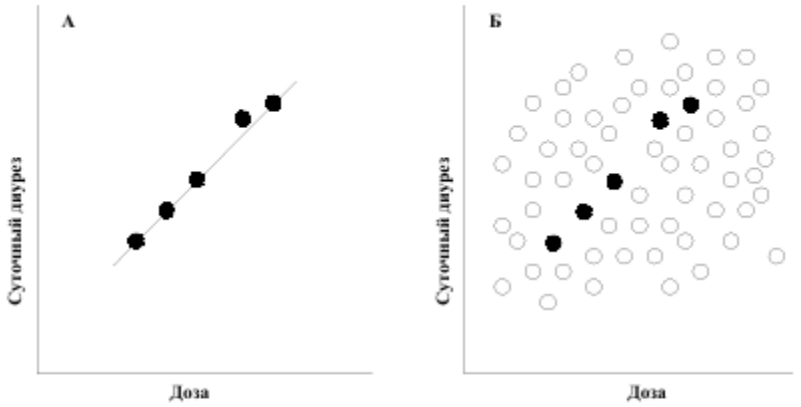

Рис. 1.2. А. У 5 добровольцев измерили суточный диурез после приема разных доз препарата (предполагаемого диуретика). Зависимость диуреза от дозы казалась бы налицо, чем больше доза – тем больше диурез. Можно ли считать диуретический эффект препарата доказанным? **Б.** Такую картину мы увидели бы, если бы могли исследовать связь дозы и диуреза у всех людей: зависимости нет в помине. Пять человек, вошедших в первоначальное исследование, помечены черным. В данном случае мнимая зависимость порождена случайностью. С помощью статистических методов можно оценить вероятность подобной ошибки.

Переменные X и Y

X — независимая переменная (предиктор).

Y — зависимая переменная (переменная отклика).

Пример: рост (X) и вес (Y).

Регрессионный анализ

Регрессия отвечает на вопрос: как именно изменяется Y при изменении X.

Модель: $\hat{y} = a + bX$.

Интерпретация коэффициентов

a — значение Y при X=0 (иногда не имеет клинического смысла).

b — наклон линии, показывает, на сколько изменится Y при изменении X на 1.

Метод наименьших квадратов (МНК)

Наилучшая линия — та, при которой сумма квадратов отклонений $\Sigma(Y - \hat{y})^2$ минимальна.

Остатки (residuals)

Разность между наблюдаемым и предсказанным значением Y.

Используются для проверки качества модели.

Основные предположения модели

1. Линейность
2. Нормальность остатков
3. Гомоскедастичность (равенство дисперсий)
4. Независимость наблюдений

Пример: рост и вес марсиан

Рост (X), см	Вес (Y), г	X^2	XY
31	7,8	961	241,8
32	8,3	1024	265,6
33	7,6	1089	250,8
34	9,1	1156	309,4
35	9,6	1225	336,0
35	9,8	1225	343,0
40	11,8	1600	472,0
41	12,1	1681	496,1
42	14,7	1764	617,4
46	13,0	2116	598,0
$\Sigma X = 369$	$\Sigma Y = 103,8$	$\Sigma X^2 = 13841$	$\Sigma XY = 3930,1$

Пример: рост и вес марсиан

$\hat{y} = -6.0 + 0.44X$

b=0.44 означает, что каждый дополнительный см роста увеличивает вес на 0.44 г.

Доверительный интервал для β

$b \pm t_{(\alpha/2)} * s_b$

Если 95% ДИ не включает 0 → связь статистически значима.

Доверительная область vs интервал прогноза

- Доверительная область — для средней линии.
- Интервал прогноза — для индивидуальных значений (всегда шире).

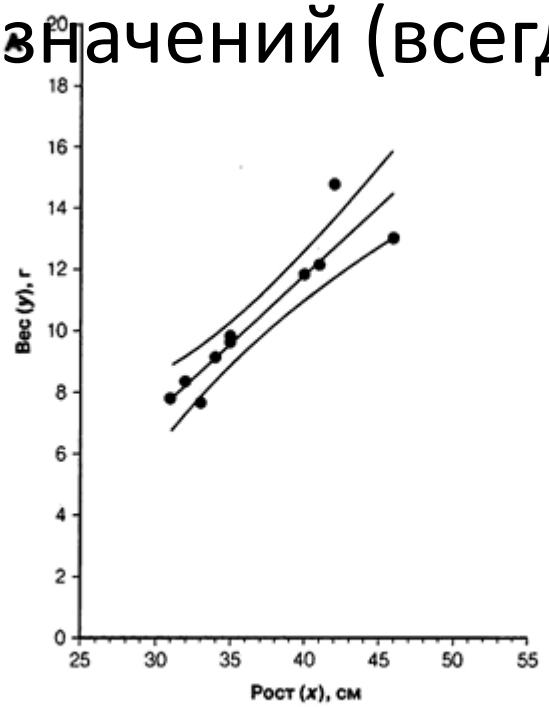

Рис. 8.7. А. 95% доверительная область для *линии регрессии* (по выборке с рис. 8.3).

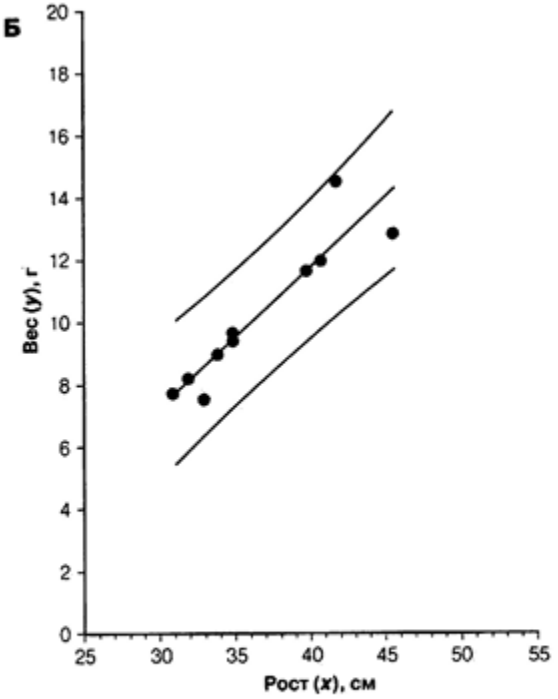

Рис. 8.7. Б. 95% доверительная область для *значений*. Если мы хотим определить вес марсианина по его росту, нам следует воспользоваться именно этой доверительной

Корреляционный анализ

Корреляция отвечает на вопрос: насколько тесно связаны переменные.

r от -1 до +1.

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

Интерпретация коэффициента корреляции

$r \approx +1$ — сильная прямая связь

$r \approx -1$ — сильная обратная связь

$r \approx 0$ — связи нет

Коэффициент детерминации r^2

Показывает долю изменчивости Y, объясняемую изменчивостью X.

Пример: $r=0.925 \rightarrow r^2=0.855 \rightarrow$ 85.5% вариации веса объясняется ростом.

Проверка значимости r

Нулевая гипотеза: $\rho = 0$.

Статистика: $t = r / \sqrt{((1 - r^2)/(n - 2))}$.

Пример расчета r

$r = 99.9 / \sqrt{(224.8 * 51.9)} \approx 0.925$

Сильная положительная связь между ростом и весом.

Пример расчета r

Рост (X), см	Вес (Y), г	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
31	7.8	-5.9	-2.58	15.222	34.81	6.6564
32	8.3	-4.9	-2.08	10.192	24.01	4.3264
33	7.6	-3.9	-2.78	10.842	15.21	7.7284
34	9.1	-2.9	-1.28	3.712	8.41	1.6384
35	9.6	-1.9	-0.78	1.482	3.61	0.6084
35	9.8	-1.9	-0.58	1.102	3.61	0.3364
40	11.8	3.1	1.42	4.402	9.61	2.0164
41	12.1	4.1	1.72	7.052	16.81	2.9584
42	14.7	5.1	4.32	22.032	26.01	18.6624
46	13.0	9.1	2.62	23.842	82.81	6.8644
$\Sigma X = 369$	$\Sigma Y = 103.8$	$\Sigma(X - \bar{X}) \approx 0$	$\Sigma(Y - \bar{Y}) \approx 0$	$\Sigma(X - \bar{X})(Y - \bar{Y}) = 99.88$	$\Sigma(X - \bar{X})^2 = 224.9$	$\Sigma(Y - \bar{Y})^2 = 51.795$

Коэффициент ранговой корреляции Спирмена

Используется для порядковых данных или при наличии выбросов.

Работает с рангами, а не исходными значениями.

Сравнение двух линий регрессии

Можно сравнить наклоны (b) и константы (a) с помощью t-теста или F-критерия.

Пример: сила и мышечная масса

Зависимость силы от массы различается у здоровых и больных артритом — разные наклоны линий регрессии.

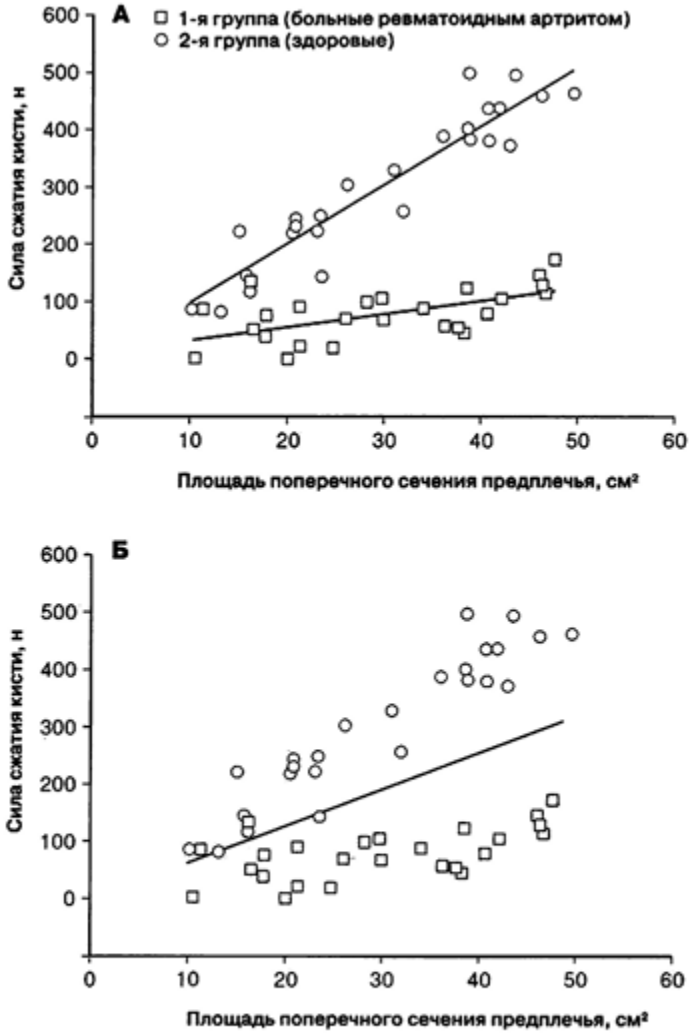

Рис. 8.9. А. Построим линии регрессии для каждой из групп и оценим разброс точек относительно этих линий. **Б.** Объединим группы и найдем линию регрессии для получившейся группы. Если разброс точек относительно этой линии значительно превышает разброс относительно двух отдельных линий, то различия линий следует считать значимыми.

Сравнение методов измерения

Корреляция ≠ согласие.

Используется метод Бланда–Алтмана.

График Бланда–Алтмана

По оси X — среднее двух измерений, по оси Y — их разность.

Оценивается смещение и пределы согласия.

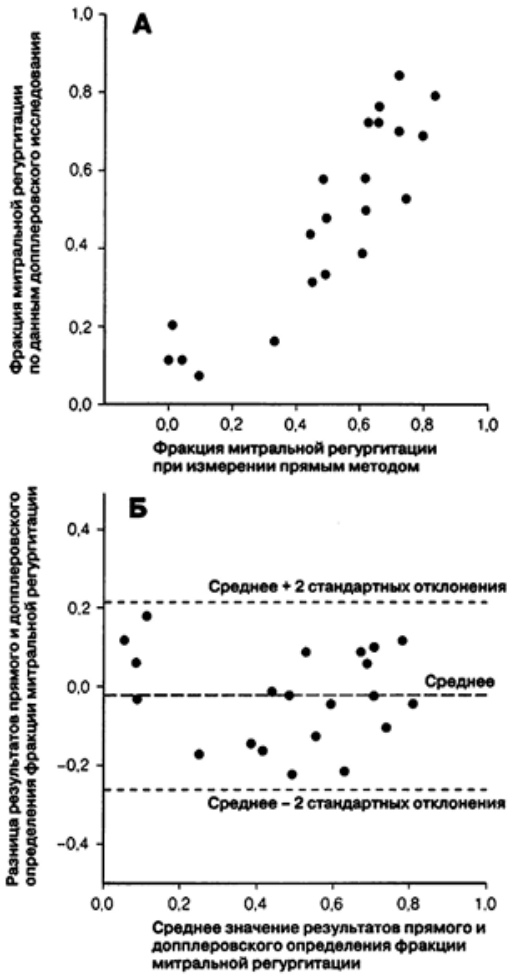

Рис. 8.15. А. Фракция митральной регургитации при измерении прямым методом и по данным допплеровского исследования. **Б.** Сравнение результатов по методу Блэнда—Алтмана.

Типичные ошибки исследователей

1. Корреляция не означает причинность.
2. Неправильный выбор метода.
3. Путаница между доверительным интервалом и прогнозом.