

# Статистика

Конспект практических занятий

Мат-Мех, ПМИ, СМ–СМ

\$Revision: 1.65 \$



# Оглавление

|   |           |
|---|-----------|
| <b>I. Оценки характеристик и параметров распределения</b>           | <b>6</b>  |
| 1. Выборка и эмпирическая случайная величина                        | 7         |
| 2. Виды признаков   | 8         |
| 3. Характеристики распределений и метод подстановки                 | 9         |
| 4. Характеристики распределений и их оценки                         | 10        |
| 4.1. Характеристики положения                                       | 10        |
| 4.2. Характеристики разброса  | 11        |
| 4.3. Анализ характера разброса                                      | 12        |
| 4.4. Характеристики зависимости                                     | 13        |
| 5. Точечная оценка параметров распределения                         | 14        |
| 5.1. Метод моментов   | 14        |
| 5.2. Метод оценки максимального правдоподобия                       | 14        |
| 6. Свойства оценок  | 16        |
| 6.1. Несмещенность  | 16        |
| 6.2. Состоятельность  | 17        |
| 6.3. Асимптотическая нормальность                                   | 19        |
| 6.4. Эффективность  | 20        |
| 7. Проверка оценок на эффективность                                 | 22        |
| 8. Построение эффективных оценок                                    | 24        |
| <b>II. Некоторые распределения, связанные с нормальным</b>          | <b>25</b> |
| 1. Распределение $\chi^2(m)$  | 26        |
| 2. Распределение Стьюдента $t(m)$                                   | 27        |
| 3. Распределение Фишера   | 28        |
| 4. Квадратичные формы от нормально распределенных случайных величин | 29        |
| <b>III. Проверка гипотез</b>  | <b>30</b> |
| 1. Построение критерия  | 32        |
| 1.1. Понятие гипотезы и критерия                                    | 32        |
| 1.2. Построение оптимальных критериев                               | 33        |
| 1.2.1. Явный вид оптимальных критериев                              | 34        |
| 1.2.2. О постановке $H_0$   | 35        |
| 1.2.3. Об отвержении гипотезы                                       | 35        |

|  |           |
|--|-----------|
| 1.3. Построение критерия при помощи статистики критерия . . . . .  | 36        |
| 1.4. Разбиение на доверительную и критические области . . . . .  | 36        |
| 1.4.1. Простая альтернатива . . . . .  | 36        |
| 1.4.2. Односторонний критерий (сложная альтернатива) . . . . .   | 37        |
| 1.4.3. Двусторонний критерий . . . . .   | 37        |
| 1.5. Схема построения критерия с помощью статистики . . . . .  | 38        |
| <b>2. Проверка гипотезы о значении мат. ожидания (<math>t</math>-критерий)</b>   | <b>41</b> |
| 2.1. $D\xi = \sigma^2 < \infty$ . . . . .  | 41        |
| 2.2. $D\xi$ неизвестна . . . . .   | 41        |
| 2.2.1. $z$ -критерий для пропорции в модели Бернулли . . . . .   | 42        |
| <b>3. Проверка гипотезы о значении дисперсии в нормальной модели (критерий <math>\chi^2</math>)</b>                    | <b>43</b> |
| 3.1. $E\xi = \mu < \infty$ . . . . .   | 43        |
| 3.2. $E\xi$ неизвестно . . . . .   | 43        |
| <b>4. Критерий <math>\chi^2</math> согласия с видом распределения</b>  | <b>45</b> |
| 4.1. Распределение с известными параметрами . . . . .  | 45        |
| 4.2. Распределение с неизвестными параметрами . . . . .  | 46        |
| 4.3. Согласие с нормальным распределением по $\chi^2$ . . . . .  | 47        |
| <b>5. Критерий Колмогорова-Смирнова согласия с видом распределения</b>   | <b>48</b> |
| 5.1. Произвольное абсолютно непрерывное распределение . . . . .  | 48        |
| 5.2. Нормальное распределение . . . . .  | 48        |
| <b>6. Критерий типа <math>\omega^2</math></b>  | <b>50</b> |
| <b>7. Визуальное определение согласия с распределением</b>   | <b>51</b> |
| 7.1. P-P plot . . . . .  | 51        |
| 7.2. Q-Q plot . . . . .  | 51        |
| <b>8. Гипотеза о равенстве распределений</b>   | <b>52</b> |
| 8.1. Двухвыборочный тест Колмогорова–Смирнова . . . . .  | 52        |
| <b>9. Равенство математических ожиданий для независимых выборок</b>  | <b>53</b> |
| 9.1. Двухвыборочный $t$ -критерий . . . . .  | 53        |
| 9.1.1. Двухвыборочный $t$ -критерий для независимых выборок с $\sigma_1^2 \neq \sigma_2^2$ (Welch $t$ -test) . . . . . | 53        |
| 9.1.2. Двухвыборочный $t$ -критерий для независимых выборок с $\sigma_1^2 = \sigma_2^2$ (pooled $t$ -test) . . . . .   | 54        |
| 9.1.3. Испытания Бернулли . . . . .  | 54        |
| 9.2. Непараметрический $t$ -критерий . . . . .   | 55        |
| 9.3. Критерии суммы рангов Wilcoxon . . . . .  | 55        |
| 9.4. Критерий Mann-Whitney ( $U$ test) . . . . .   | 55        |
| 9.5. Критерий серий (runs) . . . . .   | 56        |
| <b>10. Равенство математических ожиданий для парных (зависимых) выборок</b>  | <b>57</b> |
| 10.1. $t$ -критерий . . . . .  | 57        |
| 10.2. Непараметрический тест знаков (Sign test) . . . . .  | 57        |
| 10.3. Непараметрический критерий (Paired Wilcoxon; Wilcoxon signed-rank test) . . . . .                                | 58        |
| <b>11. Равенство дисперсии для двух распределений</b>  | <b>59</b> |
| 11.1. Критерий Фишера . . . . .  | 59        |
| 11.2. Критерий Левена (Levene's test) . . . . .  | 59        |

|   |           |
|---|-----------|
| 11.3. Критерий Brown–Forsythe . . . . .   | 59        |
| <b>IV. Доверительное оценивание</b>   | <b>60</b> |
| 1. Мотивация и определение  | 61        |
| 2. Доверительные интервалы для математического ожидания и дисперсии в нормальной модели               | 62        |
| 2.1. Доверительный интервал для $\mu$ . . . . .   | 62        |
| 2.2. Доверительный интервал для $\sigma^2$ . . . . .  | 62        |
| 3. Асимптотический доверительный интервал для математического ожидания в модели с конечной дисперсией | 64        |
| 4. Асимптотический доверительный интервал для параметра на основе MLE                                 | 65        |
| 5. Доверительный интервал для проверки гипотезы о значении параметра                                  | 66        |
| 6. Использование SE для построения доверительных интервалов   | 67        |
| 7. Доверительный интервал для двумерного параметра  | 68        |
| <b>V. Корреляционный анализ</b>   | <b>69</b> |
| 1. Вероятностная независимость  | 71        |
| 1.1. Визуальное определение независимости . . . . .   | 71        |
| 1.2. Критерий независимости $\chi^2$ . . . . .  | 71        |
| 2. Линейная / полиномиальная зависимость  | 73        |
| 2.1. О соотношении $\rho$ и коэффициента линейной регрессии . . . . .                                 | 73        |
| 2.2. Значимость коэффициента корреляции . . . . .   | 74        |
| 3. Метод наименьших квадратов (Ordinary Least Squares)  | 75        |
| 4. Корреляционное отношение   | 76        |
| 4.1. Свойства корреляционного отношения . . . . .   | 76        |
| 4.2. Выборочное корреляционное отношение . . . . .  | 77        |
| 5. Частная корреляция   | 78        |
| 6. Зависимость между порядковыми признаками   | 80        |
| 6.1. Ранговый коэффициент Спирмана . . . . .  | 80        |
| 6.1.1. Согласованность $\rho$ и $\rho_S$ . . . . .  | 81        |
| 6.2. Ранговый коэффициент Кэндалла $\tau(\xi, \eta)$ . . . . .  | 82        |
| 7. Корреляционные матрицы   | 83        |
| <b>VI. Дисперсионный анализ</b>   | <b>84</b> |
| 1. Однофакторный дисперсионный анализ (One-way ANOVA <sup>1</sup> )                                   | 85        |

---

<sup>1</sup>ANalysis Of VAriation

|   |            |
|---|------------|
| <b>2. Множественные сравнения</b>   | <b>87</b>  |
| 2.1. Single   | 88         |
| 2.2. Stepdown (Holm's algorithm)  | 88         |
| 2.2.1. Частный случай   | 89         |
| <b>3. ANOVA Post-Hoc Comparison</b>   | <b>90</b>  |
| 3.1. Least Significant Difference (LSD)   | 90         |
| 3.2. Распределение размаха  | 90         |
| 3.3. Tukey's Honest Significat Difference (HSD) Test                                  | 91         |
| 3.4. Другие критерии  | 91         |
| 3.5. Scheffé's Method   | 92         |
| 3.6. Сравнение мощностей  | 92         |
| <b>VII. Регрессионный анализ</b>  | <b>93</b>  |
| <b>1. Регрессия</b>   | <b>94</b>  |
| <b>2. Парная линейная регрессия</b>   | <b>95</b>  |
| 2.1. Модель линейной регрессии  | 96         |
| 2.2. Доверительные интервалы для $\beta_1$ и $\beta_2$                                | 96         |
| <b>3. Множественная линейная регрессия</b>  | <b>98</b>  |
| 3.1. Псевдо-обратные матрицы  | 98         |
| 3.2. Проекторы на подпространства   | 98         |
| 3.3. Ordinary and Total Least Squares   | 99         |
| 3.4. Свободный член   | 100        |
| 3.5. Стандартизованные признаки   | 100        |
| 3.6. Свойства оценки $\hat{\mathbf{b}}$   | 100        |
| 3.7. Свойства $\hat{\mathbf{b}}^{(c)}$ и $\hat{\mathbf{b}}^{(s)}$                     | 101        |
| 3.8. Сравнение оценок   | 102        |
| 3.9. Оценка $\sigma^2$  | 102        |
| 3.10. Проверка значимости коэффициентов линейной регрессии и доверительных интервалов | 103        |
| 3.10.1. Расстояние Махаланобиса   | 103        |
| 3.10.2. Доверительный эллипсоид   | 104        |
| 3.11. Значимость регрессии  | 104        |
| 3.12. О множественном коэффициенте корреляции и сепрессорах                           | 105        |
| 3.13. Взвешенная регрессия (Weighted Least Squares)                                   | 105        |
| 3.14. Гребневая (Ridge) регрессия   | 106        |
| 3.15. Анализ оценок коэффициентов   | 106        |
| 3.15.1. Корреляция между оценками коэффициентов и избыточность                        | 107        |
| 3.16. Анализ аутлаеров  | 108        |
| 3.16.1. Matrix plot   | 108        |
| 3.16.2. Deleted residuals   | 108        |
| 3.16.3. Studentized residuals   | 109        |
| 3.16.4. Расстояние по Куку  | 109        |
| 3.17. Проверка правильности и выбор модели  | 110        |
| 3.18. Доверительные интервалы   | 110        |
| 3.19. Сведение нелинейной модели к линейной   | 111        |
| 3.20. Другие странные замечания   | 111        |
| <b>A. Свойства условного математического ожидания</b>                                 | <b>112</b> |

Часть I.

# Оценки характеристик и параметров распределения

# 1. Выборка и эмпирическая случайная величина

Пусть  $\xi \sim \mathcal{P}$  — случайная величина с распределением  $\mathcal{P}$ .

**Определение.** Повторной независимой выборкой объема  $n$  (до эксперимента) называется набор

$$\mathbf{x} = (x_1, \dots, x_n), \quad x_i \sim \mathcal{P} \quad \forall i \in 1:n, \quad x_1 \perp \dots \perp x_n$$

независимых в совокупности одинаково распределенных случайных величин с распределением  $\mathcal{P}$ .

**Определение.** Повторной независимой выборкой объема  $n$  (после эксперимента) называется набор реализаций, т.е. конкретных значений  $\xi$ , случайных величин  $x_i$ :

$$\mathbf{x} = (x_1, \dots, x_n), \quad x_i \in \text{supp } \xi \quad \forall i \in 1:n.$$

**Определение.** Эмпирической случайной величиной  $\hat{\xi}_n$  называется случайная величина с дискретным распределением

$$\hat{\xi}_n \sim \hat{\mathcal{P}}_n : \begin{pmatrix} x_1 & \dots & x_n \\ 1/n & \dots & 1/n \end{pmatrix}.$$

*Замечание.* Подходящее определение выбирается по контексту.

Если  $\xi$  имеет дискретное распределение, то выборку можно *сгруппировать*; тогда получим случайную величину  $\hat{\xi}_m$  с распределением

$$\hat{\mathcal{P}}_m : \begin{pmatrix} x_1^* & \dots & x_m^* \\ \omega_1 & \dots & \omega_m \end{pmatrix} \quad \omega_i = \frac{\nu_i}{n},$$

где  $x_i^*$  — уникальные значения из выборки  $\mathbf{x}$ , а  $\nu_i$  — число  $x_i^*$  в  $\mathbf{x}$  (т.н. «абсолютная частота»; тогда  $\omega_i$  — «относительная частота»). В противном случае, можно разбить интервал всевозможных значений выборки на  $m$  подынтервалов:  $\{[e_0, e_1), \dots, [e_{m-1}, e_m)\}$  и считать число наблюдений  $\nu_i = \nu_i[e_{i-1}, e_i)$ , попавших в интервал.

**Следствие.** По ЗБЧ (теореме Бернулли),

$$\omega_i \xrightarrow{P} p_i = P(e_{i-1} \leq \xi < e_i),$$

т.е. относительная частота является хорошей оценкой вероятности на больших объемах выборки.

## 2. Виды признаков

Виды признаков случайной величины  $\xi : (\Omega, \mathcal{F}, P) \rightarrow (V, \mathfrak{A})$  характеризуются тем, что из себя представляет множество  $V$  и что можно делать с его элементами.

**Количественные признаки:**  $V \subset \mathbb{R}$

По типу операций:

- Аддитивные: заданы, т.е. имеют смысл в контексте данного признака, операции  $+$ ,  $-$
- Мультипликативные: заданы операции  $\cdot$ ,  $/$ ; признак принимает не отрицательные значения.

По типу данных:

- Непрерывные
- Дискретные

**Порядковые признаки**  $V$  — упорядоченное множество, определены отношения  $>$ ,  $=$ .

**Качественные признаки** на  $V$  заданы отношения  $=$ ,  $\neq$

**Пример.** Цвет глаз, имена, пол.



### 3. Характеристики распределений и метод подстановки

**Определение.** *Статистика* — измеримая функция от выборки.

Обобщением статистики является понятие характеристики.

**Определение.** *Характеристика* — функционал от распределения:

$$T : \{\mathcal{P}\} \rightarrow V.$$

Где  $V$  — измеримое пространство, чтобы на нём можно было завести  $\sigma$ -алгебру.

*Замечание.* Чаще всего,  $V = \mathbb{R}$ .

**Определение.** Выделяют *генеральные* характеристики  $T(\mathcal{P}) =: \theta$  и *выборочные* характеристики  $T(\hat{\mathcal{P}}_n)$ .

**Определение.** *Оценка* — выборочная характеристика  $T(\hat{\mathcal{P}}_n) =: \hat{\theta}_n$ , не зависящая от генеральной характеристики  $\theta$ .

**Следствие.** *Выражения для вычисления генеральных и выборочных характеристик отличаются только используемыми мерами ( $\mathcal{P}$  и  $\hat{\mathcal{P}}_n$  соответственно).*

**Определение.** Пусть  $\hat{\mathcal{P}}_n$  — распределение эмпирической случайной величины. Тогда *эмпирическая функция распределения* есть

$$\widehat{\text{cdf}}_{\xi}(x) = \text{cdf}_{\hat{\mathcal{P}}_n}(x) = \hat{\mathcal{P}}_n((-\infty, x)) = \int_{-\infty}^x d\hat{\mathcal{P}}_n = \sum_{x_i: x_i \leq x} \frac{1}{n} = \frac{|\{x_i \in \mathbf{x} : x_i \leq x\}|}{n}.$$

*Утверждение.* Пусть  $\widehat{\text{cdf}}_{\xi}$  — эмпирическая функция распределения,  $\text{cdf}_{\xi}$  — функция распределения  $\xi$ . Тогда, по теореме Гливенко-Кантелли,

$$\sup_x \left| \widehat{\text{cdf}}_{\xi}(x) - \text{cdf}_{\xi}(x) \right| \xrightarrow{\text{a.s.}} 0.$$

Более того, если  $\text{cdf}_{\xi}$  непрерывна, эта сходимости имеет порядок  $1/\sqrt{n}$  по теореме Колмогорова:

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{\text{cdf}}_{\xi}(x) - \text{cdf}_{\xi}(x) \right| \xrightarrow{d} \mathcal{P}_{\text{K.S.}},$$

где  $\mathcal{P}_{\text{K.S.}}$  — распределение Колмогорова-Смирнова.

*Замечание.* Поскольку  $\widehat{\text{cdf}}_{\xi}(x) = \omega_x$ , где  $\omega_x$  — частота попадания наблюдений в интервал  $(-\infty, x)$ , а  $\text{cdf}_{\xi}(x) = \mathbf{P}(\xi \in (-\infty, x))$  — вероятность того же события, то можно применить теорему Бернулли (ЗБЧ):

$$\widehat{\text{cdf}}_{\xi}(x) \xrightarrow{\mathbf{P}} \text{cdf}_{\xi}(x).$$

**Следствие.** *Значит, при достаточно больших  $n$ , в качестве интересующей характеристики  $\theta$  распределения  $\mathcal{P}$  можем брать ее оценку  $\hat{\theta}_n$  — аналогичную характеристику  $\hat{\mathcal{P}}_n$ .*

## 4. Характеристики распределений и их оценки

**Определение.** Генеральные и соответствующие им выборочные характеристики  $k$ -го момента и  $k$ -го центрального момента:

$$\begin{aligned} m_k &= \int_{\mathbb{R}} x^k d\mathcal{P} & \hat{m}_k &= \int_{\mathbb{R}} x^k d\hat{\mathcal{P}}_n = \frac{1}{n} \sum_{i=1}^n x_i^k \\ m_k^{(0)} &= \int_{\mathbb{R}} (x - m_1)^k d\mathcal{P} & \hat{m}_k^{(0)} &= \int_{\mathbb{R}} (x - \hat{m}_1)^k d\hat{\mathcal{P}}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{m}_1)^k. \end{aligned}$$

### 4.1. Характеристики положения

В качестве характеристики положения выделяется 1-й момент — математическое ожидание и выборочное среднее:

$$m_1 = E\xi, \quad \hat{m}_1 =: \bar{x} = \widehat{E\xi} = E\xi_n.$$

*Замечание.* В случае мультипликативных признаков можно посчитать среднее геометрическое; часто логарифмируют и считают среднее арифметическое.

**Определение.** Пусть  $p \in [0, 1]$  и  $\text{cdf} = \text{cdf}_P$ .  $p$ -квантилью (квантилью уровня  $p$ ) называется

$$\text{qnt}_P(p) =: z_p = \sup \{z : \text{cdf}(z) \leq p\}.$$

*Квартиль* есть квантиль уровня, кратного  $1/4$ ; *дециль* —  $1/10$ ; *перцентиль* —  $1/100$ .

*Замечание.*  $\sup$  берется для учета случая не непрерывных функций распределения.

**Определение.** Медиана есть  $1/2$ -квантиль:

$$\text{med } \xi = z_{1/2}.$$

**Определение.** Мода ( $\text{mode } \xi$ ) есть точка локального максимума плотности.

По методу подстановки можем получить аналогичные выборочные характеристики.

**Определение.** Выборочная  $p$ -квантиль есть такая точка  $\hat{z}_p$ , что она больше по значению  $|\mathbf{x}| \cdot p = np$  точек из выборки:

$$\hat{z}_p = \sup \left\{ z : \widehat{\text{cdf}}_{\xi}(z) \leq p \right\} = x_{(\lfloor np \rfloor + 1)}.$$

**Определение.** Выборочная медиана упорядоченной выборки  $\mathbf{x} = (x_{(1)}, \dots, x_{(n)})$  есть

$$\hat{z}_{1/2} = \widehat{\text{med}} = \begin{cases} x_{(k+1)} & n = 2k + 1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & n = 2k \end{cases}$$

**Определение.** Выборочная мода ( $\widehat{\text{mode}}$ ) есть значение из выборки, которое чаще всего встречается.

## 4.2. Характеристики разброса

В качестве характеристики разброса выделяется 2-й центральный момент — дисперсия и выборочная дисперсия:

$$m_2^{(0)} = D\xi \quad \hat{m}_2^{(0)} =: s^2 = \widehat{D\xi} = D\hat{\xi}_n = \begin{cases} E\left(\hat{\xi}_n - E\hat{\xi}_n\right)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ E\hat{\xi}_n^2 - \left(E\hat{\xi}_n\right)^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2\right) - \bar{x}^2. \end{cases}$$

*Замечание.* Если среднее  $E\xi = \mu$  известно, то дополнительно вводится

$$s_\mu^2 := \begin{cases} \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \\ \left(\frac{1}{n} \sum_{i=1}^n x_i^2\right) - \mu^2. \end{cases}$$

**Пример** (Оценка дисперсии оценки мат. ожидания). Пусть строится оценка мат. ожидания  $\bar{x}$ . Может интересовать точность построенной оценки. Вычислим дисперсию теоретически, после чего оценим точность по выборке:

$$D\bar{x} = D\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} \sum_{i=1}^n Dx_i = \frac{1}{n^2} \sum_{i=1}^n D\xi = \frac{D\xi}{n},$$

откуда

$$\widehat{D\bar{x}} = \frac{s^2}{n}.$$

**Пример** (Дисперсия оценки дисперсии). См. по ссылке<sup>1</sup>.

**Определение** (Энтропия). Количество информации, необходимое для выявления объекта из  $n$ -элементного множества вычисляется по *формуле Хартли*:

$$H = \log_2 n$$

(множество это следует итеративно разбивать пополам, откуда и оценка). Пусть теперь множество не равновероятно, т.е. задано дискретное распределение

$$\mathcal{P}_\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}.$$

Тогда количество информации  $H(\xi)$ , которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона и называется *энтропией*:

$$H(\xi) = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i}.$$

*Замечание.* В случае равномерного дискретного распределения, конечно,  $H = H(\xi)$ .

**Определение.** Выборочное стандартное отклонение есть

$$SD := \sqrt{\widehat{D\xi}} = s.$$

Это показатель разброса случайной величины; показатель того, насколько элементы выборки отличаются от выборочного среднего по значению.

<sup>1</sup><http://mathworld.wolfram.com/SampleVarianceDistribution.html>

SD позволяет оценивать стандартное отклонение распределения  $\xi$ .

Пусть  $\hat{\theta}_n$  — статистика. Она имеет какое-то своё распределение, стандартное отклонение которого можно также оценить.

**Определение.** *Стандартная ошибка* оценки есть

$$\text{SE}(\hat{\theta}) := \sqrt{\widehat{\text{D}\hat{\theta}}}.$$

Это показатель разброса оценки случайной величины.

*Замечание.* В частном случае  $\theta = \mathbb{E}\xi$ ,  $\hat{\theta} = \bar{\mathbf{x}}$  получаем *выборочную стандартную ошибку среднего*

$$\text{SE} := \text{SE}(\bar{\mathbf{x}}) = \sqrt{\widehat{\text{D}\bar{\mathbf{x}}}} = \sqrt{\frac{\widehat{\text{D}\xi}}{n}} = \frac{s}{\sqrt{n}}.$$

Это, в свою очередь, показатель того, насколько выборочное среднее отличается от истинного.

Пусть  $c_\gamma = \text{qnt}_{N(0,1)} \gamma$ .

**Пример** (С мостом и машинами). При возведении моста требуется, чтобы под ним могли проехать, условно, 95% машин. Чтобы эту высоту вычислить, достаточно собрать выборку высоты кузова проезжающих машин. Тогда нахождение искомой величины можно наглядно представить как выбор такой квантили гистограммы выборки, что суммирование соответствующих вероятностей даст 0.95. В предположении, что выборка из нормального распределения, с более устойчивой оценкой квантили, интервал будет иметь вид

$$(\bar{\mathbf{x}} \pm \text{SD} \cdot c_\gamma).$$

SE как показатель разброса среднего использовать по смыслу нельзя.

**Пример** (С паромом). Число машин, которое способен перевезти паром, есть Грузоподъемность/ $\mathbb{E}\xi$ , где  $\xi$  — вес машины. Поскольку оценка  $\bar{\mathbf{x}}$  всегда считается с погрешностью относительно истинного значения, интервал допустимого числа машин будет иметь вид

$$\frac{\text{Грузоподъемность}}{\bar{\mathbf{x}} \pm \text{SE} \cdot c_\gamma}.$$

### 4.3. Анализ характера разброса

**Определение.** *Коэффициент асимметрии Пирсона* («скошенности»<sup>2</sup>)

$$\gamma_3 = \text{A}\xi = \frac{m_3^{(0)}}{\sigma^3} = \frac{\mathbb{E}\xi - \text{med } \xi}{\sigma}.$$

*Замечание.* Не зависит от линейных преобразований.

**Определение.** *Коэффициент эксцесса* («крутизны», «kurtosis»):

$$\gamma_4 = \text{K}\xi = \frac{m_4^{(0)}}{\sigma^4} - 3.$$

*Замечание.* Величина  $m_4^{(0)}/\sigma^4 = 3$  соответствует стандартному нормальному распределению. Так что можно сравнивать выборку и  $\gamma_4 N(0, 1)$ .

---

<sup>2</sup> «Skewness».

*Замечание.* При замене  $z := (\xi - E\xi)/\sigma$  величину  $m_4^{(0)}/\sigma^4 = E(z^4)$  можно интерпретировать как ожидание четвертой степени центрированных и нормированных данных. Точки выборки, лежащие внутри  $E\xi \pm \sigma$  из-за малости по модулю не будут увеличивать значение коэффициента, в то время как аутлаеры будут или «тяжелые хвосты» плотности распределения будут. Поэтому  $\gamma_4$  принимает большие значения на распределениях с «тяжелыми хвостами» или выборках с некоторым количеством аутлаеров.

*Замечание.* Справедлива оценка

$$\gamma_3^2 + 4 \leq \gamma_4 + 3 \leq \infty,$$

где минимум достигается  $\text{Ber}(1/2)$ .

## 4.4. Характеристики зависимости

**Определение.** Пусть  $(\xi_1, \xi_2) \sim \mathcal{P}$  и  $(x_1, y_1), \dots, (x_n, y_n) \sim \mathcal{P}(du \times dv)$ . Тогда можно записать две другие важные характеристики: *ковариацию* и *коэффициент корреляции*:

$$\begin{aligned} \text{cov}(\xi_1, \xi_2) &= \iint_{\mathbb{R}^2} (u - m_1(u))(v - m_1(v)) \mathcal{P}(du \times dv) & \widehat{\text{cov}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}}) \\ \text{cor}(\xi_1, \xi_2) &= \frac{\text{cov}(\xi_1, \xi_2)}{\sigma_{\xi_1} \sigma_{\xi_2}} & \widehat{\text{cor}}(\mathbf{x}, \mathbf{y}) &= \frac{\widehat{\text{cov}}(\mathbf{x}, \mathbf{y})}{s(\mathbf{x})s(\mathbf{y})}. \end{aligned}$$

*Замечание (Важное).*  $\xi_1 \not\parallel \xi_2 \implies \text{cov}(\xi_1, \xi_2) = 0$ , но  $\text{cov}(\xi_1, \xi_2) = 0 \not\implies \xi_1 \parallel \xi_2$ . Необходимость и достаточность выполняется только в случае нормального распределения.

*Замечание (Проблема моментов).* Для заданной последовательности моментов  $m_1, m_2, \dots$  не обязательно существовать подходящее распределение. Помимо требований  $m_{2k} \geq 0$  и взаимосвязи между соседними моментами по неравенству Гёльдера, существенно, что ряд Тейлора по  $m_\ell$ , в который, как известно, раскладывается характеристическая функция, должен сходиться равномерно.

## 5. Точечная оценка параметров распределения

### 5.1. Метод моментов

Пусть  $\mathcal{P}(\boldsymbol{\theta})$ ,  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^\top$  — параметрическая модель. Найдем оценки для параметров  $\hat{\theta}_i$ ,  $i \in \overline{1:r}$ , для чего составим и решим систему уравнений:

$$\begin{cases} \mathbb{E}g_1(\xi) = \phi_1(\theta_1, \dots, \theta_r) \\ \vdots \\ \mathbb{E}g_r(\xi) = \phi_r(\theta_1, \dots, \theta_r) \end{cases} \implies \begin{cases} \theta_1 = f_1(\mathbb{E}g_1(\xi), \dots, \mathbb{E}g_r(\xi)) \\ \vdots \\ \theta_r = f_r(\mathbb{E}g_1(\xi), \dots, \mathbb{E}g_r(\xi)). \end{cases}$$

Примем

$$\theta_i^* = f_i(\hat{\mathbb{E}}g_1(\xi), \dots, \hat{\mathbb{E}}g_r(\xi)).$$

Часто,  $g_i(\xi) = \xi^i$ .

*Замечание.* Случается, что решение находится вне пространства параметров. На практике, если пространство параметров компактное, можно взять точку, ближайшую к полученной оценке. Однако это свидетельствует о том, что модель плохо соответствует данным.

**Пример 5.1** ( $r = 1$ ).  $\xi \sim U(0, \theta)$ .

- Оценка по 1-му моменту:  $g(\xi) = \xi$  и

$$\mathbb{E}\xi = \int_0^\theta \frac{1}{\theta} x \, dx = \frac{1}{\theta} \frac{x^2}{2} \Big|_0^\theta = \frac{\theta}{2} \implies \theta = 2\mathbb{E}\xi, \quad \theta^* = 2\bar{x}.$$

- Оценка по  $k$ -му моменту:  $g(\xi) = \xi^k$  и

$$\mathbb{E}\xi^k = \frac{1}{\theta} \int_0^\theta x^k \, dx = \frac{1}{\theta} \frac{x^{k+1}}{k+1} \Big|_0^\theta = \frac{\theta^k}{k+1} \implies \theta^* = \sqrt[k]{(k+1) \frac{1}{n} \sum_{i=1}^n x_i^k}.$$

**Пример 5.2** ( $r = 1$ ). Пусть  $\xi \sim \text{Exp}(\lambda)$ . Тогда  $\mathbb{E}\xi = \lambda$  и  $\bar{x} = \lambda$ .

**Пример 5.3** ( $r = 2$ ). Пусть  $\mathcal{P}_\xi(\theta_1, \theta_2) = \text{Bin}(m, p)$ . Тогда  $g_1(\xi) = \xi$ ,  $g_2(\xi) = (\xi - \mathbb{E}\xi)^2$  и

$$\begin{cases} \mathbb{E}\xi = mp \\ D\xi = mp(1-p) \end{cases} \quad \begin{cases} m = \frac{\mathbb{E}\xi}{p} \\ D\xi = \mathbb{E}\xi - \mathbb{E}\xi p \end{cases} \quad \begin{cases} p = \frac{\mathbb{E}\xi - D\xi}{\mathbb{E}\xi} \\ m = \frac{(\mathbb{E}\xi)^2}{\mathbb{E}\xi - D\xi} \end{cases} \implies \begin{cases} \hat{p} = \frac{\bar{x} - s^2}{\bar{x}} \\ \hat{m} = \frac{\bar{x}^2}{\bar{x} - s^2}. \end{cases}$$

### 5.2. Метод оценки максимального правдоподобия

Пусть  $\mathcal{P}_\xi(\boldsymbol{\theta})$ ,  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^\top$  — параметрическая модель.

**Определение.** Функция правдоподобия:

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = P(\mathbf{y} \mid \boldsymbol{\theta}) = \begin{cases} P_{\boldsymbol{\theta}}(x_1 = y_1, \dots, x_n = y_n) & \mathcal{P}_\xi(\boldsymbol{\theta}) \text{ дискретно;} \\ p_{\boldsymbol{\theta}}(\mathbf{y}) & \mathcal{P}_\xi(\boldsymbol{\theta}) \text{ абсолютно непрерывно.} \end{cases}$$

**Пример 5.4.** Пусть  $\xi \sim N(\mu, \sigma^2)$ . По независимости  $x_i$ ,  $p_{\theta}(\mathbf{x})$  распадается в произведение:

$$L(\theta | \mathbf{x}) = p_{\theta}(\mathbf{x}) = \prod_{i=1}^n p_{\theta}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}.$$

**Пример 5.5.**  $\xi \sim \text{Pois}(\lambda)$ ,

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda} \implies L(\theta | \mathbf{x}) = \prod_{i=1}^n \frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} = \frac{1}{\prod_{i=1}^n x_i!} \lambda^{n\bar{x}} e^{-n\lambda}.$$

*Утверждение.* Пусть  $\mathbf{x}$  — выборка. В качестве оценки максимального правдоподобия<sup>1</sup>  $\hat{\theta}_{\text{MLE}}$  следует взять

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \log L(\theta | \mathbf{x}).$$

**Пример.**  $\xi \sim \text{Pois}(\lambda)$ .

$$\ln L(\lambda | \mathbf{x}) = - \sum_{i=1}^n x_i! - n\lambda + \ln(\lambda^{n\bar{x}}) \implies \frac{\partial \log L(\lambda | \mathbf{x})}{\partial \lambda} = -n + \lambda^{-n\bar{x}} n\bar{x} \lambda^{n\bar{x}-1} = -n + \frac{n\bar{x}}{\lambda}$$

откуда

$$\frac{\partial \log L(\lambda | \mathbf{x})}{\partial \lambda} = 0 \iff -n + \frac{n\bar{x}}{\lambda} = 0, \quad n\bar{x} - n\lambda = 0, \quad \lambda = \bar{x}.$$

*Утверждение.* В условиях регулярности:

1. Существует один глобальный максимум, так что

$$\left. \frac{\partial \log L(\lambda | \mathbf{x})}{\partial \lambda} \right|_{\theta=\hat{\theta}_{\text{MLE}}} = 0.$$

2.  $\hat{\theta}_{\text{MLE}}$  обладает всеми свойствами:

- а) Состоятельность;
- б) Асимптотическая несмещенность;
- в) Асимптотическая нормальность;
- г) Эффективность.

---

<sup>1</sup>Maximum likelihood estimate (MLE).

## 6. Свойства оценок

### 6.1. Несмещенность

**Определение.** Смещение<sup>1</sup> есть

$$\text{bias } \hat{\theta}_n := E\hat{\theta}_n - \theta \quad \forall \theta \in \Theta.$$

**Определение.** Среднеквадратичная ошибка<sup>2</sup> есть

$$\text{MSE } \hat{\theta}_n := E(\hat{\theta}_n - \theta)^2.$$

*Замечание.* Поскольку

$$D\hat{\theta}_n = D(\hat{\theta}_n - \theta) = E(\hat{\theta}_n - \theta)^2 - (E(\hat{\theta}_n - \theta))^2,$$

то

$$\underbrace{E(\hat{\theta}_n - \theta)^2}_{\text{MSE}} = D\hat{\theta}_n + \underbrace{(E(\hat{\theta}_n - \theta))^2}_{\text{bias}^2}.$$

**Определение.** Оценка называется *несмещенной*, если  $\text{bias } \hat{\theta}_n = 0$ , т.е.

$$E\hat{\theta}_n = \theta.$$

**Предложение.**  $\bar{x}$  — несмещенная оценка  $E\xi$ .

*Доказательство.* Пусть  $\theta = E\xi$ ,  $\hat{\theta}_n = E\hat{\xi}_n = \bar{x}$ . Тогда

$$E\bar{x} = E\frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n E x_i = \frac{1}{n} \sum_{i=1}^n E\xi = E\xi \implies E\hat{\theta}_n = E\theta, \text{ bias } \hat{\theta}_n = 0.$$

□

**Предложение.**  $s^2$  является только асимптотически несмещенной оценкой  $D\xi$ .

*Доказательство.* Поскольку дисперсия не зависит от сдвига, обозначим  $\eta = \xi - E\xi$  и  $y_i = x_i - E\xi$ ; тогда

$$\begin{aligned} Es^2 &= E\widehat{D\xi} = E\widehat{D\eta} = E\left(\widehat{E\eta^2} - (\widehat{E\eta})^2\right) = E\left(\frac{1}{n} \sum_{i=1}^n y_i^2 - \left(\frac{1}{n} \sum_{i=1}^n y_i\right)^2\right) \\ &= E\left(\frac{1}{n} \sum_{i=1}^n y_i^2 - \frac{1}{n^2} \sum_{i,j=1}^n y_i y_j\right) = E\left(\frac{1}{n} \sum_{i=1}^n y_i^2 - \frac{1}{n^2} \sum_{i,j=1}^n y_i^2\right) = \frac{1}{n} \sum_{i=1}^n E y_i^2 - \frac{1}{n^2} \sum_{i=1}^n E y_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n E (x_i - E\xi)^2 - \frac{1}{n^2} \sum_{i=1}^n E (x_i - E\xi)^2 = \frac{1}{n} \sum_{i=1}^n D x_i - \frac{1}{n^2} \sum_{i=1}^n D x_i = D\xi - \frac{1}{n} D\xi \\ &= \frac{n-1}{n} D\xi \xrightarrow{n \rightarrow \infty} D\xi. \end{aligned}$$

□

---

<sup>1</sup>Bias.

<sup>2</sup>Mean square error (MSE).



**Определение.** Исправленная дисперсия:

$$\tilde{s}^2 := \frac{n}{n-1} s^2.$$

**Предложение.**  $\widehat{\text{cdf}}_\xi$  — несмещенная оценка  $\text{cdf}_\xi$ .

*Доказательство.* FIXME □

**Предложение.** Как правило, оценки по методу моментов смещенные.

*Доказательство.* Несмещенность означала бы выполнение для всех  $\theta \in \Theta$  равенства

$$\mathbb{E}\theta^* = \mathbb{E}\phi^{-1}(\hat{\mathbb{E}}g(\xi)) = \theta = \phi^{-1}(\mathbb{E}\hat{\mathbb{E}}g(\xi)).$$

Но часто  $\phi^{-1}$  — выпуклая, так что имеем, на самом деле, неравенство Йенсена. □

**Предложение.** В условиях регулярности, ОМП асимптотически несмещенная.

*Доказательство.* FIXME □

## 6.2. Состоятельность

**Определение.** Оценка называется *состоятельной* в среднеквадратичном смысле, если

$$\text{MSE } \hat{\theta}_n \xrightarrow{n \rightarrow \infty} 0.$$

**Определение.** Оценка называется *состоятельной*, если

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta.$$

**Предложение.** Если оценка несмещенная и состоятельная в среднеквадратичном смысле, то она состоятельная.

*Доказательство.* В самом деле, по неравенству Чебышева,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) = \mathbb{P}(|\hat{\theta}_n - \mathbb{E}\hat{\theta}_n| > \epsilon) \leq \frac{\mathbb{D}\hat{\theta}_n}{\epsilon^2} = \frac{\text{MSE } \hat{\theta}_n}{\epsilon^2} \xrightarrow{n \rightarrow \infty} 0.$$

□

**Предложение.**  $\hat{\mathbf{m}}_k$  является состоятельной оценкой  $\mathbf{m}_k$ .

*Доказательство.* Докажем для  $\hat{\mathbf{m}}_1$ . По определению выборки до эксперимента,  $x_i \sim \mathcal{P}$ . Тогда, по теореме Хинчина о ЗБЧ,

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^n x_i}{n} \xrightarrow{\mathbb{P}} \mathbf{m}_1(\mathcal{P}).$$

Для  $k$ -го момента доказывается аналогично заменой  $y_i := x_i^k$ . □

*Замечание.* Для  $\mathbf{m}_k^{(0)}$  доказательство не пройдет, потому что  $x_i$  и  $\bar{\mathbf{x}}$  не будут независимыми.

**Предложение.**  $\hat{\mathbf{m}}_k^{(0)}$  является состоятельной оценкой  $\mathbf{m}_k^{(0)}$ .

*Утверждение.* Пусть  $\xi_n \xrightarrow{\mathbb{P}} c$  и  $f \in C(U_\epsilon(c))$ . Тогда  $f(\xi_n) \xrightarrow{\mathbb{P}} f(c)$ .

## 6. Свойства оценок

*Доказательство предложения.* Докажем для  $s^2$ . Пусть  $f : (x, y) \mapsto x - y^2$ . Устроим последовательность  $(\hat{m}_2, \hat{m}_1) \xrightarrow{P} (m_2, m_1)$ . Тогда

$$f(\hat{m}_2, \hat{m}_1) = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = s^2 \xrightarrow{P} f(m_2, m_1) = D\xi.$$

Для  $m_k^{(0)}$  доказывается аналогично. □

**Предложение.**  $\bar{x}$  — состоятельная оценка  $E\xi$ .

*Доказательство.* Либо по (6.2) для  $k = 1$ , либо из того факта, что  $\text{bias } \bar{x} = 0$ , значит

$$\text{MSE } \bar{x} = D\bar{x} = \frac{D\xi}{n} \xrightarrow{n \rightarrow \infty} 0,$$

и по (6.2) получаем утверждение. □

**Предложение.**  $s^2$  — состоятельная оценка  $D\xi$ .

*Доказательство.* По (6.2) с  $k = 2$ . □

**Предложение.**  $\widehat{\text{cdf}}_\xi$  — состоятельная оценка  $\text{cdf}$  в каждой точке.

*Доказательство.* Введем случайную величину

$$y_i := \mathbf{1}_{\{x_i < x\}} = \begin{cases} 1 & x_i < x \\ 0 & x_i \geq x. \end{cases}$$

$y_i$  независимы ( $\text{cdf}_{y_i}(x) \text{cdf}_{y_j}(y) = \text{cdf}_{y_i, y_j}(x, y)$ ) и одинаково распределены; кроме того, их математическое ожидание конечно:

$$E y_i = 1 \cdot P(x_i < x) + 0 \cdot P(x_i \geq x) = P(x_i < x) = \text{cdf}_\xi(x) < \infty.$$

Тогда по теореме Хинчина о ЗБЧ,

$$\widehat{\text{cdf}}_\xi(x) = \frac{|\{x_i \in \bar{x} : x_i < x\}|}{n} = \frac{\sum_{i=1}^n y_i}{n} \xrightarrow{P} E y_i = E \mathbf{1}_{\{x_i < x\}} = P(x_i < x) = \text{cdf}(x).$$

□

*Утверждение.* Пусть  $\exists! p_0 : \text{cdf}(x) = p_0$  и  $\text{cdf}(x)$  монотонно возрастает в окрестности  $p_0$ . Тогда  $\bar{z}_{p_0} \xrightarrow{P} z_{p_0}$ , т.е. является состоятельной оценкой.

**Предложение.** Оценки, полученные по методу моментов, являются состоятельными.

*Доказательство.* Поскольку  $f_i$  — непрерывные функции и при непрерывных преобразованиях сходимости не портятся:

$$\theta^* = \phi^{-1}(\hat{E}g(\xi)) \xrightarrow{P} \phi^{-1}(Eg(\xi)) = \phi^{-1}(\phi(\theta)) = \theta.$$

□

**Предложение.** Оценки  $\hat{\theta}_{\text{MLE}}$ , полученные по методу максимального правдоподобия, являются состоятельными.

*Доказательство.* Пусть  $\theta_0$  — истинный параметр  $\mathcal{P}(\theta)$ . По УЗБЧ,

$$\frac{1}{n} \ln L(\theta | \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \ln p_{\theta}(x_i) \xrightarrow{P} E \ln p_{\theta}(x_i) = \int_{\mathbb{R}} \ln(p_{\theta}(x)) p_{\theta_0}(x) dx.$$

Навесим на обе стороны  $\operatorname{argmax}$  в условии, что это непрерывное преобразование:

$$\hat{\theta}_{\text{MLE}} \leftarrow \operatorname{argmax}_{\theta} \frac{1}{n} \ln L(\theta | \mathbf{x}) \xrightarrow{P} \operatorname{argmax}_{\theta} \int_{\mathbb{R}} \ln(p_{\theta}(x)) p_{\theta_0}(x) dx \rightarrow \theta^*.$$

Тогда в предположении непрерывности  $p_{\theta}$  по  $\theta$ ,  $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta^*$ . Покажем, что  $\theta^* = \theta_0$ . Поделим на  $p_{\theta_0}$  — константу по  $\theta$ :

$$\frac{1}{n} \sum_{i=1}^n \ln \frac{p_{\theta}}{p_{\theta_0}}(x_i) \xrightarrow{P} \int_{\mathbb{R}} \ln \left( \frac{p_{\theta}(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx = E \ln \frac{p_{\theta}}{p_{\theta_0}} \leq \ln E \frac{p_{\theta}}{p_{\theta_0}} = \ln \int_{\mathbb{R}} \frac{p_{\theta}}{p_{\theta_0}}(x) p_{\theta_0}(x) dx = \ln 1 = 0$$

по неравенству Енсена  $Eg(\xi) \leq g(E\xi)$ , для выпуклой вверх  $g(x) = \log(x)$ . Таким образом,

$$\int_{\mathbb{R}} \ln \left( \frac{p_{\theta}(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx = 0 \iff \ln \left( \frac{p_{\theta}(x)}{p_{\theta_0}(x)} \right) = 0 \iff \frac{p_{\theta}(x)}{p_{\theta_0}(x)} = 1 \iff p_{\theta}(x) = p_{\theta_0}(x)$$

для почти всех  $x$ . В предположении свойства *идентифицируемости* задачи ( $\theta_1 \neq \theta_2 \implies \mathcal{P}_{\theta_1} \neq \mathcal{P}_{\theta_2}$ ), получаем  $\theta = \theta_0$ .  $\square$

### 6.3. Асимптотическая нормальность

**Определение.** Оценка  $\hat{\theta}_n$  называется *асимптотически нормальной* оценкой параметра  $\theta$  с коэффицентом  $\sigma^2(\theta)$  если

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)).$$

**Пример.**  $\bar{x}$  — асимптотически нормальная оценка, если  $D\xi < \infty$ ,  $D\xi \neq 0$ :

$$\sqrt{n} (\bar{x} - E\xi) \xrightarrow{d} N(0, D\xi).$$

*Доказательство.* По ЦПТ,

$$\sqrt{n} (\bar{x} - E\xi) = \frac{\sum_{i=1}^n x_i - nE\xi}{\sqrt{n}} \xrightarrow{d} N(0, D\xi).$$

$\square$

**Пример.**  $\hat{m}_k$  — асимптотически нормальная оценка.

*Доказательство.* FIXME

$\square$

**Пример.**  $s^2$  и  $\tilde{s}^2$  — асимптотически нормальные оценки, если  $0 \neq D(\xi - E\xi)^2 < \infty$ :

$$\sqrt{n} (s^2 - D\xi) \xrightarrow{d} N(0, D(\xi - E\xi)^2).$$

*Доказательство.* Пусть  $\eta = \xi - E\xi$  и  $y_i = x_i - E\xi$ . Тогда

$$\begin{aligned} s^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - E\xi - (\bar{x} - E\xi))^2 = \hat{E}\eta^2 - \bar{y}^2 \\ \sqrt{n}(s^2 - D\xi) &= \sqrt{n}(\hat{E}\eta^2 - \bar{y}^2 - D\xi) = \sqrt{n}(\hat{E}\eta^2 - \underbrace{D\xi}_{=D\eta=E\eta^2}) - \sqrt{n}\bar{y}^2 \\ &= \underbrace{\frac{\sum_{i=1}^n y_i^2 - nE\eta^2}{\sqrt{n}}}_{\xrightarrow{d} N(0, D\eta^2)} - \underbrace{\bar{y}}_{\xrightarrow{d} 0} \underbrace{\sqrt{n}\bar{y}}_{\xrightarrow{d} N(0, D\xi)} \xrightarrow{d} N(0, D(\xi - E\xi)^2). \end{aligned}$$

$\square$

**Пример.**  $\widehat{\text{cdf}}_\xi$  — асимптотически нормальная оценка:

$$\sqrt{n} \left( \widehat{\text{cdf}}_\xi(x) - \text{cdf}_\xi(x) \right) \xrightarrow{d} N(0, \hat{\sigma}^2), \quad \text{где } \hat{\sigma}^2 = \text{cdf}(x)(1 - \text{cdf}(x)).$$

*Доказательство.* FIXME □

**Пример.** При определенных условиях,  $\widehat{\text{med}}\xi$  является асимптотически нормальной оценкой  $\text{med } \xi$ .

*Доказательство.* FIXME □

**Предложение.** Асимптотически нормальная оценка состоятельна.

*Доказательство.* Действительно,

$$\hat{\theta}_n - \theta = \frac{1}{\sqrt{n}} \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} 0 \cdot \eta = 0, \quad \eta \sim N(0, \sigma^2(\theta)).$$

Слабая сходимость же к константе влечет слабую сходимость по вероятности, откуда  $\hat{\theta}_n \xrightarrow{P} \theta$ . □

**Следствие.** Асимптотически нормальные оценки сходятся к оцениваемому параметру со скоростью  $1/\sqrt{n}$ .

*Замечание.* Если оценка не асимптотически нормальная, она *может* сходиться к параметру быстрее  $1/\sqrt{n}$ .

**Определение.** Пусть  $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$  — асимптотически нормальные оценки с соответствующими коэффициентами  $(\sigma^{(1)})^2(\theta)$  и  $(\sigma^{(2)})^2(\theta)$ . Говорят, что  $\hat{\theta}^{(1)}$  *лучше*  $\hat{\theta}^{(2)}$  в смысле асимптотического подхода, если

$$(\sigma^{(1)})^2(\theta) \leq (\sigma^{(2)})^2(\theta), \quad \forall \theta \in \Theta$$

(и хотя бы при одном  $\theta$  это неравенство строгое).

## 6.4. Эффективность

**Определение.** Говорят, что оценка  $\hat{\theta}^{(1)}$  *лучше*  $\hat{\theta}^{(2)}$  в среднеквадратичном смысле, если

$$\text{MSE } \hat{\theta}^{(1)} \leq \text{MSE } \hat{\theta}^{(2)}.$$

*Замечание.* Оценка  $\hat{\theta}^*$  является лучшей в среднеквадратичном смысле в классе всевозможных оценок только если она совпадает с самим оцениваемым параметром,  $\hat{\theta}^* = \theta$ . Значит, в невырожденном с точки зрения статистики случае наилучшей среднеквадратичной оценки не существует.

*Замечание.* Для несмещенных оценок определение эквивалентно, конечно,

$$D\hat{\theta}^{(1)} \leq D\hat{\theta}^{(2)}.$$

**Определение.** Поскольку в классе всех оценок наилучшей не существует, класс этот разбивается на  $\mathcal{K}_b = \left\{ \hat{\theta} \mid E\hat{\theta} = \theta + \text{bias } \theta = \theta + b \right\}$  — классы всех оценок с одинаковым смещением. Оценка  $\hat{\theta}^* \in \mathcal{K}_b$  называется эффективной, если её среднеквадратичная ошибка меньше всех других оценок в этом классе:

$$E(\hat{\theta}^* - \theta)^2 \leq E(\hat{\theta} - \theta)^2, \quad \forall \hat{\theta} \in \mathcal{K}_b, \theta \in \Theta.$$

**Определение.** Эффективная оценка в классе  $\mathcal{K}_0$  называется просто *эффективной*.

**Предложение.** Эффективная оценка в классе  $\mathcal{K}_b$  единственна.

*Доказательство.* FIXME □

**Пример.** Для  $\xi \sim U(0, \theta)$ , оценка  $\hat{\theta}_n^{(1)} = \max(\mathbf{x})$  более эффективна, чем  $\hat{\theta}_n^{(2)} = 2\bar{\mathbf{x}}$ .

*Доказательство.* FIXME □

**Пример.** Пусть  $\xi \sim \text{Pois}(\lambda)$ . Поскольку

$$\begin{aligned} D\hat{\lambda}_n &= D\bar{\mathbf{x}} = E\xi/n = \lambda/n \\ I_n(\lambda) &= n/\lambda, \end{aligned}$$

то  $\hat{\lambda}_n$  — эффективная оценка (как и ожидалась по свойствам  $\hat{\theta}_{\text{MLE}}$ ).

**Пример** (Сравнение оценок мат. ожидания симметричного распределения). Пусть  $\mathcal{P}$  симметрично — в этом случае  $\widehat{\text{med}} \xi = \bar{\mathbf{x}}$  и имеет смысл сравнить две этих характеристики.

$$\begin{aligned} D\bar{\mathbf{x}} &= \frac{D\xi}{n} \\ D\widehat{\text{med}} \xi &\sim \frac{1}{4n \text{pdf}_{N(\mu, \sigma^2)}^2(\text{med } \xi)} \quad \text{при } n \rightarrow \infty. \end{aligned}$$

Так, если  $\xi \sim N(\mu, \sigma^2)$ , то

$$\text{pdf}_{N(\mu, \sigma^2)}^2(\text{med } \xi) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{(\text{med } \xi - \mu)^2}{\sigma^2}\right\} = \frac{1}{2\pi\sigma^2},$$

откуда

$$D\widehat{\text{med}} \xi = \frac{\pi \sigma^2}{2 n} > \frac{\sigma^2}{n} = D\bar{\mathbf{x}},$$

значит  $\bar{\mathbf{x}}$  эффективнее  $\widehat{\text{med}} \xi$ .

*Замечание.* В то же время,  $\widehat{\text{med}} \xi$  более устойчива к аутлаерам, чем  $\bar{\mathbf{x}}$ , и этим лучше.

## 7. Проверка оценок на эффективность

Пусть  $\mathcal{P}_\xi(\theta)$ ,  $\theta = (\theta_1, \dots, \theta_r)^\top$  — параметрическая модель. Пусть  $r = 1$ .

**Определение.** Информанта  $n$ -го порядка:

$$S_n(\mathbf{x}, \theta) = \frac{d^n \ln L(\theta | \mathbf{x})}{d\theta^n}.$$

**Определение.** Информационное количество Фишера:

$$I_n(\theta) := -ES_2(\mathbf{x}, \theta).$$

Утверждение.

$$I_n(\theta) = ES_1^2(\mathbf{x}, \theta).$$

**Пример.**  $\xi \sim \text{Pois}(\lambda)$ .

$$S_1(\mathbf{x}, \theta) = -n + \frac{n\bar{x}}{\lambda}, \quad S_2(\mathbf{x}, \theta) = -\frac{n\bar{x}}{\lambda^2} \implies I_n(\lambda) = E \frac{n\bar{x}}{\lambda^2} = \frac{n}{\lambda^2} E\bar{x} = \frac{n}{\lambda}.$$

Замечание.

$$\ln L(\theta | \mathbf{x}) = \sum_{i=1}^n \ln p_\theta(x_i) \implies S_2 = \frac{d^2 \ln L(\theta | \mathbf{x})}{d\theta^2} = \sum_{i=1}^n (\ln p_\theta(x_i))'',$$

откуда, для повторной независимой выборки,

$$I_n(\theta) = -\sum_{i=1}^n E(\ln p_\theta(x_i))'' = n \cdot i(\theta), \quad \text{где } i(\theta) = -E(\ln p_\theta(\xi))''.$$

**Определение.**  $C \subset \mathbb{R}$  есть носитель параметрического семейства распределений  $\mathcal{P}(\theta)$ , если

$$\xi \sim \mathcal{P}(\theta) \implies P(\xi \in C) = 1, \quad \forall \theta \in \Theta.$$

**Определение.** Условие регулярности:

- Существует  $C$  — носитель распределения  $\mathcal{P}(\theta)$  такой, что  $\forall y \sqrt{p_\theta(y)}$  непрерывно дифференцируема по  $\theta$  всюду в области  $\Theta$ .
- Существует  $I_n(\theta) > 0$ , непрерывное по  $\theta$  всюду в области  $\Theta$ .

**Пример.**  $\text{Exp}(\lambda)$  — регулярное семейство;  $U(0, \theta)$  — не является регулярным.

Доказательство. FIXME

□

Утверждение. Для несмещенных оценок в условиях регулярности справедливо неравенство Рао–Крамера:

$$D\hat{\theta}_n \geq \frac{1}{I_n(\theta)}.$$

Для смещенных оценок,

$$D\hat{\theta}_n \geq \frac{(1 + \text{bias}'(\theta))^2}{I_n(\theta)}.$$

## 7. Проверка оценок на эффективность

**Следствие.** Несмещенная оценка является эффективной, если:

$$D\hat{\theta}_n = \frac{1}{I_n(\theta)}.$$

**Упражнение** (Хорошее). Показать, что  $\bar{x}$  является эффективной оценкой  $\mu$  в модели  $\xi \sim N(\mu, \sigma^2)$ .

**Определение.** Пусть  $\hat{\theta}_n$  — асимптотически несмещенная оценка. Тогда  $\hat{\theta}_n$  — *асимптотически эффективная*, если

$$D\hat{\theta}_n \cdot I_n \xrightarrow{n \rightarrow \infty} 1.$$

**Пример.** Пусть  $\xi \sim N(\mu, \sigma^2)$ . Можно посчитать, что  $s^2$  является только асимптотически эффективной оценкой  $\sigma^2$ ;  $\tilde{s}^2$  — просто эффективной.

## 8. Построение эффективных оценок

Эффективную оценку можно построить как функцию от полной и достаточной статистики.

**Определение.** Пусть  $x_i \sim \mathcal{P}_\theta$ ,  $\theta \in \Theta$ . Статистика  $T(\mathbf{x})$  называется *достаточной* для параметра  $\theta$ , если при любом  $t$  и  $B \in \mathfrak{B}(\mathbb{R}^n)$  распределение  $P(\mathbf{x} \in B \mid T = t)$  не зависит от  $\theta$ .

Таким образом, известное и фиксированное значение достаточной статистики дает всю информацию о параметре (и выборка тогда не нужна).

*Утверждение* (Факторизация Неймана–Фишера).  $T$  достаточна, тогда и только тогда, когда

$$L(\theta \mid \mathbf{x}) \stackrel{\text{ae}}{=} h(\mathbf{x})\Psi(T, \theta).$$

**Пример.** Пусть  $\xi \sim \text{Pois}(\lambda)$ , тогда

$$L(\lambda \mid \mathbf{x}) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \underbrace{\frac{1}{\prod_{i=1}^n x_i!}}_{h(\mathbf{x})} \underbrace{e^{-n\lambda} \lambda^{n\bar{x}}}_{\Psi(n\bar{x}, \lambda)} \implies T = n\bar{x}.$$

**Пример.** Пусть  $\xi \sim U(0, \theta)$ , тогда

$$L(\lambda \mid \mathbf{x}) = \prod_{i=1}^n \frac{1}{\theta} \mathbf{1}_{[0, \theta]}(x_i) = \underbrace{\frac{1}{\theta^n} \mathbf{1}(x_{(n)} < \theta)}_{\Psi(x_{(n)}, \theta)} \underbrace{\mathbf{1}(x_{(1)} \geq 0)}_{h(\mathbf{x})} \implies T = x_{(n)}.$$

**Пример.** Пусть  $\xi \sim N(\mu, \sigma^2)$ , тогда проверяется, что для  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = \Psi\left(\left(n\bar{x}^2, n\bar{x}\right), \boldsymbol{\theta}\right).$$

**Определение.** Статистика  $T$  называется *полной*, если для борелевской  $g$

$$Eg(T) = 0 \implies g(T) = 0, \quad \forall \theta \in \Theta.$$

**Следствие.** Оценка, являющаяся функцией от  $S$ , единственна в классе оценок с таким же смещением.

*Доказательство.* Пусть их две:  $\hat{\theta}_n^{(1)}(T)$  и  $\hat{\theta}_n^{(2)}(T)$ . Тогда  $E\left(\hat{\theta}_n^{(1)}(T) - \hat{\theta}_n^{(2)}(T)\right) = Eg(T) = 0$ , значит  $g(T) = 0 = \hat{\theta}_n^{(1)}(T) - \hat{\theta}_n^{(2)}(T)$ .  $\square$

**Теорема** (Рао–Блэкуэлл–Колмогоров). Если  $T$  — полная и достаточная, то оценка  $\hat{\theta}(T)$ , являющаяся функцией от  $T$ , эффективна в классе оценок с таким же смещением.

*Доказательство.* Для любой  $\hat{\theta}_n \in K_b$ ,

$$\begin{aligned} \text{MSE } \hat{\theta}_n &= E(\hat{\theta}_n - \theta)^2 = E(\hat{\theta}_n - \hat{\theta}_n(T) + \hat{\theta}_n(T) - \theta)^2 = E(\hat{\theta}_n - \hat{\theta}_n(T))^2 + E(\hat{\theta}_n(T) - \theta)^2 \\ &\geq E(\hat{\theta}_n(T) - \theta)^2 = \text{MSE } \hat{\theta}_n(T). \end{aligned}$$

Использовано равенство

$$\begin{aligned} E(\hat{\theta}_n - \hat{\theta}_n(T))(\hat{\theta}_n(T) - \theta) &= E\left(E\left\{(\hat{\theta}_n - \hat{\theta}_n(T))(\hat{\theta}_n(T) - \theta) \mid T\right\}\right) \\ &= E\left((\hat{\theta}_n(T) - \theta)E\left\{(\hat{\theta}_n - \hat{\theta}_n(T)) \mid T\right\}\right) = 0, \end{aligned}$$

потому что  $E\left\{\hat{\theta}_n \mid T\right\} = \hat{\theta}_n(T)$  по полноте  $T$ , так что и  $E\left\{(\hat{\theta}_n - \hat{\theta}_n(T)) \mid T\right\} = 0$ .  $\square$



## Часть II.

### Некоторые распределения, связанные с нормальным

# 1. Распределение $\chi^2(m)$

**Определение** (Распределение  $\chi^2(m)$ ).  $\eta$  имеет распределение  $\chi^2$  с  $m$  степенями свободы:

$$\eta \sim \chi^2(m) \iff \eta = \sum_{i=1}^m \zeta_i^2, \quad \zeta_i \sim N(0, 1), \quad \zeta_i \text{ независимы.}$$

**Свойства**<sup>1</sup>  $\chi^2(m)$

$$\begin{aligned} E\eta &= \sum_{i=1}^m E\zeta_i^2 = m \\ D\eta &= 2m \end{aligned}$$

*Утверждение.* Пусть  $\eta_m \sim \chi^2(m)$ . Тогда, по ЦПТ,

$$\frac{\eta_m - E\eta_m}{\sqrt{D\eta_m}} = \frac{\eta_m - m}{\sqrt{2m}} \xrightarrow{d} N(0, 1).$$

**Пример.**  $m = 50$ ,  $\eta_m = 80$ . Тогда

$$\frac{80 - 50}{10} = 3$$

и

$$\text{cdf}_{\chi^2(50)}(80) = 0.9955 \approx \Phi(3) = 0.9986.$$

**Предложение.**  $\chi^2(m)/m \xrightarrow{m \rightarrow \infty} 1$ .

*Доказательство.* По ЗБЧ. □

---

<sup>1</sup>Вычисление  $D\eta$ : <https://www.statlect.com/probability-distributions/chi-square-distribution>

## 2. Распределение Стьюдента $t(m)$

**Определение** (Распределение  $t(m)$ ).  $\xi$  имеет распределение Стьюдента с  $m$  степенями свободы, если

$$\xi \sim t(m) \iff \xi = \frac{\zeta}{\sqrt{\eta/m}}, \quad \zeta \sim N(0, 1), \quad \eta \sim \chi^2(m).$$

**Свойства  $t(m)$**

- При  $m = 1$  это распределение Коши.
- При  $m > 1$ ,  $E\xi = 0$  по симметричности.
- При  $m > 2$ ,  $D\xi = m/(m - 2)$ .
- При  $m > 3$ ,  $A\xi = 0$  по симметричности.
- При  $m > 4$ ,  $K\xi = 6/(m - 4)$ .

**Предложение.** *Распределение Стьюдента сходится к стандартному нормальному:*

$$t \Rightarrow N(0, 1).$$

*Соображения по поводу.*  $D\xi \rightarrow 1$ ,  $K\xi \rightarrow 0$ .

□

### 3. Распределение Фишера

**Определение.** Распределение Фишера имеет вид

$$F(m, k) = \frac{\chi^2(m)/m}{\chi^2(k)/k}.$$

*Замечание.*  $F(1, k) \sim t^2(k)$ ;  $F(m, \infty) = \chi^2(m)$  потому что  $\chi^2(k)/k \xrightarrow[k \rightarrow \infty]{} 1$ .

## 4. Квадратичные формы от нормально распределенных случайных величин

Пусть  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^\top \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p)$ ,  $\mathbf{B}$  — симметричная, неотрицательно определенная матрица. Найдем распределение  $\boldsymbol{\xi}^\top \mathbf{B} \boldsymbol{\xi}$ .

*Утверждение.* Пусть  $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p)$ ,  $\mathbf{B}, \mathbf{C}$  — симметричные матрицы размерности  $p \times p$ . Тогда  $\boldsymbol{\xi}^\top \mathbf{B} \boldsymbol{\xi} \perp \boldsymbol{\xi}^\top \mathbf{C} \boldsymbol{\xi} \iff \mathbf{BC} = \mathbf{0}$ .

**Пример** (Независимость  $\bar{\mathbf{x}}^2$  и  $s^2$ ). Запишем

$$\begin{aligned}\bar{\mathbf{x}}^2 &= \frac{1}{n^2} \left( \sum_{i=1}^n x_i \right)^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n x_i x_j = \frac{1}{n} \mathbf{x} \underbrace{\begin{pmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{pmatrix}}_{\mathbf{B}} \mathbf{x}^\top \\ s^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n} \mathbf{x} \mathbf{B} \mathbf{x}^\top = \frac{1}{n} \left( \mathbf{x} \mathbf{I}_n \mathbf{x}^\top - \mathbf{x} \mathbf{B} \mathbf{x}^\top \right) = \frac{1}{n} \mathbf{x} \underbrace{\begin{pmatrix} 1 - 1/n & \dots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \dots & 1 - 1/n \end{pmatrix}}_{\mathbf{C} = \mathbf{I}_n - \mathbf{B}} \mathbf{x}^\top.\end{aligned}$$

Таким образом,  $n\bar{\mathbf{x}}^2 = \mathbf{x} \mathbf{B} \mathbf{x}^\top$  и  $ns^2 = \mathbf{x} \mathbf{C} \mathbf{x}^\top$ . Но

$$\mathbf{BC} = \mathbf{B}(\mathbf{I}_n - \mathbf{B}) = \mathbf{B} - \mathbf{B}^2 = \mathbf{0},$$

так как

$$\mathbf{B}^2 = \begin{pmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{pmatrix}^2 = \begin{pmatrix} n \cdot 1/n & \dots & n \cdot 1/n \\ \vdots & \ddots & \vdots \\ n \cdot 1/n & \dots & n \cdot 1/n \end{pmatrix} = \mathbf{B}.$$

Значит,  $\bar{\mathbf{x}}^2 \perp s^2$ .

Видно, что  $\sigma^{-2} \boldsymbol{\xi}^\top \mathbf{I}_p \boldsymbol{\xi} \sim \chi^2(p)$ . На самом деле справедливо

*Утверждение.* Пусть  $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p)$ ,  $\mathbf{B}$  — симметричная, неотрицательно неопределенная матрица размерности  $p \times p$  и  $\text{rk } \mathbf{B} = r$ . Тогда

$$\sigma^{-2} \boldsymbol{\xi}^\top \mathbf{B} \boldsymbol{\xi} \sim \chi^2(r) \iff \mathbf{B}^2 = \mathbf{B}.$$

**Пример.** Покажем, что

$$n\sigma^{-2}s^2 \sim \chi^2(p-1).$$

Воспользуемся представлением из предыдущего примера:  $ps^2 = \mathbf{x}^\top \mathbf{C} \mathbf{x}$ . Но  $\text{rk } \mathbf{C} = \text{rk}(\mathbf{I}_p - \mathbf{B}) = p-1$ ;  $\mathbf{B}^2 = \mathbf{B}$ , значит  $p\sigma^{-2}s^2 \sim \chi^2(p-1)$ .

*Утверждение* (Cochran). Пусть  $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_p)$ ,  $\boldsymbol{\xi}^\top \boldsymbol{\xi} = \sum_i Q_i$ , где  $Q_i$  — квадратичная форма, заданная  $\mathbf{B}_i$ ,  $\text{rk } \mathbf{B}_i = r_i$ . Тогда следующие утверждения эквивалентны:

1.  $\sum r_i = p$
2.  $Q_i \sim \chi^2(r_i)$
3.  $Q_i \perp Q_j, \quad \forall i \neq j$ , т.е.  $\mathbf{B}_i \mathbf{B}_j = \mathbf{0}$ .

Часть III.

## Проверка гипотез

Этот раздел иногда называется «Confirmatory Data Analysis» в противовес «Exploratory Data Analysis», не включающему в себя понятие *гипотезы*.

# 1. Построение критерия

## 1.1. Понятие гипотезы и критерия

Пусть  $x_1, \dots, x_n \sim \mathcal{P}$ ,  $\mathcal{P}$  — множество всех распределений. О  $\mathcal{P}$  возможно делать утверждения вида  $\mathcal{P} \in \mathcal{P}' \subset \mathcal{P}$ . Стоит задача выбрать такое утверждение, что оно некоторым наилучшим образом соответствует выборке.

**Определение.** *Модель* — это предположение о выделенном классе  $\mathcal{P}_M \subset \mathcal{P}$ , которому принадлежит  $\mathcal{P}$  (допустим,  $\mathcal{P}_M = \{N(\mu, \sigma_0)\}$ , где  $\sigma_0$  — фиксированное значение). Иными словами, это утверждение о  $\mathcal{P}$ , которое считается верным и не проверяется.

**Определение.** *Гипотеза* — утверждение о  $\mathcal{P}$ , требующее проверки. Гипотеза называется *простой*, если она соответствует только одному распределению в рамках рассматриваемой модели:

$$H : \mathcal{P} = \mathcal{P}_0 \in \mathcal{P}_M$$

(например,  $\mathcal{P}_0 = N(\mu_0, \sigma_0)$ ) или *сложной*, если целому множеству:

$$H : \mathcal{P} \in \mathcal{P}' \subset \mathcal{P}_M$$

(например,  $\mathcal{P}' = \{N(\mu, \sigma_0) : \mu > 0\}$ ).

Очень часто возникает (и далее рассматривается) случай выдвижения только лишь двух гипотез:  $H_0 : \mathcal{P} \in \mathcal{P}_0 \subset \mathcal{P}$  — *нулевой*, основной и  $H_1 : \mathcal{P} \in \mathcal{P}_1 \subset \mathcal{P}$  — *альтернативной*.  $H_1$  учитывает отклонения от  $H_0$ , обнаружение которых желательно. Возможны варианты:

- простая  $H_0$  и простая  $H_1$ ;
- простая  $H_0$  и сложная  $H_1$ ;
- сложная  $H_0$  и сложная  $H_1$ .

**Определение.** *Критерий* есть отображение

$$\varphi : \mathbf{x} \mapsto \{H_0, H_1\}.$$

Критерий «решает», противоречат или не противоречат выдвинутой гипотезе выборочные данные.

**Определение.** Говорят, что гипотеза *отвергается*, если  $\varphi(\mathbf{x}) = H_1$  и *не отвергается* иначе.

Так как в заданной постановке любой критерий принимает не более двух значений, то  $\text{dom } \varphi$  разбивается на два дизъюнктивных множества  $\mathcal{A}_{\text{крит}}$  и  $\mathcal{A}_{\text{дов}}$ , называемых *критической* и *доверительной* областями, таких, что

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \mathbf{x} \in \mathcal{A}_{\text{дов}} \\ H_1 & \mathbf{x} \in \mathcal{A}_{\text{крит}}. \end{cases}$$

Поскольку выборка конечного объема позволяет делать только вероятностные заключения, со статистическим критерием ассоциированы ошибки  $i$ -ых родов.



## 1. Построение критерия

**Определение.** Говорят, что произошла *ошибка  $i$ -го рода* критерия  $\varphi$ , если критерий отверг верную гипотезу  $H_{i-1}$ . Соответствующие вероятности обозначаются

$$\alpha_i(\varphi) = P_{H_{i-1}}(\varphi(\mathbf{x}) \neq H_{i-1}).$$

Поскольку в рассмотрение введены только  $H_0$  и  $H_1$ , возможны ошибки *I-го рода* — принятие случайного различия за систематическое и *II-го рода* — принятие наблюдаемого различия за случайный эффект с соответствующими вероятностями

$$\begin{aligned}\alpha_I &= P_{H_0}(\varphi(\mathbf{x}) \neq H_0) = P_{H_0}(\mathbf{x} \in \mathcal{A}_{\text{крит}}) \\ \alpha_{II} &= P_{H_1}(\varphi(\mathbf{x}) \neq H_1) = P_{H_1}(\mathbf{x} \in \mathcal{A}_{\text{дов}}).\end{aligned}$$

*Замечание.* Если  $H_i$  — сложная гипотеза, то  $\alpha_{i+1}(\varphi)$  будет зависеть от того, на каком именно распределении  $\mathcal{P}$ , отвечающем  $H_i$ , вычисляется эта вероятность.

**Определение.** *Мощность* критерия против альтернативы это вероятность справедливо отвергнуть  $H_0$ :

$$\beta = 1 - \alpha_{II} = 1 - P_{H_1}(\varphi(\mathbf{x}) = H_0) = P_{H_1}(\varphi(\mathbf{x}) = H_1).$$

Иными словами, это способность критерия отличать  $H_0$  от  $H_1$ .

## 1.2. Построение оптимальных критериев

Рассмотрим пример.

**Пример.** Пусть  $\xi \sim N(\mu, 1)$  и  $\mathbf{x} = \{x\}$ . Выдвинем  $H_0 : \mu = 0$  против простой альтернативы  $H_1 : \mu = 1$ . Рассмотрим критерий

$$\varphi_b(\mathbf{x}) = \begin{cases} H_0 & x < b \\ H_1 & x \geq b. \end{cases}$$

(ФИХМЕ: Рисунок) Ясно, что из всего множества критериев  $\{\varphi_b\}$ , не все одинаково хорошо описывают нормальную выборку: так, тождественный  $\varphi_\infty(\mathbf{x}) \equiv H_0$  будет вести себя хуже любого  $\varphi_b$ ,  $b < \infty$ .

Выбирать оптимальный критерий можно тремя способами: минимаксным, Байесовским подходами и выбором наиболее мощного критерия.

**Определение.**  $\varphi^{(1)}$  не хуже  $\varphi^{(2)}$  в *минимаксном* смысле, если

$$\max(\alpha_I(\varphi^{(1)}), \alpha_{II}(\varphi^{(1)})) \leq \max(\alpha_I(\varphi^{(2)}), \alpha_{II}(\varphi^{(2)})).$$

Если  $\phi^*$  не хуже всех остальных в этом смысле, то он называется *минимаксным*.

**Пример.**  $\varphi_{1/2}$  минимаксный.

**Определение.** Пусть известны  $r = P(H_0)$ ,  $s = 1 - r = P(H_1)$  (или задана линейная функция потерь, равная  $r$  в случае ошибки 1-го рода и  $s$  — второго). Тогда  $\varphi^{(1)}$  не хуже  $\varphi^{(2)}$  в *байесовском* смысле, если

$$r\alpha_I(\varphi^{(1)}) + s\alpha_{II}(\varphi^{(1)}) \leq r\alpha_I(\varphi^{(2)}) + s\alpha_{II}(\varphi^{(2)}).$$

Если  $\phi^*$  не хуже всех остальных в этом смысле, то он называется *байесовским*.

*Замечание.* Короче говоря, по правилу полной вероятности это вероятность ошибки критерия:

$$P(H_0) \underbrace{P(\varphi(\mathbf{x}) = H_1 \mid H_0)}_{\text{Err}} + P(H_1) \underbrace{P(\varphi(\mathbf{x}) = H_0 \mid H_1)}_{\text{Err}} = P(\text{Err}).$$

**Пример.**  $\varphi_{1/2}$  байесовский с  $s = r$ .

## 1. Построение критерия

**Определение.** Пусть до эксперимента зафиксирован *уровень значимости*<sup>1</sup> критерия  $\alpha \in [0, 1]$ . Критерий

$$\varphi^* \in K_\alpha = \{\varphi(\mathbf{x}) \mid \alpha_I(\varphi) \leq \alpha\}$$

называется *наиболее мощным* критерием, если

$$\alpha_{II}(\varphi^*) \leq \alpha_{II}(\varphi), \quad \forall \varphi \in K_\alpha.$$

*Замечание.* Стандартные уровни значимости:  $\alpha = 0.05$  или  $\alpha = 0.01$ .

Все три подхода могут быть сведены к универсальному критерию — критерию отношения правдоподобия.

В примере 1.2 интуитивно ясно, что следует выбрать ту гипотезу, значение плотности которой в точке  $x$  больше другой. В случае, если  $n > 1$ , справедливо взять произведение плотностей — т.е. функций правдоподобия и рассматривать их отношение

$$L(\mathbf{x}) = \frac{L_2(\boldsymbol{\theta} \mid \mathbf{x})}{L_1(\boldsymbol{\theta} \mid \mathbf{x})}.$$

Выбор гипотезы затем делать по тому, больше или меньше  $L$  единицы. Однако, чтобы учесть произвольный уровень ошибки, следует сравнивать не с 1, а с константой  $c$ .

**Определение.** Пусть  $\mathcal{P}_0, \mathcal{P}_1$  либо одновременно дискретны, либо непрерывны. Пусть также  $\neg \exists c_0 : P_{H_0}(L(\mathbf{x}) = c_0) = 0$  (в противном случае критерий не сможет различить гипотезы на множестве не-нулевой меры) — т.е.,  $P_{H_0}(L(\mathbf{x}) \geq c)$  непрерывна по  $c > 0$ . Тогда *критерием отношения правдоподобия* называется

$$\varphi_c(\mathbf{x}) = \begin{cases} H_0 & L(\mathbf{x}) < c \\ H_1 & L(\mathbf{x}) \geq c. \end{cases}$$

### 1.2.1. Явный вид оптимальных критериев

*Утверждение.* В предположениях из определения, критерий отношения правдоподобия является

1. минимаксным при  $c : \alpha_I(\varphi_c) = \alpha_{II}(\varphi_c)$ ;
2. байесовским при заданных  $r, s : c = r/s$ ;
3. (лемма Неймана-Пирсона) наибольшей мощности при заданном  $\alpha : \alpha_I(\varphi_c) = \alpha$ .

**Пример.** Пусть  $\xi \sim N(\mu, 1)$ .  $H_0 : \mu = \mu_0$ ,  $H_1 : \mu = \mu_1 > \mu_0$ . Критическая область задается неравенством

$$L(\mathbf{x}) = \frac{L_2(\boldsymbol{\theta} \mid \mathbf{x})}{L_1(\boldsymbol{\theta} \mid \mathbf{x})} = \exp \left( \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2 - (x_i - \mu_1)^2 \right) \geq c,$$

упрощая которое получают

$$\bar{x} \geq \frac{\ln c}{(\mu_1 - \mu_0)n} + \frac{\mu_1 - \mu_0}{2}.$$

- Пусть критерий байесовский с  $r = 1/4$ ,  $s = 3/4$ ; тогда  $c = 1/3$ .
- Пусть критерий наибольшей мощности с заданным  $\alpha$ , критическая область задается  $\bar{x} \geq c_0$ . Тогда

$$\alpha_I = P_{H_0}(\bar{x} \geq c_0) = P_{H_0}(\sqrt{n}(\bar{x} - \mu_0) \geq \sqrt{n}(c_0 - \mu_0)) = 1 - \text{cdf}_{N(0,1)}(\sqrt{n}(c_0 - \mu_0)) = \alpha$$

если

$$c_0 = \frac{z_{1-\alpha}}{\sqrt{n}} + \mu_0.$$

<sup>1</sup>Неформально,  $\alpha$  обратно пропорциональна «строгости» критерия, выбираемой экспериментатором.

## 1. Построение критерия

- Пусть критерий минимаксный. Тогда

$$\alpha_{\Pi} = \text{cdf}_{N(0,1)}(\sqrt{n}(c_0 - \mu_1)) = 1 - \text{cdf}_{N(0,1)}(\sqrt{n}(c_0 - \mu_0)) = \text{cdf}_{N(0,1)}(\sqrt{n}(\mu_0 - c_0))$$

если

$$c_0 - \mu_1 = \mu_0 - c_0, \quad c_0 = \frac{\mu_0 + \mu_1}{2}.$$

**Определение.** Критерий называется *состоятельным* против альтернативы  $H_1$ , если  $\forall \mathcal{P}_1 \in \mathcal{P}_1$

$$\beta(\varphi, \mathcal{P}_1) = 1 - \mathbb{P}_{\mathcal{P}_1}(\varphi(\mathbf{x}) = H_0) \xrightarrow{n \rightarrow \infty} 1.$$

**Пример.** Критерий наибольшей мощности

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \bar{\mathbf{x}} < \frac{z_{1-\alpha}}{\sqrt{n}} + \mu_0 \\ H_1 & \text{иначе} \end{cases}$$

является состоятельным, потому что

$$\alpha_{\Pi}(\varphi) = \mathbb{P}_{H_1} \left( \bar{\mathbf{x}} - \frac{z_{1-\alpha}}{\sqrt{n}} < \mu_0 \right),$$

но, при верной  $H_1$ ,

$$\xi_n = \bar{\mathbf{x}} - \frac{z_{1-\alpha}}{\sqrt{n}} \xrightarrow{\mathbb{P}} \mu_1,$$

из чего следует слабая сходимостъ — сходимостъ  $\text{cdf}_{\xi_n}(x) \rightarrow \text{cdf}_{\mu_1}(x) = \mathbb{P}(\mu_1 < x)$  по всех точках; учитывая  $\text{cdf}_{\mu_1}(\mu_0) = 0$ ,

$$\alpha_{\Pi}(\varphi) = F_{\xi_n}(\mu_0) \rightarrow F_{\mu_1}(\mu_0) = 0.$$

**Определение.** Если

$\alpha_I = \alpha$  то критерий называется *точным*,

$\alpha_I \xrightarrow{n \rightarrow \infty} \alpha$  *асимптотическим*,

$\alpha_I > \alpha$  *радикальным* (т.е. отвергает гипотезу чаще, чем точный),

$\alpha_I < \alpha$  *консервативным* (если гипотеза отвергнута, то уж наверняка).

### 1.2.2. О постановке $H_0$

Задача может допускать две постановки; в этом случае, поскольку  $\alpha_I (= \alpha$  для правильно построенного критерия) контролируется экспериментатором, проверяется отрицание эффекта, который хотят подтвердить: к примеру, что новое лекарство *не* лучше старого; если  $H_0$  отвергнется, это будет означать, что новое лекарство-таки лучше старого с вероятностью не ниже, чем  $1 - \alpha$ .

**Пример** (С гранатами). FIXME

### 1.2.3. Об отвержении гипотезы

Утверждать об *отвержении* гипотезы можно с вероятностью ошибки  $\alpha$  (достаточно малой и произвольно задаваемой экспериментатором); утверждать о *принятии* гипотезы можно с вероятностью ошибки  $\alpha_{\Pi}$  — не контролируемой и потенциально довольно большой. Иными словами, попадание в доверительную область может означать как то, что  $H_0$  верна, так и то, что верна  $H_1$ , но для распознавания этого не хватило мощности. Поэтому безопасно гипотезу можно только отвергать или не отвергать. Можно и принять, если известна мощность критерия против всех возможных альтернатив, экспериментатора устраивающая.

## 1. Построение критерия

При высокой вероятности ошибки II-го рода возможна ситуация не отвержении заведомо ложной гипотезы. Это, в свою очередь, может произойти из-за маленького объема выборки (критерий не находит разницу, см. 1.4.1). Чем больше объем выборки, тем мощность больше, но возможна ситуация, когда критерий чувствителен настолько, что находит разницу там, где не должен — например, при генерации «идеальным» датчиком случайных чисел, начиная с какого-то объема заведомо истинная гипотеза может быть отвергнута из-за ошибок в точности представления чисел с плавающей точкой.

**Определение.** Критерий называется *одно- (двух-) сторонним* по тому, где находится альтернатива.

**Определение.** Критическая область называется *одно- (дву-) сторонней* по тому, где формально располагается  $\mathcal{A}_{\text{крит}}$ .

### 1.3. Построение критерия при помощи статистики критерия

**Определение.** Статистика критерия есть отображение

$$T : \mathbf{x} \mapsto y \in \mathbb{R}$$

такое, что при верной  $H_0$ ,  $T \xrightarrow{d} Q$ , где  $Q$  — полностью известное непрерывное распределение, а при верной  $H_1$  известно поведение  $T$ .

Поскольку распределение  $T$  при верной  $H_0$  известно, она должна вести себя как любая другая случайная величина из  $Q$  — попадание в некоторые области менее вероятно, чем в другие. Поэтому разумно разбить  $\text{supp } T$  по уровню значимости  $\alpha$  на  $\mathcal{A}_{\text{крит}} \sqcup \mathcal{A}_{\text{дов}}$  так, что попадание в  $\mathcal{A}_{\text{крит}}$  при верной  $H_0$  происходит с заранее зафиксированной (малой) вероятностью  $\alpha$ . Значит, если  $T(\mathbf{x}) \in \mathcal{A}_{\text{крит}}$ , то с некоторой же вероятностью можно заявлять об отвержении  $H_0$ . Таким образом,  $T$  измеряет то, насколько выборка соответствует гипотезе.

### 1.4. Разбиение на доверительную и критические области

Разберем на примере построение разбиения. Пусть  $\xi \sim N(\mu, \sigma^2)$ ,  $H_0 : \mu = \mu_0$  и фиксирован  $\alpha$ . По ??, используется статистика

$$T = z = \sqrt{n} \frac{\bar{\mathbf{x}} - a_0}{\sigma} \sim N(0, 1).$$

В зависимости от  $H_1$ , возможны варианты.

#### 1.4.1. Простая альтернатива

Пусть  $H_1 : \mu = \mu_1$ , причем  $\mu_1 > \mu_0$ . Тогда, поскольку при верной  $H_1$ ,  $E\bar{\mathbf{x}} = 1/n \cdot \sum_{i=1}^n \xi_i = n/n \cdot \mu_1$ , то

$$ET = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \implies T \sim N\left(\frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}, 1\right) \text{ при верной } H_1.$$

(дисперсия, конечно, не меняется при сдвиге).

## 1. Построение критерия

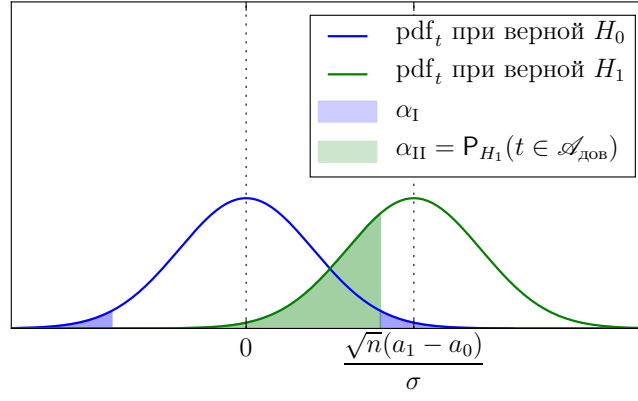


Рис. 1.1.: Плотности распределения  $z$  (неоптимальное разбиение)

Чтобы минимизировать  $\alpha_{II}$ , логично определить  $\mathcal{A}_{\text{крит}}$  только на одном хвосте — с той стороны, где находится альтернатива. Помимо этого, по рисунку видно, что минимизировать  $\alpha_{II}$  (согласившись на большую ошибку первого рода) можно сдвинув вправо центр второй плотности, увеличив  $n$ . Аналогично, чем  $\mu_1$  дальше от  $\mu_0$ , тем  $\alpha_{II}$  меньше.

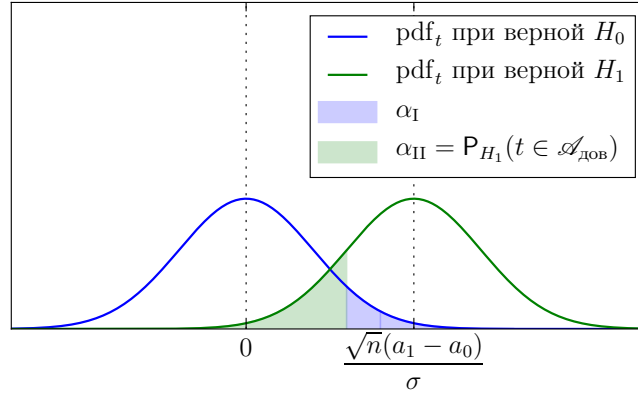


Рис. 1.2.: Плотности распределения  $z$  (оптимальное разбиение)

Таким образом,  $\mathcal{A}_{\text{крит}} = (C, +\infty)$ ,  $C = z_{1-\alpha}$ .

Разумеется, если  $\mu_1 < \mu_0$ , то  $\mathcal{A}_{\text{крит}} = (-\infty, C)$ ,  $C = z_\alpha$ .

### 1.4.2. Односторонний критерий (сложная альтернатива)

В общем случае, пусть  $H_1 : \mu = \mu_1 \ \forall \mu_1 > \mu_0$ ; тогда по ЗБЧ  $\bar{x} - \mu_0 \rightarrow \mu_1 - \mu_0 > 0$  и  $T \rightarrow +\infty$ . Следовательно, чтобы максимизировать величину  $\beta = P_{H_1}(\mathbf{x} \in \mathcal{A}_{\text{крит}}^{(\alpha)})$ , следует разместить  $\mathcal{A}$  на правом хвосте плотности:  $\mathcal{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$ . Для  $H_1 : \mu = \mu_1 \ \forall \mu_1 < \mu_0$  аналогично  $\mathcal{A}_{\text{крит}} = (-\infty, z_\alpha)$ .

### 1.4.3. Двусторонний критерий

Пусть  $H_1 : E\xi = \mu_1 \neq \mu_0$ ; тогда по ЗБЧ  $\bar{x} - \mu_0 \rightarrow \mu_1 - \mu_0$  и  $|T| \xrightarrow[n \rightarrow \infty]{} \infty$ , откуда  $\mathcal{A}_{\text{крит}} = \mathbb{R} \setminus (z_{\alpha/2}, z_{1-\alpha/2}) = \mathbb{R} \setminus (-C, C)$ , где  $C = -z_{\alpha/2}$ .

В общем виде, с использованием статистики, критерий может быть определен как

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & |T(\mathbf{x})| < C \\ H_1 & |T(\mathbf{x})| \geq C, \end{cases}$$

где критическое значение  $C$  определяется из уравнения  $\alpha = P(|T| \geq C)$ .

## 1. Построение критерия

Иногда вместо сравнения значения  $T$  с критическим вычисляют *реально достигнутый уровень значимости критерия* (« $p$ -value»).

**Определение.**  $p$ -value значения статистики  $T$  на выборке  $\mathbf{x}$  есть вероятность, взяв выборку из распределения  $H_0$ , получить по ней большее отклонение  $|T(\mathbf{x})|$  эмпирического от истинного распределения, чем получено по проверяемой выборке:

$$p\text{-value} = \alpha^* = P_{H_0}(|T| \geq |T(\mathbf{x})|).$$

Значит, критерий может быть задан и как

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \alpha^* > \alpha \\ H_1 & \alpha^* \leq \alpha. \end{cases}$$

*Замечание.*  $p$ -value обратно пропорционален «существенности» результата.

*Замечание.* Пусть  $\alpha^* = 0.05$ . Это значит, что в среднем всего лишь 5% «контрольных» выборок, удовлетворяющих основной гипотезе, будут обладать большим отклонением  $|T(\mathbf{x})|$  по сравнению с тестируемой выборкой — последняя ведет себя не хуже, чем 5% «правильных» выборок.

### 1.5. Схема построения критерия с помощью статистики

1. Фиксируют предположение относительно данных.
2. Выдвигают  $H_0$  и  $H_1$ .
  - $H_0$  формулируется согласно замечанию 1.2.2.
  - $H_1$  ставится по смыслу задачи (см. далее).
3. Выбирают подходящий критерий и статистику  $T$ .
4. Фиксируют уровень значимости  $\alpha$ .
5. Строят разбиение  $\text{im } T$  с помощью квантилей распределения  $T$  (при верной  $H_0$ ) так, чтобы  $\alpha_1 = \alpha$ ; положение квантилей выбирают из известного поведения статистики при верной  $H_1$ .
6. Считают значение статистики и принимают решение об отвержении  $H_0$  одним из способов:

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & |T(\mathbf{x})| < C \\ H_1 & |T(\mathbf{x})| \geq C, \end{cases} \quad \varphi(\mathbf{x}) = \begin{cases} H_0 & \alpha^* > \alpha \\ H_1 & \alpha^* \leq \alpha. \end{cases}$$

**Пример** (Средняя температура в холодильнике). Хотят купить холодильник, такой, чтобы температура держалась в окрестности 0. Известно количество измерений  $n = 25$  и  $\bar{\mathbf{x}} = 0.7$ .

1. Пусть  $\xi \sim N(\mu, 4)$ .
2. Выдвинута  $H_0 : E\xi = \mu_0 = 0$  — если гипотеза опровергнется, то холодильник не купят;  
 $H_1 : E\xi = \mu_1 \neq \mu_0$ .
3. Поскольку модель нормальная и известная  $\sigma^2$ , выберем статистику ?? (« $z$ -test»):

$$z = \frac{\sqrt{n}(\bar{\mathbf{x}} - \mu_0)}{\sigma} \sim N(0, 1) \text{ при верной } H_0.$$

Идеальное значение статистики — 0.

4. Зафиксируем два уровня значимости:  $\alpha^{(1)} = 0.2$  (храним петрушку) и  $\alpha^{(2)} = 0.01$  (храним дорогую красную икру).

## 1. Построение критерия

5. Построим разбиение. Поскольку  $\mu_1 \neq \mu_0$ , то  $\mathcal{A}_{\text{крит}} = \mathbb{R} \setminus (z_{\alpha/2}, z_{1-\alpha/2})$ . Для введенных уровней значимости это означает

- а)  $\mathcal{A}_{\text{крит}}^{(\alpha^{(1)})} \approx \mathbb{R} \setminus (-1.28, 1.28)$ .
- б)  $\mathcal{A}_{\text{крит}}^{(\alpha^{(2)})} \approx \mathbb{R} \setminus (-2.576, 2.576)$ .

6. Посчитаем

$$z(\mathbf{x}) = \frac{\sqrt{n}(\bar{\mathbf{x}} - \mu_0)}{\sigma} = \frac{5(0.7 - 0)}{2} = 1.75.$$

Дальнейшее принятие решения возможно на основании критического значения или  $p$ -value.

- По вычислению критического значения:

◇  $z \in \mathcal{A}_{\text{крит}}^{(\alpha^{(1)})}$ ,  $H_0$  отвергается, холодильник не покупают.

◇  $z \in \mathcal{A}_{\text{дов}}^{(\alpha^{(2)})}$ ,  $H_0$  не отвергается, холодильник, быть может, покупают.

- Можно посчитать  $p$ -value:

$$2 \cdot (1 - \text{cdf}_{N(0,1)}(1.75)) \approx 0.08.$$

Поэтому при уровне значимости  $\alpha^{(1)} = 0.2 > 0.08$   $H_0$  отвергается, а при  $\alpha^{(2)} = 0.01 < 0.08$  не отвергается.

**Пример** (С мышой). В одном из рукавов Т-образного лабиринта лежит морковка. К развилке по лабиринту бежит мышь и 7 раз из 10 поворачивает в направлении морковки. На основании этих данных хотим сделать вывод, что мышь чует морковь на расстоянии, после чего написать научную статью.

- $\xi \sim \text{Ber}(p)$ . Выдвинем гипотезу, что мышь *не* чует морковку,  $H_0 : p = p_0 = 0.5$ . Поскольку  $E\xi = p$ , воспользуемся критерием для проверки гипотезы о значении среднего с идеальным значением 0; учитывая  $D\xi = p(1-p)$ ,

$$\begin{aligned} T &= \frac{\sqrt{n}(\bar{\mathbf{x}} - p_0)}{\sqrt{p_0(1-p_0)}} \xrightarrow{d} N(0, 1). \\ &= \frac{\sqrt{10} \cdot 0.2}{0.5} \approx 1.2649 \implies p\text{-value} = 2 \cdot (1 - \text{cdf}_{N(0,1)}(1.2649)) \approx 0.2. \end{aligned}$$

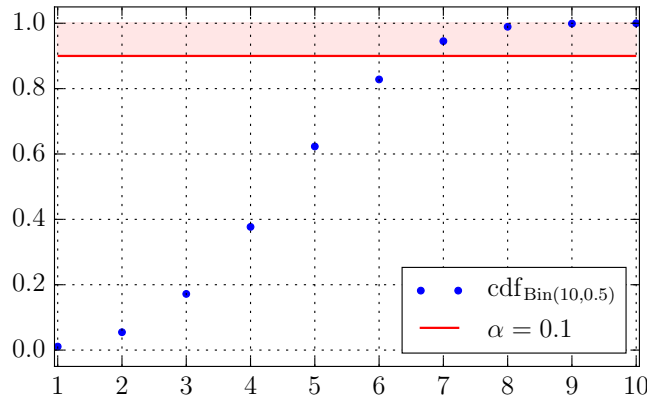
Значит, с уровнем значимости 0.2 гипотеза не отвергается. Хочется иметь, конечно, один из стандартных уровней значимости, например 0.1.

- Увеличим мощность критерия, введя альтернативную гипотезу, что мышь чует морковку (в предположении, что все мыши любят морковь и к ней бегут),  $H_1 : p_1 > p_0$ . По 1.4.1, можем устроить односторонний критерий, так что  $p$ -value теперь 0.1. Однако пользуемся асимптотическим критерием при  $n = 10$ .
- Воспользуемся точным односторонним критерием со статистикой

$$T := n\bar{\mathbf{x}} = \sum_{i=1}^n x_i \sim \text{Bin}(n, p_0)$$

и идеальным значением  $np_0$ . Тогда  $T = 10 \cdot 0.7 = 7$ . При уровне значимости  $\alpha = 0.1$  успешно попадаем в критическую область, вследствие чего  $H_0$  отвергается, и можем опубликоваться.

## 1. Построение критерия



*Замечание.* Исторически существовало два подхода к проверке гипотез: Фишера («significance test») и Неймана-Пирсона («hypothesis testing»).

**Фишер** Выдвигается  $H_0$ . Подсчитывается и сообщается точное  $p$ -value. Если результат «незначительный», не делается никаких выводов об отвержении  $H_0$ , но делается возможным дополнительный сбор данных.

**Нейман-Пирсон** Выдвигаются  $H_1, H_2$ , фиксируются  $\alpha_I, \alpha_{II}$  и  $n$ . На этом основании определяются  $\mathcal{A}_{\text{крит}}$  для каждой гипотезы. Если данные попали в  $\mathcal{A}_{\text{крит}}$   $H_1$  — предпочитается  $H_2$ , иначе  $H_2$ .

Современная теория проверки гипотез есть смесь двух этих подходов, не всегда консистентная. Вводные курсы по статистике формулируют теорию, похожую на significance testing Фишера; при повышенных требованиях к математической строгости, пользуются теорией Неймана-Пирсона.

*Замечание* (О графике  $p$ -values). Поскольку

$$\alpha_I \leftarrow \mathbf{P}_{H_0}(T \in \mathcal{A}_{\text{крит}}) = \mathbf{P}_{H_0}(p\text{-value} < \alpha),$$

то  $p$ -value по распределению стремятся к  $U(0, 1)$  при верной  $H_0$ . Это соображение позволяет визуально проверить истинность гипотезы: достаточно несколько (много) раз произвести эксперимент, для каждой выборки  $\bar{\mathbf{x}}^{(i)}$  посчитать свой  $p$ -value, построить график и убедиться, что получилась прямая. Для подсчета мощности  $\beta = \mathbf{P}_{H_1}(\mathbf{x} \in \mathcal{A}_{\text{крит}}^{(\alpha)}) = \mathbf{P}_{H_1}(p\text{-value} < \alpha)$  считать выборку с параметрами  $H_1$ , а  $T$  относительно  $H_0$ .



## 2. Проверка гипотезы о значении мат. ожидания ( $t$ -критерий)

$H_0 : E\xi = \mu = \mu_0$ . Соответствие оценки математического ожидания гипотезе удобно выражать разницей  $\bar{x} - \mu_0$  с «идеальным» значением 0. Отнормировав эту разницу, получим статистику, распределение которой известно.

### 2.1. $D\xi = \sigma^2 < \infty$

**Предложение.** Пусть  $D\xi = \sigma^2 < \infty$ ; тогда используется следующая статистика

$$t = \sqrt{n} \frac{(\bar{x} - \mu_0)}{\sigma} \xrightarrow{n \rightarrow \infty} N(0, 1)$$

*Доказательство.* По ЦПТ. □

**Предложение.** При условии нормальности данных,  $t$ -критерий называется « $z$ -критерием», причем

$$t = z \sim N(0, 1).$$

*Доказательство.*

$$z = \frac{\bar{x} - \mu_0}{\sqrt{D\bar{x}}} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \sim N(0, 1).$$

□

### Разбиение

$$H_1 : E\xi \neq \mu_0 \quad \mathcal{A}_{\text{крит}} = \mathbb{R} \setminus (z_{\alpha/2}, z_{1-\alpha/2})$$

$$H_1 : E\xi > \mu_0 \quad \mathcal{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$$

$$H_1 : E\xi < \mu_0 \quad \mathcal{A}_{\text{крит}} = (-\infty, z_{\alpha})$$

### 2.2. $D\xi$ неизвестна

**Предложение.** Пусть  $D\xi$  неизвестна; тогда используется следующая статистика

$$t = \sqrt{n-1} \frac{\bar{x} - \mu_0}{s} = \frac{\sqrt{n-1}(\bar{x} - \mu_0)}{\sqrt{n-1}/\sqrt{n} \cdot \tilde{s}} = \sqrt{n} \frac{\bar{x} - \mu_0}{\tilde{s}} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

**Предложение.** При условии нормальности данных,

$$t \sim t(n-1).$$

*Доказательство.*

$$t = \frac{\sqrt{n-1}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{n-1} \left( \frac{\bar{x} - \mu_0}{\sigma} \right)}{s/\sigma} = \frac{\left( \frac{\bar{x} - \mu_0}{\sigma} \right)}{\sqrt{\frac{s^2/\sigma^2}{n-1}}} = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} = \frac{\beta}{\sqrt{\eta/(n-1)}} \sim t(n-1),$$

поскольку

$$\beta = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \sim N(0, 1), \quad \eta = \frac{ns^2}{\sigma^2} \sim \chi^2(n-1).$$

□

## 2. Проверка гипотезы о значении мат. ожидания ( $t$ -критерий)

### Разбиение

$$H_1 : E\xi \neq \mu_0 \quad \mathcal{A}_{\text{крит}} = \mathbb{R} \setminus \left( \text{qnt}_{t(n-1)}(\alpha/2), \text{qnt}_{t(n-1)}(1 - \alpha/2) \right)$$

$$H_1 : E\xi > \mu_0 \quad \mathcal{A}_{\text{крит}} = (\text{qnt}_{t(n-1)}(1 - \alpha), \infty)$$

$$H_1 : E\xi < \mu_0 \quad \mathcal{A}_{\text{крит}} = (-\infty, \text{qnt}_{t(n-1)} \alpha)$$

*Замечание.* При нормальной аппроксимации  $\text{qnt}_{t(n-1)}$  заменить на  $N(0, 1)$ .

### 2.2.1. $z$ -критерий для пропорции в модели Бернулли

Пусть  $\xi \sim \text{Ber}(p)$ . Поскольку  $E\xi = p$ , можно воспользоваться только что введенной статистикой; учитывая  $D\xi = p(1 - p)$ ,

$$T = \sqrt{n} \frac{\bar{\mathbf{x}} - p_0}{\sqrt{p_0(1 - p_0)}} \xrightarrow{d} N(0, 1).$$

Разбиение будет таким же, как и в случае известной дисперсии.

### 3. Проверка гипотезы о значении дисперсии в нормальной модели (критерий $\chi^2$ )

Пусть  $\xi \sim N(\mu, \sigma^2)$ .  $H_0 : D\xi = \sigma^2 = \sigma_0^2$ . Соответствие оценки дисперсии гипотезе удобно выражать отношением  $s^2/\sigma_0^2$  (или  $s_\mu^2/\sigma_0^2$  если  $\mu$  известно) с «идеальным» значением 1. Домножив на  $n$ , получим статистику, распределение которой известно.

*Замечание (Важное).* Критерий работает только в нормальной модели и не становится асимптотически нормальным в ином случае!

#### 3.1. $E\xi = \mu < \infty$

**Предложение.** Пусть  $E\xi = \mu < \infty$ ; При условии нормальности данных используется следующая статистика:

$$\chi^2 = n \frac{s_\mu^2}{\sigma_0^2} \sim \chi^2(n).$$

*Доказательство.*

$$\chi^2 = \frac{ns_\mu^2}{\sigma_0^2} = \frac{n \cdot 1/n \cdot \sum_{i=1}^n (x_i - \mu)^2}{\sigma_0^2} = \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma_0} \right)^2 \sim \chi^2(n).$$

□

#### Разбиение

$$H_1 : D\xi \neq \sigma_0^2 \quad \mathcal{A}_{\text{крит}} = \mathbb{R}_+ \setminus \left( \text{qnt}_{\chi^2(n)}(\alpha/2), \text{qnt}_{\chi^2(n)}(1 - \alpha/2) \right)$$

$$H_1 : D\xi > \sigma_0^2 \quad \mathcal{A}_{\text{крит}} = (\text{qnt}_{\chi^2(n)}(1 - \alpha), \infty)$$

$$H_1 : D\xi < \sigma_0^2 \quad \mathcal{A}_{\text{крит}} = (0, \text{qnt}_{\chi^2(n)} \alpha)$$

#### 3.2. $E\xi$ неизвестно

**Предложение.** Пусть  $E\xi$  неизвестно. При условии нормальности данных используется следующая статистика:

$$\chi^2 = n \frac{s^2}{\sigma_0^2} = (n-1) \frac{\tilde{s}^2}{\sigma_0^2} \sim \chi^2(n-1).$$

*Доказательство.* См. (4).

□

*Альтернативное доказательство.* По определению запишем

$$\underbrace{D\hat{\xi}_n}_{s^2} = D(\hat{\xi}_n - \mu) = \underbrace{E(\hat{\xi}_n - \mu)^2}_{s_\mu^2} - \underbrace{(E(\hat{\xi}_n - \mu))^2}_{(\bar{x} - \mu)^2}.$$

Домножив обе части на  $n/\sigma_0^2$ , получим

$$\frac{ns^2}{\sigma_0^2} = \frac{ns_\mu^2}{\sigma_0^2} - \frac{n(\bar{x} - \mu)^2}{\sigma_0^2} = \underbrace{\frac{ns_\mu^2}{\sigma_0^2}}_{\sim \chi^2(n)} - \underbrace{\left( \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma_0} \right)^2}_{\sim \chi^2(1)} \Rightarrow \frac{ns^2}{\sigma_0^2} \sim \chi^2(n-1).$$

□

### 3. Проверка гипотезы о значении дисперсии в нормальной модели (критерий $\chi^2$ )

*Замечание.* Для строгого доказательства, нужно использовать независимость  $\bar{\mathbf{x}}^2$  и  $s^2$  (см. 4).

#### Разбиение

$$H_1 : D\xi \neq \sigma_0^2 \quad \mathcal{A}_{\text{крит}} = \mathbb{R}_+ \setminus \left( \text{qnt}_{\chi^2(n-1)}(\alpha/2), \text{qnt}_{\chi^2(n-1)}(1 - \alpha/2) \right)$$

$$H_1 : D\xi > \sigma_0^2 \quad \mathcal{A}_{\text{крит}} = (\text{qnt}_{\chi^2(n-1)}(1 - \alpha), \infty)$$

$$H_1 : D\xi < \sigma_0^2 \quad \mathcal{A}_{\text{крит}} = (0, \text{qnt}_{\chi^2(n-1)} \alpha)$$

**Упражнение.**  $s^2 = 1.44$ ,  $\bar{x} = 55$ ,  $n = 101$ . Проверить гипотезу  $\sigma_0^2 = 1.5$  в нормальной модели.

*Решение.* Воспользуемся статистикой

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = 101 \cdot 0.96 = 96.96.$$

«Идеальные» значения близки к  $E\xi_{\chi^2(100)} = 100$ , так что определим критическую область на концах плотности:

$$p\text{-value}/2 = \text{cdf}_{\chi^2(100)}(96.96) = \text{pchisq}(96.96, 100) \approx 0.43 \implies p\text{-value} \approx 0.86.$$

*Замечание.* Можно посчитать и по таблицам для нормального распределения. Раз

$$\frac{\eta_m - E\eta_m}{\sqrt{D\eta_m}} \xrightarrow[m \rightarrow \infty]{d} N(0, 1),$$

то

$$\frac{96.96 - 100}{\sqrt{200}} \approx -0.215 \implies p\text{-value}/2 = \Phi(-0.215) \approx 0.415.$$

┘

## 4. Критерий $\chi^2$ согласия с видом распределения

По выборке возможно проверить гипотезу о виде распределения случайной величины, реализацией которой является выборка.

*Утверждение.* Для проверки гипотезы согласия с видом произвольного *дискретного* распределения используется асимптотический критерий  $\chi^2$  («chi-squared test for goodness of fit»).

### 4.1. Распределение с известными параметрами

Пусть

$$H_0 : \mathcal{P} = \mathcal{P}_0, \text{ где } \mathcal{P}_0 : \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1 & \dots & p_k \end{pmatrix}.$$

Сгруппируем  $\mathbf{x}$ ; каждому  $x_i^*$  сопоставим *эмпирическую* абсолютную частоту  $\nu_i$ ; тогда  $np_i$  — *ожидаемая* абсолютная частота.

В качестве меры расхождения между эмпирическим и генеральным распределением рассматривается величина

$$\sum_{i=1}^k c_i \left( \frac{\nu_i}{n} - p_i \right)^2, \quad c_i = \frac{n}{p_i},$$

откуда записывается статистика критерия

$$T = \sum_{i=1}^k \frac{(\nu_i - np_i)^2}{np_i}$$

с идеальным значением 0.

*Утверждение.*  $T \xrightarrow{d} \chi^2(k-1)$ .

**Разбиение**  $\mathcal{A}_{\text{крит}} = \left( \text{qnt}_{\chi^2(k-1)}(1-\alpha), \infty \right)$  — гипотеза отвергается, если расстояние между предполагаемым и наблюдаемым распределениями большое.

**Упражнение.**  $n = 100$ ,

$$\begin{pmatrix} \diamond & \heartsuit & \clubsuit & \spadesuit \\ 20 & 30 & 10 & 40 \end{pmatrix}.$$

Проверить гипотезу, что колода полная.

*Решение.*  $H_0 : \mathcal{P}_\xi = \text{U}(1/4)$ . Поскольку речь идет о согласии с дискретным не параметризованным распределением, напрямую воспользуемся критерием  $\chi^2$ . Раз все  $np_i = 100 \cdot 1/4 = 25$ ,

$$\chi^2 = \sum_{i=1}^k \frac{(\nu_i - np_i)^2}{np_i} = 1 + 1 + \frac{15^2}{25} + \frac{15^2}{25} = 2 + 2 \cdot 9 = 20.$$

Так как  $\chi^2 \sim \chi^2(k-1) = \chi^2(3)$  со средним 3, и «идеальное» значение 0, определим критическую область в правом конце плотности. Из этих соображений

$$p\text{-value} = 1 - \text{cdf}_{\chi^2(3)}(20) = 1 - \text{pchisq}(20, 3) \approx 0.00017.$$

┘

## 4.2. Распределение с неизвестными параметрами

В случае сложной гипотезы  $\mathcal{P} \in \{\mathcal{P}(\theta)\}_{\theta \in \Theta}$ ,  $\theta = (\theta_1, \dots, \theta_r)^\top$ , следует найти оценку  $\hat{\theta}_{\text{MLE}}$  (или  $\hat{\theta} : \hat{\theta} \rightarrow \hat{\theta}_{\text{MLE}}$ ) по методу максимального правдоподобия. При подстановке оценок вместо истинных параметров критерий становится консервативным. Чтобы этого избежать, необходимо сделать поправку на количество параметров — отнять  $r$ . Что приятно, одна и та же поправка работает для всех распределений; в этом случае,

$$T \xrightarrow{d} \chi^2(k - r - 1).$$

**Упражнение 4.1.** 60 человек купило подарок сразу, 10 со второго раза, 20 с третьего, 10 с четвертого:

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 60 & 10 & 20 & 10 \end{pmatrix}.$$

Проверить гипотезу о том, что это выборка из геометрического распределения.

*Решение.*  $H_0 : \mathcal{P}_\xi = \text{Geom}(p)$ . Воспользуемся критерием  $\chi^2$  для параметризованного распределения  $\text{Geom}(\hat{p}_{\text{MLE}})$ .

Найдем

$$\hat{p}_{\text{MLE}} = \underset{p}{\operatorname{argmax}} \log L(\mathbf{x}; p) \iff \frac{d}{dp} \log L(\mathbf{x}; \hat{p}_{\text{MLE}}) = 0.$$

Так как  $\text{pdf}_{\text{Geom}(p)}(k) = (1-p)^k p$ ,

$$\begin{aligned} \log L(\mathbf{x}; p) &= \log \prod_{k=1}^n (1-p)^k p = \log(1-p)^{n\bar{x}} p^n = n\bar{x} \log(1-p) + n \log p \\ &= n(\bar{x} \log(1-p) + \log p) \end{aligned}$$

откуда

$$\frac{d}{dp} \log L(\mathbf{x}; p) = n \left( -\frac{\bar{x}}{1-p} + \frac{1}{p} \right) = 0 \iff 1-p-p\bar{x} = 0 \iff p = \frac{1}{1+\bar{x}}.$$

Учитывая

$$\bar{x} = 0.1 + 2 \cdot 0.2 + 3 \cdot 0.1 = 0.8,$$

найдем

$$\hat{p}_{\text{MLE}} = \frac{1}{1+0.8} \approx 0.55.$$

Посчитаем статистику  $\chi^2$ , найдя соответствующие  $p_i$ :

$$p_0 = P_{\text{Geom}(0.55)}(0) = 0.55, \quad p_1 \approx 0.26, \quad p_2 \approx 0.11, \quad p_3 \approx 0.09.$$

Тогда

$$\chi^2 = \sum_{i=1}^k \frac{(\nu_i - np_i)^2}{np_i} = \frac{25}{55} + \frac{16^2}{26} + \frac{81}{11} + \frac{1}{9} \approx 17.77.$$

Наконец, поскольку  $\chi^2 \xrightarrow[n \rightarrow \infty]{} \chi^2(k-r-1)$ ,

$$p\text{-value} = 1 - \text{cdf}_{\chi^2(2)}(17.77) \approx 0.00014.$$

┘

**Определение.** Критерий применим, если  $\alpha \rightarrow \alpha_I$ .

*Замечание.* Поскольку критерий асимптотический, с достаточной степенью точностью он применим в случае, если

#### 4. Критерий $\chi^2$ согласия с видом распределения

1.  $n \geq 50$ ;
2.  $np_i \geq 5$ .

*Замечание.* Если условие  $np_i \geq 5$  не выполняется, следует объединить состояния, например, с краев или слева направо; если в хвосте оказалось  $< 5$ , то следует присоединить к последнему.

**Пример** (С монеткой). Пусть  $n = 4040$ ,  $\#H = 2048$ ,  $\#T = 1092$ . Проверим  $H_0 : \mathcal{P} = \text{Ber}(0.5)$  с  $\alpha = 0.1$ . Условия критерия выполняются, поэтому посчитаем

$$T = \frac{(2048 - 2020)^2}{2020} + \frac{(1092 - 2020)^2}{2020} = \frac{28^2 + 28^2}{2020} \approx 0.78 \sim \chi^2(1),$$

откуда

$$p\text{-value} = 1 - \text{cdf}_{\chi^2(1)}(0.78) \approx 0.38.$$

$0.38 > 0.1$ , значит  $H_0$  не отвергается.

*Замечание.* Прохождение критерия не достаточно. Так, альтернирующая (и явно не случайная) последовательность  $\mathbf{x} = (0, 1, 0, 1, \dots)$  имеет  $T = 0$ .

### 4.3. Согласие с нормальным распределением по $\chi^2$

Для проверки гипотезы  $H_0 : \mathcal{P}_\xi = N(a, \sigma^2)$  также можно воспользоваться статистикой критерия  $\chi^2$  для сложной гипотезы. В этом случае, нужно дискретизировать нормальное распределение, так, что

$$\mathcal{P}_0 = \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1(\hat{\theta}) & \dots & p_k(\hat{\theta}) \end{pmatrix}, \quad \hat{\theta} = \hat{\theta}_{\text{MLE}}.$$

Тем не менее, нужно иметь в виду две теоретических неточности:

1. Построение  $\mathcal{P}_0$  происходит случайно, в результате объединения элементов выборки после того, как она получена.
2. Оценка параметров  $\hat{\theta}_{\text{MLE}}$  должна быть посчитана для  $\mathcal{P}_0$ , а не для исходного (нормального) распределения — не  $\bar{\mathbf{x}}, s^2$ . Однако на практике на этот момент не обращают внимания.

Существует два возможных способа дискретизации:

1. Гистограмма: одинаковые интервалы, но разные вероятности.
2. Неравные интервалы с равными вероятностями.

```
N <- length(xs)
xs <- sort(xs)
probs <- pnorm(xs, mean=mean(xs), sd=sd(xs))
i <- 1; j <- i+1
while (N * (probs[j] - probs[i]) < 5) {
  j <- j+1
}
mean(xs[i:j]) # our  $x_1^*$ ;  $n_1 = j - i + 1$ 
```

Этот способ разбиения предпочтителен, потому что:

- Можно разбить максимально часто — так, чтобы  $np_i = 5 \forall i$ , следовательно и мощность будет максимальна.
- Он оказывается точнее первого на практике.
- Получается единственное  $p$ -value.

*Замечание.* Следует иметь в виду, что этот способ не годится для непрерывных, но плохо дискретизированных данных.

## 5. Критерий Колмогорова-Смирнова согласия с видом распределения

### 5.1. Произвольное абсолютно непрерывное распределение

$H_0 : \xi \sim \mathcal{P} = \mathcal{P}_0$ .

*Утверждение.* Для проверки гипотезы согласия с видом произвольного абсолютно непрерывного распределения с известными параметрами используется асимптотический критерий Колмогорова-Смирнова со следующей статистикой:

$$D_n = \sup_{x \in \mathbf{x}} \left| \widehat{\text{cdf}}_n(x) - \text{cdf}_0(x) \right|,$$

где  $\text{cdf}_0$  — функция распределения  $\mathcal{P}_0$  нулевой гипотезы.

Альтернатива только одна:  $H_1 : \xi \not\sim \mathcal{P}_0$ ;  $\mathcal{A}_{\text{крит}} = (\text{qnt}_{\text{K-S}}(1 - \alpha), \infty)$ .

*Замечание.* Критерий является не асимптотическим, но *точным*. Значит им пользоваться и при маленьких объемах выборки (мощность, при этом, останется низкой все-равно).

*Замечание.*  $\sup_x \sqrt{n} \left| \widehat{\text{cdf}}_n(x) - \text{cdf}_0(x) \right| \xrightarrow{d} \mathcal{P}_{\text{K.S.}}$ , где  $\mathcal{P}_{\text{K.S.}}$  — распределение Колмогорова. Значит, при больших объемах выборки для такой статистики критерия можно пользоваться таблицами распределения Колмогорова.

**Упражнение.** Проверить гипотезу, что  $\mathbf{x} = (0.1, 0.2, 0.4, 0.3, 0.1)$  есть выборка из  $U[0, 1]$ .

*Решение.*  $D_n = 0.6$ ,  $p\text{-value} \approx 0.05$  (по таблицам или компьютером). Таким образом, при  $\alpha > 0.05$  гипотеза отвергается, при  $\alpha < 0.05$  — нет.

```
> ks.test(c(0.1,0.2,0.4,0.3,0.1), 'punif')
      One-sample Kolmogorov-Smirnov test
data:  c(0.1, 0.2, 0.4, 0.3, 0.1)
D = 0.6, p-value = 0.05465
```

└

*Замечание.* Критерий Колмогорова-Смирнова консервативный — значение  $p\text{-value}$  завышено. Поэтому если гипотеза отвергается, то наверняка.

### 5.2. Нормальное распределение

Пусть  $H_0 : \mathcal{P}_\xi \in \{N(\mu, \sigma^2)\}$ . Как известно, критерий Колмогорова-Смирнова используется для непрерывных непараметрических распределений. Им можно воспользоваться и для данной  $H_0$ , если вместо  $\mu, \sigma^2$  подставить соответствующие оценки — в таком случае критерий будет консервативным. По аналогии с  $\chi^2$  хотелось бы сделать поправку на количество параметров — такая поправка осуществляется путем моделирования распределения тестовой статистики. Для  $N(\mu, \sigma^2)$  и  $\text{Exp}(\lambda)$  получаем распределение  $D_n$ , не зависящее от параметров (так что поправку можно делать вне зависимости от параметров; к примеру,  $N(\mu, \sigma^2)$  можно привести к  $N(0, 1)$  непрерывным преобразованием):

**Критерий Бартлетта** есть критерий Колмогорова-Смирнова для  $H_0 : \mathcal{P}_\xi = \text{Exp}(\lambda)$ .



**Критерий Лиллиефорса**<sup>1</sup> для проверки  $H_0 : \mathcal{P}_\xi = N(\mu, \sigma^2)$  считается статистикой  $D_n$  с  $\text{cdf}_0(x) = \text{cdf}_{N(\bar{x}, s^2)}(x)$ , сходящейся к распределению Лиллиефорса (Колмогорова-Смирнова с учетом подстановки оценок).

**Критерий Шапиро-Уилка**  $T \approx \rho^2$ , т.е. ведет себя примерно как квадрат коэффициента корреляции на normal probability plot.

*Замечание.* Распределения Лиллиефорса и Колмогорова-Смирнова были получены путем моделирования.

**Пример.** В R:

```
> require('nortest')
> xx <- rt(1000, df=20)
> lillie.test(xx)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  xx
D = 0.023412, p-value = 0.2032
> shapiro.test(xx)
      Shapiro-Wilk normality test
data:  xx
W = 0.99669, p-value = 0.03409
```

## 6. Критерий типа $\omega^2$

**Определение.** Статистика

$$Q = n \int_{\mathbb{R}} (\text{cdf}_n(x) - \text{cdf}_0(x))^2 w(x) d \text{cdf}_0(x),$$

где  $w(x)$  — весовая функция.

*Замечание.* Статистика может быть проинтерпретирована как площадь разницы между соответствующими функциями распределения.

**Cramer von Mises**  $Q$  с  $w \equiv 1$ .

**Anderson-Darling**  $Q$  с

$$w(x) = \frac{1}{\text{cdf}_0(x)(1 - \text{cdf}_0(x))}.$$

*Замечание.* Весовая функция критерия Anderson-Darling присваивает большой вес значениям на хвостах распределения, поэтому сам критерий является мощным против разницы на хвостах, но и менее мощным при сдвиге.

*Замечание.* Все эти критерии точны.

*Замечание.* Распределение статистики в каждом случае не зависит от  $\text{cdf}_0$  и все эти критерии состоятельные против любой альтернативы, поэтому не очень мощные.

*Замечание.* Из всех тестов для тестирования согласия с нормальным распределением наибольшей мощностью при любых объемах выборки почти всегда обладает Shapiro-Wilk, см. [http://www.de.ufpb.br/~ulisses/disciplinas/normality\\_tests\\_comparison.pdf](http://www.de.ufpb.br/~ulisses/disciplinas/normality_tests_comparison.pdf)

## 7. Визуальное определение согласия с распределением

### 7.1. P-P plot

**Определение.** *P-P plot* есть график

$$\left\{ \left( \text{cdf}_0(x_i) + \frac{1}{2n}, \widehat{\text{cdf}}_n(x_i) \right) \right\}_{i=1}^n.$$

**Пример.** В R:

```
pp.plot <- function(xs, cdf.0=pnorm, n.knots=1000) {  
  knots <- seq(min(xs), max(xs), length.out=n.knots)  
  plot(cdf.0(knots), ecdf(xs)(knots))  
  abline(0, 1)  
}
```

### 7.2. Q-Q plot

**Определение.** *Q-Q plot* есть график

$$\left\{ \left( x_i, \text{cdf}_0^{-1} \left( \widehat{\text{cdf}}_n(x_i) + \frac{1}{2n} \right) \right) \right\}_{i=1}^n.$$

**Определение.** Частный случай Q-Q plot для  $\text{cdf}_0^{-1} = \text{cdf}_{N(0,1)}^{-1}$  называется *normal probability plot*.

**Пример.** В R:

```
qq.plot <- function(xs, qf.0=qnorm, n.ppoints=1000) {  
  qs <- ppoints(n.ppoints)  
  plot(qf.0(qs), unname(quantile(xs, probs=qs)))  
  abline(mean(xs), sd(xs))  
}
```

*Замечание.* Если  $\hat{\mathcal{P}}_n \rightarrow \mathcal{P}_\xi$ , то оба графика будут стремиться к  $y = x$ . Референсной прямой normal probability plot будет  $y = \widehat{D}\xi \cdot x + \widehat{E}\xi$ .

*Замечание.* Больше о различии Q-Q и P-P plots, см. <http://v8doc.sas.com/sashtml/qc/chap8/sect9.htm>

*Замечание.* Различные интерпретации параметров распределения по Q-Q plot можно посмотреть в интерактивном приложении: <https://xiongge.shinyapps.io/QQplots/>

## 8. Гипотеза о равенстве распределений

$$H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}.$$

Возможно рассматривать два случая:

**Независимые выборки** Две группы индивидов, на которых измеряется один и тот же признак. Формально: пусть  $\zeta \in \{1, 2\}$  — номер группы,  $\xi$  — признак. Тогда  $\xi_1 \sim \mathcal{P}_{\xi|\zeta=1}$ ,  $\xi_2 \sim \mathcal{P}_{\xi|\zeta=2}$  и  $\xi_1 \perp \xi_2$ . В этом случае выборка имеет вид

$$((x_1, x_2, \dots, x_{n_1}), (y_1, y_2, \dots, y_{n_2}))$$

или

| $\xi$ | $x_1$ | $\dots$ | $x_{n_1}$ | $y_1$ | $\dots$ | $y_{n_2}$ |
|-------|-------|---------|-----------|-------|---------|-----------|
|-------|-------|---------|-----------|-------|---------|-----------|

то есть одному признаку сопоставлено  $n_1 + n_2$  индивидов.

**Зависимые выборки** Одна группа индивидов, на каждом из которых измеряются две характеристики (либо же «до» и «после»). В этом случае выборка имеет вид

$$((x_1, y_1), \dots, (x_n, y_n))$$

или

| $\xi$  | $x_1$ | $\dots$ | $x_n$ |
|--------|-------|---------|-------|
| $\eta$ | $y_1$ | $\dots$ | $y_n$ |

то есть по строчкам стоят признаки, по столбцам — индивиды.

*Замечание.* Для одной и той же гипотезы могут существовать разные критерии; их возможно сравнить по мощности, но только если они состоятельны против одной и той же альтернативы.

*Замечание.* Непараметрические критерии хороши тем, что основаны на рангах, значит устойчивы к аутлаерам; плохи тем, что не используют всю информацию о значении — только порядок, из-за чего обладают меньшей мощностью.

### 8.1. Двухвыборочный тест Колмогорова–Смирнова

Рассматривается  $H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$  против  $H_1 : \mathcal{P}_{\xi_1} \neq \mathcal{P}_{\xi_2}$  и оба распределения абсолютно непрерывны. В качестве статистики используется

$$D = \sup_x \left| \widehat{\text{cdf}}_{\xi_1}(x) - \widehat{\text{cdf}}_{\xi_2}(x) \right|.$$

## 9. Равенство математических ожиданий для независимых выборок

### 9.1. Двухвыборочный $t$ -критерий

$$H_0 : E\xi_1 = E\xi_2.$$

**Определение.** И для зависимых, и для независимых выборок используется *двухвыборочный  $t$ -критерий*

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{D(\bar{x} - \bar{y})}} \xrightarrow{\sim} N(0, 1).$$

Пусть выборка *независима*<sup>1</sup>,  $(x_1, \dots, x_{n_1}), (y_1, \dots, y_{n_2})$ ,  $n = n_1 + n_2$ . Значит  $D(\bar{x} - \bar{y}) = D\bar{x} + D\bar{y}$ .

#### 9.1.1. Двухвыборочный $t$ -критерий для независимых выборок с $\sigma_1^2 \neq \sigma_2^2$ (Welch $t$ -test)

**Предложение.** Если дисперсия известна,  $D(\bar{x} - \bar{y}) = D\bar{x} + D\bar{y} = \sigma_1^2/n_1 + \sigma_2^2/n_2$  и

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0, 1).$$

Если данные нормальные, то

$$t \sim N(0, 1).$$

**Предложение.** Если дисперсия неизвестна,  $D(\widehat{\bar{x} - \bar{y}}) = s_1^2/n_1 + s_2^2/n_2$ , откуда

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0, 1).$$

*Замечание.* Точное распределение неизвестно, примерно равно  $t$  с дробным числом степеней свободы (что вычисляется интерполяцией по соседним степеням). Всегда ожидается, что если данные нормальны, то распределение известно. Это противоречие носит название *проблемы Беренса-Фишера*<sup>2</sup>.

#### Разбиение

$$H_1 : E\xi_1 \neq E\xi_2 \quad \mathcal{A}_{\text{крит}} = \mathbb{R} \setminus (z_{\alpha/2}, z_{1-\alpha/2})$$

$$H_1 : E\xi_1 > E\xi_2 \quad \mathcal{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$$

$$H_1 : E\xi_1 < E\xi_2 \quad \mathcal{A}_{\text{крит}} = (-\infty, z_{\alpha})$$

<sup>1</sup>Случай зависимой выборки рассматривается в другом параграфе.

<sup>2</sup>Behrens-Fisher problem.

### 9.1.2. Двухвыборочный $t$ -критерий для независимых выборок с $\sigma_1^2 = \sigma_2^2$ (pooled $t$ -test)

**Предложение.** Если дисперсия известна,

$$D(\bar{x} - \bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right),$$

откуда

$$t = \frac{\bar{x} - \bar{y}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0, 1).$$

Если данные нормальные, то

$$t \sim N(0, 1).$$

**Предложение.** Если дисперсия неизвестна,

$$t = \frac{\bar{x} - \bar{y}}{\tilde{s}_{1,2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0, 1).$$

Если данные нормальные, то

$$t \sim t(n_1 + n_2 - 2).$$

*Доказательство.* Оценку дисперсии можно найти по объединенной и центрированной выборке (т.е. если  $H_0$  верна, то  $E\xi_1 = E\xi_2$  и можно думать как про одну выборку):

$$\begin{aligned} s_{1,2}^2 &= \frac{\overbrace{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}^{\sim \chi^2(n_1-1)} + \overbrace{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}^{\sim \chi^2(n_2-1)}}{n_1 + n_2} = \frac{n_1 \cdot s_1^2}{n_1 + n_2} + \frac{n_2 \cdot s_2^2}{n_1 + n_2} \\ \tilde{s}_{1,2}^2 &= \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\tilde{s}_1^2}{n_1 + n_2 - 2} + \frac{(n_2 - 1)\tilde{s}_2^2}{n_1 + n_2 - 2}, \end{aligned}$$

где в последнем случае оценка несмещенная и  $E\tilde{s}_{1,2}^2 = \sigma^2$ . □

*Замечание.* Этот вариант более точен, чем в случае  $\sigma_1 \neq \sigma_2$ .

#### Разбиение

$$H_1 : E\xi_1 \neq E\xi_2 \quad \mathcal{A}_{\text{крит}} = \mathbb{R} \setminus \left( \text{qnt}_{t(n_1+n_2-2)}(\alpha/2), \text{qnt}_{t(n_1+n_2-2)}(1 - \alpha/2) \right)$$

$$H_1 : E\xi_1 > E\xi_2 \quad \mathcal{A}_{\text{крит}} = (\text{qnt}_{t(n_1+n_2-2)}(1 - \alpha), \infty)$$

$$H_1 : E\xi_1 < E\xi_2 \quad \mathcal{A}_{\text{крит}} = (-\infty, \text{qnt}_{t(n_1+n_2-2)} \alpha)$$

### 9.1.3. Испытания Бернулли

Пусть  $\xi_i \sim \text{Ber}(p_i)$ ,  $i \in \{1, 2\}$ . Рассмотрим  $H_0 : p_1 = p_2 = p$  против  $H_1 : p_1 \neq p_2$ . Поскольку  $E\xi_i = p_i$ , применим двух-выборочный  $t$ -критерий. Объединим выборки и запишем:

$$D(\bar{x} - \bar{y}) = \frac{\hat{p}(1 - \hat{p})}{n_1 + n_2} \implies t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n_1 + n_2}}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0, 1), \quad \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

**Разбиение** Аналогично с  $\text{qnt}_{N(0,1)}$ .

**Определение.** Выборка обладает *сбалансированным дизайном*, если  $n_1 = n_2$ .

Если дизайн сбалансирован, то

$$s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2},$$

т.е. даже если дисперсии разные, результат одинаковый. Это остается справедливым даже при  $n_1 \approx n_2$ .

## 9.2. Непараметрический $t$ -критерий

Можно использовать обычный  $t$ -критерий, но примененный к рангам.

Пусть, как и прежде, дана выборка  $(\mathbf{x}, \mathbf{y})$ . Следующие два критерия — Wilcoxon и Mann-Whitney — проверяют гипотезу  $H_0 : P(\xi_1 > \xi_2) = P(\xi_1 < \xi_2)$  или, альтернативно,  $H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$  против  $H_1 : \mathcal{P}_{\xi_1} \neq \mathcal{P}_{\xi_2}$  (что выборки получены из одной «генеральной совокупности») в случае абсолютно непрерывных распределений.

## 9.3. Критерии суммы рангов Wilcoxon

Следует сопоставить каждой выборке соответствующие её элементам ранги в *объединенной выборке*:

$$\begin{aligned} (x_1, \dots, x_{n_1}) &\mapsto (R_1, \dots, R_{n_1}) \\ (y_1, \dots, y_{n_2}) &\mapsto (T_1, \dots, T_{n_2}). \end{aligned}$$

Ясно, что если в целом элементы одной выборки окажутся больше другой, то нельзя будет говорить об их однородности. Определим

$$W_1 := \sum_{i=1}^{n_1} R_i, \quad W_2 := \sum_{i=1}^{n_2} T_i.$$

В качестве статистики можно было бы использовать либо  $W_1$ , либо  $W_2$ , однако, ни той, ни другой статистике невозможно априорно отдать предпочтение. Поэтому используется статистика

$$W := \max(W_1, W_2),$$

не имеющая аналитического выражения (но для которого посчитаны соответствующие таблицы).

Иногда в качестве статистики берут количество инверсий в объединенной выборке.

## 9.4. Критерий Mann-Whitney ( $U$ test)

Используется статистика

$$U := \max \left( n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1, n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - W_2 \right).$$

При верной  $H_0$ ,  $P(\xi_1 < \xi_2) = 1/2$ . В этом случае,

$$EU = \frac{n_1 n_2}{2}, \quad DU = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Асимптотически,

$$\frac{U - EU}{\sqrt{DU}} \xrightarrow{n_1, n_2 \rightarrow \infty} N(0, 1),$$

но для малых объемов выборки можно посчитать и точные распределения.

*Замечание.* Критерий состоятельный против альтернативы

$$H_1 : P(\xi_1 > \xi_2) \neq P(\xi_1 < \xi_2).$$

Если формы распределений одинаковы, то эта альтернатива обозначает сдвиг. Для симметричных распределений это условие обозначает равенство медиан (а для нормального — математических ожиданий). Поэтому критерий устойчив к аутлаерам, хоть и за счет небольшой ( $\approx 5\%$ ) потери мощности.

*Замечание.* Критерии Манна-Уитни и Вилкоксона *эквивалентны* — в том смысле, что выделяют один и тот же  $p$ -value. Тем не менее, проверяют они разные гипотезы ( $E\xi$  не то же, что  $\text{med } \xi$ ).

## 9.5. Критерий серий (runs)

Следует объединить выборку и в качестве статистики выбрать количество серий, т.е. подряд идущих элементов из одной выборки. Эта статистика имеет специально подобранное распределение.

*Замечание.* Все эти критерии подразумевают отсутствие повторяющихся наблюдений для избежания появления дробных рангов.



## 10. Равенство математических ожиданий для парных (зависимых) выборок

Выборка представлена набором пар  $\{(x_i, y_i)\}_{i=1}^n$ .

### 10.1. $t$ -критерий

Пусть  $\xi_1, \xi_2$  заданы на одном  $(\Omega, \mathcal{F}, P)$ . Тогда гипотезу  $H_0 : E\xi_1 = E\xi_2$  можно свести к  $H_0 : E(\xi_1 - \xi_2) = E\eta = 0$  использовать не-парный  $t$ -тест.

*Замечание* (Мощность и зависимость). Сравним статистику для сбалансированного дизайна:

- Независимая выборка

$$t_{\text{indep}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\sqrt{n}(\bar{x} - \bar{y})}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

- Зависимая выборка:

$$\begin{aligned} D(\bar{x} - \bar{y}) &= D\bar{x} + D\bar{y} - 2 \operatorname{cov}(\bar{x}, \bar{y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\rho\sqrt{D\bar{x}}\sqrt{D\bar{y}} \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\rho\frac{\sigma_1}{\sqrt{n}}\frac{\sigma_2}{\sqrt{n}} = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2), \end{aligned}$$

откуда

$$t_{\text{dep}} = \frac{\sqrt{n}(\bar{x} - \bar{y})}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}}.$$

При  $\rho > 0$ ,  $t_{\text{dep}} > t_{\text{indep}}$ . Значит, статистика чаще попадает в критическую область и критерий лучше находит различия (и мощность, следовательно, выше). Значит, тот же эксперимент на зависимых выборках мощнее.

**Пример.** Проверяют гипотезу, что белый свет лам влияет на решение задач.

- При тестировании на разных индивидах, должна быть уверенность, что они одинаковы по критичным параметрам (IQ, например).
- При тестировании на одинаковых индивидах следует составлять разные, но одинаковы по сложности задачи (второй раз одну и ту же задачу решать не займет много времени!). Мощность этого эксперимента будет выше.

### 10.2. Непараметрический тест знаков (Sign test)

$H_0 : P(\xi_1 < \xi_2) = P(\xi_1 > \xi_2)$ . Используется статистика

$$W = \sum_{i=1}^n \psi_i, \quad \psi_i = \begin{cases} 1 & x_i > y_i \\ 0 & x_i < y_i. \end{cases}$$

Если при подсчете статистики  $x_i = y_i$ , эта пара игнорируется вместе с соответствующим уменьшением объема выборки.

## 10. Равенство математических ожиданий для парных (зависимых) выборок

Пусть после удаления всех пар, таких, что  $x_i = y_i$ , объем выборки стал равен  $m$ . Тогда

$$W \sim \text{Bin}(m, 0.5)$$

и для построения разбиения можно пользоваться  $\text{qnt}_{\text{Bin}(m, 0.5)}$ .

*Замечание.* Критерий применим к порядковым признакам.

*Замечание.* Критерий очень устойчив к аутлаерам (но и очень низкомощен поэтому).

### 10.3. Непараметрический критерий (Paired Wilcoxon; Wilcoxon signed-rank test)

Увеличить мощность предыдущего критерия можно, учтя больше информации:

$$W := \sum_{i=1}^n R_i \psi_i, \quad R_i := \text{rk} |x_i - y_i|.$$

Для симметрии можно рассмотреть статистику

$$W = \sum_{i=1}^n R_i \text{sign}(x_i - y_i)$$

с идеальным значением 0. При верной  $H_0$ , распределение  $W$  не имеет простого аналитического выражения (но может быть посчитана по таблицам), при этом  $EW = 0$ ,  $DW = n(n+1)(2n+1)/6$ . Кроме того,  $W \xrightarrow{d} N(0, DW)$ , так что уже при  $n \geq 10$  можно полагать, что  $z = W/\sqrt{DW} \xrightarrow{d} N(0, 1)$  и строить разбиение соответственно.

*Замечание.* Критерий уже не применим к порядковым признакам.

## 11. Равенство дисперсии для двух распределений

$H_0 : D\xi_1 = D\xi_2, \xi_1 \perp\!\!\!\perp \xi_2, \xi_i \sim N(\mu, \sigma_i)^2$ .

### 11.1. Критерий Фишера

$H_0 : \sigma_1^2 = \sigma_2^2$ . Естественно использовать отношение  $s_1^2/s_2^2$  с идеальным значением 1. Поделив на число степеней свободы, получим статистику

$$F := \frac{\tilde{s}_1^2}{\tilde{s}_2^2} \sim F(|\mathbf{x}| - 1, |\mathbf{y}| - 1).$$

*Замечание.* При отклонении от нормальности не становится асимптотическим.

### 11.2. Критерий Левена (Levene's test)

Так как  $D\xi_i = E(\xi_i - E\xi_i)^2$ , то критерий о равенстве дисперсий можно было бы свести к критерию о равенстве математических ожиданий; в этом случае применили бы  $t$ -критерий (подразумевающий разные дисперсии) к выборкам  $\{(x_i - \bar{x})^2\}$  и  $\{(y_i - \bar{y})^2\}$ . Однако при возведении в квадрат распределение стало бы несимметричным и потребовался бы больший объем выборки. Кроме того, значительно бы усилились аутлаеры.

Вместо этого используют гипотезу  $H_0 : E|\xi_1 - E\xi_1| = E|\xi_2 - E\xi_2|$  вместе с  $t$ -критерием, подразумевающим равенство дисперсий (для нормальных данных; иначе с разными).

### 11.3. Критерий Brown–Forsythe

Критерий Brown–Forsythe — это  $t$ -критерий для гипотезы  $H_0 : E|\xi_1 - \text{med } \xi_1| = E|\xi_2 - \text{med } \xi_2|$ .

*Замечание.* Устойчив к аутлаерам из-за использования  $\text{med } \xi_i$ .

## Часть IV.

# Доверительное оценивание

# 1. Мотивация и определение

Для построенных оценок может понадобиться оценка точности. Так, даже состоятельная оценка может не быть в полном смысле «точной»: пусть  $\theta_n^* \xrightarrow{P} \theta_0$ ; тогда

$$\hat{\theta}'_n = \begin{cases} c & n < N \gg 1 \\ \hat{\theta}_n & \text{иначе} \end{cases}$$

все-равно будет, конечно, состоятельной.

$D\hat{\theta}_n$  может быть не всегда просто вычислить и использовать.

**Определение.**  $[c_1, c_2]$  — *доверительный интервал* для параметра  $\theta_0$  с уровнем доверия  $\gamma \in [0, 1]$ , если  $\forall \theta_0$

$$P(\theta_0 \in [c_1, c_2]) = \gamma, \quad \text{где } c_1 = c_1(\mathbf{x}), c_2 = c_2(\mathbf{x}) \text{ — статистики.}$$

*Замечание.* Если выборка из дискретного распределения, то  $c_1, c_2$  — тоже. Поэтому наперед заданную точность получить может не получиться; в таких случаях знак « $=$ » заменяют « $\geq$ ». Аналогично с заменой на « $\xrightarrow[n \rightarrow \infty]{}$ ».

## 2. Доверительные интервалы для математического ожидания и дисперсии в нормальной модели

**Предположение.** Пусть  $\xi \sim N(\mu, \sigma^2)$ .

### 2.1. Доверительный интервал для $\mu$

- Пусть  $\sigma^2$  известно. Свяжем  $\mu_0$  с выборкой:

$$\gamma = P(c_1 < T < c_2) = P\left(c_1 < \sqrt{n} \frac{(\bar{x} - \mu_0)}{\sigma} < c_2\right) = P\left(\mu_0 \in \left(\bar{x} - \frac{\sigma c_2}{\sqrt{n}}, \bar{x} - \frac{\sigma c_1}{\sqrt{n}}\right)\right).$$

Решений уравнения  $P(c_1 < \sqrt{n}(\bar{x} - \mu_0)/\sigma < c_2) = \Phi(c_2) - \Phi(c_1) = \gamma$  бесконечно много. Чем  $[c_1, c_2]$  короче, тем лучше. Поскольку  $\Phi$  симметрична и унимодальна,

$$\begin{aligned} c_1 &= -c_\gamma \\ c_2 &= c_\gamma, \end{aligned} \quad \text{где } c_\gamma = \text{cdf}_{N(0,1)}^{-1}\left(\gamma + \frac{1-\gamma}{2}\right) = x_{\frac{1+\gamma}{2}}.$$

Наконец,

$$P\left(\mu_0 \in \left(\bar{x} \pm \frac{\sigma}{\sqrt{n}} x_{\frac{1+\gamma}{2}}\right)\right) = \gamma.$$

- Пусть  $\sigma^2$  неизвестно. По аналогии,

$$\gamma = P\left(c_1 < \frac{\sqrt{n-1}(\bar{x} - \mu_0)}{s} < c_2\right) = P\left(\mu_0 \in \left(\bar{x} \pm \frac{c_\gamma s}{\sqrt{n-1}}\right)\right), \quad c_\gamma = \text{cdf}_{t(n-1)}^{-1}\left(\frac{1+\gamma}{2}\right)$$

и

$$P\left(\mu_0 \in \left(\bar{x} \pm \frac{\tilde{s}}{\sqrt{n}} x_{\frac{1+\gamma}{2}}\right)\right) = \gamma.$$

**Упражнение.** Пусть  $s^2 = 1.21, \bar{x} = 1.9, n = 36$ . Построить 95% доверительный интервал для  $E\xi$ .

*Решение.*

$$c_\gamma = \text{qt}(0.975, 35) \approx 2.03 \implies \left(1.9 \pm \frac{2.03 \cdot \sqrt{1.21}}{\sqrt{35}}\right) = (1.52; 2.28).$$

┘

### 2.2. Доверительный интервал для $\sigma^2$

- Пусть  $\mu$  известно. Поскольку плотность  $\chi^2$  становится все более симметричной с ростом  $n$ , примем

$$c_1 = \text{cdf}_{\chi^2(n)}^{-1}\left(\frac{1-\gamma}{2}\right), \quad c_2 = \text{cdf}_{\chi^2(n)}^{-1}\left(\frac{1+\gamma}{2}\right).$$

Тогда

$$P\left(c_1 < \frac{ns_\mu^2}{\sigma_0^2} < c_2\right) = \gamma \iff P\left(\sigma_0^2 \in \left(\frac{ns_\mu^2}{x_{(1+\gamma)/2}}, \frac{ns_\mu^2}{x_{(1-\gamma)/2}}\right)\right) = \gamma.$$

## 2. Доверительные интервалы для математического ожидания и дисперсии в нормальной модели

- Пусть  $\mu$  неизвестно. Тогда аналогично

$$P\left(\sigma_0^2 \in \left(\frac{ns^2}{x_{(1+\gamma)/2}}, \frac{ns^2}{x_{(1-\gamma)/2}}\right)\right) = \gamma,$$

где  $x_{(1\pm\gamma)/2} = \text{cdf}_{\chi^2(n-1)}^{-1}((1 \pm \gamma)/2)$ .

**Определение.** Случайная величина  $g(x_1, \dots, x_n, \theta)$  называется *центральной статистикой параметра  $\theta$* , если

1. Её распределение («центральное распределение») не зависит от распределения  $\theta$ .
2.  $G_n$  (функция распределения центрального распределения) непрерывна.
3.  $\forall z_1, z_2$  и  $\mathcal{P}_\theta$ -почти всюду

$$z_1 < g(x_1, \dots, x_n, \theta) < z_2$$

монотонно разрешимо относительно  $\theta$ , т.е.

$$\exists f_1, f_n : f_1(x_1, \dots, x_n, \theta, z_1, z_2) < \theta < f_2(x_1, \dots, x_n, \theta, z_1, z_2).$$

Рассмотрим всегда разрешимое

$$\begin{aligned} \gamma &= G_n(z_2) - G_n(z_1) = P(z_1 < g(x_1, \dots, x_n, \theta) < z_2) \\ &= P(\underbrace{f_1(z_1, z_2, x_1, \dots, x_n)}_{c_1} < \theta < \underbrace{f_2(z_1, z_2, x_1, \dots, x_n)}_{c_2}). \end{aligned}$$

### 3. Асимптотический доверительный интервал для математического ожидания в модели с конечной дисперсией

Если модель неизвестна, но известно, что  $D\xi < \infty$ , можно построить доверительный интервал для  $E\xi = \mu$ . Пусть  $\{x_i\}$  i.i.d., тогда

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \xrightarrow[n \rightarrow \infty]{} N(0, 1).$$

Если положить  $\sigma := s$ , то сходимость не испортится, потому что  $s^2$  — состоятельная оценка  $\sigma^2$ . Тогда

$$P\left(E\xi \in \left(\bar{x} \pm \frac{sc_\gamma}{\sqrt{n}}\right)\right) \xrightarrow[n \rightarrow \infty]{} \gamma, \quad c_\gamma = \text{cdf}_{t(n-1)}^{-1}\left(\frac{1+\gamma}{2}\right).$$

Альтернативно замену  $\sigma$  на  $s$  можно обосновать по теореме Slutsky.

*Утверждение (Слуцкий).* Если  $\xi_n \xrightarrow{d} \xi$ ,  $\eta_n \xrightarrow{P} c$ , то  $\xi_n + \eta_n \xrightarrow{d} \xi + c$  и  $\xi_n \eta_n \xrightarrow{d} c\xi$ .

Используя тот факт, что  $s \xrightarrow{P} \sigma$ , запишем

$$P\left(c_1 < \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \frac{\sigma}{s} < c_2\right) \xrightarrow[n \rightarrow \infty]{} \Phi(c_2) - \Phi(c_1).$$

**Пример.** Доверительный интервал для параметра  $\text{Exp}(\lambda)$ . FIXME



## 4. Асимптотический доверительный интервал для параметра на основе MLE

Если умеем находить  $\hat{\theta}_{\text{MLE}}$ , то по асимптотической нормальности,

$$\frac{\hat{\theta}_{\text{MLE}} - \mathbb{E}\hat{\theta}_{\text{MLE}}}{\sqrt{D\hat{\theta}_{\text{MLE}}}} \xrightarrow{d} N(0, 1),$$

по асимптотической несмещенности,

$$\frac{\hat{\theta}_{\text{MLE}} - \theta}{\sqrt{D\hat{\theta}_{\text{MLE}}}} \xrightarrow{d} N(0, 1),$$

и, учитывая асимптотическую эффективность ( $D\hat{\theta}_{\text{MLE}}I_n(\theta) \xrightarrow[n \rightarrow \infty]{} 1$ ), запишем статистику

$$T = (\hat{\theta}_{\text{MLE}} - \theta) \sqrt{I_n(\theta)} \xrightarrow{d} N(0, 1).$$

Чтобы по аналогии с предыдущим выразить  $\theta$  в  $P(c_1 < T < c_2) = P(|T| < c_\gamma) = \gamma$ , необходимо выразить  $\theta$  из  $I_n(\theta)$ . Для Pois и Ber это эквивалентно решению квадратного уравнения.

В общем случае, можно вместо  $\theta$  в  $I_n(\theta)$  подставить  $\hat{\theta}_{\text{MLE}}$  (при  $n \rightarrow \infty$  это не должно сильно испортить дело), откуда

$$P(|T| < c_\gamma) = \gamma \iff P(-c_\gamma < (\hat{\theta}_{\text{MLE}} - \theta) \sqrt{I_n(\theta)} < c_\gamma) = \gamma \iff P\left(\theta \in \left(\hat{\theta}_{\text{MLE}} \pm \frac{c_\gamma}{\sqrt{I_n(\theta)}}\right)\right) = \gamma,$$

где

$$T \xrightarrow{d} N(0, 1) \implies c_\gamma = \text{cdf}_{N(0,1)}^{-1}\left(\frac{1+\gamma}{2}\right).$$

**Пример.**  $\xi \sim \text{Pois}(\lambda)$ . По 5.2,  $\hat{\lambda}_{\text{MLE}} = \bar{x}$ , по 7  $I_n(\lambda) = n/\lambda = n/\bar{x}$  откуда

$$P\left(\lambda \in \left(\bar{x} \pm \text{cdf}_{N(0,1)}^{-1}\left(\frac{1+\gamma}{2}\right) \frac{\sqrt{\bar{x}}}{\sqrt{n}}\right)\right) = \gamma.$$

**Пример.**  $\xi \sim \text{Ber}(p)$ .  $p = \mathbb{E}\xi$ .  $\hat{p} = \bar{x}$ , откуда

$$P\left(p \in \left(\hat{p} \pm c_\gamma \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)\right) \xrightarrow[n \rightarrow \infty]{} \gamma.$$

*Замечание.* Этот доверительный интервал не очень хорош, потому что не принадлежит  $[0, 1]$ .

## 5. Доверительный интервал для проверки гипотезы о значении параметра

Зафиксируем  $H_0 : \theta = \theta_0$  и  $\gamma = 1 - \alpha$ , где  $\alpha$  играет роль уровня значимости. По определению доверительного интервала,  $P(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \gamma$ . Тогда

$$P(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \gamma = 1 - \alpha \implies \alpha = 1 - P(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = P(\theta \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})])$$

и  $\mathcal{A}_{\text{крит}} = \mathbb{R} \setminus [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]$ . Соответственно,

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \theta \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})] \\ H_1 & \theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})] . \end{cases}$$

Иными словам, попадание в критическую область происходит с уровнем значимости  $\alpha$ , что соответствует определению критерия.

## 6. Использование SE для построения доверительных интервалов

Пусть оценка  $\hat{\theta}_n$  имеет какое-то симметричное распределение хотя бы асимптотически. Как и для любой другой случайной величины (с симметричным распределением), доверительный интервал уровня  $\gamma$  (т.е. такой интервал, в котором лежит  $\gamma$  всех значений величины) задается как

$$E\hat{\theta}_n \pm c_\gamma \sqrt{D\hat{\theta}_n},$$

где  $c_\gamma = \text{qnt } \gamma$ . К примеру, для  $N(0, 1)$  и 95%-квантили это был бы интервал  $(-1.96; 1.96)$ , а так нужно передвинуть его на среднее и растянуть на корень из дисперсии.

Но стандартное отклонение  $\sqrt{D\hat{\theta}_n}$  распределения  $\hat{\theta}_n$  можно оценить как SE. Значит доверительный интервал будет иметь вид

$$E\hat{\theta}_n \pm c_\gamma \text{SE}.$$

## 7. Доверительный интервал для двумерного параметра

Пусть  $\hat{\theta}_1, \hat{\theta}_2 \sim N(0, 1)$ . Доверительная область с уровнем доверия  $\gamma$   $C \subset \mathbb{R}^2 : P((\hat{\theta}_1, \hat{\theta}_2) \in C) = \gamma$  может быть построена в виде квадрата или диска с центром в начале координат (по симметричности нормального распределения):

- По независимости,

$$P((\hat{\theta}_1, \hat{\theta}_2) \in C) = P(\hat{\theta}_1 \in (-x_0, x_0))P(\hat{\theta}_2 \in (-y_0, y_0)) = \gamma,$$

откуда

$$x_0 = y_0 = \text{cdf}_{N(0,1)}^{-1} \left( \frac{1 + \sqrt{\gamma}}{2} \right).$$

- Найдем

$$x : P(\underbrace{\hat{\theta}_1^2 + \hat{\theta}_2^2}_{r^2} \leq x) = \gamma.$$

$r^2 \sim \chi^2(2) = \text{Exp}(\lambda)$ , где  $\lambda = 1/2$ , поскольку  $E\chi^2(2) = 2$ . Значит,

$$P(r^2 < x) = \text{cdf}_{\text{Exp}(1/2)}(x) = 1 - e^{-1/2x} = \gamma \implies x = -2\ln(1 - \gamma),$$

и радиус получившегося круга есть

$$r = \sqrt{x} = \sqrt{-2\ln(1 - \gamma)}.$$

Находя радиус и сторону квадрата для разных  $\gamma$ , можно посчитать площади получившихся областей:

| $\gamma$ | Площадь квадрата | Площадь круга |
|----------|------------------|---------------|
| 0.99     | 31.5             | 28.94         |
| 0.9      | 15.19            | 14.47         |
| 0.8      | 10.48            | 10.11         |
| $\vdots$ | $\vdots$         | $\vdots$      |
| 0.3      | 2.26             | 2.24          |
| 0.2      | 1.41             | 1.40          |

Так как площадь круга всегда меньше площади квадрата, он является предпочтительным как доверительная область.

Часть V.

## Корреляционный анализ

**Определение.** Мера зависимости — это функционал  $r : (\xi, \eta) \mapsto x \in [-1, 1]$  со свойствами:

1.  $|r| \leq 1$ .
2.  $\xi \perp\!\!\!\perp \eta \implies r(\xi, \eta) = 0$ .
3. Если  $\xi$  и  $\eta$  «максимально зависимы», то  $r(\xi, \eta) = 1$ .

# 1. Вероятностная независимость

## 1.1. Визуальное определение независимости

- Поскольку при  $p_\eta(y_0) \neq 0$

$$\xi \perp\!\!\!\perp \eta \iff p_{\xi|\eta}(x | y_0) = \frac{p_{\xi,\eta}(x, y_0)}{p_\eta(y_0)} = p_\xi(x),$$

то срезы графика совместной плотности при фиксированном  $y_0$  после нормировки  $p_\eta(y_0)$  должны выглядеть одинаково для всех  $y_0$ .

- Для выборки независимость можно попытаться определить по *таблицам сопряженности*: сгруппируем  $\{(x_i, y_i)\}_{i=1}^n$  и сопоставим каждой уникальной паре абсолютную частоту  $\nu_{ij}$ :

$$\begin{array}{cccc} & y_1^* & \cdots & y_s^* \\ x_1^* & \nu_{11} & \cdots & \nu_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^* & \nu_{k1} & \cdots & \nu_{ks} \end{array}$$

Тогда признаки с большей чем случайной вероятностью будут независимы при пропорциональных строках / столбцах. Более формально, признаки независимы, если

$$\frac{\nu_{ij}}{\sum_k \nu_{kj}} = \frac{\nu_{ij}}{\nu_{\cdot j}} = \hat{p}_{i|j} \propto \hat{p}_{i|\ell},$$

т.е. вероятности условного распределения не зависят от выбора строки.

**Пример.** Таблица сопряженности похожей на независимую выборки:

$$\begin{array}{ccc} 1 & 3 & 2 \\ 2 & 5 & 3 \\ 9 & 20 & 11 \end{array}$$

## 1.2. Критерий независимости $\chi^2$

По определению, для двумерных дискретных распределений, независимость есть

$$\xi \perp\!\!\!\perp \eta \iff \underbrace{P(\xi = i, \eta = j)}_{p_{ij}} = \underbrace{P(\xi = i)}_{p_{i\cdot}} \underbrace{P(\eta = j)}_{p_{\cdot j}} = \underbrace{\sum_{k=1}^K P(\xi = i, \eta = k)}_{p_{i\cdot}} \cdot \underbrace{\sum_{s=1}^S P(\xi = s, \eta = j)}_{p_{\cdot j}}.$$

Проверим  $H_0 : \xi \perp\!\!\!\perp \eta$ .

*Утверждение.* ОМП оценкой будет  $\hat{p}_{i\cdot} = \nu_{i\cdot}/n$  и  $\hat{p}_{\cdot j} = \nu_{\cdot j}/n$ .

Следовательно,

$$\xi \perp\!\!\!\perp \eta \iff \hat{p}_{ij} = \frac{\nu_{ij}}{n} = \hat{p}_{i\cdot} \hat{p}_{\cdot j} = \frac{\nu_{i\cdot}}{n} \cdot \frac{\nu_{\cdot j}}{n}.$$

Это равенство удается получить редко; важно определить, не является ли это нарушение случайным.

Запишем статистику

$$\chi^2 = \sum_{i=1}^K \sum_{j=1}^S \frac{(\nu_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^K \sum_{j=1}^S \frac{(\nu_{ij} - \nu_{i\cdot}\nu_{\cdot j}/n)^2}{\nu_{i\cdot}\nu_{\cdot j}/n} \xrightarrow{d} \chi^2((k-1)(s-1))$$

Количество параметров таково, потому что если  $\xi \parallel \eta$ , то всего  $ks - 1$  параметров ( $-1$  потому что  $\sum_{ij} p_{ij} = 1$ ); если  $\xi \perp\!\!\!\perp \eta$ , то  $k + s - 2$  ( $-2$  потому что  $\sum_i p_{ij} = 1$  и  $\sum_j p_{ij} = 1$ ). Значит  $ks - 1 - k - s + 2 = (k-1)(s-1)$ .

**Пример.** Дано  $S$  кубиков. Проверить гипотезу, что кубики одинаковы.

*Решение.* FIXME ┘

*Замечание.* На маленьких выборках ( $n < 40$ ,  $np_{ij} < 5$ ) возникают проблемы со сходимостью, потому что можно объединять только столбцы / строки и каждый раз терять сразу  $S - 1$  ( $K - 1$ ) степень свободы. В этих случаях используют критерием с перестановкой<sup>1</sup> или, в случае таблиц сопряженности  $2 \times 2$ , точным критерием Фишера.

*Замечание.* Критерий верен для количественных, порядковых и качественных признаков, потому что нигде не участвуют значения из выборки.

*Замечание.* Критерий асимптотический, поэтому  $\alpha_1 \rightarrow \alpha$ .

*Замечание.* Критерий не удовлетворяет 1-му пункту определения меры зависимости ( $\chi^2 \notin [-1, 1]$ ). Это обычно исправляют так: рассматривают *среднеквадратичную сопряженность*

$$r^2 := \frac{\chi^2}{n}$$

или коэффициент сопряженности Пирсона

$$p^2 := \frac{\chi^2}{\chi^2 + n}$$

(тогда 1 никогда не достигается). Могли бы работать с  $1 - p$ -value, но так почему-то никогда не делают.

<sup>1</sup>[https://en.wikipedia.org/wiki/Resampling\\_\(statistics\)#Permutation\\_tests](https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests)



## 2. Линейная / полиномиальная зависимость

Пусть теперь  $\xi, \eta$  — количественные признаки.

**Определение.** Определим

$$\phi(x) := E\{\eta \mid \xi = x\}.$$

Тогда назовем зависимость *линейной*, если  $\phi(x)$  — линейная функция, *квадратичной* — если квадратичная и т.д.

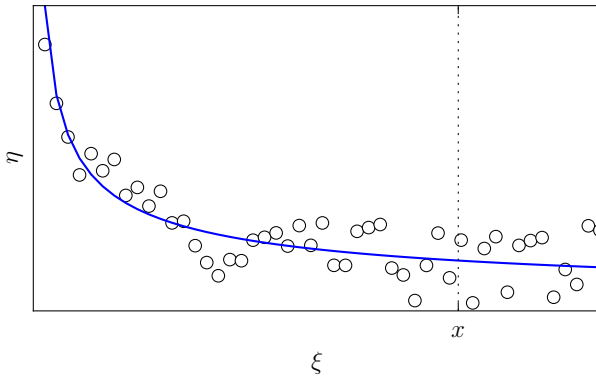


Рис. 2.1.: Нелинейная зависимость

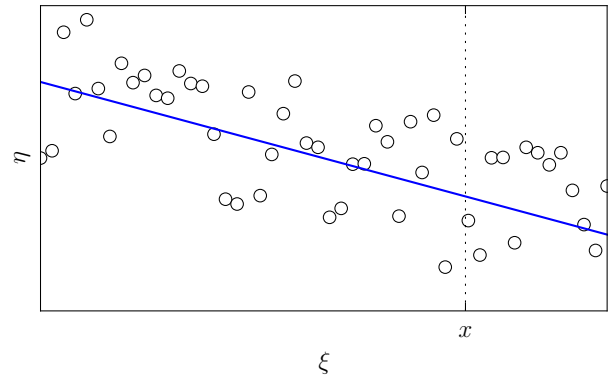


Рис. 2.2.: Линейная зависимость

**Определение.** Мера *линейной* зависимости между случайными величинами  $\xi$  и  $\eta$  есть *коэффициент корреляции Пирсона*

$$\rho = \frac{\text{cov}(\xi, \eta)}{\sqrt{D\xi}\sqrt{D\eta}}.$$

*Замечание.* Про  $\rho$  можно думать как про  $\cos$  между векторами в соответствующем пространстве.

*Замечание* (Важное).

$$\begin{aligned} \xi \perp \eta &\implies \rho = 0 \\ \xi, \eta \sim N(\mu, \sigma^2), \xi \perp \eta &\iff \rho = 0. \end{aligned}$$

**Предложение.** Для линейно зависимых данных, конечно,  $\rho = 1$ .

*Доказательство.* Пусть  $\eta = a + b\xi$ ; тогда

$$\begin{aligned} \rho(\xi, \eta) &= \frac{\text{cov}(\xi, a + b\xi)}{\sqrt{D\xi}\sqrt{D(a + b\xi)}} = \frac{E\xi(a + b\xi) - E\xi E(a + b\xi)}{\sqrt{D\xi}\sqrt{Db\xi}} = \frac{E\xi a + bE\xi^2 - E\xi E a - E\xi bE\xi}{b\sqrt{D\xi}\sqrt{D\xi}} = \\ &= \frac{aE\xi + bE\xi^2 - aE\xi - b(E\xi)^2}{bD\xi} = \frac{b(E\xi^2 - (E\xi)^2)}{bD\xi} = 1. \end{aligned}$$

□

### 2.1. О соотношении $\rho$ и коэффициента линейной регрессии

По (2), если линейная регрессия уравнением  $y = kx + b$ , то

$$k = \rho \frac{\sigma_\eta}{\sigma_\xi}.$$

В общем случае, по виду прямой линейной регрессии ничего нельзя сказать о зависимости между случайными величинами. Так, если  $\eta = a + b\xi$  есть линейная функция от  $\xi$ , то, по предыдущему,  $\rho = 1$  и

$$k = 1 \cdot \frac{\sqrt{D(a + b\xi)}}{\sqrt{D\xi}} = b$$

и прямая может иметь произвольный, в зависимости от  $b$ , наклон.

*Замечание.* В то же время, поскольку для

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} \sim N(\boldsymbol{\mu}, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_\xi^2 & \text{cov}(\xi, \eta) \\ \text{cov}(\xi, \eta) & \sigma_\eta^2 \end{pmatrix}$$

справедливо, что

$$k = \rho \frac{\sigma_\eta}{\sigma_\xi} = \frac{\text{cov}(\xi, \eta)}{\sigma_\xi \sigma_\eta} \cdot \frac{\sigma_\eta}{\sigma_\xi} = \frac{1}{\sigma_\xi^2} \text{cov}(\xi, \eta),$$

то  $k = 0 \iff \text{cov}(\xi, \eta) = 0$ , а для стандартно нормальных данных  $k = \rho = \text{cov}(\xi, \eta)$ .

## 2.2. Значимость коэффициента корреляции

**Определение.** Коэффициент корреляции *значим*, если отвергается  $H_0 : \rho = 0$ .

Чаще, чем  $H_0 : \rho = \rho_0$ , проверяют  $H_0 : \rho > \rho_0$ . Если  $\rho_0 = 0$ , то  $(\xi, \eta)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  и, по ЦПТ,

$$T = \frac{\sqrt{n-2}\hat{\rho}_n}{\sqrt{1-\hat{\rho}_n^2}} \sim t(n-2).$$

Идеальное значение — 0, два хвоста.

Если  $\rho_0 \neq 0$ , то ЦПТ не работает, тогда распределение  $\hat{\rho}$  неизвестно. Тогда применяется  $z$ -преобразование Фишера

$$z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \quad z_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}.$$

Тогда ЦПТ работает и, если  $(\xi, \eta)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,

$$T = \sqrt{n-3}(z - z_0) \xrightarrow{d} N(0, 1).$$

### 3. Метод наименьших квадратов (Ordinary Least Squares)

Пусть  $\eta, \xi \in L^2(\mathcal{F}, \mathbf{P})$  пространству  $\mathcal{F}$ -измеримых по мере  $\mathbf{P}$  функций с конечным вторым моментом и скалярным произведением  $(\eta, \xi) = \mathbf{E}\eta\xi$ , причем  $\hat{\eta} \in K = \{\phi(\xi)\} = \{\hat{\eta} : \sigma(\phi(\xi))\text{-измерима}\}$ . По свойству УМО(6.2), вектор

$$\hat{\eta}^* = \mathbf{E}\{\eta \mid \phi(\xi)\}$$

будет ортогональной проекцией  $\eta$  на  $K$ , т.е.  $(\eta - \hat{\eta}^*, \hat{\eta}) = 0 \ \forall \hat{\eta} \in K$ . Значит, он минимизирует квадрат нормы расстояния от  $\eta$  до  $K$ :

$$\hat{\eta}^* = \operatorname{argmin}_{\hat{\eta} \in K} \|\eta - \hat{\eta}\|^2 = \operatorname{argmin}_{\hat{\eta} \in K} \mathbf{E}(\eta - \hat{\eta})^2 = \mathbf{E}\{\eta \mid \phi(\xi)\}.$$

$\hat{\eta}^*$  называется наилучшим среднеквадратичным приближением в классе  $K$ .

## 4. Корреляционное отношение

Если  $K = \mathcal{L} = \{a\xi + b\}$  — линейное пространство, то теорема Пифагора принимает вид

$$D\eta = E(\eta - E\eta)^2 = \underbrace{E(\hat{\eta}^* - E\eta)^2}_{\text{объяснённая доля аппроксимации}} + \underbrace{E(\eta - \hat{\eta}^*)^2}_{\text{ошибка аппроксимации}}.$$

Откуда можно записать меру аппроксимации

$$\frac{E(\hat{\eta}^* - E\eta)^2}{D\eta} = 1 - \frac{E(\eta - \hat{\eta}^*)^2}{D\eta} = 1 - \frac{\min_{\hat{\eta} \in \mathcal{L}} E(\eta - \hat{\eta})^2}{D\eta}.$$

**Определение.** Полученная величина называется коэффициентом корреляции  $\rho^2$ :

$$\rho^2 := 1 - \frac{\min_{\hat{\eta} \in \mathcal{L}} E(\eta - \hat{\eta})^2}{D\eta}.$$

$\rho$  — коэффициент корреляции Пирсона.

**Определение.** Множественный коэффициент корреляции есть полученная величина для МНК с  $K = \mathcal{M} = \left\{ \sum_{i=1}^k b_i \xi_i + b_0 \right\}$ .

$$R^2(\eta, \xi_1, \dots, \xi_k) := 1 - \frac{\min_{\hat{\eta} \in \mathcal{M}} E(\eta - \hat{\eta})^2}{D\eta}.$$

*Замечание.*  $R^2 \geq \rho^2$ ; если же  $R^2 = \rho^2$ , то  $\xi_1, \dots, \xi_k$  все зависимы.

**Определение.** В общем случае, если  $K = \{\phi(\xi) \text{ измеримые}\}$ , то полученная величина называется *корреляционным отношением*:

$$r_{\eta|\xi}^2 := 1 - \frac{\min_{\hat{\eta} \in K} E(\eta - \hat{\eta})^2}{D\eta} = \frac{DE(\eta | \xi)}{D\eta}.$$

### 4.1. Свойства корреляционного отношения

1.  $r_{\eta|\xi}^2 \in [0, 1]$ .
2.  $\eta \perp \xi \implies r_{\eta|\xi}^2 = 0$ .
3.  $\eta = \phi(\xi) \iff r_{\eta|\xi}^2 = 1$ .
4. Вообще говоря,  $r_{\eta|\xi}^2 \neq r_{\xi|\eta}^2$ . К примеру, для любой не монотонной функции (так, чтобы не существовала обратная).
5.  $r_{\eta|\xi}^2 \geq \rho^2(\eta, \xi)$  (потому что минимум по всем функциям меньше, чем лишь по линейным, значит  $1 - \min$  больше).
6.  $(\xi, \eta)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies r_{\eta|\xi}^2 = \rho^2(\eta, \xi)$ .

## 4.2. Выборочное корреляционное отношение

По разложению дисперсии,

$$D\eta = E(\eta - E\eta)^2 = \underbrace{E(E(\eta | \xi) - E\eta)^2}_{DE(\eta|\xi)} + E(\eta - E(\eta | \xi))^2.$$

Перейдем на выборочный язык. Пусть дана выборка

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}.$$

Сгруппируем её:

$$\begin{array}{c|ccc} x_1^* & y_{11} & \dots & y_{1n_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^* & y_{k1} & \dots & y_{kn_k} \end{array}$$

Пусть  $\xi$  — дискретная случайная величина со значениями  $(x_1^*, \dots, x_k^*)$ . Тогда, учитывая

$$\bar{y}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \hat{E}(\eta | \xi = x_i^*),$$

на выборочном языке получаем (домножив на  $n$ ):

$$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}_{\text{total sum of squares}} = \underbrace{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}_{\text{межгрупповой разброс}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{\text{внутригрупповой разброс}}$$

$$ns_y^2 = ns_{y|x}^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Отсюда, так как,  $r_{\eta|\xi}^2 = DE(\eta | \xi)/D\eta$ ,

$$\hat{r}_{\eta|\xi}^2 = \hat{r}_{y|x}^2 = \frac{s_{y|x}^2}{s_y^2}.$$

## 5. Частная корреляция

**Определение.** Частная корреляция случайных величин  $\eta_1, \eta_2$  относительно  $\{\xi_1, \dots, \xi_k\}$  есть

$$\rho(\eta_1, \eta_2 \mid \{\xi_1, \dots, \xi_k\}) := \rho(\eta_1 - \hat{\eta}_1^*, \eta_2 - \hat{\eta}_2^*), \quad \text{где } \hat{\eta}_i^* = \underset{\hat{\eta}_i \in \{\sum_{i=1}^k b_i \xi_i + b_0\}}{\operatorname{argmin}} \mathbb{E}(\eta_i - \hat{\eta}_i)^2.$$

Если регрессия линейна, то

$$\rho(\eta_1, \eta_2 \mid \xi_1, \dots, \xi_k) = \rho(\eta_1 - \mathbb{E}\{\eta_1 \mid \xi_1, \dots, \xi_k\}, \eta_2 - \mathbb{E}\{\eta_2 \mid \xi_1, \dots, \xi_k\}).$$

*Замечание (Важное).* Пусть в эксперименте подсчитан ненулевой  $\rho$ . Это может означать, что один из факторов является причиной, а другой следствием; чтобы установить, что есть что, проводят эксперимент и смотрят, какой фактор в реальности влияет на какой. Это может также означать, что влияет сторонний фактор. Чтобы его исключить, считают частную корреляцию.

**Пример.** Возможна ситуация, когда  $\rho(\eta_1, \eta_2) \neq 0$ , но  $\rho(\eta_1, \eta_2 \mid \xi) = 0$ . Частная корреляция есть, по сути, корреляция на центрированных данных.

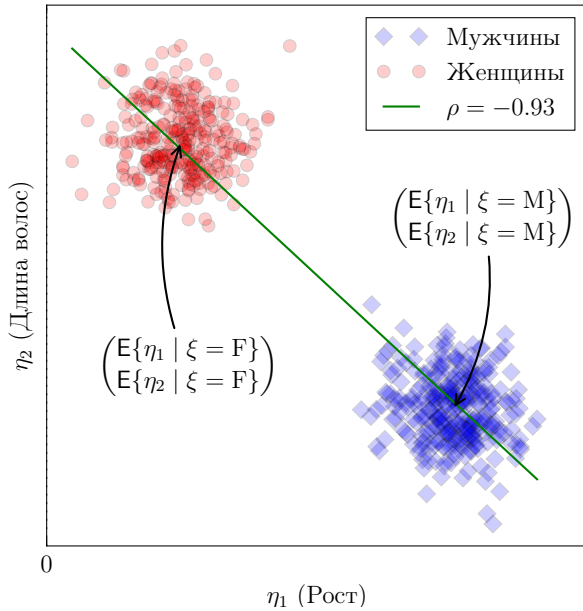


Рис. 5.1.: Исходные данные (бимодальность)

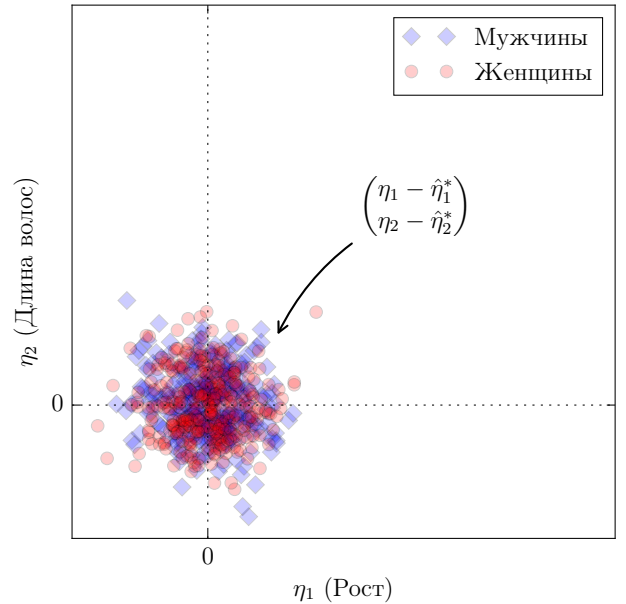


Рис. 5.2.: Центрированные данные

**Пример.** Возможна и ситуация как на (5.3), где определено  $\rho(\eta_1, \eta_2) > 0$ , но  $\rho(\eta_1, \eta_2 \mid \xi) < 0$ .

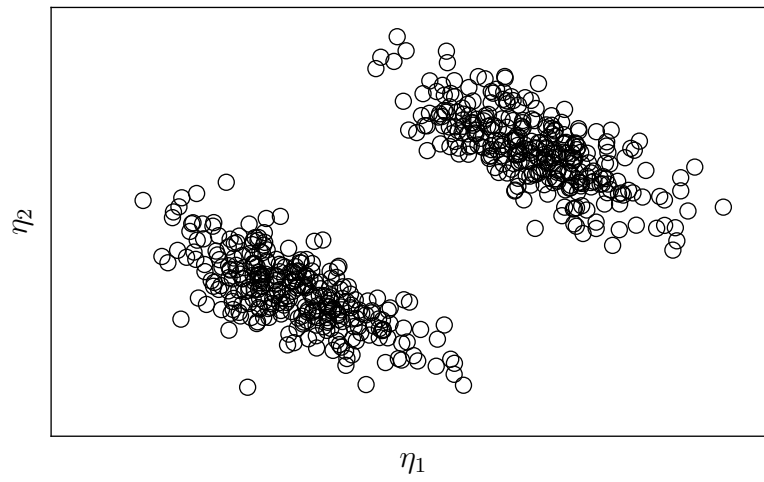


Рис. 5.3.:  $\rho(\eta_1, \eta_2) > 0$ , но  $\rho(\eta_1, \eta_2 \mid \xi) < 0$

*Замечание.* По аналогии с предыдущим примером, если  $|\text{im } \xi| \rightarrow \infty$ , то графики  $(\eta_1, \eta_2)$  при фиксированном  $\xi$  образуют эллипсоид (в этом случае с положительной корреляцией).

## 6. Зависимость между порядковыми признаками

Пусть на выборке

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} \sim \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

задан только порядок. Тогда можем считать только эмпирическую функцию распределения.

Следующие коэффициенты основаны на рангах. Ранговые характеристики хорошо работают на выборках *без повторений* (чтобы, к примеру, не возникало дробных рангов).

### 6.1. Ранговый коэффициент Спирмана

**Определение.** Ранговый коэффициент Спирмана есть

$$\rho_S = \rho(\text{cdf}_\xi(\xi), \text{cdf}_\eta(\eta)).$$

*Замечание.*  $\text{cdf}_\xi(\xi) \sim U(0, 1)$ , потому что  $P(\text{cdf}_\xi(\xi) < x) = P(\xi < \text{cdf}_\xi^{-1}(x)) = \text{cdf}_\xi(\text{cdf}_\xi^{-1}(x)) = x$ .

**Определение.** Ранг элемента из выборки есть его порядковый номер в упорядоченной выборке:

$$\text{rk } x_{(i)} = i.$$

*Обозначение.*  $\text{rk } x_{(i)} =: R_i$ ,  $\text{rk } y_{(i)} =: T_i$ .

Можем ввести эмпирическое распределение

$$\text{cdf}_{\xi_n}(x_i) = \frac{\text{rk } x_i}{n}, \quad \text{cdf}_{\eta_n}(y_i) = \frac{\text{rk } y_i}{n} = \frac{T_i}{n}.$$

Тогда будет справедливо следующее

**Определение.** Выборочный коэффициент Спирмана определяется как выборочный коэффициент корреляции Пирсона  $\hat{\rho}$ , но с заменой значений на ранги:

$$\hat{\rho}_S \begin{pmatrix} \xi_n \\ \eta_n \end{pmatrix} = \rho \begin{pmatrix} R_n \\ T_n \end{pmatrix} = \frac{1/n \cdot \sum_{i=1}^n R_i T_i - \bar{R} \bar{T}}{\sqrt{1/n \cdot \sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{1/n \cdot \sum_{i=1}^n (T_i - \bar{T})^2}}.$$

Если нет повторяющихся наблюдений, то знаменатель будет одним и тем же у всех выборок объема  $n$ , значит его можно посчитать заранее. В этом (и только этом) случае, справедлива более простая формула:

$$\hat{\rho}_S = 1 - \frac{6 \sum_{i=1}^n (R_i - T_i)^2}{n^3 - n}.$$

*Замечание.* Из последней формулы хорошо видно, что если  $x_i, y_i$  все идут в одном порядке, то  $R_i - T_i = 0 \ \forall i$  и  $\hat{\rho}_S = 1$ .

*Замечание.*  $\rho_S$  для количественных признаков есть мера монотонной зависимости:

$$\rho_S = 1 \iff (x_i > x_{i+1} \implies y_i > y_{i+1} \ \forall i)$$

(даже если зависимость нелинейная и  $\rho \neq 1$ ). Иными словами,  $\rho_S > 0$ , если  $y$  имеет тенденцию к возрастанию с возрастанием  $x$  (и  $\rho_S < 0$  иначе). Чем большее  $\rho_S$ , тем более явно выражена зависимость  $y$  от  $x$  в виде некоторой монотонной функции.



### 6.1.1. Согласованность $\rho$ и $\rho_S$

$\rho_S$  не согласована с  $\rho$  в том же смысле, что  $\rho$  и  $r_{\xi|\eta}$ .

*Утверждение.* Если данные нормальные то справедлива формула

$$\rho = 2 \sin \left( \frac{\pi}{6} \rho_S \right).$$

Значит, можем сравнить критерии между собой.

- С точностью до погрешности, по значению,  $\hat{\rho}$  и  $\hat{\rho}_S$  — это одно и то же (см. 6.1)

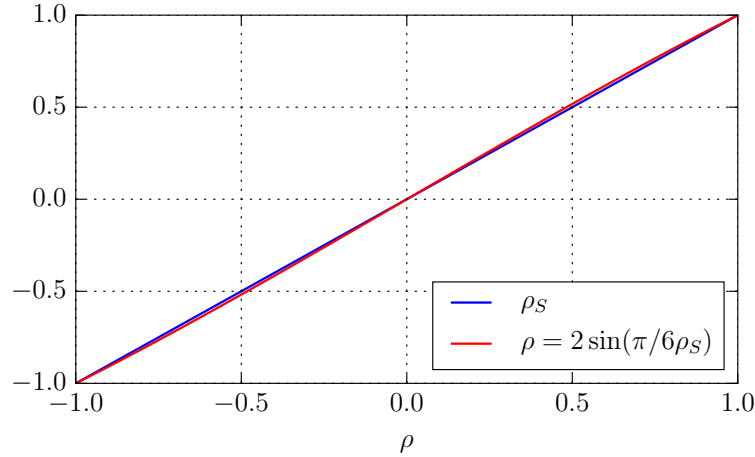


Рис. 6.1.:  $\hat{\rho} \approx \hat{\rho}_S$

- Обычный критерий оценки — выборочную дисперсию — посчитать сложно. Тем не менее, можем заметить, что  $\hat{\rho}_S$  более устойчив к аутлаерам (см. 6.2). Всегда можно добавить аутлаер такой, что  $\hat{\rho} = 0$ ;  $\hat{\rho}_S$  же поменяется не сильно. Поэтому для нормальных данных,  $\rho_S$  — это оценка, что нет аутлаеров.

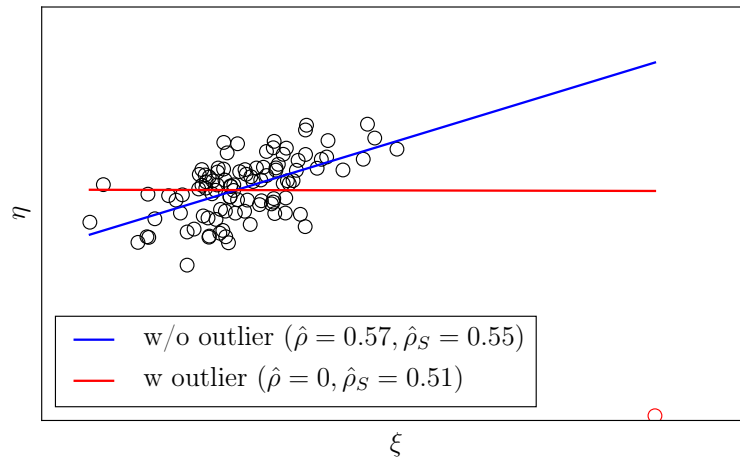


Рис. 6.2.:  $\hat{\rho}$  до и после добавления аутлаера

- Монотонным преобразованием можем всегда сделать так, чтобы  $\rho$  изменился (например, возведя в квадрат); при монотонном преобразовании, однако, не меняется  $\rho_S$  (см. 6.3). Значит, чтобы узнать  $\rho$  исходных (нормальных) данных, можно не выполнять обратного преобразования, а сразу посчитать  $\rho_S$ .

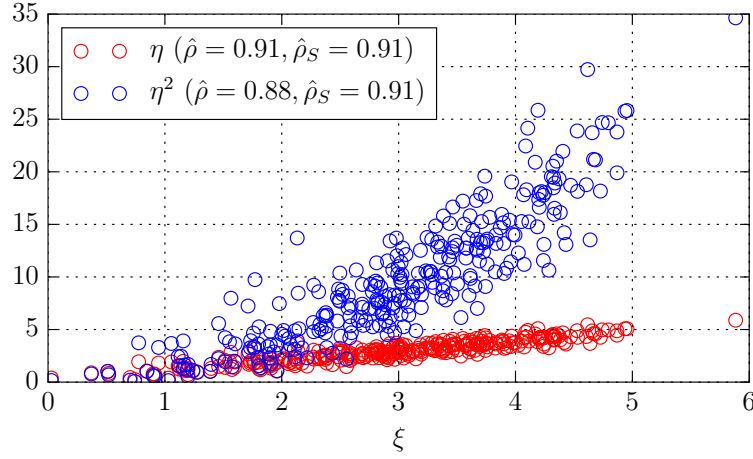


Рис. 6.3.: Монотонное преобразование нормальных данных

## 6.2. Ранговый коэффициент Кэндалла $\tau(\xi, \eta)$

**Определение.** Пусть  $(\xi_1, \eta_1)^\top \perp (\xi_2, \eta_2)^\top \sim \mathcal{P}_{\xi, \eta} \sim (\xi, \eta)^\top$ ; тогда *ранговым коэффициентом Кэндалла* называется

$$\tau(\xi, \eta) = \rho(\text{sign}(\xi_2 - \xi_1), \text{sign}(\eta_2 - \eta_1)) = P((\xi_2 - \xi_1)(\eta_2 - \eta_1) > 0) - P((\xi_2 - \xi_1)(\eta_2 - \eta_1) < 0).$$

На выборочном языке, пусть дана выборка  $(x_1, y_1), \dots, (x_n, y_n)$ ; тогда

$$\tau = \frac{\#(\text{одинаково упорядоченных пар}) - \#(\text{по-разному упорядоченных пар})}{\#(\text{комбинаций пар})},$$

где пара  $(x_i, y_i), (x_j, y_j)$  считается одинаково упорядоченной, если  $\text{sign}(x_i - x_j) = \text{sign}(y_i - y_j)$ , а  $\#(\text{комбинаций пар}) = C_n^2 = n(n-1)/2$ .

**Утверждение.** Если  $(\xi, \eta)^\top \sim N(\boldsymbol{\mu}, \Sigma)$ , то справедлива формула

$$\rho = \sin\left(\frac{\pi}{2}\tau\right).$$

Из утверждения следует, что  $\tau$  все время меньше  $\rho$  и  $\rho_S$  (по модулю).

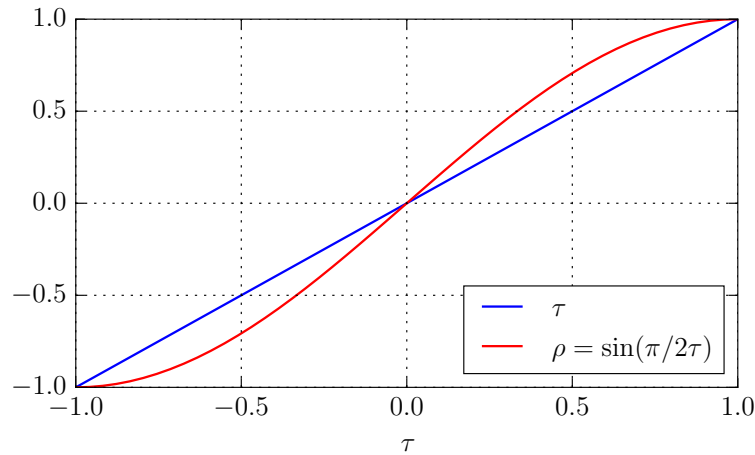


Рис. 6.4.:  $\rho$  и  $\tau$

**Пример** (Проверка ряда на тренд). Пусть  $\xi$  — номера точек, а  $\eta$  — значения ряда. Тогда  $H_0 : \tau_0 = 0$  и если  $H_0$  отвергается, то тренд присутствует.

## 7. Корреляционные матрицы

Если признаков много, то их наглядно характеризуют корреляционные матрицы. Улучшить наглядность можно переупорядочив признаки так, чтобы на диагонали матрицы стояли блоки корреляций признаков из «корреляционных плеяд».

**Определение.** Пусть  $\rho_0$ ; корреляционная плеяда есть множество признаков, таких, что их попарная корреляция больше  $\rho_0$ .

Можно выделить и несколько уровней  $\rho_i : \rho_0 < \rho_1 < \dots$ . Тогда сначала следует составить плеяду по  $\rho_0$ , затем внутри полученного по  $\rho_1$  и т.д.

Часть VI.

## Дисперсионный анализ

# 1. Однофакторный дисперсионный анализ (One-way ANOVA<sup>1</sup>)

Задача может быть поставлена двумя эквивалентными образами:

1. Пусть  $\eta_i \sim \mathcal{P}_i$ ,  $i \in 1 : k$ . Проверить гипотезу, что все распределения равны:

$$H_0 : \mathcal{P}_1 = \dots = \mathcal{P}_k.$$

2. Пусть дан двумерный вектор  $(\xi \quad \eta)^\top$ , причем  $\xi$  («фактор») принимает  $k$  значений  $A_1, \dots, A_k$ . Рассмотрим  $\eta_i \sim \mathcal{P}_i = \mathcal{P}_{\eta|\xi=A_i}$ . Проверить гипотезу

$$H_0 : \mathcal{P}_{\eta|\xi=A_1} = \dots = \mathcal{P}_{\eta|\xi=A_k}.$$

Пусть теперь  $\eta_i \sim N(\mu_i, \sigma^2)$ . Разумеется,

$$\begin{aligned} H_0 : \mu_1 = \dots = \mu_k &\iff H_0 : E\eta_1 = \dots = E\eta_k \\ &\iff H_0 : E(\eta \mid \xi = A_1) = \dots = E(\eta \mid \xi = A_k) \iff H_0 : DE(\eta \mid \xi) = 0. \end{aligned}$$

Для построения критерия, вспомним разложение дисперсии на выборочном языке:

$$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}_{Q = \hat{D}\eta} = \underbrace{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}_{Q_1 = \hat{D}E(\eta|\xi)} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{Q_2}$$

откуда в качестве критерия (следуя гипотезе) выберем  $Q_1$  с идеальным значением 0. Однако  $Q_1$  полезно отнормировать по  $Q_2$  для учета различных внутригрупповых разбросов. Чтобы получить статистику с известным распределением, вспомним, что по теореме Cochran,  $Q_1 \perp\!\!\!\perp Q_2$ ,

$$\frac{Q}{\sigma^2} \sim \chi^2(n-1), \quad \frac{Q_1}{\sigma^2} \sim \chi^2(k-1), \quad \frac{Q_2}{\sigma^2} \sim \chi^2(n-k)$$

и

$$t = \frac{Q_1/(k-1)}{Q_2/(n-k)} \sim F((k-1), (n-k)).$$

*Замечание.* Это обобщение статистики для проверки гипотезы о равенстве математических ожиданий независимых двумерных выборок с равными дисперсиями (с  $k=2$ , то есть):

$$t = \frac{\bar{x} - \bar{y}}{\tilde{s}_{1,2} \sqrt{1/n_1 + 1/n_2}}$$

с  $\tilde{s}_{1,2}^2 = Q_2/(n-2)$ . Дело в том, что статистики распределены одинаково — по определению,

$$t^2(n-2) = F(1, n-2).$$

Чтобы воспользоваться полученным критерием, должно убедиться, что дисперсии одинаковые. Как и в случае  $k=2$ , это можно проверить по тесту Левена, только многомерному, т.е. проверить равенство математических ожиданий  $E|\xi - E\xi_i| \quad \forall i \in 1 : k$ ,  $|y_{ij} - \bar{y}_i|$  — опять же, через саму ANOVA.

---

<sup>1</sup>ANalysis Of VArIation

## 1. Однофакторный дисперсионный анализ (One-way ANOVA<sup>2</sup>)

*Замечание.* Если условия нормальности нарушаются, то критерий становится асимптотическим. Тогда вместо  $F$  следует использовать  $\chi^2$ , так как  $F(k, m)/m \xrightarrow{m \rightarrow \infty} \chi^2(k)$ .

**Пример.** Пусть дана выборка вида  $\{(\xi = \text{пол}, \eta = \text{вес})\}$ . Выдвинем  $H_0$  : вес не зависит от пола. Очевидно, что  $\xi$  — категориальная случайная величина, а  $\eta$  — количественная. Значения  $\xi$  разобьют всю выборку на две (?) группы. Тогда проверка гипотезы сведется к проверке равенства распределений в двух группах,  $\mathcal{P}_{\eta|\xi=s_1} = \mathcal{P}_{\eta|\xi=s_2}$ . В предположении, что  $(\eta | \xi = s_i) \sim N(\mu_i, \sigma^2)$ , равенство распределений будет следовать из равенства математических ожиданий.

## 2. Множественные сравнения

**Пример.** Проблема множественных сравнений возникает, например, в следующих ситуациях.

- Пусть одна группа испытуемых принимает лекарство, а вторая нет. По завершению эксперимента две группы сравниваются по  $m$  показателям. Однако чем больше показателей сравнивается, тем больше вероятность того, что *хотя бы по одному* показателю будет совпадение (в силу случайности).
- Испытывают  $m = 100$  монет на честность сериями по  $n = 10$  бросков:  $\{(\xi_1^{(1)}, \dots, \xi_{10}^{(1)}), \dots, (\xi_1^{(100)}, \dots, \xi_{10}^{(100)})\}$ , иными словами  $\{\eta^{(1)}, \dots, \eta^{(100)}\}$ , где  $\eta^{(i)} \sim \text{Bin}(10, p)$ . Проверить  $m$  гипотез  $H_0^{(i)} : \eta^{(i)} \sim \text{Bin}(10, 1/2)$ ,  $i \in 1 : m$ . Зафиксируем  $\alpha = 0.05$ . Тогда, учитывая  $\text{pmf}_{\text{Bin}(10, 1/2)}(k) = C_{10}^k 2^{-10}$ ,  $P_{H_0^{(i)}}(\eta^{(i)} \geq 9) = 10 \cdot 2^{-10} + 1 \cdot 2^{-10} \approx 0.0107$ , однако уже  $P_{H_0^{(i)}}(\eta^{(i)} \geq 8) \approx 0.0546$ . Так что критерием наибольшей мощности будет  $\eta^{(i)} \geq 9$ :

$$\alpha_1 = P_{H_0^{(i)}}(H_0^{(i)} \text{ отв}) = P_{H_0^{(i)}}(\eta^{(i)} \geq 9) \approx 0.0107 \leq 0.05.$$

Но использование того же критерия для множественных сравнений сильно завышает  $\alpha_1$ :

$$\begin{aligned} P\left(\bigvee_{i=1}^{100} H_0^{(i)} \text{ отв}\right) &= 1 - P\left(\bigwedge_{i=1}^{100} H_0^{(i)} \text{ не отв}\right) = 1 - \left(1 - P(H_0^{(i)} \text{ отв})\right)^{100} \\ &= 1 - (1 - 0.0107)^{100} \approx 0.6589. \end{aligned}$$

Пусть проверяются гипотезы  $H_0^{(1)}, \dots, H_0^{(m)}$ . Возможны такие ситуации:

|             | Retain $H_0$ (критерий не значим) | Reject $H_0$ (критерий значим) |
|-------------|-----------------------------------|--------------------------------|
| True $H_0$  | # True Negative                   | # False Discovery              |
| False $H_0$ | # False Negative                  | # True Discovery               |

Используя обозначения таблички,

$$\alpha_I \approx \frac{\text{FD}}{\text{TN} + \text{FD}}, \quad \alpha_{II} \approx \frac{\text{FN}}{\text{FN} + \text{TD}}.$$

**Определение.** Family-wise error rate (FWER):

$$\text{FWER} = P(\text{хотя бы один раз отвергнута верная гипотеза}) = P\left(\bigvee_{i=1}^m H_0^{(i)} \text{ отв}\right).$$

Иными словами, FWER — это ошибка первого рода для всей совокупности экспериментов.

Требуется контролировать FWER на предзаданном уровне  $\alpha$ , т.е. чтобы  $\text{FWER} \sim \alpha$ , где  $\sim \in \{=, \leq, \rightarrow\}$ . В *слабом* смысле это осуществляется, если  $\text{FWER} \sim \alpha$  только если *все*  $H_0^{(i)}$ ,  $i \in 1 : m$  верны. В *сильном* смысле контроль FWER на уровне  $\alpha$  гарантируется для *любой* конфигурации верных и не верных  $H_0^{(j)}$ .

**Определение.** Пусть  $T_0 := \{i : H^{(i)} \text{ верна}\}$ . Тогда

$$\text{weak FWER}_T = P(\text{хотя бы один раз отвергнута верная гипотеза, если верны } H^{(i)}, i \in T)$$

т.е. если  $T_0 = T$ .

**Определение.**

$$\text{strong FWER} = \max_{T \subset \{1, \dots, m\}} \text{weak FWER}_T.$$

*Обозначение.*  $\text{FWER}_T := \text{weak FWER}_T$ .

Это осуществляется двумя процедурами:

- Single
- Stepdown

## 2.1. Single

Каждая  $H^{(i)}$  проверяется отдельно с уровнем значимости  $\alpha_1$ . Задача сводится к тому, чтобы найти такое  $\alpha_1$ , что  $\text{FWER} \leq \alpha$  для какого-то нужного предзаданного  $\alpha$ . Пусть  $T = 1 : m$ , т.е. будто все тесты верны; тогда

$$\text{FWER}_{\{1, \dots, m\}} = P\left(\bigvee_{i=1}^m H_0^{(i)} \text{ отв}\right) \leq \sum_{i=1}^m P(H^{(i)} \text{ отв}) = m\alpha_1 = \alpha \implies \alpha_1 := \frac{\alpha}{m}.$$

*Замечание.* Из-за неравенства тест консервативный, т.е.  $\text{FWER} \ll \alpha$ . Значит не максимально мощный.

$$\begin{aligned} \text{strong FWER} &= \max_{T \subset \{1, \dots, m\}} P(H^{(i)} \text{ отвергается}, i \notin T) \\ &\leq \sum_{i \notin T} P(H^{(i)} \text{ отвергается}) = |\{i : i \notin T\}| \alpha_1 = \alpha. \end{aligned}$$

**Следствие.**  $\text{FWER}$  всегда хуже strong FWER.

**Определение.** Поправка Бонферрони

$$\alpha_1 = \frac{\alpha}{m}.$$

Тест нужно проверять не с  $\alpha_1$ , а с  $\alpha/m$ . Так критерий будет консервативным (иначе — радикальным, что хуже).

**Определение.** Поправка Бонферрони для  $p$ -value:

$$p\text{-value} < \frac{\alpha}{m} \implies \text{отвергаем} \iff mp < \alpha \implies \text{отвергаем}.$$

## 2.2. Stepdown (Holm's algorithm)

Для увеличения мощности применяется «Holm's algorithm»:

1. считаются все  $p$ -value  $p_1, \dots, p_m$ ,
2. упорядочиваются:  $p_{(1)} \leq \dots \leq p_{(m)}$ .
3. если  $mp_{(1)} < \alpha$  то гипотеза отвергается, иначе и все последующие не отвергаются



4. в общем, если

$$p_{(j)} < \frac{\alpha}{m - j + 1}$$

то гипотеза отвергается, иначе и все последующие не отвергаются.

*Замечание.* Сей тест более мощный, потому что не всегда происходит умножение на  $m$ .

*Замечание.* Процедуру сложно повторить, потому что при упорядочивании гипотезы могут перемешиваться.

**Предложение.**  $\text{FWER} \leq \alpha$ .

*Доказательство.* Упорядочим  $p$ -value:  $p_{(1)} \leq \dots \leq p_{(j)} \leq \dots \leq p_{(m)}$ . Пусть  $I = \{i : H_0^{(i)} \text{ верна}\}$ ,  $m_0 = |I|$  — количество верных гипотез,  $j = \min_{k \in I} m_0$  должно «поместиться до конца»:

$$j \leq m - m_0 + 1 \implies \frac{\alpha}{m - j + 1} \leq \frac{\alpha}{m_0}.$$

Значит

$$\begin{aligned} \text{FWER}_I &\leq \mathbb{P} \left( p_{(j)} < \frac{\alpha}{m - j + 1} \right) \leq \mathbb{P} \left( p_{(j)} < \frac{\alpha}{m_0} \right) \leq \mathbb{P} \left( \min_{i \in I} p_i < \frac{\alpha}{m_0} \right) \\ &= \mathbb{P} \left( \bigvee_{i \in I} p_i < \frac{\alpha}{m_0} \right) \leq \sum_{i \in I} \mathbb{P} \left( p_i < \frac{\alpha}{m_0} \right) = m_0 \frac{\alpha}{m_0} = \alpha \end{aligned}$$

□

### 2.2.1. Частный случай

Если все гипотезы и критерии независимы, то возможно точно посчитать FWER:

$$\begin{aligned} \text{FWER}_{\{1, \dots, m\}} &= \mathbb{P} \left( \bigvee_{i=1}^m H_0^{(i)} \text{ отв} \right) = 1 - \mathbb{P} \left( \bigwedge_{i=1}^m H_0^{(i)} \text{ не отв} \right) \\ &= 1 - (1 - \alpha_1)^m = \alpha \implies \alpha_1 = 1 - \sqrt[m]{1 - \alpha} \end{aligned}$$

**Определение.** Поправка Šidák:

$$\alpha_1 = 1 - \sqrt[m]{1 - \alpha}.$$

## 3. ANOVA Post-Hoc Comparison

В случае отвержения гипотезы ANOVA, можно провести дополнительное выборочное тестирование выделенных групп.

### 3.1. Least Significant Difference (LSD)

LSD test — это просто попарный  $t$ -test:

$$t = \frac{\bar{y}_i - \bar{y}_j}{\tilde{s}_{1,\dots,k} \sqrt{1/n_i + 1/n_j}} \sim t(n - k),$$

где  $\tilde{s}_{1,\dots,k}$  — это pooled по  $k$  группам standard deviation.

*Замечание.* Его стоит применять после множественного сравнения лишь к тем группам, важность которых была зафиксирована экспериментатором до проведения множественного сравнения.

*Замечание.* Критерий радикален. Значит, если он не нашел разницу, то и другие критерии тоже не найдут.

*Замечание.* Если групп немного, то можно применить поправку Бонферрони.

### 3.2. Распределение размаха

Сопоставим  $\xi_1, \dots, \xi_n$  i.i.d. с  $\text{cdf}_{\xi_i}(x) = F(x)$  вариационный ряд  $\xi_{(1)}, \dots, \xi_{(n)}$ .

**Определение.** *Размах* есть случайная величина

$$w_n = \xi_{(n)} - \xi_{(1)}$$

с функцией распределения

$$P(w_n < w) = n \int_{\mathbb{R}} (F(x + w) - F(x))^{n-1} dF(x)$$

$(w_n < w \implies w_i < w, P(w_i < w) = F(x + w) - F(x) - n - 1$  штук таких, плюс перебор разных минимумов по  $1 : n$ ).

*Замечание.* В частном случае  $F(x) = \text{cdf}_{N(0, \sigma^2)}(x)$ ,  $\Phi(x) = \text{cdf}_{N(0, 1)}(x)$  рассматривается *стандартизированный размах*

$$P\left(\frac{w_n}{\sigma} < w\right) = n \int_{\mathbb{R}} (\Phi(x + w) - \Phi(x))^{n-1} d\Phi(x).$$

Если  $\sigma$  неизвестна, то с подставленной оценкой  $w/\tilde{s}$  называется *стюдентизированным размахом*.

*Утверждение.* Пусть  $\ell$  — некий параметр и —  $\eta^2$  такая, что  $\ell\eta^2/\sigma^2 \sim \chi^2(\ell)$ ; тогда

$$\frac{w_n}{\eta} \sim q(n, \ell),$$

где  $q$  — распределение стюдентизированного размаха. Это распределение затабулировано.

**Пример** (Проверка выборки на outliers). В нормальной модели,  $H_0$  : нет outliers. Статистика

$$\frac{x_{(n)} - x_{(1)}}{\tilde{s}} \sim q(n, n - 1)$$

потому что, естественно,

$$\frac{(n - 1)\tilde{s}^2}{\sigma^2} \sim \chi^2(n - 1).$$

*Замечание.* Полученный критерий не очень мощный — если  $H_0$  отвергается, то есть аутлаеры присутствуют, то  $x_{(n)} - x_{(1)}$  есть большая величина, но аналогично большой является и  $\tilde{s}$ , поэтому всё значение статистики вырастет незначительно по сравнению со случаем не-отвержения  $H_0$ , когда аутлаеров нет. Мощность же тем больше, чем больше (по модулю) значение статистики в случае, когда требуется отвержение  $H_0$ . Это видно из того, что  $\beta = P_{H_1}(T(\mathbf{x}) \in \mathcal{A}_{\text{крит}})$ ; но мощность, как площадь под графиком плотности  $H_1$  на критическом луче (которые располагаются на хвостах плотности  $H_0$ ), тем больше, чем дальше плотность  $H_1$  от  $H_0$ , т.е. чем больше значения статистики  $T$  в ситуации отвержения  $H_0$ .

Выход заключается в построении более устойчивых оценок для  $\sigma^2$  — например, на основе медианы и абсолютного отклонения.

### 3.3. Tukey's Honest Significance Difference (HSD) Test

**Предположение.** Модель нормальная с дисперсией  $\sigma_0^2$ , и дизайн сбалансирован:  $N(\mu_i, \sigma_0^2)$ ,  $n_0 = n_i \forall i \in 1 : k$ .

По предложению 3.2,

$$t = \frac{\bar{\mathbf{y}}^{(k)} - \bar{\mathbf{y}}^{(1)}}{\sqrt{2\tilde{s}_{1,\dots,k}^2/n_0}} \sim q(k, n - k).$$

Тогда для проверки  $H_0 : \mu_i = \mu_j$  используется HSD статистика

$$t_{ij} = \frac{|\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j|}{\tilde{s}_{1,\dots,k}\sqrt{2/n_0}},$$

а  $p$ -value считаются по  $q(k, n - k)$  (таким образом, смотрят на каждую пару  $(\bar{\mathbf{y}}_i, \bar{\mathbf{y}}_j)$  как на пару из размаха).

**Предложение.** Это точный критерий.

*Доказательство.* Действительно

$$\begin{aligned} \text{FWER}_{\{1:m\}} &= P\left(\bigvee_{i=1}^m H_0^{(i)} \text{ отв}\right) = 1 - P\left(\bigwedge_{i=1}^m H_0^{(i)} \text{ не отв}\right) = 1 - P(t_{ij} < t_\alpha \forall i, j) \\ &= 1 - P\left(\max_{i,j} t_{ij} < t_\alpha\right) = 1 - P(t_{k1} < t_\alpha) = 1 - P(t_{k1} < F^{-1}(1 - \alpha)) = 1 - (1 - \alpha) = \alpha. \end{aligned}$$

□

### 3.4. Другие критерии

**Newman-Keuls** stepdown вариант HSD.

**Tukey-Cramer HSD** вариант Tukey для несбалансированного дизайна

**Dunnett** сравнивает все группы с контрольной

### 3.5. Scheffé's Method

ANOVA гипотезу  $H_0 : \mu_1 = \dots = \mu_k$  можно записать как

$$H_0 : \sum_{i=1}^k c_i \mu_i = 0, \quad \sum_{i=1}^k c_i = 0,$$

где  $\{c_i\}_{i=1}^k$  — «контраст».

**Пример.** Пусть две группы принимают  $k$  лекарств, в том числе — первым номером — плацебо. Сравнить все лекарства с плацебо *одним сравнением* можно сравнив с ним среднее арифметическое всех лекарств, для чего положить  $c_1 = 1, c_2 = \dots c_k = -1/(k-1)$ .

Полученную сумму следует отнормировать и получить статистику

$$t = \frac{\sum_{i=1}^k c_i \bar{y}_i}{\sqrt{D\left(\sum_{i=1}^k c_i \bar{y}_i\right)}} = \frac{\sum_{i=1}^k c_i \bar{y}_i}{\sigma \sqrt{\sum_{i=1}^k c_i^2 / n_i}} \sim N(0, 1).$$

При замене  $\sigma$  на  $\tilde{\sigma}$ , получают, как обычно,  $t \sim t(n-k)$ .

Пусть  $c_1, \dots, c_d, d \leq k-1$  — наборы ортогональных контрастов. Тогда для любого вектора

$$t_j = \frac{\sum_{i=1}^k c_i^{(j)} \bar{y}_i}{\sigma \sqrt{\sum_{i=1}^k (c_i^{(j)})^2 / n_i}}, \quad j \in 1 : d.$$

Линейная комбинация нормальных векторов с ортогональными коэффициентами независима. Следовательно, можно использовать поправки Šidák.

Сколько бы ни захотелось проверить контрастов, хочется уверенности, что  $\text{FWER} \leq \alpha$ . Статистика

$$\frac{t^2}{k-1} \sim F(k-1, n-k).$$

*Замечание.* В HSD можно каждую пару рассматривать как конкретный набор контрастов. Следовательно, метод Шеффе менее мощный по сравнению с HSD (поскольку проверяет все).

### 3.6. Сравнение мощностей

Статистики всех критериев можно свести к одной с разными критическими значениями. Для примера, пусть  $k = 4, n = 20, \alpha = 0.05$ ; тогда

| Критерий                                | Критическое значение |
|---|----------------------|
| LSD                                     | 2.09                 |
| Dunnett                                 | 2.54                 |
| Bonferroni с 3-мя плановыми сравнениями | 2.63                 |
| HSD                                     | 2.8                  |
| Bonferroni с $6 = C_4^2$ сравнениями    | 2.93                 |
| Scheffé                                 | 3.05                 |

Чем больше критическое значение, тем ниже мощность, конечно.

Часть VII.

Регрессионный анализ

# 1. Регрессия

**Определение.** Регрессией  $\eta$  на  $\xi$  называется  $E\{\eta \mid \xi\}$ .

*Замечание.* Таким образом осуществляется предсказание  $\eta$  по  $\xi$  с минимальной среднеквадратичной ошибкой.

**Определение.** Функция регрессии есть  $f(x) = E\{\eta \mid \xi = x\}$ .

*Замечание.*  $f$  находится по МНК для  $K = \{\psi(\xi) : \psi\text{—измеримая}\}$ .

## Виды регрессий

- Нелинейные и линейные ( $K = \{a\xi + b\}$ );
- Парные (предсказывая величину по одной случайной величине) и множественные (по многим).

## 2. Парная линейная регрессия

**Определение.** Пусть  $\xi, \eta \in L^2$ . Парной линейной регрессией  $\eta$  по  $\xi$  называется наилучшее сред-неквадратичное приближение  $h_{\beta_1^*, \beta_2^*}(\xi) = \beta_1^* \xi + \beta_2^*$  в классе линейных по  $\xi$  функций  $K = \mathcal{L} = \{\beta_1 \xi + \beta_2\}$ . Иными словами,

$$h_{\beta_1^*, \beta_2^*}(\xi) = \operatorname{argmin}_{\beta_1, \beta_2} \|\eta - h_{\beta_1, \beta_2}(\xi)\|^2 = \mathbf{E} \{ \eta \mid h_{\beta_1, \beta_2}(\xi) \} = \operatorname{argmin}_{\beta_1, \beta_2} \underbrace{\mathbf{E}(\eta - (\beta_1 \xi + \beta_2))^2}_{\phi(\beta_1, \beta_2)} = \beta_1^* \xi + \beta_2^*.$$

*Замечание.* Найти минимум  $\phi$  можно, как обычно, решив систему  $\partial \phi / \partial \beta_i = 0$ <sup>1</sup>.

*Утверждение.*  $\beta_1^*, \beta_2^*$  таковы, что

$$\frac{h(\xi) - \mathbf{E} \eta}{\sqrt{D \eta}} = \rho \frac{\xi - \mathbf{E} \xi}{\sqrt{D \xi}}.$$

Это уравнение задает линию регрессии. Иными словами,

$$h(\xi) = \underbrace{\rho \frac{\sqrt{D \eta}}{\sqrt{D \xi}}}_{\beta_1^*} \xi + \underbrace{\mathbf{E} \eta - \rho \frac{\sqrt{D \eta}}{\sqrt{D \xi}} \mathbf{E} \xi}_{\beta_2^*}.$$

Отсюда можно получить соотношение между коэффициентом линейной регрессии  $\beta_1^* = k$  (на-клоном регрессионной прямой) и коэффициентом корреляции:

$$k = \rho \frac{\sigma_\eta}{\sigma_\xi}.$$

*Замечание.* Подстановкой проверяется, что

$$\phi(\beta_1^*, \beta_2^*) = \min_{\hat{\eta} \in K} \mathbf{E} (\eta - \hat{\eta})^2 = D \eta (1 - \rho^2),$$

откуда можно найти уже известное выражение для коэффициента корреляции Пирсона

$$\rho^2(\eta, \xi) = 1 - \frac{\phi(\beta_1^*, \beta_2^*)}{D \eta} = 1 - \frac{\min_{\hat{\eta} \in H} \mathbf{E} (\eta - \hat{\eta})^2}{D \eta}, \quad \hat{\eta} := h(\xi).$$

**Определение.** Линейная регрессия *значима*, если  $\beta_1^* \neq 0 \implies \rho \neq 0$ . Значимость регрессии эквивалентна значимости предсказания по ней.

**Определение.** Величина *sum of squares residual* есть

$$\text{SSR} = n \cdot \phi(\beta_1^*, \beta_2^*) = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \hat{y}_i = h_{\beta_1^*, \beta_2^*}(x_i).$$

<sup>1</sup>См. [https://en.wikipedia.org/wiki/Simple\\_linear\\_regression](https://en.wikipedia.org/wiki/Simple_linear_regression)

## 2.1. Модель линейной регрессии

Можно описать выборку как

$$y_i = \beta_1 x_i + \beta_2 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \epsilon_i \perp \epsilon_j.$$

$\sigma^2$  — мешающий параметр, который можно оценить через  $SSR/n$ . Но если  $\epsilon_i \sim N(0, \sigma^2)$ , то

$$\hat{\sigma}^2 = \frac{SSR}{n-2}$$

есть несмещенная оценка  $\sigma^2$ . Значит,

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(n-2).$$

*Замечание.* МНК минимизирует разницу  $y_i - \hat{y}_i$ , что на графике соответствует вертикальным отрезкам, соединяющим  $y_i$  и  $\hat{y}_i = h(x_i)$ . Это не то же, что минимизация перпендикуляров от  $y_i$  на  $h(x)$  — техники метода анализа главных компонент («PCA»).

## 2.2. Доверительные интервалы для $\beta_1$ и $\beta_2$

Как обычно, помимо точечной оценки  $\hat{\beta}_1$  и  $\hat{\beta}_2$ , интересуемся диапазоном значений, которые может принимать оценка с заданной вероятностью. Примем предположение о несмещенности оценки, т.е.  $E\hat{\beta}_i = \beta_i$ . Поскольку в модели  $y_i = \beta_1 x_i + \beta_2 + \epsilon_i$  ошибка  $\epsilon_i \sim N(0, \sigma^2)$  есть случайная величина, оценки  $\hat{\beta}_i$  — тоже становятся случайными величинами:  $\hat{\beta}_i \sim N(\beta_i, D\hat{\beta}_i)$ . В курсе регрессионного анализа доказывается<sup>2</sup>, что

$$D\hat{\beta}_1 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad D\hat{\beta}_2 = \frac{\sigma^2}{n}.$$

Кроме того,

$$SE(\hat{\beta}_1) = \sqrt{D\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{n}s_x} = \frac{\sqrt{\frac{SSR}{n-2}}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad SE(\hat{\beta}_2) = SE(\hat{\beta}_1) \cdot s_x$$

**Предложение.** *Статистика*

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t(n-2).$$

*Доказательство.* Известно,

$$t \sim t(m) \iff t = \frac{\xi}{\sqrt{\eta/m}}, \quad \xi \sim N(0, 1), \quad \eta \sim \chi^2(m).$$

Ясно, что

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim N(0, 1), \quad \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(n-2).$$

<sup>2</sup>См. [https://en.wikipedia.org/wiki/Proofs\\_involving\\_ordinary\\_least\\_squares](https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares)



Тогда

$$\frac{\left( \frac{\hat{\beta}_1 - \beta_1}{\left( \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)} \right)}{\left( \frac{\left( \frac{\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\sigma} \right)}{\sqrt{n-2}} \right)} = \frac{(\hat{\beta}_1 - \beta_1) \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\frac{\sigma}{\sqrt{\sum_{i=1}^n \epsilon_i^2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2} \sum_{i=1}^n \epsilon_i^2}} \sim t(n-2).$$

□

Используя статистику  $t$ , введем доверительные интервалы с  $c_\gamma = \text{cdf}_{t(n-2)}^{-1}((1+\gamma)/2)$ :

$$t \in (-c_\gamma, c_\gamma) \iff \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \in (-c_\gamma, c_\gamma) \iff \beta_1 \in \left( \hat{\beta}_1 - c_\gamma \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + c_\gamma \text{SE}(\hat{\beta}_1) \right).$$

Аналогично, для  $\beta_2$ :

$$\beta_2 \in \left( \hat{\beta}_2 - c_\gamma \text{SE}(\hat{\beta}_2), \hat{\beta}_2 + c_\gamma \text{SE}(\hat{\beta}_2) \right).$$

*Замечание.* На картинке доверительные интервалы изображаются в виде «рукавов» вокруг графика линейной регрессии — т.е. область всевозможных положений прямой при варьировании  $\beta_1, \beta_2$  в заданных интервалах.

**Пример.** Линейная регрессия как предсказательная модель может быть использована неправильно в следующих случаях:

- неправильная модель;
- применение к неоднородным данным (аутлаер или неоднородность);
- хотим построить предсказание в точке, далекой от данных (проблема — большая ошибка);
- не знаем какая модель там, где данных нет.

## 3. Множественная линейная регрессия

### 3.1. Псевдо-обратные матрицы

**Определение.** Матрица  $\mathbf{A}^-$  называется *псевдо-обратной*, если

1. По аналогии с  $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \implies \mathbf{A}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}$  и  $\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}$ , выполняется

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}, \quad \mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-.$$

2. (*Псевдо-обратная по Муру-Пенроузу*) По аналогии с  $\mathbf{A}^{-1} = \mathbf{A}^\top \implies (\mathbf{A}^{-1}\mathbf{A})^\top = \mathbf{A}^\top (\mathbf{A}^{-1})^\top = \mathbf{A}^{-1}\mathbf{A}$ , выполняется

$$\mathbf{A}^-\mathbf{A} = (\mathbf{A}^-\mathbf{A})^\top, \quad \mathbf{A}\mathbf{A}^- = (\mathbf{A}\mathbf{A}^-)^\top.$$

**Свойства**

1. Если столбцы  $\mathbf{A}$  линейно-независимы, то существует  $(\mathbf{A}^\top\mathbf{A})^{-1}$  и

$$\mathbf{A}^- = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top.$$

2. Пусть ищут решение  $\mathbf{X}\mathbf{b} = \mathbf{y}$  относительно  $\mathbf{b}$

- а) Если уравнение не имеет решений, то на  $\mathbf{b} = \mathbf{X}^-\mathbf{y}$  достигается минимум невязки между левой и правой частями:

$$\mathbf{b}^* = \mathbf{X}^-\mathbf{y} = \underset{\mathbf{b}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2.$$

- б) Если решение не единственно, то  $\mathbf{b} = \mathbf{X}^-\mathbf{y}$  есть решение с минимальной нормой.

### 3.2. Проекторы на подпространства

Пусть  $\mathcal{L}_d \subset \mathbb{R}^m$  — линейное подпространство размерности  $d$ , натянутое на  $\{\mathbf{p}_1, \dots, \mathbf{p}_d\}$ ,  $\mathbf{P} = [\mathbf{p}_1 : \dots : \mathbf{p}_d]$ . Тогда проектор на  $\mathcal{L}_d$  будет задан как

$$\operatorname{proj}_{\mathcal{L}_d} = \mathbf{\Pi} = \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top = \mathbf{P}\mathbf{P}^-.$$

Если  $\{\mathbf{p}_i\}_{i=1}^d$  — ортонормированная система, то

$$\mathbf{\Pi} = \mathbf{P}\mathbf{P}^\top = \mathbf{P}\mathbf{P}^-.$$

Кроме того,

$$\operatorname{proj}_{\mathcal{L}_d^\perp} = \mathbf{I}_{m \times m} - \mathbf{P}\mathbf{P}^\top.$$

(т.е., чтобы получить ортогональное пространство к проекции, нужно из исходного вектора вычесть проекцию).

**Свойства**

1.  $\Pi\Pi = \Pi$
2.  $(\mathbf{I} - \Pi)(\mathbf{I} - \Pi) = \mathbf{I} - \Pi$
3.  $\Pi^\top = (\mathbf{P}\mathbf{P}^\top)^\top = \Pi$ .

**3.3. Ordinary and Total Least Squares**

Пусть

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{1n} & \dots & x_{nk} \end{pmatrix}$$

матрица данных с  $n$  индивидами<sup>1</sup> по столбцам, каждый из которых описывается  $k$  признаками;

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

вектор наблюдений<sup>2</sup>;

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}$$

вектор неизвестных коэффициентов.

**OLS** Пусть  $\mathbf{y} \in \mathbb{R}^n$ ,  $\mathbf{X} \in \mathbb{R}^{n \times m}$ ,  $\mathbf{b} \in \mathbb{R}^m$ ,  $\text{rk } \mathbf{X} = m$ . Пусть допускаются ошибки в наблюдениях такие, что  $\mathbb{E}\epsilon_i = 0$ ,  $\epsilon_i \perp \epsilon_j$ ,  $\mathbb{D}\epsilon_i = \sigma^2 \implies \text{cov } \boldsymbol{\epsilon} = \sigma^2 \mathbf{I}$ . Тогда в модели

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon},$$

найти

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\text{argmin}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2 = \underset{\tilde{\mathbf{y}}}{\text{argmin}} \|\tilde{\mathbf{y}} - \mathbf{y}\|^2 = \mathbf{X}^- \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad \tilde{\mathbf{y}} := \mathbf{X}\mathbf{b}.$$

Откуда регрессией будет<sup>3</sup>

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \underbrace{\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{\substack{\text{proj} \\ \text{colspace } \mathbf{X}}} \mathbf{y} = \mathbf{H}\mathbf{y}.$$

Можно посчитать остатки — разницу между наблюдениями и предсказанием по регрессии:

$$\text{residuals} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}.$$

**TLS** Модель допускает ошибки  $\boldsymbol{\Delta}$  также и в  $\mathbf{X}$ ,

$$\mathbf{y} = (\mathbf{X} + \boldsymbol{\Delta})\mathbf{b} + \boldsymbol{\epsilon}$$

(где известны  $\mathbf{y}$ ,  $\tilde{\mathbf{X}} := \mathbf{X} + \boldsymbol{\Delta}$ , а  $\mathbf{X}$  — нет). Найти

$$\underset{\mathbf{b}; \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{b}}{\text{argmin}} \left( \left\| \tilde{\mathbf{X}} - \mathbf{X} \right\|_F^2 + \left\| \tilde{\mathbf{y}} - \mathbf{y} \right\|^2 \right), \quad \|\mathbf{A}\|_F^2 = \sum_{i,j} a_{ij}^2.$$

Дальше рассматривается OLS.

<sup>1</sup>Также «predictors», «regressors», «controlled variables», «explanatory variables», «features», «inputs».

<sup>2</sup>Также «regressands», «response», «explaining variables», «outcome», «experimented variables».

<sup>3</sup> $\mathbf{H}$  — «hat matrix».

### 3.4. Свободный член

Видно, что  $\mathbf{X}\mathbf{b} = \mathbf{y}$  задает СЛАУ, где каждое уравнение — прямая, проходящая через 0. Чтобы иметь возможность описывать случаи не-центрированных данных, пригодны два варианта:

1. Ввести фиктивный столбец из единиц:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times m}, \quad m = k + 1.$$

2. Центрировать признаки.

**Предложение.** Оба способа эквивалентны.

**Теорема** (О делении регрессоров). Пусть  $\mathbf{X}$  матрица данных с признаками («регрессорами») по столбцам,  $\mathbf{X} = [\mathbf{X}_1 : \mathbf{X}_2]$ ,  $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2)^\top$ ,  $\mathbf{M}_1 = \mathbf{I} - \mathbf{H}_1$ ,  $\mathbf{H}_1 = \text{proj}_{\text{colspace } \mathbf{X}_1}$ . Тогда  $\hat{\mathbf{b}}_2$  можно получить как регрессию  $\mathbf{M}_1\mathbf{y}$  на  $\mathbf{M}_1\mathbf{X}_2$ . Остатки регрессии  $\mathbf{M}_1\mathbf{y}$  будут такими же как остатки исходной.

*Доказательство.* Без доказательства. □

Пусть  $\mathbf{b} \in \mathbb{R}^m$ ,  $\hat{\mathbf{b}} = \mathbf{X}^-\mathbf{y} = (\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k)$ . Центрируем  $\mathbf{X}$ , вычитая среднее по каждому столбцу:  $\mathbf{X}^{(c)} \in \mathbb{R}^{n \times k}$ . Центрируем  $\mathbf{y}$ :  $\mathbf{y}^{(c)} \in \mathbb{R}^n$ ; тогда  $\hat{\mathbf{b}}^{(c)} = (\mathbf{X}^{(c)})^-\mathbf{y}^{(c)}$  и по теореме

$$\hat{\mathbf{b}}^{(c)} = \begin{pmatrix} \hat{b}_1^{(c)} \\ \vdots \\ \hat{b}_k^{(c)} \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_k \end{pmatrix}, \quad \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y}^{(c)} - \hat{\mathbf{y}}^{(c)}.$$

**Следствие.**

$$\hat{b}_0 = \bar{y} - \sum_{i=1}^k \hat{b}_i \bar{x}_i.$$

### 3.5. Стандартизованные признаки

Если признаки изначально измерены в разных шкалах, то коэффициенты перед признаками можно интерпретировать как «важность».

**Определение.** Чтобы стандартизировать наблюдения, следует поделить центрированные столбцы на нормы каждого столбца, получится  $\mathbf{X}^{(s)} \in \mathbb{R}^{n \times k}$ ;  $\mathbf{y}^{(s)} = \mathbf{y}^{(c)} / \|\mathbf{y}^{(c)}\|$ . Тогда

$$\hat{\mathbf{b}}^{(s)} = (\mathbf{X}^{(s)})^-\mathbf{y}^{(s)} = \left( (\mathbf{X}^{(s)})^\top \mathbf{X}^{(s)} \right)^{-1} (\mathbf{X}^{(s)})^\top \mathbf{y}^{(s)} = \hat{\boldsymbol{\beta}}, \quad \hat{\beta}_i = \frac{\|\mathbf{x}_i^{(c)}\|}{\|\mathbf{y}^{(c)}\|} \hat{b}_i.$$

Вектор  $\hat{\boldsymbol{\beta}}$  имеет такой вид, потому что по ходу вычислений два раза поделили и один раз умножили на  $\|\mathbf{x}_i^{(c)}\|$ , и умножили на  $\|\mathbf{y}^{(c)}\|$ .

### 3.6. Свойства оценки $\hat{\mathbf{b}}$

1. Несмещенность (по  $\mathbf{E}\boldsymbol{\epsilon} = \mathbf{0}$ ):

$$\mathbf{E}\hat{\mathbf{b}} = \mathbf{E}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{E}\mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{E}(\mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}) = \mathbf{b}.$$

2. Ковариационная матрица:

$$\text{cov } \hat{\mathbf{b}} = \text{cov}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{cov } \boldsymbol{\epsilon} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

3. *Состоятельность*: если оценка несмещенная и состоятельная в среднеквадратичном смысле, то она несмещенная; однако ситуация

$$\text{MSE } \hat{\mathbf{b}} = \mathbf{D} \hat{\mathbf{b}} \xrightarrow{n \rightarrow \infty} 0$$

невозможна в текущей постановке, потому что  $\mathbf{X}$  — фиксированная матрица наблюдений.

**Предложение** (О состоятельности оценки). Пусть  $\mathbf{X}_n \in \mathbb{R}^{n \times m}$  — последовательность (случайных) матриц,  $\boldsymbol{\epsilon}_n \in \mathbb{R}^n$ ; кроме того,

a) Выполняется сильная регулярность независимых переменных:

$$\frac{1}{n} \mathbf{X}_n^T \mathbf{X}_n \xrightarrow{\text{P}} \mathbf{A}, \quad \mathbf{A} \text{ невырожденная}$$

b) Ошибки независимы с регрессорами

$$\frac{1}{n} \mathbf{X}_n^T \boldsymbol{\epsilon}_n \xrightarrow{\text{P}} \mathbf{0}_m.$$

Тогда оценка

$$\hat{\mathbf{b}}_{\text{OLS},n} \xrightarrow{\text{P}} \mathbf{b}$$

является состоятельной.

*Доказательство.*

$$n \left( \mathbf{X}_n^T \mathbf{X}_n \right)^{-1} = \mathbf{A}^{-1} \implies \mathbb{E}(\hat{\mathbf{b}}_n - \mathbf{b})(\hat{\mathbf{b}}_n - \mathbf{b}) = \text{cov } \hat{\mathbf{b}}_n = \sigma^2 (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \xrightarrow{\text{P}} \mathbf{0}.$$

Значит, оценка состоятельна в среднеквадратичном, значит состоятельна.  $\square$

**Предложение** (Об асимптотической нормальности оценки). Пусть  $\{\epsilon_i\}$  i.i.d.,

$$\mathbf{A}_n = \frac{1}{\sigma} \left( \mathbf{X}_n^T \mathbf{X}_n \right)^{-1/2} \mathbf{X}_n^T.$$

$\hat{\mathbf{b}}_{\text{OLS}}$  асимптотически нормальна тогда и только тогда, когда

$$\max_i \{ \mathbf{A}_{ni1}^2, \dots, \mathbf{A}_{nin}^2 \} \xrightarrow{n \rightarrow \infty} 0$$

### 3.7. Свойства $\hat{\mathbf{b}}^{(c)}$ и $\hat{\mathbf{b}}^{(s)}$

1.  $(\mathbf{X}^{(c)})^T \mathbf{X}^{(c)}/n = \mathbf{S}_{\mathbf{xx}}$ ,  $(\mathbf{X}^{(c)})^T \mathbf{y}^{(c)}/n = \mathbf{S}_{\mathbf{xy}}$  суть выборочные ковариационные матрицы (это можно вручную расписать и убедиться); тогда в их терминах

$$\hat{\mathbf{b}}^{(c)} = (\hat{b}_1, \dots, \hat{b}_k)^T = \left( \left( \mathbf{X}^{(c)} \right)^T \mathbf{X}^{(c)}/n \right)^{-1} \left( \mathbf{X}^{(c)} \right)^T \mathbf{y}^{(c)}/n = \mathbf{S}_{\mathbf{xx}}^{-1} \mathbf{S}_{\mathbf{xy}}.$$

**Следствие.** Чем более скоррелированы признаки, тем более пропорциональны столбцы  $\mathbf{S}_{\mathbf{xx}}$  и тем более вырождена  $\mathbf{S}_{\mathbf{xx}}$ , значит «больше»  $\mathbf{S}_{\mathbf{xx}}^{-1}$ , следовательно  $\hat{\mathbf{b}}^{(c)}$  и разность  $\mathbf{y}^{(c)} - \hat{\mathbf{y}}^{(c)} = \mathbf{y}^{(c)} - \mathbf{X}^{(c)} \hat{\mathbf{b}}^{(c)}$ .

2. Ковариационная матрица:

$$\text{cov } \hat{\mathbf{b}}^{(c)} = \sigma^2 \left( \left( \mathbf{X}^{(c)} \right)^\top \mathbf{X}^{(c)} \right)^{-1} = \frac{\sigma^2}{n} \cdot \mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{n \rightarrow \infty} 0$$

3. Аналогично,

$$\hat{\mathbf{b}}^{(s)} = \mathbf{R}_{\mathbf{xx}}^{-1} \mathbf{R}_{\mathbf{xy}}$$

и

$$\text{cov } \hat{\mathbf{b}}^{(s)} = \frac{\sigma^{(s)2}}{n} \mathbf{R}_{\mathbf{xx}}^{-1}, \quad \sigma^{(s)} = \frac{\sigma}{\|\mathbf{y}^{(c)}\|}.$$

### 3.8. Сравнение оценок

По аналогии с одномерным случаем, *наилучшая оценка* — с минимально возможной дисперсией; аналог дисперсии — ковариационная матрица. Порядок вводится следующим образом:

**Определение.**  $\mathbf{A} < \mathbf{B} \iff \mathbf{A} - \mathbf{B}$  отрицательно определена, т.е.

$$\forall \boldsymbol{\gamma} \quad \boldsymbol{\gamma}^\top (\mathbf{A} - \mathbf{B}) \boldsymbol{\gamma} < 0.$$

*Замечание.* Пусть  $\boldsymbol{\gamma}^{(i)} = (0, \dots, \underbrace{1}_i, \dots, 0)^\top$ ; тогда  $a_{ii} < b_{ii}$ .

**Теорема** (Гаусс-Марков). В условиях  $\mathbb{E} \epsilon_i = 0$ ,  $\text{D} \epsilon_i = \sigma^2$ ,  $\epsilon_i \perp \epsilon_j$ ,  $\hat{\mathbf{b}}_{\text{OLS}}$  является «BLUE»: «best linear unbiased estimate».

**Следствие.** Если  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ , то

$$\hat{\mathbf{b}}_{\text{OLS}} = \hat{\mathbf{b}}_{\text{MLE}}.$$

*Доказательство.* MLE оценка есть

$$\begin{aligned} P(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}) &= \frac{1}{(2\pi)^{n/2} \sqrt{\det \sigma^2 \mathbf{I}}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\top \sigma^{-2} \mathbf{I} (\mathbf{y} - \boldsymbol{\mu}) \right\} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{X}\mathbf{b} - \boldsymbol{\mu}\|^2 \right\} \xrightarrow{\mathbf{b}} \max \end{aligned}$$

что аналогично,

$$\|\mathbf{X}\mathbf{b} - \boldsymbol{\mu}\|^2 \xrightarrow{\mathbf{b}} \min.$$

Это же есть постановка задачи OLS. □

### 3.9. Оценка $\sigma^2$

Разложение дисперсии

$$\text{D}\eta = \mathbb{E}(\eta - \mathbb{E}\eta)^2 = \underbrace{\mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)^2}_{\text{D}\mathbb{E}(\eta \mid \xi)} + \mathbb{E}(\eta - \mathbb{E}(\eta \mid \xi))^2$$

на выборочном языке будет иметь вид

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{SSTotal}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{SSRegr}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{SSError}}.$$

*Замечание.* Иногда также пишут

$$SSTotal = SSEffect + SSRResidual,$$

что ведет к неиллюзорной путанице!

Пусть  $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . Тогда, по теореме Cochran,

$$\frac{SST}{\sigma^2} \sim \chi^2(n-1), \quad \frac{SSR}{\sigma^2} \sim \chi^2(\underbrace{m-1}_k), \quad \frac{SSE}{\sigma^2} \sim \chi^2(\underbrace{n-m}_{n-k-1})$$

и  $SSE \perp SSR$ . ОМП оценка (асимптотическая!) для  $\sigma^2$  —  $SSE/n$ ; несмещенной оценкой (с поправкой на размерность) будет

$$\hat{\sigma}^2 = \frac{SSE}{n-m}.$$

*Замечание.* Утверждение про SSE справедливо всегда при нормальном распределении ошибок; про SST и SSR это верно только если  $\mathbf{b}^{(c)} = \mathbf{0}$ . Именно поэтому применяется  $F$ -критерий для проверки значимости регрессии.

### 3.10. Проверка значимости коэффициентов линейной регрессии и доверительных интервалов

**Определение.** Коэффициент  $b_i$  *значим*, если отвергается  $H_0 : b_i = 0$ . Если коэффициент значим, значит признак существенен для регрессии.

Для построения точного критерия, предполагают  $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . Значит, поскольку  $\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \beta + \epsilon)$ ,  $\hat{\mathbf{b}}$  имеет тоже нормальное распределение со средним  $\mathbf{0}$  (по несмещенности), но какой-то ковариационной матрицей:  $\hat{\mathbf{b}} \sim N(\mathbf{0}, \Sigma)$ . Тогда  $E \hat{b}_i = b_i = 0$ ,  $D \hat{b}_i = \sigma_i^2$  и

$$t = \frac{\hat{b}_i - b_i}{\sqrt{D \hat{b}_i}} = \frac{\hat{b}_i}{\sqrt{\sigma^2 ((\mathbf{X}^T \mathbf{X})^{-1})_{ii}}} = \frac{\hat{b}_i}{\sqrt{\sigma^2/n \cdot (\mathbf{S}_{xx}^{-1})_{ii}}} = \sqrt{n} \frac{\hat{b}_i}{\sigma (\mathbf{S}_{xx}^{-1})_{ii}^{1/2}} \sim N(0, 1).$$

Подставляя оценку  $\sigma$ , получают

$$t = \sqrt{n} \frac{\hat{b}_i}{\hat{\sigma} (\mathbf{S}_{xx}^{-1})_{ii}^{1/2}} = \sqrt{n} \frac{\hat{b}_i}{\sqrt{\frac{SSE}{(n-m)} (\mathbf{S}_{xx}^{-1})_{ii}^{1/2}}} = \frac{\frac{\sqrt{n} \hat{b}_i}{\sigma (\mathbf{S}_{xx}^{-1})_{ii}^{1/2}}}{\sqrt{\frac{SSE}{(n-m) \sigma^2}}} = \frac{N(0, 1)}{\sqrt{\frac{\chi^2(n-m)}{n-m}}} \sim t(n-m).$$

#### 3.10.1. Расстояние Махаланобиса

Если на прямой разброс удобно измерять стандартных отклонениях  $\sigma$ , то в многомерном пространстве аналогом такой характеристики является расстояние Махаланобиса.

**Определение.** Пусть  $\mathbf{V}$  — неотрицательно определенная симметричная матрица; тогда *расстояние Махаланобиса* есть

$$r_M^2(\mathbf{x}, \mathbf{y}; \mathbf{V}) = (\mathbf{x} - \mathbf{y})^T \mathbf{V}^{-1} (\mathbf{x} - \mathbf{y}).$$

*Замечание.* Если  $\xi \sim N(\mu, \mathbf{V})$ , то

$$\text{pdf}_{\xi}(\mathbf{x}) = C \cdot \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \mathbf{V}^{-1} (\mathbf{x} - \mu) \right\} = C \cdot \exp \left\{ -\frac{1}{2} r_M^2(\mathbf{x}, \mu; \mathbf{V}) \right\}.$$

Для любых двух  $\mathbf{x}_1, \mathbf{x}_2$  на линии уровня,  $\text{pdf}_{\xi}(\mathbf{x}_1) = \text{pdf}_{\xi}(\mathbf{x}_2)$ . Значит,  $r_M^2(\mathbf{x}_1, \mu; \mathbf{V}) = r_M^2(\mathbf{x}_2, \mu; \mathbf{V})$ , в то время, как Евклидово расстояние не обязано быть одинаковым из-за разной выраженности главных компонент. Однако  $r_M^2(\mathbf{x}, \mathbf{y}; \mathbf{I}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ . Таким образом,  $r_M^2$  — это Евклидово расстояние с поправкой на ковариацию, задаваемую  $\mathbf{V}$ .

*Замечание.* Если  $\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \mathbf{V})$ , то

$$r_M^2(\mathbf{x}, \boldsymbol{\mu}; \mathbf{V}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(m)$$

как сумма квадратов центрированных и нормированных нормальных случайных величин. Кроме того,

$$\boldsymbol{\eta} = \mathbf{V}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N(0, \mathbf{I}) \implies r_M^2(\boldsymbol{\eta}, \mathbf{0}; \mathbf{I}) = \boldsymbol{\eta}^\top \boldsymbol{\eta} \sim \chi^2(m).$$

### 3.10.2. Доверительный эллипсоид

В одномерном случае симметричного распределения, область носителя, где лежит  $\gamma$  всех значений распределения определяется равенством

$$P(|E\xi - x| < \sqrt{D\xi}c_\gamma) = \gamma.$$

Т.е. как такое множество значений, что расстояние их от среднего с учетом стандартного отклонения меньше квантиля уровня  $\gamma$ . В случае оценки среднего  $\mu_0$ , например, получают стандартное

$$P\left(\frac{|\bar{\mathbf{x}} - \mu_0|}{SE} < c_\gamma\right) = P\left(-c_\gamma < \sqrt{n} \frac{\bar{\mathbf{x}} - \mu_0}{\sigma} < c_\gamma\right), \quad \sqrt{n} \frac{\bar{\mathbf{x}} - \mu_0}{\sigma} \sim N(0, 1)$$

так что  $c_\gamma = \text{qnt}_{N(0,1)} \gamma$ .

Аналогично можно нарисовать эллипсоид, в который помещается выборка с точностью  $\gamma$ . Расстояние с учетом ковариации будет задаваться соответственно параметризованным расстоянием Махаланобиса:

$$P(r_M^2(\mathbf{x}, \boldsymbol{\mu}; \mathbf{SD}) < c_\gamma) = \gamma.$$

В случае, если  $\hat{\boldsymbol{\theta}}_n \xrightarrow{d} N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ , по предыдущему,

$$r_M^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) \sim \chi^2(m).$$

Значит,

$$P(r_M^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) < c_\gamma) = \gamma, \quad c_\gamma = \text{qnt}_{\chi^2(m)} \gamma.$$

## 3.11. Значимость регрессии

$H_0 : \mathbf{b}^{(c)} = \mathbf{0}$ . Можно проверить тремя способами:

1. Аналогично парной регрессии: критерий

$$t = r_M^2(\hat{\mathbf{b}}^{(c)}, \mathbf{0}; SE(\hat{\mathbf{b}}^{(c)})) \sim \chi^2(k)$$

а именно,

$$t = \left(\hat{\mathbf{b}}^{(c)}\right)^\top SE^{-1}(\hat{\mathbf{b}}^{(c)}) \hat{\mathbf{b}}^{(c)} = \left(\hat{\mathbf{b}}^{(c)}\right)^\top \left(\frac{\sigma^2}{n} \cdot \mathbf{S}_{\mathbf{xx}}^{-1}\right)^{-1} \hat{\mathbf{b}}^{(c)} = \frac{n \left(\hat{\mathbf{b}}^{(c)}\right)^\top \mathbf{S}_{\mathbf{xx}} \hat{\mathbf{b}}^{(c)}}{\sigma^2}.$$

Неизвестный  $\sigma^2$  следует оценить как

$$s^2 = \frac{SSE}{n - (k + 1)};$$

тогда

$$\frac{n \left(\hat{\mathbf{b}}^{(c)}\right)^\top \mathbf{S}_{\mathbf{xx}} \hat{\mathbf{b}}^{(c)} / k}{s^2} \sim F(k, n - (k + 1)).$$



2. Через ANOVA:

$$t = \frac{SSR/k}{SSE/(n - (k + 1))} \sim F(k, n - (k + 1))$$

*Замечание.* У этой статистики с предыдущей совпадает также и числитель, хотя это не оче-видно.

3. Через коэффициент детерминации регрессии: известно выражение для множественного ко-эффицента корреляции:

$$R^2(\eta, \xi_1, \dots, \xi_k) = \frac{E(\hat{\eta}^* - E\eta)^2}{D\eta}, \quad D\eta = E(\eta - E\eta)^2 = E(\hat{\eta}^* - E\eta)^2 + E(\eta - \hat{\eta}^*)^2;$$

на выборочном языке для множественной линейной регрессии получают

$$\begin{aligned} R^2 &= \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \\ \text{adjusted } R^2 &= 1 - \frac{SSE/(n - (k + 1))}{SST/(n - 1)} \xrightarrow{n \rightarrow \infty} R^2. \end{aligned}$$

*Замечание.* При удалении даже незначимого признака  $R^2$  уменьшится; adjusted  $R^2$  не обя-зательно в силу поправки  $n - (k + 1)$ , действующей как штраф за количество переменных.

Приводя к виду ANOVA-критерия, имеют

$$t = \frac{\frac{SSR}{k}}{\frac{SSE}{n - (k + 1)}} = \frac{\frac{SSR}{k} \frac{SST}{(SST - SSE + SST)SST}}{\frac{SST}{n - (k + 1)}} = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}.$$

### 3.12. О множественном коэффициенте корреляции и саппрессорах

Известно, что  $\rho(\eta, \xi)$  есть косинус угла между  $\eta$  и  $\xi$  в соответствующем пространстве. Аналогично можно думать, что  $R^2$  есть косинус между  $\eta$  и линейным пространством, натянутым на  $\xi_1, \dots, \xi_k$ :

$$R^2 = \cos^2(\eta, \mathcal{L}(\xi_1, \dots, \xi_k)).$$

Возможна ситуация, когда  $\cos^2(\eta, \mathcal{L}(\xi_1, \xi_2)) = 1 = R^2$  — т.е.  $\eta$  лежит в  $\mathcal{L}(\xi_1, \xi_2)$  (и предсказание абсолютно точно), но, тем не менее,  $\text{cov}(\xi_1, \xi_2) \gg 0$  (почти коллинеарны),  $\text{cov}(\xi_1, \eta) = 0$ ,  $0 < \text{cov}(\xi_2, \eta) \ll 1$ .  $\xi_1$  называется «саппрессором» по отношению к  $\xi_2$  (или наоборот). Подробнее, см <https://stats.stackexchange.com/a/73876>.

### 3.13. Взвешенная регрессия (Weighted Least Squares)

Пусть  $\mathbf{W}$  — симметричная, положительно определенная матрица, тогда

$$\hat{\mathbf{b}}_W = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}$$

есть «взвешенная» оценка. При  $\mathbf{W} = \mathbf{I}$ ,  $\hat{\mathbf{b}}_W = \hat{\mathbf{b}}$ , конечно.

Если  $E\epsilon = \mathbf{0}$ , то

$$E\hat{\mathbf{b}}_W = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} E\mathbf{y} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{X} \mathbf{b} = \mathbf{b}$$

и оценка несмещенная.

- Если  $\text{cov } \epsilon = \sigma^2 \mathbf{I}$ , то  $\hat{\mathbf{b}}$  — BLUE и  $\hat{\mathbf{b}}_W$  уже не лучшая.
- Если  $\text{cov } \epsilon = \mathbf{C}$  (то есть шум не белый) то нужно подобрать  $\mathbf{W}$  такую, что  $\hat{\mathbf{b}}_W$  — BLUE. Это делается операцией отбеливания: пусть всё центрированное; тогда  $\mathbf{C}^{-1/2}\epsilon$  — центрированный и нормированный белый шум, и

$$\underbrace{\mathbf{C}^{-1/2}\mathbf{y}}_{\tilde{\mathbf{y}}} = \underbrace{\mathbf{C}^{-1/2}\mathbf{X}}_{\tilde{\mathbf{X}}}\mathbf{b} + \mathbf{C}^{-1/2}\epsilon$$

откуда

$$\text{cov}(\mathbf{C}^{-1/2}\mathbf{y}) = \left(\mathbf{C}^{-1/2}\right)^T \text{cov}(\mathbf{X}\mathbf{b} + \epsilon) \mathbf{C}^{-1/2} = \mathbf{I}.$$

Так как теперь шум белый, следующая оценка будет BLUE:

$$\begin{aligned} \hat{\mathbf{b}} &= \tilde{\mathbf{X}}^{-1}\tilde{\mathbf{y}} = ((\mathbf{C}^{-1/2}\mathbf{X})^T(\mathbf{C}^{-1/2}\mathbf{X}))^{-1} (\mathbf{C}^{-1/2}\mathbf{X})^T \mathbf{C}^{-1/2}\mathbf{y} \\ &= (\mathbf{X}^T (\mathbf{C}^{-1/2})^T \mathbf{C}^{-1/2} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{C}^{-1/2})^T \mathbf{C}^{-1/2}\mathbf{y} \end{aligned}$$

Значит, следует положить  $\mathbf{W} = \mathbf{C}^{-1}$ .

Для  $\mathbf{W}$  итеративный процесс: берем начальное значение, находим коэффициент, оцениваем  $\mathbf{C}$  и т.д.

**Пример.** Стандартный случай — измерения с разной точностью, откуда

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_n^2 \end{pmatrix}.$$

Наблюдениям, таким образом, придается разный вес — чем меньше точность наблюдения, тем больше  $\sigma_i^2$  и меньший, соответственно, вес.

*Замечание.*  $\mathbf{W}$  можно также назначить и руками.

### 3.14. Гребневая (Ridge) регрессия

Чтобы бороться с вырожденностью  $\mathbf{R}_{xx}$  в оценке  $\hat{\beta} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$  рассматривают

$$\hat{\beta} = (\mathbf{R}_{xx} + \lambda \mathbf{I})^{-1}\mathbf{R}_{xy}.$$

Получается смещенная оценка, но с меньшей дисперсией. Для поиска  $\lambda$  используют кросс-валидацию.

### 3.15. Анализ оценок коэффициентов

Для анализа оценок коэффициентов можно посмотреть на попарные срезы доверительного эллипсоида. Для пары  $\hat{\beta}_i, \hat{\beta}_j$  его можно нарисовать (самостоятельно), в качестве центра взяв в качестве центра точку  $(\hat{\beta}_i, \hat{\beta}_j)^T$ , в качестве наклона главной оси и вытянутости — величину  $\text{corr}(\hat{\beta}_i, \hat{\beta}_j)$ . Если центр достаточно далек от 0, то линии уровня не должны пересекать оси, потому что гипотеза  $\beta_i = \beta_j = 0$  отвергается.

- Чем дальше от начала координат центр эллипсоида, тем больше значимость признаков.
- Чем больше корреляция тем менее адекватно центр отражает ситуацию.
- Возможны два случая: когда эллипсоид сонаправлен или перпендикулярен прямым  $y = \pm x$ ; в первом случае («хорошем») коэффициенты значимы совокупности (даже если один близок к 0, то второй вполне далек и наоборот), во втором оба коэффициента могут быть одновременно как значимы, так и нет (и, значит, и сильно, и слабо влиять на результат).

## 3.15.1. Корреляция между оценками коэффициентов и избыточность

При возрастании корреляции признаков:

- дисперсия оценок коэффициентов стремится к бесконечности;
- становится сложно оценить вклад каждого признака в регрессию.

**Пример.** Пусть  $k = 2$ ,  $\eta = b_0 + b_1\xi_1 + b_2\xi_2$ . Пусть также матрица корреляций есть

$$\mathbf{R}_{\mathbf{xx}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Тогда

$$\text{cov } \hat{\beta} = \frac{\sigma^{(s)2}}{n} \mathbf{R}_{\mathbf{xx}}^{-1} = \frac{\sigma^{(s)2}}{n} \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Значит,  $D\hat{\beta}_i \xrightarrow{\rho \rightarrow 1} \infty$ .

С этой проблемой можно бороться, удаляя подходящие признаки из анализа<sup>4</sup> по следующим критериям:

1. Множественный коэффициент корреляции

$$R^2(\xi_i; \{\xi_j, j \neq i\}).$$

Чем он больше, тем скорее  $i$ -й признак нужно удалить.

2. Допустимость  $i$ -го признака:

$$\text{tolerance}_i = 1 - R^2(\xi_i; \{\xi_j, j \neq i\}).$$

Чем он меньше, тем скорее  $i$ -й признак нужно удалить. Помимо предыдущего соотношения справедливо

$$D\hat{b}_i = \frac{\sigma^2}{\sum_{\ell=1}^n (x_{\ell} - \bar{\mathbf{x}}_i)^2} \frac{1}{\text{tolerance}_i}, \quad \frac{1}{\text{tolerance}_i} - \text{Variance Inflation Factor},$$

так что при маленькой допустимости дисперсия велика.

3. Частные корреляции

$$\rho(\xi_i, \eta \mid \{\xi_j, j \neq i\}) = \rho(\xi_i - \hat{\xi}_i, \eta - \hat{\eta})$$

Чем  $i$ -я частная корреляция больше, тем больше вклад признака в регрессию (тем менее он предпочтителен для удаления).

4. Полу-частные корреляции

$$\rho(\xi_i - \hat{\xi}_i, \eta).$$

Пусть  $\mathbf{b} = (b_0, \dots, b_{k-r}, \underbrace{b_{k-r+1}, \dots, b_k}_{r \text{ штук}})^T$ . Если  $H_0 : \mathbf{b}_{k-r+1,k} = \mathbf{0}$  не отвергается, значит последние  $r$  признаков не влияют на модель и следует выбрать более простую модель — без этих коэффициентов. Можно использовать расстояние Махаланобиса до 0 в метрике  $\text{cov}(\mathbf{b}_{k-r+1,k})$ :

$$\begin{aligned} t &= r_M^2(\hat{\mathbf{b}}_{k-r+1,k}, \mathbf{0}; \text{cov}(\mathbf{b}_{k-r+1,k})) \sim \chi^2(r) \\ &= \hat{\mathbf{b}}_{k-r+1,k}^T ((\mathbf{X}^T \mathbf{X})^{-1})_{(\text{IV})} \hat{\mathbf{b}}_{k-r+1,k} / \sigma^2, \end{aligned}$$

<sup>4</sup>Нет признака — нет проблемы.

где  $((\mathbf{X}^\top \mathbf{X})^{-1})_{(IV)}$  — IV квадрант  $(\mathbf{X}^\top \mathbf{X})^{-1}$ . Если  $\sigma^2$  неизвестна, то

$$\begin{aligned} t &= \frac{\hat{\mathbf{b}}_{k-r+1,k}^\top ((\mathbf{X}^\top \mathbf{X})^{-1})_{(IV)} \hat{\mathbf{b}}_{k-r+1,k}/r}{\hat{\sigma}^2} \sim F(r, n - (k + 1)) \\ &= \frac{(R_{1,k}^2 - R_{1,k-r}^2)/r}{(1 - R_{1,k}^2)/(n - m)}. \end{aligned}$$

Выбор оптимального набора признаков можно производить автоматически, по одному добавляя признаки («Forward stepwise regression») или убирая их («Backward»). Пусть вариант Forward. На шаге  $i$  добавляется тот признак, что максимизирует

$$R_{1,i+1}^2 - R_{1,i}^2;$$

остановиться следует, когда  $|R_{1,i+1}^2 - R_{1,i}^2|$  достаточно мало.  $H_0 : R_{1,i+1}^2 - R_{1,i}^2 = 0$ , т.е.  $b_{i+1} = 0$  перед добавленным признаком.

$$t = \frac{\hat{b}_i}{\text{SE}(\hat{b}_i)} \sim t(n - m).$$

Тогда  $k = i + 1$ ,  $r = 1$  и статистика будет иметь вид

$$t = \frac{(R_{1,i+1}^2 - R_{1,i}^2)}{(1 - R_{1,i+1}^2)/(n - (i + 2))} \sim F(1, n - (i + 2)).$$

По сути, это есть перемасштабированное значение разницы  $R_{1,i+1}^2 - R_{1,i}^2$ .

*Замечание.* Однако признак выбран «лучший» (а не случайный), значит распределение не F.

- Полное решение задачи — выбрать  $\ell$  признаков из  $k$  перебором.
- Жадный алгоритм — последовательно выбирать наиболее подходящие признаки.

*Замечание.* Если большое количество заполнить средними, то искусственно уменьшится ширина доверительных интервалов.

## 3.16. Анализ аутлаеров

### 3.16.1. Matrix plot

Аутлаеров можно найти «на глаз» при помощи стандартного matrix plot данных.

### 3.16.2. Deleted residuals

можно применить технику кросс-валидации: удалить признак, построить модель, сравнить. Если индивид является аутлаером, то наблюдение  $y_i$  на нём «перетягивает» на себя регрессионную прямую. Тогда явно «большой» будет разница

$$r_i^{(i)} = \hat{y}_i^{(i)} - y_i$$

между  $\hat{y}_i^{(i)}$  — значением регрессии на  $i$ -м индивиде без этого индивида и  $y_i$  — наблюдении на  $i$ -м индивиде.  $r_i^{(i)}$  будет «большой» также по сравнению с  $r_i = \hat{y}_i - y_i$ . Напротив, если  $i$ -й индивид аутлаером не является, то будет справедливо приближенное равенство  $r_i^{(i)} \approx r_i$ , так что графиком  $(r_i, r_i^{(i)})$  будет прямая. Deleted residuals всегда не меньше residuals, поэтому прямая  $y = x$  не получится.

## 3.16.3. Studentized residuals

Справедливо

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \implies (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

откуда

$$\text{cov}(\mathbf{y} - \hat{\mathbf{y}}) = \text{cov}(\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})^\top \text{cov} \mathbf{y} (\mathbf{I} - \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{H})$$

потому что  $\mathbf{I} - \mathbf{H}$  — матрица проектора. Тогда,

$$D(y_i - \hat{y}_i) = Dr_i = \sigma^2(1 - h_{ii}).$$

Как следствие,  $D\epsilon_i = \sigma^2 \geq Dr_i$ .

**Определение.**  $h_{ii}$  — рычаг<sup>5</sup>.

Чем больше  $i$ -й рычаг, тем больше ошибка на  $i$ -м индивиде.

**Определение.** Стандартизированные остатки:

$$\frac{r_i}{\sqrt{Dr_i}} = \frac{r_i}{\sigma\sqrt{1 - h_{ii}}}.$$

Можно рассмотреть  $\hat{\sigma}^{(i)}$  — оценку дисперсии без  $i$ -го индивида; тогда, при нормально распределенных ошибках наблюдения,

$$\frac{r_i}{\hat{\sigma}^{(i)}\sqrt{1 - h_{ii}}} \sim t(n - m - 1)$$

(«−1» потому что меньше на одного индивида).

*Замечание.* Полученную величину можно сравнивать со «средним» значением рычага

$$\frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{1}{n} \text{tr} \mathbf{H} = \frac{1}{n} \text{rk} \mathbf{H} = \frac{k + 1}{n}.$$

(как след идемпотентной матрицы, равный её рангу<sup>6</sup>: след есть сумма собственных чисел, однако у идемпотента два возможных собственных числа: 0 и 1, а кратность 1 в точности равна рангу).

## 3.16.4. Расстояние по Куку

Пусть  $\hat{\mathbf{b}}^{(i)}$  — оценка без  $i$ -го индивида. Если расстояние между  $\hat{\mathbf{b}}^{(i)}$  и  $\hat{\mathbf{b}}$  «большое», то  $i$ -й индивид есть аутлаер:

$$r_M^2(\hat{\mathbf{b}}, \hat{\mathbf{b}}^{(i)}; \text{cov} \hat{\mathbf{b}}) = (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})^\top \text{cov}^{-1}(\hat{\mathbf{b}}) (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)}) = \frac{1}{\sigma^2} (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})^\top \mathbf{X}^\top \mathbf{X} (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})$$

так что расстояние по Куку определяется как

$$\frac{(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})^\top \mathbf{X}^\top \mathbf{X} (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)}) / m}{\hat{\sigma}^2}.$$

Можно сравнить с расстоянием Махаланобиса в пространстве независимых признаков: если  $x_i$  —  $i$ -й индивид,  $\bar{\mathbf{x}}$  — вектор средних, то аутлаером можно назвать индивида, для которого велико

$$r_M^2(x_i, \bar{\mathbf{x}}; \mathbf{S}_{\mathbf{xx}}).$$

*Замечание.* Если индивид не аутлаер по Куку, но аутлаер по Махаланобису, то велика дисперсия и  $\mathbf{S}_{\mathbf{xx}}$  оценивается неправильно.

|                            | Аутлаер по Куку  | Не аутлаер по Куку                              |
|----------------------------|--|---|
| Аутлаер по Махаланобису    | Далеко от линии регрессии  | Далеко от $\bar{\mathbf{x}}$ на линии регрессии |
| Не аутлаер по Махаланобису | Близко к $\bar{\mathbf{x}}$ по одной координате и далеко по другой | Близко к $\bar{\mathbf{x}}$                     |

<sup>5</sup> «Leverage».

<sup>6</sup> <http://math.stackexchange.com/a/101515>

### 3.17. Проверка правильности и выбор модели

- Если известно, что ошибки нормально распределены (например, в случае измерений прибора), то если остатки не имеют нормального распределения, то модель не является правильной.
- Если исходные данные имеют нелинейную зависимость, то и расположение остатков по линейной регрессии на графике будет отражать характер этой зависимости.
- Модель с наименьшим количеством параметров при прочих равных является предпочтительной, поэтому если заранее известно, что среднее 0, то свободный член из модели лучше удалить.

*Замечание.* В случае нормального распределения для всех случайных величин справедливо

$$\eta = E(\eta \mid \xi_1, \dots, \xi_i) + (\eta - E(\eta \mid \xi_1, \dots, \xi_i)).$$

Поэтому получается ортогональность остатков регрессии (регрессия линейна в силу нормальности). Значит все модели «верны» и можно среди них выбрать наилучшую.

### 3.18. Доверительные интервалы

Пусть  $(1, \mathbf{z})^T \in \mathbb{R}^{k+1}$ ; тогда настоящее предсказание есть

$$\bar{y} = \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}^T \mathbf{b},$$

а его оценка<sup>7</sup>

$$\hat{y} = \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}^T \hat{\mathbf{b}}.$$

Эта оценка несмещенная,  $E\hat{y} = \bar{y}$ . Можно показать, что её дисперсия есть

$$D\hat{y} = \sigma^2 \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}^T (\mathbf{X}^T \mathbf{X})^{-1} \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} (\mathbf{z} - \bar{\mathbf{x}})^T \mathbf{S}_{\mathbf{xy}}^{-1} (\mathbf{z} - \bar{\mathbf{x}}),$$

частный случай чего выписывался в случае парной регрессии

$$D\hat{y} = \frac{\sigma^2}{n} + \frac{\sigma^2(x - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})}.$$

Доверительным интервалом (оценки того, какое среднее значение будет на выходе, если на входе  $\mathbf{z}$ ) будет

$$\bar{y} \pm c_\gamma SE = \bar{y} \pm c_\gamma \sqrt{D\hat{y}} = \bar{y} \pm c_\gamma \hat{\sigma} \sqrt{\begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}^T (\mathbf{X}^T \mathbf{X})^{-1} \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}}, \quad c_\gamma \sim t(n-m)$$

В данной модели  $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$  — значение вообще, а  $\mathbf{X}\mathbf{b}$  — среднее. Можно, поэтому, также построить ДИ для значения вообще:

$$\bar{y} \pm c_\gamma \hat{\sigma} \sqrt{\underbrace{D\boldsymbol{\epsilon}}_1 + \frac{1}{n} + \frac{1}{n} (\mathbf{z} - \bar{\mathbf{x}})^T \mathbf{S}_{\mathbf{xy}}^{-1} (\mathbf{z} - \bar{\mathbf{x}})}, \quad c_\gamma \sim t(n-m).$$

<sup>7</sup>Mean prediction.

### 3.19. Сведение нелинейной модели к линейной

Существует три базовых модели, в которых функция регрессии линейная:

1.  $\eta, \xi_1, \dots, \xi_m$  нормальные.
2.  $y = \mathbf{Xb} + \epsilon$ ,  $E\epsilon = \mathbf{0}$ ,  $\text{cov } \epsilon = \sigma^2 \mathbf{I}$ .
3.  $y = a\xi + b + \epsilon$ ,  $\xi$  принимает всего два значения (возможно, как качественный признак).

Пусть

$$\eta = \phi(\xi_1, \dots, \xi_k) + \epsilon$$

и  $\phi$  — нелинейная функция.

- $\phi$  — многочлен. Можно свести к линейной, добавляя признаки  $\xi, \xi^2, \dots$  и для этих признаков строить модель.
- $\xi$  — качественный признак. Можно ввести  $k - 1$  штук<sup>8</sup> фиктивных признаков со значениями  $\{0, 1\}$  и для них строить модель.  $A_1, \dots, A_k$  — градации  $\phi$ .

### 3.20. Другие странные замечания

- $\eta = \phi(\xi) + \epsilon$ ,  $E\epsilon = 0$ ,  $\epsilon \perp \xi$ . Если  $\phi$  не линейная, то в  $\epsilon$  войдет кусочек  $\xi$  и независимости не будет.
- Остатки всегда ортогональны т.к. проектор  $\implies$  график — горизонтальная прямая всегда
- Графике  $\hat{y}_i$  против  $\hat{y}_i - y_i$  может быть наклонной прямой в случае pairwise MD deletion (и ковариационная матрица не соответствует данным).

---

<sup>8</sup>При добавлении вектора из единиц к  $k$  признакам получается вырожденная матрица.

## А. Свойства условного математического ожидания

1.  $E\{a\xi + b\theta \mid \eta\} = aE\{\xi \mid \eta\} + bE\{\theta \mid \eta\}.$
2.  $EE\{\eta \mid \xi\} = E\eta.$
3.  $\xi \perp\!\!\!\perp \eta \implies E\{\eta \mid \xi\} = E\eta.$
4.  $\eta = f(\xi) \implies E\{\eta \mid \xi\} = E\{f(\xi) \mid \xi\} = f(\xi).$
5.  $E(\eta f(\xi) \mid f(\xi)) = f(\xi)E\{\eta \mid \xi\}.$
6.  $(\xi, \eta)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies E(\eta \mid \xi) = a\xi + b.$

*Замечание (Важное).* Таким образом, если выборка нормальная, то зависимость линейная всегда.

7.  $\operatorname{argmin}_{\hat{\eta} \in K = \{\phi(\xi)\}} E(\eta - \hat{\eta})^2 = E\{\eta \mid \xi\} = \hat{\eta}^*.$