Оглавление

1. Формы представления данных и цели анализа данных

При изучении реальных случайных явлений опытные данные упорядочиваются, представляются в компактной, наглядной или функциональной форме. Целью анализа данных является выявление и описание закономерностей, на основе которых можно проводить прогнозирование и проектирование.

2. Характеристики положения данных

Характеристикой положения данных называется число, характеризующая местоположение данных на числовой оси. Примеры:

1. Выборочное среднее: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Полусумма квартилей: $t_q = \frac{z_{1/4}+z_{3/4}}{2}$
3. Полусумма экстремальных выборочных элементов: $t_R = \frac{x_{\min}+x_{\max}}{2}$
4. Выборочная медиана: $\mathrm{med} = \begin{cases} x_{(l+1)}, n = 2l+1 \\ \frac{x_{(l)}+x_{(l+1)}}{2}, n = 2l \end{cases}$

3. Характеристики рассеяния данных

Характеристикой рассеяния данных называется число, характеризующее степень разброса данных. Примеры:

1. Выборочное среднее квадратическое отклонение: $s = \sqrt{s^2}$, где $s^2 = m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (выборочная дисперсия)
2. Выборочное среднее абсолютное отклонение: $d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \text{med}|$
3. Интерквартильная широта: $q = z_{3/4} - z_{1/4}$
4. Размах: $R = x_{\max} - x_{\min}$

4. Оптимизационный подход к построению х-к положения и рассеивания

Рассмотрим выборку $x_1, \dots, x_n \in R$, пусть $z = z(x_1, \dots, x_n)$ – хар-ка положения.

Рассмотрим $J_p(z, x_1, \dots, x_n) = (\sum_{i=1}^{n}|x_k - z|^p)^{\frac{1}{p}}, p \geq 1$. Тогда $J_1 = \sum_{i=1}^{n}|x_k - z|$, $J_2 = \sqrt{\sum_{i=1}^{n}|x_k - z|^2}$, $J_\infty = \max\limits_{1\leq i\leq n}|x_i - z|$. $z_1 = argmin\, J_1(z) = \text{med}\, x$(МНМ), $z_2 = argmin\, J_2(z) = \bar{x}$(МНК), $z_\infty = argmin\, J_\infty = z_R$, $d = J_1(\text{med}\, x, x_1, \dots x_n), \sigma = J_2(\bar{x}, x_1, \dots, x_n), R = J_\infty(z_R, x_1, \dots, x_n)$

5. Характеристики взаимосвязи данных

Характеристикой взаимосвязи данных называется число, характеризующее зависимость между данными. Примеры:

1. Выборочный коэффициент корреляции: $r = \frac{K_{XY}}{s_X s_Y}$, где $K_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
2. Выборочный квадратный коэффициент корреляции: $r_Q = \frac{(n_1+n_3)-(n_2+n_4)}{n}$, где n_1, n_2, n_3, n_4 – количества точек с координатами (x_i, y_i) попавшими в I, II, III, IV квадранты декартовой системы координат с осями $x' = x - \text{med}\, x$, $y' = y - \text{med}\, y$
3. Выборочный коэффициент ранговой корреляции

6. Характеристики экстремальных значений данных

1. Правило 3-х сигм: пусть $x \sim N(x, \mu, \sigma)$. Если $|x_k - \bar{x}| > 3\sigma$, то x_k - выброс.
2. Критерий Граббса: сравнивается $\frac{x_{(n)}-\bar{x}}{\sigma}$ и λ-порог, больше которого отбрасываем. Как правило $\lambda \in [2.5,3]$.
3. Боксплот Тьюки

7. Графическое представление данных – боксплот Тьюки

Боксплот (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей: в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. [3]

Границами ящика служат LQ и UQ, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов): X_1 и X_2 (1).

$$X_1 = LQ - \frac{3}{2}(UQ - LQ), X_2 = UQ + \frac{3}{2}(UQ - LQ)$$

8. Характеристики распределений данных: «ядерные» оценки плотности

Пусть $(x_1, ..., x_n)$ - выборка полученная по распределению с некоторой плотностью f, требуется оценить функцию f. *Ядерным оценщиком плотности* называется [4]

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{21}$$

где K - т.н. *ядро* (некоторая неотрицательная функция), $h > 0$ - сглаживающий параметр, именуемый *шириной полосы*.

Как правило используется нормальное (или гауссово) ядро, в силу его удобных математических свойств:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{22}$$

В случае если используется гауссово ядро и оцениваемая плотность является гауссовой, оптимальный выбор для h определяется т.н. *правилом Сильвермана* [4]:

$$h_n = \left(\frac{4s_n^5}{3n}\right)^{\frac{1}{5}} \approx 1.06 s_n n^{-\frac{1}{5}} \tag{23}$$

где s_n - выборочное среднеквадратичное отклонение (корень из выборочной дисперсии)

9. Что такое точечная оценка?

Пусть имеется выборка $x_1, \dots, x_n$ из некоторого семейства (параметризуемого θ) абсолютно непрерывных распределений с плотностью $f(x, \theta)$. Точечной оценкой параметра называется функция от элементов выборки $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$ приближенно равная θ (в некотором смысле).

10. Что такое статистика?

Статистикой называется борелевская функция от элементов выборки (возможно не связанная с параметром θ)

11. Какая оценка называется состоятельной, несмещенной, эффективной, робастной

- Оценка $\hat{\theta}_n$ называется состоятельной оценкой θ, если $\hat{\theta}_n \xrightarrow[P]{} \theta$.
- Оценка $\hat{\theta}_n$ называется несмещенной оценкой θ, если $\mathbf{M}\hat{\theta}_n = \theta$, это требование сужает класс состоятельных оценок т.к. если $\hat{\theta}_n$ – состоятельная оценка, то несложно показать, что $\frac{n-a}{n-b}\hat{\theta}_n$ при фиксированных a, b также состоятельная. Несмещенная оценка при этом преобразовании станет смещенной т.к. $\mathbf{M}\frac{n-a}{n-b}\hat{\theta}_n = \frac{n-a}{n-b}\mathbf{M}\hat{\theta}_n = \frac{n-a}{n-b}\theta$
- Эффективной оценкой $\hat{\theta}_{\text{эфф}}$ параметра θ для рассматриваемого распределения называется оценка из класса T – состоятельных и несмещенных оценок, имеющая минимальную дисперсию $\mathbf{D}\hat{\theta}_{\text{эфф}} = \min_T \mathbf{D}\hat{\theta}_n$
- Важно знать, насколько устойчива эффективная оценка к отклонениям от принятых допущений о генеральном распределении (по которому как мы предполагаем была получена выборка). Это свойство называется робастностью. Существуют различные отклонения от принятых гипотез о генеральном распределении. Рассмотрим т.н. “схему засорения” нормальной генеральной совокупности

$$f(x, \theta) = (1 - \varepsilon)N(x, \theta, \sigma_1) + \varepsilon N(x, \theta, \sigma_2), 0 < \varepsilon < 1, \sigma_1 \ll \sigma_2$$

$$N(x, \theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

Различные оценки θ для такой плотности распределения могут теперь сравниваться по дисперсии. Та, чья дисперсия практически не меняется при увеличении σ_2 – признается робастной.

12. Какая из двух оценок считается более эффективной?

Из двух оценок $\hat{\theta}_{1n}$ и $\hat{\theta}_{2n}$ одного параметра, одного распределения, одного класса T состоятельных и несмещенных оценок более эффективной считается та, дисперсия которой меньше.

13. Что такое эффективность, относительная эффективность, асимптотическая эффективность оценки?

Относительной эффективностью одной оценки по другой называется отношение $\frac{\mathbf{D}\hat{\theta}_{1n}}{\mathbf{D}\hat{\theta}_{2n}}$ (если $\mathbf{D}\hat{\theta}_{1n} < \mathbf{D}\hat{\theta}_{2n}$), эффективностью оценки $\hat{\theta}_{1n}$ называется $\text{eff}\,\hat{\theta}_{1n} = \frac{\inf_T \mathbf{D}\hat{\theta}_n}{\mathbf{D}\hat{\theta}_{1n}}$ (здесь T – класс состоятельных и несмещенных оценок)

Оценка $\hat{\theta}_n$ параметра θ называется асимптотически эффективной в классе T – состоятельных оценок (не обязательно несмещенных) если существует предел $\lim_{n\to\infty} \text{eff}\,\hat{\theta}_n = 1$

14. Что такое процедура складного ножа?

Общая процедура уменьшения смещения оценки. Пусть $\mathbf{M}\hat{\theta}_n = \theta + \frac{b}{n} + \frac{c}{n^2}$ и $\hat{\theta}_{k,n-1}$ – та же статистика, но построенная по выборке $(x_i)_{i=1,i\neq k}^n$ (исключили k-ый элемент), вычислим выборочное среднее $\hat{\theta}_{*,n-1} = \frac{1}{n}\sum_{k=1}^n \hat{\theta}_{k,n-1}$. Оценкой по методу "складного ножа" называется статистика $\hat{\theta}_n^* = n\hat{\theta}_n - (n-1)\hat{\theta}_{*,n-1}$, вычислим её математическое ожидание для сравнения:

$$\mathbf{M}\hat{\theta}_n^* = n\mathbf{M}\hat{\theta}_n - \frac{n-1}{n}\sum_{k=1}^n \mathbf{M}\hat{\theta}_{k,n-1} = n\theta + b + \frac{c}{n} - \frac{n-1}{n}n\left(\theta + \frac{b}{n-1} + \frac{c}{(n-1)^2}\right) =$$

$$= n\theta + b + \frac{c}{n} - (n-1)\left(\theta + \frac{b}{n-1} + \frac{c}{(n-1)^2}\right) = \theta + \frac{c}{n} - \frac{c}{n-1} = \theta - \frac{c}{n(n-1)}$$

15. Приведите примеры состоятельных оценок м.о. нормального распределения

$\bar{x}$

т.к. для нормального распределения существует м. о. и дисперсия (применяем теорему Чебышева)

$\text{med} = Mx$

т.к. в окрестности $\xi_{1/2}$ плотность нормального распределения непрерывна вместе со своей производной. В этом случае выборочная медиана как квантиль порядка ½ - состоятельная оценка генеральной медианы, а для нормального распределения м.о. совпадает с медианой.

16. Приведите примеры состоятельных оценок м.о. распределения Лапласа

$\bar{x}$

т.к. для распределения Лапласа существует м. о. и дисперсия (применяем теорему Чебышева)

$\text{med} = Mx$

как о.м.п. β равного мат. ожиданию

17. Приведите примеры состоятельных оценок м.о. равномерного распределения

$\bar{x}$

т.к. для равномерного распределения существует м. о. и дисперсия (применяем теорему Чебышева)

$$\text{med} = Na$$

т.к. в окрестности $\xi_{1/2}$ плотность равномерного распределения непрерывна вместе со своей производной. В этом случае выборочная медиана как квантиль порядка ½ - состоятельная оценка генеральной медианы, а для равномерного распределения м.о. совпадает с медианой.

$$t_R$$

как оценка максимального правдоподобия

18. Приведите примеры состоятельных оценок центра симметрии распределения Коши

Численными методами получается из выражения

$$2\sum_{i=1}^{n}\frac{x_i-\hat{\theta}_{\text{мп}}}{1+\left(x_i-\hat{\theta}_{\text{мп}}\right)^2}=0$$

это о.м.п. и потому состоятельна.

20. Приведите примеры состоятельных оценок дисперсии нормального распределения

$$s^2 \qquad (a)$$

состоятельная т.к. для нормального распределения состоятельны первый и второй выборочные моменты, дисперсия представима через первый и второй генеральные моменты, а выборочная дисперсия через аналогичные выборочные моменты

$$s^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

состоятельная т.к. $s^{*2}=\frac{n-1}{n}s^2$ и $\frac{n-1}{n}\to 1$, а s^2 состоятельная

22. Приведите примеры состоятельных оценок дисперсии распределения Лапласа

$$2d^2$$

через о.м.п., $\alpha = d^{-1} = \frac{n}{\sum_{i=1}^{n}|x_i-\text{med}|}$ (параметр масштаба), а дисперсия $\frac{2}{\alpha^2}$

26. Приведите примеры несмещенных оценок дисперсии нормального распределения.

$$s^{*2}$$

т.к. s^2 смещенная

27. Назовите состоятельные оценки начальных моментов распределений

Выборочный начальный момент порядка l:

$$a_l=\frac{1}{n}\sum_{i=1}^{n}x_i^l$$

для распределений, у которых есть α_l и α_{2l}

28. Назовите состоятельные оценки центральных моментов распределений

Выборочный центральный момент порядка l:

$$m_l=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^l$$

29. Назовите состоятельные оценки генеральных квантилей распределений

Выборочная квантиль порядка p:

$$z_p = \begin{cases} x_{([np]+1)}, np \text{ дробное} \\ x_{(np)}, np \text{ целое} \end{cases}$$

30. Что такое неравенство Рао-Крамера? В чем состоит его смысл?

(разброс выборки будет не)

Функцию $I(\theta) = \mathbf{M}\left(\frac{\partial \ln f}{\partial \theta}\right)^2 = \int_{-\infty}^{+\infty}\left[\frac{\partial \ln f}{\partial \theta}\right]^2 f(x,\theta)dx$ будем называть информацией Фишера. В свою очередь $\mathbf{D}\hat{\theta}_n \geq \frac{1}{I(\theta)}$ – неравенство Рао-Крамера, смысл которого заключается в том, что существует нижняя граница дисперсии несмещенной оценки.

31. При каком условии достигается равенство в неравенстве Рао-Крамера? Приведите примеры

Как и в неравенстве Коши-Буняковского (т.к. это единственное неравенство в доказательстве) т.е. если $\frac{\partial \ln L}{\partial \theta} = k(\theta)\left(\hat{\theta}_n - \theta\right)$. Если для заданной функции L удается получить такое представление, то оценка $\hat{\theta}_n$ будет эффективной т.е. $\hat{\theta}_n = \theta_{\text{эфф}}$

32. Сформулируйте метод максимума правдоподобия. Какова эвристическая идея этого метода?

Функцию правдоподобия (ФП) $L(x_1, \dots, x_n, \theta) = f(x_1, \theta) \cdot \dots \cdot f(x_n, \theta)$ будем рассматривать как функцию неизвестного параметра θ. Оценкой максимального правдоподобия (о.м.п) будем называть такое значение $\hat{\theta}_{\text{мп}}$ из множества допустимых значений θ для которого ФП принимает наибольшее значение при заданных $x_1, \dots, x_n$

$$\hat{\theta}_{\text{мп}} = \arg\max_{\theta} L(x_1, \dots, x_n, \theta)$$

Эвристическая идея заключается в том, что наиболее правдоподобным значением параметра является то при котором максимизируется вероятность получить именно такую выборку при n опытах. Заметим, что для абсолютно непрерывных распределений можно сказать, что

$$\mathbf{P}\big(x \in (y, y+dy)\big) = f(y)dy$$

и ФП является произведением плотностей т.е. мы как раз максимизируем вероятность события получения выборки $(x_1, \dots, x_n)$.

33. Сформулируйте метод моментов

Пусть $x_1, \dots, x_n$ выборка из генеральной совокупности с плотностью вероятности $f(x, \theta_1, \dots, \theta_m)$, где $\theta_1, \dots, \theta_m$ неизвестные параметры, подлежащие оцениванию. Метод моментов заключается в приравнивании выборочных начальных моментов a_k к соотв. генеральным α_k, являющимися функциями от неизвестных параметров. При этом количество моментов берется равным числу оцениваемых параметров

$$\alpha_k(\theta_1, \dots, \theta_m) = a_k, k = 1, \dots, m$$

Полученные уравнения решаются относительно параметров и эти решения являются искомыми оценками. Состоятельность выборочных моментов гарантирует состоятельность оценок метода моментов, для нормального распределения эти оценки также асимптотически эффективны.

34. Сформулируйте метод квантилей

Пусть $x_1, \dots, x_n$ выборка из генеральной совокупности с плотностью вероятности $f(x, \theta_1, \dots, \theta_m)$, где $\theta_1, \dots, \theta_m$ неизвестные параметры, подлежащие оцениванию. Метод квантилей заключается в

приравнивании выборочных квантилей распределения z_{p_k} к соотв. генеральным ξ_{p_k}, являющимися функциями от неизвестных параметров.

$$\xi_{p_k}(\theta_1, \dots, \theta_m) = z_{p_k}, k = 1, \dots, m$$

При этом обычно берут квантили уровня ¼, ½, ¾ т.е. нижнюю квартиль, медиану и верхнюю квартиль. Преимущество оценок в простоте вычисления, состоятельность гарантирована состоятельностью выборочных квантилей как оценок генеральных.

35. Каковы общие свойства оценок максимума правдоподобия?

1. Если существует эффективная оценка $\hat{\theta}_{\text{эфф}}$ параметра θ, то $\hat{\theta}_{\text{эфф}} = \hat{\theta}_{\text{мп}}$
2. О.м.п. состоятельны
3. О.м.п. асимптотически нормальны $N\left(\theta, \frac{1}{\sqrt{nI(\theta)}}\right)$
4. О.м.п. асимптотически эффективны $\mathbf{D}\hat{\theta}_{\text{мп}} = \frac{1}{nI(\theta)} + o\left(\frac{1}{n}\right)$
5. В общем случае могут быть смещенными

36. Каковы оценки максимума правдоподобия параметров нормального распределения?

$$m = \bar{x}, \sigma = s$$

вывод см. в Приложении

37. Каковы оценки максимума правдоподобия параметров равномерного распределения?

$$a = x_{\min}, b = x_{\max}$$

вывод см. в Приложении

38. Каковы оценки максимума правдоподобия вероятности “успеха” биномиального распределения?

$$p = \frac{1}{n^2}\sum_{k=1}^{n} x_k$$

вывод см. в Приложении

39. Какова оценка максимума правдоподобия для параметра масштаба показательного распределения?

$$\lambda = \frac{1}{\bar{x}}$$

вывод см. в Приложении

40. Каковы оценки максимума правдоподобия параметров распределения Лапласа?

$$\beta = \text{med}, \alpha = d^{-1} = \frac{n}{\sum_{i=1}^{n} |x_i - \text{med}|}$$

41. Каковы оценки метода моментов параметров нормального распределения?

$$m = \bar{x}, \sigma = s$$

вывод см. в Приложении

42. Что такое доверительный интервал?

Доверительным интервалом числовой характеристики или параметра распределения θ генеральной совокупности с доверительной вероятностью γ называется интервал (θ_1, θ_2) со случайными границами $\theta_1 = \theta_1(x_1, \dots, x_n), \theta_2 = \theta_2(x_1, \dots, x_n)$ который накрывает θ с вероятностью γ:

$$\mathbf{P}(\theta_1 < \theta < \theta_2) = \gamma$$

часто вместо γ рассматривается $\alpha = 1 - \gamma$ т.н. уровень значимости

43. Что такое интервальная оценка параметра и каковы ее отличия от точечной оценки?

Интервальная оценка параметра как сущность ↔ доверительный интервал параметра. Указывает область нахождения параметра с заданной вероятностью. Если мы по конкретной выборке получили численное значение точечной оценки, то о степени близости этого числа к оцениваемой величине мы ничего сказать не можем (кроме общих фраз о гарантиях качества близости). Интервальная оценка же позволяет задавать точность.

44. Что такое точность и надежность интервальной оценки?

- Точность: $\Delta = \frac{\theta_2 - \theta_1}{2}$
- Надежность: γ

Выигрывая в надежности, мы проигрываем в точности и наоборот.

45. Что такое критерий согласия?

Критерием значимости называется правило проверки статистической гипотезы, статистикой критерия называется статистика по значениям которой судят о справедливости статистической гипотезы. Критерием согласия называется критерий значимости, применяемый для проверки гипотезы о генеральном законе распределения.

46. Что такое ошибки первого и второго рода?

Ошибкой 1-го рода называется ошибка отвержения правильной гипотезы. Вероятность равна $\alpha = \mathbf{P}(\mathbf{Z} \in V_k / H_0)$ (условная вероятность) – уровню значимости.

Ошибкой 2-го рода называется ошибка принятия неверной гипотезы. Вероятность равна $\beta = \mathbf{P}(\mathbf{Z} \notin V_k / H_1)$.

47. Какие критерии согласия вы знаете?

- Критерий χ^2
- Критерий Колмогорова

48. Какова общая схема проверки статистических гипотез с использованием критериев согласия?

1. Выдвигается гипотеза H_0 о генеральном законе распределения с функцией распределения $F(x)$. Под конкурирующей гипотезой H_1 понимается гипотеза о справедливости одного из конкурирующих распределений.
2. Выбирается уровень значимости α
3. Выбирается статистика критерия $\mathbf{Z}$ и соотв. ей и проверяемым гипотезам критическая область V_k
4. Выбирается выборочное значение $\mathbf{Z}_\text{B}$ статистика
5. Если $\mathbf{Z}_\text{B} \in V_k$ – генеральный закон распределения не соответствует гипотетическому, выбирается конкурирующее распределение, и проверка повторяется
   Иначе гипотеза H_0 на данном этапе проверки принимается

49. Каково происхождение термина «регрессия»?

Термин введен в науку Фрэнсисом Гальтоном, которым он обозначил зависимость размеров потомков от размеров родителей. Она оказалась такой, что наблюдался регресс, т.е. рост детей наблюдался меньше роста родителей, если рост родителей был выше среднего, и наоборот.

50. Что такое задача простой линейной регрессии?

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$$

Где $x_1, \dots, x_n$ – значения фактора, а $y_1, \dots, y_n$ – наблюдаемые значения отклика, в свою очередь $\varepsilon_1, \dots, \varepsilon_n$ – независимые нормально распределенные $N(0, \sigma)$ с.в., а задача состоит в оценке параметров β_0, β_1

51. Какие методы оценивания параметров простой линейной регрессии вы знаете?

1. Метод наименьших квадратов:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min$$

расчетные формулы:

$$\hat{\beta}_1 \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

2. Метод наименьших модулей:

$$M(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min$$

является робастным

52. Как проверяется адекватность полученного решения задачи простой линейной регрессии?

Адекватность простой линейной регрессии – достаточно ли хорошо линейная функция $y = \beta_0 + \beta_1 x$ аппроксимирует экспериментальные точки. Об адекватности можно судить по поведению остатков: $\hat{\varepsilon}_i = y_i - \hat{y}_i, i - 1, \dots, n$. Однако такое суждение не всегда возможно и не формализовано математически.

1. Критерий Фишера

Пусть в каждой точке x_i проведено n_i измерений y_{ij} $(i = 1, \dots, m, j = 1, \dots, n)$ отклика Y. Вычисляется выборочное среднее в каждой точке:

$$\overline{y_i} = \frac{1}{n_i} \sum_{i=1}^{n_i} y_{ij} \, (i = 1, \dots, m)$$

И дисперсия воспроизводимости: $s_B^2 = \frac{Q_B}{n-m}, Q_B = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y_i}\right)^2$

Здесь $n - m$ – число степеней свободы статистики s_B^2.

Далее вычисляется дисперсия неадекватности: $s_H^2 = \frac{Q_H}{m-2}$,

$$Q_H = \sum_{i=1}^{m} n_i \left(\overline{y}_i - \hat{y}_i\right)^2$$

Сравниваются s_B^2 и s_H^2 с помощью отношения Фишера: $F = \frac{s_H^2}{s_B^2}$

Доказано, что статистики s_B^2, s_H^2 независимы; величина $s_B^2 \frac{n-m}{\sigma^2}$ распределена по закону λ^2 с $n-m$ степенями свободы, величина $s_H^2 \frac{n-m}{\sigma^2}$ – по закону χ^2 с $m-2$ степенями свободы, если гипотеза адекватности верна, т.е. $M\hat{y}_i = \beta_0 + \beta_1 x_i, i = 1, \dots, m$. Более того: $Ms_H^2 = \sigma^2, Ms_B^2 = \sigma^2$.

Т. е. s_B^2, s_H^2 – несмещённые оценки дисперсии.

Задаемся уровнем значимости α. Если $F = \frac{s_H^2}{s_B^2} < F_{1-\alpha}(m-2, n-m)$, то гипотеза адекватности принимается. $F_{1-\alpha}(m-2, n-m)$ – квантиль распределения Фишера.

2. Проверка наличия статистической зависимости между факторами и откликом

Если зависимость есть, то её можно описать линейной регрессией.

Отсутствие зависимости между X и Y можно проверить с помощью гипотезы $\beta_1 = 0$. Эта гипотеза принимается, если доверительный интервал при принятом уровне значимости накрывает нуль.

Наличие зависимости между X и Y можно проверить с помощью гипотезы $|\rho_{xy}| > 0$.

Приложение

Вывод о.м.п. для нормального распределения

ФП:

$$L(x_1, \dots, x_n, \theta_1, \theta_2) = \prod_{k=1}^{n} \frac{1}{\theta_2\sqrt{2\pi}} e^{-\frac{(x_k-\theta_1)^2}{2\theta_2^2}}$$

логарифмическая ФП тогда

$$\ln L = \sum_{k=1}^{n} \ln \frac{1}{\theta_2\sqrt{2\pi}} - \sum_{k=1}^{n} \frac{(x_k-\theta_1)^2}{2\theta_2^2} = n \ln \frac{1}{\theta_2\sqrt{2\pi}} - \sum_{k=1}^{n} \frac{(x_k-\theta_1)^2}{2\theta_2^2}$$

вычислим производные первого порядка и приравняем к нулю

$$\frac{\partial L}{\partial \theta_1} = \sum_{k=1}^{n} \frac{x_k-\theta_1}{\theta_2^2} = 0 \Rightarrow \sum_{k=1}^{n} x_k - n\theta_1 = 0 \Rightarrow \theta_1 = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x}$$

$$\frac{\partial L}{\partial \theta_2} = n\theta_2 \cdot \left(-\frac{1}{\theta_2^2}\right) + \frac{2}{\theta_2^3}\sum_{k=1}^{n} \frac{(x_k-\theta_1)^2}{2} = 0 \Rightarrow$$

$$\Rightarrow -\frac{n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{k=1}^{n} (x_k-\theta_1)^2 = 0 \Rightarrow \theta_2^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k-\bar{x})^2 = s^2$$

Таким образом о.м.п для нормального распределения – выборочное среднее и выборочная дисперсия т.к.

$$\frac{\partial^2 L}{\partial \theta_1^2} = -n\frac{1}{s^2} < 0$$

$$\frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} = -2\sum_{k=1}^{n} \frac{x_k-\bar{x}}{s^3} = -\frac{2}{s^3}\left(\sum_{k=1}^{n} x_k - n\bar{x}\right) = 0$$

$$\frac{\partial^2 L}{\partial \theta_2^2} = \frac{n}{s^2} - 3\frac{1}{s^4}\sum_{k=1}^{n} (x_k-\bar{x})^2 = \frac{n}{s^2} - 3\frac{n}{s^2} = -\frac{2n}{s^2}$$

и $\begin{vmatrix} -n\frac{1}{s^2} & 0 \\ 0 & -\frac{2n}{s^2} \end{vmatrix} = \frac{2n^2}{s^4} > 0$ т.е. $(\bar{x}, s)$ – максимум, определенный по критерию Сильвестра.

Вывод о.м.п. для равномерного распределения

ФП:

$$L(x_1, \dots, x_n, \theta_1) = \begin{cases} \prod_{k=1}^{n} \frac{1}{\theta_2 - \theta_1}, x_1, \dots, x_n \in [\theta_1, \theta_2] \\ 0, \text{иначе} \end{cases}$$

В данном случае видим, что максимум L достигается в случае если $\theta_2 - \theta_1$ как можно меньше и при этом $x_1, \dots, x_n \in [\theta_1, \theta_2]$, этому случаю соответствует $\theta_1 = \min\{x_1, \dots, x_n\}, \theta_2 = \max\{x_1, \dots, x_n\}$. Стоит заметить, что эта оценка крайне чувствительна к выбросам, потому-что для выборки $(1,1,1.5,1.5,1.5,2,2,2,3000)$ для которой кажется естественным взять $\theta_1 = 1, \theta_2 = 2$ получаем $\theta_1 = 1, \theta_2 = 3000$. Поэтому очень важно отсеивать выбросы.

Вывод о.м.п. для биномиального распределения

ФП в случае дискретных распределений строится как $L(x_1, \dots, x_n, \theta) = p_\theta(x_1) \cdot \dots \cdot p_\theta(x_n)$, где p_θ – функция вероятности дискретной случайной величины ξ_θ из некоторого параметрического семейства т.е. $p_\theta(x) = \mathbf{P}(\xi_\theta = x)$. Так что для биномиального распределения

$$L(x_1, \dots, x_n, \theta) = \begin{cases} \prod_{k=1}^{n} C_n^{x_k} \theta^{x_k} (1-\theta)^{n-x_k}, x_1, \dots, x_n \in \{0,1, \dots, n\} \\ 0, \text{иначе} \end{cases}$$

Как и было сказано ранее, можно рассматривать лишь случай $L > 0$, тогда логарифмическая ФП

$$\ln L = \sum_{k=1}^{n} \ln C_n^{x_k} + \sum_{k=1}^{n} \ln \theta \, x_k + \sum_{k=1}^{n} (n - x_k) \ln(1-\theta)$$

Составим уравнение правдоподобия

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{\theta} \sum_{k=1}^{n} x_k - \frac{1}{1-\theta} \sum_{k=1}^{n} (n - x_k) = 0 \Rightarrow$$
$$\Rightarrow (1-\theta) \sum_{k=1}^{n} x_k - \theta \sum_{k=1}^{n} (n - x_k) = 0 \Rightarrow$$
$$\Rightarrow \sum_{k=1}^{n} x_k - \theta n^2 = 0 \Rightarrow \hat{\theta}_{\text{мп}} = \frac{1}{n^2} \sum_{k=1}^{n} x_k$$

Интересные крайние случаи: $x_1 = \dots = x_n = 0 \Rightarrow \hat{\theta}_{\text{мп}} = 0$, если же $x_1 = \dots = x_n = n \Rightarrow \hat{\theta}_{\text{мп}} = 1$.

Вывод о.м.п. для показательного распределения

ФП в случае показательного распределения строится как

$$L(x_1, \dots, x_n, \theta) = \begin{cases} \theta^n e^{-\theta \sum_{k=1}^{n} x_k}, x_1, \dots, x_n \in [0, +\infty) \\ 0, \text{иначе} \end{cases}$$

Можно рассматривать лишь случай $L > 0$, тогда логарифмическая ФП

$$\ln L = n \ln \theta - \theta \sum_{k=1}^{n} x_k$$

Составим уравнение правдоподобия

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0 \Rightarrow \hat{\theta}_{\text{мп}} = \frac{n}{\sum_{k=1}^{n} x_k} = \frac{1}{\bar{x}}$$

Вывод о.м.м. для нормального распределения

Для нормального распределения: $\alpha_1 = m = \bar{x}$,

$$\alpha_2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} x^2 dx = \frac{1}{\sigma\sqrt{2\pi}} \sqrt{2\pi}\sigma(m^2 + \sigma^2) = m^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

итого

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = s^2$$

23. сосм. оценка станд. откл. нав

$\hat{\sigma}_{2\text{мп}} = \frac{R}{2}$ половина размаха