Глава 1

Дебильник

1.1 Многомерное нормальное распределение

def. Стандартный гауссовский вектор — случайный n-мерный вектор $Z = (Z_1, Z_2, \ldots Z_n)$, координаты которого независимы и имеют распределение $\mathcal{N}(0, 1)$.

def. Гауссовский вектор (Нормальный вектор) — вектор, для которого существует матрица $\mathbf{A} \in \mathbb{R}^{n \times m}$, стандартный гауссовский вектор $Z \in \mathbb{R}^m$, и вектор $b \in \mathbb{R}^n$ такие, что $X = \mathbf{A}Z + b$.

def. Распределение нормального вектора $X \in \mathbb{R}^n$ — $\mathcal{N}(\mu, \mathbf{\Sigma})$ или $\mathcal{N}_n(\mu, \mathbf{\Sigma})$, где $\mu = \mathbb{E}X$ и $\mathbf{\Sigma} = \operatorname{cov}(X)$.

def. Распределение хи-квадрат с n степенями свободы — распределение $\chi^2(n)$ величины $\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2$, где $Z_1, Z_2, \ldots Z_n$ — независимы $\mathcal{N}(0, 1)$ величины.

def. Распределение Стьюдента с n степенями свободы — распределение $T(n)$ величины $\frac{\sqrt{n}X}{\sqrt{Y}}$, где $X \sim \mathcal{N}(0, 1)$, $Y \sim \chi^2(n)$ и независимы.

def. Распределение Фишера со степенями свободы n и m — распределение $F(n, m)$ величины $\frac{X/n}{Y/m}$, где $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ и независимы.

1.2 Условное матожидание

def. Условное матожидание $\mathbb{E}(Y \mid X)$ случайной величины Y при условии случайной величины X — такая измеримая функция g_0 величины X, при которой $\mathbb{E}(Y - g(X))^2$ минимально для всех измеримых функций g.

Условное матожидание — ортогональная проекция Y на линейное пространство всех измеримых функций X. То есть УМО — единственная измеримая функция, которая удовлетворяет условию ортогональности:

$$\forall g\colon \mathbb{E}(Y - \mathbb{E}(Y \mid X))g(X) = 0.$$

1.3 Статистическая модель, выборка

def. Статистическая модель — множество распределений $\mathfrak{P}$, которое, по нашему мнению, адекватно приближает $\mathcal{P}_D$.

def. Данные d — реализация случайного элемента D, имеющего распределение $\mathcal{P}_D$.

Статистические модели делят на:

- параметрические, если $\mathfrak{P} = \{\mathcal{P}_0 \mid \theta \in \Theta \subset \mathbb{R}^k\}$.

  Пример: $\mathfrak{P} = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \geqslant 0\}$.

- непараметрические, если $\mathfrak{P} = \{\mathcal{P}_0 \mid \theta \in \Theta \subset V\}$, где V не обязательно конечномерное.

  Пример: $\mathfrak{P} = \{\mathcal{P}^{\otimes n} \mid \int_{\mathfrak{X}} x\mathcal{P}(dx) = 0\}$

- семипараметрические, если $\mathfrak{P} = \{\mathcal{P}_0 \mid \theta \in \Theta \subset \mathbb{R}^k \times V\}$.

  Пример: линейная регрессия $Y = X\beta + \varepsilon$, $\beta \in \mathbb{R}^k$, $\mathbb{E}\varepsilon = 0$, $\mathbb{D}\varepsilon = \sigma^2$.

Если $D = [X_1, \dots X_n]$ и X_i независимы и имеют одинаковое распределение $\mathcal{P}_X$, D называется выборкой объема n и обозначается $X_{[n]}$, $\mathcal{P}_X$ — генеральная совокупность. В этом случае модель приобретает вид $\mathfrak{P} = \{\mathcal{P}^{\otimes n} \mid \mathcal{P} \in \mathfrak{P}_X\}$, где $\mathfrak{P}_X$ — модель для $\mathcal{P}_X$.

1.4 Формула Байеса, априорное, апостериорное распределение

- Априорное распределение — наше ощущение относительно значения параметра до проведения эксперимента.
- Апостериорное распределение — ощущение после получения данных эксперимента.

def (Формула Байеса)**.** Здесь p — вероятность, d — данные, θ — параметры.

$$p(\theta \mid d) = \frac{p(d \mid \theta) \cdot p(\theta)}{p(d)}.$$

- $p(\theta \mid d)$ — апостериорное распределение,
- $p(d \mid \theta)$ — правдоподобие,
- $p(\theta)$ — априорное распределение,
- $p(d)$ — вероятность данных.

1.5 Расстояние Кульбака-Лейблера, энтропия

Пусть мы принимаем случайные символы $x_1, \ldots x_k$, вероятность появления x_i равна p_i, записываем с помощью битовой строки длины l_i. Тогда средняя длина символа равна

$$l = \sum_{i=1}^{k} p_i \cdot l_i.$$

Чтобы минимизировать l, необходимо подобрать следующие $l_i = -\log_2 p_i$. И тогда средняя длина будет равна $H(x) \coloneqq -\sum_{i=1}^{k} p_i \cdot \log_2 p_i$, эта величина называется двоичной энтропией сообщения. Аналогично можно брать любой другой логарифм, мы будем использовать натуральный.

Для непрерывной величины можно завести дифференциальную энтропию:

$$H(X) = -\int p(x) \log p(x) dx.$$

Пусть случайная величина X имеет функцию вероятности p, но мы кодируем символы, как-будто она имеет функцию вероятности q. Тогда средняя длина сообщения будет равна $-\sum_{i=1}^{k} p_i \cdot \log q_i$, эта величина называется кросс-энтропией $H(p \mid q)$ распределений p и q.

$H(p \mid q)$ всегда будет больше $H(p)$, так как $H(p)$ минимально.

def. Величина потери информации из-за использования q вместо p называется расстоянием Кульбака-Лейблера между p и q:

$$D_{KL}(p, q) = H(p \mid q) - H(p) = -\sum_{i=1}^{k} p_i \cdot \log \frac{q_i}{p_i}.$$

Для непрерывных величин все обобщается следующим образом

$$D_{KL} = -\int p_i \cdot \log \frac{q_i}{p_i}.$$

1.6 Статистика...

1.6.1 Статистика

Параметр или характеристика распределения — функционал от этого распределения.

def. Статистика — функция θ^* от данных d.

Пусть модель $\mathfrak{P}_{[n]} = \{\mathcal{P}^{\otimes n} \mid \mathcal{P} \in \mathfrak{P}\}$, искомая характеристика $\theta\colon \mathfrak{P} \to \mathbb{R}^k$.

1.6.2 Несмещенность

Чему равна оценка как случайная величина в среднем, если она равна характеристике?

def. Оценка θ^* называется

- несмещенной, если $\forall \mathcal{P} \in \mathfrak{P}\colon \mathbb{E}\theta^*(X_{[n]}) = \theta(\mathcal{P})$, где $X_{[n]} \sim \mathcal{P}^{\otimes n}$,
- асимптотически несмещенной, если $\forall \mathcal{P} \in \mathfrak{P}\colon \mathbb{E}\theta^*(X_{[n]}) \to \theta(\mathcal{P})$.

Смещение — величина $b(\theta^*) = \mathbb{E}(\theta^*(X_{[n]})) - \theta(\mathcal{P})$.

Среднеквадратичная ошибка — величина $\mathrm{MSE}(\theta^*) = \mathbb{E}\left(\theta^*(X_{[n]}) - \theta(\mathcal{P})\right)^2$.

В общем случае
$$\mathrm{MSE}(\theta^*) = \mathbb{D}\theta^*(X_{[n]} + b^2(\theta^*).$$

- Выборочное среднее как оценка матожидания — несмещенная оценка,
- Выборочная дисперсия как оценка дисперсии — асимптотически несмещенная,
- Исправленная выборочная дисперсия как оценка дисперсии — несмещенная оценка.

1.6.3 Состоятельность

def. Оценка θ^* называется

- состоятельной, если $\forall \mathcal{P} \in \mathfrak{P}\colon \theta^*(X_{[n]}) \xrightarrow{\mathbb{P}} \theta(\mathcal{P})$, где $X_{[n]} \sim \mathcal{P}^{\otimes n}$,
- сильно состоятельной, если $\theta^*(X_{[n]}) \xrightarrow{\text{п. н.}} \theta(\mathcal{P})$.

1.6.4 Асимптотическая нормальность

def. Оценка θ^* называется асимптотически нормальной с коэффициентом рассеивания (или просто дисперсией) $\sigma^2\big(\theta(\mathcal{P})\big) > 0$, если

$$\sqrt{n}\left(\theta^*(X_{[n]}) - \theta(\mathcal{P})\right) \xrightarrow{d} \eta \sim \mathcal{N}(0, \sigma^2(\theta^*(\mathcal{P}))).$$

В многомерном случае рассматривается ковариационная матрица вместо дисперсии.

- Выборочная дисперсия и второй момент — асимптотически нормальная оценка.
- Из асимптотической нормальности следует состоятельность.

1.6.5 Эффективность

Рассмотрим класс оценок $K = \{\hat{\theta}\}$ параметра θ.

def. Оценка $\theta^* \in K$ называется эффективной в классе K, если *для любой другой оценки* $\hat{\theta} \in K$ *и для любого исследуемого параметра* $\theta \in \Theta$ *выполняется*

$$\mathrm{MSE}_\theta(\theta^*) \leqslant \mathrm{MSE}_\theta(\hat{\theta}).$$

Класс несмещенных оценок

$$K_0 = \{\hat{\theta} \mid \mathbb{E}\hat{\theta} = \theta, \forall \theta \in \Theta\}.$$

def. Эффективная оценка θ^*, если эффективна в классе K_0.

def. Асимптотически эффективной в классе K, если для любой оценки $\hat{\theta} \in K$ и для любого $\theta \in \Theta$ выполняется

$$\overline{\lim_{n\to\infty}} \frac{\mathrm{MSE}(\theta^*)}{\mathrm{MSE}(\hat{\theta})} \leqslant 1.$$

1.6.6 Робастность

def. Робастность — свойство оценки быть устойчивой к хвостам распределения.

Пусть F — распределение, $\{G_n\}$ — последовательность распределений, что

$$|F - G_n| \coloneqq \sup_x |F(x) - G_n(x)| \to 0.$$

def. Характеристика θ обладает качественной робастностью, если $\theta(G_n) \to \theta(F)$

Пусть также δ_x — вырожденное распределение в точке x.

def. Загрязненное распределение — смесь $F_{x,\varepsilon} = (1-\varepsilon)F + \varepsilon\delta_x$.

def. Функция влияния характеристики θ — величина

$$IF(x) = \lim_{\varepsilon\to 0+} \frac{\theta(F_{x,\varepsilon}) - \theta(F)}{\varepsilon}.$$

def. Характеристика θ называется B-робастной или инфинитезимально робастной, если $IF(x)$ ограничена.

def. Асимптотическая толерантность характеристики θ —

$$\tau = \inf\bigl\{\varepsilon \mid \sup_x |\theta(F_{x,\varepsilon} - \theta(F)| = \infty\bigr\}.$$

1.6.7 Достаточность

def. Статистика $T(x) = \{T_1(x), \dots, T_m(x))\}$ называется достаточной, если для всех

- $\theta \in \Theta$,
- $B \in \mathfrak{P}(\mathbb{R}^n)$ и
- $t = (t_1, \dots, t_m)$

условная вероятность $\mathbb{P}(X_{[n]} \in B \mid T(X_{[n]}) = t)$ не зависит от θ.

То есть информация о θ в выборке полностью содержится в значении $T(x_{[n]})$.

thm (факторизации)**.** *$T(x)$ достаточна, согда существуют функции g и h, что*

$$p(X_{[n]} = x_{[n]} \mid \theta) = g(T(x_{[n]}), \theta) h(x_{[n]}),$$

где p — вероятность или плотность.

1.6.8 Полнота

def. Статистика T называется полной, если для любой *измеримой g верно следствие*

$$\forall \theta \in \Theta \colon \mathbb{E} g(T(X_{[n]})) \equiv 0 \quad \Longrightarrow \quad g(T(X_{[n]})) \overset{п.н.}{=} 0.$$

1.7 Теоремы Колмогорова-Блэкуэлла-Рао и Лемана-Шеффе

thm (Колмогорова-Блэкуэлла-Рао). *Пусть θ^* — оценка параметра θ, T — достаточная статистика. Тогда*

$$\mathrm{MSE}(\theta^*) \geqslant \mathrm{MSE}(\mathbb{E}(\theta^* \mid T)).$$

thm (Лемана-Шеффе). *Пусть θ^* — оценка параметра θ, T — достаточная и полная статистика. Тогда $\mathbb{E}(\theta^* \mid T)$ — единственная эффективная оценка в классе оценок со смещением $b(\theta^*)$.*

1.8 Доверительный интервал

Пусть есть модель $\mathfrak{P}_{[n]} = \{\mathcal{P}^{\otimes n} \mid \mathcal{P} \in \mathfrak{P}\}$ и $\theta\colon \mathfrak{P} \to \mathbb{R}^k$ — искомая характеристика.

def. Доверительный интервал (точный доверительный интервал) с уровнем доверия γ — пара статистик (θ^*_L, θ^*_R), такая что для любого $\mathcal{P} \in \mathfrak{P}$ и $X_{[n]} \sim \mathcal{P}^{\otimes n}$

$$\mathbb{P}\left(\theta^*_L(X_{[n]}) \leqslant \theta(\mathcal{P}) \leqslant \theta^*_R(X_{[n]})\right) = \gamma.$$

Интервал называется

- асимптотическим, если

$$\mathbb{P}\left(\theta^*_L(X_{[n]}) \leqslant \theta(\mathcal{P}) \leqslant \theta^*_R(X_{[n]})\right) \xrightarrow{n\to\infty} \gamma.$$

- центральным, если

$$\mathbb{P}\left(\theta^*_L(X_{[n]}) > \theta(\mathcal{P})\right) = \mathbb{P}\left(\theta^*_R(X_{[n]}) < \theta(\mathcal{P})\right).$$

- левым, если

$$\mathbb{P}\left(\theta^*_L(X_{[n]}) > \theta(\mathcal{P})\right) = 0.$$

- правым, если

$$\mathbb{P}\left(\theta^*_R(X_{[n]}) < \theta(\mathcal{P})\right) = 0.$$

1.9 Бутстреп

1.9.1 Параметрический бутстреп

Если работаем с параметрической моделью, можем заменить $X = X(\theta)$ не на X^*, а на $X(\theta^*)$ и сэмплировать из этого распределения.

1.9.2 Непараметрический бутстреп

Рецепт

1. изготовим N выборок $x^*_{[n],1}, \dots, x^*_{[n],N}$ из эмпирического распределения (рандом с возвращением)

2. вычисляем $\theta^b_i = \theta^*(x^*_{[n],i})$, получаем бутстреповскую выборку $\theta^b_{[N]}$,

3. по бутстреповской выборке оцениваем, что нужно.

Ограничения

- θ^* — plug-in оценка

- θ^* — достаточно гладкая (обычно дифференцируема)

- у X достаточно много моментов (обычно конечная дисперсия)

- нужно генерировать большие выборки

- на очень больших данных трудозатратен

- на маленьких данных велика неустранимая ошибка

1.10 Гипотеза, альтернатива...

Пусть $\mathfrak{P}$ — модель.

1.10.1 Гипотеза и альтернатива

def. Гипотеза — утверждение вида $H\colon \mathcal{P}_X \in \mathfrak{P}_0 \subset \mathfrak{P}$.

Если $|\mathfrak{P}_0| = 1$, гипотеза называется простой, иначе сложной.

Нулевая гипотеза — гипотеза H_0, которую мы хотим проверить. Проверка гипотезы — процесс принятия решения о том, противоречит ли она наблюдаемой выборке данных.

Альтернатива — гипотеза H_1, которая отражает, какие отклонения от нулевой гипотезы нам интересны.

1.10.2 Критерий

def. Нерандомизированный критерий (критерий) — отображение $\varphi\colon d \to \{\text{принимаем}, \text{отвергаем}\} = \{H_0, H_1\} = \{0, 1\}$.

Часто критерий устроен так: имеется
- статистика критерия T и
- критическое множество C, и

$$\varphi(d) = [T(d) \in C] = [d \in T^{-1}(C)].$$

def. Рандомизированный критерий — отображение $\varphi\colon d \to [0, 1]$. Значение на данных d определяется как реализация случайной величины $D(\varphi(d))$.

Пусть мы согласны отвергать нулевую гипотезу пр условии, что она верна, но хотим делать это не очень часто. Пусть зафиксирован уровень значимости

$$\alpha \coloneqq \mathbb{P}(\varphi(D) = 1 \mid H_0),$$

который обычно является *параметром критерия*, то есть, задавая его, мы определяем *критическое множество* C_α такое, что

$$\mathbb{P}(T(D) \in C_\alpha \mid H_0) = \alpha.$$

Таким образом, для одного критерия определено семейство критических областей $\{C_\alpha \mid \alpha \in [0, 1]\}$, где обычно $C_\alpha \subset C_{\alpha'}$, если $\alpha < \alpha'$.

def. Уровень значимости — параметр критерия, который регулирует, насколько часто мы будем отвергать нулевую гипотезу при условии, что она верна.

1.10.3 p-value

Хотим оценить, насколько гипотеза противоречит наблюдаемым данным.

<u>**def**</u>**.** p-value — характеристика противоречия гипотезы наблюдаемым данным:

$$\text{p-value} \coloneqq \arg\min\{\alpha \in [0,1] \mid T(d) \in C_\alpha\}.$$

Другими словами, p-value — минимальное значение уровня значимости для данного значения статистики критерия, при котором H_0 может быть отвергнута.

Чем меньше p-value, тем больше гипотеза противоречит данным.

1.10.4 Ошибки разных родов

<u>**def**</u>**.** Ошибка первого рода — событие $\varphi(D) = 1 \mid H_0$. Если уровень значимости совпадает с вероятностью ошибки первого роба, То критерий называется точным.

Уровень значимости — вероятность ошибки первого рода.

<u>**def**</u>**.** Ошибка второго рода β — событие $\varphi(D) = 0 \mid H_1$, не отклонили нулевую гипотезу при условии, что была верна альтернатива.

Мощность критерия — вероятность $1 - \beta$ отклонить H_0 при условии, что верна H_1.

Для заданного уровня значимости мы хотим иметь как можно более мощный критерий.

1.10.5 Свойства критериев

<u>**def**</u>**.** Несмещенность — мощность всегда не меньше ошибки первого рода, критерий не отдает предпочтение альтернативе. $1 - \beta \geqslant \alpha$ для всех простых гипотез из $\mathfrak{P}_0$ и простых альтернатив из $\mathfrak{P}_1$.

<u>**def**</u>**.** Состоятельность — $\beta \xrightarrow{n\to\infty} 0$ для всех простых альтернатив из $\mathfrak{P}_1$.

<u>**def**</u>**.** Асимптотичность — $\alpha \xrightarrow{n\to\infty}$ для всех простых гипотез из $\mathfrak{P}_0$.

def. Наиболее мощный критерий для данного уровня значимости α_0 и простой альтернативы — такой критерий φ_1, что для любого критерия φ_2 такого, что $\alpha(\varphi_2) \leqslant \alpha_0$:

$$\beta(\varphi_1) \leqslant \beta(\varphi_2).$$

1.10.6 Размер эффекта

Во многих случаях важна не только информация о p-value, но и величина наблюдаемого эффекта. Размеры эффекта бывают разные, использование того или иного размера эффекта зависит от контекста.

Вместо сравнения p-value с уровнем значимости для принятия статистического решения можно считать размер эффекта, сравнивать с минимальным практически интересным.

1.11 Постановка гипотезы согласия. Критерии Колмогорова и Андерсона-Дарлинга

1.11.1 Постановка гипотезы согласия

def. Гипотеза согласия — гипотеза о соответствии эмпирического распределения теоретическому распределению вероятностей.

Критерии для гипотез согласия бывают

- общие — применимые к любому предполагаемому распределению выборки,
- специальные — применимые к гипотезам, формулирующие согласие с определенным свойством распределений;

- для простых гипотез,
- для сложных гипотез.

1.11.2 Критерий Колмогорова

Сравнивает эмпирическое и истинное распределение. Для простой гипотезы.

Пусть F_0 непрерывна на $\mathbb{R}$. Определим статистику Колмогорова:

$$D_n(x_{[n]}) = \sup_{x\in\mathbb{R}} |F_n^* - F_0(x)|.$$

- Если H_0 верна, то $D_n\left(X_[n]\right) \xrightarrow{\text{п.н.}} 0$;
- Если H_0 неверна, то $D_n\left(X_{[n]}\right) \xrightarrow{\text{п.н.}} \sup_{x\in\mathbb{R}} |F_X(x) - F_0(x)| > 0$.

1.11.3 Критерий Андерсона-Дарлинга

Для простой гипотезы.

Определим статистику критерия Андерсона-Дарлинга:

$$\begin{aligned} A^2 &= n\int_{\mathbb{R}} \frac{\left(F_n^*(x) - F_0(x)\right)^2}{F_0(x)\left(1 - F_0(x)\right)} dF_0(x) = \\ &= -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln F_0(X_{(i)} \mid \theta) + \ln\left(1 - F_0(X_{(n+1-i)} \mid \theta\right)\right] \end{aligned}$$

Статистика A^2 при выполнении H_0 и непрерывности F_0 подчиняется табличному распределению. $C_\alpha = (a_{1-\alpha}^2, \infty)$.

1.12 Постановка гипотезы о параметрах, проверка через доверительные интервалы, z-test, t-test, бутстреп из нулевой гипотезы

1.12.1 Постановка гипотезы о параметрах

Пусть θ — параметр ($X \sim F(x,\theta)$) или характеристика ($\theta = \varphi(F_x)$) распределения.

Нулевая гипотеза: $H_0\colon \theta = \theta_0$. Типичные альтернативы:

- $H_1\colon \theta = \theta_1 \neq \theta_0$,
- $H_>\colon \theta > \theta_0$,
- $H_<\colon \theta < \theta_0$,
- $H_{\neq}\colon \theta \neq \theta_0$.

Усредненный рецепт:

1. Выбираем оценку θ^* параметра θ, распределение которой приближенно известно при данном θ.

2. В зависимости от альтернативы строим критическое множество:

   - $H_1\colon \theta = \theta_1 > \theta_0$ или $H_>$, то $C_\alpha = (\theta^*_{1-\alpha}, \infty)$ [1] — правое критическое множество;
   - $H_1\colon \theta = \theta_1 < \theta_0$ или $H_<$, то $C_\alpha = (-\infty, \theta^*_\alpha)$ — левое критическое множество;
   - $H_{\neq}$, то $C_\alpha = (-\infty, \theta^*_{\frac{\alpha}{2}} \cup (\theta^*_{1-\frac{\alpha}{2}}, \infty)$ — двустороннее критическое множество.

3. Если $\theta^* \in C_\alpha$, то гипотезу можно отклонить, иначе — нельзя.

1.12.2 Проверка через доверительные интервалы

Усредненный рецепт:

1. В зависимости от альтернативы строим доверительный интервал с уровнем доверия $\gamma = 1 - \alpha$:

   - $H_1\colon \theta = \theta_1 > \theta_0$ или $H_>$, то (θ^*_L, ∞) — правый доверительный интервал;
   - $H_1\colon \theta = \theta_1 < \theta_0$ или $H_<$, то $(-\infty, \theta^*_R)$ — левый доверительный интервал;
   - $H_{\neq}$, то (θ^*_L, θ^*_R) — центральный доверительный интервал.

[1] Здесь θ^*_x — квантиль уровня x распределения $\theta^* \mid H_0$

1.12.3 z-test

Пусть $X \sim \mathcal{N}(\mu, \sigma^2)$, μ неизвестно, σ^2 известно.

Если H_0 верна, то $Z = \frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} \sim \mathcal{N}(0,1)$ и $\overline{X} \sim N(\mu_0, \frac{\sigma^2}{n})$.

В зависимости от альтернативы подбираем критическую область:

	$\theta_1 > \theta_0,\ H_>$	$\theta_1 < \theta_0,\ H_<$	$H_{\neq}$
C_α	$(\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty)$	$(-\infty, \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}})$	$\mathbb{R} \setminus (\mu_0 \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$
p-value	$1 - \Phi^{-1}(z)$	$\Phi^{-1}(z)$	$2(1 - \Phi^{-1}(\lvert z \rvert))$

Таблица 1.1: Критическая область для альтернативы

1.12.4 t-test

Пусть $X \sim \mathcal{N}(\mu, \sigma^2)$, μ неизвестно, σ^2 неизвестно.

Если H_0 верна, то $T = \frac{\sqrt{n}(\overline{X}-\mu_0)}{s} \sim T(n-1)$.

В зависимости от альтернативы подбираем критическую область:

	$\theta_1 > \theta_0,\ H_>$	$\theta_1 < \theta_0,\ H_<$	$H_{\neq}$
C_α	$(\mu_0 + t_{1-\alpha}\frac{s}{\sqrt{n}}, \infty)$	$(-\infty, \mu_0 + t_\alpha \frac{s}{\sqrt{n}})$	$\mathbb{R} \setminus (\mu_0 \pm t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$
p-value	$1 - T^{-1}(t)$	$T^{-1}(t)$	$2(1 - T^{-1}(\lvert t \rvert))$

Таблица 1.2: Критическая область для альтернативы

1.12.5 Бутстреп из нулевой гипотезы

Пусть мы хотим проверить гипотезу $H_0 \colon \mathbb{E}X = \theta_0$.

Рецепт:

1. Назначим каждому наблюдению x_i в выборке вероятность p_i.
2. Из пар (x_i, p_i) изготовим дискретное распределение F_p^*.

3. Подберем p_i так, чтобы с одной стороны $\overline{x} = \theta_0$, а с другой p_i максимизировали правдоподобие выборки $\mathcal{L}(p \mid x_{[n]}) = p_1 p_2 \dots p_n$.

4. Бутстрепим кучу выборок из получившегося F_p^*, считаем по ним выборочное среднее.

5. Построим критическое множество в зависимости от альтернативы и проверим, лежит ли в нем выборочное среднее исходной выборки.

1.13 Постановка гипотезы однородности, ранговые критерии, permutation.

1.13.1 Гипотеза однородности

Пусть $D = (x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m)$ и

- $\mathsf{X}_{\text{нез}}$ — $X_1, \dots, X_n$ независимы и имеют одну функцию распределения $F(x)$,

- $\mathsf{Y}_{\text{нез}}$ — $Y_1, \dots, Y_n$ независимы и имеют одну функцию распределения $G(x)$,

- FG_{c} — F и G непрерывны.

Гипотеза однородности: $H_0\colon F = G$.

Альтернативы:

- неоднородности: $H_{\neq}\colon \exists x\ F(x) \neq G(x)$;

- доминирования $H_{\geqslant}\colon \forall x\ F(x) \geqslant G(x) \wedge \exists x\ F(x) > G(x)$;

- правого сдвига $H_{\rightarrow}\colon \forall x\ F(x) = G(x + \theta) \wedge \theta > 0$;

- масштаба $H_{\leftrightarrow}\colon \forall x\ F(x) = G(x\theta) \wedge 1 \neq \theta > 0$;

1.13.2 Ранговые критерии

Критерий Уилкоксона ранговых сумм

Используется для проверки гипотезы H_0 против $H_{\geqslant}$ и $H_{\rightarrow}$.

Идея: Если H_0 верна, то $Y_{(i)}$ распределены в вариационном ряду Z равномерно.

Статистика критерия: $W = \sum_{i=1}^{m} R(Y_i)$.

$$W \in \left[\frac{m(m+1)}{2}, mn + \frac{m(m+1)}{2}\right].$$

$$C_\alpha = \left(c_\alpha, mn + \frac{m(m+1)}{2}\right).$$

Критерий Манна-Уитни

Используется для проверки гипотезы H_0 против $H_{\geqslant}$ и $H_{\rightarrow}$.

Идея аналогичная.

Статистика критерия: $U = \sum_{i=1}^{n} \sum_{j=1}^{m} [x_i < y_j]$.

Нетрудно видеть, что $U = W - \frac{m(m+1)}{2}$, поэтому $U \in [0, mn]$.

$$C_\alpha = (c_\alpha, mn).$$

Ценность этих критериев в том, что можно проверять выборки из величин, сравнимых только качественно.

Критерий знаковых рангов Уилкоксона

Пусть

- $\mathsf{E}_{\text{нез}}$ — $E_1, \ldots, E_n$ независимы,
- $\mathsf{E}_{\text{сим}}$ — $E_1, \ldots, E_n$ распределены одинаково и симметричны относительно нуля.

Рассматриваем вариационный ряд величин $|z_i|$.

Статистика критерия: $T = \sum_{i=1}^n R(|z_i|)[z_i < 0]$.

Для маленьких n квантили смотрим в таблице, для больших можем использовать Монте-Карло или аппроксимацию:

$$\frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \to \eta \sim \mathcal{N}(0,1).$$

1.13.3 Permutation

Пусть выполнено $\mathsf{E}_{\text{нез}}$ и $\mathsf{E}_{\text{сим}}$.

1. Случайно умножаем z_i на 1 и -1 ;
2. Для полученного вектора считаем медиану;
3. Повторяем так N раз;
4. Считаем, сколько медиан меньше $\mathrm{med}(z)$ и делим на N.

Пусть $y^p_{[n]}$ — случайная перестановка $y_{[n]}$. Если верна H_0, то все выборки $(x, y^p)_{[n]}$ «равновероятны». План-капкан:

1. генерируем случайную перестановку y^p,
2. вычисляем значение требуемой статистики $r(x, y^p)$,
3. повторяем N раз,
4. считаем, какая доля оказалась меньше, чем $r(x, y)$.

Пусть есть статистика $\theta^*(x_{[n]}, y_{[m]})$. Мы можем представить ее в виде $\theta^* = \theta^*(z_v, u)$.

Пусть u^p — случайная перестановка u. Тогда, если верна H_0, то

$$\mathbb{P}\left(\theta^*(z_v, u^p) < \theta^*(z_v, u)\right) = \frac{\#\{u^p \mid \theta^*(z_v, u^p) < \theta^*(z_v, u)\}}{\binom{n+m}{n}}.$$

Статистика критерия:

$$C(x_{[n]}, y_{[m]}) = \frac{\#\{u^p \mid \theta^*(z_v, u^p) < \theta^*(z_v, u)\}}{\binom{n+m}{n}},$$

$$C_\alpha \in \left\{(1-\alpha, 1),\ (0, \alpha),\ (0, \tfrac{\alpha}{2}) \cup (1 - \tfrac{\alpha}{2}, 1)\right\}.$$

Killer Feature: ошибка первого рода в точности равна α.

1.14 Дисперсионного анализ, корреляционного анализ, таблицы сопряженности.

ANOVA — ANalysis Of VAriance.

1.14.1 Однофакторный дисперсионный анализ

- имеется несколько выборок $x^1_{[n_1]}, \ldots, x^k_{[n_k]}$,
- которые являются наблюдениями случайных величин

  $$X_{i,j} = \mu + \beta_j + \varepsilon_{i,j}, \quad i = 1, \ldots n, j = 1, \ldots k,$$

  где μ — общее среднее, β_j — систематическая ошибка (или эффект фактора) выборки $x^j_{[n_j]}$ и $\varepsilon_{i,j}$ — случайная ошибка.

Постановка задачи дисперсионного анализа

Пусть $\mu_j = \mu + \beta_j$, $N = n_1 + \ldots + n_j$ и

- $\mathsf{E}_{\text{нез}}$ — все ошибки $\varepsilon_{i,j}$ независимы,
- $\mathsf{E}_{\text{с}}$ — все ошибки $\varepsilon_{i,j}$ имеют одинаковое непрерывное распределение.

Гипотеза $H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k$.

Альтернатива $H_1\colon \exists i, j\colon \mu_i \neq \mu_j$.

Критерий Андерсона-Дарлинга для ANOVA

Пусть выборки $X^i_{[n_i]}$ независимы и имеют распределение F_i.

Гипотеза $H_0\colon F_1 = F_2 = \ldots = F_k$.

Альтернатива $H_1\colon \exists i, j\colon F_i \neq F_j$.

$$A^2_{k,N} = \sum_{i=1}^{k} n_i \int_{\mathbb{R}} \frac{\left(F^2_{n_i}(x) - H_N(x)\right)^2}{H_N(x)\left(1 - H_N(x)\right)} dH_N(x),$$

где $H_N = \frac{1}{n}\sum_{i=1}^{k} n_i F^*_{n_i}$ — эмпирическая функция распределения объединенной выборки $Z = \left(X_{[n_1]}, \ldots, X_{[n_k]}\right)$.

Если повторений нет, то можно переписать следующим образом:

$$A^2_{k,N} = \frac{1}{N}\sum_{j=1}^{k}\frac{1}{n_j}\sum_{i=1}^{N-1}\frac{(Nc_{i,j} - jn_j)^2}{i(N-i)},$$

где $c_{i,j}$ — количество наблюдений $X^j_{[n_j]}$ меньших $Z_{(i)}$.

1.14.2 Постановка задачи корреляционного анализа

Гипотеза независимости Имеется выборка $(x, y)_{[n]}$ — реализация $(X, Y)_{[n]}$, при этом X_i имеет функцию распределения F_X, а Y_i имеет F_Y.

Гипотеза $H_0\colon F_{X,Y}(x, y) = F_X(x)F_Y(y)$.

Предполагаем гипотезу независимости, но проверять будем отсутствие корреляции.

1.14.3 Таблица сопряженности

Три схемы, в которых они возникают:

- Гипотеза однородности: строка i — реализация случайной величины x_i с вероятностями $\mathbb{P}(x_i = y_j) = q_{i,j}$, $\sum q_{i,j} = 1$ и с заданным числом наблюдений n (то есть $\forall i\colon n = n_{i,+}$).

$$H_I\colon q_{i,j} = q_{+,j}, \quad q_{+,j} = \frac{1}{k}\sum_{i} q_{i,j}.$$

	y_1	...	y_l	
x_1	$n_{1,1}$	...	$n_{1,l}$	$n_{1,+}$
...	...	...	...	...
x_k	$n_{k,1}$	...	$n_{k,l}$	$n_{k,+}$
	$n_{+,1}$	...	$n_{+,l}$	$n_{+,+}$

Пример: k кубиков, каждый подбросили n раз, $n_{i,j}$ — количество выпадений числа j у кубика i.

- Гипотеза независимости: вся таблица — реализация случайной величины ξ с $\mathbb{P}\left(\xi = (x_i, y_j)\right) = p_{i,j}$, где $\sum_{i,j} p_{i,j} = 1$, и с $n_{*,*}$ наблюдениями.

$$H_{II}\colon \mathbb{P}\left(\xi = (x_i, y_j)\right) = \mathbb{P}(\xi_x = x_i)\mathbb{P}(\xi_y = y_j).$$

  Пример: выборка двумерной случайной величины $(x, y)_{[n]}$, для которой построена гистограмма с ячейками $\delta_i^x \times \delta_j^y$.

- Гипотеза мультипликативности: каждая ячейка — реализация случайной величины.

  Важный частный случай: когда $n_{i,j}$ независимы и имеют распределение Пуассона с параметрами $\lambda_{i,j}$. Тогда их сумма тоже имеет распределение Пуассона.

$$H_{III}\colon \lambda_{i,j} = \frac{a_i b_j}{c}, \quad a_i = \sum_j \lambda_{i,j}, b_j = \sum_i \lambda_{i,j}, c = \sum_{i,j} \lambda_{i,j}.$$

  Пример: $n_{i,j}$ — количество заболевших с диагнозом i в районе j за некоторый фиксированный промежуток времени.

Все три гипотезы проверяются с помощью критерия хи-квадрат.

Статистика критерия:

$$\xi^2 = n_{+,+} \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,+} n_{+,j}}{n_{+,+}}\right)^2}{n_{i,+} n_{+,j}}.$$

Если гипотеза H_I, H_{II} или H_{III} верна, то

$$\xi^2 \xrightarrow{d} \eta \sim \xi^2\left((k-1)(l-1)\right),$$

$$C_\alpha = (\xi^2_{1-\alpha}, \infty).$$

1.15 Линейная регрессия: постановка, теорема Гаусса-Маркова, базовые свойства, беды с регрессией.

1.15.1 Регрессионный анализ

- Данные: многомерная выборка $(X, Y)_{[n]}$, $X_i \in \mathbb{R}^k$, $Y_i \in \mathbb{R}$ и семейство функционалов $\{f(\cdot \mid \beta) \mid \beta \in B\}$.
- $Y \coloneqq Y_{[n]}$ — зависимая переменная, target.
- $X \coloneqq X_{[n]}$ — факторы, фичи.

Считаем, что $y_i \approx f(X_i, \beta_0)$, то есть $y_i = f(X_i, \beta_0) + \varepsilon_i$, где ε_i — шум, обладающий какими-то свойствами, например, $\mathbb{E}\varepsilon_i = 0$.

Хотим по (X, Y) найти наилучшую в каком-то смысле оценку β^* параметра β_0. Смысл задается функционалом качества $Q(\beta)$.

1.15.2 Линейная регрессия

Модель:

- $D = (X, Y)$
- $Y \in \mathbb{R}^n$, $Y = X\beta_0 + \varepsilon$,
- $X \in \mathbb{R}^{n \times k}$ фиксирована и известна, $\operatorname{rank} X = k$,
- $\beta_0 \in \mathbb{R}^k$ фиксирован и неизвестен,
- $\varepsilon \in \mathbb{R}^n$, $\mathbb{E}\varepsilon = 0$,
- $\mathbb{D}\varepsilon_i = \sigma^2$ — гомоскедастичностьб
- $\forall i \neq j\colon\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ — некоррелированность.

$$\operatorname{cov}(\varepsilon) = \sigma^2 I.$$

Ищем оценку β^* в виде $\arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n (y_i - X_{i,*}\beta)^2$. Она называется оценкой метода наименьших квадратов или МНК-оценкой.

thm (Гаусс-Марков)**.** *Если X имеет ранг k, ошибки гомоскедастичны и некоррелированы, то*

- β^* *— несмещенная оценка* β_0*,*
- $\operatorname{cov}(\beta^*) = \sigma^2(X^\top X)^{-1}$,
- β^* *— эффективная оценка в классе несмещенный линейных оценок*[2]*.*

Базовые свойства

- $\frac{\beta_i^* - \beta_i}{s(\beta_i^*)} \to \mathcal{N}(0,1)$, где $s^2(\beta_i^* = \hat\sigma \left(X^\top X\right)^{-1}_{i,i}$ и $\hat\sigma^2 = \frac{RSS}{(n-k)}$,
- $\hat\sigma^2$ является несмещенной и состоятельной оценкой σ^2.

Дисперсия target — сумма дисперсии предсказания и дисперсии ошибки:

$$\mathbb{D}y_i = \mathbb{D}\left(f(X_i \mid \beta_0) + \varepsilon_i\right) = \mathbb{D}\left(f(X_i \mid \beta_0)\right) + \mathbb{D}\varepsilon_i.$$

- $TSS = \sum(Y_i - \overline{Y})^2$ — total sum of squares (типа $n\mathbb{D}y_i$),
- $ESS = \sum(Y_i^* - \overline{Y})^2$ — explained sum of squares (типа $n\mathbb{D}\left(f(X_i \mid \beta_0)\right)$),
- $RSS = \sum(Y_i - Y_i^*)^2$ — residual sum of squares (типа $n\mathbb{D}\varepsilon_i$),

$$TSS = ESS + RSS.$$

Беды с регрессией

Беды с предположениями

- Неверная спецификация модели: Y не линейно выражаются через X. Смотрим на график остатков против предсказания. Надо чтобы было линейно. Можно переделать модель.
- Непостоянная дисперсия остатков: дисперсия (ковариация) зависит от X_i.
- Корреляция остатков: возникает, когда наблюдения близки во времени или пространстве.

[2] Линейные оценки — оценки вида $\beta = f(X)y$, оптимальность означает, что $\forall c \in \mathbb{R}^k : c^\top \operatorname{cov}(\beta^*)c \leqslant c^\top \operatorname{cov}(\beta^*)c$

Беды с данными

- Выбросы: X_i типичный, а Y_i нетипичный.
- Разбалансировка: X большой и Y большой.
- Мультиколлинеарность: k факторов, но $\operatorname{rank} X < k$, есть линейные зависимости или нестрогая $\operatorname{rank} X = k, \operatorname{cond} \gg 1$.