Семинар 2

Суслова Ирина

12 февраля 2024

1 Порядковые статистики

Определение 1.1.
Пусть дано вероятностное пространство $(\Omega, \mathcal{F}, \mathbb{P})$, на котором определены элементы выборки $X_1, \ldots, X_n$ и $x_i = X_i(\omega)$, $\omega \in \Omega$. Пронумеруем последовательность $\{x_i\}$ в порядке неубывания: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Тогда функция $X_{(k)}(\omega) = x_{(k)}$ называется k-ой порядковой статистикой

$$X_{(1)} = \min(X_1, \ldots, X_n),\ X_{(n)} = \max(X_1, \ldots, X_n).$$

Задача 1.1. Найти распределение k-ой порядковой статистики

Решение:

$$F_{X_k} = \mathbb{P}(X_{(k)} < x)$$

Рассмотрим событие $X_{(k)} < x$. Заметим, что k-ое по величине значение выборки меньше x тогда, когда как минимум k элементов выборки меньше x.

$$\mathbb{P}(X_k < x) = \mathbb{P}(\xi < x) = F_\xi(x)$$
$$\mathbb{P}(X_k \geq x) = \mathbb{P}(\xi \geq x) = 1 - F_\xi(x)$$

Тогда т.к выборка одинаковы распределены и независимы

$$F_{X_{(k)}}(x) = \sum_{m=k}^{n} C_n^m F_\xi^m(x)(1 - F_\xi(x))^{n-m}$$

Если у ξ есть плотность распределения, то и $X_{(k)}$ обладает плотностью распределения.

$$f_{X_{(k)}} = \lim_{\epsilon \to 0} \frac{F_{X_{(K)}}(x+\epsilon) - F_{X_{(k)}}(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\mathbb{P}(X_{(k)} \in [x, x+\epsilon))}{\epsilon}$$

Для вычисления предела достаточно разложить числитель по степеням ϵ и оставить только слагаемые пропорциональные ϵ.
Рассмотрим событие, что $X_{(k)}$ попало в необходимый интервал. Это может случиться тогда, когда

- либо k-1 значение меньше x, одно значение принадлежит интервалу $[x, x+\epsilon)$, остальные больше либо равны $x+\epsilon$
- k-2 значения меньше x, 2 элемента выборки попали в нужный интервалы, а все остальные $\geq x+\epsilon$
- и.т.д

Однако, нам подходит только первый случай. Рассмотрим второй случай. Пусть какие-то 2 значения попали в нужный нам интервал, допустим X_1, X_2. Рассмотрим вероятность

$$\mathbb{P}(X_1 \in [x, x+\epsilon), X_2 \in [x, x+\epsilon)) = \mathbb{P}(X_1 \in [x, x+\epsilon))\mathbb{P}(X_2 \in [x, x+\epsilon)) = f_\xi(x)\epsilon f_\xi(x)\epsilon \approx \epsilon^2$$

. Поэтому этот случай нас не интересует. Для случаев 3 и далее мы получим аналогичные результаты со степенью $\epsilon > 2$. таким образом, нам интересует только первый случай.

$$f_{X_{(k)}} = \lim_{\epsilon \to 0} \frac{C_n^1 \mathbb{P}(\xi \in [x, x+\epsilon)) C_{n-1}^{k-1} \mathbb{P}^{k-1}(\xi < x) C_{n-k}^{n-k} \mathbb{P}^{n-k}(\xi \geq x+\epsilon]}{\epsilon} = \\ = n f_\xi(x) C_{n-1}^{k-1} F_\xi^{k-1}(x)(1-F_\xi(x))^{n-k}$$

Стоит запомнить: Если $\xi \in U(0,1)$, то $X_{(k)} \in \text{Beta}(k, n-k+1)$

2 Проверка статистических гипотез

Гипотезой называется любое утверждени о распределеии случайной величиной
Примеры: Дана случайная величина ξ

1. Утверждение $\xi \in N(0,1)$ является гипотезой
2. Утверждение $\xi \in N(\mu, \sigma^2)$ также является гипотезой
3. Можно выдвинуть несколько гипотез $\text{H}_1 : N(0,1)$, $\text{H}_2 : N(0,2)$

Стоит отметить, что установить верность той или иной гипотезы невозможно. Задача теории проверки гипотез состоит в минимизации этих ошибок

Определение 2.1.
Множество $\Omega \subseteq \mathbb{R}^n$ всех значений выборки называется *выборочным пространством*

Пусть F_ξ - низвестная функция распределения случайной величины ξ, принадлежащая некоторому множеству априори допустимых распределений $\mathcal{F}$ (например, множество всех возможных распределений)

Определение 2.2.
Любое утверждение о принадлежности F_ξ какому-либо подмножеству$\mathcal{F}' \subset \mathcal{F}$называется *гипотезой* и обозначается, напримет, так:

$$\mathrm{H}: \ F_\xi \in \mathcal{F}' \subset \mathcal{F}$$

Определение 2.3.
Если $\mathcal{F}'$ состоит из одного элемента, то гипотеза H называется простой. Иначе - сложной

Например, гипотеза $\xi \in N(0,1)$ является простой, а гипотеза $\mathrm{H}_1 : N(\mu, \sigma^2)$ - сложной.
На проверку может быть выдвинуто несколько гипотез, при этом некоторые из них могут быть простыми, а некоторые - сложным. Бывает, что выдвинута одна гипотеза. В этом случа на самом деле подразумевается, что гипотез две, просто вторая гипотеза по умолчанию дополняет множество из основной гипотезы до множества всех априори допустимых гипотез и ее не пишут.

Определение 2.4.
Пусть выдвинуто r гипотез $\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_r$. *Статистическим критерием* называется измеримая функция $\delta : \Omega \to \{H_1, \ldots, H_r\}$.
Где Ω - выборочное пространство

То есть каждой выборке статистический критерий сопоставляет какую-то гипотезу. Задание измеримой функции δ равносильно разбиению выборочного пространства на r непересекающихся подмножеств $\Omega_1, \ldots, \Omega_r$ таких, что при попадании выборки x в область Ω_1 принимается гипотеза H_1, при попадании в область Ω_2 принимается гипотеза H_2 и так далее

Определение 2.5.
Пусть дано r простых гипотез H_1,$\mathrm{H}_2, \ldots$,H_r. Вероятность

$$\alpha_i(\delta) = \mathbb{P}_i(X \notin \Omega_i) = \mathbb{P}_i(\delta(X) \neq \mathrm{H}_i)$$

называется *вероятность ошибки i-го рода.* Индекс i под символом вероятноти означает, что вероятность подсчитывается в случае, когда выборка распределена по закону гипотезы H. Другими словами, вероятность ошибки i-го рода - это вероятность отклонить i-ю гипотезу, если она на самом деле верна

Определение 2.6.
В случае двух гипотез H_1 и H_2 множество Ω_2 называют *критической областью* гипотезы H_1

В случае двух простых гипотез вероятность ошибки первого рода обозначается символом α, а вероятность ошибки второго рода символом β. Гипотеза H_1 принято называть основной гипотезой, а H_2 - альтернативной гипотезой или просто альтернативой. При этом ошибка первого рода $\alpha(\delta) =$

$\mathbb{P}_1(\delta(X) \neq \mathrm{H}_1) = \mathbb{P}_1(\delta(X) = \mathrm{H}_2)$, то есть ошибка первого рода-вероятность принять альтернативу (=отвергнуть основную гипотезу), если на самом деле верна. Ошибка второго рода $\beta(\delta) = \mathbb{P}_2(\delta(X) \neq \mathrm{H}_2) = \mathbb{P}_2(\delta(X) = \mathrm{H}_1)$ - вероятность принять основную гипотезу, если на самом деле верна альтернатива

Определение 2.7.
Если $\alpha(\delta) \leq \alpha_0$, то говорят, что критерий δ имеет *уровень значимости* α_0.

Определение 2.8.
Пусть F(x) - некоторая функция распределения. Тогда любое решение уравнения $F(x) = p \in (0,1)$*, если оно существует, называется p-квантилем.* Если решение не существует, то p-квантилем для непрерывной слева функции F(x) называется $x = \sup\{y : \ F(y) \leq p\}$.

3 Критерий согласия Колмогорова

Условия.
Дана выборка $x_1, x_2, \ldots x_n$ и непррывная функция $F(x)$. Выдвинута простая гипотеза:

$$\mathrm{H}_1 : \ F_\xi(x) = F(x)$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.
Алгоритм

1. Составить эмпирическую функцию распределения на данной выборке
2. Вычислить реализацию статистики Колмогорова-Смирнова:

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$$

3. Выбрать в качестве критческой области $\Omega_2 = \{x \in \Omega : \ D_n \geq t_\alpha\}$, и найти t_α из условия на уровень значимости:

$$\mathbb{P}_1(X \in \Omega_2) = \mathbb{P}_1(D_n \geq t_\alpha) = \alpha$$

, то есть найти $(1-\alpha)$-квантиль распределения случайной величины D_n.

4. Принять решение по следующей схеме:

$$\mathrm{H}_1 \text{ отвергается} \iff D_n \geq t_\alpha$$

Замечание 1. Статистика Колмогорова-Смирнова равна максимальному отклонению эмпирической функции распределения, построенной по выборке, от гипотетической F(x). Согласно теореме Гливенко, если гипотеза H_1 верна, то это отклонение почти наверное стремится к нулю с ростом объема

выборки. Таким образом если гипотеза верна, то мы ожидаем, что величина D_n будет небольшой. Поэтому в качестве критической области выбираются достаточно большие значения этой статистики, а граница для этих значений определяется из условия на уровень значимости.

Замечание 2. Критерий используется только для непрерывных функций F(x).

Замечание 3. Если значения n достаточно большие ($n \geq 20$), то благодаря теореме Колмогорова можно воспользоваться приближением

$$\mathbb{P}_1(D_n \geq t_\alpha) = \mathbb{P}_1(\sqrt{n}D_n \geq \sqrt{n}t_\alpha) \approx 1 - \mathrm{K}(\sqrt{n}t_\alpha) = \alpha$$

и находить t_α через квантиль распределения Колмогорова

Замечание 4. Если мы отклонили гипотезу H_1, то мы можем гарантировать, что вероятность нашей ошибки не превышает уровень значимости α. Если же гипотезу мы не отклоняем, то утверждать, что она верная, нельзя, т.к. мы не знаем вероятность ошибки такого утверждения. Поэтому в случае неотклонения гипотезы H_1 мы говорим: «данные гипотезе не противоречат».

Замечание 5. На практике, зная выборку, статистику D_n рассчитывают по формулам

$$D_n = \max\{D_n^+, D_n^-\},$$

$$D_n^+ = \max_{1\leq k\leq n}\left(\frac{k}{n} - F(X_{(k)})\right). \; D_n^- = \max_{1\leq k\leq n}\left(F(X_{(k)}) - \frac{k-1}{n}\right).$$

Задача 3.1 (Задача 1)**.**
Дана выборка $x = (0.1, 0.9, 0.3, 0.4, 0.7)$ из непрерывного распределения. На уровне значимости $\alpha = 0.05$ проверить гипотезу о равномерном на отрезке (0,1) распределении измеряемой случайной величины:

$$\mathrm{H}_1 : \; U(0,1)$$

Решение:
Функция распределения $U(0,1)$ есть

$$F(x) = x, \;\; x \in (0,1).$$

Вычислим реализацию статистики D_n. Для этого воспользуемся формулами

$$D_n = \max\{D_n^+, D_n^-\},$$

$$D_n^+ = \max_{1\leq k\leq n}\left(\frac{k}{n} - F(X_{(k)})\right). \; D_n^- = \max_{1\leq k\leq n}\left(F(X_{(k)}) - \frac{k-1}{n}\right).$$

В нашем случае

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}) = (0.1, 0.3, 0.4, 0.7, 0.9),$$

$$D_n^+ = \max\left\{\frac{1}{5} - \frac{1}{10}, \frac{2}{5} - \frac{3}{10}, \frac{3}{5} - \frac{4}{10}, \frac{4}{5} - \frac{7}{10}, 1 - \frac{9}{10}\right\} = \frac{1}{5},$$

$$D_n^- = \max\left\{\frac{1}{10}, \frac{3}{10} - \frac{1}{5}, \frac{4}{10} - \frac{2}{5}, \frac{7}{10} - \frac{3}{5}, \frac{9}{10} - \frac{4}{5}\right\} = \frac{1}{10},$$

$$D_n = \max\left\{\frac{1}{5}, \frac{1}{10}\right\} = \frac{1}{5}$$

Теперь заглянем в таблицу квантилей и найдем квантиль $t_\alpha = 0.563$. Видим, что $D_n < t_\alpha$, поэтому ответ: "Данные гипотезе не противоречат"