

Теория вероятностей и математическая статистика, Коллоквиум IV

Версия от 09.06.2021 14:17

Содержание

1.	Выборка, оценка, статистика. Несмещенность, состоятельность, асимптотическая нормальность и эффективность оценок. Пример отсутствия несмещенной оценки. Отсутствие эффективной оценки в классе всех оценок. Единственность эффективной оценки. Состоятельность асимптотической нормальной оценки.	2
1.1.	Выборка, оценка, статистика.	2
1.2.	Несмещенность, состоятельность, асимптотическая нормальность и эффективность оценок.	2
1.3.	Пример отсутствия несмещенной оценки.	2
1.4.	Отсутствие эффективной оценки в классе всех оценок.	2
1.5.	Состоятельность асимптотической нормальной оценки.	2
2.	Метод моментов и его состоятельность. Метод максимального правдоподобия. Энтропия и состоятельность оценки максимального правдоподобия.	2
3.	Информация Фишера и неравенство Рао-Крамера. Критерий равенства в неравенстве Рао-Крамера.	2
4.	Доверительные интервалы. Различные методы построения доверительных интервалов (с помощью неравенств на вероятность больших отклонений, с помощью центральной статистики, с помощью асимптотически нормальной оценки). Примеры.	2
4.1.	Доверительные интервалы	2
4.2.	Различные методы построения доверительных интервалов (с помощью неравенств на вероятность больших отклонений, с помощью центральной статистики, с помощью асимптотически нормальной оценки)	3
5.	Построение точных доверительных интервалов для параметров нормального распределения.	5
6.	Проверка гипотез. Ошибки 1-го и 2-го рода. Уровень значимости и мощность статистического критерия. Пример построения критерия с помощью доверительного интеграла. Нижняя оценка суммы вероятностей ошибок 1-го и 2-го рода.	5
7.	Теорема Неймана-Пирсона и пример её применения.	5
7.1.	Теорема Неймана-Пирсона.	5
7.2.	Пример применения теоремы Неймана-Пирсона.	5
8.	Эмпирическая функция распределения. Теорема Гливенко-Кантелли.	6

1. Выборка, оценка, статистика. Несмещенность, состоятельность, асимптотическая нормальность и эффективность оценок. Пример отсутствия несмещенной оценки. Отсутствие эффективной оценки в классе всех оценок. Единственность эффективной оценки. Состоятельность асимптотической нормальной оценки.

1.1. Выборка, оценка, статистика.

1.2. Несмещенность, состоятельность, асимптотическая нормальность и эффективность оценок.

1.3. Пример отсутствия несмещенной оценки.

1.4. Отсутствие эффективной оценки в классе всех оценок.

1.5. Состоятельность асимптотической нормальной оценки.

2. Метод моментов и его состоятельность. Метод максимального правдоподобия. Энтропия и состоятельность оценки максимального правдоподобия.

3. Информация Фишера и неравенство Рао-Крамера. Критерий равенства в неравенстве Рао-Крамера.

4. Доверительные интервалы. Различные методы построения доверительных интервалов (с помощью неравенств на вероятность больших отклонений, с помощью центральной статистики, с помощью асимптотически нормальной оценки). Примеры.

4.1. Доверительные интервалы

Знать, что оценка $\hat{\theta}_n(X)$ состоятельна (сходится по вероятности к θ) это, конечно, круто, но особо много информации о ней нам не даёт. Нам хотелось бы знать как быстро она куда-то там сходится – хотим для фиксированного $\alpha \in (0, 1)$ и фиксированного $\varepsilon > 0$ знать такой номер n , что $P_\theta(|\hat{\theta}_n(X) - \theta| < \varepsilon) > 1 - \alpha$.

Определение. $(\hat{\theta}_1(X), \hat{\theta}_2(X))$ — доверительный интервал уровня доверия $1 - \alpha$, если

$$P_\theta(\theta \in (\hat{\theta}_1(X), \hat{\theta}_2(X))) \geq 1 - \alpha$$

$$P_\theta(\hat{\theta}_1(X) \leq \theta \leq \hat{\theta}_2(X)) \geq 1 - \alpha$$

Определение. Последовательность оценок $\hat{\theta}_1^n(X), \hat{\theta}_2^n(X)$ образует асимптотический доверительный интервал, если $\liminf_{n \rightarrow \infty} P_\theta(\hat{\theta}_1^n(X) \leq \theta \leq \hat{\theta}_2^n(X)) \geq 1 - \alpha$

Пример. Пусть есть выборка из случайных величин с нормальным распределением $X_j \sim \mathcal{N}(\theta, 1)$.

Знаем, что $\bar{X}_n = \frac{X_1 + \dots + X_n}{n} \xrightarrow{P_\theta} \theta$ (ЗБЧ) — среднее хорошо приближает θ .

Посмотрим на разность эмпирического среднего и реальной θ : $\frac{X_1 + \dots + X_n}{n} - \theta = \frac{\overbrace{(X_1 - \theta)}^{\sim \mathcal{N}(0,1)} + \dots + \overbrace{(X_n - \theta)}^{\sim \mathcal{N}(0,1)}}{n} \sim \mathcal{N}(0, \frac{1}{n})$
 $\Rightarrow \sqrt{n}(\frac{X_1 + \dots + X_n}{n} - \theta) \sim \mathcal{N}(0, 1)$

Теперь по таблице значений функции распределения нормального закона найдём квантили $z_{\frac{\alpha}{2}}$ и $z_{1-\frac{\alpha}{2}}$: $\Phi(z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}, \Phi(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

\Rightarrow

$$P_\theta(z_{\frac{\alpha}{2}} \leq \sqrt{n}(\frac{X_1 + \dots + X_n}{n} - \theta) \leq z_{1-\frac{\alpha}{2}}) = \Phi(z_{1-\frac{\alpha}{2}}) - \Phi(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

$$P(\frac{z_{\frac{\alpha}{2}}}{\sqrt{n}} - \bar{X}_n \leq -\theta \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} - \bar{X}_n) = 1 - \alpha$$

$$P(\bar{X}_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \theta \leq \bar{X}_n - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}) = 1 - \alpha$$

Заметим, что мы взяли симметричный интервал: $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$. В таком случае наш интервал принимает вид:

$$(\bar{X}_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}). \text{ В таком случае длина этого интервала равна } O(\frac{1}{\sqrt{n}})$$

Но зачем мы решили взять симметричный интервал? Вспомним, что мы от него хотим: минимальной длины. А какой интервал на графике нормального распределения будет захватывать нужную площадь и при этом быть самым коротким среди всех? Правильно, симметричный с центром в пике колокола нормального распределения.

4.2. Различные методы построения доверительных интервалов (с помощью неравенств на вероятность больших отклонений, с помощью центральной статистики, с помощью асимптотически нормальной оценки)

1. Неравенства Чебышёва или Чернова

$$X_1, \dots, X_n \sim \text{Bern}(\theta), P(X_j = 1) = \theta$$

Чебышёв:

$$P_\theta(|\bar{X}_n - \theta| \geq \varepsilon) \leq \frac{\mathbb{D}X_1}{n\varepsilon^2} = \frac{\theta(1-\theta)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2} = \alpha \implies \varepsilon = \frac{1}{\sqrt{2n\alpha}} \implies P_\theta(\bar{X}_n - \frac{1}{\sqrt{2n\alpha}} < \theta < \bar{X}_n + \frac{1}{\sqrt{2n\alpha}}) \geq 1 - \alpha.$$

$$\text{Чернов: } P_\theta(|\bar{X}_n - \theta| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{4}} = \alpha$$

$$-\frac{n\varepsilon^2}{4} = \ln \frac{\alpha}{2}$$

$$\varepsilon = 2\sqrt{-\frac{\ln \frac{\alpha}{2}}{n}}$$

$$\implies P_\theta(\bar{X}_n - 2\sqrt{-\frac{\ln \frac{\alpha}{2}}{n}} < \theta < \bar{X}_n + 2\sqrt{-\frac{\ln \frac{\alpha}{2}}{n}}) \geq 1 - \alpha$$

Заметим, что в обеих оценках мы получили, что длина интервала равна $O(\frac{1}{\sqrt{n}})$, но несложно заметить, что константа Чернова значительно лучше, чем у Чебышёва.

2. Метод центральной статистики

Определение. $V(X, \theta)$ называется центральной статистикой, если:

$$(a) \text{ её распределение не зависит от } \theta: P_\theta(V(X, \theta) \leq t) = F(t)$$

$$(b) \forall X: \theta \mapsto V(X, \theta) \text{ — монотонная}$$

Пусть у нас есть такая статистика. Вопрос: как с её помощью строить доверительные интервалы? Предельно просто: подберём числа t_1 и t_2 таким образом, чтобы $P_\theta(t_1 \leq V(X, \theta) \leq t_2) \geq 1 - \alpha$. Мы можем так сделать, потому что распределение V не зависит от θ . Теперь поскольку при любом X наша функция монотонна, то данная оценка равносильна тому, что $P_\theta(\hat{\theta}_1(X) \leq \theta \leq \hat{\theta}_2(X)) \geq 1 - \alpha$ — чисто из-за монотонности по θ .

Пример. $X_j \sim \mathcal{U}(0, \theta) \implies \theta^{-1}X_j \sim \mathcal{U}(0, 1)$. Это уже центральная статистика, однако она зависит всего от одного элемента выборки. Рассмотрим $X_{(n)} = \max_{1 \leq j \leq n} X_j$: $P_\theta(\theta^{-1}X_{(n)} \leq t) = P_\theta(\max_{1 \leq j \leq n} \theta^{-1}X_j \leq t) = \prod_{j=1}^n P_\theta(\underbrace{\theta^{-1}X_j}_{\sim \mathcal{U}(0,1)} \leq t) = t^n$

Теперь грубо попробуем оценить, куда там наша статистика попадает:

$$P_\theta(\underbrace{t}_{t_1} \leq \theta^{-1}X_{(n)} \leq \underbrace{1}_{t_2}) = 1 - t^n = 1 - \alpha \implies t = \alpha^{\frac{1}{n}}$$

Теперь попробуем вытащить отсюда θ :

$$P_\theta(\alpha^{\frac{1}{n}} \leq \theta^{-1}X_{(n)} \leq 1) = 1 - \alpha$$

$$P_\theta(\frac{\alpha^{\frac{1}{n}}}{X_{(n)}} \leq \theta^{-1} \leq \frac{1}{X_{(n)}}) = 1 - \alpha$$

$$P_\theta(\underbrace{X_{(n)}}_{\hat{\theta}_1(X)} \leq \theta \leq \frac{X_{(n)}}{\underbrace{\alpha^{\frac{1}{n}}}_{\hat{\theta}_2(X)}}) = 1 - \alpha$$

Теперь посмотрим на длину полученного доверительного интервала:

$$(\alpha^{-\frac{1}{n}} - 1)X_{(n)}$$

Что мы можем сказать про $\alpha^{-\frac{1}{n}} - 1$? Разложим это дело по Тейлору:

$$\alpha^{-\frac{1}{n}} - 1 \sim e^{-\frac{\ln \alpha}{n}} - 1 \sim \frac{-\ln \alpha}{n} = \underline{O}\left(\frac{1}{n}\right) \rightarrow \infty$$

Получается длина доверительного интервала с ростом количества элементов выборки стремится к нулю. Получается мы построили что-то более менее разумное.

Часто в роли центральной статистики можно взять следующую лабуду: $V(X, \theta) = -\sum_{j=1}^n \ln F_{\theta}(X_j)$ — это сумма независимых распределений, поэтому достаточно показать что одно не зависит от θ — тогда в силу независимости сумма тоже будет не зависеть от θ :

$P_{\theta}(-\ln F_{\theta}(X_j) \leq) = P_{\theta}(F_{\theta}(X_j) \geq e^{-t}) = P_{\theta}(X_j \geq F_{\theta}^{-1}(e^{-t})) = 1 - F_{\theta}(F_{\theta}^{-1}(e^{-t})) = 1 - e^{-t}$, а это экспоненциальное распределение. Сумма экспоненциальных распределений это Гамма распределение $\implies V(X, \theta) = \Gamma(n, 1)$

3. Построение асимптотических доверительных интервалов

Пусть у нас есть $\hat{\theta}_n(X)$ — асимптотически нормальная оценка θ с асимптотической дисперсией $\sigma^2(\theta)$. Это значит, что

$$\frac{\sqrt{n}(\hat{\theta}_n(X) - \theta)}{\sigma(\theta)} \xrightarrow{d_{\theta}} Z \sim \mathcal{N}(0, 1)$$

Теперь мы хотим получить доверительный интервал. Если бы у нас $\sigma(\theta)$ была константой, то мы могли бы уже привычно взять там квантили нормального распределения, туды сюды и получить интервал:

$$P_{\theta}(t_1 \leq \frac{\sqrt{n}(\hat{\theta}_n(X) - \theta)}{\sigma(\theta)} \leq t_2) \rightarrow \Phi(t_2) - \Phi(t_1) = 1 - \alpha.$$

Тогда мы могли бы просто взять такие $\Phi(t_2) = 1 - \frac{\alpha}{2}$ и $\Phi(t_1) = \frac{\alpha}{2}$ и получить, что

$$P_{\theta}(\hat{\theta}_n(X) - \frac{t_{1-\frac{\alpha}{2}}\sigma(\theta)}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n(X) + \frac{t_{1-\frac{\alpha}{2}}\sigma(\theta)}{\sqrt{n}}) \rightarrow 1 - \alpha$$

Но тут есть проблема — у нас слева и справа есть $\sigma(\theta)$ в числителе, что совершенно ломает корректность статистики, мы ведь хотим чтобы штуки слева и справа от θ в неравенстве не зависели от θ . Как это решать? Очень просто, перейти от $\sigma(\theta)$ к состоятельной оценке $\sigma(\theta)$. Возможны следующие случаи:

(a) σ — непрерывная функция

Тогда $\sigma(\hat{\theta}_n(X)) \xrightarrow{P_{\theta}} \sigma(\theta)$ и мы можем везде в наших рассуждениях заменить $\sigma(\theta)$ на $\sigma(\hat{\theta}_n(X))$ и сходимость сохранится:

$$P_{\theta}(\hat{\theta}_n(X) - \frac{t_{1-\frac{\alpha}{2}}\sigma(\hat{\theta}_n(X))}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n(X) + \frac{t_{1-\frac{\alpha}{2}}\sigma(\hat{\theta}_n(X))}{\sqrt{n}}) \rightarrow 1 - \alpha$$

(b) Изначально было ЦПТ

$$\hat{\theta}_n(X) = \overline{X_n}, \sigma^2(\theta) = \mathbb{D}_{\theta}X$$

В таком случае мы можем использовать выборочную дисперсию в качестве состоятельной оценки дисперсии:

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X_n})^2 \text{ — выборочная дисперсия}$$

(c) Можно поправить нашу асимптотическую дисперсию: подобрать такую функцию φ , что

$$\sqrt{n}(\varphi(\hat{\theta}_n(X)) - \varphi(\theta)) \rightarrow \underbrace{\mathcal{N}(0, 1)}_{=\varphi'(\theta) \cdot \mathcal{N}(0, \sigma^2(\theta))} \implies \varphi'^2(\theta) \sigma^2(\theta) = 1$$

5. Построение точных доверительных интервалов для параметров нормального распределения.
6. Проверка гипотез. Ошибки 1-го и 2-го рода. Уровень значимости и мощность статистического критерия. Пример построения критерия с помощью доверительного интеграла. Нижняя оценка суммы вероятностей ошибок 1-го и 2-го рода.
7. Теорема Неймана-Пирсона и пример её применения.

7.1. Теорема Неймана-Пирсона.

Пусть гипотеза H_0 утверждает, что плотность выборки – это f_0 , а альтернативная гипотеза H_1 утверждает, что плотность выборки – это f_1 .

Предположим, что $\forall \alpha \in [0, 1] \exists t := t(\alpha) : P_0(f_1(x) \geq t f_0(x)) = \alpha$.

Теорема (Неймана-Пирсона). В такой постановке наиболее мощный критерий уровня значимости α имеет вид $K_{t(\alpha)} := \{f_1(x) \geq t(\alpha)f_0(x)\}$.

Доказательство. Пусть S – тоже критерий уровня значимости α : $P_0(X \in S) \leq \alpha = P_0(X \in K_{t(\alpha)})$. Хотим сравнить $P_1(X \in K_{t(\alpha)}) - P_1(X \in S)$. Хотим, чтобы это было больше либо равно нулю. Это и будет означать, что у нас критерий наиболее мощный.

$$P_1(X \in K_{t(\alpha)}) - P_1(X \in S) = \int_{K_{t(\alpha)}} f_1 dx - \int_S f_1 dx = [\text{можем выкинуть пересечение, так как на пересечении эти интегралы просто сократятся}] = \int_{K_{t(\alpha)} \setminus S} f_1 dx - \int_{S \setminus K_{t(\alpha)}} f_1 dx.$$

Заметим, что на $S \setminus K_{t(\alpha)}$ выполнено $f_1 < t(\alpha)f_0$, так как это взято из дополнения к $K_{t(\alpha)}$, где по условию выполняется $f_1(x) \geq t(\alpha)f_0(x)$. Поэтому имеем: $\int_{K_{t(\alpha)} \setminus S} f_1 dx - \int_{S \setminus K_{t(\alpha)}} f_1 dx \geq t(\alpha) \int_{K_{t(\alpha)} \setminus S} f_0 dx - t(\alpha) \int_{S \setminus K_{t(\alpha)}} f_0 dx = [\text{снова добавим пересечение и вынесем } t(\alpha)] = t(\alpha) \cdot \left(\int_{K_{t(\alpha)}} f_0 dx - \int_S f_0 dx \right) = t(\alpha) \cdot (P_0(X \in K_{t(\alpha)}) - P_0(X \in S)) \geq 0$ из построения критерия S ($P_0(X \in S) \leq \alpha = P_0(X \in K_{t(\alpha)})$).

Получили: $P_1(X \in K_{t(\alpha)}) - P_1(X \in S) \geq 0$, что и требовалось доказать. ■

7.2. Пример применения теоремы Неймана-Пирсона.

Пример. Пусть у нас выборка из нормального закона $N(\theta, 1)$. Пусть наша гипотеза H_0 говорит, что $\theta = \theta_0$, а альтернативная гипотеза H_1 говорит, что $\theta = \theta_1 > \theta_0$.

$$f_1(X) = \frac{1}{\sqrt{2\pi}^n} \cdot \exp\left(-\frac{1}{2} \sum_{j=1}^n (X_j - \theta_1)^2\right)$$

$$f_0(X) = \frac{1}{\sqrt{2\pi}^n} \cdot \exp\left(-\frac{1}{2} \sum_{j=1}^n (X_j - \theta_0)^2\right)$$

Зададим критерий K_t из теоремы Неймана-Пирсона (ничего в 0 не обращается – сразу можем поделить):

$$K_t = \left\{ \frac{f_1}{f_0} \geq t \right\} = \left\{ \exp\left(\frac{1}{2} \sum_{j=1}^n [(X_j - \theta_0)^2 - (X_j - \theta_1)^2]\right) \geq t \right\} = [\text{логарифмируем, раскрываем скобки, умножаем на два}] = \left\{ \sum_{j=1}^n [2X_j(\theta_1 - \theta_0) + n(\theta_0^2 - \theta_1^2)] \geq 2 \ln t \right\} = \left\{ (\theta_1 - \theta_0) \overline{X}_n \geq \frac{\ln t}{n} - \frac{(\theta_0^2 - \theta_1^2)}{2} \right\} = [\text{по условию } \theta_1 > \theta_0 \Rightarrow \text{поделим}]$$

$$= \left\{ \overline{X}_n \geq \frac{\frac{\ln t}{n} - \frac{(\theta_0^2 - \theta_1^2)}{2}}{\theta_1 - \theta_0} \right\}$$

Таким образом пришли к тому, что $K_t = \left\{ \frac{f_1}{f_0} \geq t \right\}$ равносильно множеству $\tilde{K}_s = \{\overline{X_n} \geq s\}$. Равносильно в том смысле, что для каждого t мы можем подобрать $s(t)$, что множество K_t совпадает с $\tilde{K}_{s(t)}$. Теперь будем искать критические множества именно в таком виде (для удобства).

Должно выполняться: $P_0(X \in K_t) = \alpha \Leftrightarrow P_0(X \in \tilde{K}_{s(t)}) = \alpha$. А что это за вероятности? Это вероятность $P_{\theta_0}(\overline{X_n} \geq s) = \alpha$. То есть, $P_{\theta_0}(\sqrt{n}(\overline{X_n} - \theta_0) \geq \sqrt{n}(s - \theta_0)) = \alpha$, где $\sqrt{n}(\overline{X_n} - \theta_0) \sim N(0, 1)$, поэтому тут просто написано, что $1 - \Phi(\sqrt{n}(s - \theta_0)) = \alpha$.

Значит, выбираем квантиль нормального закона уровня $1 - \alpha$: $Z_{1-\alpha} = \sqrt{n}(s - \theta_0) \Rightarrow s = \theta_0 + \frac{Z_{1-\alpha}}{\sqrt{n}}$. Выразили s .

Таким образом, наше критическое множество $\left\{ \overline{X_n} \geq \theta_0 + \frac{Z_{1-\alpha}}{\sqrt{n}} \right\}$. Это критерий уровня значимости α .

Теперь посчитаем мощность (это же самый мощный критерий):

$$P_{\theta_1} \left(\overline{X_n} \geq \theta_0 + \frac{Z_{1-\alpha}}{\sqrt{n}} \right) = P_{\theta_1} (\sqrt{n}(\overline{X_n} - \theta_1) \geq \sqrt{n}(\theta_0 - \theta_1) + Z_{1-\alpha}) = 1 - \Phi(\sqrt{n}(\theta_0 - \theta_1) + Z_{1-\alpha}).$$

Заметим, что если объём выборки n устремить к бесконечности, то точка, в которой мы берём Φ стремится к минус бесконечности (так как $(\theta_0 - \theta_1) < 0$ по условию), поэтому мощность стремится к 1.

По теореме Неймана-Пирсона выписанная мощность максимальна.

8. Эмпирическая функция распределения. Теорема Гливленко-Кантелли.