Лекция 7 (от 14.10)

Методы поиска доверительных интервалов

1. Метод центральной функции

Пусть $G(X,\theta)$ — функция, распределение которой известно и не зависит от θ (*центральная функция*). Возьмем $\alpha_1, \alpha_2 \in (0,1)$ т. ч. $\alpha_2 - \alpha_1 = \alpha$ и g_j — α_j-квантиль распределения $G(X,\theta)$. Тогда $S(X) = \{\theta \in \Theta | g_1 \leqslant G(X,\theta) \leqslant g_2\}$ — доверительная область уровня доверия α.
Действительно, $P_\theta(\theta \in S(X)) = P_\theta(g_1 \leqslant G(X,\theta) \leqslant g_2) = \alpha_2 - \alpha_1 = \alpha.$

Пример: $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, σ известно. Построить точные доверительные интервалы для θ.

$\triangle$ Заметим, что $X_i - \theta \sim \mathcal{N}(0, \sigma^2)$, следовательно, $\overline{X} - \theta \sim \mathcal{N}(0, \frac{\sigma^2}{n})$.
$G(X,\theta) = \sqrt{n}\dfrac{\overline{X}-\theta}{\sigma} \sim \mathcal{N}(0,1)$ — центральная функция. Будем обозначать через z_p p-квантили распределения $\mathcal{N}(0,1)$. Тогда

$$P_\theta\left(-z_{\frac{1+\alpha}{2}} \leqslant \sqrt{n}\frac{\overline{X}-\theta}{\sigma} \leqslant z_{\frac{1+\alpha}{2}}\right) = \alpha \implies P_\theta\left(\overline{X} - \frac{z_{\frac{1+\alpha}{2}}\sigma}{\sqrt{n}} \leqslant \theta \leqslant \overline{X} + \frac{z_{\frac{1+\alpha}{2}}\sigma}{\sqrt{n}}\right) = \alpha.$$

Ответ: $\left(\overline{X} \pm \dfrac{z_{\frac{1+\alpha}{2}}\sigma}{\sqrt{n}}\right)$.

Пусть $\alpha = 0.95 \implies z_{\frac{1+\alpha}{2}} = z_{0.975} \approx 1.96 \approx 2$. $n = 100, \overline{x} = 5, \sigma = 1$. Тогда реализация интервала $(5 \pm 2/10) = (4.8, 5.2)$.

2. Асимптотические доверительные интервалы

> **Определение:** Пусть $X = (X_1, X_2, \ldots)$ — выборка неограниченного размера из распределения $P \in \{P_\theta | \theta \in \Theta\}$. Последовательность пар статистик $(T_1^{(n)}(X_1,\ldots,X_n), T_2^{(n)}(X_1,\ldots,X_n))$ называется *асимптотическим доверительным интервалом* уровня доверия α, если
>
> $$\forall \theta \in \Theta \liminf_{n\to\infty} P_\theta(T_1^{(n)}(X_1,\ldots,X_n) \leqslant \theta \leqslant T_2^{(n)}(X_1,\ldots,X_n)) \geqslant \alpha.$$
>
> Он называется *точным*, если
>
> $$\forall \theta \in \Theta \lim_{n\to\infty} P_\theta(T_1^{(n)} \leqslant \theta \leqslant T_2^{(n)}) = \alpha.$$

Метод построения асимптотического доверительного интервала:

1. Пусть $\widehat{\theta}$ — а.н.о θ с асимпт. дисперсией $\sigma^2(\theta)$.

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)).$$

2. Поделим все на $\sigma(\theta)$:

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

Из теоремы Александрова

$$P_\theta\left(\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma(\theta)} \leqslant z_{\frac{1+\alpha}{2}}\right) \to \alpha.$$

Проблема: $\sigma(\theta)$ может плохо зависеть от θ.

3. Пусть $\widehat{\sigma}$ — состоятельная оценка $\sigma(\theta)$. Тогда

$$\sqrt{n}\frac{\widehat{\theta}-\theta}{\widehat{\sigma}} = \underbrace{\sqrt{n}\frac{\widehat{\theta}-\theta}{\sigma(\theta)}}_{\xrightarrow{d_\theta}\mathcal{N}(0,1)} \cdot \underbrace{\frac{\sigma(\theta)}{\widehat{\sigma}}}_{\xrightarrow{P_\theta}1 \text{ (th о насл. сх-тей)}}.$$

По лемме Слуцкого $\sqrt{n}\frac{\widehat{\theta}-\theta}{\widehat{\sigma}} \xrightarrow{d_\theta} \mathcal{N}(0,1)$.

4. $P_\theta\left(\frac{\sqrt{n}(\hat{\theta}-\theta)}{\widehat{\sigma}} \leqslant z_{\frac{1+\alpha}{2}}\right) \to \alpha$. Получаем интервал $\left(\widehat{\theta} \pm \frac{z_{\frac{1+\alpha}{2}}\widehat{\sigma}}{\sqrt{n}}\right)$ — точный асимптотический доверительный интервал уровная доверия α.

5. Откуда взять $\widehat{\sigma}$?
Если $\sigma(\theta)$ непрерывна, то по теореме о наследовании сходимостей $\widehat{\sigma} = \sigma(\widehat{\theta})$ — состоятельная оценка $\sigma(\theta)$.

Пример:

1. $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2),\ \sigma$ неизвестна. Построить асимптотический доверительный интервал уровня доверия α для θ. $\triangle\quad \overline{X}$ — а.н.о θ с асимпт. дисперсией σ^2. S — состоятельная оценка θ. Получаем интервал $\left(\overline{X} \pm z_{\frac{1+\alpha}{2}}\frac{S}{\sqrt{n}}\right)$. $\quad\square$
2. $X_1, \ldots, X_n \sim Pois(\theta)$. Построить асимптотический доверительный интервал уровня доверия α для θ.
$\triangle\quad \overline{X}$ — а.н.о θ с асимпт. дисперсией $\sigma^2(\theta) = \theta$. $\sqrt{\overline{X}}$ — состоятельная оценка $\sigma(\theta) = \sqrt{\theta}$. Получаем интервал $\left(\overline{X} \pm z_{\frac{1+\alpha}{2}}\sqrt{\frac{\overline{X}}{n}}\right)$. $\quad\square$

Замечание: При $n = 30$ условие ЦПТ применимо с хорошей точностью. Поэтому при $n \geqslant 30$ имеет смысл пользоваться асимптотическими доверительными интервалами.

3.2. Точые доверительные интервалы в нормальной модели

Пусть $X = (X_1, \ldots, X_n) \sim \mathcal{N}(a, \sigma^2)$.

1. Интервал для a, если σ известна

Уже получили: $\left(\overline{X} \pm z_{\frac{1+\alpha}{2}}\frac{S}{\sqrt{n}}\right)$.

2. Интервал для σ, если a известно

$\dfrac{X_i - \theta}{\sigma} \sim \mathcal{N}(0,1)$

$G(X,\theta) = \sum\limits_{i=1}^{n} \left(\dfrac{X_i - a}{\sigma}\right)^2 \sim \chi^2_n$ — центральная функция *(распределение хи-квадрат с n степенями свободы)*

$$P_\theta\left(\chi^2_{n,\frac{1-\alpha}{2}} \leqslant \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - a)^2 \leqslant \chi^2_{n,\frac{1+\alpha}{2}}\right) = \alpha$$

Получаем интервал $\left(\sqrt{\dfrac{\sum(X_i-a)^2}{\chi^2_{n,\frac{1+\alpha}{2}}}}, \sqrt{\dfrac{\sum(X_i-a)^2}{\chi^2_{n,\frac{1-\alpha}{2}}}}\right)$.

3. Интервал для a, если σ неизвестна

> **Теорема:** Пусть $X = (X_1, \ldots, X_n) \sim \mathcal{N}(a, \sigma^2)$. Тогда:
>
> 1. Статистики $\overline{X}$ и S^2 независимы
> 2. $\dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$
> 3. $\sqrt{n-1}\dfrac{\overline{X}-a}{S} \sim T_{n-1}$ — *распределение Стьюдента с $n-1$ степенями свободы.*

$\triangle$ 1), 2) — позже

3) $\sqrt{n}\dfrac{\overline{X}-a}{\sigma} \sim \mathcal{N}(0,1)$; $\dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$

Свойство распределения Стьюдента: если $\xi \sim \mathcal{N}(0,1), \eta \sim \chi^2_k$ — независимые с.в., то $\zeta = \frac{\xi}{\sqrt{\eta/k}} \sim T_k$. Следовательно:

$$\frac{\sqrt{n}\dfrac{\overline{X}-a}{\sigma}}{\sqrt{\dfrac{nS^2}{\sigma^2}\cdot\dfrac{1}{n-1}}} = \sqrt{n-1}\frac{\overline{X}-a}{S} \sim T_{n-1}. \quad \square$$

$G(X,\theta) = \sqrt{n-1}\dfrac{\overline{X}-\theta}{S}$ — центральная функция.

Получаем интервал $\left(\overline{X} \pm T_{n-1,\frac{1+\alpha}{2}}\dfrac{S}{\sqrt{n-1}}\right)$.

Замечание: При больших n интервал почти совпадает с интервалом из пункта 1.

4. Интервал для σ, если a неизвестно

$G(X,\sigma) = \dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$ — центральная функция. Аналогично п.2 получаем интервал $\left(\sqrt{\dfrac{nS^2}{\chi^2_{n,\frac{1+\alpha}{2}}}}, \sqrt{\dfrac{nS^2}{\chi^2_{n,\frac{1-\alpha}{2}}}}\right)$.

> **Теорема (о разложении гауссовского вектора):** Пусть $\xi = (\xi_1, \dots, \xi_n) \sim \mathcal{N}(a, \sigma^2 I_n)$, $\mathbb{R}^n = \mathcal{L}_1 \oplus \dots \oplus \mathcal{L}_k$ — разложение в прямую сумму ортогональных подпространств, $\eta_j = \operatorname{proj}_{\mathcal{L}_j} \xi$ — проекция на $\mathcal{L}_j$. Тогда:
>
> 1. $\eta_1, \dots, \eta_k$ независимы в совокупности;
> 2. $E\eta_j = \operatorname{proj}_{\mathcal{L}_j} a$;
> 3. $\frac{1}{\sigma^2}\|\eta_j - E\eta_j\|^2 \sim \chi^2_{d_j}$, где $d_j = \dim \mathcal{L}_j$.

Доказательство:
Выберем ортонормированный базис в $\mathbb{R}^n$ следуюзим образом:

$$\underbrace{e_1, e_2, \dots}_{\text{базис в } \mathcal{L}_1} \underbrace{\dots\dots}_{\text{базис в } \mathcal{L}_2} \cdots \underbrace{\dots e_n}_{\text{базис в } \mathcal{L}_k}.$$

Обозначим:

- I_j — набор индексов, соответствующий базису в $\mathcal{L}_j$;
- $B = (e_1, \dots, e_n) \in \mathbb{R}^{n \times n}$ — ортогональная матрица;
- $\zeta_i = \langle \xi, e_i \rangle = e^T \xi$ — проекция на e_i.

Получаем:

$$\zeta = \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{pmatrix} = \begin{pmatrix} e_1^T \xi \\ \vdots \\ e_n^T \xi \end{pmatrix} = B^T \xi$$

$$\xi = \sum_{i=1}^{n} \langle \xi, e_i \rangle \cdot e_i = \sum_{i=1}^{n} \zeta_i e_i = (e_1 \dots e_n) \cdot \zeta$$

$\xi = B\zeta$

- $E\zeta = EB^T\xi = B^T E\xi = B^T a$
- $D\zeta = DB^T\xi = BD\xi B^T = B\sigma^2 I_n B^T = \sigma^2 \underbrace{BB^T}_{=I_n} = \sigma^2 I_n$

Вывод: ζ — гауссовский вектор с независимыми компонентами.

$$\eta_j = \operatorname{proj}_{\mathcal{L}_j} \xi = \sum_{i \in I_j} \langle \xi, e_i \rangle e_i = \sum_{i \in I_j} \zeta_i e_i.$$

Компоненты вектора ζ в разных η_j не пересекаются, следовательно, $\eta_1, \dots, \eta_k$ независимы в совокупности — утв. 1 доказано;

$E\eta_j = \sum_{i \in I_j} \langle E\xi, e_i \rangle e_i = \sum_{i \in I_j} \langle a, e_i \rangle e_i = \operatorname{proj}_{\mathcal{L}_j} a$ — утв. 2 доказано;

$$\frac{1}{\sigma^2}\|\eta_j - E\eta_j\|^2 = \frac{1}{\sigma^2}\left\| \sum_{i \in I_j} \langle \xi - a, e_i \rangle e_i \right\|^2 = \sum_{i \in I_j} \underbrace{\left(\frac{\zeta_i - E\zeta_i}{\sigma} \right)^2}_{\sim \mathcal{N}(0,1) \text{ и незав.}} \sim \chi^2_{\dim \mathcal{L}_j}. \quad \square$$

Доказательство пп. 1-2 из предыдущей теоремы:

1. $$\mathbb{R}^n = \mathcal{L} \oplus \mathcal{L}^\perp, \text{ где } \mathcal{L} = \left\langle \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\rangle.$$

$$\operatorname{proj}_{\mathcal{L}} X = \operatorname*{arg\,min}_{c \in \mathbb{R}} \left\| X - \begin{pmatrix} c \\ c \\ \vdots \\ c \end{pmatrix} \right\|^2 = \operatorname*{arg\,min}_{c \in \mathbb{R}} \sum_{i=1}^{n} (X_i - c)^2 = \begin{pmatrix} \overline{X} \\ \overline{X} \\ \vdots \\ \overline{X} \end{pmatrix}.$$

$$\operatorname{proj}_{\mathcal{L}^\perp} X = X - \operatorname{proj}_{\mathcal{L}} X = \begin{pmatrix} X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix}.$$

По теореме о разложении гауссовского вектора X и $(X_1 - \overline{X}, \dots, X_n - \overline{X})$ независимы, а S^2 зависит только от $(X_1 - \overline{X}, \dots, X_n - \overline{X})$. Вывод: $\overline{X}$ и S^2 независимы.

2. Докажем, что $\dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$:

$$\frac{1}{\sigma^2} \| \operatorname{proj}_{\mathcal{L}^\perp} X - E \operatorname{proj}_{\mathcal{L}^\perp} X \| = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

по теореме о разложении гауссовского вектора. $\square$