Лекція 4. Статистичне виведення в R. Оцінки та асимптотика

Данило Тавров

01.03.2023

План лекції

Вступні зауваги

- Ми продовжуємо розгляд основних понять із теорії ймовірностей і статистики та їхніх реалізацій в R
- Сьогодні згадаємо першу половину основних понять зі статистики
  - Ми пригадаємо, що таке статистична оцінка та які її властивості
  - Ми з'ясуємо, чому нас у першу чергу цікавлять асимптотичні властивості оцінок
- Корисними матеріалами є:
  - Фундаментальна книжка *All of Statistics* (Larry Wasserman), розділи 6, 9 (викладено на диску в загальному каталозі з літературою)
  - Книжка *Introduction to Econometrics* (Bruce Hansen) (розділи 6, 8) (викладено на диску в загальному каталозі з літературою)
  - Книжка *Using R for Introductory Statistics* (John Verzani), розділ 7 (викладено на диску в каталозі з лекцією)
- Матеріал цієї лекції частково базується на конспекті лекцій із дисципліни ECON 141 *Econometrics: Math Intensive* (University of California, Berkeley) авторства Віри Семенової та Данила Таврова

Параметричне і непараметричне виведення

- Типовим питанням статистичного виведення є: «Маючи на руках вибірку $X_1, \ldots, X_n \sim F$, що можна сказати про F?»
- **Статистичною моделлю** (statistical model) вважатимемо множину розподілів $\mathcal{F}$
- **Параметричною моделлю** (parametric model) є статистична модель $\mathcal{F}$, яку можна параметризувати скінченною кількістю параметрів
  - Наприклад, нормальний розподіл є параметричною моделлю:
$$\mathcal{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \ \mu \in \mathbb{R}, \sigma > 0 \right\}$$
  - У загальному випадку маємо деяку модель
$$\mathcal{F} = \{ f(x; \theta) : \theta \in \Theta \} ,$$
    де Θ є **простором параметрів** (parameter space)
- **Непараметричною моделлю** (nonparametric model) є модель $\mathcal{F}$, яку неможливо параметризувати скінченним числом параметрів
  - Тобто ми нічого не знаємо про DGP і припускаємо, що це деякий розподіл
  - Тоді потрібно оцінити цілу функцію такого розподілу, тобто **незліченну** кількість параметрів (значень функції в кожній точці)
- У нашому курсі ми розглянемо (деякі!) методи виведення як параметричні, так і непараметричні

Приклади параметричного виведення

- Нехай $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}\,(p)$
  - Тоді задача стоїть — оцінити параметр p
- Нехай $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$
  - Тоді задача стоїть — оцінити параметри μ і σ^2
  - Задача може стояти — оцінити **тільки** параметр μ
  - Тоді σ буде вважатися **завадним параметром** (nuisance parameter)

Приклади непараметричного виведення

- Нехай $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$
  - Задача може стояти — оцінити $F \in \mathcal{F}$, де $\mathcal{F}$ — множина всіх можливих функцій розподілу
  - Задача може стояти — оцінити $f = F'$, де $f \in \mathcal{F}$ — множина всіх щільностей розподілу таких, що, наприклад, $\int_{\mathbb{R}} (f''(x))^2 \, dx < \infty$
    - Вимагаємо певної гладкости
- Нехай $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$
  - Задача може стояти оцінити **функціонал** (functional) від функції розподілу F, що відповідає $\mathbb{P}_X$, або від самого $\mathbb{P}_X$
  - Наприклад, сподівання $\mathbb{E}[X] = \int X \, d\mathbb{P}_X$
  - Або медіану $M(F) = F^{-1}(0.5)$

Приклад напівпараметричного виведення

- Нехай маємо вибірку $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$
  - Наприклад, X_i відповідає артеріальному тиску
  - А Y_i відповідає тривалості життя
- У цьому контексті X називають **предиктором** (predictor), **регресором** (regressor), **ознакою** (feature) або **незалежною змінною** (independent variable)
- У цьому контексті Y називають **результатом** (outcome), **відгуком** (response) або **залежною змінною** (dependent variable)
- Ми називаємо $r(x) = \mathbb{E}[Y \mid X = x]$ **функцією регресії** (regression function)
- Якщо $r \in \mathcal{F}$, де $\mathcal{F}$ має скінченну кількість параметрів (напр., множина всіх прямих), то маємо **параметричну регресійну модель** (parametric regression model)
- Якщо ж $\mathcal{F}$ є класом деяких функцій, які неможливо описати скінченним числом параметрів, то маємо **непараметричну регресійну модель**
- Статистичне виведення в цьому сенсі може мати таку інтерпретацію:
  - Якщо потрібно спрогнозувати Y для нового пацієнта з певним тиском X, то матимемо задачу прогнозування
  - Якщо Y дискретна (напр., «вижив» або «помер»), то маємо задачу класифікації
  - Якщо ж потрібно оцінити функцію r (напр., щоб проаналізувати зв'язок між змінними), то маємо задачу регресійного аналізу
- Часто ми не накладаємо жодних обмежень на розподіл X_i, але розглядаємо тільки параметричні регресійні моделі
  - Тоді такі моделі називають **напівпараметричними** (semiparametric)

План лекції

Типовий приклад, де постає потреба оцінювання

- Нехай маємо деяку популяцію, наприклад, усіх зростів людей на планеті Земля
- Нехай $X =$«зріст випадково вибраної людини»
- Нехай стоїть задача оцінити середній зріст у популяції
  - Тобто фактично чому дорівнює $\mathbb{E}\left[X\right]$
- Вочевидь, ми можемо мати доступ тільки до деякої вибірки
- Тому потрібно порахувати деяку **оцінку** $\mathbb{E}\left[X\right]$ на основі наявної вибірки
- Вибірку $X_1, X_2, \dots, X_n$ можна розглядати як:
  - n незалежних реалізацій X
  - **Що те ж саме** (але простіше для аналізу): реалізації n **незалежних та однаково розподілених** (independent and indentically distributed, i.i.d.) величин $X_1, X_2, \dots, X_n$
  - У цьому контексті незалежність означає, що реалізація X_i не впливає на реалізацію $X_j, i \neq j$
  - Однаковий розподіл означає, що всі реалізації походять з однакового розподілу
  - Таку вибірку також називають **репрезентативною** (representative)

Термінологія

- У загальному випадку ми хочемо оцінити (можливо, багатовимірний) параметр θ деякої моделі[1]
  - Цей параметр є фіксованою константою!
- **Статистикою** (statistic) називають будь-яку функцію від вибірки
  - Статистика є випадковою величиною!
  - Це тому, що вибірка має випадкову природу: інші вибірка — інше значення
- Статистикою не може бути функція, яка залежить від невідомого параметра!
- Приклади статистик:
  - Середнє вибіркове: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Вибіркова дисперсія: $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
  - Перше спостереження: X_1
  - Медіана
  - Середнє вибіркове всіх значень, за виключенням найменших 10% і найбільших 10%
- Приклади величин, які **не є** статистиками:
  - Центроване вибіркове середнє $\overline{X} - \mu_X$ (μ_X невідоме)
  - Z-оцінка: $Z = \dfrac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}}$ (μ_X, σ_X невідомі)
- **Оцінкою** (estimator) $\widehat{\theta}$ є статистика, яку використовують для оцінювання θ
  - Це є випадкова величина!
- Ми будемо відрізняти оцінку від **значення оцінки** (estimate) — її реалізації
  - Така реалізація (для конкретної вибірки) є константою!

[1] Ці самі поняття стосуються і непараметричного виведення, але так трішки простіше

Властивості оцінок

- Дуже грубо всі властивості оцінок можна розділити на:
  - **Властивості на малих вибірках** (small-sample properties), які описують поведінку оцінки для деякої вибірки фіксованого розміру n
  - **Властивості на великих вибірках**, або **асимптотичні** (large-sample, asymptotic properties), які описують поведінку оцінки, коли $n \to \infty$
  - Прямування до нескінченности потрібно сприймати як синонім фрази «коли n достатньо велике»
  - У різних контекстах «достатньо великими» можуть бути різні n
- Основними властивостями на малих вибірках, які нас цікавлять, є незміщеність та ефективність
- Основними асимптотичними властивостями, які нас цікавлять, є спроможність та, ширше, асимптотичний розподіл
  - Особливо якщо можна довести, що він буде нормальний
- У нашому курсі ми робимо **акцент на великих наборах даних**, тому нас у першу чергу цікавлять саме асимптотичні властивості оцінок

Незміщені оцінки

- Як правило, ми хочемо, щоб наша оцінка була «точною» в тому сенсі, що вона якнайближче відповідає популяційному параметру
- Кажучи неформально, якби ми могли зібрати велику кількість вибірок, то середнє підрахованих оцінок повинно було б бути близьким до справжнього значення параметра
- Формальніше, оцінку $\hat{\theta}$ називають **незміщеною** (unbiased), якщо

$$\mathbb{E}\left[\hat{\theta}\right] = \theta \tag{2.1}$$

- Тоді **зміщенням** (bias) є (сподівана) різниця між оцінкою та параметром:

$$\text{Bias} = \mathbb{E}\left[\hat{\theta}\right] - \theta \tag{2.2}$$

Приклад аналізу оцінки на незміщеність

- Нехай $U_1, U_2, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \mathsf{U}([0; b]) = U$, параметр b **невідомий**
- Розгляньмо оцінку $\hat{b} = \max\{U_1, U_2, \ldots, U_n\}$
- Як відомо з теорії ймовірностей, з урахуванням незалежности всіх величин,

$$F_{\hat{b}}(x) = \mathbb{P}_U(U_1 \leq x) \cdot \ldots \cdot \mathbb{P}_U(U_n \leq x) = (F_U(x))^n = \begin{cases} 0\,, & x < 0 \\ \left(\dfrac{x}{b}\right)^n\,, & 0 \leq x < b \\ 1\,, & b \leq x \end{cases}$$

- Тоді щільність такого розподілу дорівнює відповідній похідній:

$$f_{\hat{b}}(x) = \begin{cases} \dfrac{nx^{n-1}}{b^n}\,, & 0 \leq x \leq b \\ 0\,, & \text{інакше} \end{cases}$$

- Отже $\mathbb{E}\left[\hat{b}\right] = \displaystyle\int_0^b x \cdot \frac{nx^{n-1}}{b^n}\,dx = \frac{n}{n+1} \cdot b$
- Тому така оцінка є **зміщеною**, і Bias $= \dfrac{n}{n+1} \cdot b - b = -\dfrac{b}{n+1} \neq 0$
- Натомість, **не**зміщеною буде оцінка $\hat{b}_{\text{corrected}} = \frac{n+1}{n}\hat{b}$
- До речі, іншою незміщеною оцінкою b може бути $\tilde{b} = 2\overline{U} \equiv \frac{2}{n}\sum_{i=1}^n U_i$
  - Перевірте це самостійно

Ефективні оцінки

- Оцінка невідомого параметра повинна бути не тільки незміщеною, а ще й зосередженою навколо параметра
- Іншими словами, ми хочемо, щоб оцінка мала **малу дисперсію**
- На практиці в нас є доступ тільки до однієї вибірки (як правило)
  - Отже якщо дисперсія велика, то немає жодної впевнености, що пораховане значення близьке до θ
- Проте така постановка питання не є адекватною
  - Зрештою, $\widehat{\theta} \equiv 0$ має найменшу дисперсію — нульову
  - Але це безсенсовна оцінка
  - Тому доречно говорити про порівняння дисперсій оцінок у **деякому класі**
- Розгляньмо клас незміщених оцінок
- Тоді оцінка $\widehat{\theta}$ **ефективніша** (more efficient) від оцінки $\widetilde{\theta}$, якщо
  - $\operatorname{Var}\left(\widehat{\theta}\right) \leq \operatorname{Var}\left(\widetilde{\theta}\right)$ для всіх θ
  - $\operatorname{Var}\left(\widehat{\theta}\right) < \operatorname{Var}\left(\widetilde{\theta}\right)$ щонайменше для одного θ
- Кажемо, що оцінка $\widehat{\theta}$ є **найефективніша** (the most efficient), або просто **ефективна**, у деякому класі, якщо вона має найнижчу дисперсію серед оцінок цього класу

Приклад ефективної оцінки

- Нехай оцінкою сподівання[2] $\mathbb{E}[X] \equiv \theta$ є статистика $\tilde{\theta} = X_1$
- Ця оцінка ефективніша від оцінки $\breve{\theta} = \frac{X_1 + X_2}{2}$
- Справді, обидві оцінки незміщені, оскільки $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[X] = \theta$
- Проте

$$\text{Var}\left(\breve{\theta}\right) = \frac{\sigma_X^2}{2} < \sigma_X^2 = \text{Var}\left(\tilde{\theta}\right), \quad \forall \theta \in \mathbb{R}$$

- Далі ми побачимо, що вибіркове середнє є ефективним у класі лінійних оцінок

[2] $\sigma_X^2 < \infty$

Середньоквадратична похибка

- Для визначення, наскільки далекою є оцінка від справжнього параметра, використовують **середньоквадратичну похибку** (mean squared error, MSE):

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right] \tag{2.3}$$

- Якщо розписати MSE, дістанемо

$$\mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right] = \mathrm{Var}\left(\widehat{\theta}\right) + \mathrm{Bias}^2\left(\widehat{\theta}\right)$$

- Зокрема, якщо оцінка незміщена, то її MSE дорівнює дисперсії
- На практиці доволі часто виринають оцінки, які **є зміщеними**, але які мають **меншу дисперсію**, ніж відповідні незміщені аналоги
- Для таких оцінок MSE буде менша
- Тобто незміщеність **не є дуже принциповою** характеристикою
- Значно важливіше — щоб оцінка була **спроможною**

Спроможні оцінки (1)

- Кажучи неформально, оцінка $\hat{\theta}$ є **спроможною** (consistent), якщо її розподіл «концентрується» навколо θ зі збільшенням n
- Формальніше, оцінка $\hat{\theta}$ є спроможною, якщо вона збігається до θ за ймовірністю: $\hat{\theta}_n \xrightarrow{p} \theta$

Визначення 2.1

Ми кажемо, що послідовність випадкових величин X_n **збігається** до випадкової величини X **за ймовірністю** (converges in probability), якщо

$$\forall \varepsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}_{X_n, X} \left(|X_n - X| > \varepsilon \right) = 0 \tag{2.4}$$

□

- Інтерпретація збіжности за ймовірністю така: зі збільшенням розміру вибірки n імовірність відхилення оцінки $\hat{\theta}$ від справжнього значення θ прямує до 0
- Тобто якщо n «дуже велике», то підраховане для деякої вибірки значення $\hat{\theta}$ можна вважати близьким до θ

Спроможні оцінки (2)

- Оскільки нас у цьому курсі цікавлять асимптотичні властивості оцінок, то аналіз оцінки на спроможність є ключовим
  - Якщо оцінка не є спроможною, у ній немає жодного практичного сенсу
  - Тому потрібно обов'язково доводити спроможність будь-якої оцінки, із якою ми працюємо
- Якщо оцінка незміщена, то довести її спроможність можна доволі просто за допомогою нерівности Чебишова
- Згадаймо цю нерівність для випадкової величини X:

$$\forall \varepsilon > 0 \quad \mathbb{P}_X\left(|X - \mathbb{E}\left[X\right]| \geq \varepsilon\right) \leq \frac{\mathrm{Var}\left(X\right)}{\varepsilon^2}$$

- Оскільки для незміщної оцінки $\mathbb{E}\left[\hat{\theta}\right] = \theta$, маємо

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(\left|\hat{\theta} - \theta\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}\left(\hat{\theta}\right)}{\varepsilon^2}$$

- Відтак достатньо показати, що $\mathrm{Var}\left(\hat{\theta}\right) \to 0$, коли $n \to \theta$
- Власне, саме так доводять так званий **закон великих чисел** (ЗВЧ)

Спроможні оцінки (3) — Закон великих чисел

Теорема 2.2 ((Слабкий) Закон великих чисел (Weak law of large numbers))

- Нехай маємо послідовність незалежних випадкових величин $X_1, X_2, \dots$
- Нехай їхні сподівання існують, скінченні й дорівнюють $\mathbb{E}\left[X_i\right] = \mu_i, i \geq 1$
- Нехай їхні дисперсії також існують, скінченні й рівномірно обмежені, тобто $\mathrm{Var}\left(X_i\right) \leq C < \infty, i = 1, 2, \dots$
- Тоді

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_n - \mathbb{E}\left[X_n\right]\right) \xrightarrow{p} 0 \tag{2.5}$$

- Існують інші формулювання ЗВЧ, де до послідовности випадкових величин висувають трішки інші вимоги:
  - Усі величини незалежні й мають однаковий розподіл зі скінченним сподіванням $\mathbb{E}\left[X_i\right] = \mu, i = 1, 2, \dots$
  - Усі величини незалежні, дисперсії скінченні й $\sum_{n=1}^{\infty}\frac{\mathrm{Var}\left(X_n\right)}{n^2} < \infty$
- Для нашого курсу слабкого ЗВЧ цілком достатньо[3]
- Деталі див. у Конспекті лекцій із теорії ймовірностей, Розд. 14

[3] Хоча існує й **посилений** (strong) ЗВЧ, де замість збіжности за ймовірністю використовується збіжність майже напевно

Спроможні оцінки (4)

- Отже, маючи деяку вибірку $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$, можемо за ЗВЧ стверджувати, що

$$\frac{1}{n}\sum_{i=1}^{n}(X_n - \mathbb{E}[X_n]) \xrightarrow{p} 0 \quad \Rightarrow \quad \overline{X} \xrightarrow{p} \mathbb{E}[X]$$

- У схожий спосіб можна показати, наприклад, що

$$\frac{1}{n}\sum_{i=1}^{n} X_n^2 \equiv \overline{X^2} \xrightarrow{p} \mathbb{E}\left[X^2\right]$$

- Але корисних оцінок подібної структури можна придумати не так багато

Спроможні оцінки (5)

- Тому на практиці для доведення спроможности корисно застосовувати таку теорему

Теорема 2.3 (Теорема про неперервне відображення, ТНВ (Continuous mapping theorem, CMT))

- Нехай маємо послідовність випадкових векторів $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$
- Нехай $g : \mathbb{R}^k \to \mathbb{R}^n$ неперервна майже напевно
- Тоді $g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{X})$

- Наприклад, застосування цієї теореми безпосередньо дає такі результати, як
  - $\left(\overline{X}\right)^2 \xrightarrow{p} \left(\mathbb{E}\left[X\right]\right)^2$
  - $\dfrac{1}{\overline{X}} \xrightarrow{p} \dfrac{1}{\mathbb{E}\left[X\right]}$ тощо

План лекції

Середнє вибіркове (1)

- Розгляньмо властивості чи не найчастіше використовуваної оцінки — оцінки сподівання $\mathbb{E}\left[X\right]$ деякої випадкової величини X
- Нехай $X_i \overset{\text{i.i.d.}}{\sim} X, i = 1, \ldots, n$, $\mathbb{E}\left[X\right] = \mu$, $\text{Var}\left(X\right) = \sigma^2$
- Тоді **середнє вибіркове** (mean) $\overline{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$ має корисні властивості
- Воно незміщене:

$$\mathbb{E}\left[\overline{X}\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \mu$$

- Воно спроможне за ЗВЧ: $\overline{X} \xrightarrow{p} \mu$

Середнє вибіркове (2)

- Дисперсією $\overline{X}$ (з урахуванням незалежности величин!) є

$$\operatorname{Var}\left(\overline{X}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}\left(X_i\right) = \frac{\sigma^2}{n}$$

- Звісно, це менше від, скажімо, $\operatorname{Var}\left(X_1\right) = \sigma^2$
- Проте в загальному випадку середнє вибіркове не є ефективною оцінкою[4]
  - Згідно з теоремою Рао-Блеквелла (Rao–Blackwell theorem), ефективність оцінки $\hat{\theta}$ можна підвищити
  - Для цього потрібно взяти умовне сподівання $\mathbb{E}\left[\hat{\theta} \mid T\right]$, обумовивши так званою **достатньою** статистикою T
  - Зокрема, для рівномірного розподілу достатніми статистиками є мінімум і максимум
  - Відтак ефективність $\overline{X}$ можна підвищити, обумовивши його мінімумом і максимумом
  - Наприклад, якщо $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} \mathrm{U}\left((0;\theta)\right)$, то $\hat{\theta} = \frac{n+1}{2n}\max\left\{X_1, \ldots, X_n\right\}$ має найменшу дисперсію серед усіх незміщених оцінок

[4] Нижченаведені факти не є обов'язковими до розуміння, їх наведено з міркувань цікавости

Ефективність середнього вибіркового в класі лінійних оцінок (1)

- Можемо показати, що середнє вибіркове є ефективним у класі **лінійних** (linear) оцінок:

$$\mathcal{L} = \left\{ \sum_{i=1}^{n} w_i X_i \equiv \mathbf{w}^\top \mathbf{X} \ , \ \mathbf{w} \in \mathbb{R}^n \right\}$$

  - Тут $\mathbf{w}^\top = (w_1, w_2, \ldots, w_n)$ є вектором вагових коефіцієнтів
- Приклади лінійних оцінок:
  - Середнє вибіркове: $\mathbf{w}^\top = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$
  - Перше спостереження: $\mathbf{w}^\top = (1, 0, \ldots, 0)$
- Цілком очевидно, що всі оцінки $\widehat{\theta} \in \mathcal{L}$ незміщені тоді й тільки тоді, коли

$$\mathbb{E}\left[\widehat{\theta}\right] = \mathbb{E}\left[\sum_{i=1}^{n} w_i X_i\right] = \left(\sum_{i=1}^{n} w_i\right)\mu = \mu$$

- Тобто коли вагові коефіцієнти нормовані: $\sum_{i=1}^{n} w_i = 1$[5]

[5] Окрім, звісно, випадку $\mu = 0$, але тоді вагові коефіцієнти можуть бути довільні, тобто в тому числі й нормовані

Ефективність середнього вибіркового в класі лінійних оцінок (2)

- Дисперсією лінійної оцінки, з урахуванням незалежности, є

$$\text{Var}\left(\hat{\theta}\right) = \text{Var}\left(\sum_{i=1}^{n} w_i X_i\right) = \left(\sum_{i=1}^{n} w_i^2\right) \sigma_X^2$$

- Отже, ефективною незміщеною оцінкою буде така оцінка, яка є розв'язком такої мінімізаційної задачі:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} w_i^2$$

$$\text{s.t.} \sum_{i=1}^{n} w_i = 1$$

- Нескладно показати[6], що оптимальним розв'язком є вектор $\mathbf{w}^* = \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)$
- Оскільки цей вектор відповідає вибірковому середньому, воно є **найліпшою лінійною незміщеною оцінкою** (best linear unbiased estimator, BLUE) сподівання

[6] Лагранжіан дорівнює $L = \sum_{i=1}^{n} w_i^2 - \lambda \left(\sum_{i=1}^{n} w_i - 1\right)$, і умова першого порядку дає $2w_j = \lambda$, $j = 1, \dots, n$. Оскільки $1 = \sum_{i=1}^{n} w_i = \frac{n\lambda}{2}$, маємо $\lambda = \frac{n}{2}$, а відтак $w_j = \frac{1}{n}$ for each j

Ефективність середнього вибіркового в класі лінійних оцінок (3)

- До речи, якщо $\operatorname{Var}(X_i) = \sigma_i^2$, тобто дисперсії неоднакові, то мінімізаційна програма набуває виду

$$\min_{\mathbf{w}} \sum_{i=1}^{n} w_i^2 \sigma_i^2$$
$$\text{s.t.} \sum_{i=1}^{n} w_i = 1$$

- Тоді її розв'язок буде

$$w_i = \frac{1}{C\sigma_i^2}\ , \quad i = 1, \dots, n\ , \quad C = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

Середнє вибіркове як оцінка найменших квадратів (1)

- Також можемо показати, що середнє вибіркове постає в зовсім іншому контексті
  - Ці міркування будуть для нас корисні, коли ми розглядатимемо регресійний аналіз
- Нехай маємо вибірку $X_1, \dots, X_n$
- Нехай нам потрібно з'ясувати значення m^* таке, що сумарна відстань кожного $X_i, i = 1, \dots, n$, до нього є мінімальною
- Замість того, щоб рахувати відстань через модуль, ми візьмемо квадрат відхилення
  - У цьому випадку сумарна «відстань» буде гладкою функцією
- Ми маємо таку оптимізаційну задачу:

$$m^* = \arg\min_m \sum_{i=1}^n (X_i - m)^2$$

- Виконуючи стандартні маніпуляції, маємо:

$$\begin{aligned}
\sum_{i=1}^n (X_i - m)^2 &= \sum_{i=1}^n \left(X_i - \overline{X} + \overline{X} - m\right)^2 \\
&= \sum_{i=1}^n \left(X_i - \overline{X}\right)^2 + 2\left(\overline{X} - m\right) \sum_{i=1}^n \left(X_i - \overline{X}\right) + n\left(\overline{X} - m\right)^2
\end{aligned}$$

Середнє вибіркове як оцінка найменших квадратів (2)

- Можна помітити, що

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n} X_i - \overline{X} = 0$$

- Відтак середній доданок пропадає, і залишається

$$\sum_{i=1}^{n}\left(X_i - m\right)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - m\right)^2$$

- Сума $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ не залежить від m
- Отже весь вираз мінімізовано, коли $m = \overline{X}$

Ілюстрація властивостей середнього вибіркового (1)

- Розгляньмо вибірку $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X = \text{Exp}\,(2)$
- Сподіванням повинно бути $\mathbb{E}\left[X\right] = 0.5$, проте ми цього не знаємо і хочемо оцінити
- Проілюструймо властивості $\overline{X}$ як оцінки сподівання за допомогою **симуляцій за методом Монте-Карло** (Monte Carlo simulations)[7]
- Для цього:
  - Згенеруймо випадкову вибірку розміром n
  - Обчислімо значення $\overline{X}$ для цієї вибірки
  - Повторімо цей процес «велику» кількість разів (наприклад, 10 000)
- Розгляньмо вибірки трьох різних розмірів: $n = 10, 100, 1000$
- Для порівняння розгляньмо оцінку X_1, яка є незміщеною для $\mathbb{E}\left[X\right]$, але явно не є спроможною ($X_1 \xrightarrow{p} X_1$)

[7] Відповідну назву запропонував угорсько-американський математик Джон фон Нойманн (John von Neumann, 1903–1957) для опису першої симуляції такого роду, яка мала місце для дослідження поведінки нейтронів у науковій лабораторії Лос-Аламос під час Другої світової війни. Назва «Монте-Карло», вочевидь, походить від назви курорту в Монако, відомого своїми гральними закладами

Ілюстрація властивостей середнього вибіркового (2)

- Для імплементації методу Монте-Карло в R потрібно написати функцію, яку будемо повторювати
- У нашому випадку ця функція генерує вибірку та обчислює на її основі дві оцінки

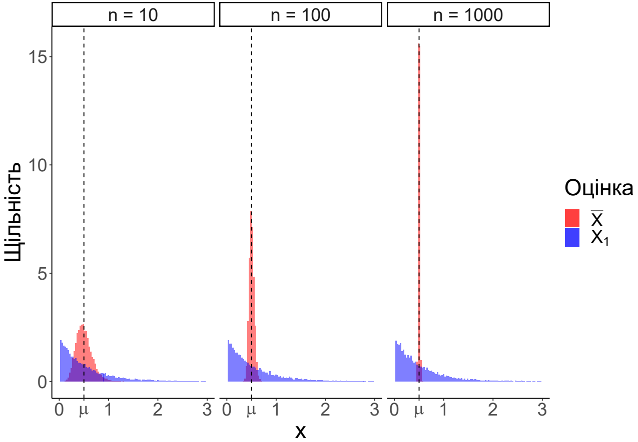
```
estimate_expectation <- function(n){
  x <- rexp(n, rate = 2)

  result <- c(mean(x), x[1], n)
  names(result) <- c("mean", "first", "n")

  return(result)
}
```

- Тепер потрібно викликати цю функцію багато разів (за допомогою `replicate`) для різних n і сформувати датафрейм

```
T <- 10000

df <- NULL
for (n in c(10, 100, 1000)){
  df <- rbind(df, as_tibble(t(replicate(T, estimate_expectation(n)))))
}
```

  - Ми використали функцію транспонування `t`, оскільки `replicate` повертає результат кожного запуску як стовпець, а нам треба рядок
  - Ми конвертуємо результати застосування `estimate_expectation` n разів у датафрейм
  - Ми нарощуємо в циклі великий датафрейм, для кожного n дописуючи до поточного новий шматок

Ілюстрація властивостей середнього вибіркового (3)



- Як можна бачити, хоча обидві оцінки незміщені, тільки оцінка $\overline{X}$ є спроможною
- Також видно, що $\overline{X}$ має меншу дисперсію від X_1
- Можете самостійно модифікувати код і перевірити, чи є оцінка $\frac{n+1}{2n}\max\{X_1, \ldots, X_n\}$ ефективною для випадку $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{U}\left((0;\theta)\right)$

Вибіркова дисперсія (1)

- Для оцінки дисперсії можна було б використати вираз $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$
- Проте це не є статистикою, адже μ невідоме
- Але μ можна, у свою чергу, оцінити як $\overline{X}$, що веде до такої оцінки:

$$\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

- Простими, але нудними маніпуляціями можна показати, що

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\left(\mu-\overline{X}\right)^2$$

- Зокрема, можна далі показати, що $\mathbb{E}\left[\hat{\sigma}^2\right]=\frac{n-1}{n}\sigma^2$
- Тобто $\hat{\sigma}^2$ є зміщеною
- Незміщеною оцінкою є **вибіркова дисперсія** (sample variance):

$$s_X^2 \equiv \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \tag{3.1}$$

Вибіркова дисперсія (2)

- Можна показати, що (3.1) є спроможною
- По-перше, $\dfrac{n}{n-1} \to 1$, коли $n \to \infty$
- Тому $s_X^2 = \frac{n}{n-1}\hat{\sigma}^2$ прямуватиме «туди ж», куди й $\hat{\sigma}^2$
- Тоді можна було б застосувати ТНВ і сказати, що

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \xrightarrow{p} \mathbb{E}\left[\left(X_i - \mu\right)^2\right] = \sigma^2$$

- Проте ми **не можемо** собі цього дозволити, адже випадкові величини $X_1 - \overline{X}$, ..., $X_n - \overline{X}$ **не є незалежні**
  - Це очевидно на такому простому міркуванні
  - І $X_1 - \overline{X}$, і $X_2 - \overline{X}$ містять у собі як X_1, так і X_2
  - Це тому, що $\overline{X}$ є функцією від $X_1, X_2, \ldots, X_n$

Вибіркова дисперсія (3)

- Проте ми знаємо, що

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\mu - \overline{X}\right)^2 \xrightarrow{p} \mathbb{E}\left[(X_i-\mu)^2\right] - (\mu-\mu)^2 = \sigma^2 - 0 = \sigma^2$$

- А відтак остаточно

$$s_X^2 = \frac{n}{n-1}\cdot\hat{\sigma}^2 \xrightarrow{p} 1\cdot\sigma^2 = \sigma^2$$

- За цією ж логікою спроможною також є оцінка середньоквадратичного відхилення:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \xrightarrow{p} \sigma$$

План лекції

Вступні зауваги

- Як ми вже з'ясували, оцінка $\widehat{\theta}$ деякого параметра θ є випадковою величиною
- Підрахувавши її для наявної вибірки, ми всього лише маємо одну реалізацію
- Якщо оцінка спроможна, то ми можемо вважати, що ця реалізація доволі близька до θ
- Але щоб з'ясувати, **наскільки** вона близька, нам потрібно знати **розподіл** $\widehat{\theta}$
- Знання цього розподілу дає змогу оцінити, наскільки далеко ця реалізація може відхилятися від справжнього θ
- У цьому курсі нас переважно цікавитимуть методи аналізу даних з **доволі великими** вибірками
- Відтак нас у першу чергу цікавлять **асимптотичні** розподіли оцінок
- Основним інструментом, який дає нам змогу встановити асимптотичний розподіл, є **центральна гранична теорема** (ЦГТ)
- Щоб її сформулювати, пригадаймо, що таке збіжність за розподілом

Збіжність за розподілом

Визначення 4.1

- Послідовність випадкових величин X_n з функціями розподілу F_n **збігається** до випадкової величини X із функцією розподілу F **за розподілом** (converges in distribution), якщо
$$\lim_{n \to \infty} F_n(x) = F(x) \tag{4.1}$$
для всіх точок x, у яких F неперервна
- Це позначають як $X_n \Rightarrow X$ або $X_n \xrightarrow{d} X$

□

- Тобто маємо збіжність послідовности $X_1, X_2, \ldots$ у тому розумінні, що їхні функції розподілів прямують до деякої граничної функції розподілу
- Або, кажучи неформально, можемо говорити про «приблизний» розподіл X_n для деякого «великого» n
- Ми позначатимемо це через $X_n \overset{\mathrm{a}}{\sim} X$
- Між збіжностями за ймовірністю та за розподілом існує зв'язок
  - Можна довести, що якщо $X_n \xrightarrow{p} X$, то $X_n \xrightarrow{d} X$
  - Зворотне твердження в **загальному** випадку **не є** справедливим
  - Проте можна довести, що якщо $X_n \xrightarrow{d} c$, $c \in \mathbb{R}$, то $X_n \xrightarrow{p} c$

Центральна гранична теорема

Теорема 4.2 (Центральна гранична теорема Ліндеберга-Леві (Lindeberg-Lévy central limit theorem)[8])

- Нехай маємо послідовність $X_1, X_2, \ldots$ **незалежних** величин з **однаковим** розподілом зі **скінченними** сподіванням μ та дисперсією σ^2
- Тоді

$$Z_n \equiv \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim N(0,1) \tag{4.2}$$

- Неформально, для «дуже великого n» Z_n має приблизно стандартний нормальний розподіл
  - Тобто $Z_n \overset{\text{a}}{\sim} N(0,1)$
- Альтернативний для (4.2) запис можна дістати, якщо домножити зліва й справа на σ:

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

  - Він чіткіше дає змогу побачити, якою є асимптотична дисперсія

[8] Ярл Ліндеберг (Jarl Waldemar Lindeberg, 1876–1932) — фінський математик. Поль Леві (Paul Pierre Lévy, 1886–1971) — французький математик

Приклад застосування ЦГТ до вибіркових середніх

- Якщо в (4.2) домножити на весь знаменник і додати μ зліва і справа, то дістанемо
$$\overline{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$
- Отже за ЗВЧ ми знаємо, що $\overline{X} \xrightarrow{p} \mu$
- А ЦГТ додатково вказує, який розподіл має $\overline{X}$ «на шляху до» μ
- Для того, щоб виводити асимптотичні розподіли для інших оцінок, відмінних від $\overline{X}$, корисно розглянути декілька додаткових теорем

Декілька важливих теорем

- По-перше, для збіжности за розподілом справедлива ТНВ, як і для збіжности за ймовірністю
- Також корисною є така теорема

Теорема 4.3 (Теорема Слуцького (Slutsky's Theorem)[9])

- Нехай маємо дві послідовності випадкових величин $X_n \xrightarrow{d} X$ і $Y_n \xrightarrow{p} c$
- Тоді $X_n + Y_n \xrightarrow{d} X + c$
- Також $X_n Y_n \xrightarrow{d} Xc$
- Якщо $c \neq 0$, то $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$

- Разом ці теореми дають змогу встановлювати факт збіжности за розподілом

[9] Євген Слуцький (1880–1948) — український радянський математик, економіст і статистик

Асимптотичний розподіл вибіркової дисперсії (1)

- Як приклад, з'ясуймо, який асимптотичний розподіл має вибіркова дисперсія
- Для спрощення аналізу подамо її в такий спосіб: $s_X^2 = \frac{1}{n}s_X^2 + \frac{n-1}{n}s_X^2$
- Потрібно розглянути

$$\sqrt{n}\left(s_X^2 - \sigma^2\right) = \frac{1}{\sqrt{n}}s_X^2 + \sqrt{n}\left(\frac{n-1}{n}s_X^2 - \sigma^2\right) = \frac{1}{\sqrt{n}}s_X^2 + \sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right)$$

- За ЗВЧ, $s_X^2 \xrightarrow{p} \sigma^2$, і також $\frac{1}{\sqrt{n}} \xrightarrow{p} 0$
- Отже, за теоремою Слуцького, $\frac{1}{\sqrt{n}}s_X^2 \xrightarrow{d} \sigma^2 \cdot 0 = 0$
- З урахуванням раніше показаних викладок,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - \left(\mu - \overline{X}\right)^2$$

- Розберімося спочатку з другим доданком
  - Можна переписати $\sqrt{n}\left(\mu - \overline{X}\right)^2 = \sqrt{n}\left(\mu - \overline{X}\right)\left(\mu - \overline{X}\right)$
  - За ЗВЧ та ТНВ, $\mu - \overline{X} \xrightarrow{p} 0$
  - За ЦГТ, $\sqrt{n}\left(\mu - \overline{X}\right) \xrightarrow{d} N(0,1)$
  - Отже, за теоремою Слуцького, $\sqrt{n}\left(\mu - \overline{X}\right)^2 \xrightarrow{d} 0$

Асимптотичний розподіл вибіркової дисперсії (2)

- Відтак маємо

$$\sqrt{n}\left(s_X^2 - \sigma^2\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - \sigma^2\right)$$

- Проте $\mathbb{E}\left[\left(X_i - \mu\right)^2\right] = \sigma^2$ за визначенням
- Отже застосовна ЦГТ:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - \sigma^2\right) \xrightarrow{d} N\left(0, \mathsf{Var}\left((X_i - \mu)^2\right)\right)$$

- Відтак остаточно

$$s_X^2 \overset{\text{a}}{\sim} N\left(\sigma^2, \frac{\mathsf{Var}\left((X_i - \mu)^2\right)}{n}\right) = N\left(\sigma^2, \frac{\mathbb{E}\left[(X_i - \mu)^4\right] - \sigma^4}{n}\right)$$

- Між іншим, очевидно, що це справедливо, тільки якщо четвертий момент $\mathbb{E}\left[X_i^4\right]$ скінченний

Дельта-метод

- Іще один спосіб виводити асимптотичний розподіл застосовний, якщо нас цікавить розподіл **функції** від оцінки, розподіл якої відомий

Теорема 4.4 (Дельта-метод (Delta-method))

- Нехай маємо $X_1, X_2, \ldots$, і сподівання та дисперсії кожної з них скінченні й дорівнюють μ і σ^2 відповідно
- Нехай виконується $\frac{X_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$
- Нехай функція $g\left(X_n\right)$ диференційовна, і її похідна в точці μ не дорівнює 0
- Тоді

$$\frac{g(X_n) - g(\mu)}{\sigma g'(\mu)/\sqrt{n}} \xrightarrow{d} N(0,1) \tag{4.3}$$

- Або ж

$$g(X_n) \overset{a}{\sim} N\left(g(\mu), \frac{\sigma^2 \left(g'(\mu)\right)^2}{n}\right)$$

- Наприклад, оскільки $s_X = \sqrt{s_X^2}$, і $\mathbb{E}\left[s_X^2\right] = \sigma^2$, то

$$s_X \overset{a}{\sim} N\left(\sigma, \frac{\mathbb{E}\left[(X_i - \mu)^4\right] - \sigma^4}{n} \cdot \left(\frac{1}{2\sqrt{\sigma^2}}\right)^2\right) = N\left(\sigma, \frac{\mathbb{E}\left[(X_i - \mu)^4\right] - \sigma^4}{4\sigma^2 n}\right)$$

Ілюстрація ЦГТ в R (1)

- Розгляньмо вибірку $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} X = \text{Exp}\left(4\right)$
- Тоді $\mathbb{E}\left[X\right] = \frac{1}{4} = 0.25$, $\text{Var}\left(X\right) = \frac{1}{16} = 0.0625$
- Проведімо симуляцію за методом Монте-Карло для вибірок трьох різних розмірів: $n = 10, 100, 1000$
- Для вибірки кожного розміру підрахуймо $\overline{X}$ та s_X^2 та проаналізуймо їхній розподіл
- Ми вже знаємо, що $\overline{X} \overset{\text{a}}{\sim} N\left(\frac{1}{4}, \frac{1}{16n}\right)$
- Також ми знаємо, що $s_X^2 \overset{\text{a}}{\sim} N\left(\frac{1}{16}, \frac{\mathbb{E}\left[(X_i - 0.25)^4\right] - \frac{1}{16^2}}{n}\right)$
- Можна акуратно порахувати, що $\mathbb{E}\left[(X_i - 0.25)^4\right] = \frac{9}{4^4}$
- Відтак $s_X^2 \overset{\text{a}}{\sim} N\left(\frac{1}{16}, \frac{1}{32n}\right)$
- З аналогічних міркувань та з урахуванням дельта-методу, $s_X \overset{\text{a}}{\sim} N\left(\frac{1}{4}, \frac{1}{8n}\right)$

Ілюстрація ЦГТ в R (2)

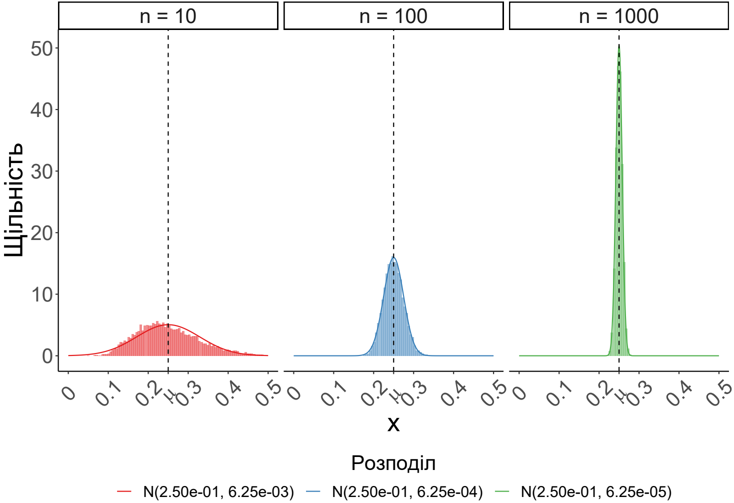
- Функція, яка рахує вибіркові статистики:

```
sample_statistics <- function(n, rate){
  x <- rexp(n, rate = rate)

  result <- c(mean(x), sd(x), (sd(x))^2, n)
  names(result) <- c("mean", "sd", "var", "n")

  return(result)
}
```

- Викликаємо багато разів для різних n і формуємо датафрейм:

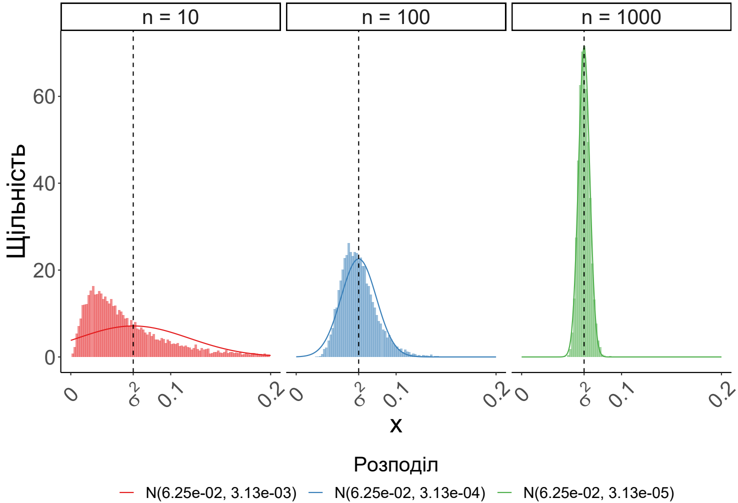
```
T <- 10000
ns <- c(10, 100, 1000)
lambda <- 4
df <- NULL
for (n in ns){
  df <- rbind(df, as_tibble(t(replicate(T, sample_statistics(n, lambda)))))
}
```

Ілюстрація ЦГТ в R (3) — розподіл $\overline{X}$



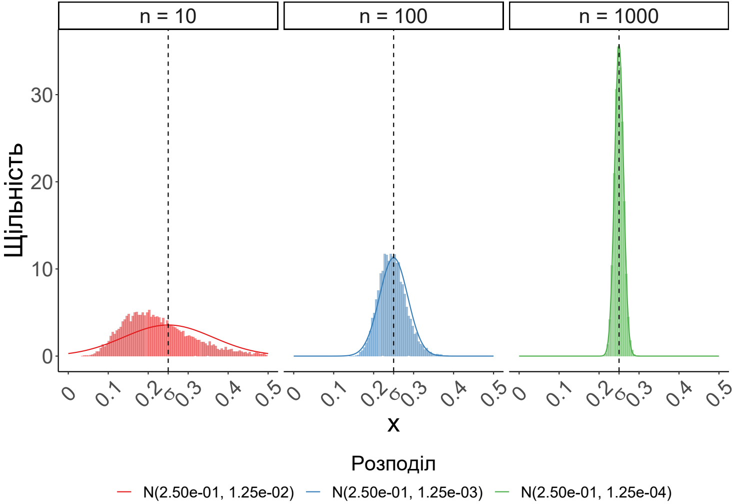
- Згідно з ЦГТ, розподіли прямують до нормального (і дуже швидко)
- Оцінка $\overline{X}$ незміщена для μ і спроможна (дисперсія все менша й менша)
- Теоретичні асимптотичні розподіли підтверджуються симуляцією

Ілюстрація ЦГТ в R (4) — розподіл s_X^2



- Згідно з ЦГТ, розподіли прямують до нормального
- Але для малих n розподіли далекі від нормального!

Ілюстрація ЦГТ в R (5) — розподіл s_X



Стандартна похибка

- Якщо $\widehat{\theta} \overset{a}{\sim} N\left(\mu, \mathrm{Var}\left(\widehat{\theta}\right)\right)$, то **стандартною похибкою** (standard error) є

$$\mathrm{se}\left(\widehat{\theta}\right) = \sqrt{\mathrm{Var}\left(\widehat{\theta}\right)} \tag{4.4}$$

- Стандартну похибку варто відрізняти від вибіркового середньоквадратичного відхилення $\sqrt{s_X^2}$
  - Останнє є оцінкою середньоквадратичного відхилення X, $\sigma = \sqrt{\mathrm{Var}\left(X\right)}$
  - Стандартна похибка є середньоквадратичним відхиленням $\widehat{\theta}$
- **Будь-яка оцінка, подана без відповідної їй стандартної похибки, є НЕПОТРЕБОМ!!!**
- Не маючи жодної інформації про середньоквадратичне відхилення $\widehat{\theta}$, неможливо оцінити, наскільки далеким від істини є підраховане значення оцінки

Способи підрахунку стандартних похибок (1)

- Ідеальною є ситуація, коли ми знаємо асимптотичну дисперсію точно
  - Тоді достатньо взяти корінь і дістати середньоквадратичне відхилення
  - У попередньому прикладі ми саме так і зробили, оскільки це був повністю контрольований симульований експеримент
- На практиці, якщо ми маємо вибірку $X_1, \dots, X_n \sim X$ з невідомого розподілу, усе, що ми в найліпшому випадку знаємо, це що $\overline{X} \overset{a}{\sim} N\left(\mathbb{E}\left[X\right], \frac{\mathrm{Var}(X)}{n}\right)$
  - Так, ми знаємо, що це буде саме асимптотичний **нормальний** розподіл
  - Але нам невідома $\mathrm{Var}\left(X\right)$!
- Тому на практиці найчастіше потрібно **оцінювати** асимптотичну дисперсію, і тоді замість стандартної похибки ми використовуємо її оцінку: $\widehat{\mathrm{se}}\left(\hat{\theta}\right) = \sqrt{\widehat{\mathrm{Var}}\left(\hat{\theta}\right)}$
- Зверніть увагу на таке:
  - Нехай $\frac{\hat{\theta}-\theta}{\sqrt{\mathrm{Var}\left(\hat{\theta}\right)}/\sqrt{n}} \xrightarrow{d} N(0,1)$
  - Також нехай $\widehat{\mathrm{Var}}\left(\hat{\theta}\right) \xrightarrow{p} \mathrm{Var}\left(\hat{\theta}\right)$, тобто наша оцінка дисперсії є спроможною
  - Тоді, за теоремою Слуцького, також $\frac{\hat{\theta}-\theta}{\sqrt{\widehat{\mathrm{Var}}\left(\hat{\theta}\right)}/\sqrt{n}} \xrightarrow{d} N(0,1)$
  - Це означає, що якщо ми можемо оцінити дисперсію $\mathrm{Var}\left(\hat{\theta}\right)$ спроможною оцінкою $\widehat{\mathrm{Var}}\left(\hat{\theta}\right)$, то «все буде в порядку»

Способи підрахунку стандартних похибок (2)

- Отже якщо можна показати, що розподіл $\hat{\theta}$ буде асимптотично нормальний $N\left(\theta, \text{Var}\left(\hat{\theta}\right)\right)$ із деякою дисперсією $\text{Var}\left(\hat{\theta}\right)$, її оцінку можна дістати, замінивши всі моменти на вибіркові аналоги
  - Тобто замінити всі сподівання на вибіркові середні тощо
  - За ЗВЧ та ТНВ такі оцінки будуть спроможними
  - В англомовній літературі такі оцінки називають **plug-in estimators**
- Якщо вираз асимптотичної дисперсії **не можна** (або дуже важко) вивести теоретично, то оцінку можна дістати шляхом симуляцій
- Якщо можна симулювати весь експеримент, можна застосувати метод Монте-Карло
  - Тоді на основі симульованого розподілу можна визначити середньоквадратичне відхилення
- Якщо ж експеримент відтворити неможливо, тобто якщо в нас наявна тільки одна-єдина вибірка, то можна застосувати метод **бутстреп** (bootstrap)
  - Він дає змогу проводити «фейкові» симуляції на основі однієї-єдиної вибірки
  - Ми до нього повернемося в Лекції 6

Ілюстрація підрахунків стандартних похибок (1)

- Розгляньмо простий приклад на основі даних пасажирів «Титаніка»
  - На перший погляд може здатися, що це окрема популяція, тому будь-які статистики будуть точними
  - Проте згідно з джерелом даних, це всього лише вибірка з 891 пасажира з-поміж 1309
- Спробуймо оцінити (безумовну) ймовірність уціліти
  - Тобто не враховуючи жодних інших характеристик пасажирів
  - Така ймовірність навряд чи цікава сама по собі
  - Але вона відіграє свою ілюстративну роль
- Оскільки змінна `Survived` є бінарною, фактично маємо справу з величиною $X \sim \text{Bern}(p)$
  - Тоді $p = \mathbb{P}_X (X = 1) = \mathbb{E}[X]$
  - А відтак $\widehat{p} = \overline{X}$

```
p_est <- passengers %>% filter(!is.na(Survived)) %>% summarise(p_hat = mean(Survived))
p_est
```

```
## # A tibble: 1 x 1
##   p_hat
##   <dbl>
## 1 0.384
```

- Саме по собі значення 0.384 **жодного** сенсу немає
- Потрібно оцінити стандартну похибку

- Згадаймо, що $\overline{X} \overset{\text{a}}{\sim} N\left(\mathbb{E}\left[X\right], \frac{\mathrm{Var}(X)}{n}\right) = N\left(p, \frac{p(1-p)}{n}\right)$
  - Кількість спостережень близько тисячі дає підстави застосувати ЦГТ
- Отже $\widehat{\mathrm{se}}\left(\hat{p}\right) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$
  - Ми замінили всі параметри на їхні вибіркові аналоги (тобто p замінили на $\hat{p}$)
  - Звісно, можна було б використати функцію `var` і порахувати вибіркову дисперсію `Survived`, результат буде дуже подібний

```
p_est <- passengers %>% filter(!is.na(Survived)) %>%
  summarise(p_hat = mean(Survived),
            p_se = sqrt(p_hat * (1 - p_hat) / n()))
p_est
```

```
## # A tibble: 1 x 2
##   p_hat   p_se
##   <dbl>  <dbl>
## 1 0.384 0.0163
```

- Тепер можемо бачити, що, згідно з правилом трьох сигм, приблизно 95% значень зосереджено в інтервалі $[0.384 - 2 \cdot 0.016; 0.384 + 2 \cdot 0.016] = [0.352; 0.416]$
- Ми до цих міркувань повернемося в наступній лекції, коли будемо розглядати довірчі інтервали

Ілюстрація підрахунків стандартних похибок (3)

- Розгляньмо тепер таку характеристику, як **шанси** (odds): $odds = \frac{p}{1-p}$
- Наприклад, якщо $p = \frac{1}{4}$, то $odds = \frac{p}{1-p} = \frac{1}{3}$, тобто шанси один до трьох, що подія станеться
- Для нашого прикладу оцінкою $odds$ є $\widehat{odds} = \frac{\hat{p}}{1-\hat{p}}$
- За ЗВЧ і ТНВ вона є спроможною
- Стандартну похибку такої оцінки можна знайти за допомогою дельта-методу
  - Справді, маємо функцію $g(x) = \frac{x}{1-x}$
  - Її похідна дорівнює $g'(x) = \frac{1}{(1-x)^2}$
  - Усі інші умови Теореми 4.4 також виконуються
- Відтак

$$\widehat{odds} \overset{a}{\sim} N\left(odds, \frac{\mathrm{Var}\,(X)}{n} \cdot \left(\frac{1}{(1-p)^2}\right)^2\right) = N\left(odds, \frac{p}{n(1-p)^3}\right)$$

- У нашому випадку це можна обчислити так:

```
p_est <- passengers %>% filter(!is.na(Survived)) %>%
  summarise(p_hat = mean(Survived),
            p_se = sqrt(p_hat * (1 - p_hat) / n()),
            odds_hat = p_hat / (1 - p_hat),
            odds_se = sqrt(p_hat / ((1 - p_hat)^3 * n())))
p_est
```

```
## # A tibble: 1 x 4
##   p_hat   p_se odds_hat odds_se
##   <dbl>  <dbl>    <dbl>   <dbl>
## 1 0.384 0.0163    0.623  0.0429
```

Ілюстрація підрахунків стандартних похибок (4)

- Для автоматизації цього процесу можна використати функцію `deltamethod` із пакету `msm`

```
p_est <- passengers %>% filter(!is.na(Survived)) %>%
  summarise(p_hat = mean(Survived),
            p_se = sqrt(p_hat * (1 - p_hat) / n()),
            odds_hat = p_hat / (1 - p_hat),
            odds_se = deltamethod(~ x1 / (1 - x1), mean = p_hat, cov = p_se^2))
p_est
```

```
## # A tibble: 1 x 4
##   p_hat   p_se odds_hat odds_se
##   <dbl>  <dbl>    <dbl>   <dbl>
## 1 0.384 0.0163    0.623  0.0429
```

- Зверніть увагу, що для застосування дельта-методу потрібна (принаймні асимптотична) нормальність розподілу $\widehat{p}$ та добра якість лінійної апроксимації похідної g'
  - Інакше відповідна стандартна похибка була далека від істини

Ілюстрація ЦГТ в R (1)

- Розгляньмо вибірку $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} X = \text{Exp}\,(4)$
- Тоді $\mathbb{E}\,[X] = \frac{1}{4} = 0.25$, Var $(X) = \frac{1}{16} = 0.0625$
- Проведімо симуляцію за методом Монте-Карло для вибірок трьох різних розмірів: $n = 10, 100, 1000$
- Для вибірки кожного розміру підрахуймо $\overline{X}$ та s_X^2 та проаналізуймо їхній розподіл
- Ми вже знаємо, що $\overline{X} \overset{\text{a}}{\sim} N\left(\dfrac{1}{4}, \dfrac{1}{16n}\right)$
- Також ми знаємо, що $s_X^2 \overset{\text{a}}{\sim} N\left(\dfrac{1}{16}, \dfrac{\mathbb{E}\left[(X_i - 0.25)^4\right] - \frac{1}{16^2}}{n}\right)$
- Можна акуратно порахувати, що $\mathbb{E}\left[(X_i - 0.25)^4\right] = \dfrac{9}{4^4}$
- Відтак $s_X^2 \overset{\text{a}}{\sim} N\left(\dfrac{1}{16}, \dfrac{1}{32n}\right)$
- З аналогічних міркувань та з урахуванням дельта-методу, $s_X \overset{\text{a}}{\sim} N\left(\dfrac{1}{4}, \dfrac{1}{8n}\right)$