Лекція 6. Бутстреп

Данило Тавров

2023-03-15

Вступні зауваги

- Сьогодні ми розглянемо надзвичайно цікавий і потужний статистичний інструментарій — **бутстреп** (bootstrap)
  - Етимологія терміна походить з ідіоми «pull oneself up by one's bootstraps»
  - Це можна перекласти як «витягти самого себе за шнурівки»
- Метафора доволі промовиста, адже ми зможемо здійснювати статистичне виведення, не маючи жожного уявлення про теоретичний розподіл аналізованих оцінок!
- Корисними матеріалами є:
  - Фундаментальна книжка *All of Statistics* (Larry Wasserman), розділи 7–8 (викладено на диску в загальному каталозі з літературою)
  - Книжка *An Introduction to the Bootstrap* (Bradley Efron, R.J. Tibshirani) , розділи 2, 4–7, 10, 12 (викладено на диску в загальному каталозі з літературою)

Ідея та призначення бустрепу

- Авторство цієї ідеї належить американському статистику Бредлі Ефрону (Bradley Efron)
  - Він опублікував її в статті *Bootstrap methods: Another look at the jackknife* (The Annals of Statistics, 7 (1), 1–26) у 1979 р.
- Бутстреп дає змогу виконувати статистичне виведення не за допомогою **теоретичних розрахунків**, а за допомогою **симуляцій**
- Зокрема, якби ми мали доступ до DGP і могли генерувати вибірки з його допомогою, ми б могли оцінювати розподіли будь-яких цікавих для нас параметрів за методом Монте-Карло
  - Наприклад, ми маємо монетку, але не знаємо, чи вона правильна
  - Ми можемо підкидати її багато разів, і встановити ймовірність випадку герба як емпіричну частку гербів
  - ЗВЧ гарантує близькість цих двох чисел
- Проте на практиці ми маємо доступ тільки до **однієї** вибірки, і не можемо генерувати нових
- Використовуючи бутстреп, можна **навіть на основі однієї вибірки** оцінити стандартні похибки, довірчі інтервали оцінок невідомих параметрів, тестувати гіпотези з ними тощо
- Застосування бутстрепу не обмежується теоретичною складністю аналізу властивостей деякої оцінки або припущеннями, які висуваються до DGP
- Навпаки, дослідник може бути агностиком щодо конкретного розподілу, який мають дані, але при цьому здійснювати статистичне виведення

План лекції

Емпірична функція розподілу (1)

- Так чи інакше, статистичне виведення передбачає, що існує деякий DGP — розподіл $\mathbb{P}_X$, який мають дані з популяції, та відповідна йому функція розподілу F
- Замість популяції ми спостерігаємо випадкову вибірку $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$, яку позначатимемо через $\mathbf{X}$
- На її основі можна визначити **емпіричну функцію розподілу** $\hat{F}$:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} \tag{1.1}$$

- Тобто для кожного x це просто **частка** всіх спостережень, які не перевищують x
- Тривіальне застосування ЗВЧ дає $\hat{F}(x) \xrightarrow{p} F(x)$
  - Це справді очевидно, адже $\mathbb{E}_F[\mathbb{1}\{X_i \leq x\}] = \mathbb{P}_F(X_i \leq x) = F(x)$
- Більше того, теорема Гливенка-Кантеллі (Glivenko-Cantelli theorem)[1] каже, що

$$\sup_x \left| F(x) - \hat{F}(x) \right| \overset{\text{м.н.}}{\to} 0 \tag{1.2}$$

- Тобто ця збіжність є рівномірною

[1] Валерій Гливенко (1896–1940) — український математик. Франческо Кантеллі (Francesco Paolo Cantelli, 1875–1966) — італійський математик

Емпірична функція розподілу (2)

- Так само, як F однозначно пов'язана з розподілом $\mathbb{P}_X$, так і $\hat{F}$ однозначно пов'язана зі своїм розподілом, який позначатимемо $\hat{\mathbb{P}}_X$
- Якщо уважно подивитися, то фактично $\hat{F}$ визначає дискретний рівномірний розподіл $\hat{\mathbb{P}}_X$, тобто $X \sim \hat{\mathbb{P}}_X$, якщо

$$\hat{\mathbb{P}}_X\left(X = X_i \mid \mathbf{X}\right) = \frac{1}{n}, \quad i = 1, \dots, n$$

- Для дискретних величин імовірність A дорівнює сумі ймовірностей елементів A:

$$\hat{\mathbb{P}}_X\left(X \in A \mid \mathbf{X}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i \in A\right\}$$

- Отже

$$\hat{F}(x) \equiv \hat{\mathbb{P}}_X\left(X \leq x \mid \mathbf{X}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i \leq x\right\}$$

- Це те саме, що було на попередньому слайді, тобто $\hat{\mathbb{P}}_X$ справді є дискретний рівномірний

Загальний принцип plug-in оцінок

- Будь-яку характеристику DGP, яку ми хочемо оцінити, можна подати як **статистичний функціонал** (statistical functional)
  - Ми їх часто називаємо параметрами, хоча ці моделі необов'язково є параметричними
- Це фактично функція від $\mathbb{P}_X$, яку будемо позначати через $T(\mathbb{P}_X)$
- Наприклад, сподівання можна записати так: $\mathbb{E}_{\mathbb{P}_X}[X] = T(\mathbb{P}_X) = \int X \, d\mathbb{P}_X$
  - Тут індекс явно вказує, за яким розподілом ми рахуємо
  - Це буде потрібно далі, щоб не плутати позначення
- А медіану — як $M = T(\mathbb{P}_X) = F^{-1}(0.5)$, адже F і $\mathbb{P}_X$ пов'язані однозначно
- **Plug-in оцінка** (plug-in estimator) деякого $\theta = T(\mathbb{P}_X)$ полягає в тому, що замість $\mathbb{P}_X$ використовують $\hat{\mathbb{P}}_X$:

$$\hat{\theta} = T\left(\hat{\mathbb{P}}_X\right) \tag{1.3}$$

- Наприклад, нехай $\theta = T(\mathbb{P}_X) = \mathbb{E}_{\mathbb{P}_X}[h(X)]$, тоді

$$\hat{\theta} = T\left(\hat{\mathbb{P}}_X\right) = \mathbb{E}_{\hat{\mathbb{P}}_X}[h(X)] = \int h(X) \, d\hat{\mathbb{P}}_X = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

  - Чому саме так?
  - Будь-який **дискретний** розподіл абсолютно неперервний відносно **лічної** міри!
  - Ми про це говорили в Лекції 3

Приклади plug-in оцінок (1)

- Якщо $\theta = \text{Var}_{\mathbb{P}_X}(X) = \mathbb{E}_{\mathbb{P}_X}[X^2] - (\mathbb{E}_{\mathbb{P}_X}[X])^2$, то відповідна plug-in оцінка дорівнює

$$\text{Var}_{\hat{\mathbb{P}}_X}(X) \equiv \hat{\sigma}^2 = \mathbb{E}_{\hat{\mathbb{P}}_X}[X^2] - (\mathbb{E}_{\hat{\mathbb{P}}_X}[X])^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2$$

  - Зверніть увагу, що це зміщена оцінка
  - Зазвичай ми користуємося варіантом із $n-1$ у знаменнику дробу
- Нехай маємо X із розподілом, який має скінченні сподівання μ та дисперсією σ^2
  - Тоді **коефіцієнт асиметрії** (skewness) дорівнює

$$\text{Skew}_{\mathbb{P}_X}(X) = \mathbb{E}_{\mathbb{P}_X}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\int (x-\mu)^3 \, d\mathbb{P}_X}{\left(\int (x-\mu)^2 \, d\mathbb{P}_X\right)^{3/2}} \tag{1.4}$$

  - Його plug-in оцінкою буде

$$\text{Skew}_{\hat{\mathbb{P}}_X}(X) = \frac{\mathbb{E}_{\hat{\mathbb{P}}_X}\left[\left(X - \mathbb{E}_{\hat{\mathbb{P}}_X}[X]\right)^3\right]}{\left(\mathbb{E}_{\hat{\mathbb{P}}_X}\left[\left(X - \mathbb{E}_{\hat{\mathbb{P}}_X}[X]\right)^2\right]\right)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{\mathbf{X}}\right)^3}{\hat{\sigma}^3}$$

Приклади plug-in оцінок (2)

- Нехай $Z = (X, Y)^\top$ і нехай $\rho = T(\mathbb{P}_Z) = \frac{\mathbb{E}_{\mathbb{P}_Z}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\text{Var}_{\mathbb{P}_Z}(X)\cdot\text{Var}_{\mathbb{P}_Z}(Y)}}$ — коефіцієнт кореляції
- Цей коефіцієнт можна формально записати як

$$T(\mathbb{P}_Z) = a(T_1(\mathbb{P}_Z), T_2(\mathbb{P}_Z), T_3(\mathbb{P}_Z), T_4(\mathbb{P}_Z), T_5(\mathbb{P}_Z))$$

  - Тут $T_1(\mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_Z}[X]$
  - $T_2(\mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_Z}[Y]$
  - $T_3(\mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_Z}[XY]$
  - $T_4(\mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_Z}[X^2]$
  - $T_5(\mathbb{P}_Z) = \mathbb{E}_{\mathbb{P}_Z}[Y^2]$
  - $a(t_1, t_2, t_3, t_4, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}$
- Замінюючи початковий розподіл на емпіричний, дістаємо таку plug-in оцінку (вибірковий коефіцієнт кореляції):

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{\mathbf{X}})(Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^n (X_i - \overline{\mathbf{X}})^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{\mathbf{Y}})^2}}$$

Приклади plug-in оцінок (3)

- Нарешті, нехай випадкова величина X має строго зростаючу функцію розподілу F зі щільністю f
  - Тоді p-ий квантиль ($0 < p < 1$) можна визначити як $T(F) = F^{-1}(p)$
  - Відтак plug-in оцінкою буде вибірковий квантиль $T\left(\hat{F}\right) = \hat{F}^{-1}(p)$
  - Якщо $\hat{F}$ не має оберненої, кладуть $\hat{F}^{-1}(p) = \inf\left\{x : \hat{F}^{-1}(x) \geq p\right\}$

Спроможність plug-in оцінок (1)

- Нас цікавить, щоб $\hat{\theta} = \theta(\hat{\mathbb{P}}_X) \xrightarrow{p} \theta(\mathbb{P}_X)$
- Із теореми Гливенка-Кантеллі випливає, що в певному сенсі $\hat{\mathbb{P}}_X$ прямує до $\mathbb{P}_X$
- Теоретично застосування ТНВ дало б підстави стверджувати, що й $\theta(\hat{\mathbb{P}}_X) \xrightarrow{p} \theta(\mathbb{P}_X)$
- Тому проблема може полягати в тому, коли саме відображення θ не є «неперервним» у $\mathbb{P}_X$
- Розгляньмо простий контрприклад
- Нехай маємо деякий розподіл $\mathbb{P}_X$
- Для $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$ нас цікавить параметр

$$\theta(\mathbb{P}_X) = \mathbb{1}\left\{\mathbb{P}_X\left(X_1 = X_2\right) > 0\right\}$$

- Вочевидь, якщо $\mathbb{P}_X$ відповідає неперервній випадковій величині, то $\theta(\mathbb{P}_X) = 0$ завжди
- Проте $\theta(\hat{\mathbb{P}}_X) = 1$, якою б великою не була вибірка
- Отже збіжности за ймовірністю в цьому випадку не спостерігатиметься

Спроможність plug-in оцінок (2)

- Аналогічні проблеми виникають, коли θ лежить на межі деякої області значень
- Наприклад, нехай відомо, що $\mathbb{E}_{\mathbb{P}_X}[X] \geq 0$
- Тоді можна задати $\hat{\theta} = \max\left\{\overline{\mathbf{X}}, 0\right\}$, щоб значення не були від'ємні
- Тоді якщо $\theta = 0$ насправді, то plug-in оцінка цього параметра не буде спроможною
- Теоретичне доведення доволі складне, але потрібно пам'ятати, що існують такі унікальні ситуації, коли plug-in оцінки не є найліпші

Обчислення стандартних похибок plug-in оцінок

- Інколи ми знаємо, як порахувати стандартну похибку $T\left(\hat{\mathbb{P}}_X\right)$ зі статистичної теорії
  - Наприклад, асимптотично для вибіркових середніх чи дисперсій
- Проте часто це зробити неможливо (розподіл оцінки невідомий)
  - Або принаймні складно (напр., асимптотичний розподіл вибіркових квантилів залежить від невідомої щільности)
- Тоді для оцінювання стандартної похибки plug-in оцінок можна використовувати бутстреп
  - Також бутстреп можна застосовувати відразу до побудови довірчих інтервалів, але про це пізніше

План лекції

Основна ідея

- Підбиймо проміжні підсумки
- Ми маємо DGP з розподілом $\mathbb{P}_X$, нас цікавить деякий параметр $\theta = T(\mathbb{P}_X)$
- У нас на руках є деяка вибірка $\mathbf{X} = (X_1, \dots, X_n)$, $X_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$
- Оцінкою параметра є деяка статистика $\hat{\theta} = \hat{\theta}(\mathbf{X})$
  - Це може бути plug-in оцінка $\hat{\theta} = T\left(\hat{\mathbb{P}}_X\right)$
  - Але необов'язково
- Нас цікавить дисперсія цієї оцінки $\text{Var}_{\mathbb{P}_X}\left(\hat{\theta}(\mathbf{X})\right)$
  - Ми підкреслюємо, що $\hat{\theta}$ залежить від розподілу $\mathbb{P}_X$, оскільки $X_i \sim \mathbb{P}_X$
  - Наприклад, для $\hat{\theta} = \overline{\mathbf{X}}$ дисперсія $\text{Var}_{\mathbb{P}_X}\left(\overline{\mathbf{X}}\right) = \frac{\text{Var}_{\mathbb{P}_X}(X_i)}{n}$
- **Ідея бутстрепу** полягає в тому, що $\mathbb{P}_X$ замінюють на $\hat{\mathbb{P}}_X$ і оцінюють $\text{Var}_{\mathbb{P}_X}\left(\hat{\theta}(\mathbf{X})\right)$ як $\text{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$
  - Так само, як $X_i \sim \mathbb{P}_X$, так і $X_i^* \sim \hat{\mathbb{P}}_X$ (про це далі)
- Для простих ситуацій на кшталт $\overline{\mathbf{X}}$ цього достатньо:

$$\text{Var}_{\hat{\mathbb{P}}_X}\left(\overline{\mathbf{X}^*}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}_{\hat{\mathbb{P}}_X}\left(X_i^*\right) = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i^* - \overline{\mathbf{X}^*}\right)^2\right)$$

- У загальному випадку $\text{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$ наближують за допомогою **симуляції Монте-Карло** (Monte Carlo simulation)

Бутстреп-оцінка дисперсії

- Згідно з ЗВЧ та ТНВ, відомо, що вибіркова дисперсія прямує до популяційної за ймовірністю
- Отже можна апроксимувати $\mathrm{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}\right)$ у такий спосіб

1. Згенерувати B бутстреп-вибірок $\mathbf{X}_1^*, \dots, \mathbf{X}_B^*$
   - Кожна вибірка $\mathbf{X}_b^* = (X_{1,b}^*, \dots, X_{n,b}^*)$, $X_{i,b}^* \overset{\text{i.i.d.}}{\sim} \hat{\mathbb{P}}_X$
   - Оскільки $\hat{\mathbb{P}}_X$ є (умовним) дискретним рівномірним розподілом $\hat{\mathbb{P}}_X\left(X = X_i \mid \mathbf{X}\right) = \frac{1}{n}$, нова вибірка фактично є нічим іншим, як **вибіркою** n **значень із $\mathbf{X}$ із повтореннями**
   - Наприклад, якщо $\mathbf{X} = (X_1, X_2, X_3, X_4)$, то однією з бутстреп-вибірок може бути $\mathbf{X}_b^* = (X_3, X_1, X_2, X_1)$
2. Обчислити $\hat{\theta}_b^* = \hat{\theta}(\mathbf{X}_b^*), b = 1, \dots, B$
3. Оцінити дисперсію $\mathrm{Var}_{\mathbb{P}_X}\left(\hat{\theta}\right)$ як вибіркову дисперсію

$$\mathrm{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right) \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*\right)^2 \tag{2.1}$$

   - Тут $\approx$ використано в тому сенсі, що вираз справа прямує за ймовірністю до виразу зліва
4. Оцінити стандартну похибку як $\mathrm{se}_{\mathbb{P}_X}\left(\hat{\theta}\right) \approx \sqrt{\mathrm{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right)}$

Зв'язок реального світу і «світу бутстрепу»

	Реальний світ	Світ бутстрепу
Розподіл	$\mathbb{P}_X$ (невідомий!)	$\hat{\mathbb{P}}_X$ (на основі вибірки $\mathbf{X}$)
Параметр	$\theta = T(\mathbb{P}_X)$ (невідомий!)	$\hat{\theta} = T\left(\hat{\mathbb{P}}_X\right)$
Дані	$\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathbb{P}_X$ **(єдина вибірка)**	$\mathbf{X}^* = (X_1^*, \dots, X_n^*)^\top \sim \hat{\mathbb{P}}_X$ **(можемо генерувати довільну кількість)**
Оцінка	$\hat{\theta} = \hat{\theta}(\mathbf{X})$	$\hat{\theta}^* = \hat{\theta}(\mathbf{X}^*)$
Дисперсія	$\mathrm{Var}_{\mathbb{P}_X}\left(\hat{\theta}\right)$	$\mathrm{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$
Оцінка дисперсії	$\mathrm{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right) \xrightarrow{p}$ $\mathrm{Var}_{\mathbb{P}_X}\left(\hat{\theta}\right)$	$\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2 \xrightarrow{p}$ $\mathrm{Var}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$

Оцінювання зміщення оцінок (1)

- У схожий спосіб можна оцінити **зміщення** (bias) деякої оцінки

$$\text{Bias}_{\mathbb{P}_X}\left(\hat{\theta}(\mathbf{X})\right) = \mathbb{E}_{\mathbb{P}_X}\left[\hat{\theta}(\mathbf{X})\right] - \theta$$

- (Тим більше, що plug-in оцінки часто є зміщені)
- Plug-in оцінкою зміщення в природний спосіб буде величина

$$\text{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right) = \mathbb{E}_{\hat{\mathbb{P}}_X}\left[\hat{\theta}(\mathbf{X}^*)\right] - T\left(\hat{\mathbb{P}}_X\right) \tag{2.2}$$

- У простих випадках ми можемо порахувати відповідні plug-in оцінки зміщень
- Нехай $\theta = \mathbb{E}_{\mathbb{P}_X}[X]$, $X \sim \mathbb{P}_X$, тоді $\hat{\theta}(\mathbf{X}) = \overline{\mathbf{X}}$
- Зміщення дорівнює
  $\text{Bias}_{\mathbb{P}_X}\left(\hat{\theta}(\mathbf{X})\right) = \mathbb{E}_{\mathbb{P}_X}\left[\hat{\theta}(\mathbf{X})\right] - \theta = \mathbb{E}_{\mathbb{P}_X}[X_i] - \mathbb{E}_{\mathbb{P}_X}[X_i] = 0$
- Тоді $\text{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right) = \mathbb{E}_{\hat{\mathbb{P}}_X}\left[\hat{\theta}(\mathbf{X}^*)\right] - \hat{\theta}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\hat{\mathbb{P}}_X}[X_i^*] - \overline{\mathbf{X}}$
- Оскільки $\hat{\mathbb{P}}_X$ є (умовним) дискретним розподілом, $\mathbb{E}_{\hat{\mathbb{P}}_X}[X_i^*] = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{\mathbf{X}}$
- Остаточно $\text{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right) = \overline{\mathbf{X}} - \overline{\mathbf{X}} = 0$
- Іншими словами, plug-in оцінка зміщення є нульовою, що не дивно, адже й сама оцінка $\hat{\theta}(\mathbf{X}) = \overline{\mathbf{X}}$ мала нульове зміщення

Оцінювання зміщення оцінок (2)

- Розгляньмо тепер параметр $\theta = \mathrm{Var}_{\mathbb{P}_X}(X)$
- Її plug-in оцінкою є $\hat{\theta}(\mathbf{X}) = \mathrm{Var}_{\hat{\mathbb{P}}_X}(X) \equiv \hat{\sigma}^2 = \frac{1}{n}\left(X_i - \overline{\mathbf{X}}\right)^2$
  - Ми знаємо, що вона зміщена
  - $\mathrm{Bias}_{\mathbb{P}_X}\left(\hat{\sigma}^2\right) = -\frac{1}{n}\mathrm{Var}_{\mathbb{P}_X}(X)$
- Порахуймо тепер оцінку цього зміщення $\mathrm{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\sigma}^2(\mathbf{X}^*)\right)$
- Оскільки $\hat{\mathbb{P}}_X$ є (умовним) дискретним, його дисперсією є

$$\mathrm{Var}_{\hat{\mathbb{P}}_X}(X_i^*) = \mathbb{E}_{\hat{\mathbb{P}}_X}\left[(X_i^*)^2\right] - \left(\mathbb{E}_{\hat{\mathbb{P}}_X}[X_i^*]\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{\mathbf{X}}\right)^2 = \hat{\sigma}^2$$

- А отже $\mathbb{E}_{\hat{\mathbb{P}}_X}\left[(X_i^*)^2\right] = \mathrm{Var}_{\hat{\mathbb{P}}_X}(X_i^*) + \left(\mathbb{E}_{\hat{\mathbb{P}}_X}[X_i^*]\right)^2 = \hat{\sigma}^2 + \left(\overline{\mathbf{X}}\right)^2$
- Також, через незалежність, маємо $\mathrm{Cov}_{\hat{\mathbb{P}}_X}(X_1^*, X_2^*) = 0$
  - Тобто $\mathbb{E}_{\hat{\mathbb{P}}_X}[X_1^* X_2^*] = \mathbb{E}_{\hat{\mathbb{P}}_X}[X_1^*]\,\mathbb{E}_{\hat{\mathbb{P}}_X}[X_2^*] = \left(\overline{\mathbf{X}}\right)^2$
- Тоді можна показати, що

$$\mathbb{E}_{\hat{\mathbb{P}}_X}\left[\hat{\sigma}^2(\mathbf{X}^*)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\hat{\mathbb{P}}_X}\left[\left(X_i^* - \overline{\mathbf{X}^*}\right)^2\right] = \left(1 - \frac{1}{n}\right)\hat{\sigma}^2$$

- Відтак $\mathrm{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\sigma}^2(\mathbf{X}^*)\right) = \mathbb{E}_{\hat{\mathbb{P}}_X}\left[\hat{\sigma}^2(\mathbf{X}^*)\right] - \hat{\sigma}^2 = -\frac{1}{n}\hat{\sigma}^2$
- Це дуже схоже на зміщення початкової оцінки

Оцінювання зміщення оцінок (3)

- У більшості інших випадків обчислити $\text{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$ складно
- Тому його апроксимують за допомогою симуляції Монте-Карло бутстреп-вибірок та обчислення відповідного вибіркового середнього:

$$\text{Bias}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}(\mathbf{X}^*)\right) \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* - T\left(\hat{\mathbb{P}}_X\right)$$

  - Тут $\approx$ використано в тому сенсі, що вираз справа прямує за ймовірністю до виразу зліва
- **Важливе зауваження**: оцінка параметра θ в принципі може бути не plug-in оцінкою (напр., ми оцінюємо дисперсію незміщеною оцінкою)
- Але в (2.2) $\hat{\theta}(\mathbf{X})$ повинна бути **саме plug-in оцінкою**
  - Трішки конкретніше, нехай маємо оцінку $\tilde{\theta}(\mathbf{X})$, яка не є plug-in оцінкою, і оцінку $\hat{\theta}(\mathbf{X})$, яка є plug-in оцінкою
  - Тоді (2.2) матиме вигляд

$$\text{Bias}_{\hat{\mathbb{P}}_X}\left(\tilde{\theta}(\mathbf{X}^*)\right) = \mathbb{E}_{\hat{\mathbb{P}}_X}\left[\tilde{\theta}(\mathbf{X}^*)\right] - \hat{\theta}(\mathbf{X})$$

- Це тому, що ми підставляємо $\hat{\mathbb{P}}_X$ замість $\mathbb{P}_X$ **усюди**
  - У тому числі в $\theta = T(\mathbb{P}_X)$, щоб дістати $\hat{\theta} = T(\hat{\mathbb{P}}_X)$

Корекція зміщення (1)

- У будь-якому випадку корекцію зміщення (bias correction) можна виконати, віднявши (оцінку) зміщення від оцінки параметра:

$$\hat{\theta}^{BC} = \hat{\theta}(\mathbf{X}) - \left(\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^* - \hat{\theta}(\mathbf{X}) \right) = 2\hat{\theta}(\mathbf{X}) - \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

- Проте потрібно бути дуже акуратними
  - Як відомо, $\text{MSE}(\hat{\theta}) = \text{Var}\left(\hat{\theta}\right) + \text{Bias}^2\left(\hat{\theta}\right)$
  - А тому нас цікавить не те, зміщена оцінка чи ні, а наскільки велику MSE вона має
  - Зокрема, $\text{MSE}(\hat{\theta}) = \text{Var}\left(\hat{\theta}\right)\left(1 + \left(\frac{\text{Bias}^2\left(\hat{\theta}\right)}{\text{Var}\left(\hat{\theta}\right)}\right)^2\right)$
  - Тобто головне, щоб зміщення було невеликим **відносно** стандартної похибки
  - В окремих випадках може так статися, що скоригована оцінка має вищу стандартну похибку
- Щоб перевірити, чи призводить корекція зміщення до збільшення стандартної похибки, можна застосувати **подвійний бутстреп** (double bootstrap)

Корекція зміщення (2)

- Для підрахунку MSE початкової оцінки $\hat{\theta}$ ми генеруємо бутстреп-вибірки $\mathbf{X}_1^*, \dots, \mathbf{X}_B^*$, рахуємо для кожної з них $\hat{\theta}_b^*$ і обчислюємо

$$MSE\left(\hat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}(\mathbf{X})\right)^2$$

- Для підрахунку MSE скоригованої оцінки $\hat{\theta}^{BC}$ ми B' разів генеруємо бутстреп-вибірки $\mathbf{X}_1^{*,k}, \dots, \mathbf{X}_B^{*,k}, k = 1, \dots, B'$
  - На основі вибірок $\mathbf{X}_1^{*,k}, \dots, \mathbf{X}_B^{*,k}$ обчислюємо $\hat{\theta}_k^{BC}$
  - І тоді

$$MSE\left(\hat{\theta}^{BC}\right) = \frac{1}{B'}\sum_{k=1}^{B'}\left(\hat{\theta}_k^{BC} - \hat{\theta}(\mathbf{X})\right)^2$$

Ілюстрація бутстрепу (1)

- Розгляньмо DGP $\mathbb{P}_X \sim \text{Exp}\,(4)$
- Нехай маємо вибірку $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$
- Нас цікавлять оцінки сподівання та дисперсії $\mathbb{P}_X$
  - Відомо, що $\mathbb{E}_{\mathbb{P}_X}[X] = \frac{1}{4} = 0.25$, $\text{Var}_{\mathbb{P}_X}(X) = \frac{1}{16} = 0.0625$
  - Plug-in оцінками сподівання та дисперсії будуть $\overline{X}$ та $\hat{\sigma}^2$ відповідно
- Розгляньмо вибірку розміру $n = 500$
- На минулій лекції ми з'ясували, що в цьому випадку розподіли відповідних оцінок будуть майже нормальні
  - Зокрема, $\overline{X} \overset{\text{a}}{\sim} N\left(\frac{1}{4}, \frac{1}{16n}\right) = N(0.25, 6.25 \cdot 10^{-5})$
  - А $\hat{\sigma}^2 \overset{\text{a}}{\sim} N\left(\frac{n-1}{n} \cdot \frac{1}{16}, \frac{(n-1)^2}{n^2} \cdot \frac{1}{32n}\right) \approx N\left(0.062, 3.12 \cdot 10^{-5}\right)$
- Можемо перевірити, чи буде мати приблизно такий розподіл бутстреп-оцінка для різних $B = 100, 500, 1000$

Ілюстрація бутстрепу (2)

- Генеруємо бутстреп-вибірки та рахуємо вибіркові статистики

```
n <- 100
lambda <- 4
set.seed(100)

x <- rexp(n, rate = lambda)

df <- NULL
for (B in c(100, 500, 1000)){
  x_ast <- replicate(B, sample(x, replace = TRUE))
  means <- colMeans(x_ast)
  vars <- colSums((x_ast - means)^2 / n)

  df <- rbind(df, tibble(mean = means, var = vars, B = B))
}
```

- Стандартними похибками на основі асимптотичного нормального розподілу є $\text{se}\left(\overline{X}\right) \approx 0.0079$ і $\text{se}\left(\hat{\sigma}^2\right) \approx 0.0056$
- Стандартними похибками на основі бутстреп-розподілу є відповідні середньоквадратичні відхилення

```
df %>% group_by(B) %>% summarize(se_mean = sd(mean), sd_var = sd(var))
```

```
## # A tibble: 3 x 3
##       B se_mean sd_var
##   <dbl>   <dbl>  <dbl>
## 1   100  0.0225 0.0107
## 2   500  0.0227 0.0102
## 3  1000  0.0231 0.0105
```

- Як можна бачити, ці значення дуже близькі до справжніх

Ілюстрація бутстрепу (3)

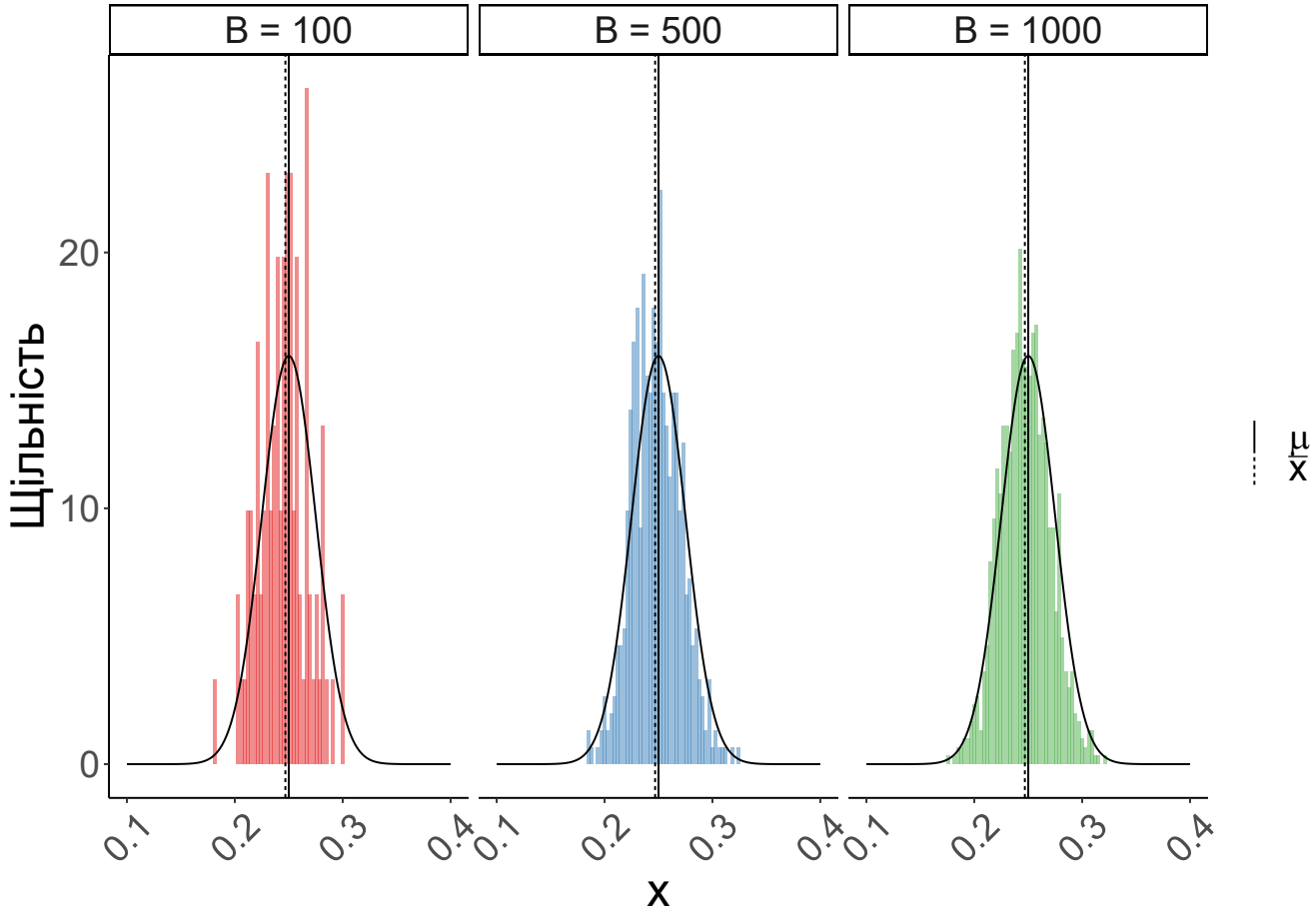
- Зміщеннями на основі бутстреп-розподілу є відповідні різниці між вибірковими середніми та оцінками

```
df %>% group_by(B) %>% summarize(bias_mean = mean(mean) - mean(x),
                                 bias_var = mean(var) - (n - 1) / n * var(x))
```

```
## # A tibble: 3 x 3
##       B bias_mean  bias_var
##   <dbl>     <dbl>     <dbl>
## 1   100 -0.00177  -0.00133
## 2   500  0.00146   0.000909
## 3  1000  0.000287  0.000697
```

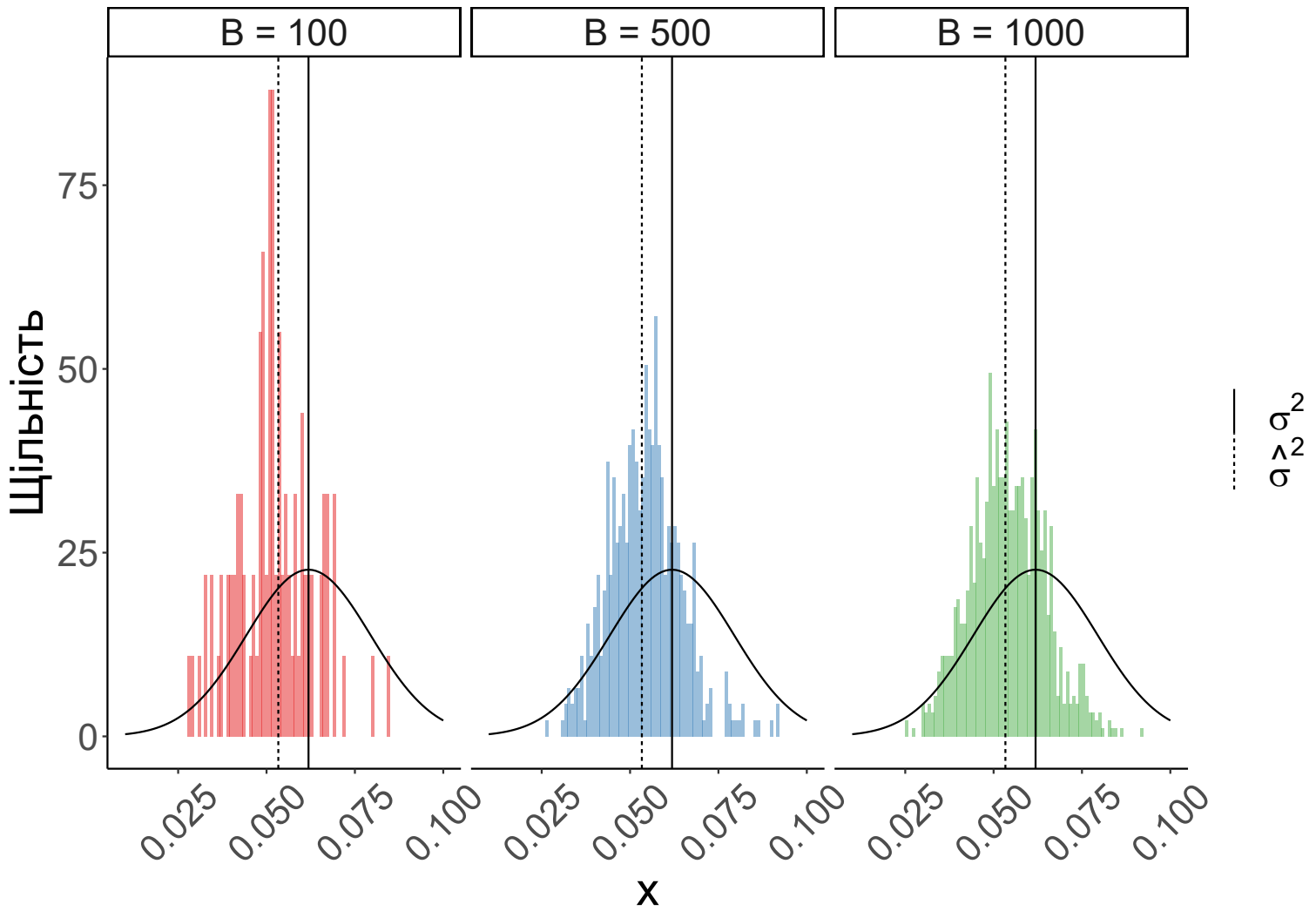
- Як можна бачити, у цьому випадку обидві оцінки є фактично незміщені
  - Вибіркове середнє — за теорією
  - Вибіркова дисперсія — тому що $-\frac{1}{n}\hat{\sigma}^2 \approx -5.3 \times 10^{-4}$

Ілюстрація бутстрепу (4) — розподіл $\overline{X}$



- Можна помітити, що оцінка $\overline{X}$ є незміщеною
  - Справді, відповідні бутстреп-розподіли центровані навколо $\overline{X}$
  - Яке є справжнім сподіванням для $\hat{\mathbb{P}}_X$
- Також зі збільшенням B підвищується якість гістограм бутстреп-оцінок
- А самі гістограми прямують до нормального розподілу з центром в $\overline{X}$ та дисперсією, дуже подібною на $\mathrm{Var}_{\mathbb{P}_X}\left(\overline{X}\right)$

Ілюстрація бутстрепу (5) — розподіл $\hat{\sigma}^2$



- Висновки аналогічні попередньому випадку

Ілюстрація бутстрепу з функцією `boot` (1)

- Усі ці розрахунки можна повторити, використовуючи функцію `boot` із пакету `boot`
- Спочатку потрібно створити функцію з двома параметрами — дані та вектор індексів

```
boot_mean_var <- function(x, indices){
  n <- length(x)
  return(c(mean(x[indices]), (n - 1)/n * var(x[indices])))
}
```

- Тепер можемо запустити весь процес

```
n <- 100
Bs <- c(100, 500, 1000)
lambda <- 4
set.seed(100)
x <- rexp(n, rate = lambda)

boot_result_meanvar_1 <- boot(x, statistic = boot_mean_var, R = Bs[1])
boot_result_meanvar_2 <- boot(x, statistic = boot_mean_var, R = Bs[2])
boot_result_meanvar_3 <- boot(x, statistic = boot_mean_var, R = Bs[3])
```

- Вихідний об'єкт має клас `boot` і містить, серед іншого, такі корисні поля:
  - `t0` — значення $\hat{\theta}(\mathbf{X})$
  - `t` — матриця з усіма $\hat{\theta}_b^*$

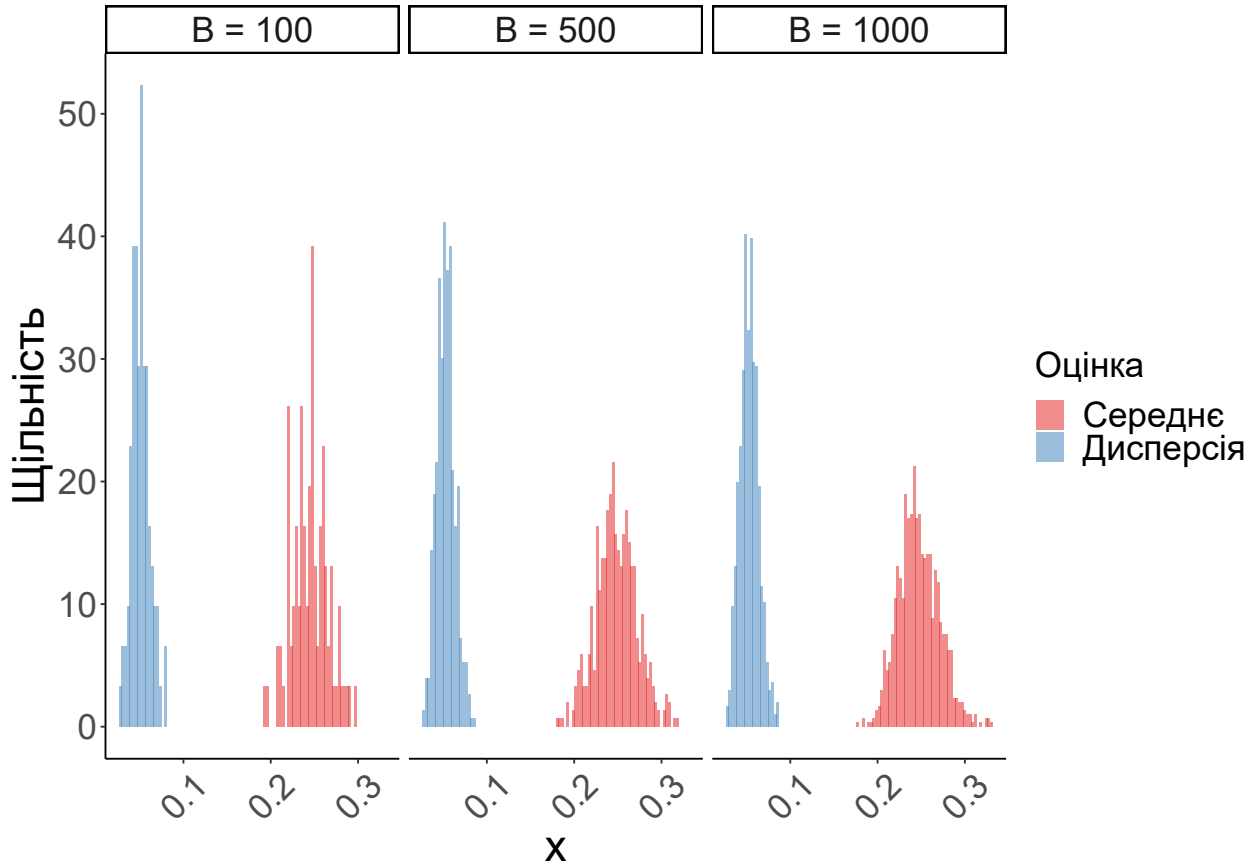
Ілюстрація бутстрепу з функцією `boot` (2)

- Так, для нашого прикладу маємо:

```
df_boot <- tibble(stat = c(boot_result_meanvar_1$t[, 1], boot_result_meanvar_1$t[, 2],
                           boot_result_meanvar_2$t[, 1], boot_result_meanvar_2$t[, 2],
                           boot_result_meanvar_3$t[, 1], boot_result_meanvar_3$t[, 2]),
                  type = c(rep("mean", Bs[1]), rep("var", Bs[1]),
                           rep("mean", Bs[2]), rep("var", Bs[2]),
                           rep("mean", Bs[3]), rep("var", Bs[3])),
                  B = c(rep(Bs[1], 2*Bs[1]), rep(Bs[2], 2*Bs[2]), rep(Bs[3], 2*Bs[3])))
```

Ілюстрація бутстрепу з функцією `boot` (3)

- Відповідна візуалізація:



- Можна бачити, що розподіли схожі на здобуті вище
- Також можемо бачити, що ці розподіли прямують до нормальних
  - Згадайте, що це розподіли бутстреп-оцінок $\hat{\theta}(\mathbf{X}^*)$, а не наших початкових
- Є підстави вважати, що й розподіли початкових оцінок асимптотично нормальні
  - Це ми знаємо з теорії!

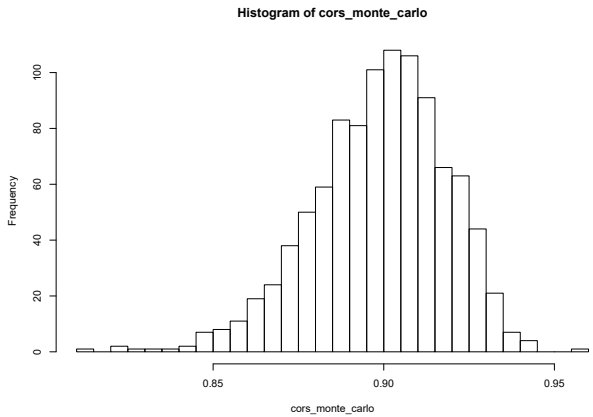
Застосування бутстрепу до багатовимірного розподілу (1)

- До цього моменту ми розглядали ситуацію, коли є наявний деякий DGP $\mathbb{P}_X$, за допомогою якого згенеровано вибірку $X_1, \dots, X_n$
  - На основі цієї вибірки ми обчислюємо деяку статистику $\hat{\theta}$
  - За допомогою бутстрепу ми обчислюємо її стандартну похибку
  - Для цього ми замінюємо $\mathbb{P}_X$ на емпіричний аналог $\hat{\mathbb{P}}_X$
  - За допомогою $\hat{\mathbb{P}}_X$ генеруємо B вибірок, для кожної з яких рахуємо значення $\hat{\theta}^*$
  - Середньоквадратичне відхилення $\hat{\sigma}$ для набору значень $\hat{\theta}^*$ вважаємо приблизно рівним $\widehat{\text{se}}_{\mathbb{P}_X}\left(\hat{\theta}\right)$
- У цю ж схему вкладається й обчислення статистик на основі вибірок із **багатовимірних розподілів**
  - Наприклад, для обчислення кореляції між випадковими величинами X та Y, які утворюють випадковий вектор $(X, Y)^\top$
  - Тоді DGP є **спільний розподіл** $\mathbb{P}_{XY}$ цього вектора
  - Нехай маємо вибірку $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$, яку можна подати як матрицю $n \times 2$
  - Емпіричним аналогом буде розподіл $\hat{\mathbb{P}}_{XY}$, який кожному **рядку** матриці зіставляє ймовірність $\frac{1}{n}$
  - Тоді генерування бутстреп-вибірок потрібно здійснювати шляхом випадкового вибору **рядків** матриці з повтореннями

Застосування бутстрепу до багатовимірного розподілу (2)

- Розгляньмо бутстреп для вибіркової кореляції для багатовимірного нормального розподілу з кореляцією 0.9
- Оскільки ми не знаємо теоретичного асимптотичного розподілу, то ми його просимулюємо за допомогою методу Монте-Карло

```
n <- 100
Bs <- c(100, 500, 1000)
cor_true <- 0.9
set.seed(100)
Sigma <- matrix(c(1, cor_true,
                  cor_true, 1),
                nrow = 2, ncol = 2, byrow = TRUE)

boot_cor <- function(X, indices = 1:nrow(X)){
  return(cor(X[indices, ])[1, 2])
}

cors_monte_carlo <- replicate(1000, boot_cor(MASS::mvrnorm(n, mu = c(0, 0), Sigma = Sigma)))
hist(cors_monte_carlo, breaks = 30)
```



- Як можна бачити, розподіл не є нормальним (скошений уліво)

Застосування бутстрепу до багатовимірного розподілу (3)

- Тепер обчислімо бутстреп-кореляції:

```
set.seed(100)
X <- MASS::mvrnorm(n, mu = c(0, 0), Sigma = Sigma)
boot_result_cor_1 <- boot(X, statistic = boot_cor, R = Bs[1])
boot_result_cor_2 <- boot(X, statistic = boot_cor, R = Bs[2])
boot_result_cor_3 <- boot(X, statistic = boot_cor, R = Bs[3])

df_boot <- tibble(cor = c(boot_result_cor_1$t, boot_result_cor_2$t, boot_result_cor_3$t),
                  B = c(rep(Bs[1], Bs[1]), rep(Bs[2], Bs[2]), rep(Bs[3], Bs[3])))
```

- Якщо на вхід `boot` подати матрицю або датафрейм, то вона сприйматиме кожний **рядок** як спостереження
  - Це дуже зручно
- Можна вказати аргумент `simple = TRUE`, щоб генерувати індекси нової вибірки окремо для кожної вибірки
  - Це може зекономити пам'ять, якщо початковий датасет та R дуже великі

Застосування бутстрепу до багатовимірного розподілу (4)

- Можемо порівняти стандартні похибки і зміщення
- На основі Монте-Карло ми бачимо, що «теоретичні» такі значення

```
mean(cors_monte_carlo)
```

```
## [1] 0.8987286
```

```
sd(cors_monte_carlo)
```

```
## [1] 0.0196371
```

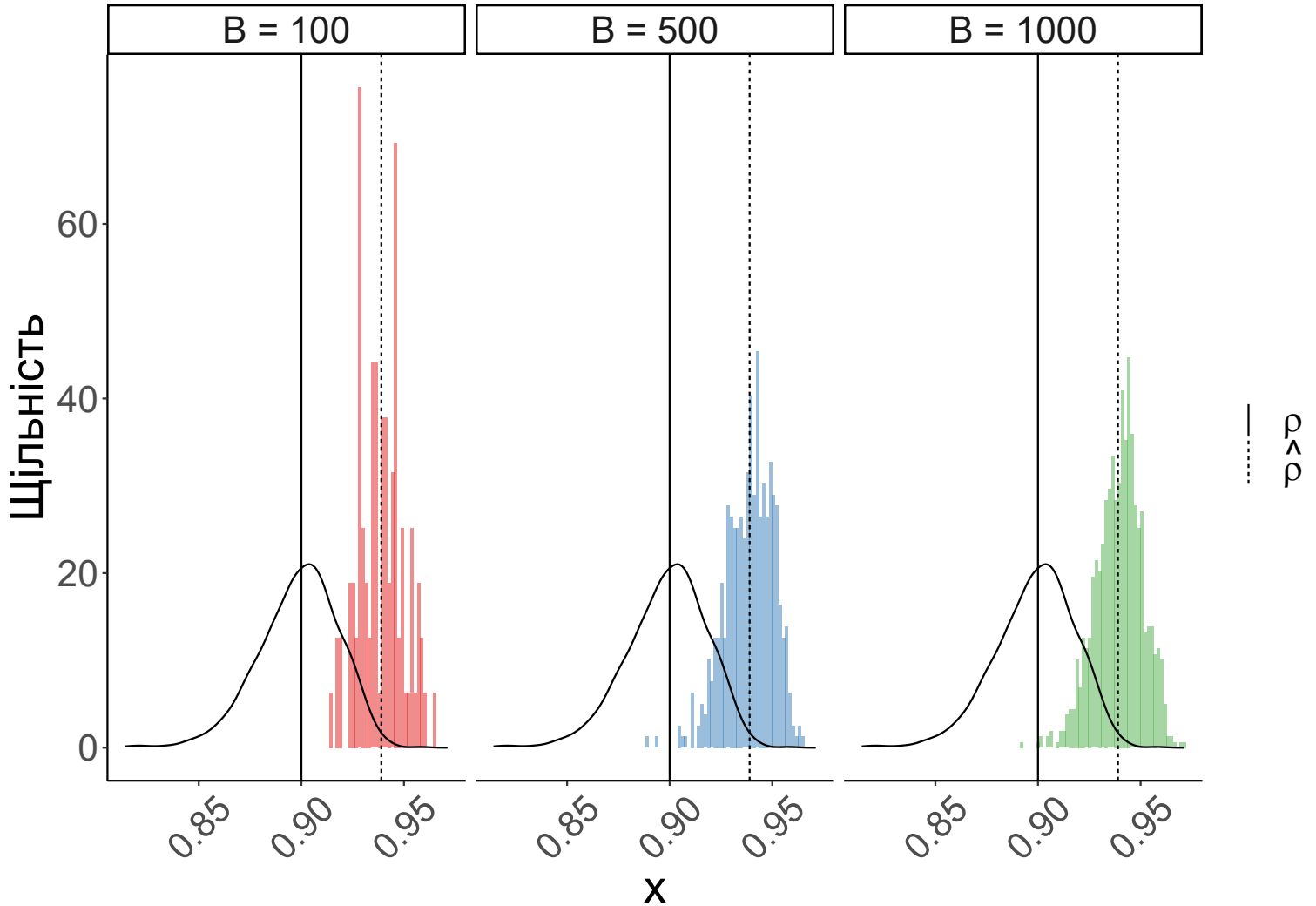
- Тобто в принципі оцінка є незміщена
  - А її стандартна похибка приблизно дорівнює 0.02
- На основі бутстреп-розподілу маємо такі результати:

```
df_boot %>% group_by(B) %>% summarize(se_cor = sd(cor), bias_cor = mean(cor) - boot_cor(X))
```

```
## # A tibble: 3 x 3
##       B se_cor   bias_cor
##   <dbl>  <dbl>      <dbl>
## 1   100 0.0109 -0.000222
## 2   500 0.0115 -0.0000594
## 3  1000 0.0114  0.000544
```

- Вони дуже подібні до «теоретичних»

Застосування бутстрепу до багатовимірного розподілу (5)



- Тут чорним наведено оцінку щільности розподілу з симуляції Монте-Карло
- Можна бачити, що бутстреп-розподіл подібний до справжнього

Застосування бутстрепу до складних статистик (1)

- Розгляньмо бутстреп для такої цікавої статистики
- Нехай нас цікавить інвестування фіксованої суми грошей у два фінансові інструменти з прибутковістю X та Y відповідно
  - X та Y випадкові
- В актив X інвестуватимемо частку α коштів, а в актив Y — частку $1 - \alpha$
- Ми хочемо підібрати таке α, щоб **дисперсію** всього портфеля було **мінімізовано**:

$$\alpha = \arg\min_a \operatorname{Var}\left(aX + (1-a)Y\right)$$

- Можна показати[2], що

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

- Вочевидь, plug-in оцінкою цієї статистики буде

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

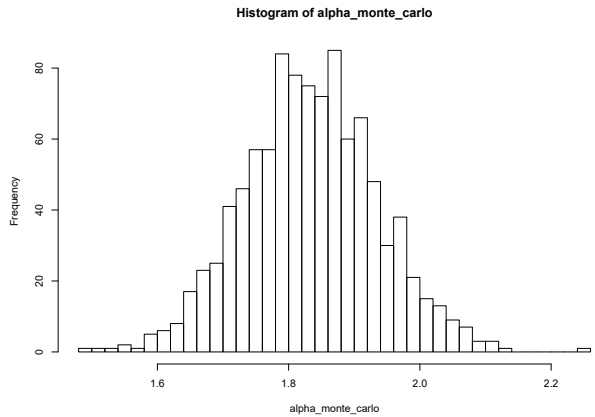
- Цілком очевидно, що розподіл такої статистики є, м'яко кажучи, непростим

[2] Покажіть, це справді нескладно!

Застосування бутстрепу до складних статистик (2)

- Ми його просимулюємо за допомогою методу Монте-Карло

```
n <- 100
Bs <- c(100, 500, 1000)
set.seed(100)
Sigma <- matrix(c(0.5, 0.75,
                  0.75, 1.3),
                nrow = 2, ncol = 2, byrow = TRUE)
alpha_true <- (1.3 - 0.75) / (0.5 + 1.3 - 2*0.75)

boot_alpha <- function(X, indices = 1:nrow(X)){
  x <- X[indices, 1]
  y <- X[indices, 2]
  return((var(y) - cov(x, y)) / (var(x) + var(y) - 2*cov(x, y)))
}

alpha_monte_carlo <- replicate(1000, boot_alpha(MASS::mvrnorm(n, mu = c(0, 0), Sigma = Sigma)))
hist(alpha_monte_carlo, breaks = 30)
```



- Справжнє значення дорівнює $\alpha = 1.833$
- Ми використовуємо вбудовані статистики `var` та `cov`, оскільки коефіцієнт $\frac{n-1}{n}$ усе одно буде скорочено

Застосування бутстрепу до багатовимірного розподілу (3)

- Тепер обчислімо значення бутстреп-оцінок:

```
set.seed(100)
X <- MASS::mvrnorm(n, mu = c(0, 0), Sigma = Sigma)
boot_result_alpha_1 <- boot(X, statistic = boot_alpha, R = Bs[1])
boot_result_alpha_2 <- boot(X, statistic = boot_alpha, R = Bs[2])
boot_result_alpha_3 <- boot(X, statistic = boot_alpha, R = Bs[3])

df_boot <- tibble(alpha = c(boot_result_alpha_1$t, boot_result_alpha_2$t, boot_result_alpha_3$t),
                  B = c(rep(Bs[1], Bs[1]), rep(Bs[2], Bs[2]), rep(Bs[3], Bs[3])))
```

- На основі Монте-Карло ми бачимо, що «теоретичні» значення стандартних похибок і зміщень такі:

```
mean(alpha_monte_carlo)
```

```
## [1] 1.833659
```

```
sd(alpha_monte_carlo)
```

```
## [1] 0.1016502
```

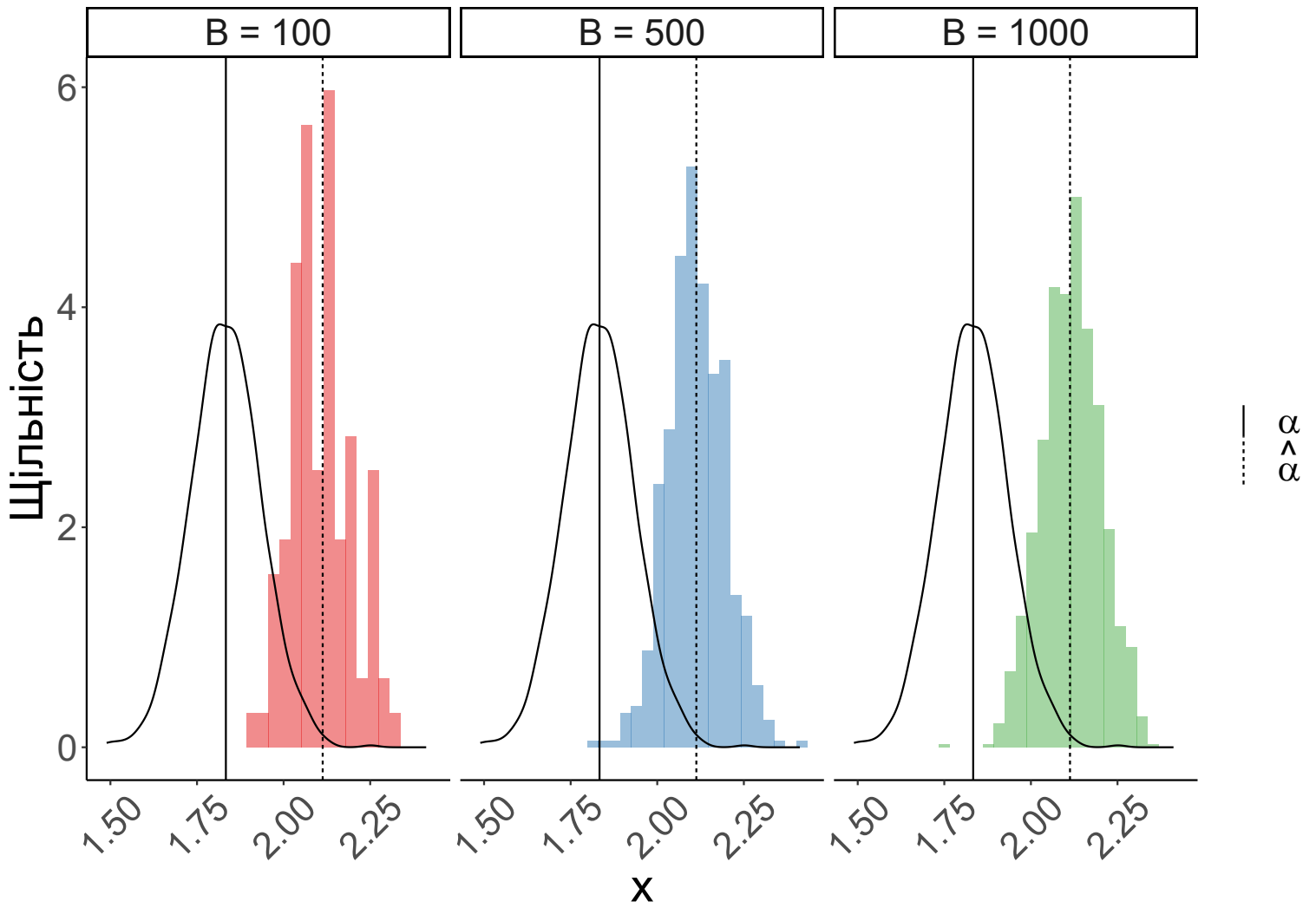
- Тобто в принципі оцінка є незміщена
- На основі бутстреп-розподілу маємо такі результати:

```
df_boot %>% group_by(B) %>% summarize(se_alpha = sd(alpha), bias_alpha = mean(alpha) - boot_alpha(X))
```

```
## # A tibble: 3 x 3
##       B se_alpha bias_alpha
##   <dbl>    <dbl>      <dbl>
## 1   100   0.0868  -0.00393
## 2   500   0.0850  -0.000218
## 3  1000   0.0854   0.00383
```

- Вони дуже подібні до «теоретичних»

Застосування бутстрепу до багатовимірного розподілу (4)



Особливості застосування бутстрепу для декількох незалежних вибірок (1)

- Нехай тепер стоїть задача оцінити **різницю** деяких статистик двох (або більше) **незалежних** вибірок **різного розміру**
  - Розгляньмо приклад різниці медіан
- У цьому випадку DGP складається з двох **окремих** розподілів, $\mathbb{P}_X$ та $\mathbb{P}_Y$ таких, що $X \sim \mathbb{P}_X$, $Y \sim \mathbb{P}_Y$, $X \perp\!\!\!\perp Y$
- Відтак для генерування бутстреп-вибірок потрібно замінити сам такий DGP на емпіричний
  - Тобто розглянути $\hat{\mathbb{P}}_X$ та $\hat{\mathbb{P}}_Y$ як емпіричні аналоги на основі відповідних вибірок $\mathbf{X}$ та $\mathbf{Y}$...
  - ...і генерувати нові вибірки $\mathbf{X}^*$ і $\mathbf{Y}^*$ **незалежно** одну від одної
  - Але кожну як випадковий вибір із повтореннями з $\mathbf{X}$ та $\mathbf{Y}$ відповідно

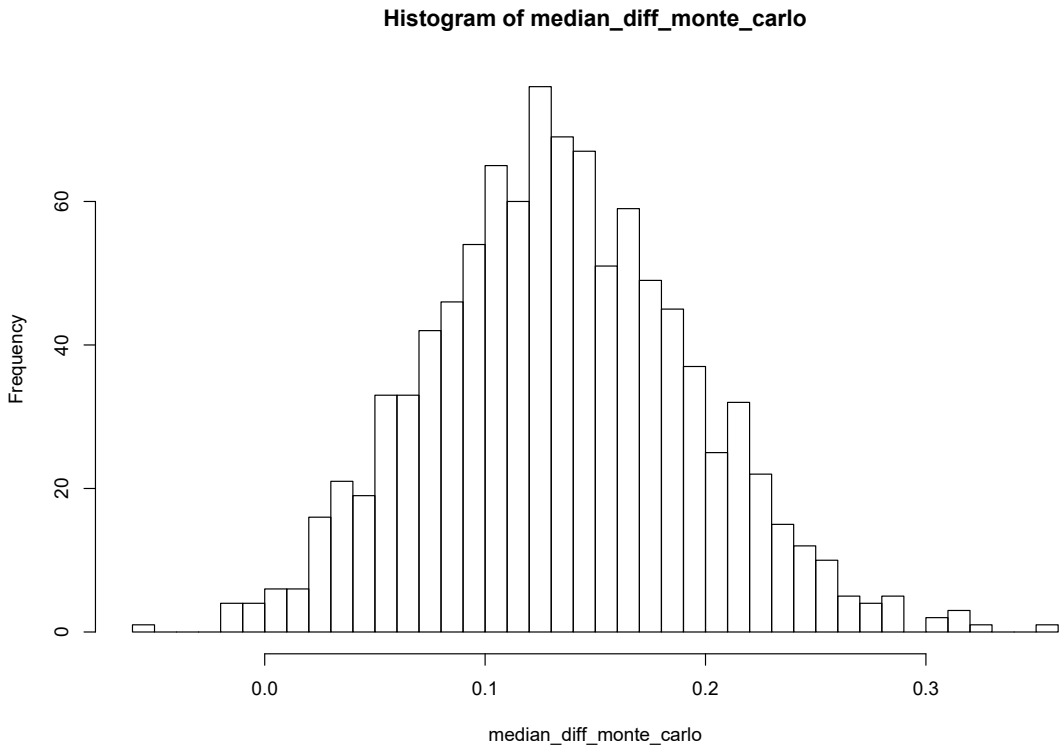
Особливості застосування бутстрепу для декількох незалежних вибірок (2)

- Розгляньмо різницю медіан двох розподілів — гамма $X \sim \text{Gamma}(2, 2)$ та бета $Y \sim \text{Beta}(2, 1)$
  - У цьому випадку справжня різниця двох медіан дорівнює $M_X - M_Y \approx 0.132$
- Просимулюймо асимптотичний розподіл за допомогою методу Монте-Карло

```
n <- 200
Bs <- c(100, 500, 1000)
set.seed(100)
shape <- 2
rate <- 2
a <- 2
b <- 1
median_diff_true <- qgamma(0.5, shape = 2, rate = 2) - qbeta(0.5, 2, 1)

boot_median_diff <- function(x, indices){
  n_sample1 <- 1:table(x$sample)[1]
  indices_sample1 <- indices[n_sample1]
  indices_sample2 <- indices[-(n_sample1)]

  m1 <- median(x[indices_sample1, 1])
  m2 <- median(x[indices_sample2, 1])

  return(m1 - m2)
}
```

Особливості застосування бутстрепу для декількох незалежних вибірок (3)

```
median_diff_monte_carlo <- replicate(
  1000, median(rgamma(n, shape = shape, rate = rate)) - median(rbeta(n, a, b))
)
hist(median_diff_monte_carlo, breaks = 30)
```



- Як можна бачити, розподіл є близький до нормального
- Це очікувано, адже обидві медіани повинні мати асимптотично нормальний розподіл

Особливості застосування бутстрепу для декількох незалежних вибірок (4)

- Тепер обчислімо бутстреп-різниці:

```
x <- rgamma(n, shape = shape, rate = rate)
y <- rbeta(n, a, b)
dat <- data.frame(values = c(x, y), sample = rep(0:1, each = n))

boot_result_median_diff_1 <- boot(dat, statistic = boot_median_diff,
                                  R = Bs[1], strata = dat$sample)
boot_result_median_diff_2 <- boot(dat, statistic = boot_median_diff,
                                  R = Bs[2], strata = dat$sample)
boot_result_median_diff_3 <- boot(dat, statistic = boot_median_diff,
                                  R = Bs[3], strata = dat$sample)

df_boot <- tibble(median_diff = c(boot_result_median_diff_1$t,
                                  boot_result_median_diff_2$t,
                                  boot_result_median_diff_3$t),
                  B = c(rep(Bs[1], Bs[1]), rep(Bs[2], Bs[2]), rep(Bs[3], Bs[3])))
```

Особливості застосування бутстрепу для декількох незалежних вибірок (5)

- Можемо порівняти стандартні похибки і зміщення
- На основі Монте-Карло ми бачимо, що «теоретичні» такі значення

```
mean(median_diff_monte_carlo)
```

```
## [1] 0.1343268
```

```
sd(median_diff_monte_carlo)
```

```
## [1] 0.06021779
```

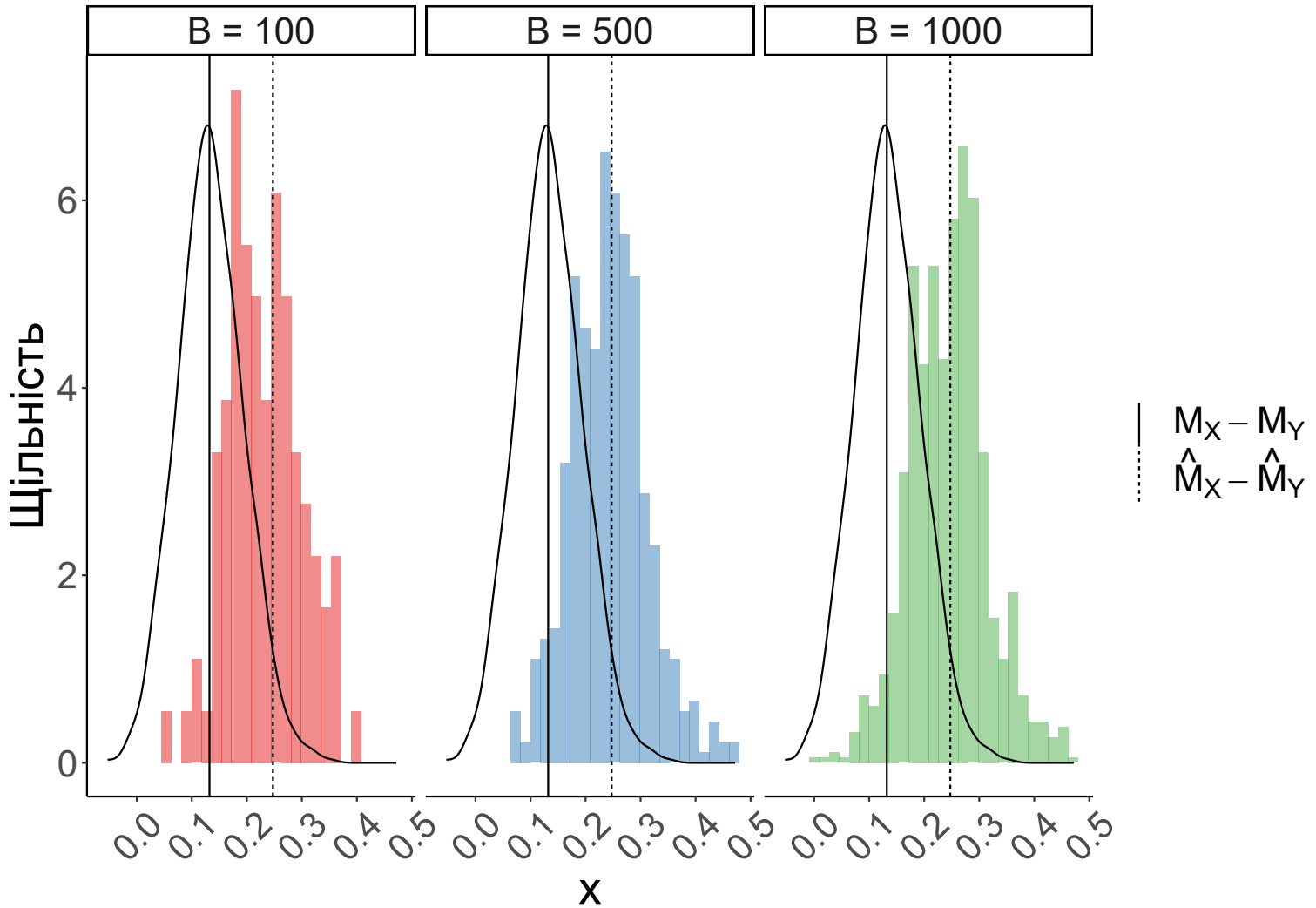
- Тобто в принципі оцінка є незміщена
- На основі бутстреп-розподілу маємо такі результати:

```
df_boot %>% group_by(B) %>%
  summarize(se_median_diff = sd(median_diff),
            bias_median_diff = mean(median_diff) - (median(x) - median(y)))
```

```
## # A tibble: 3 x 3
##       B se_median_diff bias_median_diff
##   <dbl>          <dbl>            <dbl>
## 1   100         0.0685         -0.0157
## 2   500         0.0697         -0.00531
## 3  1000         0.0707         -0.00292
```

- Вони дуже подібні до «теоретичних»

Особливості застосування бутстрепу для декількох незалежних вибірок (6)



Мотивація

- До цього ми переважно займалися питаннями бутстреп-оцінювання стандартної похибки деякої оцінки
- Але самі по собі стандартні похибки мають мало користи, адже значно важливіше мати довірчий інтервал
- Звісно, якщо (асимптотичний) розподіл деякої оцінки є нормальний, то стаднартні похибки можна використати для побудови такого інтервалу
- Проте ми вже бачили випадки, коли розподіли оцінок не є нормальними
- Понад те, часто на практиці корисніше бути агностиками і не покладатися на асимптотичні властивості оцінок
- Тому далі розглянемо способи побудови довірчих інтервалів за допомогою бутстрепу

Пивотальні довірчі інтервали (1)

- Розгляньмо величину $R_n = \hat{\theta}(\mathbf{X}) - \theta$
  - Такі величини називають **пивотальними** (pivotal), оскільки їхні розподіли не залежать від невідомих параметрів
  - Але це не є статистика, бо вона сама залежить від невідомого θ
- Нехай функція розподілу R_n дорівнює $H(r) = \mathbb{P}_X\left(R_n \leq r\right)$
  - Ми її не знаємо
- Тоді нехай $a = H^{-1}\left(\frac{\alpha}{2}\right), b = H^{-1}\left(1 - \frac{\alpha}{2}\right)$
- Легко показати, що довірчий інтервал

$$C_{1-\alpha} = \left[\hat{\theta}(\mathbf{X}) - b; \hat{\theta}(\mathbf{X}) - a\right] \tag{3.1}$$

  покриває θ з імовірністю $1 - \alpha$
- Справді,

$$\begin{aligned}\mathbb{P}_X\left(\theta \in \left[\hat{\theta}(\mathbf{X}) - b; \hat{\theta}(\mathbf{X}) - a\right]\right) &= \mathbb{P}_X\left(a \leq \hat{\theta}(\mathbf{X}) - \theta \leq b\right) \\ &= \mathbb{P}_X\left(R_n \leq b\right) - \mathbb{P}_X\left(R_n \leq a\right) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha\end{aligned}$$

Пивотальні довірчі інтервали (2)

- Проте на практиці ми не знаємо справжнього розподілу H
  - Принаймні тому, що θ невідоме
  - Також можуть бути проблеми, якщо розподіл $\hat{\theta}(\mathbf{X})$ складний або невідомий
- Тоді можна застосувати **бутстреп** і замінити H на $\hat{H}$:

$$\hat{H}(r) = \hat{\mathbb{P}}_X \left(\hat{\theta}(\mathbf{X}^*) - \hat{\theta}(\mathbf{X}) \leq r \right)$$

- Нехай $r^*_\alpha = \hat{H}^{-1}(\alpha)$ — α-квантиль $\hat{H}$
- Із теорії ймовірностей відомо, що якщо q є квантилем X, то $h(q)$ є квантилем $h(X)$, якщо h монотонно зростаюча
- Звідси випливає, що $r^*_\alpha = q_{\hat{\theta}^*,\alpha} - \hat{\theta}(\mathbf{X})$, де $q_{\hat{\theta}^*,\alpha}$ — α-квантиль **емпіричного** розподілу $\hat{\theta}^*_1, \dots, \hat{\theta}^*_B$
- Звідси

$$\hat{a} = r^*_{\frac{\alpha}{2}} = q_{\hat{\theta}^*,\frac{\alpha}{2}} - \hat{\theta}(\mathbf{X})$$

$$\hat{b} = r^*_{1-\frac{\alpha}{2}} = q_{\hat{\theta}^*,1-\frac{\alpha}{2}} - \hat{\theta}(\mathbf{X})$$

Пивотальні довірчі інтервали (3)

- І тоді наш **пивотальний довірчий інтервал** (pivotal confidence interval) дорівнює

$$\hat{C}^P_{1-\alpha} = \left[\hat{\theta}(\mathbf{X}) - \hat{b}; \hat{\theta}(\mathbf{X}) - \hat{a}\right] = \left[2\hat{\theta}(\mathbf{X}) - q_{\hat{\theta}^*, 1-\frac{\alpha}{2}}; 2\hat{\theta}(\mathbf{X}) - q_{\hat{\theta}^*, \frac{\alpha}{2}}\right] \tag{3.2}$$

  - Також такі інтервали інколи називають **базовими бутстреп-інтервалами** (basic bootstrap intervals)
- Можна довести, що для «адекватних» $T(\mathbb{P}_X)$ цей інтервал спроможний для $C_{1-\alpha}$ з (3.1)
- Якщо оцінка не є plug-in оцінкою, наприклад, якась $\tilde{\theta}(\mathbf{X})$, то $\hat{H}(r) = \hat{\mathbb{P}}_X\left(\tilde{\theta}(\mathbf{X}^*) - T\left(\hat{\mathbb{P}}_X\right) \leq r\right)$ з відповідними наслідками

Студентизовані пивотальні довірчі інтервали (1)

- Розгляньмо тепер пивотальну величину $Z_n = \frac{\hat{\theta}(\mathbf{X}) - \theta}{\text{se}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}\right)}$ з (невідомою) функцією розподілу G[3]
- Тоді, за аналогією з попереднім випадком, нас повинен цікавити інтервал

$$C_{1-\alpha} = \left[\hat{\theta}(\mathbf{X}) - \text{se}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}\right) \cdot G^{-1}\left(1 - \frac{\alpha}{2}\right); \hat{\theta}(\mathbf{X}) - \text{se}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}\right) \cdot G^{-1}\left(\frac{\alpha}{2}\right)\right]$$

- Оскільки G невідомий, ми застосовуємо бутстреп і замінюємо його на емпіричний $\hat{G}$:

$$\hat{G}(r) = \hat{\mathbb{P}}_X\left(\frac{\hat{\theta}(\mathbf{X}^*) - \hat{\theta}(\mathbf{X})}{\text{se}_{\hat{\mathbb{P}}_X^b}\left(\hat{\theta}(\mathbf{X}^*)\right)} \leq r\right)$$

[3] Далі вважатимемо, що $\hat{\theta}$ є plug-in оцінкою

Студентизовані пивотальні довірчі інтервали (2)

- Тут $\text{se}_{\hat{\mathbb{P}}^b_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$ є оцінкою стандартної похибки **не** $\hat{\theta}(\mathbf{X})$**, а саме** $\hat{\theta}(\mathbf{X}^*)$
  - Тобто в загальному випадку потрібно для кожної бустреп-вибірки $\mathbf{X}^*_b, b = 1, \dots, B$ **додатково** запускати бутстреп
  - Тобто потрібно генерувати нові вибірки $\mathbf{X}^*_{b,1}, \dots, \mathbf{X}^*_{b,B^*}$ шляхом випадкового вибору з повторенням **уже не з** $\mathbf{X}$**, а з** $\mathbf{X}^*$
  - Тоді можна обчислити $\text{se}_{\hat{\mathbb{P}}^b_X}\left(\hat{\theta}(\mathbf{X}^*)\right)$ як вибіркове середньоквадратичне відхилення оцінок на нових вибірках
  - Кількість B^* вибірок у «внутрішньому» бутстрепі може бути меншою від B
  - Як правило, для оцінювання квантилів потрібно більше вибірок, ніж для стандартних похибок
- Тоді остаточно

$$\hat{C}^{PS}_{1-\alpha} = \left[\hat{\theta}(\mathbf{X}) - \text{se}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}\right) \cdot \hat{G}^{-1}\left(1 - \frac{\alpha}{2}\right); \hat{\theta}(\mathbf{X}) - \text{se}_{\hat{\mathbb{P}}_X}\left(\hat{\theta}\right) \cdot \hat{G}^{-1}\left(\frac{\alpha}{2}\right)\right] \quad (3.3)$$

- Можна показати, що цей інтервал також спроможний
- Згідно з *An Introduction to the Bootstrap*, студентизовані інтервали найліпше працюють для мір центральної тенденції (сподівань, медіан тощо)
  - З іншими статистиками потрібно бути обережними

Персентильні довірчі інтервали (1)

- Проста ідея побудови довірчого інтервалу полягає в таких міркуваннях
- Нехай $\hat{\theta} \xrightarrow{d} N\left(\theta, \text{Var}\left(\hat{\theta}\right)\right)$
- Тоді (асимптотичний) довірчий інтервал рівня $1-\alpha$ дорівнює

$$\left[\hat{\theta} - z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{se}}\left(\hat{\theta}\right); \hat{\theta} + z_{\frac{\alpha}{2}} \cdot \widehat{\text{se}}\left(\hat{\theta}\right)\right]$$

- Розгляньмо випадкову величину $\hat{\theta}^* \sim N\left(\hat{\theta}, \text{Var}\left(\hat{\theta}\right)\right)$
- Тоді **персентильним довірчим інтервалом** (percentile confidence interval) є

$$C_{1-\alpha}^{Perc} = \left[q_{\hat{\theta}^*, \frac{\alpha}{2}}; q_{\hat{\theta}^*, 1-\frac{\alpha}{2}}\right] \tag{3.4}$$

Персентильні довірчі інтервали (2)

- Якщо розподіл бутстреп-оцінок приблизно нормальний, то персентильний інтервал не сильно відрізнятиметься від асимптотичного нормального
- Але якщо розподіл бутстреп-оцінок не є нормальним, що дає підстави стверджувати, що персентильний інтервал може бути адекватним?
- Уявімо, що існує монотонне відображення m таке, що $U = m(T)$, $U \sim N(m(\theta), c^2)$
  - Це відображення нам невідоме
- Нехай $U_b^* = m\left(\hat{\theta}_b^*\right)$
- Тоді, оскільки m монотонне, аналогічне співвідношення існуватиме й між квантилями відповідних розподілів: $q_{U_b^*,\alpha} = m\left(q_{\hat{\theta}_b^*,\alpha}\right)$
- А оскільки $U \sim N(m(\theta), c^2)$, маємо, що α-квантилем U є величина $m(\theta) - z_\alpha c$
- Звідси випливає, що $q_{U_b^*,\frac{\alpha}{2}} = m(\theta) - z_{\frac{\alpha}{2}} c \approx U - z_{\frac{\alpha}{2}} c$
  - А $q_{U_b^*,1-\frac{\alpha}{2}} \approx U + z_{\frac{\alpha}{2}} c$

Персентильні довірчі інтервали (3)

- Тому остаточно маємо

$$\begin{aligned}
\hat{\mathbb{P}}_X\left(C^{Perc}_{1-\alpha} \ni \theta\right) &= \hat{\mathbb{P}}_X\left(q_{\hat{\theta}^*_b,\frac{\alpha}{2}} \le \theta \le q_{\hat{\theta}^*_b,1-\frac{\alpha}{2}}\right) \\
&= \hat{\mathbb{P}}_X\left(m\left(q_{\hat{\theta}^*_b,\frac{\alpha}{2}}\right) \le m(\theta) \le m\left(q_{\hat{\theta}^*_b,1-\frac{\alpha}{2}}\right)\right) \\
&= \hat{\mathbb{P}}_X\left(q_{U^*_b,\frac{\alpha}{2}} \le m(\theta) \le q_{U^*_b,1-\frac{\alpha}{2}}\right) \\
&\approx \mathbb{P}_X\left(U - z_{\frac{\alpha}{2}}\, c \le m(\theta) \le U + z_{\frac{\alpha}{2}}\, c\right) \\
&= \mathbb{P}_X\left(-z_{\frac{\alpha}{2}} \le \frac{U - m(\theta)}{c} \le z_{\frac{\alpha}{2}}\right) = 1-\alpha
\end{aligned}$$

- Що цікаво в цих викладках — нам непотрібно **знати** перетворення m
- Достатньо тільки припускати, що воно в принципі існує

Персентильні довірчі інтервали (4)

- Простий приклад для ілюстрації цих міркувань
- Нехай нам потрібно оцінити $\theta = e^{\mu}$, де μ — сподівання нормального розподілу
  - Цілком очевидно, що plug-in оцінка $\hat{\theta} = e^{\overline{X}}$ **не буде** мати нормальний розподіл
  - Тому асимптотичний нормальний інтервал буде неточний
- Проте якщо взяти логаритм: $m\left(\hat{\theta}\right) = \overline{X}$, то відповідні оцінки **будуть** мати нормальний розподіл
- Персентильний довірчий інтервал дає змогу автоматично враховувати перетворення на кшталт m, якщо вони справді існують
- Також інша корисна властивість персентильних інтервалів — збереження розмаху
  - Наприклад, ми знаємо, що коефіцієнт кореляції повинен лежати на проміжку $[-1; 1]$
  - Персентильний довірчий інтервал ніколи не вискочить за ці межі
    - (а) оцінка $\hat{\rho}$ не може за них вискочити
    - (б) межі довірчого інтервалу будуються як квантилі $\hat{\rho}^*$, тому також не можуть вискочити за межі $[-1; 1]$

Метод BC$_a$

- Пивотальні (студентизовані) довірчі інтервали мають добрі теоретичні властивості покриття, але на практиці можуть поводити себе погано
- Персентильні довірчі інтервалі стійкіші в цьому сенсі, але можуть бути занадто консервативні
- Існує ще один популярний метод побудови довірчих бутстреп-інтервалів — **метод BC$_a$** (від bias-corrected and accelerated)
- Ми не будемо детально розглядати теоретичні засади цього методу
- Якщо коротко, то суть полягає в тому, що потрібно побудувати персентильний інтервал, але вибрати не квантилі $\frac{\alpha}{2}$ і $1 - \frac{\alpha}{2}$, а квантилі, які рахують за окремими формулами
- Нові квантилі повинні коригувати зміщення бутстреп-оцінки та швидкість зміни стандартної похибки $\widehat{\theta}$ зі зміною параметра θ

Порівняння різних методів

- Можна показати, що асимптотичний нормальний, пивотальний і персентильний інтервали є інтервалами **першого порядку** точности (first-order accurate), а студентизований пивотальний і BC_a — **другого** (second-order accurate)
  - Мається на увазі, що зі збільшенням n справжнє покриття інтервалу прямує до α зі швидкістю $O(n^{-1/2})$ для інтервалів першого порядку, і зі швидкістю $O(n^{-1})$ — для другого
- Студентизований інтервал, хоча й другого порядку точности, але не є інваріантним до перетворень статистики
  - Якщо будувати інтервал не для $\hat{\theta}$, а для $g\left(\hat{\theta}\right)$, то між цими інтервалами не буде прямого зв'язку
- Персентильний інтервал і BC_a-інтервал є інваріантними до перетворень статистики
  - Інтервал для, скажімо, $\ln\left(\hat{\theta}\right)$ можна дістати застосуванням логаритму до меж інтервалу для $\hat{\theta}$

Рекомендації щодо значення B

- Можна сформулювати такі загальні рекомендації щодо числа бутстреп-вибірок B[4]
  - Для оцінювання параметрів достатньо брати $B = 200$
  - Для довірчих інтервалів рівня 0.9 потрібно брати B від 1000 до 2000
  - Для довірчих інтервалів рівня 0.99 потрібно брати B понад 5000
- У загальному випадку, можна взяти мале B і поступово збільшувати, якщо результати не є задовільні
  - Це цілком допустимо, на відміну від p-hacking!
  - Ми не проводимо жодних додаткових аналізів, ми просто уточнюємо параметри одного й того самого статистичного процесу

[4] Florent Buisson, *Data Analysis with R & Python. Customer-Driven Data for Real Business Results*, p. 153

Побудова довірчих бутстреп-інтервалів в R

- Для автоматизації обчислення довірчих інтервалів можна використовувати функцію `boot.ci` з пакету `boot`
- На вхід цієї функції потрібно подати такі обов'язкові аргументи:
  - Перший аргумент — результат роботи функції `boot` (об'єкт класу `boot`)
  - `conf` — рівень інтервалу (за замовчуванням `conf = 0.95`)
  - Тип інтервалу — можна вказувати такі типи:
    - `norm` — асимптотичний нормальний інтервал
    - `basic` — пивотальний інтервал
    - `stud` — студентизований пивотальний інтервал
    - `perc` — персентильний інтервал
    - `bca` — BC$_a$-інтервал
    - `all` — усі відразу (значення за замовчуванням)
- Усі інші аргументи можна залишити за замовчуванням, деталі можна знайти в офіційній документації
- Єдиний нюанс — для побудови студентизованих пивотальних інтервалів потрібно вказати оцінки стандартних похибок кожної $\widehat{\theta}_b^*$
  - Для цього потрібно передбачити, щоб `statistic`, використана у виклику функції `boot`, другим результатом повертала цю оцінку, і вказати аргумент `index` (див. далі)
  - Альтернативно можна вказати аргумент `var.t`

Ілюстрація для прикладу з середніми та вибірковими дисперсіями (1)

- Перепишімо нашу функцію для підрахунку середніх і дисперсій, щоб вона повертала оцінку дисперсії
- Для цих статистик дисперсію можна обчислити точно
- Але з педагогічних міркувань ми для всіх наших прикладів оцінку дисперсії $\widehat{\theta}_b^*$ робитимемо за допомогою бутстрепу

```
boot_mean_var_with_sd <- function(x, indices, estimate_var = TRUE, R_for_sd = 200){
  n <- length(x)

  mean_bar <- mean(x[indices])
  var_bar <- (n - 1)/n * var(x[indices])

  if (estimate_var){
    boot_out <- boot(x[indices], statistic = boot_mean_var_with_sd,
                     R = R_for_sd, estimate_var = FALSE)

    return(c(mean_bar, var(boot_out$t[, 1]), var_bar, var(boot_out$t[, 2])))
  }
  else {
    return(c(mean_bar, var_bar))
  }
}
```

- Ми використали 200 бутстреп-вибірок для оцінювання стандартної похибки, адже саме так рекомендують у книжці *An Introduction to the Bootstrap*

Ілюстрація для прикладу з середніми та вибірковими дисперсіями (2)

- Тепер можемо порахувати всі інтервали

```
n <- 200
B <- 2000
lambda <- 4
set.seed(100)

x <- rexp(n, rate = lambda)

boot_result_meanvar <- boot(x, statistic = boot_mean_var_with_sd, R = B)

print("Довірчі бутстреп-інтервали для середнього:")
```

```
## [1] "Довірчі бутстреп-інтервали для середнього:"
```

```
boot.ci(boot_result_meanvar, index = c(1, 2))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_result_meanvar, index = c(1, 2))
##
## Intervals :
## Level      Normal              Basic             Studentized
## 95%   ( 0.1936,  0.2519 )   ( 0.1928,  0.2512 )   ( 0.1954,  0.2539 )
##
## Level     Percentile            BCa
## 95%   ( 0.1932,  0.2516 )   ( 0.1964,  0.2566 )
## Calculations and Intervals on Original Scale
```

```
print("Асимптотичний нормальний довірчий інтервал для середнього:")
```

```
## [1] "Асимптотичний нормальний довірчий інтервал для середнього:"
```

```
sd_mean <- sqrt(var(x) / n)
c(mean(x) + qnorm(0.025)*sd_mean, mean(x) + qnorm(0.975)*sd_mean)
```

```
## [1] 0.1929040 0.2514982
```

Ілюстрація для прикладу з середніми та вибірковими дисперсіями (3)

```r
print("Довірчі бутстреп-інтервали для вибіркової дисперсії:")
```

```
## [1] "Довірчі бутстреп-інтервали для вибіркової дисперсії:"
```

```r
boot.ci(boot_result_meanvar, index = c(3, 4))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_result_meanvar, index = c(3, 4))
##
## Intervals :
## Level      Normal              Basic             Studentized
## 95%   ( 0.0319,  0.0580 )   ( 0.0315,  0.0572 )   ( 0.0337,  0.0615 )
##
## Level     Percentile            BCa
## 95%   ( 0.0318,  0.0575 )   ( 0.0336,  0.0607 )
## Calculations and Intervals on Original Scale
```

```r
print("Асимптотичний нормальний довірчий інтервал для вибіркової дисперсії:")
```

```
## [1] "Асимптотичний нормальний довірчий інтервал для вибіркової дисперсії:"
```

```r
sd_var <- sqrt((mean((x - mean(x))^4) - var(x)^2) / n)
c(var(x) + qnorm(0.025)*sd_var, var(x) + qnorm(0.975)*sd_var)
```

```
## [1] 0.03161860 0.05775564
```

Ілюстрація для прикладу з середніми та вибірковими дисперсіями (4)

- Можемо порівняти (асимптотичне) покриття відповідних довірчих інтервалів
- Для цього здійснімо симуляцію за методом Монте-Карло для $T = 100$
- Генеруємо всі вибірки (код не показано, див. вихідний .rmd-файл)
- Та обчислюємо покриття інтервалу кожного типу

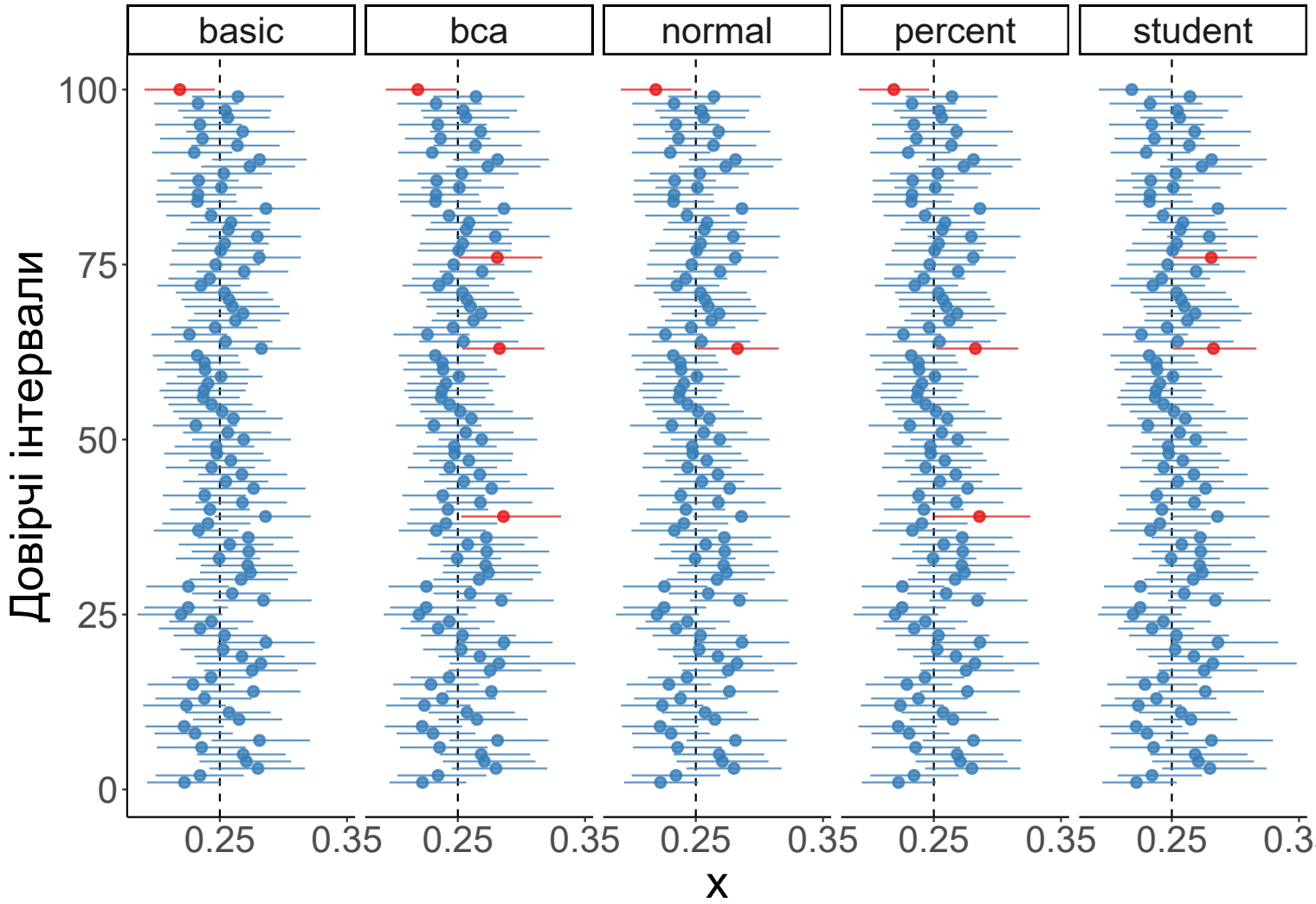
```
df_mean_var %>% group_by(type) %>%
  summarise(coverage_mean = mean(covered_mean),
            coverage_var = mean(covered_var))
```

```
## # A tibble: 5 x 3
##   type    coverage_mean coverage_var
##   <chr>           <dbl>        <dbl>
## 1 basic            0.99         0.92
## 2 bca              0.96         0.94
## 3 normal           0.98         0.93
## 4 percent          0.97         0.93
## 5 student          0.98         0.96
```

- Як можна бачити, найліпшими є студентизовані та BC$_a$ інтервали
- До того ж для вибіркових дисперсій всі інтервали мають гірше покриття, ніж для середніх
  - Це пояснюється недостатньою вибіркою n та недостатнім числом повторень бутстрепу B
  - Але враховуючи, що $n = 200$, що в принципі є дуже малим значенням, результати дуже добрі

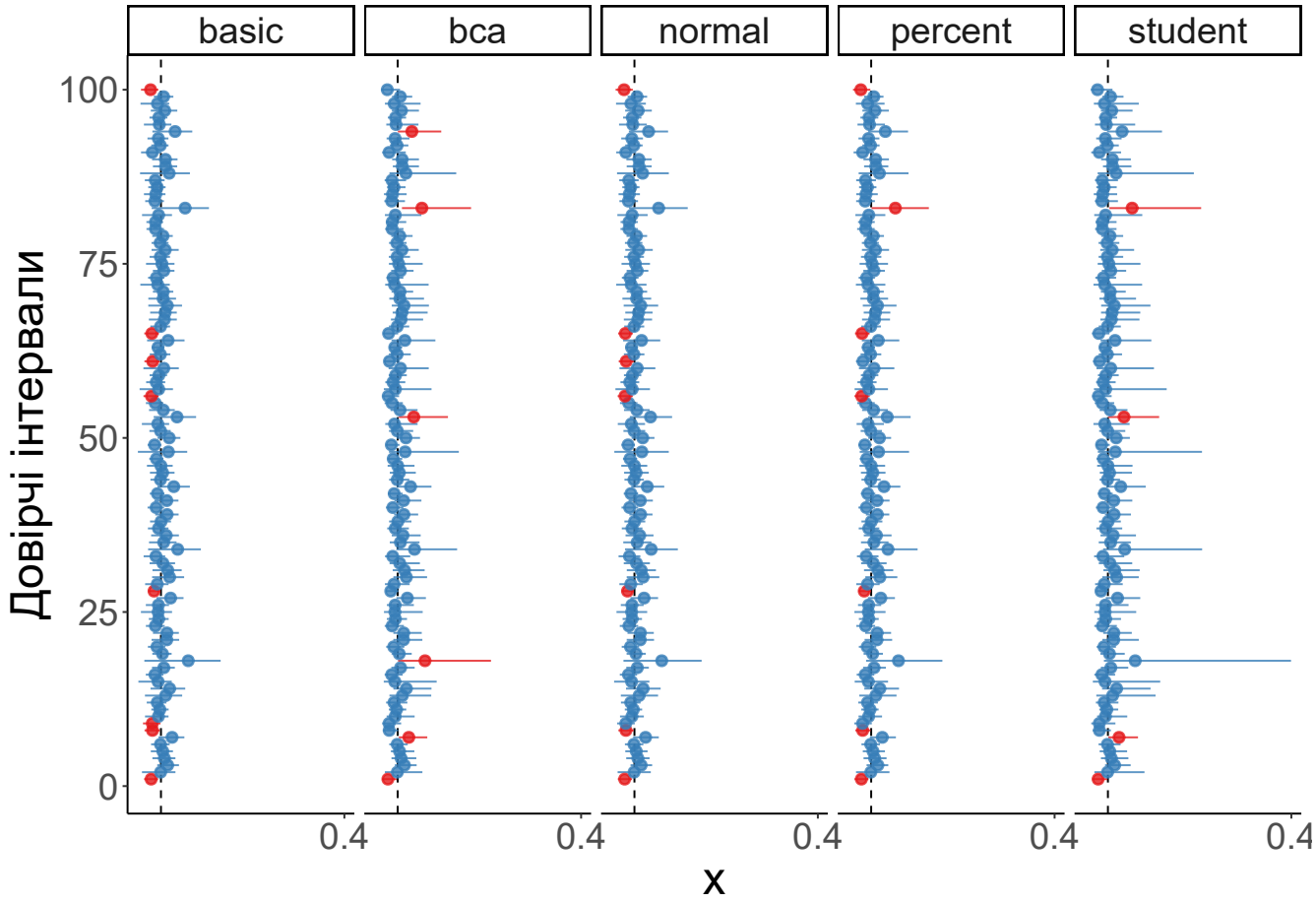
Ілюстрація для прикладу з середніми та вибірковими дисперсіями (4)

- Графічна ілюстрація для середніх



Ілюстрація для прикладу з середніми та вибірковими дисперсіями (5)

- Графічна ілюстрація для вибіркових дисперсій



- Окремі інтервали виявилися доволі асиметричними
  - Це пояснюється тим, що для цих конкретних вибірок розподіл вибіркових дисперсії був далекий від нормального (скошений управо)
  - Особливо це кидається у вічі для студентизованих інтервалів, бо там відбувається ділення на (доволі) малу стандартну похибку

Ілюстрація для прикладу з кореляціями (1)

- Аналогічні ілюстрації можна зробити для інших статистик, що ми розглядали вище
- Перепишімо нашу функцію для підрахунку кореляцій, щоб вона повертала оцінку дисперсії

```
boot_cor_with_sd <- function(X, indices, estimate_var = TRUE){
  cor_bar <- cor(X[indices, ])[1, 2]

  if (estimate_var){
    boot_out <- boot(X[indices, ], statistic = boot_cor_with_sd, R = 200, estimate_var = FALSE)

    return(c(cor_bar, var(boot_out$t[, 1])))
  }
  else {
    return(cor_bar)
  }
}
```

Ілюстрація для прикладу з кореляціями (2)

- Тепер можемо порахувати всі інтервали

```
n <- 100
B <- 2000
cor_true <- 0.9
Sigma <- matrix(c(1, cor_true,
                  cor_true, 1),
                nrow = 2, ncol = 2, byrow = TRUE)
set.seed(100)

X <- MASS::mvrnorm(n, mu = c(0, 0), Sigma = Sigma)

boot_result_cor <- boot(X, statistic = boot_cor_with_sd, R = B)

print("Довірчі бутстреп-інтервали для вибіркової кореляції:")
```

```
## [1] "Довірчі бутстреп-інтервали для вибіркової кореляції:"
```

```
boot.ci(boot_result_cor, index = c(1, 2))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_result_cor, index = c(1, 2))
##
## Intervals :
## Level      Normal              Basic              Studentized
## 95%   ( 0.9163,  0.9612 )   ( 0.9191,  0.9648 )   ( 0.9110,  0.9582 )
##
## Level     Percentile            BCa
## 95%   ( 0.9132,  0.9588 )   ( 0.9095,  0.9568 )
## Calculations and Intervals on Original Scale
```

Ілюстрація для прикладу з кореляціями (3)

- Порівняймо (асимптотичне) покриття відповідних довірчих інтервалів

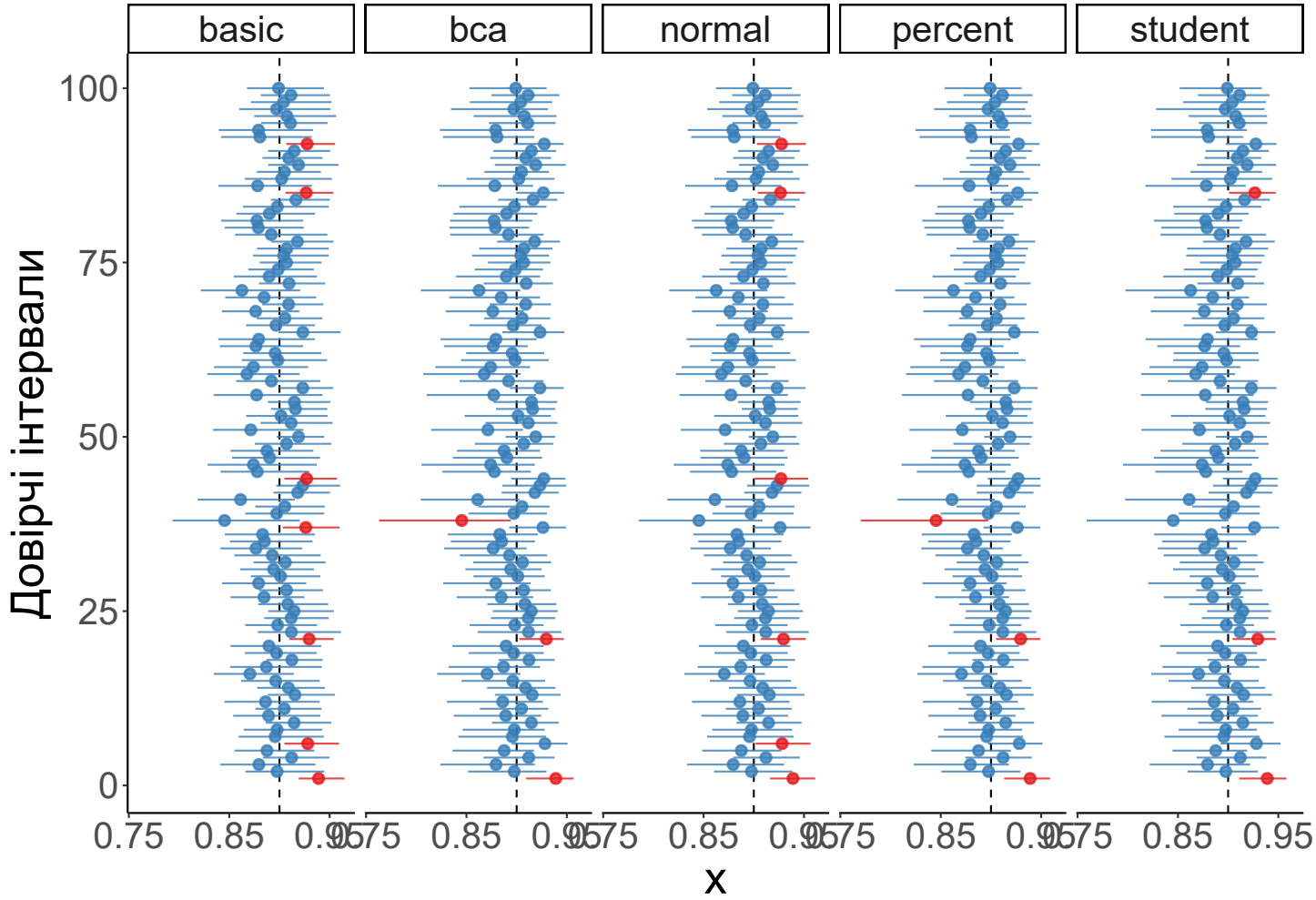
```
df_cor %>% group_by(type) %>%
  summarise(coverage = mean(covered))
```

```
## # A tibble: 5 x 2
##   type    coverage
##   <chr>      <dbl>
## 1 basic       0.93
## 2 bca         0.97
## 3 normal      0.94
## 4 percent     0.97
## 5 student     0.97
```

- Як можна бачити, найліпшими є студентизовані, персентильні та BC_a інтервали

Ілюстрація для прикладу з кореляціями (4)

- Графічна ілюстрація для кореляцій



Ілюстрація для прикладу з різницями медіан (1)

- Перепишімо нашу функцію для підрахунку різниць медіан, щоб вона повертала оцінку дисперсії

```
boot_median_diff_with_sd <- function(x, indices, n_sample1,
                                     estimate_var = TRUE, R_for_sd = 200){
  indices_sample1 <- indices[1:n_sample1]
  indices_sample2 <- indices[-(1:n_sample1)]

  m1 <- median(x[indices_sample1, 1])
  m2 <- median(x[indices_sample2, 1])

  median_diff_bar <- m1 - m2

  if (estimate_var){
    boot_out <- boot(x[indices, ], statistic = boot_median_diff_with_sd,
                     R = R_for_sd, strata = x[, 2],
                     n_sample1 = n_sample1, estimate_var = FALSE)

    return(c(median_diff_bar, var(boot_out$t[, 1])))
  }
  else {
    return(median_diff_bar)
  }
}
```

Ілюстрація для прикладу з різницями медіан (2)

- Тепер можемо порахувати всі інтервали

```
n <- 200
B <- 2000
shape <- 2
rate <- 2
a <- 2
b <- 1
median_diff_true <- qgamma(0.5, shape = 2, rate = 2) - qbeta(0.5, 2, 1)
set.seed(100)

x <- rgamma(n, shape = shape, rate = rate)
y <- rbeta(n, a, b)
dat <- cbind(c(x, y), rep(0:1, each = n))

boot_result_median_diff <- boot(dat, statistic = boot_median_diff_with_sd,
                                R = B, strata = dat[, 2], n_sample1 = n)

print("Довірчі бутстреп-інтервали для різниці вибіркових медіан:")
```

```
## [1] "Довірчі бутстреп-інтервали для різниці вибіркових медіан:"
```

```
boot.ci(boot_result_median_diff, index = c(1, 2))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_result_median_diff, index = c(1, 2))
##
## Intervals :
## Level      Normal              Basic              Studentized
## 95%   (-0.0019,  0.2513 )   (-0.0006,  0.2533 )   (-0.0112,  0.2527 )
##
## Level     Percentile            BCa
## 95%   (-0.0102,  0.2436 )   (-0.0094,  0.2454 )
## Calculations and Intervals on Original Scale
```

Ілюстрація для прикладу з різницями медіан (3)

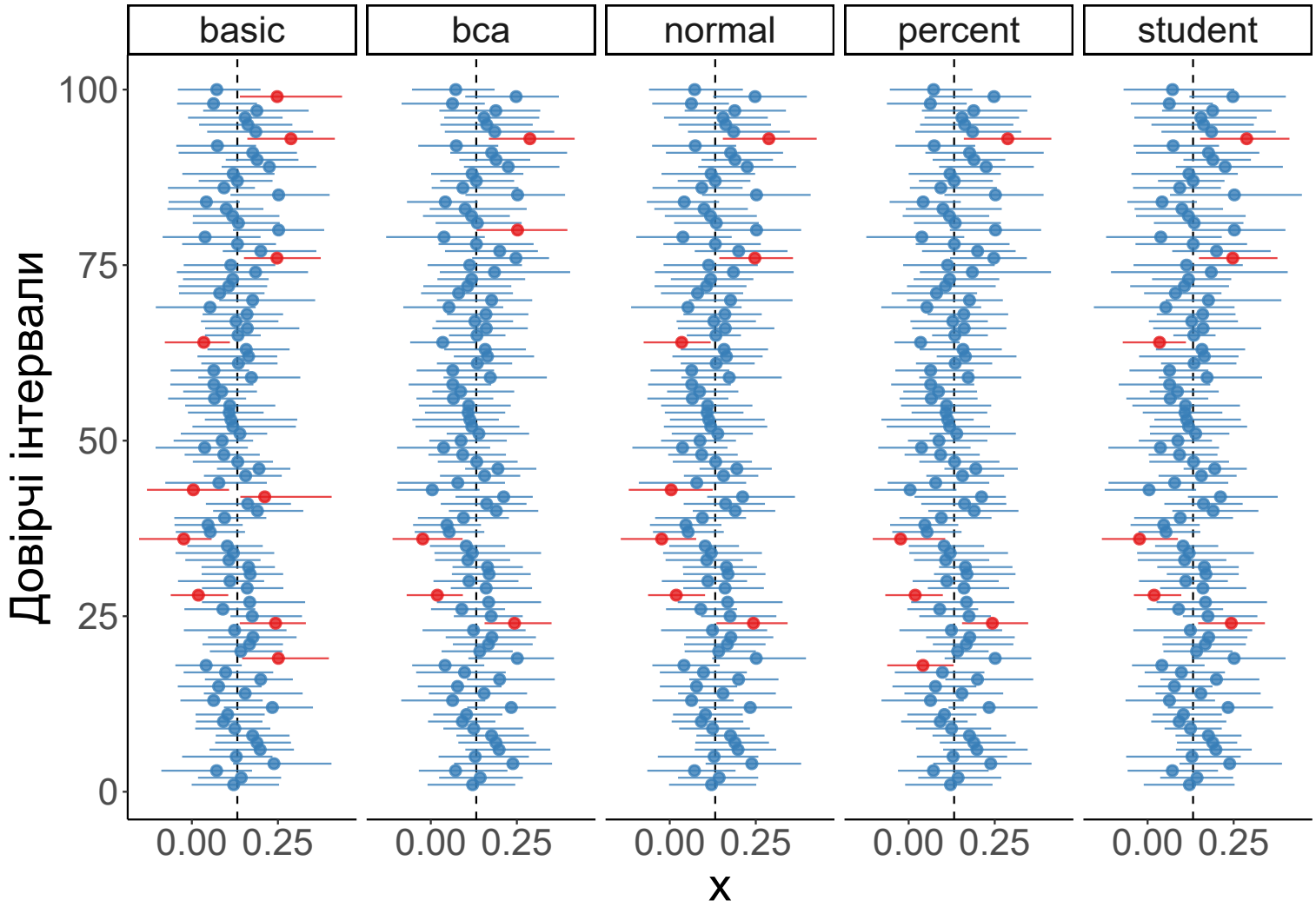
- Порівняймо (асимптотичне) покриття відповідних довірчих інтервалів

```
df_median_diff %>% group_by(type) %>%
  summarise(coverage = mean(covered))
```

```
## # A tibble: 5 x 2
##   type    coverage
##   <chr>      <dbl>
## 1 basic       0.9
## 2 bca         0.95
## 3 normal      0.93
## 4 percent     0.95
## 5 student     0.94
```

- Як можна бачити, найліпшими є студентизовані, персентильні та BC_a інтервали

Ілюстрація для прикладу з різницями медіан (4)



Бутстреп для даних про пасажирів «Титаніку» (1)

- На попередній лекції нас цікавили різниці у віці між уцілілими пасажирами та загиблими
  - Ми розглядали дві популяції $X_1, \dots, X_n \sim X$ (вік уцілілих) та $Y_1, \dots, Y_m \sim Y$ (вік загиблих)
- Ми встановили, що в даних немає достатньо підстав, щоб відкинути $H_0 : \mu_X - \mu_Y \leq 0$ vs. $H_1 : \mu_X - \mu_Y > 0$
- Також ми казали, що якби гіпотеза була «дорівнює 0» vs. «не дорівнює 0», то ми її відкинули б
- Ми можемо перевірити це, побудувавши довірчі інтервали за допомогою бутстрепу

Бутстреп для даних про пасажирів «Титаніку» (2)

- Спочатку напишімо відповідну функцію

```
boot_mean_diff_with_sd <- function(x, indices, n_sample1,
                                   estimate_var = TRUE, R_for_sd = 200){
  indices_sample1 <- indices[1:n_sample1]
  indices_sample2 <- indices[-(1:n_sample1)]

  m1 <- mean(x[indices_sample1, 1])
  m2 <- mean(x[indices_sample2, 1])

  mean_diff_bar <- m1 - m2

  if (estimate_var){
    boot_out <- boot(x[indices, ], statistic = boot_mean_diff_with_sd,
                     R = R_for_sd, strata = x[, 2],
                     n_sample1 = n_sample1, estimate_var = FALSE)

    return(c(mean_diff_bar, var(boot_out$t[, 1])))
  }
  else {
    return(mean_diff_bar)
  }
}
```

Бутстреп для даних про пасажирів «Титаніку» (3)

- А тепер порахуймо самі інтервали

```
set.seed(100)
B <- 2000

dat <- as.matrix(passengers %>% filter(!is.na(Age)) %>% select(Age, Survived) %>%
                   arrange(Survived))

boot_result_mean_diff <- boot(dat, statistic = boot_mean_diff_with_sd,
                              R = B, strata = dat[, 2], n_sample1 = sum(dat[, 2] == 0))

print("Довірчі бутстреп-інтервали для різниці вибіркових середніх:")
```

```
## [1] "Довірчі бутстреп-інтервали для різниці вибіркових середніх:"
```

```
boot.ci(boot_result_mean_diff, index = c(1, 2))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_result_mean_diff, index = c(1, 2))
##
## Intervals :
## Level      Normal              Basic             Studentized
## 95%   ( 0.109,  4.382 )   ( 0.091,  4.414 )   ( 0.193,  4.266 )
##
## Level     Percentile            BCa
## 95%   ( 0.151,  4.474 )   ( 0.089,  4.310 )
## Calculations and Intervals on Original Scale
```

Бутстреп для даних про пасажирів «Титаніку» (4)

- Можемо це порівняти з результатом застосування функції `t.test`

```
t.test(Age ~ Survived, data = passengers)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Survived
## t = 2.046, df = 598.84, p-value = 0.04119
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09158472 4.47339446
## sample estimates:
## mean in group 0 mean in group 1
##        30.62618        28.34369
```

- Бачимо, що в цьому прикладі метод BC_a дав найліпший результат
  - Як і звичайний базовий, як і асимптотичний нормальний
  - Бо статистика доволі «проста»
- Але якщо згадати, що мова про вік, який переважно вимірюється цілими числами, то стає зрозуміло, що ці різниці не дуже суттєві
- Принаймні нуль точно не входить у жодний інтервал
- Отже різниця у віці статистично значуща