Лекція 5. Статистичне виведення в R. Довірчі інтервали та тестування гіпотез

Данило Тавров

08.03.2023

Вступні зауваги

- Ми продовжуємо розгляд основних понять із теорії ймовірностей і статистики та їхніх реалізацій в R
- Сьогодні згадаємо другу половину основних понять зі статистики
  - Ми пригадаємо, що таке довірчий інтервал і як його будувати
  - Ми пригадаємо, як тестувати гіпотези та інтерпретувати відповідні результати
- Корисними матеріалами є:
  - Фундаментальна книжка *All of Statistics* (Larry Wasserman), розділи 6, 10 (викладено на диску в загальному каталозі з літературою)
  - Книжка *Introduction to Econometrics* (Bruce Hansen) (розділи 13–14) (викладено на диску в загальному каталозі з літературою)
  - Книжка *Using R for Introductory Statistics* (John Verzani), розділи 8–10 (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

Що було на попередній лекції (1)

- Минулого разу ми з Вами з'ясували, що для оцінювання деякого (невідомого) параметра θ DGP, який породив наші дані, використовують оцінку $\widehat{\theta}$
- Ми з'ясували, що будь-яка адекватна оцінка повинна бути спроможною: $\widehat{\theta} \xrightarrow{p} \theta$
  - Тоді підрахована для конкретної вибірки $X_1, X_2, \ldots, X_n$ *реалізація* цієї оцінки буде близька до θ
- Також ми говорили, що обов'язково потрібно наводити «ступінь цієї близькости»
  - Просто так порахувати якусь статистику і сказати «ось вона» — непрофесіонально і антинауково!
- Тому нас цікавлять розподіли оцінок
  - А оскільки в цьому курсі ми працюємо з великими наборами даних, то нас цікавлять **асимптотичні** розподіли оцінок
- Нам дуже подобається, коли асимптотичний розподіл є нормальний
  - Бо з нормальним розподілом легко працювати
  - Багато оцінок мають асимптотично нормальний розподіл (напр., усі оцінки за методом максимальної правдоподібности та багато інших)

Що було на попередній лекції (2)

- Якщо ми знаємо (асимптотичну) дисперсію оцінки, $\operatorname{Var}\left(\hat{\theta}\right)$, ми можемо обчислити її **стандартну похибку** як $\operatorname{se}\left(\hat{\theta}\right) = \sqrt{\operatorname{Var}\left(\hat{\theta}\right)}$
- Якщо ми **знаємо** дисперсію теоретично, але **не можемо її підрахувати** (бо не знаємо якихось параметрів), то тоді можна **оцінити** стандартну похибку: $\widehat{\operatorname{se}}\left(\hat{\theta}\right) = \sqrt{\widehat{\operatorname{Var}}\left(\hat{\theta}\right)}$
    - Оцінка дисперсії $\widehat{\operatorname{Var}}\left(\hat{\theta}\right)$ повинна бути спроможною
    - Найпростіше цього досягти, замінивши всі параметри на їхні вибіркові аналоги
    - Тобто сподівання — на вибіркові середні тощо
- Нарешті, може бути так, що ми **не знаємо** дисперсії, або навіть усього розподілу
    - У цьому випадку потрібно застосовувати бутстреп
    - Це є предмет наступної лекції

План лекції

Потреба в довірчих інтервалах

- Отже ми вже добре усвідомлюємо, що будь-яку оцінку $\widehat{\theta}$ потрібно супроводжувати її стандартною похибкою $\mathrm{se}\left(\widehat{\theta}\right)$ або оцінкою такої $\widehat{\mathrm{se}}\left(\widehat{\theta}\right)$
- Якщо теоретично відомо, що оцінка $\widehat{\theta}$ має асимптотично нормальний розподіл, то стандартна похибка дає змогу оцінити справжній розкид значень за правилом трьох сигм
- Якщо ж розподіл не є нормальний, то стандартна похибка все одно корисна, але вже не має такої якісної інтерпретації
- Тому в загальному випадку більшу користь має так званий **довірчий інтервал** (confidence interval)
  - Або, якщо параметр багатовимірний, **довірча множина** (confidence set)

Визначення довірчого інтервалу

Визначення 1.1

- Нехай маємо деяку статистичну модель із параметром $\theta \in \Theta \subseteq \mathbb{R}^k$
- Тоді **довірчою множиною** C_n **рівня** $1-\alpha$ ($1-\alpha$ level confidence set) буде множина така, для якої виконується

$$\mathbb{P}\left(C_n \ni \theta\right) \geq 1-\alpha\ , \quad \theta \in \Theta \tag{1.1}$$

- Якщо $\theta \in \mathbb{R}$, то маємо довірчий інтервал $C_n = [a; b]$

□

- Варто звернути увагу, що границі інтервалу є статистиками: $a = a(X_1, \ldots, X_n)$, $b = b(X_1, \ldots, X_n)$
- Кажемо, що довірчий інтервал **покриває** θ, якщо $C_n \ni \theta$
- А значення $1-\alpha$ називаємо **покриттям** (coverage)
- На практиці **дуже** часто використовують $\alpha = 0.05$
  - Як писав Роналд Фішер[1], "It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance"
  - Із того часу всі так і роблять, хоча обґрунтування немає!

[1] Sir Ronald Aylmer Fisher (1890–1962) — британський статистик

Інтерпретація довірчого інтервалу

- Варто відразу розставити всі акценти щодо коректної інтерпретації довірчого інтервалу
- Довірчий інтервал C_n є **випадковою величиною**!
- Розгляньмо **типові помилки** інтерпретації
- Інколи можна почути, що якщо ми порахували, наприклад, 95% довірчий інтервал $C_n = [1.4; 1.6]$, то $\theta \in [1.4; 1.6]$ з імовірністю 95%
  - Це **нісенітниця**, адже θ є фіксованим значенням, а не випадковою величиною
- Тоді можна сказати: добре, $[1.4; 1.6] \ni \theta$ з імовірністю 95%
  - Але це ще більша **нісенітниця**!
  - Адже $[1.4; 1.6]$ також є фіксованим інтервалом
  - Тому θ або належить йому, або ні
- Правильна інтерпретація така:
  - Довірчий інтервал C_n із покриттям $1 - \alpha$ **як випадкова величина** покриває θ з імовірністю $1 - \alpha$
  - Тобто якби ми утворювали нові вибірки, і для кожної з них рахували свій C_n, частка $1 - \alpha$ з них містили б θ
  - Тобто, якщо $C_n = [1.4; 1.6]$, то в принципі θ **може взагалі там не лежати**, бо саме ця вибірка одна з тих, що C_n не покриває θ
  - Але ми усвідомлює цей ризик і вважаємо, що покриває

Підрахунок довірчого інтервалу (1)

- Для того, щоб знайти довірчий інтервал, потрібно знати, як рахувати ймовірність $\mathbb{P}\left(C_n \ni \theta\right)$
- Найпростіша і часто використовувана на практиці ситуація виринає, коли оцінка $\widehat{\theta}$ має асимптотичний нормальний розподіл $\widehat{\theta} \overset{\text{a}}{\sim} N\left(\theta, \text{Var}\left(\widehat{\theta}\right)\right)$
- Тоді

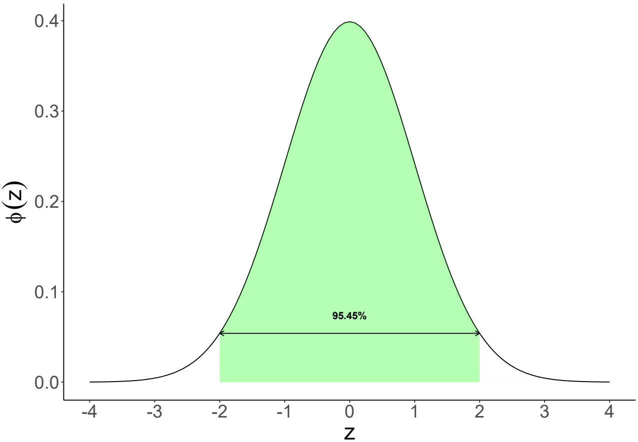
$$C_n = \left[\widehat{\theta} - z_{1-\alpha/2} \cdot \text{se}\left(\widehat{\theta}\right); \widehat{\theta} + z_{1-\alpha/2} \cdot \text{se}\left(\widehat{\theta}\right)\right] \tag{1.2}$$

  - Тут $z_a = \Phi^{-1}(a)$ — квантиль стандартного нормального розподілу
- Це справді так, адже

$$\begin{aligned}
\mathbb{P}\left(C_n \ni \theta\right) &= \mathbb{P}\left(\widehat{\theta} - z_{1-\alpha/2} \cdot \text{se}\left(\widehat{\theta}\right) \leq \theta \leq \widehat{\theta} + z_{1-\alpha/2} \cdot \text{se}\left(\widehat{\theta}\right)\right) \\
&= \mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\text{se}\left(\widehat{\theta}\right)} \leq z_{1-\alpha/2}\right) \\
&= F\left(z_{1-\alpha/2}\right) - F\left(-z_{1-\alpha/2}\right) \\
&\overset{d}{\to} \Phi\left(z_{1-\alpha/2}\right) - \Phi\left(-z_{1-\alpha/2}\right) \\
&= 1 - \frac{\alpha}{2} - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha
\end{aligned}$$

- Тут F — функція розподілу відповідного дробу

Підрахунок довірчого інтервалу (2)



- Якщо нас цікавить 95% інтервал, то потрібно взяти $z_{0.025} = -z_{0.975} \approx -1.96$ і $z_{0.975} \approx 1.96$
- На практиці часто для швидкої оцінки інтервалу можна використовувати значення 2 і (подумки) рахувати інтервал $\hat{\theta} \pm 2 \cdot \text{se}\left(\hat{\theta}\right)$
  - У статтях та звітах потрібно подавати коректні інтервали без цих округлень

Ілюстративний приклад (1)

- Обчислімо 95% довірчий інтервал для середньої вартости квитка на «Титанік» для жінок, X
- Пригадаймо, що середнє вибіркове $\overline{X}$ має асимптотичний розподіл $\overline{X} \overset{a}{\sim} N\left(\mathbb{E}\left[X\right], \frac{\mathrm{Var}(X)}{n}\right)$
    - Ми не знаємо розподіл X, а тому повинні оцінити $\mathrm{Var}\left(X\right)$
- Отже оцінка стандартної похибки дорівнює $\widehat{\mathrm{se}}\left(\overline{X}\right) = \sqrt{s_X^2}$
- Тоді в R відповідний довірчий інтервал можна побудувати так:

```
ci <- passengers %>% filter(Sex == "female") %>%
  summarise(mean = mean(Fare),
            sd = sd(Fare),
            n = n(),
            a = mean(Fare) + qnorm(0.025) * sd(Fare) / sqrt(n()),
            b = mean(Fare) + qnorm(0.975) * sd(Fare) / sqrt(n()))
ci
```

```
## # A tibble: 1 x 5
##    mean    sd     n     a     b
##   <dbl> <dbl> <int> <dbl> <dbl>
## 1  44.5  58.0   314  38.1  50.9
```

- Як можна бачити, середня вартість квитка для жінок склала $\overline{X} = 44.48$, а вибіркове середньоквадратичне відхилення дорівнює $s_X = 57.998$

Ілюстративний приклад (2)

- Відтак згідно з (1.2) 95% довірчий інтервал дорівнює

$$C_n = \left[\overline{X} - 1.96\frac{s_X}{\sqrt{n}}; \overline{X} + 1.96\frac{s_X}{\sqrt{n}}\right] = [38.065; 50.895]$$

- Отже справжнє значення середньої вартости квитка може бути будь-яке з цього інтервалу
  - Ми **не можемо казати**, що справжнє середнє лежить у цьому інтервалі з імовірністю 95%!!!
  - Ми тільки можемо казати, що будь-яке значення з цього інтервалу нічим не гірше від 44.48

Підрахунок довірчого інтервалу з квантилями t-розподілу (1)

- Нехай $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$
- Тоді $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- Тоді Z-оцінка $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ **точно**
- У класичній статистиці доводять, що $t = \frac{\overline{X}-\mu}{s_X/\sqrt{n}} \sim t_{n-1}$ **точно**
    - Статистику t з очевидних міркувань називають t**-статистикою** (t statistic)
- Іншими словами, якщо ми маємо нормальну вибірку, але замість (невідомої нам) дисперсії σ^2 використовуємо її оцінку s_X^2, то відповідне відношення має t-розподіл
- Тоді можна будувати довірчий інтервал

$$C_n = \left[\hat{\theta} - q_{1-\alpha/2}\text{se}\left(\hat{\theta}\right); \hat{\theta} + q_{1-\alpha/2}\text{se}\left(\hat{\theta}\right)\right]$$

- Тут q_a — a-ий квантиль t-розподілу з $n-1$ ступенями свободи
- Такий інтервал для малих вибірок буде коректніший від нормального, який ми розглядали вище

Підрахунок довірчого інтервалу з квантилями t-розподілу (2)

- На практиці ми (а) не знаємо, чи є розподіл нормальний, (б) маємо всі підстави підозрювати, що розподіл **не є** нормальний
- Понад те, якщо розмір вибірки «достатньо» великий, і асимптотичний розподіл нормальний, то можна застосовувати нормальні довірчі інтервали
- Проте прийнято вважати, що якщо ми використовуємо замість $\operatorname{Var}\left(\hat{\theta}\right)$ її оцінку $\widehat{\operatorname{Var}}\left(\hat{\theta}\right)$, усе одно ліпше використовувати t-розподіл замість нормального
- Зокрема, в нашому прикладі маємо

```
ci <- ci %>% mutate(a_t = mean + qt(0.025, df = n - 1) * sd / sqrt(n),
                    b_t = mean + qt(0.975, df = n - 1) * sd / sqrt(n))
ci
```

```
## # A tibble: 1 x 7
##    mean    sd     n     a     b   a_t   b_t
##   <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl>
## 1  44.5  58.0   314  38.1  50.9  38.0  50.9
```

- Як можна бачити, інтервали доволі близькі між собою
- А якщо згадати, що одиниця виміру — фунти, то відмінність стає просто нікчемною

Підрахунок довірчого інтервалу з квантилями t-розподілу (3)

- В R такі інтервали можна генерувати за допомогою функції `confint`
- Спочатку потрібно оцінити *лінійну модель* за допомогою функції `lm`
  - Під лінійною моделлю мається на увазі, що ми одну величину виражаємо через лінійну комбінацію інших: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$
  - Тут ε — деяка похибка
  - Тоді коефіцієнти $\beta_0, \ldots, \beta_n$ моделі шукають методом найменших квадратів
  - Тобто мінімізують $(Y - \beta_0 - \beta_1 X_1 - \ldots - \beta_n X_n)^2$
  - Ми цим дуже багато займатимемося в рамках регресійного аналізу
- Як ми казали в попередній лекції, вибіркове середнє саме по собі є оцінкою найменших квадратів
- Тому суто формально можна розглянути лінійну модель $X = \beta_0 + \varepsilon$ і знайти $\beta_0 = \overline{X}$ як мінімізатор виразу $(X - \beta_0)^2$
- В R це можна порахувати в такий спосіб:

```
model <- lm(Fare ~ 1, data = passengers %>% filter(Sex == "female"))
confint(model, level = 0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 38.03996 50.91968
```

- Зліва від ~ стоїть залежна змінна
  - Справа від ~ стоїть 1, що дає R зрозуміти, що жодних інших незалежних змінних немає, а тільки довільна константа
- Цей результат повністю відповідає підрахованому вручну для випадку з t-розподілом

Симуляція за методом Монте-Карло (1)

- Можна проілюструвати відповідність асимптотичного довірчого інтервалу нашим очікуванням
- Для цього виконаймо симуляцію за методом Монте-Карло для вибірок трьох різних розмірів $n = 10, 100, 1000$
- Вибірки будемо генерувати з розподілу Exp (2)
- Для кожної вибірки ми рахуватимемо три різні довірчі інтервали для $\overline{X}$:
  - Інтервал на основі Z-оцінки та нормальних квантилів
  - Інтервал на основі t-статистики та нормальних квантилів
  - Інтервал на основі t-статистики та квантилів t-розподілу
- Генеруємо всі вибірки:

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

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

  return(result)
}

T <- 100
lambda <- 2
mu_true <- 1/lambda
sigma_true <- 1/lambda
df <- NULL
for (n in c(10, 100, 1000)){
  df <- rbind(df, as_tibble(t(replicate(T, sample_statistics(n, lambda)))))
}
df <- df %>% mutate(index = rep(1:T, 3))
```

Симуляція за методом Монте-Карло (2)

- Обчислюємо довірчі інтервали

```
df <- df %>% mutate(a_t = mean + qt(0.025, df = n - 1)  * sd / sqrt(n),
                    b_t = mean + qt(0.975, df = n - 1) * sd / sqrt(n),
                    covered_t = a_t < mu_true & mu_true < b_t,
                    a_z = mean + qnorm(0.025)  * sigma_true / sqrt(n),
                    b_z = mean + qnorm(0.975) * sigma_true / sqrt(n),
                    covered_z = a_z < mu_true & mu_true < b_z,
                    a_z_asy = mean + qnorm(0.025)  * sd / sqrt(n),
                    b_z_asy = mean + qnorm(0.975) * sd / sqrt(n),
                    covered_z_asy = a_z_asy < mu_true & mu_true < b_z_asy)
```

- Можемо порахувати емпіричні покриття

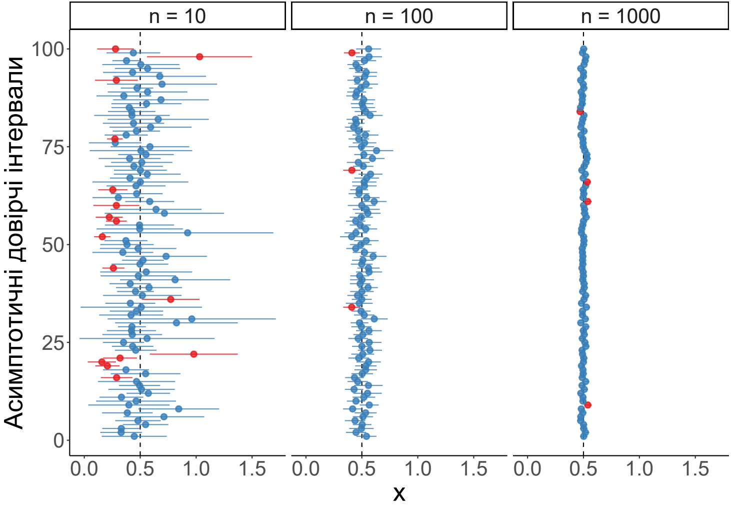
```
df %>% group_by(n) %>%
  summarise(coverage_z = mean(covered_z),
            coverage_t = mean(covered_t),
            coverage_z_asy = mean(covered_z_asy))
```

```
## # A tibble: 3 x 4
##       n coverage_z coverage_t coverage_z_asy
##   <dbl>      <dbl>      <dbl>          <dbl>
## 1    10       0.91       0.88           0.84
## 2   100       0.96       0.97           0.97
## 3  1000       0.95       0.96           0.96
```

- Як можна бачити, найліпшим є варіант, коли ми достеменно знаємо дисперсію
- Але оскільки на практиці це неможливо, ми використовуємо стандартну похибку
- Зі збільшенням n покриття наближається до 95%
- Для малих вибірок варіант із t-розподілом трішки ліпший
- Але на великих вибірках ці відмінності зовсім несуттєві

Симуляція за методом Монте-Карло (3)

- Графічна ілюстрація



Довірчі інтервали в складних випадках

- Для низки оцінок неможливо або дуже важко встановити наявність асимптотичного нормального розподілу
  - Наприклад, для коефіцієнтів кореляцій
  - Чи дуже складних функцій від параметрів, що унеможливлює застосування дельта-методу
- А навіть якщо цей розподіл і можна встановити, оцінити стандартну похибку неможливо або дуже важко
  - Наприклад, стандартна похибка вибіркового квантиля залежить від (невідомої) щільности розподілу
- У таких випадках можна використовувати бутстреп
  - Бутстреп можна використовувати для оцінки стандартної похибки, якщо розподіл асимптотично нормальний
  - Бутстреп можна використовувати для побудови самого інтервалу безвідносно до форми асимптотичного розподілу
- Ми про це говоритимемо в наступній лекції

План лекції

1. Довірчі інтервали

2. Тестування гіпотез

Загальні поняття про гіпотези

- Як ми вже знаємо, значення оцінки $\widehat{\theta}$ деякого параметра θ обов'язково повинно супроводжуватися відповідною стандартною похибкою $\widehat{\mathrm{se}}\left(\widehat{\theta}\right)$ чи довірчим інтервалом
  - Це дає змогу зрозуміти, наскільки близько до справжнього параметра лежить значення нашої оцінки
- Інший спосіб з'ясувати, яке **насправді** значення має параметр θ, передбачає тестування відповідної гіпотези
- **Нульовою гіпотезою** (null hypothesis) H_0 є деяке твердження про параметр θ, яке ми хочемо перевірити
  - Наприклад, $H_0 : \theta < 2$ або $H_0 : \mu_X = \sigma_X$
  - Твердження на кшталт $H_0 : \overline{X} = 2$ є нісенітницею, адже ми можемо гіпотезувати тільки відносно сталих чисел, а не випадкових величин
  - У загальному випадку $H_0 : \theta \in \Theta_0$ для деякої множини $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^k$
- **Альтернативною гіпотезою** (alternative hypothesis) H_1 є доповнення H_0
  - Тобто $H_1 : \theta \in \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$
  - Наприклад, $H_1 : \theta \geq 2$ або $H_1 : \mu_X \neq \sigma_X$ для прикладів вище
- Альтернативні гіпотези можуть бути:
  - **Однобічні**: якщо $H_0 : \theta \geq \theta_0$, то $H_1 : \theta < \theta_0$ або аналогічно з протилежними знаками
  - **Двобічні**: $H_0 : \theta = \theta_0$, то $H_1 : \theta \neq \theta_0$

Загальні поняття про тести

- **Тестом** (test) називають деяке правило, згідно з яким можна з'ясувати, відкидати нульову гіпотезу чи ні
- Як правило, будують **тестову статистику** (test statistic) $T = T(X_1, \ldots, X_n)$
- Тоді тест можна сформулювати так: відкинути H_0, тільки якщо $T \in R$
  - Тут R є **критичною областю** (critical region) або **областю відкидання** (rejection region)
- Наприклад, може стояти питання про вплив участи в деякій навчальній програмі на рівень доходів її випускників
  - Ми можемо розглянути дві популяції: учасників програми $X_1, \ldots, X_n$ та тих, що не брали участи, $Y_1, \ldots, Y_m$
  - Нехай θ дорівнює різниці в середніх доходах між популяціями
  - Тоді нас цікавить $H_0 : \theta \leq c$ vs. $H_1 : \theta > c$, де c — певна грошова сума
  - Тоді тест *можна* сформулювати так: якщо $T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \overline{X} - \overline{Y} > c$, відкинути H_0, інакше залишити
  - Тобто $R = \{x : x > c\}$
  - Звісно, для того, щоб такий простий тест мав сенс, потрібно, щоб учасники програми походили з тієї ж популяції, що й не учасники, а відповідні вибірки були достатньо великі
  - Інакше може так статися, що в навчальній програмі беруть участь найбільш здібні учасники, і тоді весь сенс втрачається

Різні типи помилок

- Здійснюючи тестування, ми не можемо не помилятися
- **Помилка I роду** (type I error) має місце, коли ми (помилково) відкидаємо істинну H_0
- **Помилка II роду** (type II error) має місце, коли ми (помилково) **не** відкидаємо хибну H_0

	H_0 істинна	H_1 істинна
H_0 не відкинуто	Правильне рішення	Помилка II роду
H_0 відкинуто	Помилка I роду	Правильне рішення

- Помилки двох родів **не є** симетричними
  - Так, помилка I роду для нашого прикладу означає, що ми ухвалили рішення про дієвість програми, яка насправді безкорисна
  - На її впровадження може бути витрачено ресурси, які можна було б використати деінде
  - З іншого боку, помилка II роду може призвести до (помилкового) рішення відмовитися від справді корисної програми
  - Тоді буде втрачено потенційні можливості, але принаймні жодних наявних ресурсів витрачено не буде
- Що гірше — стратити невинуватого чи відпустити на волю злочинця?..

Потужність тесту

- В ідеальному випадку тест ніколи не помилятиметься
- Але на практиці це фактично нереально, тому ми намагаємося мінімізувати ймовірність помилки
- Імовірність відкинути H_0 називають **функцією потужности** (power function) тесту:

$$\beta(\theta) = \mathbb{P}_\theta\left(T \in R\right) \tag{2.1}$$

- Тут індекс θ означає, що ми застосовуємо розподіл випадкової величини T за умови, що справжнє значення невідомого параметра дорівнює θ
- **Рівнем значущости** (significance level) тесту називають величину

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

- Ми кажемо, що тест має деякий рівень, наприклад, 0.05, якщо $\alpha \leq 0.05$
- Рівень значущости можна інтерпретувати як (найгіршу) імовірність допустити помилку I роду
- На практиці задають потрібний рівень значущости α і підбирають тест, який матиме найвищу потужність для $\theta \in \Theta_1$
- Ми в ці питання заходити не будемо, а просто розглянемо деякі основні використовувані на практиці тести

Інтерпретація рішення про неможливість відкидання H_0 (1)

- Дуже важливо розуміти, що ми ніколи не можемо «підтвердити» H_0
  - Максимум, що ми можемо — виявити відсутність підстав для її спростування
  - Тому якщо $T \notin R$, ми кажемо, що ми не можемо відкинути H_0, **але це не означає, що вона істинна**!
- Це не просто семантичні міркування
- Справді, ми фіксуємо рівень значущости α, як правило, на доволі низькому рівні, для мінімізації помилки I роду: $\beta(\theta_0) \leq \alpha$
- Але це означає, що для значень θ **близько до** справжнього θ_0 потужність тесту також може бути дуже низькою: $\beta(\theta) \approx \alpha$
  - Або навіть і не близько, але все одно не дуже високою (наприклад, 40% або 50%)
- Тобто ми можемо не відкинути H_0, навіть якщо вона хибна, з доволі високою ймовірністю
- У той же час, якщо ми таки відкинули H_0, це має сенс, оскільки ймовірність помилки гарантовано низька

Інтерпретація рішення про неможливість відкидання H_0 (2)

- Для нашого прикладу нездатність відкинути H_0 може свідчити про неефективність програми
  - …або про те, що вплив програми низький у порівнянні з розмахом доходів у популяції
  - …або про те, що розмір вибірки недостатній
  - Тобто неможливо стверджувати, що програма неефективна
  - У нас просто немає підстав казати, що вона **є** ефективною
- Саме тому ми, як правило, формулюємо H_0 у термінах «небажаного» результату, який ми хочемо спростувати
  - Наприклад, що програма неефективна
  - Якщо ми відкинемо H_0, ми будемо знати, що різниця в середніх доходах додатна і програма дієва
  - Якби ми сформулювали навпаки, і **не відкинули** $H_0 : \theta > c$, то це **не означало б**, що програма дієва

Тест Волда (1)

- Розгляньмо надзвичайно поширений на практиці **тест Волда** (Wald test)[2]
- Нехай $\theta \in \Theta \subseteq \mathbb{R}$, а $\hat{\theta}$ — його оцінка з оціненою стандартною похибкою $\widehat{\text{se}}\left(\hat{\theta}\right)$
- Нехай маємо гіпотезу $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
- Нехай
$$T = \frac{\hat{\theta} - \theta_0}{\widehat{\text{se}}\left(\hat{\theta}\right)} \xrightarrow{d} N(0, 1)$$
  - Це нам так хочеться, але це справедливо, тільки якщо $\hat{\theta}$ має асимптотично нормальний розподіл
  - Це потрібно доводити для кожної окремої оцінки $\hat{\theta}$
- Тоді тест Волда рівня α такий: відкинути H_0 тоді й тільки тоді, коли:
  - Для двобічної альтернативи маємо $|T| > z_{1-\alpha/2}$
  - Для однобічної альтернативи виду $H_1 : \theta > \theta_0$ маємо $T > z_{1-\alpha}$
  - Для однобічної альтернативи виду $H_1 : \theta < \theta_0$ маємо $T < z_\alpha$

[2] Абрагам Волд (Abraham Wald, 1902–1950) — американський статистик угорського походження

Тест Волда (2)

- З'ясуймо, чому саме так
- Якщо H_0 істинна, то $T \xrightarrow{d} N(0,1)$
- Відтак імовірність відкинути H_0, якщо вона істинна, дорівнює

$$\begin{aligned}
\mathbb{P}_{\theta_0}\left(\frac{\left|\hat{\theta}-\theta_0\right|}{\widehat{\mathsf{se}}\left(\hat{\theta}\right)} > z_{1-\alpha/2}\right) &= \mathbb{P}_{\theta_0}\left(\frac{\hat{\theta}-\theta_0}{\widehat{\mathsf{se}}\left(\hat{\theta}\right)} < -z_{1-\alpha/2}\right) + \mathbb{P}_{\theta_0}\left(\frac{\hat{\theta}-\theta_0}{\widehat{\mathsf{se}}\left(\hat{\theta}\right)} > z_{1-\alpha/2}\right) \\
&= F_{\theta_0}\left(-z_{1-\alpha/2}\right) + 1 - F_{\theta_0}\left(1 - z_{\alpha/2}\right) \\
&\xrightarrow{d} \Phi\left(-z_{1-\alpha/2}\right) + 1 - \Phi\left(1 - z_{\alpha/2}\right) \\
&= 2\left(1 - \Phi\left(z_{1-\alpha/2}\right)\right) = \alpha
\end{aligned}$$

- Аналогічно можна показати й для однобічних альтернатив

Тест Волда (3)

- Можна також обчислити функцію потужности тесту
- Нехай справжнім значенням параметра є $\theta' \neq \theta_0$
- Для цього розгляньмо уявну ситуацію, коли ми можемо обчислити $\mathrm{se}\left(\hat{\theta}\right)$, використовуючи θ_0 як справжнє значення параметра

$$
\begin{aligned}
\beta\left(\theta'\right) &= \mathbb{P}_{\theta'}\left(\frac{\left|\hat{\theta}-\theta_0\right|}{\mathrm{se}\left(\hat{\theta}\right)} > z_{1-\alpha/2}\right) \\
&= \mathbb{P}_{\theta'}\left(\frac{\hat{\theta}-\theta_0}{\mathrm{se}\left(\hat{\theta}\right)} < -z_{1-\alpha/2}\right) + \mathbb{P}_{\theta'}\left(\frac{\hat{\theta}-\theta_0}{\mathrm{se}\left(\hat{\theta}\right)} > z_{1-\alpha/2}\right) \\
&= \mathbb{P}_{\theta'}\left(\frac{\hat{\theta}-\theta'}{\mathrm{se}\left(\hat{\theta}\right)} < \frac{\theta'-\theta_0}{\mathrm{se}\left(\hat{\theta}\right)} - z_{1-\alpha/2}\right) + \mathbb{P}_{\theta'}\left(\frac{\hat{\theta}-\theta'}{\mathrm{se}\left(\hat{\theta}\right)} > \frac{\theta'-\theta_0}{\mathrm{se}\left(\hat{\theta}\right)} + z_{1-\alpha/2}\right) \\
&\xrightarrow{d} \Phi\left(\frac{\theta'-\theta_0}{\mathrm{se}\left(\hat{\theta}\right)} - z_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\theta'-\theta_0}{\mathrm{se}\left(\hat{\theta}\right)} + z_{1-\alpha/2}\right)
\end{aligned}
$$

- Цілком очевидно, що якщо $\theta' = \theta_0$, ми просто дістанемо $\beta(\theta_0) = \alpha$, що ми й так знаємо

Тест Волда (4)

- Можемо просимулювати це за методом Монте-Карло за допомогою такого прикладу
- Генеруємо вибірку з $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}\,(2)$
- Нас цікавить оцінка параметра λ
- За методом максимальної правдоподібности, $\hat{\lambda} = \left(\overline{X}\right)^{-1}$
- Також за методом максимальної правдоподібности $\text{Var}\left(\hat{\lambda}\right) = \frac{\lambda^2}{n}$
    - Отже $\widehat{\text{se}}\left(\hat{\lambda}\right) = \frac{\hat{\lambda}}{\sqrt{n}}$
- Ми хочемо протестувати гіпотезу $H_0 : \lambda = \lambda'$ vs. $H_1 : \lambda \neq \lambda'$ для $\lambda' \in [1.5; 2.5]$ (для ілюстрації) на рівні $\alpha = 0.05$
- Відтак тестова статистика дорівнює $T = \frac{\hat{\lambda} - \lambda'}{\widehat{\text{se}}\left(\hat{\lambda}\right)}$

Тест Волда (5)

- Симулюємо тестування для одного λ':

```
power_simulation <- function(ns, lambda, lambda_prime, T){
  mle_exp <- function(n, rate){
    x <- rexp(n, rate = rate)

    lambda_hat <- 1 / mean(x)

    result <- c(lambda_hat, lambda_hat / sqrt(n), n)
    names(result) <- c("lambda_hat", "se_hat", "n")

    return(result)
  }

  df <- NULL
  for (n in ns){
    df <- rbind(df, as_tibble(t(replicate(T, mle_exp(n, lambda)))))
  }

  df <- df %>% mutate(test_stat = (lambda_hat - lambda_prime) / se_hat,
                      rejected = abs(test_stat) > qnorm(0.975)) %>%
    group_by(n) %>% summarise(p_mc = mean(rejected), sd = sqrt(p_mc*(1 - p_mc))) %>%
    mutate(se = sd / sqrt(n),
           lambda_prime = lambda_prime,
           p_theory = pnorm((lambda_prime - lambda) / (lambda / sqrt(n)) - qnorm(0.975)) +
             1 - pnorm((lambda_prime - lambda) / (lambda / sqrt(n)) + qnorm(0.975))) %>%
    pivot_longer(cols = starts_with("p"),
                 names_to = "type", names_prefix = "p_",
                 values_to = "prob")

  return(df)
}
```

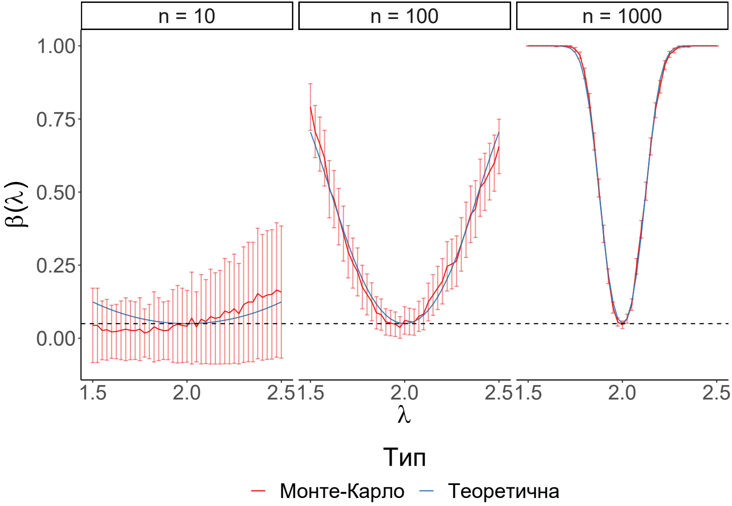
Тест Волда (6)

- Рахуємо тестові статистики і проводимо тестування

```
ns <- c(10, 100, 1000)
lambda <- 2
T <- 1000
df <- NULL
for (lambda_prime in seq(1.5, 2.5, by = 0.025)){
  df <- rbind(df, power_simulation(ns, lambda, lambda_prime, T))
}
```

Тест Волда (7)

- Графічна ілюстрація:



- Теоретична потужність зростає з n, а емпірична наближається до теоретичної
- Потужність у точці $\lambda = 2$ дорівнює 0.05, як і планувалося

Еквівалентність тесту Волда та довірчих інтервалів

- Можна показати, що тест Волда відкидає гіпотезу $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ тоді й тільки тоді, коли

$$\theta_0 \notin \left[\hat{\theta} - \widehat{\mathsf{se}}\left(\hat{\theta}\right) \cdot z_{1-\alpha/2}; \hat{\theta} + \widehat{\mathsf{se}}\left(\hat{\theta}\right) \cdot z_{1-\alpha/2}\right]$$

- Це в певному сенсі самоочевидно, адже тест відкидає H_0 тоді й тільки тоді, коли $|T| > z_{1-\alpha/2}$
- Відтак він **не відкидає** H_0 тоді й тільки тоді, коли $|T| \leq z_{1-\alpha/2}$, тобто

$$-z_{1-\alpha/2} \leq T \leq z_{1-\alpha/2}$$

- Власне, довірчі інтервали й утворюють у такий спосіб, шляхом **інверсії тесту** (test inversion)
  - Ви можете самостійно в аналогічний спосіб збудувати довірчі інтервали (чи то пак, промені) і для однобічних тестів Волда

t-тест

- Нехай тест Волда використовується для перевірки середнього вибіркового $\overline{X}$
- Як ми розглядали раніше, якщо нам невідома справжня дисперсія, ми використовуємо її оцінку, і тоді тестова статистика дорівнює

$$T = \frac{\overline{X} - \theta_0}{s_X / \sqrt{n}}$$

- Якщо X має нормальний розподіл, то тоді $T \sim t_{n-1}$, і тест Волда має назву **t-теста** (t-test)
- На практиці ми не знаємо, який саме розподіл має X, а якщо n дуже велике, то різниця між t-розподілом і стандартним нормальним розподілом несуттєва
- Незважаючи на це, у цьому випадку все одно за замовчуванням виконують t-тест
    - Вважається, що t-тест стійкий до ненормальности розподілу
- Це схоже на нашу вищенаведену дискусію про довірчі інтервали з квантилями нормального розподілу та квантилями t-розподілу

p-значення (1)

- Підхід до тестування, за якого ми спочатку задаємо рівень значущости α, а потім визначаємо, яка повинна бути область відкидання, є доволі нудним
- Натомість ми можемо просто підрахувати значення тестової статистики T і з'ясувати, для якого **найменшого** α тест відкине H_0
- Таке значення називають p**-значенням** (p-value):

$$p = \inf \{\alpha : T \in R_\alpha\} \tag{2.2}$$

- Тут R_α — область відкидання для тесту рівня α
- У випадку, коли тест рівня α передбачає відкидання H_0 тоді й тільки тоді, коли $T \geq c_\alpha$, маємо $p = \inf \{\alpha : T \geq c_\alpha\}$
- З іншого боку, оскільки тест має рівень α, найвища ймовірність помилки I роду дорівнює α
  - Тобто $\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta (T \geq c_\alpha)$
- Через монотонну спадність найменше значення α досягається, коли c_α є найбільшим

p-значення (2)

- Для деякої конкретної вибірки $\mathbf{x} = (x_1, \ldots, x_n)$ таким значенням є $T(\mathbf{x})$ — **реалізація** T для цієї вибірки
- Справді, якщо використати будь-яке значення, вище від $T(\mathbf{x})$, то тест не зможе відкинути H_0
- Іншими словами, у цьому випадку

$$p = \mathbb{P}_{\theta_0} \left(T \geq T(\mathbf{x}) \right) \tag{2.3}$$

- Аналогічно для інших двох тестів:
  - Якщо тест відкидає тоді й тільки тоді, коли $T \leq c_\alpha$, то $p = \mathbb{P}_{\theta_0} \left(T \leq T(\mathbf{x}) \right)$
  - Якщо тест відкидає тоді й тільки тоді, коли $|T| \geq c_\alpha$, то $p = \mathbb{P}_{\theta_0} \left(|T| \geq |T(\mathbf{x})| \right)$
- p-значення має інтерпретацію ймовірности зустріти реалізацію тестової статистики, **щонайменше таку ж екстремальну**, як і підрахована для вибірки $\mathbf{x}$

Інтерпретація p-значень (1)

- Що менше p-значення, то більше підстав ми маємо, щоб відкинути H_0
  - Тобто якщо ми її відкинемо, імовірність помилки буде щонайбільше p
- Проте часто їх інтерпретація є некоректною
- Можна почути, що p — це ймовірність того, що H_0 є істинна
  - Це нісенітниця
- Часто в дослідженнях можна побачити, що та чи та оцінка супроводжується певними позначками
  - Як правило, $*$ означає, що $p < 0.1$
  - Як правило, $**$ означає, що $p < 0.05$
  - Як правило, $***$ означає, що $p < 0.01$
- Наразі спостерігається тенденція не наводити таких позначок, а концентрувати увагу читача на інтерпретації довірчих інтервалів

Інтерпретація p-значень (2)

- Також потрібно розуміти, що оцінка може бути **статистично значущою** (statistically significant)
  - Тобто мати мале p-значення з погляду деякого тесту
- Але при цьому вона може бути **малою**
- Часто дослідники намагаються подати як визначний здобуток статистично значущі результати, які на практиці мають дуже малий вплив
  - Наприклад, різниця в доходах може бути дуже статистично значущою, але дорівнювати 1 грн
- Тому довірчі інтервали мають більшу практичну цінність, особливо в контексті тесту Волда

Приклад тесту Волда (1)

- Розгляньмо питання, чи є вцілілі молодші (в середньому) від загиблих пасажирів
- Розгляньмо дві популяції $X_1, \ldots, X_n \sim X$ (вік уцілілих) та $Y_1, \ldots, Y_m \sim Y$ (вік загиблих)
- Припустімо, що вони між собою **незалежні**
    - Тобто вважаємо, що пасажири могли вціліти або загинути **абсолютно випадково**
    - Це дуже сильне припущення, і в регресійному аналізі ми побачимо, що з цим можна зробити
- Тоді нас цікавить тестування гіпотези $H_0 : \mu_X - \mu_Y \leq 0$ vs. $H_1 : \mu_X - \mu_Y > 0$
- Відповідними оцінками будуть $\hat{\mu}_X = \overline{X}$ та $\hat{\mu}_Y = \overline{Y}$
- Обидві оцінки асимптотично мають нормальний розподіл: $\hat{\mu}_i \overset{\text{a}}{\sim} N\left(\mu_i, \text{Var}\left(\hat{\mu}_i\right)\right)$, $i = X, Y$
- З урахуванням незалежности вибірок, $\hat{\mu}_X - \hat{\mu}_Y \sim N\left(\mu_X - \mu_Y, \text{Var}\left(\hat{\mu}_X\right) + \text{Var}\left(\hat{\mu}_Y\right)\right)$
    - Відтак $\widehat{\text{se}}\left(\hat{\mu}_X - \hat{\mu}_Y\right) = \sqrt{\widehat{\text{Var}}\left(\hat{\mu}_X\right) + \widehat{\text{Var}}\left(\hat{\mu}_Y\right)}$
- Тоді тестова статистика дорівнюватиме

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y - 0}{\widehat{\text{se}}\left(\hat{\mu}_X - \hat{\mu}_Y\right)} = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\widehat{\text{Var}}\left(\hat{\mu}_X\right) + \widehat{\text{Var}}\left(\hat{\mu}_Y\right)}}$$

- А p-значення дорівнюватиме $p = \mathbb{P}_{\mu_X - \mu_Y = 0}\left(T \leq T(\mathbf{x})\right) = \Phi(T(\mathbf{x}))$

Приклад тесту Волда (2)

- Розгляньмо реалізацію в R

```
estimates <- passengers %>% filter(!is.na(Age)) %>% group_by(Survived) %>%
  summarise(mean_hat = mean(Age),
            var_hat = var(Age) / n())

mean_hat_s <- estimates %>% filter(Survived == 1) %>% pull(mean_hat)
mean_hat_ns <- estimates %>% filter(Survived == 0) %>% pull(mean_hat)

var_hat_s <- estimates %>% filter(Survived == 1) %>% pull(var_hat)
var_hat_ns <- estimates %>% filter(Survived == 0) %>% pull(var_hat)

se <- sqrt(var_hat_s + var_hat_ns)

T <- (mean_hat_s - mean_hat_ns) / se

p_value <- pnorm(T, lower.tail = FALSE)
conf.int <- c(mean_hat_s - mean_hat_ns - qnorm(0.95)*se, Inf)

mean_hat_s
```

```
## [1] 28.34369
```

```
mean_hat_ns
```

```
## [1] 30.62618
```

```
p_value
```

```
## [1] 0.9796233
```

```
conf.int
```

```
## [1] -4.117439       Inf
```

- Як можна бачити, у даних **немає достатньо підстав**, щоб відкинути H_0
  - Тобто нічого конкретного ми сказати не можемо
  - Але оскільки p-значення дуже велике, то можна вважати, що вік загиблих не є (статистично значущо) менший від уцілілих

Приклад тесту Волда (3)

- Схожого результату можна досягнути за допомогою функції `t.test`

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

```
##
##  Welch Two Sample t-test
##
## data:  Age by Survived
## t = 2.046, df = 598.84, p-value = 0.9794
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 4.120282
## sample estimates:
## mean in group 0 mean in group 1
##        30.62618        28.34369
```

- Оскільки значення `0` іде раніше від `1`, R вважає, що H_0 у цьому випадку полягає в тому, що середній вік для `Survived = 0` **більший** від середнього віку `Survived = 1`
  - А H_1, відповідно, що **менший** (`less`)
  - Це відповідає нашій ситуації
- Зверніть увагу, що цей тест має назву **t-тесту Велча** (Welch's t-test)[3]
  - На дуже великих вибірках (тобто асимптотично) він подібний до t-тесту, тому ми на цьому не зупиняємося
- Можна вказувати, яку природу має альтернативна гіпотеза
  - За замовчуванням маємо значення аргументу `alternative = "two.sided"`
  - Для однобічних гіпотез можна вказувати `"less"` або `"greater"`
  - Також можна вказувати аргумент `mu`, що відповідає гіпотетичній різниці (за замовчуванням дорівнює 0)

[3] Бернард Велч (Bernard Lewis Welch, 1911–1989) — британський статистик

Приклад тесту Волда (4)

- До речи, якби ми взяли $H_0 : \mu_X - \mu_Y = 0$ vs. $H_1 : \mu_X - \mu_Y \neq 0$, то ми **відкинули б** цю гіпотезу
  - Перевірте це самостійно!
- Але що це *в дійсності* означає?
- Що між вцілілими та загиблими в середньому суттєво відрізняється вік?
- Ми не враховували інших факторів
  - Цілком очевидно, що стать, клас квитка тощо також повинні відігравати свою роль
  - Тобто вибірки вцілілих та загиблих явно **не можуть** бути незалежні
  - Якби ми тестували якісь ліки і випадково роздавали ліки і плацебо, тоді можна було б щось коректно порівнювати
  - Потрібні інші методи, які розглядатимемо далі
- Більше того, навіть якби вибірки були справді незалежні, що ми власне можемо показати?
  - Цілком очевидно, що між уцілілістю та віком не існує причиново-наслідкового зв'язку
  - Зв'язок може **гіпотетично** існувати в інший бік (що дуже сумнівно)
  - Але тоді t-тест застосувати просто так не вийде

Приклад тесту Волда (5)

- Розгляньмо ситуацію, коли нас цікавить, чи є між імовірностями вціліти для осіб двох статей статистично значуща різниця
  - Тут уже можна щось припускати про **вплив** стати на ймовірність уціліти
- Як і в попередньому випадку, вважатимемо, що маємо дві популяції $X_1, \ldots, X_n \sim X$ та $Y_1, \ldots, Y_m \sim Y$
  - X_i — вижив чоловік (1) або ні (0)
  - Y_i — вижила жінка (1) або ні (0)
- Нас цікавить тестування гіпотези $H_0 : p_X - p_Y = 0$ vs. $H_1 : p_X - p_Y \neq 0$
  - Тут p_X — імовірність уціліти для чоловіків, а p_Y — для жінок
- Вважатимемо знову, що вони між собою **незалежні**
  - Тобто будь-який випадковий пасажир може бути або чоловіком, або жінкою
  - І за іншими його характеристиками неможливо вгадати його стать
  - Це дуже **неправдоподібно**, але ходімо далі
- Відповідними оцінками будуть $\widehat{p}_X = \overline{X}$ та $\widehat{p}_Y = \overline{Y}$
  - Обидві оцінки асимптотично мають нормальний розподіл: $\widehat{p}_i \overset{\text{a}}{\sim} N\left(p_i, \text{Var}\left(\widehat{p}_i\right)\right)$, $i = X, Y$
  - З урахуванням незалежности вибірок, $\widehat{p}_X - \widehat{p}_Y \sim N\left(p_X - p_Y, \text{Var}\left(\widehat{p}_X\right) + \text{Var}\left(\widehat{p}_Y\right)\right)$
  - Відтак $\widehat{\text{se}}\left(\widehat{p}_X - \widehat{p}_Y\right) = \sqrt{\widehat{\text{Var}}\left(\widehat{p}_X\right) + \widehat{\text{Var}}\left(\widehat{p}_Y\right)}$

Приклад тесту Волда (6)

- Проте на цей раз маємо справу з величинам Бернуллі (можуть дорівнювати тільки або 0, або 1)
  - Тому $\text{Var}\left(\hat{p}_i\right) = \frac{p_i(1-p_i)}{n}, i = X, Y$
  - Відтак $\widehat{\text{Var}}\left(\hat{p}_i\right) = \frac{\hat{p}_i(1-\hat{p}_i)}{n}, i = X, Y$
- Тоді тестова статистика дорівнюватиме

$$T = \frac{\hat{p}_X - \hat{p}_Y - 0}{\widehat{\text{se}}\left(\hat{p}_X - \hat{p}_Y\right)} = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n}}}$$

- А p-значення дорівнюватиме $p = \mathbb{P}_{p_X - p_Y = 0}\left(|T| \leq |T(\mathbf{x})|\right) = 2\Phi(T(\mathbf{x}))$

Приклад тесту Волда (7)

- Для нашого випадку з пасажирами маємо

```
estimates <- passengers %>% group_by(Sex) %>%
  summarise(p_hat = mean(Survived),
            var_hat = p_hat * (1 - p_hat) / n(),
            n = n(),
            n_surv = sum(Survived == 1))

p_hat_1 <- estimates %>% filter(Sex == "male") %>% pull(p_hat)
p_hat_2 <- estimates %>% filter(Sex == "female") %>% pull(p_hat)

var_hat_1 <- estimates %>% filter(Sex == "male") %>% pull(var_hat)
var_hat_2 <- estimates %>% filter(Sex == "female") %>% pull(var_hat)

se <- sqrt(var_hat_1 + var_hat_2)

T <- (p_hat_1 - p_hat_2) / se

p_value <- 2*pnorm(-abs(T))
conf.int <- c(p_hat_1 - p_hat_2 - qnorm(0.975)*se, p_hat_1 - p_hat_2 + qnorm(0.975)*se)

p_hat_1
```

```
## [1] 0.1889081
```

```
p_hat_2
```

```
## [1] 0.7420382
```

```
p_value
```

```
## [1] 5.187071e-78
```

```
conf.int
```

```
## [1] -0.6111119 -0.4951483
```

- p-значення фактично дорівнює 0, тобто можемо відкинути гіпотезу про однаковість середніх

Приклад тесту Волда (8)

- Але, знову ж таки, це за умови, що вибірки були незалежні
  - У чому є сумніви
- Навіть якщо припустити незалежність, то на ймовірність уціліти впливає **не тільки** стать
  - Тобто **не можна** стверджувати, що різниця -0.55 з'явилася **виняткова** за рахунок впливу стати
  - Що з цим робити, вивчатимемо далі в нашому курсі
- Схожого результату можна досягнути за допомогою функції `prop.test`

```
prop.test(estimates$n_surv, n = estimates$n)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  estimates$n_surv out of estimates$n
## X-squared = 260.72, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.4926894 0.6135708
## sample estimates:
##    prop 1    prop 2
## 0.7420382 0.1889081
```

- У нашому випадку ми подали на її вхід:
  - Вектор кількостей уцілілих по двох категоріях
  - Вектор загального числа спостережень по двох категоріях

Інші приклади простеньких тестів Волда

- За аналогією з порівнянням середніх вибіркових також можна порівнювати медіани
  - Асимптотична нормальність дає підстави застосовувати тест Волда
  - Стандартні похибки можна шукати за допомогою бутстрепа
- Можуть бути ситуації, що вибірки $X_1, \ldots X_n$ та $Y_1, \ldots, Y_n$ є **паровані** (paired)
  - Тобто кожне спостереження i відповідає одному об'єкту, а X і Y — це його дві різні характеристики
  - Наприклад, дохід до і після навчання на програмі
  - Тоді потрібно розглянути різниці $D_i = X_i - Y_i$ і виконати тест Волда для цих різниць
  - В R для цього можна використати функцію `t.test`, передавши аргумент `paired = TRUE`
  - Вона використовує t-розподіл, але для великих вибірок результати будуть дуже близькі до асимптотичного тесту Волда

Тест χ^2 для перевірки мультиномного розподілу (1)

Визначення 2.1

- Нехай n об'єктів можна незалежно одне від одного віднести до однієї з k категорій
- Нехай у категорію j об'єкт можна віднести з імовірністю p_j, $\sum_{j=1}^k p_j = 1$
- Нехай X_j — число об'єктів у категорії j, $\sum_{j=1}^k X_j = n$
- Тоді $\mathbf{X} = (X_1, \dots, X_k)^\top$ має **мультиномний розподіл** (multinomial distribution) із параметрами n і $\mathbf{p} = (p_1, \dots, p_k)^\top$: $\mathbf{X} \sim \mathrm{Mult}_k\left(n, \mathbf{p}\right)$

□

- Можна вивести спільну функцію ймовірности такого випадкового вектора:

$$
\begin{aligned}
\mathbb{P}_{\mathbf{X}}\left(X_1 = n_1, \dots, X_k = n_k\right) &= \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} \\
&\times \prod_{j=1}^k \mathbb{1}\left\{n_j \in \mathbb{Z}^+\right\} \cdot \mathbb{1}\left\{n_1 + \dots + n_k = n\right\}
\end{aligned}
\tag{2.4}
$$

Тест χ^2 для перевірки мультиномного розподілу (2)

- Розгляньмо **тест χ^2 Пірсона** (Pearson's χ^2 test)[4]
- Маємо нульову гіпотезу $H_0 : \mathbf{p} = \mathbf{p}_0$ vs. $H_1 : \mathbf{p} \neq \mathbf{p}_0$
- Тестовою статистикою є

$$T = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j} = \sum_{j=1}^{k} \frac{(X_j - np_{0,j})^2}{np_{0,j}} \tag{2.5}$$

  - $\mathbf{p}_0 = (p_{0,1}, \dots, p_{0,k})$ — вектор теоретичних імовірностей за умов виконання H_0
  - E_j — теоретичне сподівання X_j, яке за умов виконання H_0 для мультиномного розподілу дорівнює $\mathbb{E}\left[X_j\right] = np_{0,j}$
- Можна показати (але ми цього робити не будемо), що, за умов виконання H_0, $T \xrightarrow{d} \chi^2_{k-1}$
- Відтак тест рівня α полягає в тому, що H_0 відкидають, якщо $T > q_{\chi^2_{k-1},\alpha}$
  - Тобто якщо тестова статистика перевищила відповідний квантиль розподілу χ^2 із $k-1$ ступенями свободи
- А p-значення дорівнює $\mathbb{P}\left(T > t\right)$, де t — реалізація тестової статистики в наявних даних

[4] Карл Пірсон (Karl Pearson, 1857–1936) — британський статистик

Тест χ^2 для перевірки мультиномного розподілу (3)

- Практична цінність такого тесту сильно зменшується через те, що потрібно точно знати, який саме розподіл ми перевіряємо
  - У багатьох випадках ліпше виконувати непараметричні оцінки на кшталт оцінки щільности тощо
- Проте є одне практичне застосування цього тесту, яке може мати обмежену користь
- Нехай маємо таблицю спряжености з r рядками та c стовпцями
  - Елементи таблиці позначмо через $n_{ij}, i = 1, \dots, r, j = 1, \dots, c$
  - Маржинальні значення позначмо через $n_i = \sum_{j=1}^{c} n_{ij}$ та $n_j = \sum_{i=1}^{r} n_{ij}$
  - Загальну суму значень позначмо через $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$
- Нас цікавить перевірити, чи сумісні емпіричні частоти в таблиці з гіпотезою H_0 про **незалежність** відповідних змінних
- Якби змінні були незалежні, то ми мали б $\frac{n_{ij}}{n} = \frac{n_i n_j}{n^2}$
- Відтак теоретичними сподіваннями за умови H_0 в цьому випадку будуть значення $E_{ij} = \frac{n_i n_j}{n}$
- Тоді тестова статистика перетворюється в

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

- Асимптотично $T \xrightarrow{d} \chi^2_{(r-1)(c-1)}$

Тест χ^2 для перевірки мультиномного розподілу (4)

- Розгляньмо таку таблицю спряжености для пасажирів «Титаніку»

```
cont_tab <- xtabs(~ Sex + Survived, data = passengers)
cont_tab
```

```
##         Survived
## Sex        0   1
##   female  81 233
##   male   468 109
```

- Нас цікавить протестувати, чи є такий розподіл випадковим, чи осіб деякої статі непропорційно багато через уцілілих

```
margin_rows <- rowSums(cont_tab)
margin_cols <- colSums(cont_tab)

Eij <- margin_rows %*% t(margin_cols) / sum(cont_tab)

T <- sum((cont_tab - Eij)^2 / Eij)

pchisq(q = T, df = 1, lower.tail = FALSE)  # df = (2 - 1)*(2 - 1)
```

```
## [1] 3.711748e-59
```

- Тобто ми бачимо, що H_0 **явно** не може бути істинна
- Точно такого результату можна досягти за допомогою функції `chisq.test`

```
chisq.test(cont_tab, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_tab
## X-squared = 263.05, df = 1, p-value < 2.2e-16
```

- Якщо вказати `correct = TRUE` (це значення за замовчуванням), то статистика рахуватиметься з `(abs(cont_tab - Eij) - 0.5)^2` у чисельнику

Типові помилки у використанні статистичного виведення

- Тестування гіпотез дуже поширене
  - Можливо, навіть більше, ніж хотілося б
  - У багатьох випадках довірчі інтервали інформативніші й корисніші
- Відтак часто виникають помилкові й непрофесіональні інтерпретації[5]
- Уявімо, що $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$
  - Тобто нам хочеться знайти свідчення того, що $\theta \neq 0$, а відтак спостерігається якийсь цікавий ефект
  - Наприклад, що навчальна програма підвищує дохід
- Розгляньмо деякі **помилкові** судження
- Якщо $p < 0.05$, то існує ненульовий ефект (або, що ще гірше, те, що ми знайшли, і є справжній ефект)
  - Довірчий інтервал може бути, наприклад, $[0.02; 4]$, тобто формально нуль не входить, але все одно дуже близько
- Якщо $p \approx 0$, то ефект дуже посутній
  - Насправді це всього лише каже, що довірчий інтервал буде дуже вузький
  - Але на величину ефекту це не впливає
- Якщо для чоловіків довірчий інтервал є $[0.2; 3.2]$, а для жінок — $[-0.2; 2.8]$, то ефект є для чоловіків, але не для жінок
  - Це нісенітниця, адже ці інтервали повністю сумісні з ситуацією, що ефект для чоловіків насправді дорівнює 0.5, а для жінок, наприклад, — 2.7

[5] Цей і наступний слайд створено на основі окремих лекцій Вілла Фітіана (Will Fithian) з University of California, Berkeley

Деякі загальні коментарі щодо тестування гіпотез

- У загальному випадку, потрібно розуміти, що тестування гіпотез вимагає дуже чіткого усвідомлення **припущень**, які висуваються до даних, **імовірнісної моделі**, яка описує ці дані, та **конкретного тесту**
  - Наприклад, якщо асимптотичний розподіл не є нормальний, то тест Волда безсенсовний, хоча R залюбки нам порахує все, що завгодно
- Навіщо ми в принципі тестуємо речі типу $\theta = \theta_0$, якщо ми знаємо, що неперервна випадкова величина ніколи не може **точно** дорівнювати якомусь дійсному числу?
  - У принципі, можна тестувати і гіпотези виду $H_0 : \theta \in [a; b]$ vs. $H_1 : \theta \notin [a; b]$, просто це не так поширено
  - Виходячи зі знаку тестової статистики, ми можемо принаймні сказати, у який бік від 0 лежить справжнє значення параметру
- Часто непрофесіональні дослідники використовують довірчі інтервали та p-значення, присвоюючи їм невластиві ймовірнісні інтерпретації
  - Наприклад, що якщо 95% інтервал $C = [0.5; 0.8]$, то параметр лежить у цьому інтервалі з імовірністю 95%
  - Це нісенітниця, але тільки в частотному підході
  - У Беєсівському підході така інтерпретація справді можлива, але ми ці методи тут не розглядаємо
  - Цікаві міркування на цю тему можна прочитати тут і в статті, викладеній на диску

p-hacking (1)

- У науковій літературі, особливо в галузях соціальних наук, медицини та економіки, набуло поширення так зване явище p-hacking
- Як відомо, для тесту рівня 0.05 помилка I роду стається з імовірністю 5%
- Тобто якби ми проводили на одних даних багато тестів, кожний 20-ий давав би статистично значущий результат там, де його насправді немає
- Дослідники можуть зловживати цим, пробуючи різні підходи, поки p-значення якогось конкретного тесту не стане достатньо низьким
- Такий підхід до аналізу даних є **безвідповідальним**

p-hacking (2)

- По-перше, для тестування багатьох різних гіпотез існують спеціальні методи **множинного тестування** (multiple testing)
- По-друге, усе поширенішою стає практика, коли спочатку дослідник чітко викладає суть свого дослідження, експерименту, який він збирається провести, гіпотези, які він очікує побачити в даних, та процедуру тестування
  - Тільки після того, як рецензенти затвердять такий план, дослідник збирає дані та проводить тестування
  - Навіть якщо в результаті гіпотезу **не буде** відкинуто, статтю все одно публікують
- Спочатку потрібно чітко сформулювати, що саме ми хочемо проаналізувати, і тільки потім використовувати дані для підтвердження своїх гіпотез
  - Принципово неправильно формулювати гіпотези на основі ознайомлення з самими даними
  - Якщо підкидати монетку багато разів і помітити, що герб випадає в половині разів, то тестування гіпотези $H_0 : \mathbb{P}\left(\text{герб}\right) = 0.5$ просто безглуздо!!!
  - Треба спочатку поставити завдання — дізнатися ймовірність випаду герба, а потім протестувати цю гіпотезу на даних

Множинне тестування (1)

- На практиці постійно виникає потреба тестувати не якусь одну, а декілька гіпотез
  - Особливо критичним це є, коли потрібно проводити тисячі різних тестів одночасно
  - Наприклад, щоб виявити, який із генів впливає на певний результат тощо
- Проблема полягає в тому, що якщо тестувати гіпотези окремо, то помилка I роду ставатиметься частіше, ніж потрібно
- Справді, нехай потрібно протестувати гіпотезу $H_0 : \mu_1 = 0$ і $\mu_2 = 0$ vs. $H_1 : \mu_1 \neq 0$ чи $\mu_2 \neq 0$ (у тому числі одночасно)
- Нас цікавить, наприклад, рівень значущости $\alpha = 0.05$
- Якщо використовувати t-статистики T_1 та T_2 відповідно та відкидати гіпотези окремо одну від одної, то рівень такого тесту буде неправильний
- Тоді маємо

$$\mathbb{P}_{H_0}\left(|T_1| > 1.96 \text{ чи } |T_2| > 1.96\right) > \mathbb{P}_{H_0}\left(|T_1| > 1.96\right) = 0.05$$

- Зокрема, якщо $T_1 \perp\!\!\!\perp T_2$, то можна швидко підрахувати, що $\mathbb{P}_{H_0}\left(|T_1| > 1.96 \text{ чи } |T_2| > 1.96\right) = 0.0975$
  - Тобто ймовірність помилки майже удвічі вища!
- Щоб імовірність помилки I роду була адекватною, потрібно вносити певні коригування в процес тестування

Множинне тестування (2)

- Можна застосувати популярний, проте доволі консервативний і не дуже потужний **метод Бонферроні** (Bonferroni method)[6]
- Нехай маємо m тестів H_{0i} vs. H_{1i}, $i = 1, \ldots, m$
- Нехай $p_1, \ldots, p_m$ — p-значення, що відповідають цим тестам
- Суть методу полягає в тому, що потрібно відкидати не ті гіпотези, для яких $p_i < \alpha$, а ті, для яких $p_1 < \frac{\alpha}{m}$
- Доведення цього твердження дуже просте:

$$\begin{aligned}\mathbb{P}\left(\text{щонайменше одна помилка I роду}\right) &= \mathbb{P}\left(\bigcup_{i=1}^{m} \text{помилка I роду в гіпотезі } i\right) \\ &\leq \sum_{i=1}^{m} \mathbb{P}\left(\text{помилка I роду в гіпотезі } i\right) \\ &= \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha\end{aligned}$$

- Існують і інші методи зі схожою ідеєю, проте ми їх розглядати в цьому курсі не будемо

[6] Карло Еміліо Бонферроні (Carlo Emilio Bonferroni, 1892–1960) — італійський математик

Множинне тестування (3)

- Інша група методів принципово відмінна
- Замість того, щоб мінімізувати ймовірність допустити помилку I роду, вони намагаються контролювати так звану **частоту хибних відкриттів** (false discovery rate, FDR)
- Нехай m_0 — число нульових гіпотез, які є істинними, і нехай $m_1 = m - m_0$
- Тоді можна класифікувати тести за таким принципом:

	H_0 істинна	H_1 істинна	Разом
H_0 не відкинуто	U	T	$m - R$
H_0 відкинуто	V	S	R
Разом	m_0	m_1	m

- Можна розглянути поняття **частки хибних відкриттів** (false discovery proportion, FDP):

$$\mathrm{FDP} = \frac{V}{R} \cdot \mathbb{1}\left\{R > 0\right\} \tag{2.6}$$

  - Це частка хибних відкидань
- Тоді

$$\mathrm{FDR} = \mathbb{E}\left[\mathrm{FDP}\right] \tag{2.7}$$

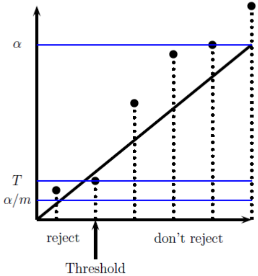
Множинне тестування (4)

- Розгляньмо одну з найпоширеніших процедур для контролювання FDR — **метод Беньяміні-Хохберга** (Benjamini-Hochberg method)[7]
- Нехай $p_1, \dots, p_m$ упорядковано за зростанням
- Нехай $\ell_i = \alpha \frac{i}{m}$
- Нехай $R = \max\{i : p_i < \ell_i\}$
- Тоді потрібно відкинути всі гіпотези, для яких $p_i \leq T = p_R$

[7] Йоав Беньяміні (Yoav Benjamini, 1949) та Йосеф Хохберг (Yosef Hochberg) — ізраїльські математики

Множинне тестування (5)

- Графічна ілюстрація (Wasserman, Fig. 10.6)



- Якби ми тестували гіпотези окремо, ми повинні були б відкинути всі, для яких p-значення менші від α (5 із 6)
- Якби ми використовували метод Бонферроні, ми б відкинули всі гіпотези, для яких p-значення менші від $\frac{\alpha}{m}$ (жодну)
- Нарешті, за методом BH ми спочатку визначаємо поріг T як **останнє** p-значення, що лежить під відповідною прямою $\frac{\alpha}{m}i$
  - А потім відкидаємо всі гіпотези, для яких p-значення менші від цього T (тільки дві)
- Цей підхід має вищу потужність, ніж метод Бонферроні та подібні до нього

Множинне тестування (6)

- В R ці, та багато інших, ідей реалізовано у функції `p.adjust` із базового пакету `stats`
- На вхід потрібно подати вектор p-значень та вказати потрібний метод
- На виході буде вектор скоригованих p-значень
- Ідея такого підходу в тому, що можна, грубо кажучи, порівнювати p-значення з $\frac{\alpha}{m}$, а можна mp — з α
- Розгляньмо це на прикладі порівняння середніх цін на діаманти різного кольору, вага яких менша від 0.25 карата, із відповідного датасету
- Спочатку підрахуймо всі можливі тести за допомогою `t.test` та зберімо відповідні p-значення у вектор[8]

```
diamonds_ttests <- diamonds %>%
  filter(carat < 0.25) %>%
  select(price, color) %>% group_by(factor(color)) %>%
  summarise(across(price, list)) %>%
  tidyr::expand(nesting(color1 = `factor(color)`, price1 = price),
                nesting(color2 = `factor(color)`, price2 = price)) %>%
  filter(color1 != color2) %>%
  rowwise %>% mutate(key = paste0(sort(c(color1, color2)), collapse = "-")) %>%
  distinct(key, .keep_all = TRUE) %>%
  select(color1, color2, price1, price2) %>%
  mutate(p_value = t.test(price1, price2)$p.value) %>% ungroup()
```

[8] Код модифіковано з цього сайту

Множинне тестування (7)

- Тепер можемо застосувати декілька методів корекції

```
diamonds_ttests %>%
  mutate(p_value_bonferroni = p.adjust(p_value, method = "bonferroni"),
         p_value_BH = p.adjust(p_value, method = "BH"),
         reject = p_value < 0.05,
         reject_bonferroni = p_value_bonferroni < 0.05,
         reject_BH = p_value_BH < 0.05) %>%
  select(-c("price1", "price2")) %>%
  print(n = 21)
```

```
## # A tibble: 21 x 8
##    color1 color2   p_value p_value_bonferroni p_value_BH reject reject~1 rejec~2
##    <fct>  <fct>      <dbl>              <dbl>      <dbl> <lgl>  <lgl>    <lgl>
##  1 D      E      0.733               1          0.733    FALSE  FALSE    FALSE
##  2 D      F      0.0939              1          0.123    FALSE  FALSE    FALSE
##  3 D      G      0.192               1          0.224    FALSE  FALSE    FALSE
##  4 D      H      0.0114              0.240      0.0343   TRUE   FALSE    TRUE
##  5 D      I      0.00652             0.137      0.0228   TRUE   FALSE    TRUE
##  6 D      J      0.0528              1          0.101    FALSE  FALSE    FALSE
##  7 E      F      0.0651              1          0.105    FALSE  FALSE    FALSE
##  8 E      G      0.0576              1          0.101    FALSE  FALSE    FALSE
##  9 E      H      0.000763            0.0160     0.00801  TRUE   TRUE     TRUE
## 10 E      I      0.00522             0.110      0.0219   TRUE   FALSE    TRUE
## 11 E      J      0.0564              1          0.101    FALSE  FALSE    FALSE
## 12 F      G      0.00235             0.0495     0.0124   TRUE   TRUE     TRUE
## 13 F      H      0.0000121           0.000253   0.000253 TRUE   TRUE     TRUE
## 14 F      I      0.00169             0.0355     0.0118   TRUE   TRUE     TRUE
## 15 F      J      0.0425              0.893      0.0992   TRUE   FALSE    FALSE
## 16 G      H      0.241               1          0.267    FALSE  FALSE    FALSE
## 17 G      I      0.0274              0.575      0.0718   TRUE   FALSE    FALSE
## 18 G      J      0.0728              1          0.109    FALSE  FALSE    FALSE
## 19 H      I      0.0920              1          0.123    FALSE  FALSE    FALSE
## 20 H      J      0.104               1          0.128    FALSE  FALSE    FALSE
## 21 I      J      0.310               1          0.326    FALSE  FALSE    FALSE
## # ... with abbreviated variable names 1: reject_bonferroni, 2: reject_BH
```

- Як можна бачити, після корекції не всі гіпотези потрібно відкинути

Множинне тестування (8)

- Такого ж результату можна було б досягти за допомогою функції `pairwise.t.test`

```
diamonds_small <- diamonds %>% filter(carat < 0.25)
pairwise.t.test(diamonds_small$price, factor(diamonds_small$color),
                p.adjust.method = "BH", pool.sd = FALSE)
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  diamonds_small$price and factor(diamonds_small$color)
##
##   D       E       F       G       H       I
## E 0.73279 -       -       -       -       -
## F 0.12329 0.10512 -       -       -       -
## G 0.22404 0.10087 0.01236 -       -       -
## H 0.03428 0.00801 0.00025 0.26677 -       -
## I 0.02284 0.02194 0.01183 0.07183 0.12329 -
## J 0.10087 0.10087 0.09918 0.10927 0.12841 0.32564
##
## P value adjustment method: BH
```

- Проте попередній підхід загальніший і працюватиме для тестів різного роду