**Stepik Course Part 1**

**Module 1**

**Lecture 1 Генеральна сукупність і вибірка**

**Генеральна сукупність** – це всі значення. **Вибірка** – це певна сукупність значень. **Репрезентативність** – це ступінь схожості вибірки на генеральну сукупність.

Вибірка:

* **проста випадкова вибірка** (simple random sample);
* **стратифікована вибірка** (stratified sample) – спочатку генеральна вибірка ділиться на декілька страт (груп) по певних параметрах;
* **групова вибірка** (cluster sample) – спочатку генеральна вибірка ділить на кластери, які дуже схожі між собою;

**Lecture 2 Типи змінних**

Типи змінних вибірки:

* **кількісні**:

1. **неперервні** – змінна може приймати будь-яке значення на певному проміжку;
2. **дискретні** – значення може приймати тільки певні значення;

* **якісні** (**номінативні**) – використовуються для розподілу на групи, при цьому цифри не несуть математичного змісту, наприклад у вибірці чоловіки отримають цифру 1, а жінки 2.
* **рангові** (частковий випадок якісної змінної) – розподіл вибірки по рангу, наприклад результат забігу;

**Lecture 3 Міри тенденцій**

Два типа обов’язкових статистик:

* **міри центральної тенденції**:

1. **мода** – значення, яке зустрічається найчастіше (може бути декілька мод, якщо однакова кількість), найкраща для асиметричних гістограм;
2. **медіана** – значення середнього елемента, якщо парна кількість то два середніх просумувати і поділити на два (елементи записують по зростанню);
3. **середнє або мат.очікування** – це середнє значення від усіх - суму поділити на кількість (для генеральної сукупності позначають грецькою буквою **µ** (мю), для вибірки символом ), найкраща для симетричних гістограм і не підходить для асиметричних;
4. **усічене середнє** – це середнє значення з відкинутими 5-10% найбільших і найменших значень, якщо є аномальні величини;

**Властивості мір центральної тенденцій**:

1. µx+c = µx + c
2. µx\*c = µx \* c
3. = 0 – сума всіх відхилень рівна 0;

* **міри мінливості**:

1. **розмах** (**range**) – різниця між максимальним і мінімальним значенням;
2. **міжквартильний розмах** (**interquartile range**) – різниця між найбільшим і найменшим значенням з відкинутими 25% найбільших і найменших значень, також міжквартильний розмах записується як різниця між першим і третім квартилями;
3. **відхилення** (**deviation**) – різниця між середнім значенням та конкретно взятим значенням, сума таких різниць рівна нулю;
4. **дисперсія** (**variance**) – сума квадратів всіх відхилень поділено на кількість значень (зазвичай позначають буквою **D**), для вибірки ділиться на (n-1);
5. **середньоквадратичне відхилення (стандартне відхилення)** – корінь від дисперсії (для генеральної сукупності позначають грецькою буквою **σ** (**сигма**), для вибірки буквою **S** або **sd** - **standard deviation**), для вибірки береться дисперсія вибірки (n-1);

**Властивості мір мінливостей**:

1. Dx+c = Dx
2. sdx+c = sdx
3. Dx\*c = Dx \* c2
4. sdx\*c = sdx \* c

**Lecture 4 Квартилі розподілу та графік box-plot**

**Квартиль** – це 3 точки, які ділять дані на 4 рівні частини. Другий квартиль співпадає з медіаною.

**box-plot** – графік, центральна лінія ящика показує медіану, нижня і верхня лінії показують 1 і 3 квартилі. Ширина (висота по суті) ящика – різниця між 1 і 3 квартилем – міжквартильний розмах (більшість значень вибірки). Вуса мають розмір до півтора розміру ширини ящика (міжквартильного розмаху). Нижній вус – спостережуваний мінімум значення, верхній вус – спостережуваний максимум значення. Окремими точками відображаються викиди.

**Lecture 5 Нормальний розподіл**

**Нормальний розподіл** – це **унімодальний** (одна вершина на гістрограмі) **симетричний** розподіл, який підкоряється правилу 3-х сігм (схожий а капелюх).

**Стандартизація** (**z - перетворення**) – перетворення даних в стандартну Z-шкалу, де середнє (µ = 0) і дисперсія (D = 1). Щоб зробити z – перетворення потрібно від кожного значення вибірки відняти середнє і поділити на стандартне відхилення:

Правило 3-х сігм:

* 68% всіх значень попадають в діапазон µ.
* 95% в діапазоні µ.
* 99% в діапазоні µ.

Приклад: Считается, что значение IQ (уровень интеллекта) у людей имеет нормальное распределение со средним значением равным 100 и стандартным отклонением равным 15 (M = 100, sd = 15).

Какой приблизительно процент людей обладает IQ > 125?

Результат: Рахуємо .

Дальше в таблицях дивимось значення для z = 1.66. В нашому випадку використовуємо другу таблиця із списку нижче і отримуємо результат 0.0485 = 4.8% - тобто ймовірність зустріти людину з IQ > 125 буде 4.8%

<http://users.stat.ufl.edu/~athienit/Tables/Ztable.pdf> - процент значення, які не перевищують Z.

<http://www.normaltable.com/ztable-righttailed.html> - процент значення, які перевищують Z.

<http://davidmlane.com/hyperstat/z_table.html> – онлайн калькулятор для розрахунку проценту.

**Lecture 6 Центральна гранична теорема**

**Центральна гранична теорема -** якщо із генеральної сукупності взятивибірки однакової величини і порахувати середнє значення для всіх вибірок і побудувати розподіл цих вибіркових середніх, то такий розподіл буде нормальним, із середнім, яке співпадає з середнім генеральної сукупності (буде дуже близьким).

Стандартне відхилення для такого розподілу називається **стандартною помилкою середнього** і показує наскільки в середньому вибіркові значення відхиляються від середнього у генеральній сукупності.

**Стандартна помилка середнього**: , n – кількість значень у вибірці.

Чим більший розмір вибірки і кількість вибірок, тим ближчі значення вибіркових середніх до середнього генеральної сукупності. Чим менше стандартна помилка середнього, тим рідше вибіркові середні будуть сильно відхилятися від середнього в генеральної сукупності.

Якщо кількість значень у вибірці більше 30 (число спостережень) і ця вибірка є репрезентативна (значення вибрані випадково), то у формулі стандартної помилки ми можемо підкласти значення стандартного відхилення даної вибірки замість стандартного відхилення генеральної сукупності:

На основі даних тільки цієї вибірки ми можемо припустити як будуть вести себе всі вибіркові середні для інших вибірок – вони б розмістились з відхиленням (se) від певного невідомого стандартного відхилення генеральної сукупності (µ).

**Lecture 7 Довірчі інтервали для середнього**

**Довірчий інтервал** – це інтервал на вибірці, центр якого знаходить в точці середнього значення вибірки (), а сам інтервал в зоні () для 95% ймовірності.

Із ймовірністю 95% середнє значення генеральної сукупності належить розрахованому довірчому інтервалу.

Для 99% ймовірності формула довірчого інтервалу матиме вигляд: ()

**Lecture 8 Ідея статистичного висновку, p-рівень значущості**

**Нульова гіпотеза** (Н0) – змін не відбулось, середнє вибірки дорівнює середньому генеральної сукупності (значення відрізняється через випадковість).

**Альтернативна гіпотеза** (Н1) – зміни відбулись, середнє вибірки не дорівнює середньому генеральної сукупності.

**Ідея статистичного висновку** – припускаємо гіпотезу Н0, то середні вибірок () розмістились би нормально навколо середнього генеральної сукупності (µ) зі стандартним відхиленням ( ). Дальше потрібно визначити наскільки вибіркове середнє віддалилось від середнього генеральної сукупності в одиницях стандартного відхилення.

Для цього робиться z-перетворення:

Далі потрібно розрахувати ймовірність відхилення середнього за допомогою p-рівня значимості. Якщо p < 0.05 – можна приймати альтернативну гіпотезу, а якщо більше то не достатньо підстав відхилити нульову гіпотезу.

<https://gallery.shinyapps.io/dist_calc/> - показує ймовірність відхилення від середнього в одиницях стандартного відхилення (p).

p – рівень значущості – дозволяє сказати про ймовірність Н0.

Помилки статистичного виду:

* **Помилки першого роду** – відкинули Н0, а вона була правильна, результати були випадкові;
* **Помилки другого роду** – не відкинули Н0, а правильною була Н1;

**Module 2 Порівняння середніх**

**Lecture 9 T-розподілення**

**Розподілі Стьюдента (t-distribution)** – кількість об’єктів у вибірці мала (n < 30) і середнє генеральної сукупності (σ) не відоме. Даний розподіл являється унімодальним і симетричним. На відмінку від нормального розподілу в розподілі Стьюдента «вищі хвости», тобто вибіркові значення з більшою ймовірністю попадають за межі ±2σ від µ. Форма такого розподілу (графіка) залежить від числа ступенів свободи (df = n - 1). Чим більше df тим більше розподіл схожий на нормальний.

Фактично нам зазвичай не відоме середнє генеральної сукупності (σ), тому для перевірки гіпотез використовується та ж формула, що і для знаходження z – значення, тобто t – розподілення буде: , де

Розрахунок p – рівня значимості для t – розподілу:

<https://gallery.shinyapps.io/dist_calc/>

<https://vavilovva.shinyapps.io/dist_calc/> - вказати «Distribution» в позицію «t» і Degrees of freedom (df) = n - 1;

**Lecture 10 Порівняння двох середніх; t-критерій Стьюдента**

**t-критерій Стьюдента (t-test)** – дозволяє порівняти дві вибірки між собою (два вибіркові середні).

Н0: дві вибірки з генеральної сукупності однакові: µ1 = µ2

Н1: дві вибірки з генеральної сукупності не однакові: µ1 ≠ µ2

Рахується як різниця середніх значень двох вибірок поділено на стандартну помилку цієї різниці: ,, при

Далі розрахуємо p–рівень значимості при знайдених t та df.

Вимоги до вибірок при використанні t-критерія Стьюдента:

* **гомогенність дисперсії** – однакові дисперсії для обох вибірок. Перевірити це можна за допомогою критерію Левене або критерію Фішера;
* **нормальність** **розподілу** двох вибірок для малого об’єму вибірок (n < 30);

**Довірчий інтервал** можна порахувати більш коректно, якщо знати t-розподіл, особливо при n < 30. В формулі зони довірливого інтервалу (), де , замість (1.96) потрібно підкласти значення з таблиці нижче, при заданому (df = n - 1) і потрібній точності (95% чи іншій):

<https://www.medcalc.org/manual/t-distribution.php>

**Lecture 11 Перевірка розподілу на нормальність, QQ-Plot**

**Перевірка розподілу на нормальність** – наскільки наший розподіл відрізняться від нормального. Для цього потрібно побудувати гістограму частот вибірки і поверх накласти криву ідеального нормального розподілу.

**QQ Plot (Normal Q-Q Plot)** – показує наскільки добре значення вибірки (sample quantiles) відповідають передбачуваним (Theoretical Quantiles). Точки над прямою показують значення, які вищі за нормальне розподілення, а під прямою – нижче за нормальне розподілення.

Для перевірки нормальності розподілу використовують спеціальні тести:

* **тест колмогорова-смирнова**;
* **тест шапиро-уилка**;

Дані тести дозволяють перевірити, що вибірка взята з генеральної сукупності, де розподіл нормальний (p > 0.05 для нормального розподілу генеральної сукупності).

**U-критерій Манна-Уїтні** – непараметричний аналог t – критерію. Даний критерій не настільки чутливий до викидів.

**Lecture 12 Однофакторний дисперсійний аналіз**

**Дисперсійний аналіз (analysis of variance (ANOVA))** – дозволяє порівняти будь-яку кількість груп (вибірок) між собою (вибіркові середні). Змінна, яка розділяє генеральну сукупність на групи (**номінативна змінна** з декількома градаціями) називається **незалежною змінною**, а **кількісна змінна**, за ступенем вираженості якої ми порівнюємо групи – **залежна змінна**.

**SST (total sum of squares anova)** – загальна сума квадратів, яка характеризує наскільки сильна мінливість даних, без врахування розподілу їх на групи.

– середнє для всіх значень.

**SST** – рахується як сума квадратів відхилень всіх значень від загального середнього.

**df (для SST)** – число ступенів свободи, рахується: df = N – 1, де N – загальна кількість об’єктів

**SSB (sum of squares between)** – сума квадратів між групами, рахується як сума квадратів відхилення середнього групи від загального середнього, помножено на кількість елементів в групі.

**df (для SSB - міжгрупова)** = m – 1, де m – кількість груп.

**SSW (sum of squares within)** – сума квадратів всередині груп, рахується як сума квадратів відхилень всіх значень від середнього кожної групи.

**df (для SSW - внутрішньо групова)** = N – m, де N – загальна кількість об’єктів, m – кількість груп.

Якщо SSB > SSW – групи об’єктів сильно відрізняються, а якщо SSB < SSW – групи об’єктів не відрізняються.

**F** - **основний статистичний показник дисперсійного аналізу**. Розраховується по формулі:

– для Н0 (всі вибірки рівні) має прямувати до нуля, відповідно (F) буде дуже малим.

Для знаходження ймовірності (p) використаємо:

<https://gallery.shinyapps.io/dist_calc/> - вказати «Distribution» в позицію «F», df - міжгрупова ступінь свободи, df(2) – внутрішньо групова ступінь свобод та (a) – значення розрахованого F.

При p < 0.05 відкидаємо нульову гіпотезу про рівність і приймаємо альтернативну.

Теоретичний розподіл F-значення в дисперсійному аналізі не є нормальним, а підпорядковується розподілу Фішера (F - distribution).

**Lecture 13 Множинні порівняння в ANOVA**

Для великої кількості вибірок (3 і більше) число вибірок, які порівнюються попарно рахується по формулі: , де n – кількість вибірок (гіпотез).

**Проблема множинних порівнянь** – якщо з генеральної сукупності вибирати велику кількість вибірок (гіпотез) і попарно порівнювати t-критерієм Стьюдента, то ймовірність побачити значні відмінності (p < 0.05) стрімко зростає, хоча всі вибірки з однієї генеральної сукупності.

Для вирішення даної проблеми вводять різні поправки, зокрема **поправку Бонферроні** – потрібно поділити ймовірність (p = 0.05) на кількість порівнянь (k) – це і буде наша статистично достовірна ймовірність. Поправка Бонферроні дуже консервативна, з нею дуже складно отримати статистично значиму відмінність.

**Критерій Тьюки** – порівнює попарно всі комбінації середніх, але по іншому рахується стандартна помилка.

**Lecture 14 Багатофакторний ANOVA**

**Багатофакторний ANOVA** – дозволяє порівняти велику кількість вибірок по декільком параметрам (**незалежні змінні**). Параметр, який змінюється в залежності від вибору незалежних змінних називається **залежною змінною** (наприклад: підсумковий бал - залежна змінна, факультет і курс - незалежні змінні).

**Module 3**

**Lecture 15 Поняття кореляції**

**Кореляція** – показує взаємозв’язок двох кількісних змінних.

**Діаграма розсіювання або точкова** – кожна вісь характеризує свою ознаку – наприклад «великі і швидкі» об’єкти будуть знаходитись в правому верхньому куті.

Види кореляцій:

* **позитивна кореляція** – при рості одного параметра, росте і інший;
* **негативна кореляція** – при рості одного параметра, падає інший;

**Коефіцієнт кореляції** **Пірсона** – кількісний показник сили і напрямку взаємодії двох змінних. Рахується по формулі: , де cov – **коваріація**.

rxy – приймає значення в діапазоні [-1, 1]:

* rxy = -1 – негативна кореляція;
* rxy = 1 – позитивна кореляція;
* rxy = 0 – змінні між собою не зв’язані;

Чим ближче значення коефіцієнта кореляції до -1 або 1, тим сильніша взаємодія між змінними.

<http://rpsychologist.com/d3/correlation/> - візуально показує зв'язок змінних при різних значеннях параметра кореляції.

**Коефіцієнт детермінації** **(R2)** – квадрат коефіцієнту кореляції. Показує в якій мірі дисперсія однієї змінної впливає на дисперсію іншої. Приймає значення в діапазоні [0, 1].

**Н0 для коефіцієнту кореляції** – rxy = 0 – зв’язку між змінними немає.

**Н1 для коефіцієнту кореляції** – rxy ≠ 0 – зв'язок між змінними є.

df = N - 2

**Lecture 16 Умови застосування коефіцієнта кореляції**

Умови застосування коефіцієнта кореляції:

* характер зв’язку між змінними має бути **лінійним** (описати силу і напрямок взаємозв’язку двох змінних можна **лінією**) і **монотонним**;
* **нормальний розподіл** обох змінних (симетрія і унімодальність);
* **відсутність викидів**;

Якщо розподіл не нормальний або зв'язок змінних не лінійний, то можна використати не параметричні аналоги коефіцієнту Пірсона:

* коефіцієнт кореляцій Спірмана;
* коефіцієнт кореляції Кендалла (тау Кендалла);

**Помилка кореляції** – позитивна або негативна кореляції між змінними не обов’язково говорять про причинно-наслідкові зв’язки між ними.

**Lecture 17 Регресія з однією незалежною змінною**

**Регресійний аналіз** – аналіз впливу незалежної змінної або змінних на залежну.

**Одновимірний регресивний аналіз** – дослідження взаємодії двох змінних. Незалежна змінна відображається на осі Х і називається **предиктором** (**регресором**), а залежна – на осі Y.

**Лінія регресії** – відображає максимально точно взаємозв’язок двох змінних, має формулу: ,

де b0 – intercept (де пряма пересікає вісь Y), b1 - slope (показує напрямок і кут нахилу (по знаку) до вісі X).

**Метод найменших квадратів** – метод знаходження оптимальних параметрів лінії регресії (параметри b0 і b1), таких, що сума квадратів залишків (квадратів відстаней точок до ідеальної лінії регресії) була мінімальна, має формулу: , ,

де - середнє значення змінної Y, - середнє значення змінної X.

b0 – показує значення змінної (y) при x = 0, а b1 – відношення змінних (y) до (x).

Якщо коефіцієнт кореляції між двома змінними дорівнює нулю, і обидві змінні представлені в z – значеннях, то рівняння регресії набуде вигляду y = 0.

**Lecture 18 Гіпотеза про значимість взаємозв'язку і коефіцієнт детермінації**

Н0 (для rxy = 0) – b1 = 0.

Н1 – b1 ≠ 0, то rxy ≠ 0.

**Коефіцієнт лінійної регресії** – якщо провести велику кількість експериментів при rxy = 0, то вибіркові значення b1 розподілились би нормально навколо нуля, рахується по формулі: , де set – стандартна помилка розподілу цього коефіцієнту; df = N-2, де N – кількість вибірок.

**t-критерій** – перевіряє гіпотезу, що коефіцієнт b1 відмінний від нуля (чи є зв'язок між змінними).

**Коефіцієнт детермінації (R2)** – доля залежної змінної Y, яка пояснюється регресійною моделлю. Показує в якій мірі дисперсія однієї змінної впливає на дисперсію іншої. Приймає значення в діапазоні [0, 1]. Рахується по формулі:

,

де ssres – сума квадратів залишків (квадратів відстаней точок до ідеальної лінії регресії), sstotal – загальну сума квадратів (квадратів відстаней точок до лінії середнього значення по Y).

При R2 ≈ 1 – майже 100% мінливості залежної змінної Y пояснюється її взаємодією з незалежною змінною X. Тобто регресійна модель дуже добре пояснює і описує поведінку і мінливість залежної змінної.

Чи вищий R2, тим краще регресійна модель пояснює і передбачає значення залежної змінної.

R2 = 0, то і коефіцієнт b1 (slope) також дорівнює нулю.

**Lecture 19 Умови застосування лінійної регресії з одним предиктором**

Умови застосування лінійної регресії з одним предиктором:

* **лінійний** взаємозв'язок змінних X і Y;
* **нормальний** розподіл залишків;
* **гомоскедастичність** – незалежність дисперсії випадкових складових від номера спостереження (постійна мінливість залишків на всіх рівнях незалежної змінної);

<https://gallery.shinyapps.io/slr_diag/> - перевірка на умови застосування лінійної регресії.

**Lecture 20 Застосування регресійного аналізу та інтерпретація результатів**

Якщо коефіцієнт b1 (slope) не рівний нулю, то це не завжди означає, що ми відхиляємо нульову гіпотезу про відсутність зв’язку.

**Lecture 21 Передбачення значень залежної змінної**

**Лінія тренду** – регресійна пряма, дозволяє передбачити значення залежної змінної при конкретних значеннях незалежної змінної.

Поведінка змінної являється лінійною тільки на певному відрізку (наприклад відношення віку до росту буде лінійним до приблизно 20 років).

Умови для передбачення змінної:

* Регресійна модель має добре описує взаємозв’язок двох змінних:

1. **Лінійність** взаємозв’язку;
2. **Нормальний** розподіл залишків;
3. Постійна мінливість залишків - **гомоскедантичність**;

* Обмеження значень незалежної змінної в певних межах (на великих значеннях поведінка залежної змінної може стати не лінійною);

**Lecture 22 Регресійний аналіз з декількома незалежними змінними**

Множинна регресія дозволяє дослідити вплив відразу декількох незалежних змінних на одну залежну змінну.

Формула для множинної регресії:

Для двох незалежних змінних апроксимація множинної регресії матиме вигляд площини в трьохвимірному просторі.

Методом найменших квадратів знаходяться оптимальні значення b0, b1, … , bn таких, що сума квадратів залишків буде мінімальною.

Умови застосування множинної регресії (вимоги до даних):

* **Лінійна** залежність змінних;
* **Нормальний** розподіл залишків;
* **Гомоскедастичність**;
* Перевірка на **мультиколінеарність**;
* (Бажано) **Нормальний розподіл** всіх змінних (залежної і незалежних);

Множинна регресія дає можливість подивитись на вплив однієї незалежної змінної на залежну при умові, що інші незалежні змінні являються константами (наприклад розмір ноги впливає на інтелект, тому що з віком росте розмір ноги та інтелект, але при множинній регресії вік береться за константу і значимої залежності між розміром ноги і інтелектом не буде).

**Lecture 23 Вибір найкращої моделі**

**Мультиколінеарність** – дуже сильний взаємозв’язок між якимись з незалежних змінних.

Незалежна змінна, яка має сильну кореляцію з іншими незалежними змінними може **погіршувати** регресивну модель опису взаємодії змінних.

**Підбір оптимальної моделі опису взаємодії змінних**:

1. Внести в модель всі змінні і розрахувати виправлений коефіцієнт детермінації Adj. R-squared;
2. По черзі видаляти різні змінні і рахувати Adj. R-squared;
3. Для початку варто видалити незалежну змінну, які має сильну кореляцію з іншими незалежними змінними;
4. Таким чином знайти модель з найкращим Adj. R-squared;
5. Із з найденої моделі по черзі видаляти різні змінні і рахувати Adj. R-squared;
6. Продовжуємо, поки не знайдемо найбільше Adj. R-squared;

**Lecture 24 Класифікація: логістична регресія і кластерний аналіз**

**Логістична регресія** – метод пошуку взаємозв’язку між залежною змінною, яка має тільки дві градації (два можливих значення) і незалежною змінною (предиктором).

**Кластерний аналіз** – метод розбиття заданої вибірки на підмножини, що називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Представляє ієрархічну структуру даних.