Контроль набутих знань
практика 9

Текстовий опис завдання:

Розглянемо вибірку пасажирів Титаніка (train.csv). Наша ціль перевірити, чи середній вік пасажирів 30 років. Проблема полягає в тому, що для значної частини суб'єктів у вибірці вік невідомий, що може значно вплинути на адекватність висновків. Для вирішення проблеми використано підхід "Multiple data imputation", який полягає в заповнені порожніх значень змінної випадковими величинами, що розподілені аналогічно досліджуваній змінній. Процес повторюється кілька разів після чого проводиться аналіз усіх ітерацій.

Кроки виконання:

1. Підготовка:
   1.1. Створіть код в SAS Studio з назвою у форматі - lastname_cw9.sas;
   1.2. Створіть у SAS Studio макрозмінні для свого SAS ID та прізвища.
   1.3. Створіть бібліотеку(назва бібліотеки=прізвище інтерна), в якій будете зберігати лише фіналізовані датасети (проміжні датасети зберігайте в директорії WORK).
2. Попередній аналіз:
   2.1. Імпортуйте датасет train.csv.
   2.2. За допомогою PROC SQL порахуйте кількість записів з порожнім полем віку та загальну кількість записів. Збережіть результати в макрозмінні з назвами "missing", та "total" відповідно. А також, обрахуйте відсоток порожніх записів відносно вибірки та збережіть результат в макрозмінній "percent".
   2.3. За допомогою процедури PROC MEANS. Збережіть значення статистик середнього та стандартного відхилення в макрозмінні з назвами "mean", "stddev" відповідно.
   2.4. За допомогою PROC UNIVARIATE побудуйте гістограму змінної Age з кривою нормального розподілу.
3. Перший крок до "Multiple data imputation" для однієї змінної:
   3.1. В дата кроці замініть порожні значення Age на випадкові числа з відповідними параметрами розподілу (обраховані в пункті 2.3). Додайте змінну "imputation" на надайте їй значення 1, оскільки це перша ітерація. (Зверніть увагу, Age вважаємо нормально розподіленою; Age не може набувати від'ємних значень).
   3.2. Повторіть пункт 2.4 з імпутованим датасетом. Упевніться, що після імпутації гістограма і крива не зазнали значних змін *(див. додаток).*
4. "Multiple data imputation" для однієї змінної:
   4.1. Створіть макрос, %single_var_mi(inds, var, nimps=10);
   Обов'язкові параметри: inds – вхідний датасет;
   var – числова змінна, у якій заповнюватимуться порожні значення;
   Опціональні параметри: nimps – кількість ітерацій;

   4.2. Макрос обраховує середнє та стандартне відхилення змінної var.
   4.3. Потім nimps разів заповнює порожні значення var випадковими числами (аналогічно пункту 3).

4.4. Результат роботи макроса – датасет "imput_var" вигляду:

imputation	&var.	Інші змінні з &inds.
1	xx	
1	xx	
1	xx	
…	xx	
2	xx	
2	xx	
2	xx	
…	xx	
&nimps.	xx	
&nimps.	xx	
&nimps.	xx	

5. Використання макросу:
   5.1. Запустіть макрос single_var_mi з параметрами inds=train, var=age, nimps=5.
   5.2. Застосуйте PROC TTEST на отриманому датасеті для перевірки нульової гіпотези H_0: mean(Age)=30; (використайте "by" для окремого аналізу кожної ітерації). Результати збережіть в датасет "ttest_result", що матиме наступний вигляд:

imputation	variable	t_statistic	p
1	Age	xxx	0.xxx
2	Age	xxx	0.xxx
…	Age	xxx	0.xxx

   5.3. Виведіть в лог макрозмінні "missing", "total", "percent ", "mean", "stddev" у вигляді ноутів.
6. Результат роботи макроса " imput_var", та датасет "ttest_result" експортуйте в csv.

7. Додаткове завдання:
   7.1. Виконайте "Multiple data imputation" за допомогою: PROC MI -> PROC TTEST -> PROC MIANALYZE; (для виконання цього завдання необхідно аналізувати принаймні дві змінні, тому це завдання виконати для змінних Age, Fair).

7.2. Інтерпретуйте результати роботи процедури MIANALYZE. Які висновки можна зробити стосовно гіпотези про середній вік в 30 років.

Матеріали:

1. https://documentation.sas.com/doc/en/statug/15.2/statug_ttest_gettingstarted01.htm
2. https://stefvanbuuren.name/fimd/sec-nutshell.html (1.4.1-1.4.2)
3. https://documentation.sas.com/doc/en/pgmsascdc/v_062/procstat/procstat_univariate_syntax09.htm
4. https://www.listendata.com/2015/03/multiple-imputation-with-sas.html

Додаток:

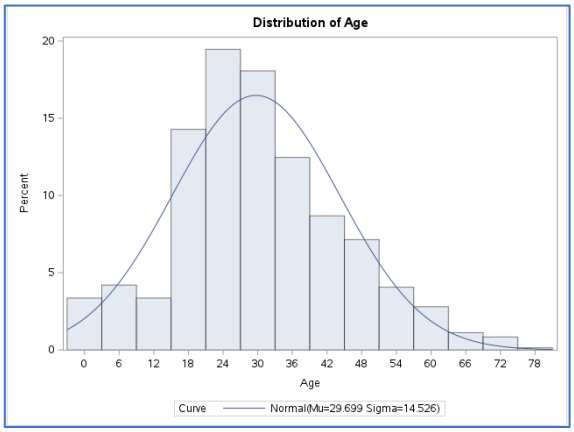

Рисунок 1 Гістограма початкових даних

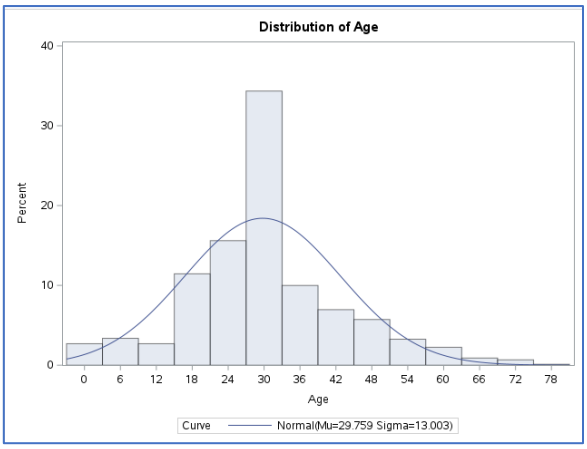

Рисунок 2 Пропущені значення заповненні середнім. (гістограма зазнала значних змін)

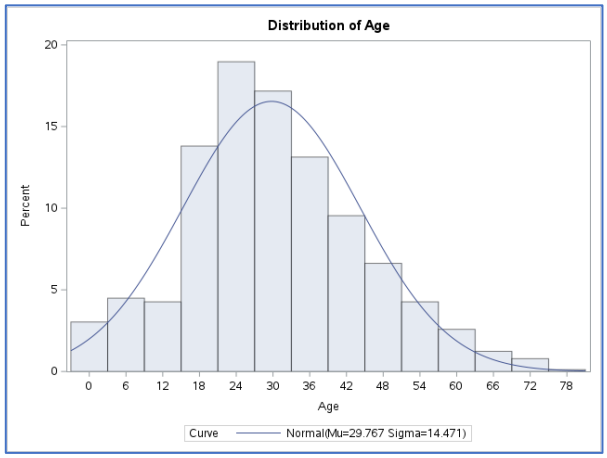

Рисунок 3 Пропущенні значення заповнені випадковими числами з аналогічним розподілом (гістограма подібна до початкової)