ỨNG DỤNG PYTHON TRONG XỬ LÝ VÀ KIỂM ĐỊNH SỐ LIỆU THỐNG KÊ

1. Giới thiệu

Trong nghiên cứu khoa học sức khỏe, việc xử lý và phân tích số liệu không chỉ nhằm mô tả đặc điểm mẫu nghiên cứu mà còn để kiểm định giả thuyết và rút ra kết luận có giá trị suy luận. Python, với hệ sinh thái thống kê mạnh mẽ, cho phép thực hiện toàn bộ quy trình phân tích số liệu một cách minh bạch, tái lập và phù hợp với chuẩn mực nghiên cứu quốc tế.

Chuyên đề này trình bày có hệ thống các phương pháp xử lý và kiểm định số liệu thường dùng, bao gồm: thống kê mô tả, kiểm định tỷ lệ và kiểm định trung bình, kèm theo các câu lệnh Python minh họa chuẩn hóa..

2. Môi trường và thư viện phân tích

Trong phân tích thống kê bằng Python, các thư viện sau được sử dụng như tiêu chuẩn:

```
import pandas as pd
import numpy as np
from scipy import stats
from statsmodels.stats.proportion import proportions_ztest
```

Dữ liệu nghiên cứu được giả định lưu dưới dạng bảng trong DataFrame df.

3. Thống kê mô tả biến định tính: Tần số và tỷ lệ

Ví dụ bạn khảo sát điểm PHQ-9 (trầm cảm), GAD-7 (lo âu), DASS-21 (stress/lo âu/trầm cảm), PSQI (chất lượng giấc ngủ) hoặc chỉ số sinh học như cortisol huyết tương: việc xác định các biến này có tuân phân phối chuẩn hay không sẽ giúp chọn so sánh trung bình giữa nhóm, mô hình hồi quy, hay tương quan phù hợp (tham số hay phi tham số).

3.1.Tần số (Frequency)

Tần số phản ánh số lượng quan sát của từng mức biến và thường được trình bày cho các biến định tính như giới tính, tình trạng bệnh, mức độ tuân thủ điều trị.

Tính tần số: df['gioi_tinh'].value_counts()

Bảng tần số chéo (Crosstab): pd.crosstab(df['gioi_tinh'], df['benh'])

3.2. Tỷ lệ (Proportion / Percentage)

Tỷ lệ giúp chuẩn hóa tần số, cho phép so sánh giữa các nhóm có quy mô khác nhau và thường được trình bày dưới dạng n (%).

Tỷ lệ phần trăm toàn mẫu:

df['gioi_tinh'].value_counts(normalize=True) * 100

Tỷ lệ theo nhóm (tỷ lệ hàng):

pd.crosstab(df['gioi_tinh'], df['benh'], normalize='index') * 100

4. Thống kê mô tả biến định lượng: Trung bình và độ phân tán

4.1. Trung bình và độ lệch chuẩn

Đối với biến định lượng phân phối gần chuẩn, trung bình và độ lệch chuẩn là các tham số mô tả quan trọng nhất.

df['tuoi'].mean()

df['tuoi'].std()

df['tuoi'].describe()

4.2. Trung bình theo nhóm

df.groupby('gioi_tinh')['tuoi'].mean()

Kết quả này thường được sử dụng làm cơ sở cho các kiểm định so sánh trung bình.

5. Kiểm định thống kê đối với tỷ lệ

5.1. Kiểm định Chi-square (χ^2)

Mục đích: So sánh tỷ lệ giữa hai hoặc nhiều nhóm độc lập.

bang = pd.crosstab(df['gioi_tinh'], df['benh'])

chi2, p, dof, expected = stats.chi2_contingency(bang)

Báo cáo: χ^2 = … ; df = … ; p = …

5.2. Kiểm định Fisher's Exact

Áp dụng khi cỡ mẫu nhỏ hoặc tần số kỳ vọng < 5 (bảng 2×2).

oddsratio, p = stats.fisher_exact(bang)

5.3. Kiểm định một tỷ lệ so với giá trị chuẩn

Ví dụ: So sánh tỷ lệ tuân thủ điều trị quan sát với tỷ lệ chuẩn 70%.

count = 65

nobs = 100

value = 0.70

z_stat, p = proportions_ztest(count, nobs, value=value)

5.4. So sánh hai tỷ lệ độc lập

count = np.array([40, 55])

nobs = np.array([80, 100])

z_stat, p = proportions_ztest(count, nobs)

6. Kiểm định thống kê đối với trung bình

6.1. Kiểm định t-test một nhóm

stats.ttest_1samp(df['hba1c'], popmean=7)

6.2. Kiểm định t-test hai nhóm độc lập

nam = df[df['gioi_tinh'] == 'Nam']['tuoi']

```
nu  = df[df['gioi_tinh'] == 'Nu']['tuoi']
stats.ttest_ind(nam, nu, equal_var=False)
```

6.3. Kiểm định t-test cặp (trước – sau)

```
stats.ttest_rel(df['truoc'], df['sau'])
```

6.4. Phân tích phương sai ANOVA một chiều

```
stats.f_oneway(
   df[df['nhom'] == 1]['tuoi'],
   df[df['nhom'] == 2]['tuoi'],
   df[df['nhom'] == 3]['tuoi'])
```

7. Kiểm định phi tham số

Áp dụng khi dữ liệu không phân phối chuẩn hoặc có ngoại lệ.

7.1. Mann–Whitney U

```
stats.mannwhitneyu(nam, nu)
```

7.2. Wilcoxon signed-rank

```
stats.wilcoxon(df['truoc'], df['sau'])
```

7.3. Kruskal–Wallis

```
stats.kruskal(
   df[df['nhom'] == 1]['tuoi'],
   df[df['nhom'] == 2]['tuoi'],
   df[df['nhom'] == 3]['tuoi'])
```

8. Chuẩn trình bày trong bài nghiên cứu khoa học

"Số liệu được xử lý và phân tích bằng Python (Pandas, SciPy, Statsmodels). Thống kê mô tả được trình bày dưới dạng tần số, tỷ lệ (%), trung bình ± độ lệch chuẩn. So sánh tỷ lệ sử dụng kiểm định Chi-square hoặc Fisher's Exact. So sánh trung bình sử dụng t-test hoặc ANOVA; khi dữ liệu không phân phối chuẩn áp dụng các kiểm định phi tham số tương ứng. Mức ý nghĩa thống kê được chọn là $p < 0,05$."

9. Kết luận

Python là một nền tảng phân tích thống kê toàn diện, đáp ứng đầy đủ yêu cầu xử lý, kiểm định và suy luận số liệu trong nghiên cứu khoa học sức khỏe hiện đại. Việc chuẩn hóa quy trình phân tích bằng Python giúp nâng cao tính khoa học, minh bạch và khả năng tái lập của nghiên cứu.

TÀI LIỆU THAM KHẢO

1. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585(7825):357–362. doi:10.1038/s41586-020-2649-2.

2. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17(3):261–272. doi:10.1038/s41592-019-0686-2.

3. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1991.

Tác giả bài viết: Nguyễn Thái Bình