statisticka_praca

September 13, 2023

1 Štatistická práca

Patrik Broček

1.1 Úvod

V mojej štatistickej práci budem analyzovať dáta o nakupovaní v supermarketoch získané z : https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset. Dáta však **nie sú plne realistické**, pretože pre každý jednotlivý **nákup** je v datasete uvedená iba **jedna zakúpená položka**, jej množstvo a cena. Z tohto dôvodu budem celý čas uvažovať len svet v ktorom sa dá v obchode kúpiť iba jedna položka a všetky závery ktoré odvodím budú platné len v pod touto podmienkou.

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from prettytable import PrettyTable

data = pd.read_csv("dataset/supermarket/archive/customer_shopping_data_parsed.
 ↪csv")
print(data.head())
```

```
   gender  age   category payment_method     total
0  Female   28   Clothing           Card   7502.00
1    Male   21      Shoes           Card   5401.53
2    Male   20   Clothing           Cash    300.08
3  Female   66      Shoes           Card  15004.25
4  Female   49  Cosmetics           Cash     40.66
```

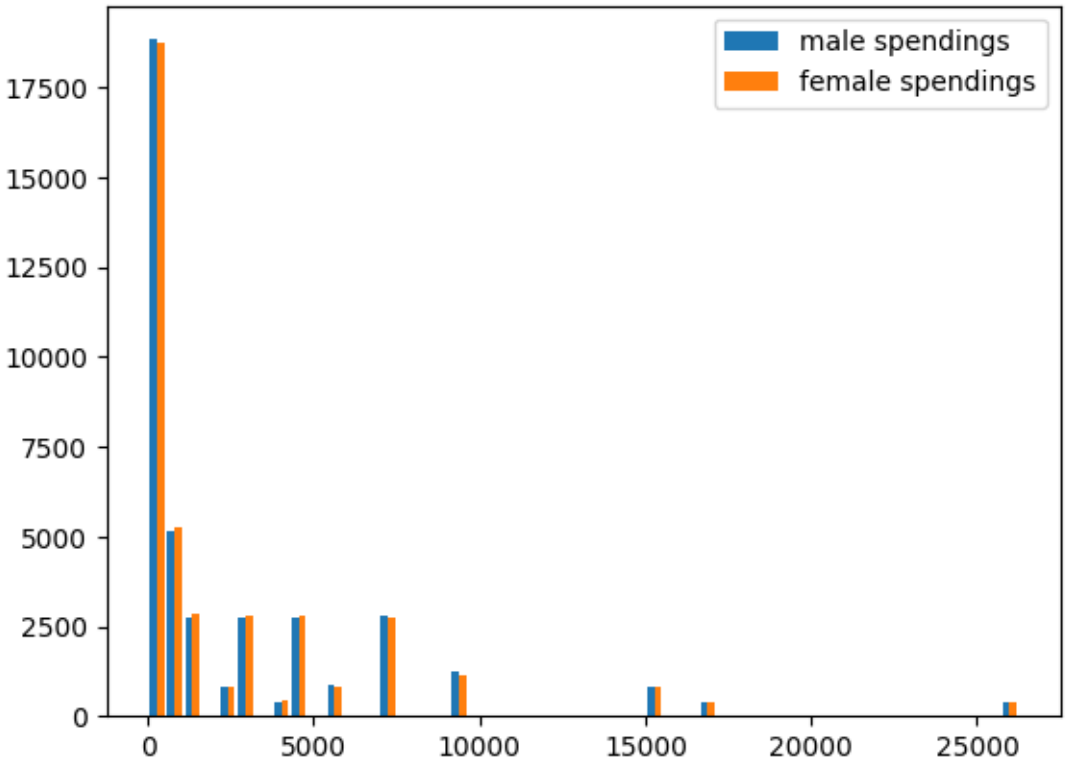
1.2 Ženy míňajú na nákupoch viac ako muži

Nemám potuchy čo s tým, ale šak nejak to skúsim nehehe

[ ]:
```
males = data[data["gender"] == "Male"]
females = data[data["gender"] == "Female"]
```

[ ]:
```
# Graph
bins = np.linspace(data["total"].min(), data["total"].max(), 50)
plt.hist([males["total"], females["total"]], bins, label=["male spendings",␣
 ↪"female spendings"])
plt.legend(loc='upper right')

# Table
table = PrettyTable(); table.field_names = males["total"].describe().index
table.add_row(males["total"].describe().values)
table.add_row(females["total"].describe().values)
table.add_column("gender", ["male", "female"])
print(table)
```

```
+--------+-------------------+------------------+------+--------+--------+---
----------------+---------+--------+
|  count |        mean       |        std       | min  |  25%   |  50%   |
75%         |   max   | gender |
+--------+-------------------+------------------+------+--------+--------+---
----------------+---------+--------+
| 39975.0 | 2534.0502371482175 | 4216.352328763888 | 5.23 | 130.75 | 600.17 |
2700.7200000000003 | 26250.0 |  male  |
| 39975.0 | 2498.821610006254  | 4187.396113572586 | 5.23 | 130.75 | 600.17 |
2700.7200000000003 | 26250.0 | female |
+--------+-------------------+------------------+------+--------+--------+---
----------------+---------+--------+
```



Z histogramu a aj z tabuľky to zatiaľ vyzerá, že by sme hypotézu mohli ľahko vyvrátiť. Zároveň však histogram vyzerá, že dáta s ktorými pracujem nie sú normálne distribuované. Toto ešte overím pomocou knižnice :

$ H_0 $ Dáta sú normálne distribuované $ H_1 $ Dáta nie sú normálne distribuované $ = 0.05 $

[ ]:
```
_, p = stats.normaltest(data["total"])
print("p value is : ", p)
```

```
p value is :  0.0
```

$ p $ hodnota je oveľa menšia ako $ $ takže nulovú hypotézu môžem zamietnuť. Dáta teda nie sú normálne rozdelené a tak budem musieť použiť neparametrický test.

Aby som zistil, či dáta pochádzajú z rovnakého rozdelenia, vykonám U-test. Pre oba testy mi budú platiť nasledovné hypotézy :

$ H_0 $: Muži a ženy utrácajú v obchode rovnako $ H_1 $: Ženy utrácajú v obchode viac $ = 0.05 $

1.2.1 U-test

[ ]:
```
_, p_value = stats.mannwhitneyu(males["total"], females["total"])
print("p_value: ", p_value)
```

```
p_value:  0.6378679368451287
```

$ p $ hodnota je podľa očakávania vyššia ako $ $ a tak nulovú hypotézu zamietnuť nemôžem.

1.2.2 Median test

Podľa očakávania je $ p $ hodnota vyššia ako požadovaná $ $ hodnota a tak nemôžem zamietnuť nulovú hypotézu. Týmto som zistil, že dáta útrat mužov a žien pravdepodobne pochádzajú z rovnakej distribúcie. Skúsim vykonať ešte mediánový test, ktorým zistím, či pochádzajú z distribúcií s rovnakým mediánom.

[ ]:
```
result = stats.median_test(males["total"], females["total"])
print("p value is : ", result.pvalue)
```

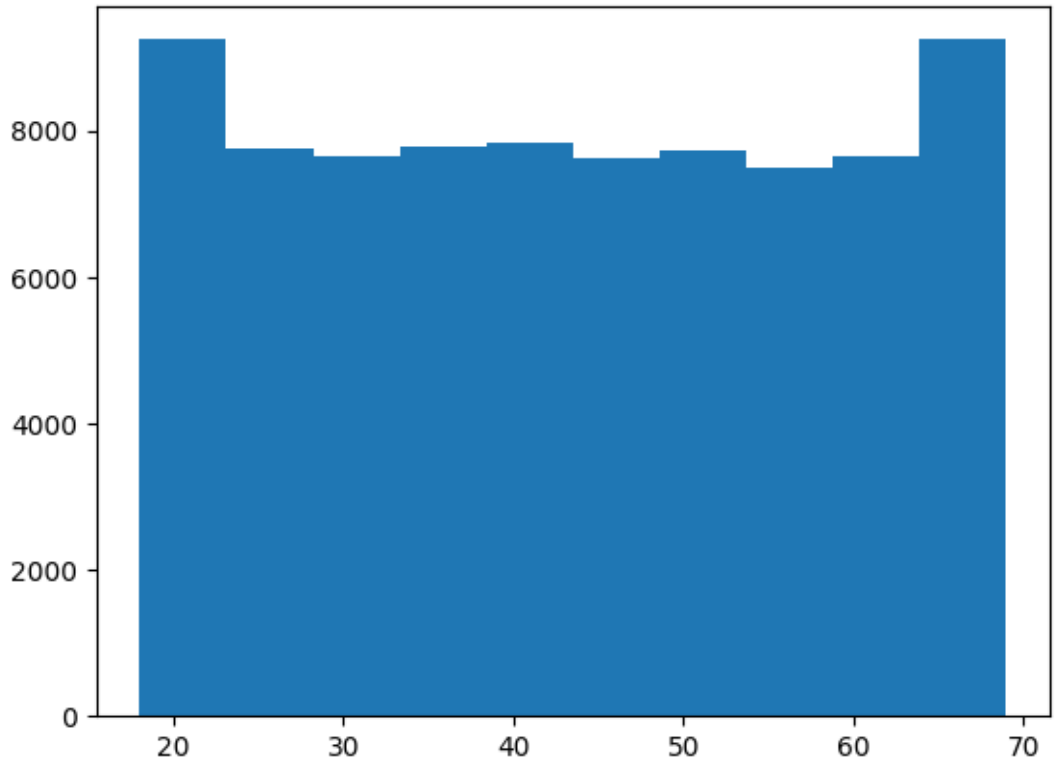
```
p value is :  0.7182381769504835
```

Aj podľa tohto testu nemôžeme zamietnuť nulovú hypotézu a teda môžem vyvodiť záver, že muži aj ženy utrácajú v obchode rovnako.

2 Mladí ľudia nakupujú kartou viac ako starší

V tejto hypotéze skúsim overiť, či existuje závislosť medzi medzi vekom a spôsobom platby. Konkrétne sa domnievam, že mladí ľudia používajú kartu viac ako starší ľudia. Vek mám v dátach zastúpený celkom uniformne a tak ďalej nijak upravovať nebudem

[ ]:
```
plt.hist(data["age"])
plt.show()
```

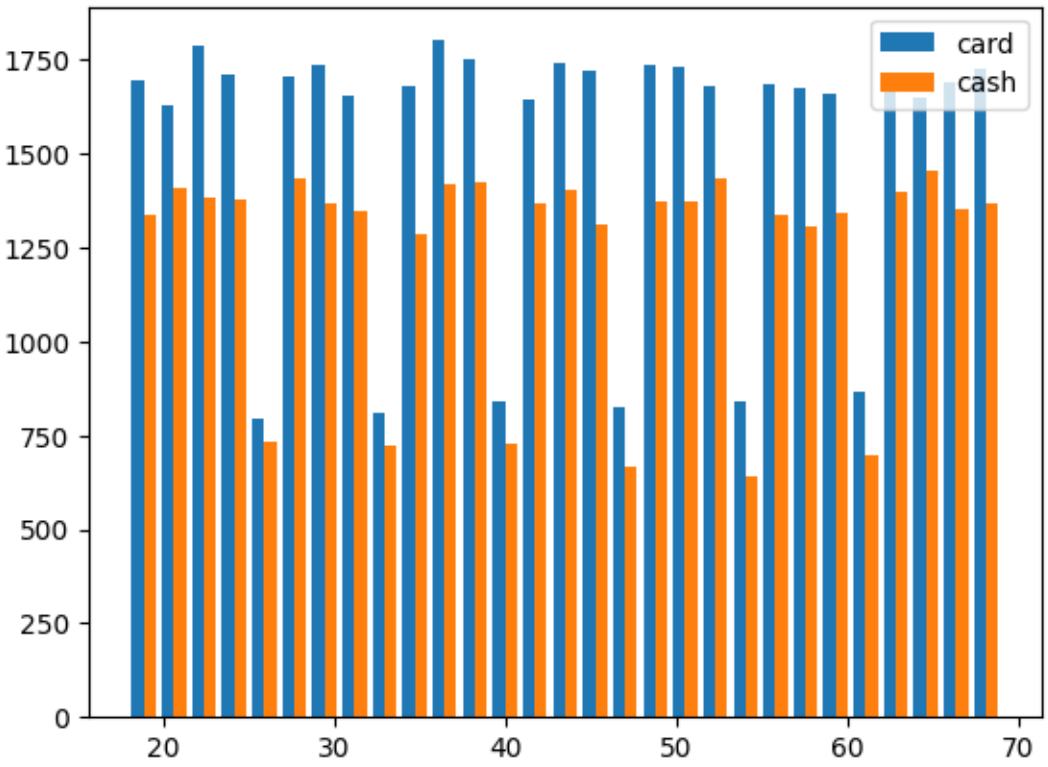


[ ]:
```
card = data[data["payment_method"] == "Card"]
cash = data[data["payment_method"] == "Cash"]

# Table
table = PrettyTable(); table.field_names = card["age"].describe().index
table.add_row(card["age"].describe().values)
table.add_row(cash["age"].describe().values)
table.add_column("payment_method", ["card", "cash"])
print(table)

# Graph
bins = np.linspace(data["age"].min(), data["age"].max(), 30)
plt.hist([card["age"], cash["age"]], bins, label=["card", "cash"])
plt.legend(loc='upper right')
```

```
+---------+------------------+-------------------+------+------+------+------
+------+----------------+
|  count  |        mean      |        std        | min  | 25%  | 50%  | 75%  |
max  | payment_method |
+---------+------------------+-------------------+------+------+------+------
+------+----------------+
| 44144.0 | 43.46830826386372 | 14.982816172093813 | 18.0 | 30.0 | 43.0 | 56.0 |
```

```
69.0 |      card      |
| 35806.0 | 43.43869742501257 | 15.022222463327042 | 18.0 | 30.0 | 43.0 | 56.0 |
69.0 |      cash      |
+---------+-------------------+--------------------+------+------+------+------
+------+----------------+
```

[ ]: <matplotlib.legend.Legend at 0x7f23ece42320>



Z dát je jasne vidno, že platba kartou je populárnejšia ako platba v hotovosti, avšak okrem toho vyzerjú byť dáta rozdelené takmer identicky. Histogram vyzerá tak, že by sme aj túto domnienku mali zamietnuť, pretože pre každú vekovú skupinu je platba kartou nepopulárna o takmer rovnakú hodnotu.

2.0.1 Chi-kvadrát

Aby som zistil, či sú dáta na sebe skutočne závislé použijem chi-kvadrát test nezávislosti. Ten porovnáva na základe početnosti a tak si dáta najskôr kategorizujem do skupín.

Nulová hypotéza $ H_0 $: Neexistuje závislosť medzi vekom a spôsobom platby (mladí nakupujú kartou rovnako veľa ako starší) $ H_1 $: Existuje závislosť medzi vekom a spôsobom platby $ = 5% $

[ ]:
```
from ages import categorize_ages
from statistics import median

categories = ["young", "young middle", "old middle" ,"old"]
interval = np.linspace(data["age"].min(), data["age"].max(), len(categories) +␣
↪1)
categorized_data = categorize_ages(interval, data)
```

[ ]:
```
# age_counts = new_data["age"].value_counts()
# payment_counts = new_data["payment_method"].value_counts()
# print(age_counts)
# print(payment_counts)
```

[ ]:
```
crosstab = pd.crosstab(categorized_data["payment_method"],␣
↪categorized_data["age"])
crosstab
```

[ ]:
```
age                18     31     44     57
payment_method
Card            11057  11109  11036  10942
Cash             9044   9018   8823   8921
```

[ ]:
```
output = stats.chi2_contingency(crosstab)
output
```

[ ]:
```
Chi2ContingencyResult(statistic=1.5075424056579085, pvalue=0.680530644338637,
dof=3, expected_freq=array([[11098.66846779, 11113.02424015, 10965.04935585,
10967.25793621],
       [ 9002.33153221,  9013.97575985,  8893.95064415,  8895.74206379]]))
```