Exercise Sheet 2 – Probability and Statistics

08/09/2021

Exercise 1

We roll a fair dice three times. What is the probability that two even numbers won't appear in consecutive rolls?

ALl possible combinations: _ _ _ = 6 6 6 = 216 Even 2 4 6 Events possible OEO, EOE = 18

Probability = $\frac{18}{216} = \frac{1}{12} = \mathbf{0.083}$

Exercise 2

All possible combinations for 4 babys: _ _ _ _ = 2 2 2 2 = 2^4 = 16 We can have:

Posibilities

- GGBB
- BBGG
- BGGB
- GBBG
- GBGB
- BGBG

Thus: Probability: $\frac{6}{16} = \frac{3}{8} = \mathbf{0.375}$

Exercise 3

In a family with 3 children, of whom at least one is a girl, what is the probability that the family has at least 2 daughters?

All possible combinations when boy is defined B _ _ = B 2 2 = 2^2 = 4 possibilities

Posibilities:

- BB
- GG
- BG
- GB Thus: Probability = $\frac{3}{4} = \mathbf{0.75}$

Exercise 4

Exercise 4 If A, B are events in sample space S such that P(A) = P(B|A) = 0.3 and P(A|B) = 0.6, what is the value of P(B)?

P(A|B) = P(B|A) · P(A) / P(B)

0.6 = 0.3 · 0.3 / P(B)

0.6 = 0.09 / P(B)

P(B) = 0.09/0.6

P(B) = **0.15**

Exercise 5

Almost 10% of all root canal therapies fail. Let X denote the probability variable of number of fails in 4 attempts at root canal.

p = Probability that root canal fails = 0.1

n = numbers of attempts = 4

Attempts are independent so we can use binominal distribution.

1. Write the probability distribution function of the variable X

P(X) = $\frac{n!}{x!(n-x)}p^x\ (1-p)^{n-x}$

P(X) = $\frac{4!}{x!(4-x)}0.1^x\ (1-0.1)^{4-x}$

2. What is the probability that at least one attempt is successful?
This probability that at zero attempt of root canal fails.

P(X = 0) = $\frac{4!}{0!(4-0)}0.1^0\ (1-0.1)^{4-0} = 0.9^4 = 0.65$

P (X = 0) = 0.65

3. How many successful therapies are expected in 20 attempts?

Succesful terapies in 20 attempts = n · (1-p) = 20 · 0.9 = 18

Exercise 6

If we have three identical boxes such that the first box contains three white pins and five black pins, the second box contains four white and four black pins, and the third box contains six white and two black pins. If we take two pins of the opposite color from a selected box, what is the probability that the second box is chosen?

We have boxes:

- B1 = {3W, 5B}
- B2 = {4W, 4B}
- B3 = {6W, 2B}

P(B1) = P(B2) = P(B3)

Let E1, E2 and E3 be the event that a ball is chosen from a box B1, B2 and B3 respectively.

Case white ball EW

P(E1) = P(E2) = P(E3) = 1/3

P(EW|E1) = $\frac{3}{8}$, P(EW|E2) = $\frac{4}{8}$, P(EW|E3) = $\frac{6}{8}$

Case black ball EB

P(E1) = P(E2) = P(E3) = $\frac{1}{3}$

P(EB|E1) = $\frac{5}{8}$, P(EB|E2) =$\frac{4}{8}$, P(EB|E3) = $\frac{2}{8}$

Now by Bayes‘ Theorem for white

P(E2/EW) = $\frac{P(E2) \cdot P(EW|E2)}{P(E1) \cdot P(EW|E1) + P(E2) \cdot P(EW|E2) + P(E3) \cdot P(EW|E3)}$

P(E2/EW) = $\frac{(\frac{1}{3} \cdot \frac{4}{8})}{(\frac{1}{3} \cdot \frac{3}{8}) + (\frac{1}{3} \cdot \frac{4}{8}) + (\frac{1}{3} \cdot \frac{6}{8})}$ = 0.19

Now by Bayes‘ Theorem for black

P(E2/EB) = $\frac{P(E2) \cdot P(EB|E2)}{P(E1) \cdot P(EB|E1) + P(E2) \cdot P(EB|E2) + P(E3) \cdot P(EB|E3)}$

P(E2/EB) = $\frac{(\frac{1}{3} \cdot \frac{4}{8})}{(\frac{1}{3} \cdot \frac{5}{8}) + (\frac{1}{3} \cdot \frac{4}{8}) + (\frac{1}{3} \cdot \frac{2}{8})}$ = 0.21

To get to know if two pins of the opposite color from a selected boxwe have to multiply probabilities thus: P(EW|E2 & EB|E2) = 0.19 * 0.21 = **0.039**

Exercise 7

Theory

1. The probability P of event E is described as P(E) = P(E) $\geq$ 0; where
2. P($\emptyset$) = 1;

We start with:

- A $\cup$ B = (B A) $\cup$ A
- (B A) $\cap$ A = $\emptyset$
- P(A $\cup$ B) = P(B A) + P(A)
- A $\cup$ B = B then P(B) = P(B A) + P(A)
- From theory 1 = P(B A) $\geq$ 0, and we can conclude that P(B) $\geq$ P(A)

Exercise 8

Mean = $\frac{45+70+94+78+61+18+19+34+48+73+56+46+32+95+47+39+96+58+1+23+30+50+21+47+80+23+38+33+5+39+}{30}$ = 46.633333

Sorted sample from probability distribution: [1, 5, 18, 19, 21, 23, 23, 30, 32, 33, 34, 38, 39, 39, 45, 46, 47, 47, 48, 50, 56, 58, 61, 70, 73, 78, 80, 94, 95, 96]

- **Median** = 45.5
- **Mode** = 47
- **75% quantile** = 60.250000
- **Variance** = 655.1367816091955
- **Standard deviation** = 25.595639894505382
- **95% confidence interval** = (37.07576432247458, 56.19090234419208)
- **90% confidence interval** = (38.69313753342548, 54.57352913324119)

Exercise 8

Theory

1. The probability P of event E is described as $P(E) = P(E) \geq 0$; where
2. $P(\emptyset) = 1$;

We start with:

- $A \cup B = (B \;\; A) \cup A$
- $(B \;\; A) \cap A = \emptyset$
- $P(A \cup B) = P(B \;\; A) + P(A)$
- $A \cup B = B$ then $P(B) = P(B \;\; A) + P(A)$
- From theory $1 = P(B \;\; A) \geq 0$, and we can conclude that $P(B) \geq P(A)$

Exercise 9

```python
import numpy as np
import matplotlib.pyplot as plt  # To visualize
import pandas as pd  # To read data
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

```python
df = pd.read_csv ('Real estate.csv')
```

```python
#print (df)

df.describe()

#Not that not all columns give meaning to summary

#summary table

summary = []

columns = df.columns.values.tolist()

column_names = ["Column", "Mean", "Median", "Quantile 75", "Variance", "SD"]
summary = pd.DataFrame(columns = column_names)
summary['Column'] = columns

for x in range(df.shape[1]):
    data = (df.iloc[:, [x]])
    summary['Mean'][x] = (data.mean()[0])
    summary['Median'][x] = (data.median()[0])
    summary['Quantile 75'][x]= (data.quantile(.75)[0])
    summary['Variance'][x] = (data.var()[0])
    summary['SD'][x] = (data.std()[0])
```

```python
    print(summary)
```

```
                                   Column     Mean   Median Quantile 75  \
0                                      No    207.5    207.5      310.75
1                     X1 transaction date  2013.15  2013.17     2013.42
2                            X2 house age  17.7126     16.1       28.15
3  X3 distance to the nearest MRT station  1083.89  492.231     1454.28
```

```
4        X4 number of convenience stores   4.0942        4         6
5                            X5 latitude   24.969  24.9711   24.9775
6                           X6 longitude  121.533  121.539   121.543
7              Y house price of unit area  37.9802    38.45      46.6

      Variance         SD
0      14317.5    119.656
1    0.0795055   0.281967
2      129.789    11.3925
3  1.59292e+06    1262.11
4      8.67633    2.94556
5  0.000154013  0.0124102
6  0.000235536  0.0153472
7      185.137    13.6065
```

```
##2
df.isnull().values.any()
X = df.drop('Y house price of unit area',axis=1)
y = df['Y house price of unit area']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                    random_state=101)

#test size represent the proportion of the dataset to include in the test split.
#random_state Controls the shuffling applied to the data before applying the split -> simply sets seed
#without specified the random_state in the code, then every time code run(execute) a new random value
#is generated and the train and test datasets
#would have different values each time.

model= LinearRegression()
model.fit(X_train, y_train)
LinearRegression()
pd.DataFrame(model.coef_, X.columns, columns=['Coeficient'])
y_pred=model.predict(X_test)


houseToPredict = pd.DataFrame()
prediction = {'No': 1, 'X1 transaction date': 2014, 'X2 house age': 10,
              'X3 distance to the nearest MRRT station' : 1200, 'X4 number of convenience stores': 5,
              'X5 latitude': 24.93,
              'X6 longitude': 121.54}

prediction = {'No': 1, 'X1 transaction date': 2014, 'X2 house age': 10,
              'X3 distance to the nearest MRRT station' : 1200, 'X4 number of convenience stores': 5,
              'X5 latitude': 24.93,
              'X6 longitude': 121.54}




houseToPredict = houseToPredict.append(prediction, ignore_index = True)
predictedPrice=model.predict(houseToPredict)
```

```
print("Predicted house price (with given criteria) is ", predictedPrice)
```

```
Predicted house price (with given criteria) is  [36.42730457]
```

Exercise 10

By changing the sample size we modify the shape of histogram. For large values of sample size, the distributions of the count X and the sample proportion are approximately normal. This follows from the **Central Limit Theorem.**