Edit View Run Kernel Settings Help

Markdown

Jupyte

Inferential Statistics

Key Terms:

- **Random Experiment**: An action/experiment that leads to one of several possible outcomes. Ex: Tossing a coin, rolling a dice
- **Trial**: Each time when a random experiment is conducted.
- **Sample space**: A list of all possible outcomes of an experiment.
- **Event** : An event is a possible outcome of a random experiment. It is a subset of sample space.
- **Random Variable**: A variable which takes one of the values from the outcome of an event defined based on the random experiment.
- **Probability**: How likely is that an event can occur.

Let A be an event.

- $P(A) = \frac{FavorableNumberOfOutcomes}{TotalNumberOfOutcomes}$
- $0 <= P(A) <= 1$

Examples:

1. Consider a random experiment of tossing an unbiased coin.
   - Then the sample space is $S = \{H, T\}$.
   - Let A be an event of getting a Head.
   - Then the random variable, say X, takes two possible values 0 and 1, each with probability 1/2.

Search fractal

Then the random variable, say X, takes two possible values 0 and 1, each with probability 1/2.

2. Consider a random experiment of tossing two unbiased coins simultaneously.
   - Then the sample space is $S = \{HH, HT, TH, TT\}$.
   - Let A be an even of getting at least one Tail.
   - Then the random variable, say X, takes the values 0 (HH), 1 (HT or TH) and 2 (TT). Probabilities will be 1/4, 2/4 and 1/4 respectively

Question:

- Consider a random experiment of rolling two dice at a time. What is the probability that sum of the values is 8?

Solution:

- Sample space for rolling two dice at a time will contain 36 combinations.
- Sample space for the event of getting sum of the values as 8: {(2,6), (3, 5), (4, 4), (5,3), (6,2)}
- Hence the probability is $\frac{5}{36}$

Few rules of Probability

- The probability of complementary event A' of A is given by $P(A') = 1 - P(A)$
- If A and B are any two events, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- When two events are mutually exclusive/disjoint then $P(A \cap B) = 0$
- **Multiplication theorem:** If A and B are independent events then $P(A \cap B) = P(A)P(B)$

- **Joint Probability:**
  - Joint probability is a statistical measure that calculates the likelihood of two events occurring together and at the same point in time.
  - It can only be applied to situations where more than one observation can occur at the same time.
  - For example, from a deck of 52 cards, the joint probability of picking up a card that is both red and 6 is P(6 ∩ red) = 2/52 = 1/26, since a deck of cards has two red sixes—the six of hearts and the six of diamonds. Because the events "6" and "red" are independent in this example, you can also use the following formula to calculate the joint probability:

    $$P(6 \cap red) = P(6) \times P(red) = \frac{4}{52}\frac{26}{52} = \frac{1}{26}$$

- **Marginal Probability:**
  - It is the probability of an event irrespective of the outcome of another variable.
  - The probability of an event for one random variable, irrespective of the outcome of another random variable.
  - For example, the probability of X=A for all outcomes of Y.

    $$P(X = A) = \sum P(X = A, Y = y_i) \forall i$$

- **Conditional Probability:**
  - The probability of event A given that event B has already occured is $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Examples:

1. Consider a random experiment of rolling a die. What is the probability that getting an even number or a number divisible by 3?

Solution:

- When a die is rolled, the sample space is {1, 2, 3, 4, 5, 6}
- Let A be an event of getting an even number. Then favourable outcomes are {2, 4, 6}.
- Let B be an event of getting a number divisible by 3. Then favourable outcomes are {3, 6}
- Events A and B are not mutually exclusive.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}$

2. Consider a random experiment of rolling a die. What is the probability that getting a en even number and a number divisible by 3?

Solution:

- When a die is rolled, the sample space is {1, 2, 3, 4, 5, 6}
- Let A be an event of getting an even number. Then favourable outcomes are {2, 4, 6}.
- Let B be an event of getting a number divisible by 3. Then favourable outcomes are {3, 6}
- $P(A \cap B) = P(A)P(B) = \frac{3}{6} \cdot \frac{2}{6} = \frac{1}{6}$

3. Consider a random experiment of rolling a die. Let A be the event that shows an outcome is an odd number and suppose B the event that shows the outcome is les than or equal to 3. Then what is the probability A given B, P(A|B)?

Solution:

- When a die is rolled, the sample space is {1, 2, 3, 4, 5, 6}
- Let A be an event of getting an odd number. Then favourable outcomes are {1, 3, 5}.
- Let B be an event of getting a number ≤ 3. Then favourable outcomes are {1,2,3}
- Favourable outcomes for both A and B are {1, 3}
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$

4. Consider the table given below showing probabilities of people owning pets. What is the probability a randomly selected person is male, given that they own a pet?

	Have pets	Do not have pets	Total
Male	0.41	0.08	0.49
Female	0.45	0.06	0.51
Total	0.86	0.14	1

Solution:

- Let M is for male and PO stands for pet owner, so the formula becomes: $P(M|PO) = P(M \cap PO)/P(PO)$
- From the table, it is obvious that $P(M \cap PO) = 0.41$ and $P(PO) = 0.86$
- So, $P(M|PO) = \frac{0.41}{0.86} = 0.4777$

Creating

- Frequency Tables
- Two-way Tables
- Two-way Table with joint Probability
- Two-way Table with marginal Probability
- Two-way Table with conditional Probability
- Correlation

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
cars_data = pd.read_csv('UsedCarsPrice.csv',index_col=0, na_values=["??", "????"])
cars_data.head(10)
```

[2]:

	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight
0	13500	23.0	46986.0	Diesel	90.0	1.0	0	2000	three	1165
1	13750	23.0	72937.0	Diesel	90.0	1.0	0	2000	3	1165
2	13950	24.0	41711.0	Diesel	90.0	NaN	0	2000	3	1165
3	14950	26.0	48000.0	Diesel	90.0	0.0	0	2000	3	1165
4	13750	30.0	38500.0	Diesel	90.0	0.0	0	2000	3	1170
5	12950	32.0	61000.0	Diesel	90.0	0.0	0	2000	3	1170

ZIM - IND

File Edit View Run Kernel Settings Help

Markdown

[3]:
```
cars_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1436 entries, 0 to 1435
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Price      1436 non-null   int64
 1   Age        1336 non-null   float64
 2   KM         1421 non-null   float64
 3   FuelType   1336 non-null   object
 4   HP         1430 non-null   float64
 5   MetColor   1286 non-null   float64
 6   Automatic  1436 non-null   int64
 7   CC         1436 non-null   int64
 8   Doors      1436 non-null   object
 9   Weight     1436 non-null   int64
dtypes: float64(4), int64(4), object(2)
memory usage: 123.4+ KB
```

[4]:
```
cars_data.shape
```

[4]:
```
(1436, 10)
```

```
cars_data.isna().sum()
```

```
Price      0
Age      100
KM        15
```

ZIM - IND
Live

Scanned with CamScanner

```
[6]: cars_data.Doors.value_counts()
```

```
[6]: 5        673
     3        621
     4        137
     2          2
     three      1
     four       1
     five       1
     Name: Doors, dtype: int64
```

Frequency Table using crosstab()

- To compute a simple cross tabulation of one or more factors
- By default, computes a frequency table of factors

```
[7]: # Number of cars having various fuel types
     pd.crosstab(index=cars_data['FuelType'], columns = 'count', dropna=True)
```

[7]:

col_0	count
FuelType	
CNG	15
Diesel	144
Petrol	1177

Two way tables

- To look at the frequency distribution of gearbox types with respect to different fuel types of the cars
- 0 indicates Manual gear box, 1 indicates automatic gearbox

[8]:
```
#Two-way table of fuel type vs gear type (automatic or manual)
pd.crosstab(index=cars_data['Automatic'],
            columns = cars_data['FuelType'], dropna=True)
```

[8]:

FuelType	CNG	Diesel	Petrol
Automatic			
0	15	144	1104
1	0	0	73

Two Way table with joint probability ¶

- Joint probability is the likelihood of two independent events happening at the same time.
- Convert the table values from numbers to proportions to get joint probability table

[14]:
```
#joint probability
pd.crosstab(index=cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = True,
```

[14]:
```
#joint probability
pd.crosstab(index=cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = True,

            dropna=True)
```

[14]:

FuelType	CNG	Diesel	Petrol
Automatic			
0	0.011228	0.107784	0.826347
1	0.000000	0.000000	0.054641

Interpretation:

- Probability of a car with manual gearbox and fuel type CNG is 0.011228
- Probability a petrol car with automatic gearbox is 0.054641

Questions:

1. Assume that you have picked one car randomly from the dataset. What is the probability that it is a diesel car? (marginal probability)
2. What is the probability of a car is of automatic gear?
3. What is the prob of a petrol car is of manual gear?

Search

fractal

2. What is the probability of a car is of automatic gear?

3. What is the prob of a petrol car is of manual gear?

Answers:

1. Marginal probabaility of diseal car irrespective of automatic or not - 0.107784

[15]:
```
# alternative code:
cars_data['FuelType'].value_counts(normalize=True)
```

[15]:
```
Petrol    0.880988
Diesel    0.107784
CNG       0.011228
Name: FuelType, dtype: float64
```

[ ]:

Two Way table with Marginal probability

- Marginal probability is the probability of the occurrence of the single event.
- Gives the row-sums and column-sums on Joint probability table

File Edit View Run Kernel Settings Help

Markdown

[16]:
```
#row sum and col sum joint probability


pd.crosstab(index=cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = True,
            margins= True,
            dropna=True)
```

[16]:

FuelType	CNG	Diesel	Petrol	All
Automatic				
0	0.011228	0.107784	0.826347	0.945359
1	0.000000	0.000000	0.054641	0.054641
All	0.011228	0.107784	0.880988	1.000000

Interpretation:

- Probability of cars having manual gear box when the fuel type are CNG or Diesel or Petrol is 0.945
- Probability of cars having Fuel Type as Petrol, whether it is automatic or manual, is 0.88 (or, 88% of the cars are petrol cars)

Two Way table with Conditional probability

- Conditional probability is the probability of an event (A), given that another event (B) has already occurred.

Question:

- Given the type of gear box, compute the probability of different fuel type. For example, what is the probability that a car is a Petrol car, given that it is of automatic gea box?
- Set *normalize* argument as *index*

```
#normalize
pd.crosstab(index=cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = 'index',
            margins= True,
            dropna=True)
```

FuelType	CNG	Diesel	Petrol
Automatic			
0	0.011876	0.114014	0.874109
1	0.000000	0.000000	1.000000
All	0.011228	0.107784	0.880988

Interpretation:

- Observe that the row-sum is 1.
- Given that the gearbox type is manual,
  - the probability of the car being CNG Fuel Type is 0.011876
  - the probability of the car being Petrol Fuel Type is 0.874109

- Given that the gearbox type is manual,
  - the probability of the car being CNG Fuel Type is 0.011876
  - the probability of the car being Petrol Fuel Type is 0.874109
- Given that the gearbox type is automatic,
  - the probability of the car being Petrol Fuel Type is 1

Question:

- Given the FuelType, compute the probability of different gear box types.
- Set *normalize* argument as *columns*

[18]:
```
pd.crosstab(index=cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = 'columns',
            margins= True,
            dropna=True)
```

[18]:

FuelType	CNG	Diesel	Petrol	All
Automatic				
0	1.0	1.0	0.937978	0.945359
1	0.0	0.0	0.062022	0.054641

Interpretation:

- Observe that the column-sum is 1.
- Given that the Fuel type is CNG,

1	0.0	0.0	0.062022	0.054641

Interpretation:

- Observe that the column-sum is 1.
- Given that the Fuel type is CNG,
  - the probability of the car being automatic is 0
  - the probability of the car is of manual gear is 1
- Given that the Fuel type is Petrol,
  - the probability of the car being automatic is 0.054641

]:

Exercises:

1. Consider a random experiment of tossing 3 unbiased coins.
   - What is the probability of getting 3 heads?
   - If given that an event that shows the first toss was heads, then what is the probability of three heads?
2. Consider the previous dataset of UsedCarPrice. Analyze the following:
   - How many metalic cars are of automatic gear box?
   - How many cars are with manual gear box and of non-metalic color body?
   - What is the probability that the car is of metalic colored body, given that it has automatic gear box?