simpsons_paradox

November 20, 2017

1 Simpson's Paradox

Use `admission_data.csv` for this exercise.

```
In [1]: # Load and view first few lines of dataset
        import pandas as pd

        df = pd.read_csv('admission_data.csv')
        df.head()
```

Out[1]:

	student_id	gender	major	admitted
0	35377	female	Chemistry	False
1	56105	male	Physics	True
2	31441	female	Chemistry	False
3	51765	male	Physics	True
4	53714	female	Physics	True

```
In [19]: num_students = df.shape[0]
         num_students
```

```
Out[19]: 500
```

1.0.1 Proportion and admission rate for each gender

```
In [13]: # Proportion of students that are female
         df_f = df[df["gender"] == "female"]
         df_f.shape[0]/num_students
```

```
Out[13]: 0.514
```

```
In [14]: # Proportion of students that are male
         df_m = df[df["gender"] == "male"]
         df_m.shape[0]/num_students
```

```
Out[14]: 0.486
```

```
In [16]: num_males = df_m.shape[0]
         num_males
```

```
Out[16]: 243

In [17]: num_females = df_f.shape[0]
         num_females

Out[17]: 257

In [27]: # Admission rate for females
         num_females_admit = df_f[df_f["admitted"] == True].shape[0]
         num_females_admit/num_females

Out[27]: 0.28793774319066145

In [28]: # Admission rate for males
         num_males_admit = df_m[df_m["admitted"] == True].shape[0]
         num_males_admit/num_males

Out[28]: 0.48559670781893005
```

1.0.2 Proportion and admission rate for physics majors of each gender

```
In [44]: df_female_phys = df_f[df_f["major"] == "Physics"]
         df_female_phys.head()

Out[44]:     student_id  gender    major  admitted
         4         53714  female  Physics      True
         47        54442  female  Physics      True
         59        27446  female  Physics      True
         66        29216  female  Physics     False
         85        30213  female  Physics     False

In [45]: df_male_phys = df_m[df_m["major"] == "Physics"]
         df_male_phys.head()

Out[45]:    student_id gender    major  admitted
         1        56105   male  Physics      True
         3        51765   male  Physics      True
         6        25946   male  Physics      True
         8        55247   male  Physics     False
         9        35838   male  Physics      True

In [46]: num_female_phys = df_female_phys.shape[0]
         num_female_phys

Out[46]: 31

In [47]: num_male_phys = df_male_phys.shape[0]
         num_male_phys

Out[47]: 225
```

```
In [54]: # What proportion of female students are majoring in physics?
         num_female_phys / num_females
```

```
Out[54]: 0.12062256809338522
```

```
In [55]: # What proportion of male students are majoring in physics?
         num_male_phys / num_males
```

```
Out[55]: 0.9259259259259259
```

```
In [52]: # Admission rate for female physics majors
         df_female_phys_admit = df_female_phys[df_female_phys["admitted"] == True]
         num_female_phys_admit = df_female_phys_admit.shape[0]
         num_female_phys_admit / num_female_phys
```

```
Out[52]: 0.7419354838709677
```

```
In [53]: # Admission rate for male physics majors
         df_male_phys_admit = df_male_phys[df_male_phys["admitted"] == True]
         num_male_phys_admit = df_male_phys_admit.shape[0]
         num_male_phys_admit / num_male_phys
```

```
Out[53]: 0.5155555555555555
```

1.0.3 Proportion and admission rate for chemistry majors of each gender

```
In [56]: df_female_chem = df_f[df_f["major"] == "Chemistry"]
         df_female_chem.head()
```

```
Out[56]:     student_id  gender      major  admitted
         0         35377  female  Chemistry     False
         2         31441  female  Chemistry     False
         5         50693  female  Chemistry     False
         7         27648  female  Chemistry      True
         11        53708  female  Chemistry      True
```

```
In [57]: df_male_chem = df_m[df_m["major"] == "Chemistry"]
         df_male_chem.head()
```

```
Out[57]:      student_id gender      major  admitted
         22         42508   male  Chemistry     False
         84         35357   male  Chemistry     False
         109        41460   male  Chemistry     False
         129        47442   male  Chemistry     False
         199        27315   male  Chemistry     False
```

```
In [58]: num_female_chem = df_female_chem.shape[0]
         num_female_chem
```

```
Out[58]: 226
```

In [59]:
```
num_male_chem = df_male_chem.shape[0]
num_male_chem
```

Out[59]: 18

In [60]:
```
# What proportion of female students are majoring in chemistry?
num_female_chem / num_females
```

Out[60]: 0.8793774319066148

In [61]:
```
# What proportion of male students are majoring in chemistry?
num_male_chem / num_males
```

Out[61]: 0.07407407407407407

In [63]:
```
# Admission rate for female chemistry majors
df_female_chem_admit = df_female_chem[df_female_chem["admitted"] == True]
num_female_chem_admit = df_female_chem_admit.shape[0]
num_female_chem_admit / num_female_chem
```

Out[63]: 0.22566371681415928

In [64]:
```
# Admission rate for male chemistry majors
df_male_chem_admit = df_male_chem[df_male_chem["admitted"] == True]
num_male_chem_admit = df_male_chem_admit.shape[0]
num_male_chem_admit / num_male_chem
```

Out[64]: 0.1111111111111111

1.0.4 Admission rate for each major

In [65]:
```
# Admission rate for physics majors
(num_female_phys_admit + num_male_phys_admit) / (num_female_phys + num_male_phys)
```

Out[65]: 0.54296875

In [66]:
```
# Admission rate for chemistry majors
(num_female_chem_admit + num_male_chem_admit) / (num_female_chem + num_male_chem)
```

Out[66]: 0.21721311475409835

In [ ]: