Calculating Errors

Here are two datasets that represent two of the examples you have seen in this lesson.

One dataset is based on the parachute example, and the second is based on the judicial example. Neither of these datasets is based on real people.

Use the exercises below to assist in answering the quiz questions at the bottom of this page.

In [1]:
```
import numpy as np
import pandas as pd

jud_data = pd.read_csv('judicial_dataset_predictions.csv')
par_data = pd.read_csv('parachute_dataset.csv')
```

In [2]:
```
jud_data.head()
```

Out[2]:

	defendant_id	actual	predicted
0	22574	innocent	innocent
1	35637	innocent	innocent
2	39919	innocent	innocent
3	29610	guilty	guilty
4	38273	innocent	innocent

In [3]:
```
par_data.head()
```

Out[3]:

	parachute_id	actual	predicted
0	3956	opens	opens
1	2147	opens	opens
2	2024	opens	opens
3	8325	opens	opens
4	6598	opens	opens

1. Above, you can see the actual and predicted columns for each of the datasets. Using the **jud_data**, find the proportion of errors for the dataset, and furthermore, the percentage of errors of each type. Use the results to answer the questions in quiz 1 below.

In [4]:
```
jud_data[jud_data['actual'] != jud_data['predicted']].shape[0]/jud_data.shape[0] # Number of errors
```

Out[4]:
```
0.042152958945489497
```

In [5]:
```
jud_data.query("actual == 'innocent' and predicted == 'guilty'").count()
[0]/jud_data.shape[0] # Type 1 errors
```

Out[5]: 0.001510366607167376

In [6]:
```
jud_data.query("actual == 'guilty' and predicted == 'innocent'").count()
[0]/jud_data.shape[0] # Type 2 errors
```

Out[6]: 0.040642592338322119

In [7]:
```
# If everyone was predicted to be guilty, then every actual innocent
# person would be a type I error.

# Type I = pred guilty, but actual = innocent
jud_data[jud_data['actual'] == 'innocent'].shape[0]/jud_data.shape[0]
```

Out[7]: 0.45159961554304545

In [8]:
```
#If everyone has prediction of guilty, then no one is predicted inncoent
#Therefore, there would be no type 2 errors in this case

# Type II errs = pred innocent, but actual = guilty
0
```

Out[8]: 0

2. Above, you can see the actual and predicted columns for each of the datasets. Using the **par_data**, find the proportion of errors for the dataset, and furthermore, the percentage of errors of each type. Use the results to answer the questions in quiz 2 below.

In [9]:
```
par_data[par_data['actual'] != par_data['predicted']].shape[0]/par_data.
shape[0] # Number of errors
```

Out[9]: 0.039972551037913875

In [10]:
```
par_data.query("actual == 'fails' and predicted == 'opens'").count()[0]/
par_data.shape[0] # Type 1 errors
```

Out[10]: 0.00017155601303825698

In [11]:
```
par_data.query("actual == 'opens' and predicted == 'fails'").count()[0]/
par_data.shape[0] # Type 2 errors
```

Out[11]: 0.039800995024875621

In [12]:
```
# If every parachute is predicted to fail, what is the proportion
# of type I errors made?

# Type I = pred open, but actual = fail
# In the above situation since we have none predicted to open,
# we have no type I errors

0
```

Out[12]: 0

In [13]:
```
# If every parachute is predicted to fail, what is
# the proportion of Type II Errors made?

# This would just be the total of actual opens in the dataset,
# as we would label these all as fails, but actually they open

# Type II = pred fail, but actual = open
par_data[par_data['actual'] == 'opens'].shape[0]/par_data.shape[0]
```

Out[13]: 0.9917653113741637

In [ ]: