



Estatística Aplicada

Amostragem Estratificada

Prof. Me. Max Gabriel Steiner

AMOSTRAGEM ESTRATIFICADA

- ✓ *População: 100 pessoas*
- ✓ *60 homens e 40 mulheres*
- ✓ *Amostra de 10% da população: 10 pessoas*
- ✓ *Homens representam 60% da população*
- ✓ *Mulheres representam 40% da população*
- ✓ *Amostra 10 pessoas*
- ✓ *Homens: 6*
- ✓ *Mulheres: 4*

✓ Amostra estratificada

✓
1s



```
1 from sklearn.model_selection import StratifiedShuffleSplit
```

```
[ ] 1 dataset['income'].value_counts()
```

<=50K	24720
>50K	7841

Name: income, dtype: int64

<=50K	24720
>50K	7841

Name: income, dtype: int64

```
[ ] 1 7841 / len(dataset), 24720 / len(dataset)
      (0.2408095574460244, 0.7591904425539756)
```

```
[ ] 1 0.2408095574460244 + 0.7591904425539756
      1.0
```



```
1 split = StratifiedShuffleSplit(test_size=0.1)
2 for x, y in split.split(dataset, dataset['income']):
3     df_x = dataset.iloc[x]
4     df_y = dataset.iloc[y]
```



```
1 split = StratifiedShuffleSplit(test_size=0.1)
2 for x, y in split.split(dataset, dataset['income']):
3     df_x = dataset.iloc[x]
4     df_y = dataset.iloc[y]
```

```
1 df_x.shape, df_y.shape
```

```
((29304, 16)), (3257, 16)
```



```
[ ] 1 df_y.head()
```

	age	workclass	final-weight	education	education-num	marital-status	occupation	relationship
25406	26	Private	122206	HS-grad	9	Never-married	Craft-repair	Other-relative
3639	39	Private	218490	Bachelors	13	Married-civ-spouse	Sales	Husband
2916	38	Private	37028	Some-college	10	Divorced	Sales	Unmarried
25166	27	Private	303954	Bachelors	13	Married-civ-spouse	Sales	Husband
15429	36	Private	149833	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife

df_y.head()

	l-ht	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loos	hour-per-week	native-country	income	grupo
	38	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	60	United-States	>50K	47
	60	Assoc-voc	11	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	1902	45	United-States	<=50K	287
	77	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	65	United-States	>50K	151
	92	HS-grad	9	Never-married	Craft-repair	Not-in-family	White	Male	0	0	40	United-States	>50K	257
	06	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	Black	Male	0	0	40	United-States	<=50K	62



```
1 100 / len(dataset)
```



```
0.0030711587481956942
```

```
[ ] 1 split = StratifiedShuffleSplit(test_size=0.1)
    2 for x, y in split.split(dataset, dataset['income']):
    3     df_x = dataset.iloc[x]
    4     df_y = dataset.iloc[y]
```

```
[99] 100 / len(dataset)
```

```
0.0030711587481956942
```

```
[100] split = StratifiedShuffleSplit(test_size=0.0030711587481956942)
      for x, y in split.split(dataset, dataset['income']):
          df_x = dataset.iloc[x]
          df_y = dataset.iloc[y]
```

```
df_x.shape, df_y.shape
```

```
((32461, 16), (100, 16))
```

```
[98] df_y.head()
```

	age	workclass	final-weight	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain
4847	34	Private	253438	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	White	Male	0
28987	57	Private	64960	Assoc-voc	11	Married-civ-spouse	Prof-specialty	Husband	White	Male	0

```
[ ] 1 df_y['income'].value_counts()
```

```
<=50K    76
```

```
>50K     24
```

```
Name: income, dtype: int64
```

<>



```
[91] from sklearn.model_selection import StratifiedShuffleSplit
```

```
[92] dataset['income'].value_counts()
```

```
<=50K    24720  
>50K      7841  
Name: income, dtype: int64
```

```
7841 / len(dataset), 24720 / len(dataset)
```

```
(0.2408095574460244, 0.7591904425539756)
```

```
[95] 0.2408095574460244 + 0.7591904425539756
```

```
1.0
```

```
[99] 100 / len(dataset)
```

```
0.0030711587481956942
```

```
[100] split = StratifiedShuffleSplit(test_size=0.0030711587481956942)
```



CENTRO UNIVERSITÁRIO
UNISATC



  /UNISATC