Data_Bias-_Fairness_Gerrymandering

November 2, 2024

1 Data Bias: Fairness Gerrymandering

In this exercise you will slip into the role of data scientists that are requested as data experts for a judicial dispute. The scenario in dispute is as follows:

A woman of color applied for a job at the company *MajorEngine*, but got rejected. She suspects that she got turned down for racist and sexist reasons, *i.e.* because she is a woman of color. *MajorEngine* refutes this claim and provides employment records in court in order to disprove the claims.

```
[1]: import pandas as pd
     import matplotlib.pyplot as plt

     # load the data from the file 'hiring_records_MajorEngine.csv' and inspect the␣
      ↪first rows with the pandas function 'head'
     data = pd.read_csv("./hiring_records_MajorEngine.csv")
     data.head()
```

```
[1]:   gender      race
     0   male     white
     1 female     white
     2 female     white
     3   male     white
     4   male  hispanic
```

1.0.1 Task 1

Slip into the role of a data scientist hired by *MajorEngine* in order to show that based on the employment records

(a) the company has no racist hiring policy, and

(b) has no strongly sexist hiring policy. Note that according to the 2020 U.S. census, the perfect, expected percentage of white employees would be 61.6%.

Use bar charts to convey your findings to a lay person and write a comment that explains your figure in favor of *MajorEngine*.

Hint: While exploring the dataset, look at the ratio of white employees vs. non-white employees, and the ratio of male employees vs. non-male employees. It can also be useful to create a plot of the ideal distribution as comparison.

[28]:

```
# Part (a): show that MajorEngine has no strongly racist hiring policy
employ_num = data.shape[0]
races = data["race"].unique()

race_percentage = {}

for race in races:
    race_percentage[race] = (data[data["race"] == race].shape[0]) /␣
 ↪float(employ_num)

print(race_percentage)
```

```
{'white': 0.7, 'hispanic': 0.171, 'black': 0.108, 'asian': 0.021}
```

The result seems to imply a slight racist trend according to the 2020 US census, but according to a newer study from 2021, the hiring policy could be valued as a bit more racist. But the problem about this argumentation is though, that even if there are more white employees than the ethnicity statistics imply, the reality is, that with white people being privileged, there also is a trend of white-people working in jobs with more academical qualifications than non-white people.

In summary this means, that depending on the context of this being an academic job or not, the results could imply a less racist employment scheme, because the general distribution of different races should not be assumed to be the same when talking about different academic qualifications, because non-white people have a harder time getting the same qualifications as white-people because the system itself is racist and historically spoken, non-white people have lived in worse environments in America (i.e. slums etc.).

[29]:

```
# Part (b): Show that MajorEngine has no sexist hiring policy

employ_num = data.shape[0]
genders = data["gender"].unique()

gender_percentage = {}

for gender in genders:
    gender_percentage[gender] = (data[data["gender"] == gender].shape[0]) /␣
 ↪float(employ_num)

print(gender_percentage)
```

```
{'male': 0.5, 'female': 0.497, 'non-binary': 0.003}
```

These results seem to imply that the hiring policy is not sexist, since the distribution of males and females was 49.5% to 50.5% in the US (2021, source).
I have not found any statistics of how many people identify as non-binary in the US, so I can't check if the hiring policy is queer-phobic, but at this low sample-size there are many factors that could contextualize the percentage (f.e. non-binary people being registered with a binary gender).

1.0.2 Task 2

Slip into the role of a data scientist that works pro bono in order to demonstrate that *MajorEngine* has exhibited a bias in the past and thus is likely to have treated the woman of color unfairly.

Use a confusion matrix to convey your findings to a lay person.

Hint: While superficially, the argumentation form task 1 may seem sound, you have the sneaking suspicion that you should look at the two attributes 'race' and 'gender' in combination instead of separately.

Second hint: You may create a makeshift confusion matrix by creating another pandas dataframe of the four intersectional values and renaming columns and index.

[45]:

```
employ_num = data.shape[0]
races = data["race"].unique()
genders = data["gender"].unique()

confusion_matrix = pd.DataFrame(columns=["gender", "race", "expected␣
 ↪percentage", "real percentage", "difference"])

# we'll use values calculated in task 1 to simplicity assume that there is no␣
 ↪one-dimensional biased hiring policy.
expected_gender = {'male': 0.5, 'female': 0.497, 'non-binary': 0.003}
expected_race = {'white': 0.7, 'hispanic': 0.171, 'black': 0.108, 'asian': 0.
 ↪021}

size = 0

for gender in genders:
    for race in races:
        # we'll assume for this that the distribution should be independent, if␣
 ↪it is, then we have bias.
        expected_percentage = expected_gender[gender] * expected_race[race]
        real_percentage = (data[(data["gender"] == gender) & (data["race"] ==␣
 ↪race)]).shape[0] / float(employ_num)

        confusion_matrix.loc[size] = [
            gender,
            race,
            expected_percentage,
            real_percentage,
            real_percentage - expected_percentage
        ]
        size += 1

print(confusion_matrix)
```

```
        gender      race  expected percentage  real percentage  difference
0         male     white             0.350000            0.200   -0.150000
```

```
1         male  hispanic             0.085500             0.171    0.085500
2         male     black             0.054000             0.108    0.054000
3         male     asian             0.010500             0.021    0.010500
4       female     white             0.347900             0.497    0.149100
5       female  hispanic             0.084987             0.000   -0.084987
6       female     black             0.053676             0.000   -0.053676
7       female     asian             0.010437             0.000   -0.010437
8   non-binary     white             0.002100             0.003    0.000900
9   non-binary  hispanic             0.000513             0.000   -0.000513
10  non-binary     black             0.000324             0.000   -0.000324
11  non-binary     asian             0.000063             0.000   -0.000063
```

The results imply that indeed there seems to be a bias against white men and non-white women, which looks like the company is hiring white women to hire women and non-white men to hire non-white people. This is unfair for white men as well as non-white women. Because of the low sample-size I am ignoring the results for non-binary people.