EE 325 Programming Assignment

Vedant Bhardwaj

September 2, 2024

1 Part 2

1.1 Is the IITB UG population a good sample of India at the granularity of states? If not, how badly is it skewed in favour of some states? De ne your own measure for skew (remember, no cheating or being inspired by anything that Google/ChatGPT/... can throw up) and apply to the data that you have. Justify the measure.

We can answer this question by comparing the population distribution across each state of IITB vs. India. Ideally, the ratio of them should be close to one but (a **HUGE** but) it doesn't always mean that if the ratio is one or close to one then our data isn't skewed (**Why?**). Because the mean of those ratios can still be one but one of them can be as high as 3.0 and the other as low as 0.25, this would mean our data is skewed towards the state with a 3.0 ratio.

If the state population ratio is not close to one for every state then our data is skewed and will be more skewed towards the state with a higher state population ratio in IITB than other states.

So we designed a parameter that measures the skewness of the population in the dataset. We are calculating skewness by comparing the probability ratios of state i in IITB and India.

$p_i = Populationofstateiin IITB/1500$

$q_i = Populationofstateiin India/Totalpopulationof India$

Now if our data is not skewed then for a given state i p_i/q_i should be close to 1 and variance low. So my parameter for skewness is:-

$\sum \frac{p_i}{q_i}/23$

Why 23? 23 is the number of states. So apparently our one parameter is the average of p_i/q_i values and **the other parameter would be the variance of the ratios.** As you can see my p_i/q_i values shoot up for places like Maharashtra, Andhra Pradesh, and Goa, this tells us that our data is skewed towards these few states because ideally it was supposed to be uniform but it isn't. Also a look over variance can also help us to comment on this.

	State Codes	State	Population	prob_of_state_IITB	prob_of_state_India	Pi/Qi	var
0	AP	Andhra Pradesh	49577103	0.090000	0.041076	2.191060	1.152851
1	BR	Bihar	104099452	0.041333	0.086249	0.479231	0.407197
2	CG	Chhattisgarh	25545198	0.015333	0.021165	0.724470	0.154356
3	GA	Goa	1458545	0.004667	0.001208	3.861711	7.531514
4	GJ	Gujarat	60439692	0.053333	0.050076	1.065048	0.002736
5	HR	Haryana	25351462	0.028667	0.021004	1.364794	0.061228
6	HP	Himachal Pradesh	6864602	0.004667	0.005688	0.820511	0.088114
7	JH	Jharkhand	32988134	0.012667	0.027332	0.463444	0.427594
8	KA	Karnataka	61095297	0.031333	0.050619	0.619001	0.248352
9	KE	Kerala	33406061	0.010667	0.027678	0.385386	0.535772
10	MP	Madhya Pradesh	72626809	0.060667	0.060173	1.008198	0.011914
11	MH	Maharashtra	112374333	0.280000	0.093105	3.007348	3.572091
12	OD	Odisha	41974219	0.008667	0.034777	0.249208	0.753671
13	PB	Punjab	27743338	0.008667	0.022986	0.377039	0.548061
14	RJ	Rajasthan	68548437	0.099333	0.056794	1.749002	0.398983
15	TN	Tamil Nadu	72147030	0.024000	0.059776	0.401500	0.512442
16	TL	Telangana	35003674	0.084000	0.029002	2.896400	3.165015
17	UP	Uttar Pradesh	199812341	0.065333	0.165550	0.394644	0.522305
18	UK	Uttarakhand	10086292	0.007333	0.008357	0.877532	0.057513

Figure 1: p_i/q_i values

Our skew-mean value is 1.11 which doesn't help us infer anything about the skewness of the data but variance is 0.99 which is high considered such small valued samples.

1.2 Considering the population and the per capita income of the states, is the distribution of the student body among the states/regions fair? Once again, define your own measure for skewfairness and apply to the data that you have. Justify the measure.

Skewfairness metric The question asks us to check for the fairness in the population (state-wise) distribution of IITB concerning the distribution of per capita income and population. So we can basically think of it as a multiple linear regression problem.

Where y_i is a fraction of people of $state_i$ from the total IITB students. And other two independent variables are first - fraction of per capita income of people of state i (we will consider it as x_i and the fraction of people of state i from the total population of India (we will consider it as q_i.

So considering our skewfairness depends on the per capita income of state i and the population of state i, this implies I have to find the fairness in the relationship of the population of IITB and the per capita income of state i and the population of state i. So we have to first find a way to indirectly calculate the correlation between per capita income and population of IITB states.

Now the question is what is correlation and how to find it? Correlation is basically a strength of how two values are strongly linearly correlated.

```
Correlation Matrix:
                     Fraction_PerCapita  Fraction_Population  \
Fraction_PerCapita             1.000000            -0.443796
Fraction_Population           -0.443796             1.000000
IITB_Pop_Fraction              0.019784             0.443617

                     IITB_Pop_Fraction
Fraction_PerCapita            0.019784
Fraction_Population           0.443617
IITB_Pop_Fraction             1.000000
```

Figure 2: Correlation matrix

```
IITB_Pop_Fraction             1.000000
                            OLS Regression Results
==============================================================================
Dep. Variable:      IITB_Pop_Fraction   R-squared:                       0.255
Model:                            OLS   Adj. R-squared:                  0.181
Method:                 Least Squares   F-statistic:                     3.427
Date:                Sun, 01 Sep 2024   Prob (F-statistic):             0.0525
Time:                        21:45:14   Log-Likelihood:                 35.989
No. Observations:                  23   AIC:                            -65.98
Df Residuals:                      20   BIC:                            -62.57
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
```

Figure 3: Regression results

The more it is close to one the more linear the relationship between variables. In linear regression, we use **Pearson coefficient number**. But since it is an MLR problem we will be having a correlation matrix whose elements have Pearson coefficient number between all the variables (dependent-independent, independent-independent)

How to get Pearson coefficient no?

$r = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$

Now we can conclude that the IITB state population distribution had skew-fairness of only 0.4436 when compared to the state population distribution of India and only 0.0019 of fairness when comparing to per capita distribution states across India.

1.3 How strongly does the state affect the JEE rank, the graduating CPI, and the first salary?

For comparing the correlation between various parameters and state of origin, literacy rate can be directly related. **How?** We are here at IITB, the best engineering college. Now comparing state of origin sounds vague. So we decided to compare state with some other parameter and then use that parameter to check how state affects jee rank, cpi, and first salary.

There are various parameters we can use like GDP, per capita income, and literacy but the most strongly correlated is literacy. So we chose literacy as a parameter to correlate state with JEE rank, CPI, and first salary.

Here's the code block for it.

```
import pandas as pd
```

```
Correlation between mean Rank and Literacy: 0.24
Correlation between mean CPI and Literacy: -0.39
Correlation between mean First Salary and Literacy: -0.22
```

Figure 4: Correlation comaparison

```
jee_data_path = '/content/JEEDemographics.csv'
jee_df = pd.read_csv(jee_data_path)

state_info_path = '/content/StateInfo.csv'
state_info_df = pd.read_csv(state_info_path)

state_means = jee_df.groupby('Origin').agg({'Rank': 'mean', 'CPI': 'mean', 'First_Salary':

merged_df = pd.merge(state_means, state_info_df, left_on='Origin', right_on='State Codes')

correlation_rank = merged_df['Rank'].corr(merged_df['Literacy %'])
correlation_cpi = merged_df['CPI'].corr(merged_df['Literacy %'])
correlation_salary = merged_df['First_Salary'].corr(merged_df['Literacy %'])

print(f"Correlation between mean Rank and Literacy: {correlation_rank:.2f}")
print(f"Correlation between mean CPI and Literacy: {correlation_cpi:.2f}")
print(f"Correlation between mean First Salary and Literacy: {correlation_salary:.2f}")
```

(i) Sates v/s JEE Rank Correlation = 0.24 It means that state of origin has a moderately negative effect on JEE Rank. A correlation of 0.24 suggests that there is a slight tendency for JEE Rank to increase as the literacy rate of state increases, but this relationship is not strong. The relationship is not strictly linear.

This makes sense right? because **look at Andhra Pradesh** it has literacy rate of 66.9 percent but still top rankers are from Andhra Pradesh on the other hand there are other states with overall better accuracy but with high jee ranks. This overall induces a negative correlation in the dataset. (**Negative correlation here means that expected correlation was that if literacy increases accuracy decreases but it is negative of that, that's why negative of that correlation**)

(ii) States v/s Graduating CPI Correlation = -0.39 It means that state of origin has a moderately negative effect on Graduating CPI. A correlation of 0.39 suggests that there is a moderate tendency for Graduating CPI to decrease as the literacy rate of state increases. It also indicates that our data is biased as we have taken literacy rates into consideration and literacy rates should have a positive correlation with CPI but quantitatively it is following the opposite trend.

This also follows the trend which is **caused by the contamination in the dataset caused by outliers** (example AP, Rajasthan).

(iii) States v/s First Salary Correlation = -0.22 It means that state of origin has a weak negative effect on first Salary. A correlation of 0.22 suggests that there is a weak tendency for First Salary to decrease as the literacy rate of state increases. This also shows the biased nature of our data as first salary should increase along with literacy rate.

It is also because some of the values can alter the results drastically. For example, literacy rate of Andhra Pradesh is 66.9

1.4 How strongly does the family income affect the JEE rank, the graduating CPI, and the first salary?

(i) Family income v/s JEE Rank Correlation = -0.30 It means that family income has a weak negative effect on JEE Rank. A correlation of 0.30 suggests that there is a slight tendency for JEE rank to decrease as the family income increases, but this relationship is relatively weak. It makes sense as high family income will provide more quality education to the child and hence, less ranks will show in the results.

(ii) Family income v/s Graduating CPI Correlation = 0.16 It means that family income has a weak positive effect on Graduating CPI. A correlation of 0.16 suggests that there is a slight tendency for Graduating CPI to increase as the family income increases, but this relationship is also relatively weak. Graduating CPI will depend more on the students' performance over the years of his/her so family income doesn't affect it too much.

(iii) Family income v/s First Salary Correlation = 0.61 It means that family income has a strong positive effect on Graduating CPI. A correlation of 0.61 suggests that there is a strong tendency for First Salary to increase as the family income increases, and this relationship is strong and approaching linearity. High family income generally indicates to well-mannered families and their children got a good quality education as well as exposure of what is being happened around them so they may learn more than other students and got good skills which are required to grab high first salary in the workplace.

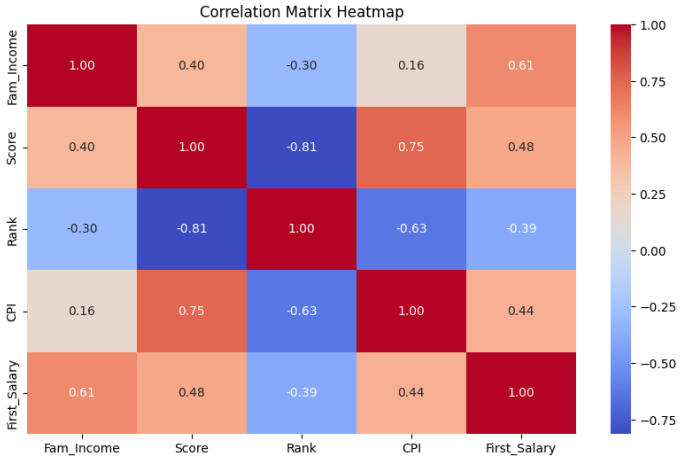

Figure 5: Enter Caption