

# Untitled

ARCHIVE

## Hypothesis Testing and ANOVA

The null hypothesis is whether the number of internet users differs for different income class. There is no association between the number of internet users and income classes.

The alternative hypothesis states that there is an association between the number of internet users and income classes.

where the variable ipp means income per person

- 1 – income per person less than 1,000
- 2 – income per person between 1,000 and 4,000
- 3 – income per person between 4,000 and 7,000
- 4 – income per person between 7,000 and 10,000
- 5 – income per person greater than 10,000

From the results of the ANOVA procedure, we have a F value of 83.64 and a P-value of 0.0001. Since the P-value of 0.0001 is less than our significance level of 0.05, we reject the null hypothesis. Therefore, we can conclude that there is an association between the number of internet users and the income classes.

Since our test is significant at  $P < 0.0001$ , we proceed to run the post hoc test. Using the Duncan multiple paired comparison, the income rates 3 and 4 are not significantly different. Income group 5 has significantly more number of internet users in 2010 than income groups 4, 3, 2 and 1.

```

1 LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
2 Data new;set mydata.gapminder;
3 keep country incomeperperson employrate lifeexpectancy internetuserate ipp er le;
4
5 Label incomeperperson="2010 GDP per capita in constant 2000$"
6 label employrate="2007 percent of Population age 15 and above employed"
7 label lifeexpectancy="2011 life expectancy at birth in years"
8 label internetuserate="2010 number of internet users";
9
10 /*data management for incomeperperson*/
11 if incomeperperson <= 1000 then ipp = "1";
12 if incomeperperson > 1000 and incomeperperson <= 4000 then ipp = "2";
13 if incomeperperson > 4000 and incomeperperson <= 7000 then ipp = "3";
14 if incomeperperson > 7000 and incomeperperson <= 10000 then ipp = "4";
15 if incomeperperson > 10000 then ipp = "5";
16
17 /*data management for employrate*/
18 if employrate < 20 then er = "1";
19 if employrate >= 20 and employrate < 40 then er = "2";
20 if employrate >= 40 and employrate < 60 then er = "3";
21 if employrate >= 60 and employrate < 80 then er = "4";
22 if employrate >= 80 then er = "5";
23
24 /*data management for lifeexpectancy*/
25 if lifeexpectancy < 45 then le = "1";
26 if lifeexpectancy >= 45 and lifeexpectancy < 55 then le = "2";
27 if lifeexpectancy >= 55 and lifeexpectancy < 65 then le = "3";
28 if lifeexpectancy >= 65 and lifeexpectancy < 75 then le = "4";
29 if lifeexpectancy >= 75 then le = "5";
30
31 run;
32
33 proc sort;by COUNTRY;
34 proc freq;tables ipp er le;
35 run;
36
37 /*univariate graphs*/
38 PROC GCHART; VBAR ipp/Discrete type=PCT width=30;
39 run;
40 PROC GCHART; VBAR er/Discrete type=PCT width=30;
41 run;
42 PROC GCHART; VBAR le/Discrete type=PCT width=30;
43 run;
44
45 /*bivariate graph*/
46 PROC GPLOT; PLOT incomeperperson*employrate;
47 run;
48
49 proc anova; class ipp;
50 model internetuserate=ipp;
51 means ipp/duncan;
52
53 run;

```

Never miss a post!



anyasosamuel

Untitled

Follow

# The ANOVA Procedure

Class Level Information					
Class	Levels	Values			
ipp	5	1	2	3	4 5

Number of Observations Read	213
Number of Observations Used	192

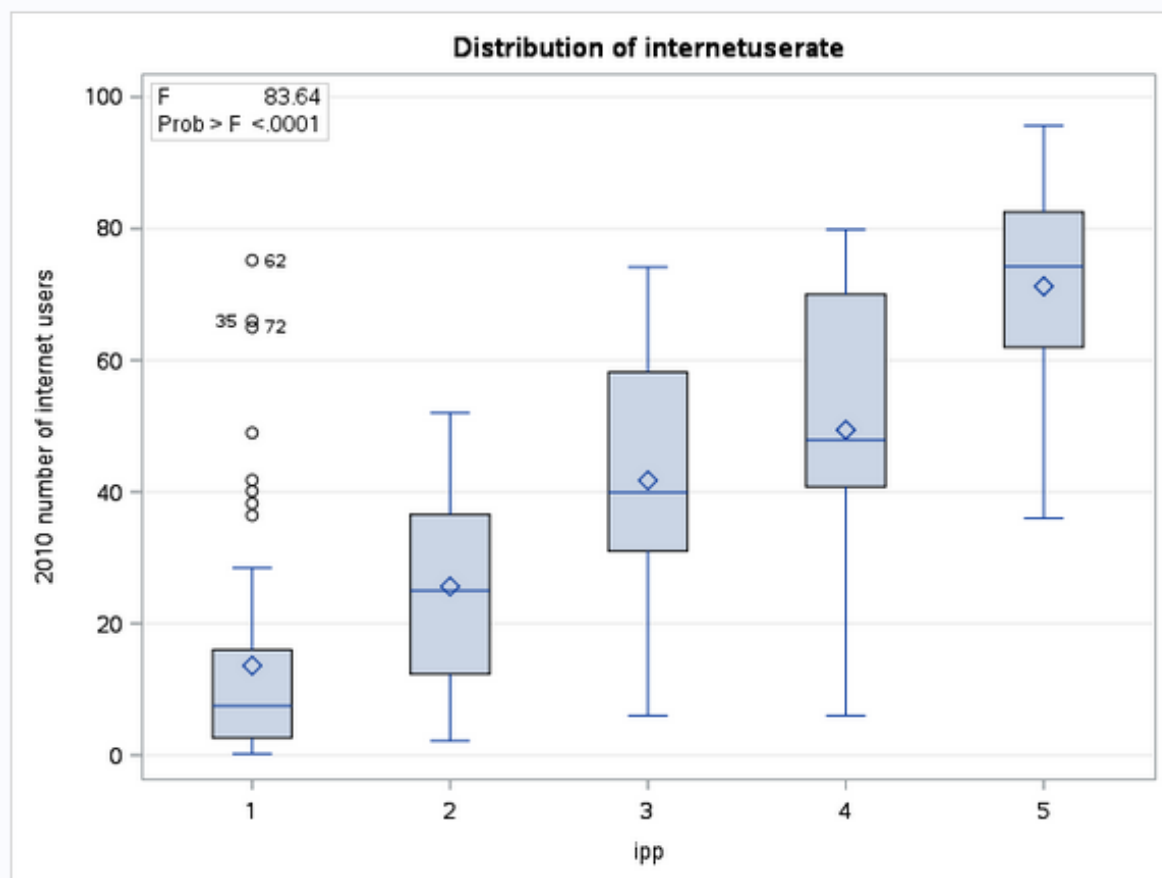
## The ANOVA Procedure

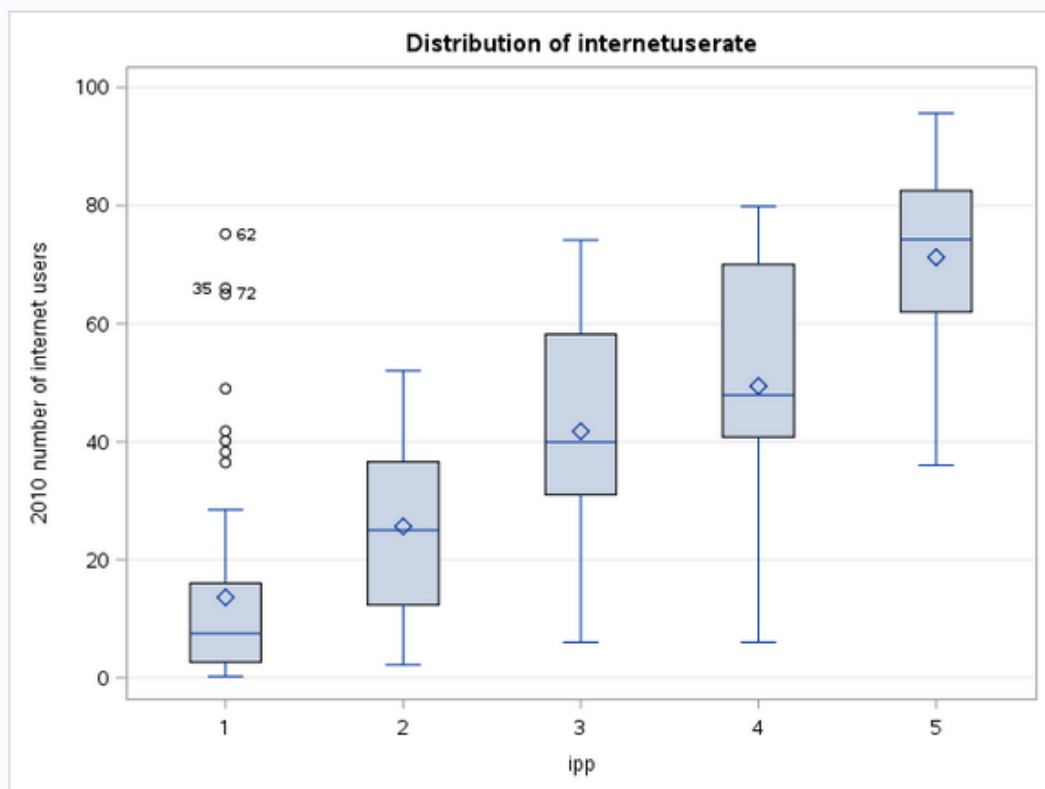
Dependent Variable: internetuserate 2010 number of internet users

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	94553.6911	23638.4228	83.64	<.0001
Error	187	52849.4533	282.6174		
Corrected Total	191	147403.1444			

R-Square	Coeff Var	Root MSE	internetuserate Mean
0.641463	47.17919	16.81123	35.63272

Source	DF	Anova SS	Mean Square	F Value	Pr > F
ipp	4	94553.69105	23638.42276	83.64	<.0001





**The ANOVA Procedure**  
**Duncan's Multiple Range Test for internetuserate**

**Note:** This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	187
Error Mean Square	282.6174
Harmonic Mean of Cell Sizes	23.77993

**Note:** Cell sizes are not equal.

Number of Means	2	3	4	5
Critical Range	9.62	10.12	10.46	10.71

**Means with the same letter are not significantly different.**

Duncan Grouping	Mean	N	ipp
A	71.246	45	5
B	49.421	9	4
B			
B	41.742	24	3
C	25.627	53	2
D	13.616	61	1

# Creating data graphs

```
1 LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
2 Data new;set mydata.gapminder;
3 keep country incomeperperson employrate lifeexpectancy ipp er le;
4
5 Label incomeperperson="2010 GDP per capita in constant 2000$"
6 label employrate="2007 percent of Population age 15 and above employed"
7 label lifeexpectancy="2011 life expectancy at birth in years";
8
9 /*data management for incomeperperson*/
10 if incomeperperson <= 1000 then ipp = "1";
11 if incomeperperson > 1000 and incomeperperson <= 4000 then ipp = "2";
12 if incomeperperson > 4000 and incomeperperson <= 7000 then ipp = "3";
13 if incomeperperson > 7000 and incomeperperson <= 10000 then ipp = "4";
14 if incomeperperson > 10000 then ipp = "5";
15
16 /*data management for employrate*/
17 if employrate < 20 then er= "1";
18 if employrate >= 20 and employrate < 40 then er = "2";
19 if employrate >= 40 and employrate < 60 then er = "3";
20 if employrate >= 60 and employrate < 80 then er = "4";
21 if employrate >= 80 then er = "5";
22
23 /*data management for lifeexpectancy*/
24 if lifeexpectancy < 45 then le = "1";
25 if lifeexpectancy >= 45 and lifeexpectancy < 55 then le = "2";
26 if lifeexpectancy >= 55 and lifeexpectancy < 65 then le = "3";
27 if lifeexpectancy >= 65 and lifeexpectancy < 75 then le = "4";
28 if lifeexpectancy >= 75 then le = "5";
29
30 run;
31
32 proc sort;by COUNTRY;
33 proc freq;tables ipp er le;
34 run;
35
36 /*univariate graphs*/
37 PROC GCHART; VBAR ipp/Discrete type=PCT width=30;
38 run;
39 PROC GCHART; VBAR er/Discrete type=PCT width=30;
40 run;
41 PROC GCHART; VBAR le/Discrete type=PCT width=30;
42 run;
43
44 /*bivariate graph*/
45 PROC GPLOT; PLOT incomeperperson*employrate;
46 run;
```

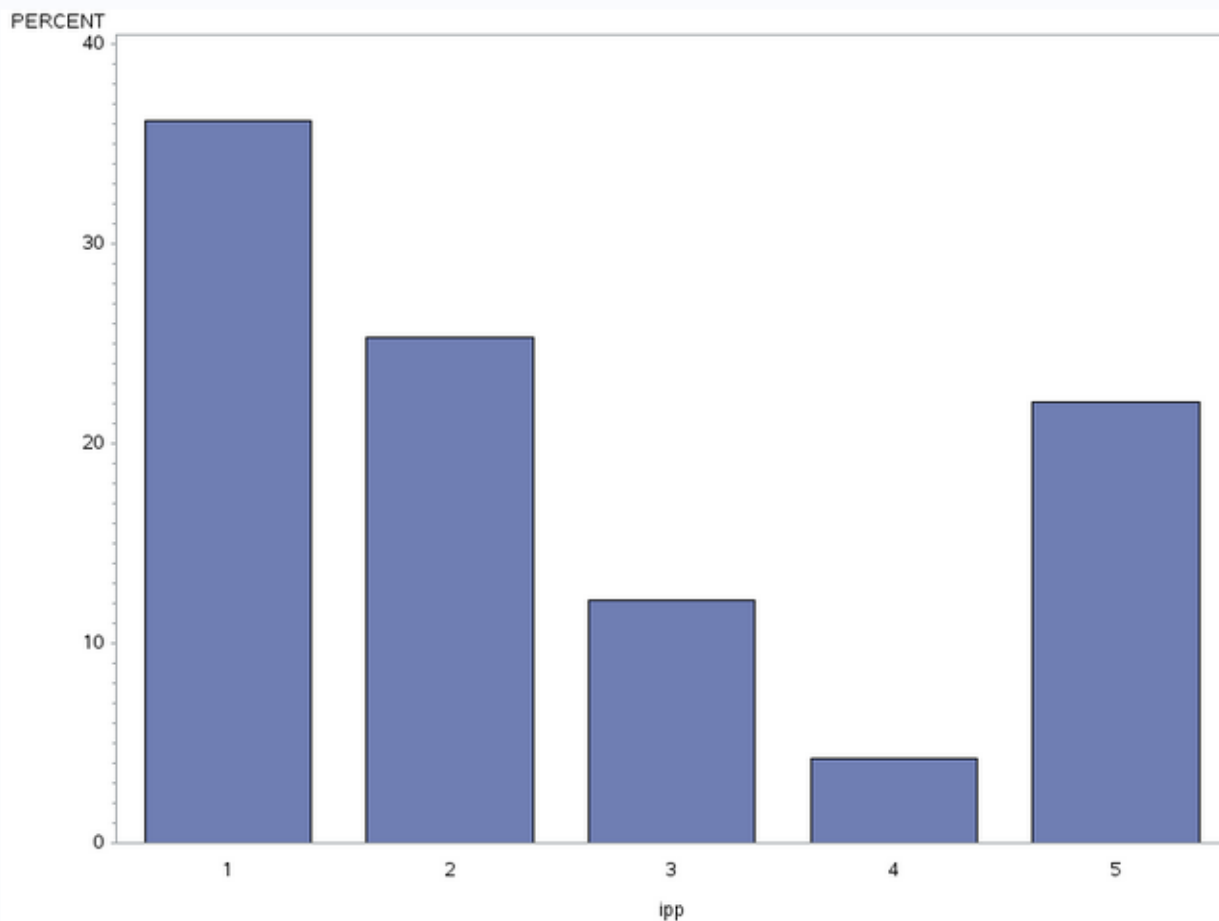
The following are meaning of the variables;

ipp - income per person

er - employment rate

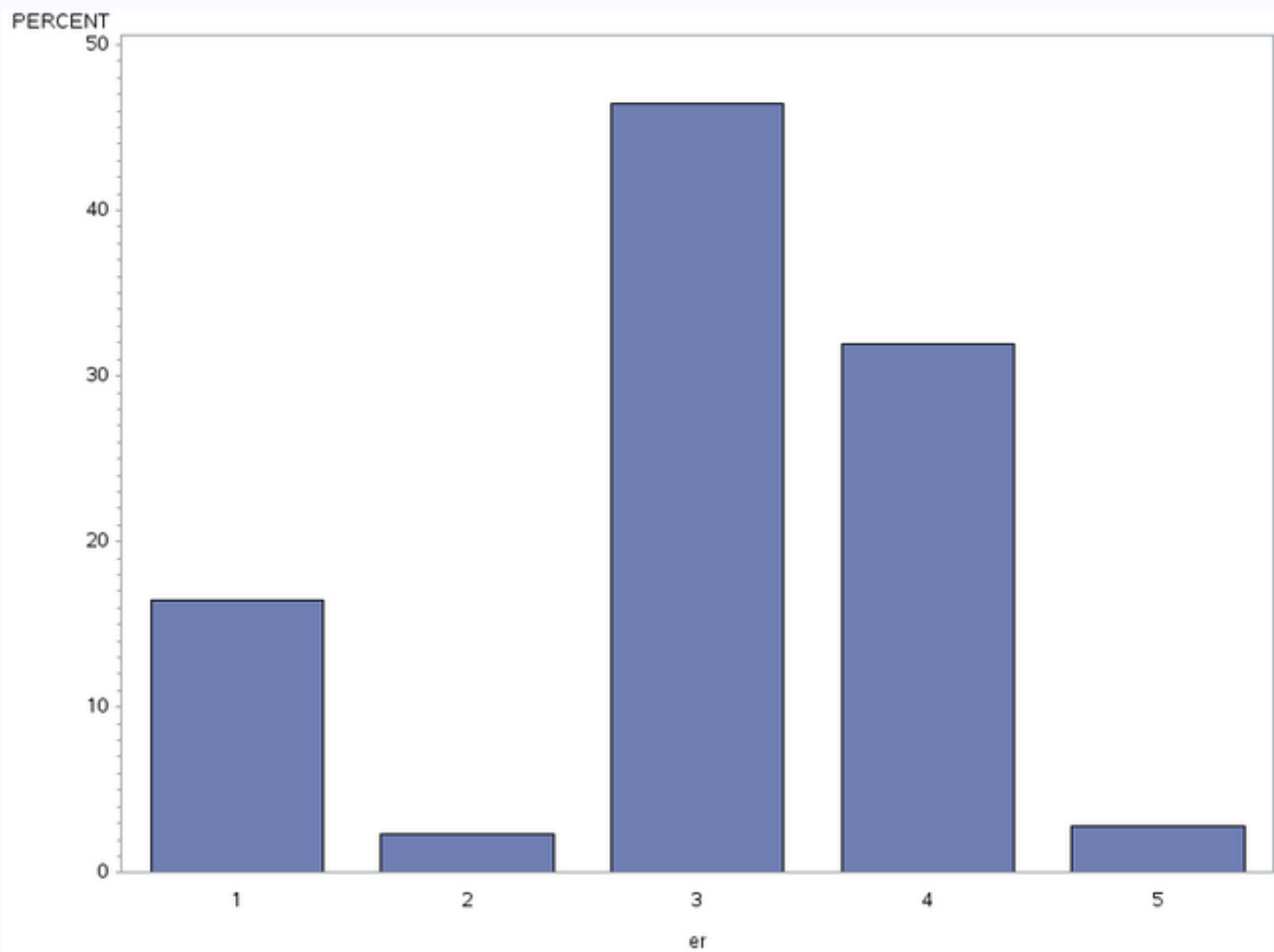
le - life expectancy

**The univariate graph of the income per person:**



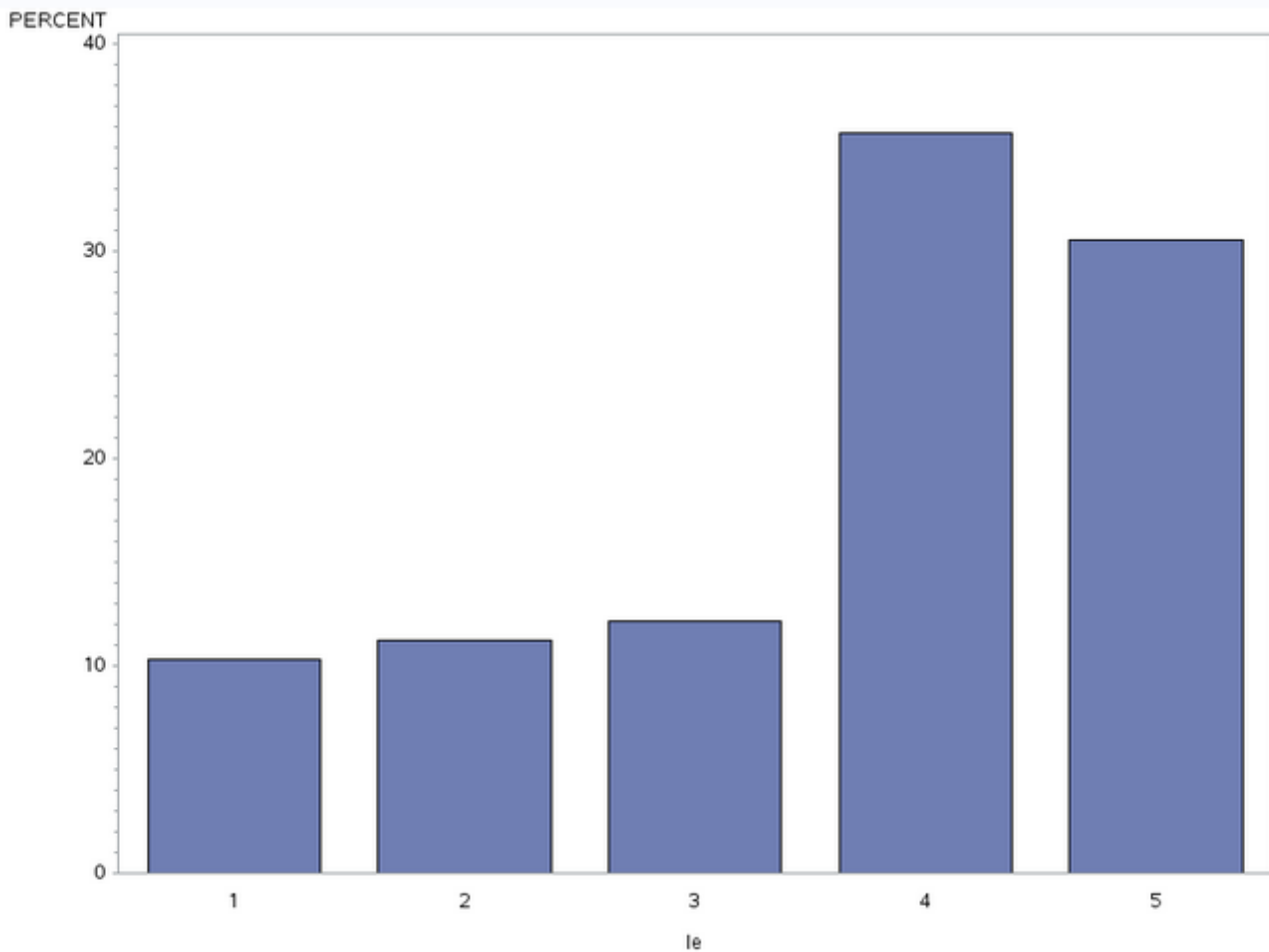
This graph is unimodal, with its highest peak at income less than \$1,000 per person. It seems to be skewed to the right however the skewness is not pronounced.

**The univariate graph of the employment rate:**

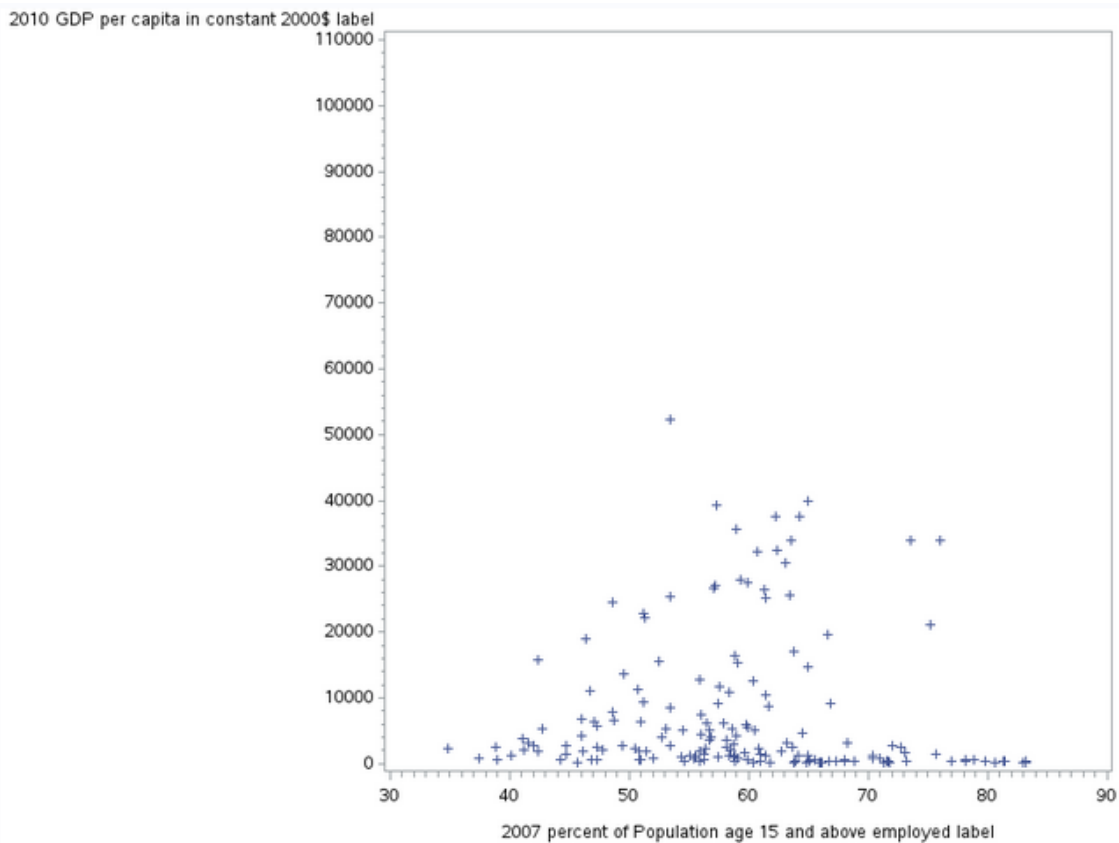


This graph is unimodal, with its highest peak at the median category of 40 to 60% employment rate. It seems to be skewed to the left as there are lower frequencies in lower categories than the higher categories.

**The univariate graph of the life expectancy:**



This graph is unimodal, with its highest peak at 65 to 75 years. It seems to be skewed to the left as there are higher frequencies in the categories with higher ages than in the categories with lower ages.





The graph above plots the 2010 GDP per capita in constant \$2000 against the 2007 percentage of population age 15 and above that are employed. The scatter graph shows that with increased percentage of 2007 population age 15 and above that are employed, there is a gradual increase in the 2010 GDP per capita in constant \$2000.

## Assignment 3: Making Data Management Decisions

```
1 LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
2 Data new;set mydata.gapminder;
3 keep country incomeperperson employrate lifeexpectancy ipp er le;
4
5 Label incomeperperson="2010 GDP per capita in constant 2000$"
6 label employrate="2007 percent of Population age 15 and above employed"
7 label lifeexpectancy="2011 life expectancy at birth in years";
8
9 /*data management for incomeperperson*/
10 if incomeperperson <= 1000 then ipp = "1";
11 if incomeperperson > 1000 and incomeperperson <= 4000 then ipp = "2";
12 if incomeperperson > 4000 and incomeperperson <= 7000 then ipp = "3";
13 if incomeperperson > 7000 and incomeperperson <= 10000 then ipp = "4";
14 if incomeperperson > 10000 then ipp = "5";
15
16 /*data management for employrate*/
17 if employrate < 20 then er= "1";
18 if employrate >= 20 and employrate < 40 then er = "2";
19 if employrate >= 40 and employrate < 60 then er = "3";
20 if employrate >= 60 and employrate < 80 then er = "4";
21 if employrate >= 80 then er = "5";
22
23 /*data management for lifeexpectancy*/
24 if lifeexpectancy < 45 then le = "1";
25 if lifeexpectancy >= 45 and lifeexpectancy < 55 then le = "2";
```

```

8
9 /*data management for incomeperperson*/
10 if incomeperperson <= 1000 then ipp = "1";
11 if incomeperperson > 1000 and incomeperperson <= 4000 then ipp = "2";
12 if incomeperperson > 4000 and incomeperperson <= 7000 then ipp = "3";
13 if incomeperperson > 7000 and incomeperperson <= 10000 then ipp = "4";
14 if incomeperperson > 10000 then ipp = "5";
15
16 /*data management for employrate*/
17 if employrate < 20 then er= "1";
18 if employrate >= 20 and employrate < 40 then er = "2";
19 if employrate >= 40 and employrate < 60 then er = "3";
20 if employrate >= 60 and employrate < 80 then er = "4";
21 if employrate >= 80 then er = "5";
22
23 /*data management for lifeexpectancy*/
24 if lifeexpectancy < 45 then le = "1";
25 if lifeexpectancy >= 45 and lifeexpectancy < 55 then le = "2";
26 if lifeexpectancy >= 55 and lifeexpectancy < 65 then le = "3";
27 if lifeexpectancy >= 65 and lifeexpectancy < 75 then le = "4";
28 if lifeexpectancy >= 75 then le = "5";
29
30 proc sort;by COUNTRY;
31 proc freq;tables incomeperperson employrate lifeexpectancy ipp er le;
32 run;

```

ipp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	77	36.15	77	36.15
2	54	25.35	131	61.50
3	26	12.21	157	73.71
4	9	4.23	166	77.93
5	47	22.07	213	100.00

er	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	35	16.43	35	16.43
2	5	2.35	40	18.78
3	99	46.48	139	65.26
4	68	31.92	207	97.18
5	6	2.82	213	100.00

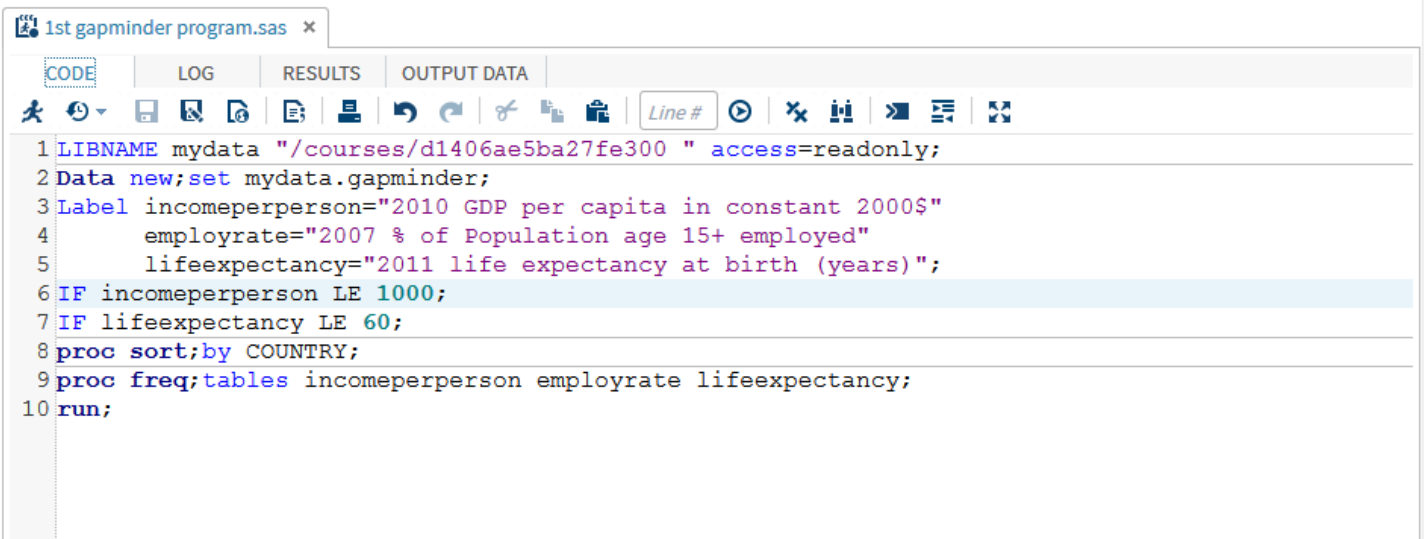
le	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	22	10.33	22	10.33
2	24	11.27	46	21.60
3	26	12.21	72	33.80
4	76	35.68	148	69.48
5	65	30.52	213	100.00

I summarized the incomeperperson, employrate and lifeexpectancy and created three new variables: ipp, er and le. From the frequency table created, it was observed that for the ipp variable, the most frequent outcome was 1(36.15), meaning that most countries have a GDP per capita less than 1000.

Next, for the er variable, we also observe that the most frequent outcome was 3 (46.48), which means that about half of the countries have an employment rate of 40%-60%.

Lastly, for the le variable, we observe that the most frequent outcome was 4 (35.68), which means that about one-third of the countries have life expectancy of 65-75years.

# 1st Gapminder Program



```
1 LIBNAME mydata "/courses/d1406ae5ba27fe300 " access=readonly;
2 Data new;set mydata.gapminder;
3 Label incomeperperson="2010 GDP per capita in constant 2000$"
4       employrate="2007 % of Population age 15+ employed"
5       lifeexpectancy="2011 life expectancy at birth (years)";
6 IF incomeperperson LE 1000;
7 IF lifeexpectancy LE 60;
8 proc sort;by COUNTRY;
9 proc freq;tables incomeperperson employrate lifeexpectancy;
10 run;
```

2010 GDP per capita in constant 2000\$				
incomeperperson	Frequency	Percent	Cumulative Frequency	Cumulative Percent
103.77585724	1	3.23	1	3.23
115.3059959	1	3.23	2	6.45
155.03323123	1	3.23	3	9.68
161.3171371	1	3.23	4	12.90
180.083376	1	3.23	5	16.13
184.14179659	1	3.23	6	19.35
220.89124792	1	3.23	7	22.58
239.51874937	1	3.23	8	25.81
268.3317903	1	3.23	9	29.03
269.89288112	1	3.23	10	32.26
275.88428653	1	3.23	11	35.48
276.20041296	1	3.23	12	38.71
285.22444925	1	3.23	13	41.94
320.77188995	1	3.23	14	45.16
338.26639123	1	3.23	15	48.39
354.59972629	1	3.23	16	51.61
377.03969946	1	3.23	17	54.84
377.42111326	1	3.23	18	58.06
389.76363425	1	3.23	19	61.29
411.50144725	1	3.23	20	64.52
432.22633697	1	3.23	21	67.74

2007 % of Population age 15+ employed

employrate	Frequency	Percent	Cumulative Frequency	Cumulative Percent
45.700000763	1	3.03	1	3.03
46.900001526	1	3.03	2	6.06
48.700000763	1	3.03	3	9.09
50.900001526	1	3.03	4	12.12
54.5	1	3.03	5	15.15
55.700000763	1	3.03	6	18.18
56.299999237	1	3.03	7	21.21
59.099998474	1	3.03	8	24.24
59.900001526	1	3.03	9	27.27
60.400001526	1	3.03	10	30.30
61	1	3.03	11	33.33
63.799999237	1	3.03	12	36.36
63.900001526	1	3.03	13	39.39
65.599998474	1	3.03	14	42.42
65.900001526	1	3.03	15	45.45
66	2	6.06	17	51.52
66.199996948	1	3.03	18	54.55
66.800003052	1	3.03	19	57.58
68.900001526	1	3.03	20	60.61
71.300003052	1	3.03	21	63.64
71.599998474	1	3.03	22	66.67
74.600000000	1	3.03	23	69.70

2011 life expectancy at birth (years)				
lifeexpectancy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
47.794	1	3.13	1	3.13
48.132	1	3.13	2	6.25
48.196	1	3.13	3	9.38
48.397	1	3.13	4	12.50
48.398	1	3.13	5	15.63
48.673	1	3.13	6	18.75
49.025	1	3.13	7	21.88
49.553	1	3.13	8	25.00
50.239	1	3.13	9	28.13
50.411	1	3.13	10	31.25
51.219	1	3.13	11	34.38
51.384	1	3.13	12	37.50
51.444	1	3.13	13	40.63
51.61	1	3.13	14	43.75
51.879	1	3.13	15	46.88
54.097	1	3.13	16	50.00
54.116	1	3.13	17	53.13
54.21	1	3.13	18	56.25
54.675	1	3.13	19	59.38
55.377	1	3.13	20	62.50
55.439	1	3.13	21	65.63
55.442	1	3.13	22	68.75

After running my first program using the Gapminder dataset in SAS, I was able to make some insightful deductions. The data comprises of continuous data drawn across from different categories, unless the specified variable is grouped or limited to a certain range, the frequency table would not provide an ideal summary for the data. Hence, I decided to limit the 2010 Gross Domestic Product per capita in constant 2000 US\$ to figures less than 1,000. Considering this,

the output table provided a better summary.

Also, I limited the 2011 life expectancy at birth (years) to figures less than 60 years to show a better summary. Restricting my data variables to these figures, logically, my research question would now shed more light on developing countries. As most developing countries have GDP per capita less than 1,000 and life expectancy less than 60.

## Evaluating the association between income and employment rate

After carefully looking through the datasets, I have decided to work with the Gapminder dataset. My growing interest in social, economic and health indicators across the world has influenced my choice for the dataset.

After carefully studying the Gapminder codebook, associations of income is the topic that I am particularly interested in. I will use these variables for this study; incomeperperson and employrate. This study aims at verifying whether economic capability plays a vital role in determining the health status and other health-related problems among the various age groups.

I have chosen to evaluate the association between income and employment rate. The codebook variables that would be used for this evaluation still remains the incomeperperson and employrate.

Also, I have decided to study the association among life expectancy, alcohol consumption and number of people living with HIV. I add to my codebook variables lifeexpectancy, alconsumption, and HIVrate

As part of literature review, Park BH et al deduced that the health status of the aged is related more closely to the individual's wealth than income.

Roy K, et al discusses the successive controls for education, income, and property ownership narrows the gender gap in both health and healthcare utilization, significant differentials still persist.

Pollack CE, et al propounds that health studies should include wealth as an important socioeconomic status indicator. Failure to measure wealth may result in under estimating the contribution of socioeconomic status indicator to health. Validation is needed for simpler approached to measuring wealth that would be feasible in health studies.

After studying the above literature, I developed a hypothesis which states whether the employment rate is dependent on the income per person. This hypothesis is formulated around the variables incomeperperson and employrate.

## CODEBOOK VARIABLES

Variable Name

Description of Indicator Main Source

incomeperperson 2010 Gross Domestic Product per capita in constant 2000 US\$. The inflation but not the differences in the cost of living between countries has been taken into account. World Bank Work Development

Indicators

employrate

2007 total employees age 15+ (% of population) Percentage of total population, age above 15, that has been employed during the given year.

International Labour

Organization

alcoholconsumption 2008 alcohol consumption per adult (age 15+), litres Recorded and estimated average alcohol consumption, adult (15+) per capita consumption in litres pure alcohol

WHO

lifeexpectancy

2011 life expectancy at birth (years)

The average number of years a newborn child would live if current mortality patterns were to stay the same.

1. Human Mortality Database,

2. World Population Prospects:

3. Publications and files by

history prof. James C Riley

4. Human Lifetable Database

HIVrate 2009 estimated HIV Prevalence % - (Ages 15-49) Estimated number of people living with HIV per 100 population of age group 15-49.

UNAIDS online database

## REFERENCES

Park BH, et al; "Associations of income and wealth with health status in the Korean elderly". J Prev Med Public Health (2009).

Pollack CE, et al; "Should health studies measure wealth? A systematic review". Am J Prev Med. (2007).

Roy K et al; "Influence of socioeconomic status, wealth and financial empowerment on the gender differences in health and healthcare utilization in later life: evidence in India". soc sci med (2008).



