UNIVERSITI TEKNOLOGI MALAYSIA

Semester 2020/2021

Subject : Probability & Statistical Data Analysis SECI2143

Section : 07

Project 2

Submission : 30/06/2021

Group name: *DATA GENERATION*

	Name	Matric number
1	TOYA LAZMIN KHAN	(A20EC0284)
2	ADNAN SHAFI	(A20EC0255)
3	G M SHAHEEN SHAH SHIMON	(A20EC0266)

Lecturer: Dr. Nies Hui Wen

Presentation video: https://youtu.be/IUxUOmyyPcs

Table of Contents

Life Expectancy of humans in developing countries

Project 2

Introduction

For completing the fulfillment of our project2, all of us got into finding the best dataset that will also match with the kind of tests we will be conducting and also, we will be able to learn something informative through it. After checking many datasets, we came across the life expectancy of humans in developing countries dataset. This dataset caught out eyes as all the group members in our course belonged to a developing country and we instantly chose this topic.

The title of our project is Life Expectancy of humans in developing countries. As per Roser et al., (2013) a person lives 72.6 years, but it is a world statistics . We were interested into knowing the life expectancy of the developing counties and what are the core factors that lead to the high life expectancy or low life expectancy of the developing countries.

Dataset

We made 2 excels files for our convenience. First, the dataset we used was Life expectancy of developing countries which was our population, and the second csv file was the sample we extracted from the Life expectancy of developing countries csv files. The sample we used for our project were 17 developing countries from North America.

Country	Country_C	Year	Continent	Region	Status	Income_Le	GNI	Life_Expec	Adult_Mor	Infant_Dea	Alcohol	Expenditur	Total_Exp	Measles	BMI	Polio	Diphtheria	GDP_Per_C	Population	thinness_	thinness_5	Income_Co	Sc
Algeria	DZA	2014	Africa	Northern A	Developing	Upper Mid	5490	75.4	11	21.9	0.01	357.8986	6.547214	0	58.4	95	95	5493.026	16.42216	6	5.8	0.741	
Angola	AGO	2014	Africa	Sub-Sahara	Developing	Upper Mid	5010	51.7	348	60.5	8.33	131.7519	2.434129	11699	22.7	68	64	5408.41	21.59338	8.5	8.3	0.527	
Argentina	ARG	2014	South Ame	Latin Amer	Developing	Upper Mid	12330	76.2	118	10.8	7.93	1089.5	8.258233	1	62.2	92	94	12334.8	15.70566	1	0.9	0.825	
Armenia	ARM	2014	Europe	Western A	Developing	Upper Mid	4140	74.6	12	13.5	3.91	406.5929	10.17827	13	54.1	95	93	3986.232	102.0801	2.1	2.1	0.739	
Azerbaijan	AZE	2014	Europe	Western A	Developing	High Incon	7690	72.5	119	25	0.01	432.8713	5.467744	0	51.5	97	94	7891.313	115.3488	2.8	2.9	0.752	
Bahamas (	BHS	2014	North Ame	Latin Amer	Developing	High Incon	28000	75.4	16	12.3	9.45	1689.116	5.891466	0	63.8	96	96	29445.03	38.17872	2.5	2.5	0.789	
Bahrain	BHR	2014	Asia	Western A	Developing	High Incon	24070	76.8	7	6.6	1.57	1100.991	4.406818	46	62.9	98	98	24989.4	1733.329	6.1	6	0.82	
Banglades	BGD	2014	Asia	Southern A	Developing	Lower Mid	1110	71.4	132	32.1	0.01	28.83413	2.496011	289	17.7	97	97	1118.854	1224.593	18.1	18.6	0.57	
Barbados	BRB	2014	North Ame	Latin Amer	Developing	High Incon	15630	75.4	1	13.3	8.82	1177.756	7.106533	0	53.7	95	94	16489.07	659.0349	3.8	3.7	0.793	
Belarus	BLR	2014	Europe	Eastern Eu	Developing	High Incon	7590	72	199	3.2	13.94	447.7013	5.387934	64	61.7	97	97	8318.513	46.69317	1.9	2	0.796	
Belize	BLZ	2014	North Ame	Latin Amer	Developing	High Incon	4250	70	177	13.9	6.58	273.1342	5.673932	0	5.1	95	95	4705.603	15.41841	3.4	3.4	0.705	
Benin	BEN	2014	Africa	Sub-Sahara	Developing	Lower Mid	1270	59.7	252	65.5	0.01	34.40216	3.645507	786	25.2	74	78	1291.41	91.22661	7.1	6.9	0.475	
Bolivia (Pl	BOL	2014	South Ame	Latin Amer	Developing	Lower Mid	2830	74	189	26.4	3.62	184.1956	5.964915	0	51.9	99	98	3081.879	9.749985	1.2	1.1	0.666	
Bosnia and	BIH	2014	Europe	Southern E	Developing	High Incon	5300	77.2	89	5.5	4.03	492.1351	9.475169	3000	55.3	86	86	5329.635	69.64848	2.4	2.4	0.742	
Botswana	BWA	2014	Africa	Sub-Sahara	Developing	High Incon	7520	65.1	268	31	0.01	419.4948	5.597932	1	37.3	96	95	7780.638	3.826466	6.7	6.4	0.697	
Brazil	BRA	2014	South Ame	Latin Amer	Developing	High Incon	12120	74.8	144	14.5	7.32	1009.829	8.396054	876	55.3	96	93	12112.59	24.43284	2.7	2.7	0.747	
Brunei Dar	BRN	2014	Asia	South-east	Developing	High Incon	43090	77.6	8	8.6	0.01	795.0394	1.914368	1	4.2	99	99	41726.78	78.1222	5.7	5.2	0.863	
Burkina Fa	BFA	2014	Africa	Sub-Sahara	Developing	Low Incon	760	59.3	268	61.1	0.01	39.59185	5.624916	343	18.8	91	91	792.8468	64.27623	8.2	7.7	0.398	
Burundi	BDI	2014	Africa	Sub-Sahara	Developing	Low Incon	260	59.1	297	48	0.01	20.43534	6.889277	0	18.2	95	95	274.8579	385.1943	7.4	7.3	0.404	
Cabo Verd	CPV	2014	Africa	Sub-Sahara	Developing	Lower Mid	3360	73	117	17.9	0.01	188.2107	5.327222	0	3.5	95	95	3588.674	130.6295	6.8	6.7	0.643	
Cambodia	KHM	2014	Asia	South-east	Developing	Low Incon	1020	68.3	179	28.6	0.01	73.29932	6.701577	0	18.9	87	88	1093.496	86.51025	1.9	11	0.553	
Cameroon	CMR	2014	Africa	Sub-Sahara	Developing	Lower Mid	1490	56.7	366	59.8	0.01	71.24244	4.534279	831	28.4	86	87	1542.62	47.04767	5.7	5.7	0.507	
Central Afi	CAF	2014	Africa	Sub-Sahara	Developing	Low Incon	390	58	437	92.1	0.01	18.30316	4.853249	210	22.1	47	47	424.4488	7.248053	8.4	8.3	0.345	
Chad	TCD	2014	Africa	Sub-Sahara	Developing	Low Incon	980	52.6	362	77.7	0.01	44.02822	4.273169	1275	18.7	44	37	1020.288	10.77624	8.7	8.5	0.39	
Chile	CHL	2014	South Ame	Latin Amer	Developing	High Incon	14940	83	83	6.9	7.16	1154.848	7.807299	0	63.1	95	95	14671	23.68936	0.8	0.8	0.841	
China	CHN	2014	Asia	Eastern As	Developing	High Incon	7510	75.8	86	9.9	5.78	361.7244	4.773227	52628	31.9	99	99	7678.599	145.3174	3.7	3	0.723	
Colombia	COL	2014	South Ame	Latin Amer	Developing	High Incon	8200	74.6	144	13.9	4.38	569.8973	7.14658	0	57.2	9	9	8114.084	43.07518	2.1	1.9	0.72	
Comoros (	COM	2014	Africa	Sub-Sahara	Developing	Lower Mid	1490	63.2	23	55.6	0.01	68.41656	7.584681	0	23.7	87	87	1513.83	408.0521	6.8	6.6	0.497	

Figure 1: Dataset of life expectancy of humans in developing countries 2014.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Country	Country_C	Year	Continent	Region	Status	Income_Le	GNI	Life_Expe	Adult_Mor	Infant_De	Alcohol	Expenditu	Total_Exp	Measles	BMI	Polio	Diphtheria	GDP_Per_	Population	thinness _	thinness_	Income_C	Schooling	Education	Unemployme
2	Bahamas	BHS	2014	North Ame	Latin Ame	Developin	High Incor	28000	75.4	16	12.3	9.45	1689.116	5.891466	0	63.8	96	96	29445.03	38.17872	2.5	2.5	0.789	12.6	0.796	13.8
3	Barbados	BRB	2014	North Ame	Latin Ame	Developin	High Incor	15630	75.4	1	13.3	8.82	1177.756	7.106533	0	53.7	95	94	16489.07	659.0349	3.8	3.7	0.793	15.3	0.812	12.17
4	Belize	BLZ	2014	North Ame	Latin Ame	Developin	High Incor	4250	70	177	13.9	6.58	273.1342	5.673932	0	5.1	95	95	4705.603	15.41841	3.4	3.4	0.705	12.8	0.71	8.244
5	Costa Rica	CRI	2014	North Ame	Latin Ame	Developin	High Incor	10090	79.5	96	7.9	3.45	837.9087	7.881752	1	59.5	91	91	10547.15	93.17617	1.7	1.7	0.768	13.9	0.785	9.059
6	Cuba	CUB	2014	North Ame	Latin Ame	Developin	High Incor	6840	79	93	4.3	4.37	856.0169	12.14123	0	6.7	99	99	7133.342	109.9766	3.4	3.3	0.772	14	0.764	2.7
7	Dominicar	DOM	2014	North Ame	Latin Ame	Developin	High Incor	6330	73.6	154	26.6	5.92	371.2853	5.847229	0	57.9	9	91	6608.828	215.3973	3.3	3.2	0.712	13.2	0.724	6.719
8	El Salvado	SLV	2014	North Ame	Latin Ame	Developin	Lower Mid	3400	73.3	181	13.7	2.52	276.2992	7.681528	0	55.4	93	94	3589.042	303.1462	1.6	1.6	0.676	13.2	0.66	4.154
9	Guatemal	GTM	2014	North Ame	Latin Ame	Developin	Lower Mid	3570	71.7	187	24.6	1.88	219.1399	5.942374	0	49.9	65	73	3779.642	148.5961	1.2	1.2	0.614	10.7	0.643	2.717
10	Haiti	HTI	2014	North Ame	Latin Ame	Developin	Lower Mid	1410	63.1	245	54.3	0.01	63.88166	7.797636	0	48.8	55	48	831.8739	383.6163	3.9	3.9	0.487	9.1	0.488	14.033
11	Honduras	HND	2014	North Ame	Latin Ame	Developin	Lower Mid	2000	74.5	149	17.1	2.87	183.2359	8.170272	0	5.2	97	97	2206.052	78.73104	2.1	2.1	0.618	11.4	0.606	5.488
12	Jamaica	JAM	2014	North Ame	Latin Ame	Developin	High Incor	4910	75.8	133	13.9	3.83	254.0866	5.232599	0	53.5	93	92	4833.79	264.274	1.8	1.7	0.727	12.8	0.72	13.74
13	Mexico	MEX	2014	North Ame	Latin Ame	Developin	High Incor	10500	76.6	122	14.4	5.26	595.8328	5.630386	3	62.8	87	87	10922.38	63.90164	1.6	1.5	0.754	13.1	0.754	4.809
14	Nicaragua	NIC	2014	North Ame	Latin Ame	Developin	Lower Mid	1830	74.5	148	17.3	3.55	158.2309	8.009811	0	53.2	99	98	1934.07	49.97505	1.8	1.7	0.636	11.6	0.64	4.519
15	Panama	PAN	2014	North Ame	Latin Ame	Developin	High Incor	11190	77.6	119	15	6.74	883.6766	6.910641	0	57.1	8	8	12796.07	52.51528	1.9	1.8	0.78	13	0.779	2.723
16	Saint Lucia	LCA	2014	North Ame	Latin Ame	Developin	High Incor	8840	75	139	18.2	9.97	433.3609	4.956199	0	46.7	99	99	9843.916	289.2148	4.3	4.3	0.723	13.1	0.728	23.124
17	Saint Vinc	VCT	2014	North Ame	Latin Ame	Developin	High Incor	6390	73.1	157	15.7	7.01	293.0375	4.403226	0	53.1	97	98	6684.807	280.4026	3.5	3.4	0.72	13.3	0.715	19.394
18	Trinidad a	TTO	2014	North Ame	Latin Ame	Developin	High Incor	18560	71.1	171	18.1	6.94	1060.558	5.227505	0	46	94	92	20270.86	264.0337	5.8	6	0.778	12.7	0.788	2.224
19																										
20																										
21																										

Figure 2: Dataset for the Life expectancy of humans in North America 2014

In first dataset, we have the mean life expectancy of all the developing countries in 2014. and in our second dataset, we have the mean life expectancy of developing countries in north America. In our datasets, we have information about many factors that effects the life expectancy. We have GNI, income level of the people, adult mortality, infant death, alcohol, expenditure in heath per capital, total expenditure of GDP, Vaccinations against measles, hepatitis B, polio, diphtheria, education index, schooling, unemployment level. There are many more data in our dataset, but we used only certain datasets to conduct our analysis.

From a certain source (provided in reference), we found out that an average human lives 72.6 years) worldwide. We used hypothesis testing 1 to find out if the mean Life expectancy worldwide, was different from the Life expectancy of developing countries. The variables we used for the test were the mean life expectancy of humans worldwide, mean life expectancy of humans in developing countries in North America. We used a 0.05 significance level to test this hypothesis. We thought the result would end up with us rejecting the null hypothesis because developing countries would have lower life expectancy then worldwide as they are still developing.

We used correlation to find relation between life expectancy and expenditure on health per capital. For this analysis, we were hoping to find a positive correlation. Through correlation we also found the relation between education index and adult mortality. For this analysis, we were optimistic we will find a negative correlation.

We ran a regression analysis to find if infant deaths have an impact on life expectancy. We were not expecting to not find an impact as infant deaths have diminished and as the dataset is of 2014 and infant deaths have lessened since the past few decades.

In our dataset we have two nominal data of “Income level” = “high income” and “Low income”. We want to test whether developing countries from our sample dataset has the income level with equal proportion .To find this out, we ran a chi squared test, which is called the goodness of fit

test on equal probabilities. We used the observed frequency of “high income” and “low income” of the people in North American developing countries. The level of significance used was 0.05.

Data Analysis

Hypothesis testing sample 1

We ran a 1 sample hypothesis test to find if the world population mean was different from the sample mean.

Variance is unknown so we find variance from R programming

The information about the variables is below:

World Population mean =72.6, Population mean = 74.1, Sample mean = 69.4,

Standard deviation = 3.82 (from r programming), sample number = 17

H0: $\mu = 0$

H1: $\mu \neq 0$

The formula used:

t = 1.59

critical value = -2.12, 2.12

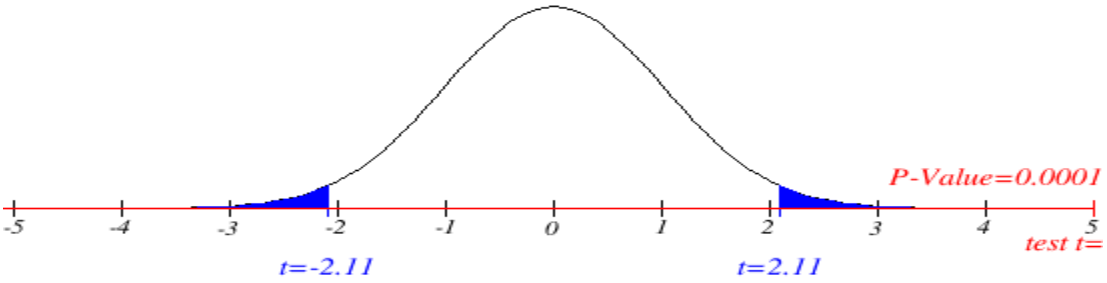

Figure 3 : Graph for the hypothesis t test

As our t value falls in the non rejection region, we cannot reject the null hypothesis. The results says that there is not enough evidence that the worldwide life expectancy is different from the sample life expectancy. So , The average life expectancy worldwide has no difference to the life expectancy of developing countries

Correlation

We conducted correlation test to find the relation between life expectancy and expenditure on health per capital.

The information about the variables is below:

X: life expectancy y: expenditure on health per capital

We found out that life expectancy and expenditure on health per capital has a positive weak correlation of 0.4389

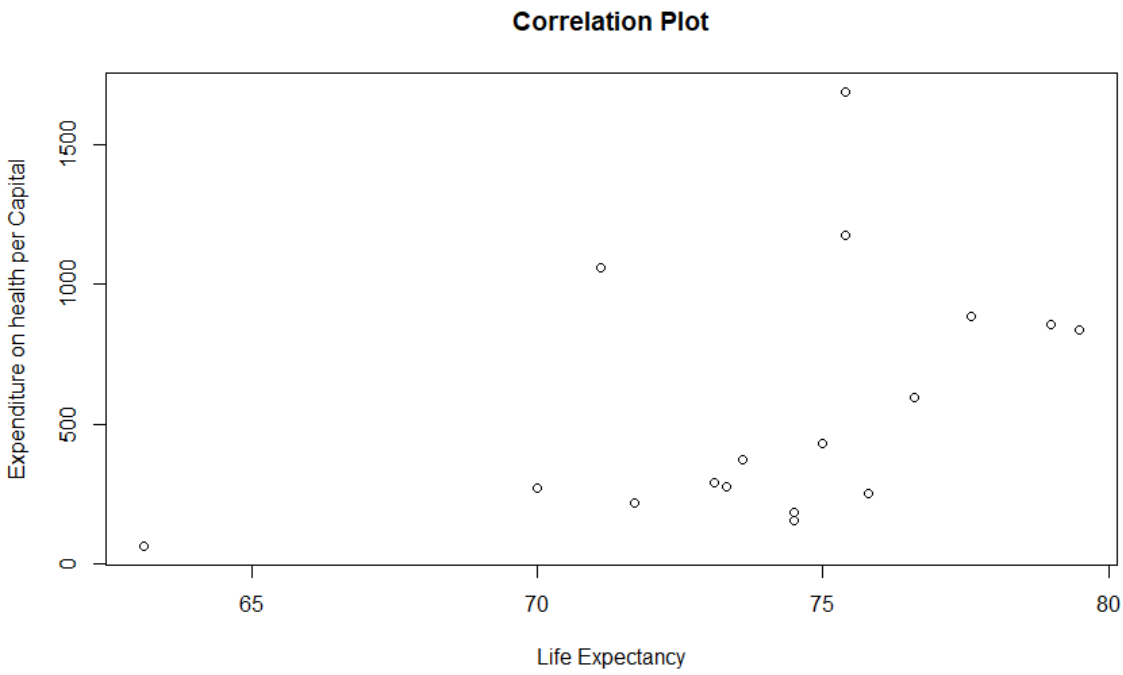

Figure 4: correlation of life expectancy vs expenditure on health per capital

We also conducted correlation test to find the relation between education index and adult mortality. The information about the variables is below:

x: Education_index y: Aduly_Mortality

We found that , Education and Adult mortality has a negative strong correlation which is -0.752981.

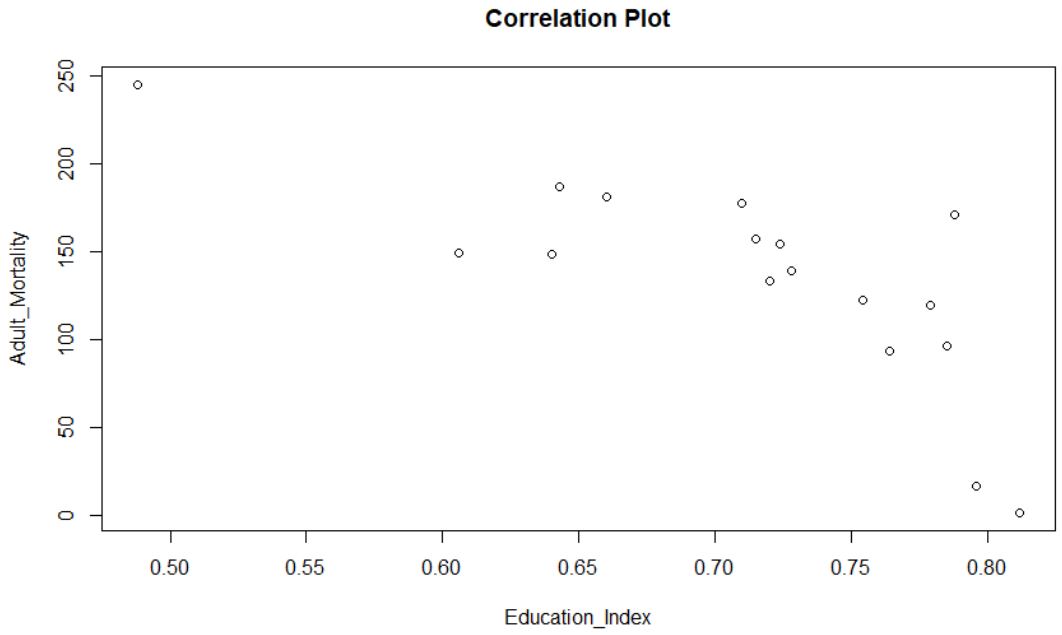

Figure 5: Education_index vs Aduly_Mortality

Regression

We conducted regression test to analysis to find if infant deaths have an impact on life expectancy.

And we were surprised to find out that infant death does have a impact on the adult mortality. We can conclude from the graph that life expectancy decreases in a country as more infant deaths happen.

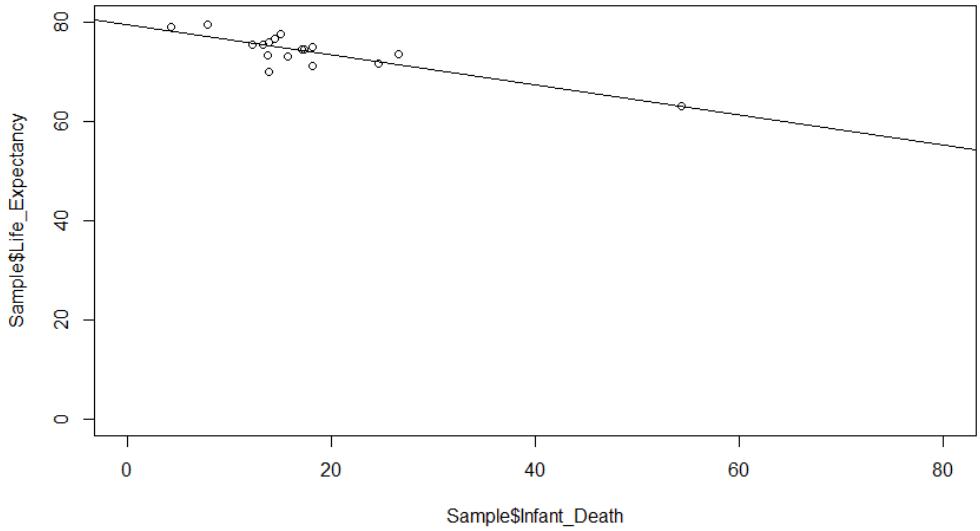

Figure 6: Infant_Death vs Life_expectancy

Goodness of fit (goodness of fit test)

We conducted the test to see whether the people of North American developing countries have the equal proportion of income level.

Income level	
High income	Low income
12	5

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

So,

Significance level(alpha)=0.05

H0: p1=p2

H1: At least 1 of the proportions is different from others.

Using the formula,

Chi square = 2.8824.

The critical region for chi square is 3.8415.

So, test statistic does not fall within the critical region, so we fail to reject the null hypothesis and we accept the fact that the people of North American developing countries have equal proportion of income level.

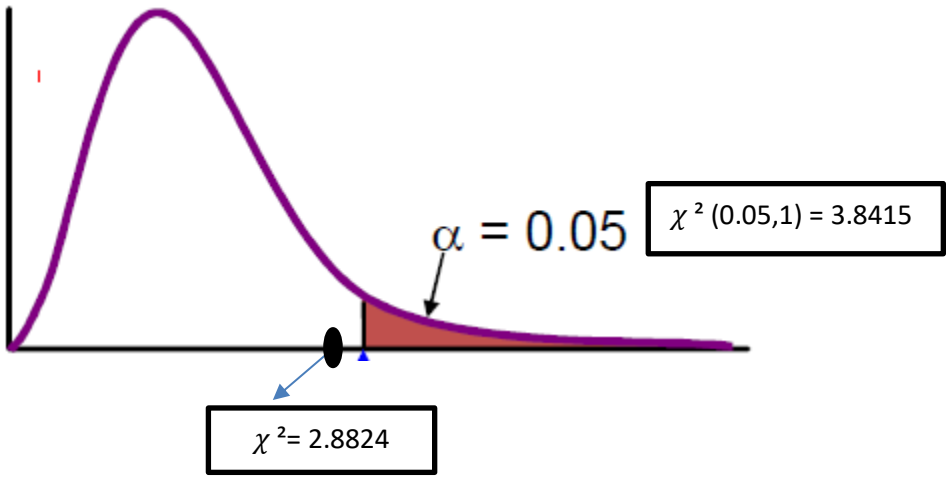

Figure 7: The chi square graph

$\chi^2 = 2.8824 < \chi^2(0.05,1) = 3.8415$ so fail to reject H0

Findings

After doing the data analysis we were surprised to find out that some of our expectations from the results did not match.

For the hypothesis test sample 1,

We found that the dataset life expectancy was not different from the worldwide life expectancy we found on the internet. We were confused about this finding but later we figured out that The world consists of three types of countries, developed , underdeveloped and developing. And so the average of life expectancies worldwide is close to developing countries because developing countries falls in the middle of both developed and underdeveloped countries.

For correlation tests ,

i) Between expenditure on health per capital and life expectancy, our expectations met , it showed that the more money spent on health , the better the life expectancy . It is true as frequent hospital visits and monthly check ups can find any hidden disease that might be affecting the body

ii) Between adult mortality and education, we found a negative correlation where we say that the more educated a country is the less people die . It might be because being

educated helps people take the right decisions about their health and also eat health and take care of themselves.

For regression test,

We found that infant deaths had an impact on the whole life expectancy . We were surprised at our findings as we thought we are in an age where infant mortality as diminished a lot but instead we see the infants deaths have cause an impact on life expectancy. Infant death is one cause for low life expectancy in a country.

For the chi squared test

We wanted to know difference in income can be a turning factor for life expectancies so we ran this test to see if the income of North American developing countries same or not.We found that it was the same so income of a country is not a factor for causing any changes in life expectancy. People of North American developing countries has the same proportion of income, high income and low

Conclusion

From this project we learned how to separate a certain amount of data from dataset , we learned a new programming language , which is called R programming . First , we had no idea about the language and thought it would not be seemingly possible to learn a programming language in such a short time but the language was so interesting that it took us one night to have a good prowess in it . Other than that, we were able to implement the data analysis processes practically than only learning it through text books .

Through our findings we learn that the world is progressing very quickly and life expectancy of people have been increasing in developing countries as developing country life expectancy is not much different from the worldwide life expectancy , thanks to the ever growing knowledge in science. All people in a country should be educated because it saves lives and people should try to go see a doctor every year if they can afford.

Infants' health care should be increased as thousands of infants die each year without even getting to see the world and effecting the life expectancy of a whole country.

Presentation Video Link

https://youtu.be/IUxUOmyyPcs

Appendix

We have attached the R script, the original dataset and also the modified sample dataset in our zip file .

References

Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2013, May 23). *Life Expectancy*. Our World in Data.

https://ourworldindata.org/life-

expectancy#:~:text=The%20United%20Nations%20estimate%20a%20global%20average%20life,Norway%20with%20a%20life%20expectancy%20of%2072.3%20years.