DISPARITY IN LIFE EXPECTANCIES IN THE UNITED STATES

Math 3007 (Spring 2018) Project Paper

Abstract

Globally and locally there are discrepancies in life expectancy. There are reasons behind these disparities, which may help uncover mistreatment of certain groups or help us increase quality of life. The first step to achieving this goal is locating the issues. We assessed whether people are expected to live longer in different regions of the United States. Using the dataset from the Global Health Data Exchange (GHDx), which contained the average life expectancies from 3142 counties across the United States, we apply tests for normality, variance, and independence. Then, we applied multiple t-tests to determine whether there was a disparity in life expectancy. It was found that the South has a shorter life expectancy than the North and West. In addition, we investigated possible factors for the disparity using linear regression testing. Using data from the United States 2010 Census, we found that life expectancy is not affected by living in rural areas. However, using another dataset from GHDx, we discovered that smoking affects life expectancy by one day per one percent increase in daily smokers. Future analysis of other possible factors for the disparity is needed to accurately outline the reasons behind the lower life expectancy in the South.

By: Federico Klinkert, Amy Feng, Eric Flores

Motivation

The motivation for this paper is to first determine if a certain region of the United States is expected to live longer than the others. The study will try to discuss the implications of why one region may tend to have a higher life expectancy. The goal will conclude with general challenges and questions life expectancy should address. Our approach to the life expectancy analysis focuses on conceptual and empirical work that shows the South tends to have a shorter life expectancy than the other regions with error involved.

Introduction

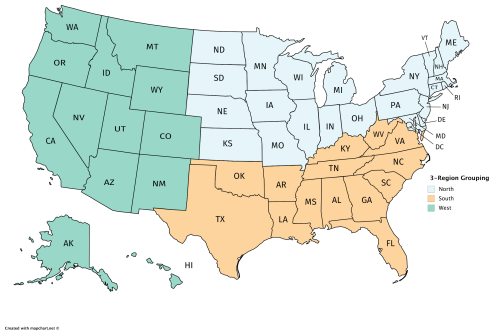
The United States is diverse in its people, resources, regulations, and more. However, the imbalanced distributions of these elements are impacting the US residents negatively. We try to investigate unfair treatments or conditions that are lowering the quality of life for groups in the United States. While there are many ways of going about uncovering these conditions, we started by trying to find the groups that might be unfairly treated. We used statistical procedures such as t-testing to analyze the life expectancy of the 3142 counties in the US. Although quality of life and life expectancies are not equivalent, we believe that the reasons behind low life expectancies are tied to reasons that lower quality of life. By observing if there are any disparities in life expectancy, we can pinpoint our focus on comparing these groups with the rest of the US. Using another procedure called linear regression, we can build a model that can help us predict the reasons behind lower life expectancy. Our results are the building blocks to improving the quality of life for the whole country.

Method and Statistical Tools

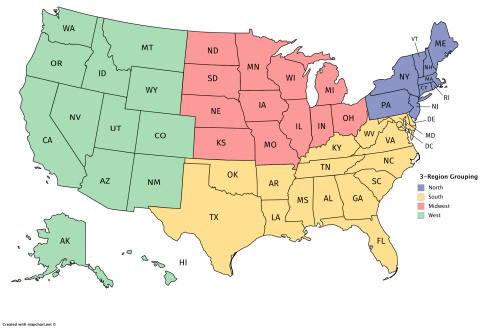
First, we had to organize the life expectancy data from the GHDx (Global Health Data Exchange) database. The data was originally made up of the life expectancy of every county from the United States from every five years starting from 1980 to 2014. (2014 is the exception since 2015 had not happened when the data was collected). We averaged the life expectancy of each year

for each county. Next, we grouped these averages into three regions: South, North, and West. Then we grouped the data into four regions: South, Northeast, Midwest, and West. By creating two different divisions of the United States, we had to apply all the procedures below twice.

3-Region Groupings (North, South, & West)



4-Region Groupings (Northeast, South, Midwest, West)



Next, we needed to check several assumptions before comparing the average life expectancy of each region using the t-test. First, we had to apply a D'Agostino and Pearson's k^2 test for normality on each region because this test is Python implementation of the normality test as opposed the chi-square test for normality that we learned in class. This test measures how far the skewness and kurtosis (size of tails) is from a normal distribution using the following formula for the test's statistic (the k-square test statistic)[1].

$$K^2 = Z_1(g_1)^2 + Z_2(g_2)^2$$

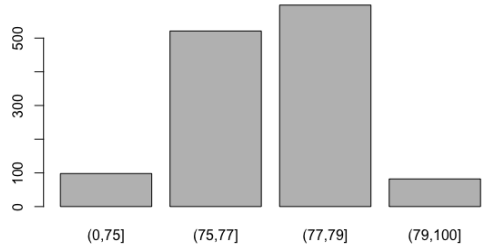
g_1 is the measurement of skewness
g_2 is the measurement of kurtosis
Z_1 is a transformation to make the g_1 statistic close to standard normal as possible
Z_2 is a transformation to make the g_2 statistic close to standard normal as possible

K^2 is approximately Chi^2 with two degrees of freedom, so we can use the Chi-Square table of p-values to determine the K^2's p-value. If this p-value is lower than our significance level we can reject the null hypothesis, which states that the distribution is normal, and assume our alternative hypothesis, which states that the distribution is not normal, is true.

[1] We are unable to explain fully D'Agostino and Pearson's k-square test for normality due to time limitation. Please read the references cited for more information.

If the normality test stated that some populations were not normal, we would need to observe the PDF's (Probability Distribution Functions) of each of the seven regions. We would make sure that the PDF's were not heavily skewed to one direction. Also, we would make sure that the sample size of each population was large enough against non-normality. We defined large enough to mean over 200 samples. Then, we need to ensure that there are no extreme outliers by using boxplots. If there are, we would need to justify removing them to continue with our testing. Once these observations are met, we can assume that the regions are approximately normal due to the robustness of the t-test. If the observations were not met, then we would have to use another dataset or find alternatives to the t-test.

Then, we must check if each pair of regions in the three and four division are independent or dependent. If they are dependent, we must use a dependent/paired t-test. If they are independent, we can apply the two sample t-test. We then proceeded to apply a chi-squared test for independence. The chi-squared test for independence has as null hypothesis that the data is independent, and the alternative hypothesis is that the data is not independent. All the tests were performed in pairs to account for the independence between regions. In order to perform the test, we had to divide the data into the 'correct' bins. The distribution of the bins have to describe the actual distribution on the data. The chi square test for independence has as null hypothesis that the data is independent, and the alternative hypothesis is that the data is not independent. The North region was divided into bins that cover a range of 2 years, and from there all the other regions were grouped into the same bins. The bins for the north region had the following histogram, which captured the actual distribution.



The expected count were calculated using the formula $p_{ij} = P(R_i)P(C_j)$, and the statistic was calculated using the formula $d_1 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$

Another procedure that needs to done before applying multiple independent[2] t-tests on those pairs of regions is checking if the variances are equal or unequal. If the variances are equal, we apply Student's t-test. If the variances are unequal, we apply Welch's t-test for that single pair of regions. We apply Levene's test to check the homogeneity of variances. The null hypothesis is all the variances are equal to each other. Alternatively, at least of the variances is different. First, we have to check our assumptions, which are approximately normal and independent distributions. Also, the Levene test is robust against non-normality.

Next, we will evaluate the test statistic, W, using the following formula.

$$W = \frac{(N-g)\sum_{k=1}^{g} n_k (Z_k - \bar{Z})^2}{(g-1)\{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(Z_{ki} - \bar{Z}_k)^2\}}$$

where

$$Z_{ki} = |Y_{ki} - \bar{Y}_k|$$

$$\bar{Z}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} Z_{ki}$$

$$\bar{Z} = \frac{1}{N}\sum_{k=1}^{g}\sum_{i=1}^{n_k} Z_{ki}$$

$$\bar{Y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} Y_{ki}$$

g= # of groups with possible different means and standard deviations
nk = # of subjects in each group
Yki = mean age in each county
Yk = Sample Mean
Zk = group means
Zki = overall mean

We determined whether to use the mean of the i-th subgroup, the median of the i-th subgroup, or the trimmed mean of the i-th subgroup, which determines the robustness and power of the test. We can use the trimmed mean if the graphs were Cauchy distributions. We can use the median if the

[2] Note that we do not have to check if the variances are equal for a dependent t-test because in the paired sample t-test, we are using the sample standard deviation of the differences between every pair in the two populations.

graphs resembled chi-squared distributions. Lastly, we can use the mean if our graphs were moderately-tailed. The test is dependent upon the underlying distribution. We reject the null hypothesis if W > F (upper critical value of F distribution with k-1 and N-k degrees of freedom at an alpha = 0.01 significance level, with N = number of subjects in all groups and k = number of groups). If the variances were not equal in the three populations and the four populations (we reject the null hypothesis in both cases), we conclude with testing the pairs of variances defined by the two-sample Levene test or *F* procedure. We will use the F procedure if the Levene results are inconsistent and fail.

The two-sample *F* procedure takes independent samples from two populations. If we have two independent populations, the ratio is also independent, so F has the F distribution with m-1 and n-1 degrees of freedom, where m and n are population sizes. The null hypothesis is that the variances are equal, which is similar to the Levene test. The P-value is defined by using ratios of the sample variances.

$$P = P\left(\frac{S_Y^2}{S_X^2} \leq \frac{s_Y^2}{s_X^2} \,\middle|\, \sigma_X^2 = \sigma_Y^2\right) = P\left(\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq \frac{s_Y^2}{s_X^2}\right) = P(F \leq f),$$

The P value will determine which variances can be assumed to be equal amongst the pairs of regions.

After calculating which versions of the t-test to apply, we started to apply t-tests between every pair of regions in their division. The dependent t-test for paired samples is similar to the one-sample t-test, but we are using the differences between the two populations as the one-sample. To calculate the t-statistic we use the following formula,

$$t = \frac{\bar{X}_D - \mu_0}{\frac{s_D}{\sqrt{n}}}.$$

X_D is the mean of the differences between the pairs
S_D is the standard deviation of the differences between the pairs
μ_0 is the true mean of the differences between the pairs
n is the number of pairs

Our null hypothesis in the test is that $\mu_0 = 0$. Our first alternative hypothesis in the test is that $\mu_0 > 0$. Our second alternative hypothesis in the test is that $\mu_0 < 0$. Our degrees of freedom is n - 1.

Using the degrees of freedom and the t statistic, we can use the t-test p-value table to find our p-value. If our p-value/2 is less than our significance level, we can reject the null hypothesis and say that there is a difference between the life expectancy of the two populations. Otherwise, we say that there is no difference between the life expectancies. If we reject the null and the t-statistic value is greater than 0, we know that the first population's mean is larger than the second population's mean. If we reject the null, but the t-statistic is less than 0, then the first population's mean was less than that of the second population.

If our pairs of regions are independent and our variances are equal, we calculate our t-tests using the following formula,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}.$$

X_1 is the mean of the first population
X_2 is the mean of the first population
S_X is the standard deviation of the first population
S_Y is the standard deviation of the first population
n_1 is the sample size of the first population
n_2 is the sample size of the first population

Our null hypothesis is that the two regions have the same mean. Our first alternative hypothesis in the test is that the first population's mean is larger than the second population's mean. Our second hypothesis in the test is that the first population's mean is less than the second

population's mean. Our degrees of freedom is $n_1 + n_2 - 2$. Now the rest of the procedure of comparing the p-values and t-statistic follows the same outline as the dependent t-test.[3]

Next, if our two populations are independent, but the variances are unequal, we apply the Welch's t-test. The null and the alternative hypothesis is the same as the two-sample t-test. The procedure is almost exactly the same as the two-sample independent t-test, but we use the following formula for the t-statistic and the degrees of freedom,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

where

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \qquad \mathrm{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

X_1 is the mean of the first population
X_2 is the mean of the first population
S_1 is the standard deviation of the first population
S_2 is the standard deviation of the first population
n_1 is the sample size of the first population
n_2 is the sample size of the first population

Lastly, we combined the conclusions from the t-tests into one observation about the three division and another about the four division.

Because we applied multiple tests for our assumptions, we had to calculate our type I error, so that we can understand the validity of our results. To calculate our type I error across multiple tests, we calculated the Familywise Error Rate (the formula is listed below).

$$FWER = 1 - (1 - \alpha)^c$$

where α = the significance level used for each test, c = the number of tests applied

If our error is large, we need to consider removing the tests needed for meeting the assumptions. In addition, we can apply the Bonferroni Correction to lower the type I error at the

[3] See two paragraphs above to see the outline of the rest of the procedure.

expensive of our power. The correction requires us to change our significance level to our original significance level divided by the number of tests we applied. Therefore, to reject test T_i

$$P(T_i \ passes \mid H_o) \ \leq \ \frac{a}{n} \ \ for \ 1 \leq i \ \leq n$$

where α = the original significance level used for each test,
n = the number of tests applied,
H_o is the null hypothesis for test T_i

After the significance level is changed, we must reevaluate t-test p-values to see whether our conclusions have changed.

If any discrepancies have been found, we can formulate a linear regression model to help explain the disparities in the US life expectancy. Since previous tests concluded that the rest of the country has the same mean and variance, we proceed to group the rest of the country except for the south. We found data on the percentage of people living in a rural area and the percentage of people that smoke daily for over 3,000 counties. Two bivariate linear regressions were applied based on the following regression line:

$$LE \ = \ \beta_0 + \beta_1 smok \ + \beta_2 rur \ + u$$

LE: life expectancy
β_0: intercept (life expectancy not affected by smoking and rural population)
β_1: slope that accounts for the effect of smoking
β_2: slope that accounts for the effect of living in a rural area

In order to be able to conduct the bivariate linear regression, the following assumptions have to be met:

Linear relationship: the regression has to be linear in coefficients; this regression is going to be performed through a F-test comparing the current model (unrestricted) and a model with just an intercept (restricticted). The F-statistic is distributed $F_{1,n}$ can be calculated as

$$F \ = \ \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/df_{ur}}$$

The null hypothesis indicates that the model is not linear in coefficients and the alternative hypothesis concludes that the model is in fact linear and the regression can be considered to provide a relationship between the regressors and the explained variable.

Constant variance in the residual terms: the OLS assumes that the variance of the error terms is constant across time and equal to σ^2. If the variance is not constant, the regression will be unbiased, but it will not be efficient. The standard errors of the coefficient will be affected and the model will yield the incorrect p-values for the coefficients. In order to check if the variance of the linear regression is constant, the Breusch-Pagan test will be applied. The Breusch-Pagan test runs an auxiliary regression on the residual terms of the original linear regression. The R^2 multiplied by n will provide the coefficient of determination and it will be χ^2_k distributed, where k is the number of regressors.

$$\widehat{u} = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_k x_k + v$$

$$H_0: \gamma_0 = \gamma_1 = \ldots = \gamma_k = 0$$

$$H_1: \textit{The variance is not constant}$$

In the case the variance is not constant, the Eicker's heteroscedasticity-consistent estimator has to be applied to obtain efficient standard errors. The variance of the each coefficient equals:

$$\widehat{Var(x)} = \frac{1}{n}\sum_{i=i}^{n} X_i X'_i \widehat{u}^2_i$$

Significance of Coefficients: We then can proceed to test if the coefficients are statistically significant. If a coefficient is not statistically significant, a new regression without that variable should be performed since it does not have any effect on life expectancy.

Residual Analysis: After all the past assumptions are met, we need to check the residuals. The residuals have to be normally distributed with zero mean. This part is going to be done graphically by looking at the PDF of the residuals.

Findings[4]

Note: All finding below use a significance level of 0.01.

Grouping Results:

Sample Size

Region-3	Sample Size	4-Region	Sample Size
North	1299	**Northeast**	217
South	1395	**South**	1422
West	448	**West**	448
		Midwest	1055

When choosing the divisions of the United States, we had to be cautious of the sample size of each region. We did not want our sample size for any region to be too small or else we could not use that division. In addition, we struggled to find balanced sized divisions that were interesting and intuitive ways of dividing the United States. However, despite our desire for more equal regions, we found that the three and four divisions we choose to be large enough sizes and informative ways of dividing the United States.

Normality Results:

K^2 Test Statistics

3-Region	K^2 Statistic	4-Region	K^2 Statistic
North	527.09	**Northeast**	23.50
South	6.46	**South**	5.71

[4] Because statistical tests are dependent on the data and have a chance of being wrong, when we make our conclusion on any test, we acknowledge that we can only say "there is evidence for our conclusion".

West	61.76	**West**	61.76
		Midwest	425.79

P-Values

3-Region	P value	4-Region	P value
North	3.50e-115	**Northeast**	7.91e-06
South	0.04	**South**	0.058
West	3.89e-14	**West**	3.886e-14
		Midwest	3.47e-93

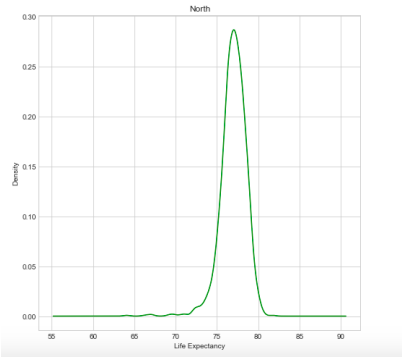
Conclusion on Normality:

3- Region	Normality	4-Region	Normality
North	No	**Northeast**	No
South	Yes	**South**	Yes
West	No	**West**	No
		Midwest	No

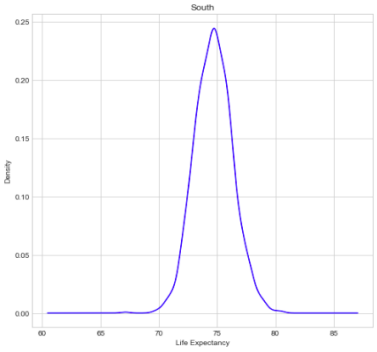
Another struggled that we faced was that only the south had a normal distribution, which means we should not use the t-tests. However, if our sample sizes are large, our distributions are not skewed, and there are no extreme outliers, we can continue with the t-test because the t-test is robust under non-normality as long as these conditions are met. From the first table, since all the sample sizes for each region is larger than 200, which we define as large, this sample size condition has been met.

3-Region PDFs

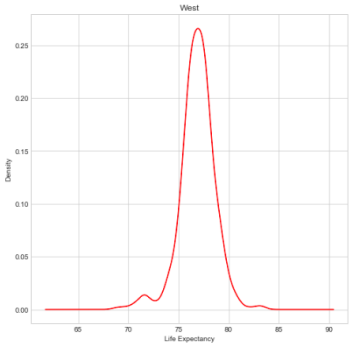
NORTH



SOUTH

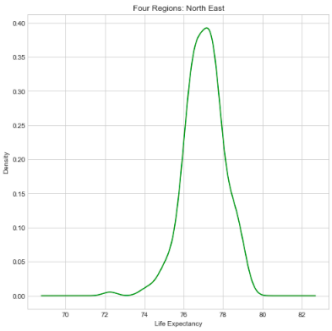


WEST

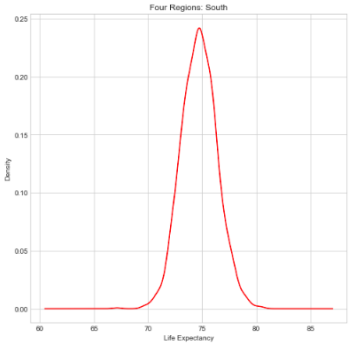


4-Regions PDFs

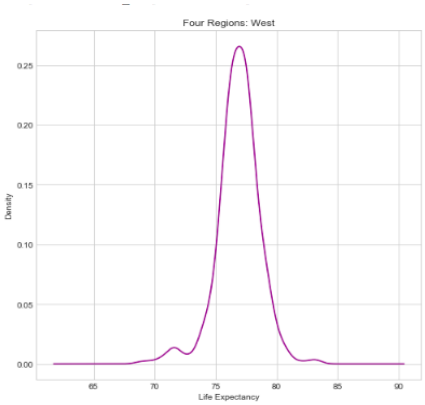
NORTH



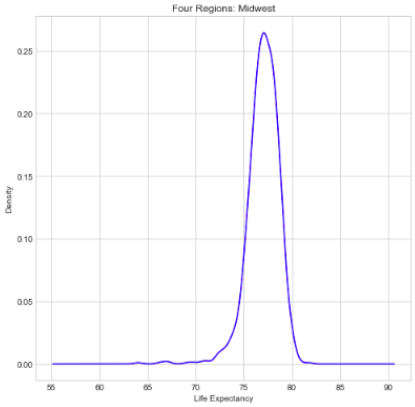
SOUTH



WEST

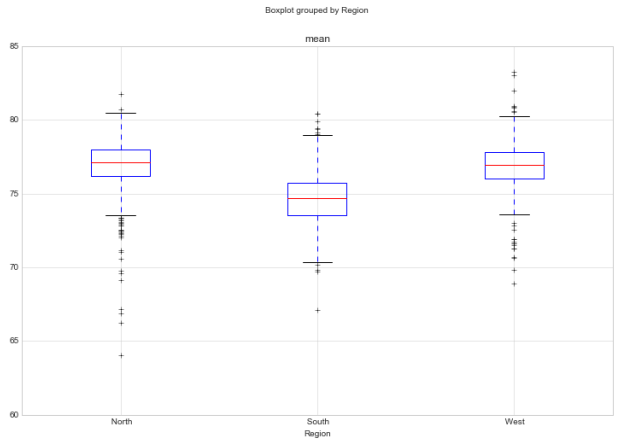


MIDWEST



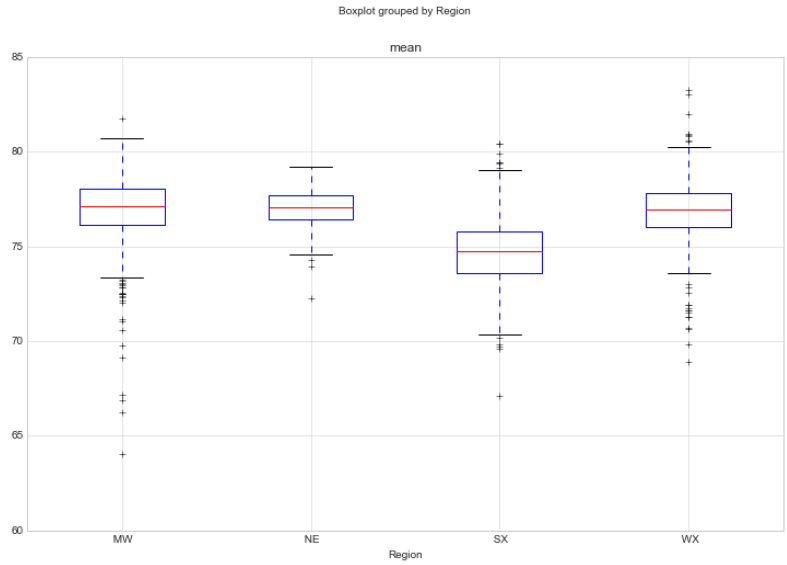
We can see that the PDFs for each region are not skewed and almost symmetric. Therefore, the shape condition has been met.

3-Regions Boxplots



North South West

4-Regions Boxplots



Midwest Northeast South West

We can see outliers from the boxplots (the black tick marks). However, all the values ranged from 60 to 85 including the outliers. Because our data is so concreted in a few values in the 70s, we have many outliers. However, we did not believe that any of our outliers were extreme because they were at most 10 years different from most means, which were most commonly 75.

Thus, our outlier condition has been met. Now, we can assume that our distributions are approximately normal and using t-tests are still appropriate.

Variance Results:

3 Regions

W Statistic and P-Value for 3 Regions

W Statistic = 7.28587	P-Value = 0.00069
F Critical Value = 4.612	Conclude: Reject null hypothesis, variances are not equal

Two-Sample Levene Tests

Comparison	P	Conclusion
North vs. South	0.000157497572208	Var(north)=/Var(south)
North vs. West	0.0314788294262	fail to reject Var(north)=Var(west)
West vs. South	0.73131395576	fail to reject Var(west)=Var(south)

Conclude: Levene tests yielded inconsistent results.

Two-Sample *F* Tests

Comparison	P	Conclusion
North vs. South	5.34E-244	Var(north)=/Var(south)

North vs. West	0.161240009	fail to reject Var(north)=Var(west)
West vs. South	3.31E-111	Var(west)=/Var(south)

4 Regions

W Statistic and P-Value for 4 Regions

W Statistic = 15.18210	P-Value = 8.28333e-10
F Critical Value = 3.788	Conclude: Reject null hypothesis, variances are not equal

Two-Sample Levene Tests

Comparison	P Value	Conclusion
Northeast vs South	3.85438703105e-13	Var(NE)=/Var(S)
Northeast vs West	5.75361397964e-08	Var(NE)=/Var(W)
Northeast vs Mid-West	1.60299521133e-07	Var(NE)=/Var(MW)
South vs West	0.58372772807	Var(S)=Var(W)
South vs Mid-West	0.0250332802403	Var(S)=Var(MW)
West vs Mid-West	0.31814798095	Var(W)=Var(MW)

We conclude that both groups have different variances. For the group of four populations, the Northeast is different from the rest of the populations. Whereas in the group of three populations, the South had a different variance from the rest of the populations. There was an inconsistency in

the Levene test used in the three regions, and we needed to perform the F test to see if we could gather any other information.

Independence Results:

3-Regions: for the 3-regions division, all of the test have the same degrees of freedom.

$df = (r - 1)(c - 1) = (4 - 1)(2 - 1) = 3$

$\chi^2_{3,0.01} = 11.3$

Life Expectancy	Observed North	Expected North	Observed South	Expected South
$(0, 75]$	98	438.47	811	470.53
$(75, 77]$	521	479.47	473	514.53
$(77, 79]$	598	338.13	103	362.87
$(79, \infty)$	82	42.93	7	46.07

$d_1 = 972.18$

We conclude that there is some relationship between geographical location and life expectancy in the North and South regions.

Life Expectancy	Observed West	Expected West	Observed South	Expected South
$(0, 75]$	41	207.22	811	644.78
$(75, 77]$	195	162.47	473	505.53
$(77, 79]$	174	67.37	103	209.63
$(79, \infty)$	38	10.94	7	34.06

$d_1 = 496.17$

We conclude that there is some relationship between geographical location and life expectancy in the West and South regions.

Life Expectancy	Observed North	Expected North	Observed West	Expected West
$(0, 75]$	98	103.35	**41**	35.65

$(75, 77]$	521	532.39	**195**	183.61
$(77, 79]$	598	574.03	**174**	197.97
$(79, \infty)$	82	89.23	**38**	30.77

$d_1 = 8.22$

We conclude that there is not a relationship between geographical location and life expectancy in the North and West regions.

Overall, the North and West regions are independent while all other test concluded that the data is dependent and there is some relationship between geographical location and life expectancy across regions.

4-Regions: for the 4-regions division, all of the test have the same degrees of freedom.

$df = (r - 1)(c - 1) = (4 - 1)(2 - 1) = 3$

$\chi^2_{3,0.01} = 11.3$

Life Expectancy	**Observed North**	**Expected North**	**Observed West**	**Expected West**
$(0, 75]$	6	15.4	41	31.6
$(75, 77]$	99	95.9	195	198.1
$(77, 79]$	108	92.0	174	189.0
$(79, \infty)$	4	13.7	38	28.3

$d_1 = 22.9$

We conclude that there is some relationship between geographical location and life expectancy in the North and West regions.

Life Expectancy	**Observed North**	**Expected North**	**Observed Midwest**	**Expected Midwest**
$(0, 75]$	6	15.9	87	77.1
$(75, 77]$	99	86.3	407	419.7
$(77, 79]$	108	101.0	484	491.0

$(79, \infty)$	4	13.8	77	67.2

$d_1 = 18.6$

We conclude that there is some relationship between geographical location and life expectancy in the North and Midwest regions.

Life Expectancy	**Observed South**	**Expected South**	**Observed West**	**Expected West**
$(0, 75]$	817	652.4	41	205.6
$(75, 77]$	488	519.4	195	163.6
$(77, 79]$	109	215.2	174	67.8
$(79, \infty)$	8	35.0	38	11.0

$d_1 = 486.8$

We conclude that there is some relationship between geographical location and life expectancy in the West and South regions.

Life Expectancy	**Observed South**	**Expected South**	**Observed Midwest**	**Expected Midwest**
$(0, 75]$	817	519.0	87	385.0
$(75, 77]$	488	513.8	407	381.2
$(77, 79]$	109	340.4	484	252.6
$(79, \infty)$	8	48.8	77	36.2

$d_1 = 854.4$

We conclude that there is some relationship between geographical location and life expectancy in the Midwest and South regions.

Life Expectancy	**Observed West**	**Expected West**	**Observed Midwest**	**Expected Midwest**
$(0, 75]$	41	38.2	87	89.8
$(75, 77]$	195	179.4	407	422.6
$(77, 79]$	174	196.1	484	461.9
$(79, \infty)$	38	34.3	77	80.7

$d_1 = 6.36$

We conclude that there is not a relationship between geographical location and life expectancy in the Midwest and West regions.

Applying the test for the North and South regions yielded a expected observation smaller than 5, and the last 2 binds had to be combined.

$df = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$

$\chi^2_{2,0.01} = 9.21$

Life Expectancy	Observed North	Expected North	Observed South	Expected South
$(0, 75]$	6	108.4	817	710.6
$(75, 77]$	99	77.7	488	509.3
$(77, \infty)$	112	30.9	117	202.1

$d_1 = 375.73$

We conclude that there is some relationship between geographical location and life expectancy in the North and South regions.

Overall, the West and Midwest regions are independent while all other test concluded that the data is dependent and there is some relationship between geographical location and life expectancy across regions.

T-Test Results:

If the populations were *independent* and variances were *equal*,

T-Statistic

3-Region Comparison	T Statistic	4-Region Comparison	T Statistic
North vs South	27.41	**Northeast vs South**	18.89
North vs West	0.73	**Northeast vs West**	0.46
South vs West	-22.90	**Northeast vs Midwest**	4.56
		South vs West	-13.79
		South vs Midwest	-14.44

	West vs Midwest	2.64

P-Value/2

3-Region Comparison	P-value/2	4-Region Comparison	P-value/2
North vs South	8.00e-121	**Northeast vs South**	9.47e-59
North vs West	0.23	**Northeast vs West**	0.32
South vs West	5.31e-92	**Northeast vs Midwest**	3.35e-06
		South vs West	2.14e-36
		South vs Midwest	3.90e-39
		West vs Midwest	0.0043

If the populations were *independent* and the variances were *unequal*,

T-Statistic

3-Region Comparison	T Statistic	4-Region Comparison	T Statistic
North vs South	27.41	**Northeast vs South**	18.89
North vs West	0.73	**Northeast vs West**	0.46
South vs West	-22.90	**Northeast vs Midwest**	4.56
		South vs West	-13.79
		South vs Midwest	-14.44
		West vs Midwest	2.64

P-Value/2

3-Region Comparison	P-value/2	4-Region Comparison	P-value/2
North vs South	8.03e-120	**Northeast vs South**	5.50e-56
North vs West	0.23	**Northeast vs West**	0.32
South vs West	9.93e-92	**Northeast vs Midwest**	3.37e-06
		South vs West	3.22e-36
		South vs Midwest	1.63e-38
		West vs Midwest	0.0043

If the populations were *dependent*

T-Statistic

3-Region Comparison	T Statistic	4-Region Comparison	T Statistic
North vs South	26.21	**Northeast vs South**	17.62
North vs West	0.69	**Northeast vs West**	0.43
South vs West	-24.50	**Northeast vs Midwest**	4.79
		South vs West	-14.84
		South vs Midwest	-14.26
		West vs Midwest	2.61

P-Value/2

3-Region Comparison	P-value/2	4-Region Comparison	P-value/2

North vs South	1.036e-92	**Northeast vs South**	5.65e-44
North vs West	0.24	**Northeast vs West**	0.33
South vs West	5.79e-85	**Northeast vs Midwest**	1.53e-06
		South vs West	4.21e-35
		South vs Midwest	2.85e-33
		West vs Midwest	0.0049

Conclusion on T-Tests

3-Region Comparison	Mean Conclusion	4-Region Comparison	Mean Conclusion
North vs South	Reject. North>South	**Northeast vs South**	Reject. Northeast > South
North vs West	No Reject. North=West	**Northeast vs West**	No Reject. Northeast = West
South vs West	Reject. South<West	**Northeast vs Midwest**	Reject. Northeast > Midwest
		South vs West	Reject. South < West
		South vs Midwest	Reject. South < Midwest
		West vs Midwest	Reject. West > Midwest

For all the possibilities of t-tests (unequal variance, equal variance, independent, or dependent), our conclusions are the same. For the three regions, we can conclude that the South has the lowest life expectancy. The North and West have the same life expectancy. For the four regions, the South had the lowest life expectancy again. The Midwest had the second lowest. The Northeast and the West tied for the highest life expectancy.

Error Analysis Results:

We did a total of 12 tests for the three regions (3 Normality Tests + 3 Independence Tests + 3 Variance Tests + 3 T-Tests). We did a total of 22 tests for the four regions (4 Normality Tests + 6 Independence Tests + 6 Variance Tests + 6 Normality Test). Using the FWER formula, our error is 11.36% for the three regions and 19.84% for the four regions.

Originally, we planned to use the ANOVA test instead of multiple t-tests to lower our error because we could have used two ANOVA tests instead of 9 t-tests. However, the ANOVA test did not fit our problem. The ANOVA does not give specific information about which region would have a lower life expectancy. The test would only test whether there was a difference in at least one of the region's life expectancies.

With such a large error and the need to apply multiple t-tests, we were faced with the challenge of reducing our error. We lowered the error by observing that we did not need to do a normality test to reach our conclusions. We can simply observe the PDFs, outliers, and sample size to see that the regions are approximately normal. In addition, we did not need the results for the variance and independence tests because there was a very small change to the p-value and t-test statistic for whichever one of the t-test we applied. With these observations, we can see that our error should be lower and that the new error is 2.97% for the three regions and 5.85% for the four regions. However, these error rates are still much higher than our original target of 1%. We applied the Bonferroni Correction. Using this correction, we lower our significance level for every three region test to 0.01/3 and for every four region test to 0.01/6. Now, our FWER is 0.01.[5]

Since the significance level for all the tests have changed, we have to reanalysis our p-values to see if any conclusions have changed. We did find a change between the West and Midwest because the p-value is 0.0049, at 0.01 significance level, we could reject the null hypothesis. However, using our new significance level of 0.01/6, this p-value is larger so we cannot reject the null hypothesis. Therefore, our new conclusion is that the south is the only region with a lower life expectancy than the other regions. We should be cautious about this change in conclusion. Lowering the type I error may have allowed us to reach the correct conclusion;

[5] The proof for why the new error FWER is equal to the original significance after the correction can be found here.

however, since the Bonferroni Correction does not protect against the loss of power, we might have come to an incorrect conclusion because our power was too low to reject the null hypothesis.

Linear Regression Results:

The bivariate linear regressions gave the following output:

South: $\widehat{LE} = \widehat{80.654} - \widehat{0.0028}smok - \widehat{0.0003}rur$

Rest of the country: $\widehat{LE} = \widehat{82.53} - \widehat{0.0029}smok - \widehat{0.0002}rur$

First, We need to check for linearity in the coefficients.

Region	Distribution	F-statistic	p-value
South	$F_{1,1387}$	327.8	0.000
Rest of the country	$F_{1,1732}$	650.7	0.000

In both cases, we can reject the null hypothesis and conclude that both models are linear in coefficients. We then can proceed to test for constant variance using the Breusch-Pagan test.

Region	Distribution	BP-statistic	p-value
South	χ^2_2	0.4443	0.8008
Rest of the country	χ^2_2	49.363	0.0000

The Breusch-Pagan test concludes that the variance of the error terms in the South is constant, but it is not across the rest of the country. The Eicker's heteroskedasticity consistent estimators have to be applied to obtained the correct standard errors to perform t-test on the coefficients. We can proceed to analyse the significance of coefficients.

South			
Coefficient	Standard Error	t-value	p-value
β_0	0.1776	454.16	0.000

β_1	0.0001	-22.58	0.000
β_2	0.0002	-1.66	0.097

Rest of the country			
Coefficient	Standard Error	t-value	p-value
β_0	0.1737	475.26	0.000
β_1	0.0002	-18.36	0.000
β_2	0.0002	1.23	0.217

In both cases, the slope of the impact of rural population on life expectancy is not statistically significant, and can be taken out of both regressions since it does not have any effect on life expectancy.

We have to perform simple linear regressions on life expectancy and smoking:

South: $\widehat{LE} = \widehat{80.62} - \widehat{0.0029}smok$

Rest of the country: $\widehat{LE} = \widehat{82.56} - \widehat{0.0029}smok$

First, We need to check for linearity in the coefficients.

Region	Distribution	F-statistic	p-value
South	$F_{1,1388}$	652.1	0.000
Rest of the country	$F_{1,1733}$	1298	0.000

In both cases, we can reject the null hypothesis and conclude that both models are linear in coefficients. We then can proceed to test for constant variance using the Breusch-Pagan test.

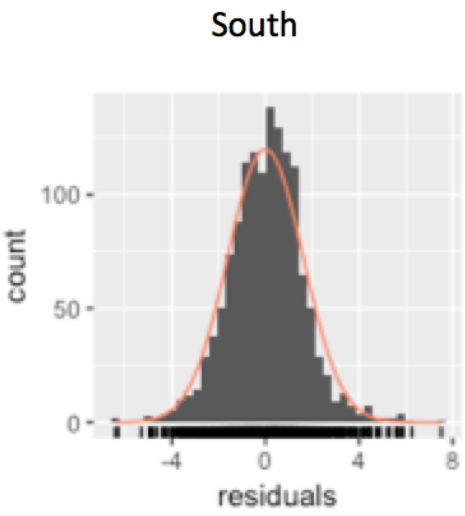
Region	Distribution	BP-statistic	p-value
South	χ^2_1	0.2811	0.596
Rest of the country	χ^2_1	44.472	0.0000

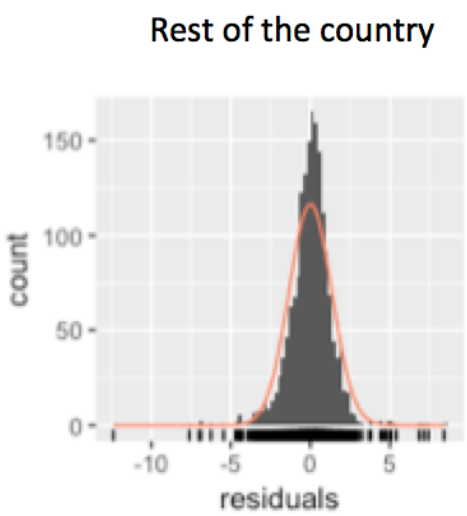
The Breusch-Pagan test concludes that the variance of the error terms in the South is constant, but it is not across the rest of the country. The Eicker's heteroskedasticity consistent estimators have to be applied to obtained the correct standard errors to perform t-test on the coefficients. We can proceed to analyse the significance of coefficients.

South			
Coefficient	Standard Error	t-value	p-value
β_0	0.1075	768.01	0.000
β_1	0.0001	-36.03	0.000

Rest of the country			
Coefficient	Standard Error	t-value	p-value
β_0	0.1818	454.01	0.000
β_1	0.0001	-19.61	0.000

In both cases, both the intercept and the slope are statistically significant, so we can proceed to check the residuals.





It can be observed that the residuals are normally distributed with mean close to zero. In general, both models are a good fit of our data.

Conclusion

We can conclude that the South has a lower life expectancy than the rest of the United States. In addition, there is no difference between the North, Northeast, West, and Midwest in terms of life expectancy.

Both regressions concluded that the percentage of people living in a rural area does not have an effect on the life expectancy. Our intuition before running the linear regression was that people in rural area do not have easy access to hospitals but have a healthier lifestyle, so we were not sure if rural area was going to have an effect on life expectancy. The linear regression yielded the same slope for smoking in both regressions, which indicates that smoking has the same effect on life expectancy across the United States, and the lower life expectancy in the South is not attributed to this factor. Although smoking was statistically significant, its effect were pretty small. According to the linear regression, the expected difference between a county where 0% of the population does not smoke and a county where 100% of the population smokes daily it's only of 0.29 years, which is about three and a half months. Overall, smoking slightly lowers the life expectancy in a county, and the disparities of life expectancy in the South are attributed to different

factors. This can be seen since both R^2 are around 40%, which indicates that smoking only explain 40% of the variability of life expectancy.

Since we were not able to include all the variables that explain the disparities of life expectancy, our estimates are biased since omitting a variable implies that the covariance between the explanatory variables and the residual errors is not equal to zero. We still proceeded with this regression since the access to free data for every county in the United States is not very accessible.

Future Work

In our analysis, we divided the United States into three and four regions. However, we can divide the United States further given that our data is large. In addition, we can investigate different divisions such as time-zones or political party alignment for a given election. Locating more disparities by testing different regions will help us uncover more disparity that might not be as apparent.

In addition, we used the Bonferroni Correction to reduce our type I error. However, the correction is criticized because it allows the power to be lowered. If we were given more time, we would like to use more complicated procedures for lowering the FWER such as Tukey's procedure (for pairwise tests), Holm's step-down procedure (1979), or Hochberg's step-up procedure. These procedures are known to be better at maintaining the power of the tests. The Holm-Bonferroni Method is a Sequential correction, whereas the Bonferroni Correction is a single step correction. Sequential corrections are adaptive to the p-values, whereas the single steps are generalized for every test. In the Holm-Bonferroni Method, we still apply the Bonferroni Correction, but we start the test with the lowest p-value and end with the test that has the highest p-value. Moreover, we were unable to calculate the overall power of our tests because the topic is still in research. However, if we had time, we would want to follow one of the papers on calculating the cumulative power, so that we can get a better sense of the validity of our results.

In the realm of uncovering the reasons behind the disparity in the South, we want to find or collect on data about other factors such obesity rates, number of hospitals, and number of guns for

each county. If we had more data, we could build a more complex and accurate linear regression model to explain the life expectancy disparities in the US.

References

https://en.wikipedia.org/wiki/Student%27s_t-test#Assumptions
https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mstats.normaltest.html
https://en.wikipedia.org/wiki/Kurtosis
https://en.wikipedia.org/wiki/Student%27s_t-test#Independent_two-sample_t-test
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html
https://en.wikipedia.org/wiki/Student%27s_t-test#Dependent_t-test_for_paired_samples
https://en.wikipedia.org/wiki/Student%27s_t-test#Equal_or_unequal_sample_sizes,_unequal_variances
https://www.merriam-webster.com/dictionary/life%20expectancy
https://www.merriam-webster.com/thesaurus/disparity
https://vizhub.healthdata.org/subnational/usa
https://www.npr.org/sections/health-shots/2017/05/08/527103885/life-expectancy-can-vary-by-20-years-depending-on-where-you-live
https://www.census.gov/newsroom/press-releases/2016/cb16-210.html
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html
https://en.wikipedia.org/wiki/Family-wise_error_rate
http://www.statisticshowto.com/familywise-error-rate/
http://mathworld.wolfram.com/BonferroniCorrection.html
https://stats.stackexchange.com/questions/73646/how-do-i-test-that-two-continuous-variables-are-independent
http://www.biostathandbook.com/gtestgof.html
https://projecteuclid.org/download/pdfview_1/euclid.ss/1270041260
https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Levene_Test_of_Variances-Simulation.pdf
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm
https://workresearch.aut.ac.nz/news-and-events/news/all-news/quality-of-life-symposium-an-interdisciplinary-discussion
https://en.wikipedia.org/wiki/Bonferroni_correction
https://fordham.blackboard.com/bbcswebdav/pid-2534677-dt-content-rid-8602445_1/courses/MATH3007L01201820/Math3007_LectureNotes%20copy%2816%29.pdf