Stats 501 – Final Project

Study of Life Expectancy Variations among US States

Ajay Shaan Shanmugam

Siddharth Chandrasekaran

Abhishek Kumar

1. Introduction:

We aimed to figure out if life expectancy upon birth differs among the various US states. We were curious because we wanted to know if there is a statistically significant difference between life expectancies depending on where we reside in the United States of America. This is a particularly important analysis because historically there has been fluctuations in life expectancies based on various socio-political and geographical factors within a country. We believe that this study would help us better understand how to allocate national financial resources for healthcare between the different states of the USA.

Healthcare policies among the various states are very different and this analysis would provide some insights on which of the states have healthcare policies that are very effective and the ones that need improvement. We could also observe from the visualizations how life expectancies vary within each state as well.

2. Data:

We used the dataset from the Center for Disease Control and Prevention acquired as a part of U.S. Small-area Life Expectancy Estimates Project – USALEEP. Here's the link below.

https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html#life-expectancy

The dataset contains anonymized geographic identifiers, life expectancy at birth for 2010-2015, and flags describing whether the estimates were based exclusively on observed data, a combination of observed and predicted values, or exclusively predicted values. Such a dataset in the form of a CSV file was available for every state except Maine and Wisconsin. They were excluded from this study because they only had 5 years of geocoded death records (2011-2015). Each CSV file contains the following fields – An anonymized tract ID for every region in the state, mean life expectancy in years for a region and the standard deviation of the same. We used entries that were both observed and predicted for our research.

For the purpose of our study, we decided to do ANOVA, and therefore we had to make the assumption that the mean life expectancies of all regions in a state form the sample which collectively represents the entire state. As part of pre-processing, we extracted these mean life expectancies as sample units for a state for each state and created a new dataset containing these mean life expectancies in one field and the corresponding state in the other. This was done to prepare the data for performing Analysis of Variance in R.

After importing the dataset into R, we generated the summary of descriptive statistics. We could immediately notice that the number of samples varies between the states. To use ANOVA we need to check if each group follows a normal distribution. We ensured that the number of samples in each group is sufficiently high enough for an approximately normal distribution to be observed. The table below is an excerpt from the summary of descriptive statistics.

State <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
AK	65.7	77.075	79.35	81.000	86.9	78.916	3.4871	296
AL	63.6	72.600	74.60	77.100	88.2	74.814	3.5251	1101
AR	65.8	73.400	75.50	77.600	89.6	75.627	3.3734	658
AZ	65.8	76.100	78.40	80.500	90.8	78.365	3.3653	1299
CA	64.4	77.900	80.40	82.400	93.4	80.231	3.4172	7516
CO	67.3	77.400	79.70	81.700	89.5	79.466	3.4399	1075
CT	68.9	78.000	80.40	82.300	89.1	80.101	3.2951	783
DC	63.2	72.300	76.80	80.500	90.7	76.408	5.1820	165
DE	68.2	75.800	77.65	79.950	86.1	77.794	3.2265	198
FL	61.1	75.800	78.50	81.000	91.6	78.375	3.9894	3762

Although there were outliers comprising of life expectancies in certain regions, they were few and therefore we didn't bother weeding them out of the dataset before starting our study.

3. Analysis:

As said earlier, in order to determine presence of statistical significance in difference in mean life expectancies across multiple groups (US states), we decided to use ANOVA instead of pairwise comparison using t tests because we didn't want to end up increasing the type I error for our experiment. We began by forming the null and alternate hypotheses as follows.

$H_0 = \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_a = Not all μ_x are equal, or, at least one is different from the others

Where 1, 2, 3… k represent the different US states and x belongs to this range.

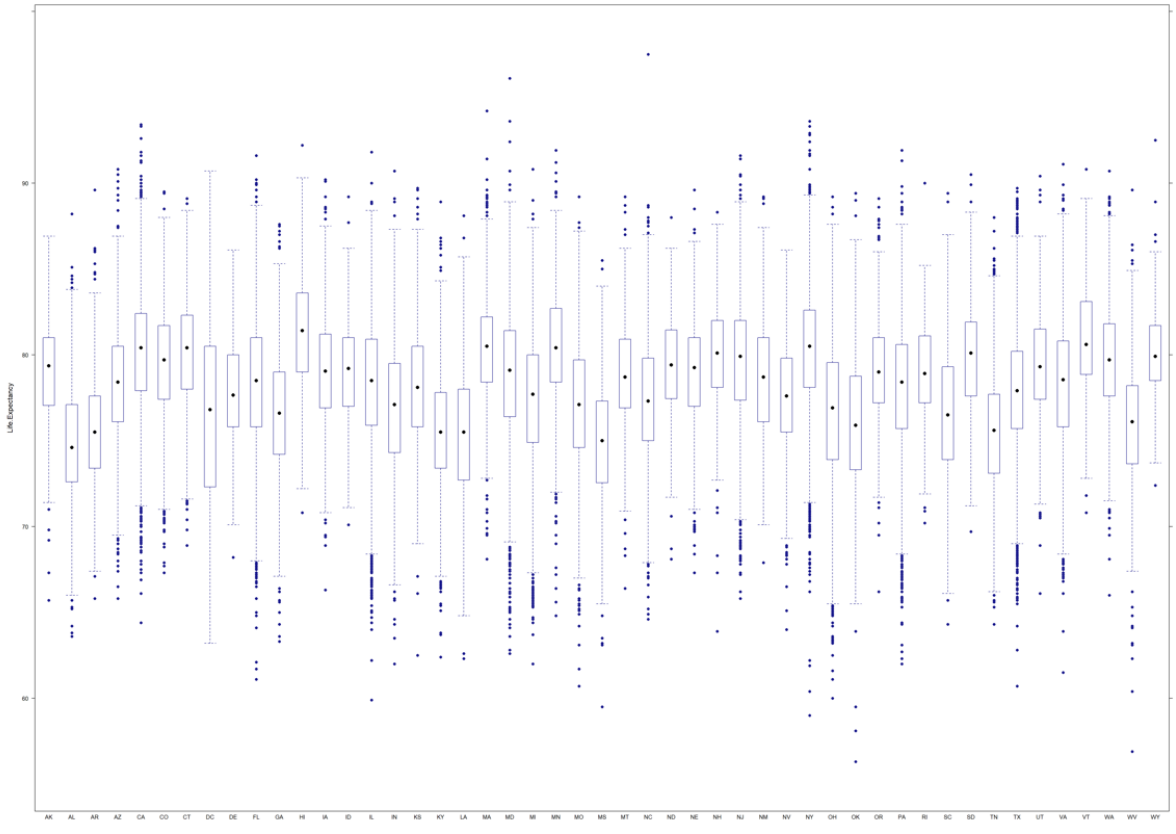
Here, our test statistic is the F ratio which is the ratio of MST and MSE where MST is the mean sum of squares of treatment and MSE is the mean sum of squares of the residuals. Our test statistic follows the F distribution on (k-1), (n-k) degrees of freedom where k refers to the number of groups (States) which is 49 and n refers to the total number of samples which is 65810. Please be reminded that the states of Maine and Wisconsin are not included.

Before proceeding, we made sure to check assumptions required for the ANOVA statistical test.

1. Independence assumption: We assume that the mean life expectancy for each region in each state was determined from independent and randomly selected samples. For our test, we use all these mean life expectancies for every state.

2. Equal variance: We constructed boxplots of life expectancies of different states side by side and compared the variances between groups. We could see that the variances were

more or less equal. We further verified that the ratio of highest standard deviation and lowest standard deviation is less than 3.



3. Normal population distribution: We assume that the population follows a normal distribution.

We could then perform the ANOVA in R. The first tabulation below describes the variables in the ANOVA table and their definitions. The second table summarizes the results.

Source	Df	SS	MS	F=ratio
Factor (treatment)	k-1	SST	MST=SST/(k-1)	MST/MSE
Error	n-k	SSE	MSE=SSE/(n-k)	
Total	n-1	SS		

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## State         48 165881    3456     256 <2e-16 ***
## Residuals   65761 886206      13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen, the test statistic F ratio is large and consequently, the p value for this test statistic is negligibly small (2e-16) and is less than the default significance level α = 0.05. Therefore, the

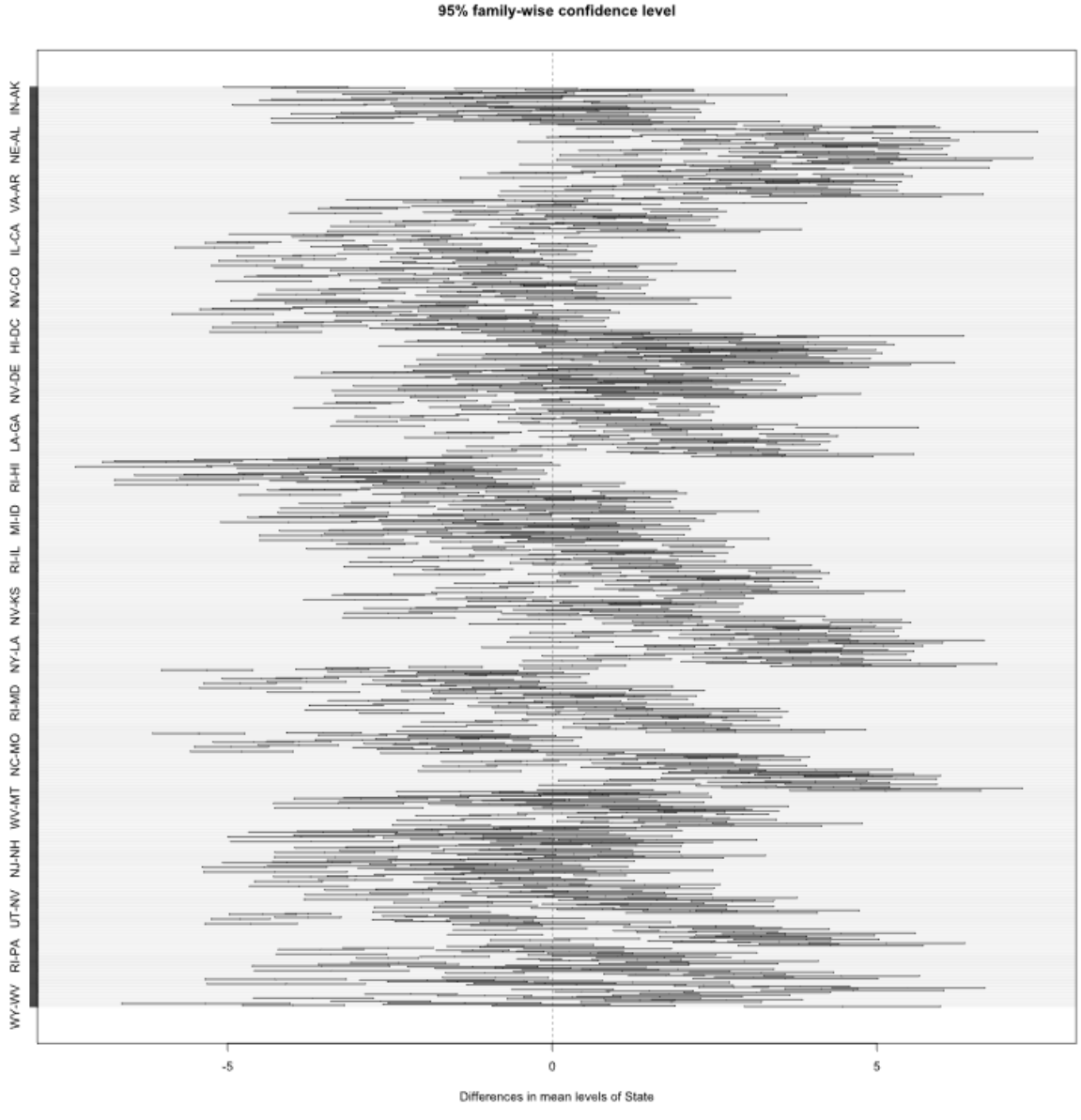
null hypothesis can be rejected, and this experiment clearly shows statistically significant difference in mean life expectancies between the US states.

In order to determine the state pairs whose difference in mean life expectancies are contributing to the rejection of the null hypothesis, we explored further and performed the Tukey test. The Tukey test is very similar to the T test, except that it minimizes family-wise error rate, and is defined by the formula:

$$q_s = \frac{Y_A - Y_B}{SE}$$

Where Y_A and Y_B are the larger and smaller means respectively, and SE is the standard error of the sum of means. The Tukey test's assumptions are the same as that of ANOVA.

The following graph plots the confidence intervals of difference in mean life expectancies between every pair of states. Please note that not all labels in Y axis are displayed for aesthetical reasons.



We could see from the 95% confidence intervals that some state pairs don't have significant difference in mean life expectancies and most of them do. To better understand this, we collected

the results of the Tukey test in a CSV file and studied it. It could be seen that a quarter of these pairwise comparisons had a p value greater then our default significance level of 0.05 and three quarters of them had a p value less then 0.05. This means that a three quarter of these pairwise comparisons between states has statistically significant difference in mean life expectancies.

4. Conclusions:

From the ANOVA test, we could determine that the mean life expectancies between the US states differ with 95% statistical significance. With the post-hoc Tukey test, we could determine that approximately three fourths of pairwise comparisons of mean life expectancies of US states differ with 95% statistical significance. For example, according to the Tukey test, there's a positive difference in means between Massachusetts and Alabama (5.5262) with a p value of approximately 0 and a positive difference in means between California and Alabama (5.4170) with a p value of approximately 0.

The sample dataset we have for each state contains records where in each record describes expected life expectancy of people in a small region (pertaining to a pin code) in a state. Even though these records do not concern individuals' lifetimes, but are means within tiny regions of the state, we used these mean life expectancies of each region in a state as a sample unit collectively which describes the state's sample. We think it's worth exploring the option of getting random, independent individual lifetime values as a sample for a state for each state. Also, the dataset we used corresponds to the years from 2010 to 2015. If we could have found an updated dataset or had the resources to make one, the results from the analysis could have been more accurate and relevant to the present day.

As a different experiment design strategy to determine that the mean life expectancies between the US states differ with statistical significance, we could have also experimented with non-parametric alternatives to ANOVA such as the Kruskal Wallis test which doesn't require us to make an assumption that our data comes from a particular distribution.

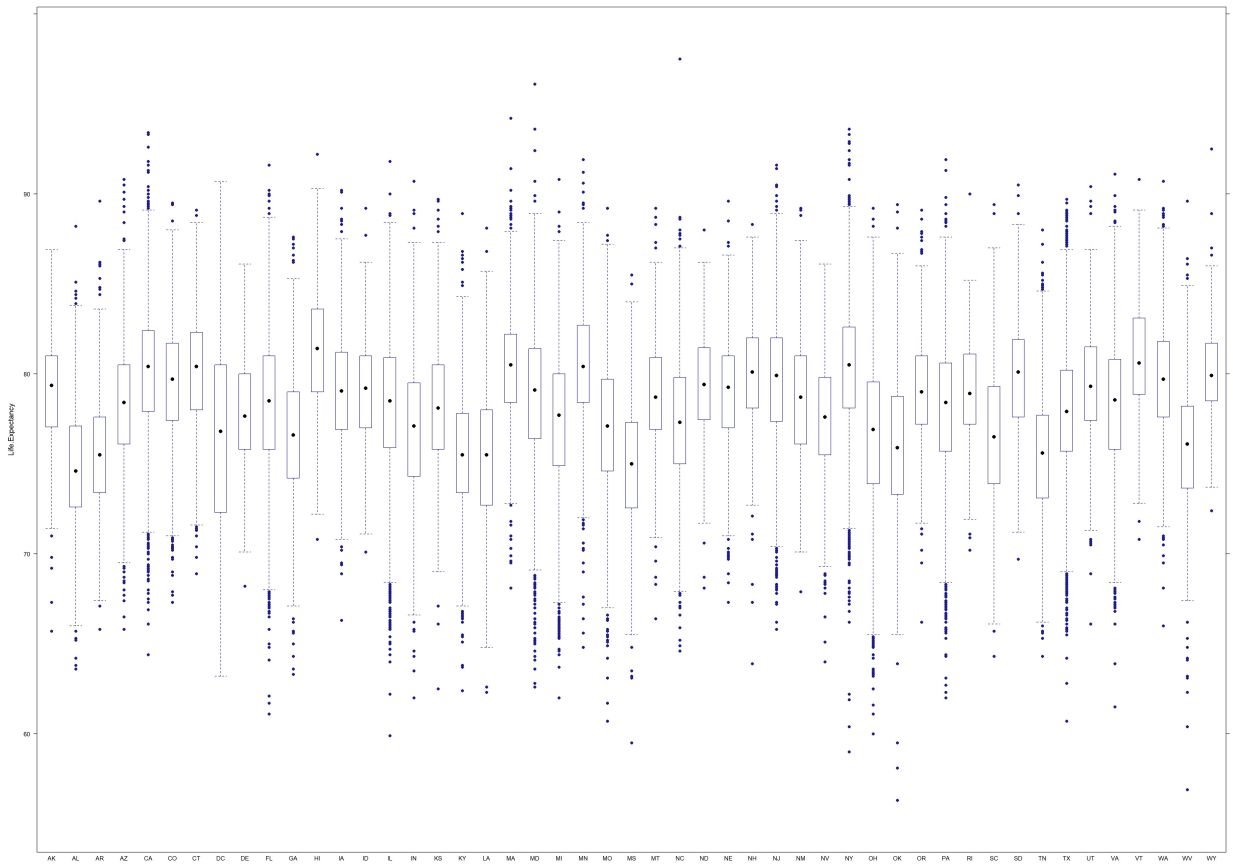
In conclusion, we could find a clear variation in life expectancies across states in the US. The study could be further extended by trying to understand different factors that affect life expectancy in different states.

5. Code Appendix:

Analysis of Variance in Life Expectancies Between States of USA

Analysis of variance

```
lf_df = read.csv("life_expectancy_data.csv")
bwplot(Life.Expectancy ~ State, data=lf_df)
```



The summary statistics can be calculated using `favstats()`.

```
favstats(Life.Expectancy ~ State, data=lf_df)
```

```
##    State  min     Q1 median     Q3  max   mean     sd    n missing
## 1     AK 65.7 77.075  79.35 81.000 86.9 78.916 3.4871  296       0
## 2     AL 63.6 72.600  74.60 77.100 88.2 74.814 3.5251 1101       0
## 3     AR 65.8 73.400  75.50 77.600 89.6 75.627 3.3734  658       0
## 4     AZ 65.8 76.100  78.40 80.500 90.8 78.365 3.3653 1299       0
## 5     CA 64.4 77.900  80.40 82.400 93.4 80.231 3.4172 7516       0
## 6     CO 67.3 77.400  79.70 81.700 89.5 79.466 3.4399 1075       0
## 7     CT 68.9 78.000  80.40 82.300 89.1 80.101 3.2951  783       0
## 8     DC 63.2 72.300  76.80 80.500 90.7 76.408 5.1820  165       0
## 9     DE 68.2 75.800  77.65 79.950 86.1 77.794 3.2265  198       0
## 10    FL 61.1 75.800  78.50 81.000 91.6 78.375 3.9894 3762       0
## 11    GA 63.3 74.200  76.60 79.000 87.6 76.615 3.5359 1774       0
## 12    HI 70.8 79.000  81.40 83.575 92.2 81.315 3.4195  282       0
## 13    IA 66.3 76.900  79.05 81.200 90.2 79.064 3.6116  798       0
## 14    ID 70.1 77.000  79.20 81.000 89.2 79.086 2.9253  286       0
## 15    IL 59.9 75.900  78.50 80.900 91.8 78.165 4.0467 2919       0
## 16    IN 62.0 74.300  77.10 79.500 90.7 76.765 4.0193 1443       0
## 17    KS 62.5 75.800  78.10 80.500 89.7 78.051 3.6583  723       0
## 18    KY 62.4 73.400  75.50 77.800 88.9 75.564 3.5981 1045       0
## 19    LA 62.3 72.700  75.50 78.000 88.1 75.373 3.7491 1043       0
## 20    MA 68.1 78.400  80.50 82.200 94.2 80.340 3.1661 1384       0
## 21    MD 62.6 76.400  79.10 81.400 96.1 78.699 4.3980 1302       0
## 22    MI 62.0 74.900  77.70 80.000 90.8 77.340 3.9684 2582       0
## 23    MN 64.8 78.400  80.40 82.700 91.9 80.469 3.5567 1277       0
## 24    MO 60.7 74.625  77.10 79.700 89.2 76.957 3.9364 1318       0
## 25    MS 59.5 72.550  75.00 77.300 85.5 75.027 3.6176  631       0
## 26    MT 66.4 76.900  78.70 80.900 89.2 78.819 3.5537  255       0
## 27    NC 64.6 75.000  77.30 79.800 97.5 77.397 3.5377 1971       0
## 28    ND 68.1 77.450  79.40 81.450 88.0 79.397 3.4926  195       0
## 29    NE 67.3 77.025  79.25 81.000 89.6 79.027 3.4470  498       0
## 30    NH 63.9 78.125  80.10 81.975 88.3 79.956 3.3545  282       0
## 31    NJ 65.8 77.375  79.90 82.000 91.6 79.591 3.6611 1892       0
## 32    NM 67.9 76.100  78.70 81.000 89.2 78.547 3.4786  462       0
## 33    NV 64.0 75.500  77.60 79.800 86.1 77.549 3.4279  594       0
## 34    NY 59.0 78.100  80.50 82.600 93.6 80.327 3.6492 4547       0
## 35    OH 60.0 73.900  76.90 79.525 89.2 76.567 4.0831 2772       0
## 36    OK 56.3 73.300  75.90 78.750 89.4 75.870 4.0480  991       0
## 37    OR 66.2 77.200  79.00 81.000 89.1 79.088 3.0263  785       0
## 38    PA 62.0 75.700  78.40 80.600 91.9 78.061 3.7666 3060       0
## 39    RI 70.2 77.200  78.90 81.100 90.0 79.085 2.9445  224       0
## 40    SC 64.3 73.900  76.50 79.300 89.4 76.563 3.6844 1018       0
## 41    SD 69.7 77.600  80.10 81.900 90.5 79.788 3.7272  209       0
## 42    TN 64.3 73.100  75.60 77.700 88.0 75.530 3.5854 1352       0
## 43    TX 60.7 75.700  77.90 80.200 89.7 77.864 3.4375 4709       0
## 44    UT 66.1 77.400  79.30 81.500 90.4 79.366 3.2199  536       0
## 45    VA 61.5 75.800  78.55 80.800 91.1 78.327 3.9341 1668       0
## 46    VT 70.8 78.850  80.60 83.100 90.8 81.022 3.4401  175       0
## 47    WA 66.0 77.600  79.70 81.800 90.7 79.675 3.3342 1370       0
## 48    WV 56.9 73.650  76.10 78.200 89.6 75.691 4.0929  467       0
## 49    WY 72.4 78.500  79.90 81.700 92.5 80.161 3.0085  118       0
```

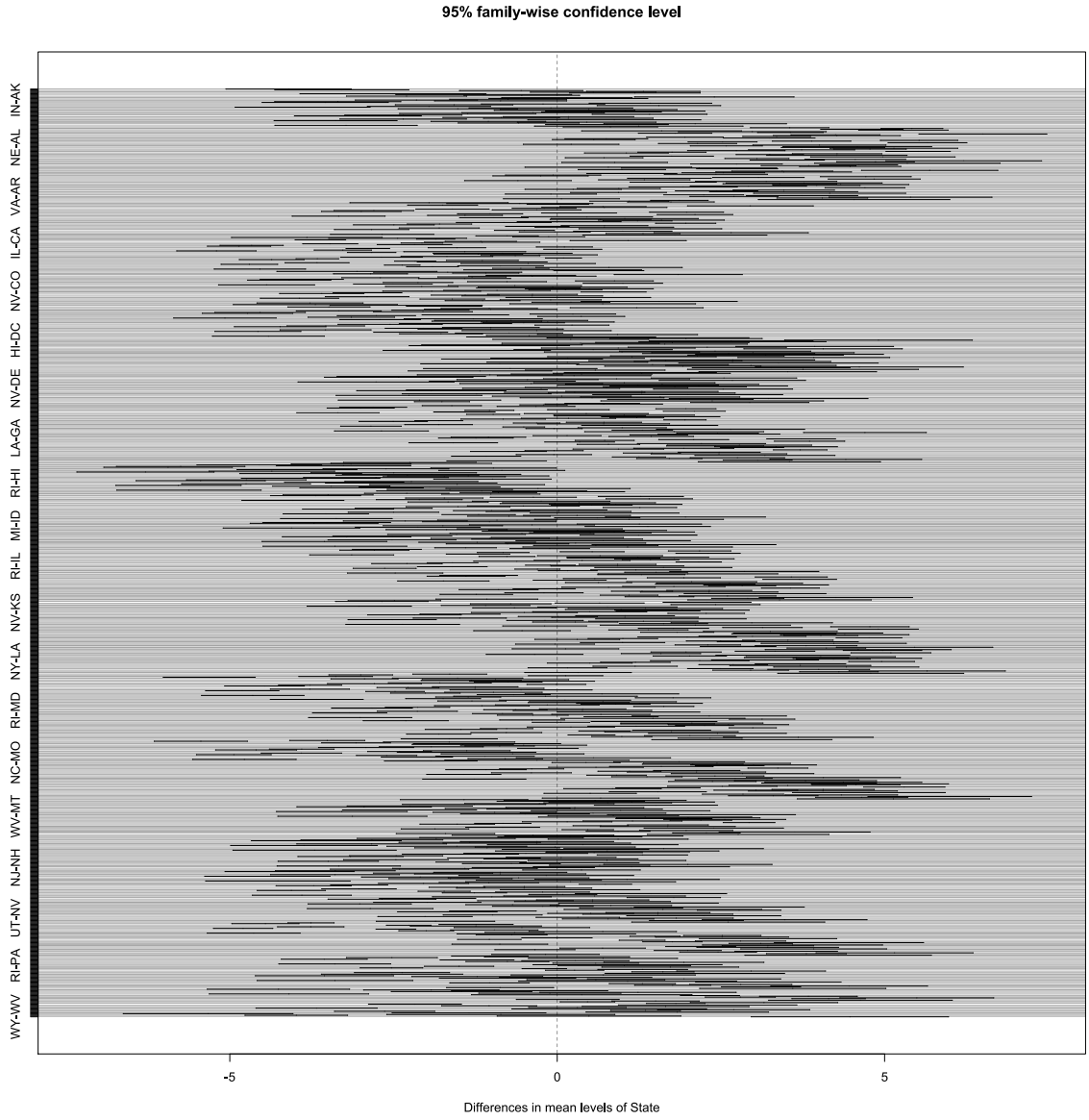
The `aov()` function can be used to fit an analysis of variance model.

```
aovmod = aov(Life.Expectancy ~ State, data=lf_df)
summary(aovmod)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## State          48 165881    3456     256 <2e-16 ***
## Residuals   65761 886206      13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model has 3 degrees of freedom for the model (numerator) and 28 degrees of freedom for the error (denominator). The `xpf()` function can replicate the calculation of the exact p-value

```
tuk = TukeyHSD(aovmod)
plot(tuk)
```




```
tukey_results = as.data.frame(tuk[1:1])
write.csv(tukey_results, 'tukey_results.csv')
```