Group Project: Anova Hypothesis of Insurance Charges

STAT 301-01 Spring 2025

April 12, 2025

Group 8: Nicholas Dillner, Kyle Smith

Introduction:

The insurance dataset provides an opportunity to explore potential factors that affect insurance costs. This set contains 1,338 records and includes numeric variables such as age, BMI, number of children, and insurance charges (in dollars). The dataset also includes categorical variables such as sex that has two levels (male, female), smoking status with two levels (yes, no), and region that contains four levels (southwest, southeast, northwest, northeast). This analysis will focus on two research questions:

Question 1: Do average insurance charges for smoking and nonsmoking differ significantly across the four regions? Additionally, we will examine the interaction between region and smoking status and to clarify regional effects within smoker and non-smokers.

Question 2: Do average insurance charges differ significantly between smokers and non-smokers?

Because of our research questions, many of the listed variables will not be needed. Sex, BMI, number of children, and age are unimportant for our research questions. As such, from the dataset, we will be focused on insurance charge (our Y variable) smoking status for our second question, and the various regions for our first question. Thus, our dependent variable is charges a continuous variable, our independent variables are region a categorical variable with four levels and smoker a categorical variable with two levels. Table 1.0 provides a summary of the dependent and independent variables.

DV	Charges
Min	1121.87
Median	9382.03
Max	63770.43
Average	13282.97
Count	1338
IV1	**Region**
Southwest	325
Northwest	325
Southeast	364
Northeast	324
IV2	**Smoker**
Yes	274
No	1064

Table 1.0

Hypothesis for question 1:

H0: μNorthwest&Smoker = μNortheast&Smoker = μSouthwest&Smoker = μSoutheast&Smoker

Ha: At least two means are significantly different

Hypothesis for question 2:

H0: μSmoker = μNonSmoker

HA: μSmoker ≠ μNonsmoker

Methods:

For our analysis for question 1, we used a Two-Way Anova to test the main effects of region and to see if the regions were different, and the effect smoking has on charges, as well as their interaction. This approach was chosen because it addresses several research questions while

looking at potential interactions between the variables. The dataset was checked confirming no null or missing values within the 1,338 records.

We could use the results from two-way for question 2, but this could complicate things when the question is much simpler to check. A simple t-test and a graph will do here.

First we checked to see if there were any null values. While there didn't seem to be any visually, code was ran to see if any existed and there were none that came up.

For the first question, we tested for an interaction between all regions to see if they were significantly different, and to see if there was an interaction between smoking and regions. If there is an interaction, we split the data between smokers and nonsmokers across all regions to see if there is a significant difference between both results. If there are differences between regions and smokers we would expect the smoking graph to show significance, while the nonsmoking one would show there is no significance. If nonsmoking has significant differences, it's probable that the region in general just has higher prices for all customers or there is another factor unaccounted for. If only smoking is significantly different, we can expect no significant differences between nonsmoking, and significant differences for smoking.

A One-Way Anova test was performed on each subgroup to determine regional differences for smokers and non-smokers. For significant test results, Tukey's HSD test was used to identify pairwise differences. All Anova tests in this analysis assume normality and homogeneity of variances with a significance level of $\alpha = 0.05$. Note that there *are* outliers in our overall data set.

Results:

Our first task to filter data:

```
filter_data_test <- insurance_data %>% filter(if_any(everything(), ~ is.na(.)))
print(filter_data_test) #Returns 0 rows. There is nothing that is NA.
```

Figure 1.0

We can conclude that there are 0 rows returned, thus there are no NA values. We can continue with our question.

Figure 1.1

For our first question we checked to see if there were different means between the regions in general.

H0: μNorthwest = μNortheast = μSouthwest = μSoutheast

Ha: At least two means are significantly different

Our first test concluded that there was a significant difference between regions and all groups of smokers and nonsmokers with an F-stat of 2.97 and a P value of 0.0309. Note that if we had tested a 99%, there would not have been a significant difference at all, but we could still have tested for interaction because other factors could have been affecting the fact they are the same.

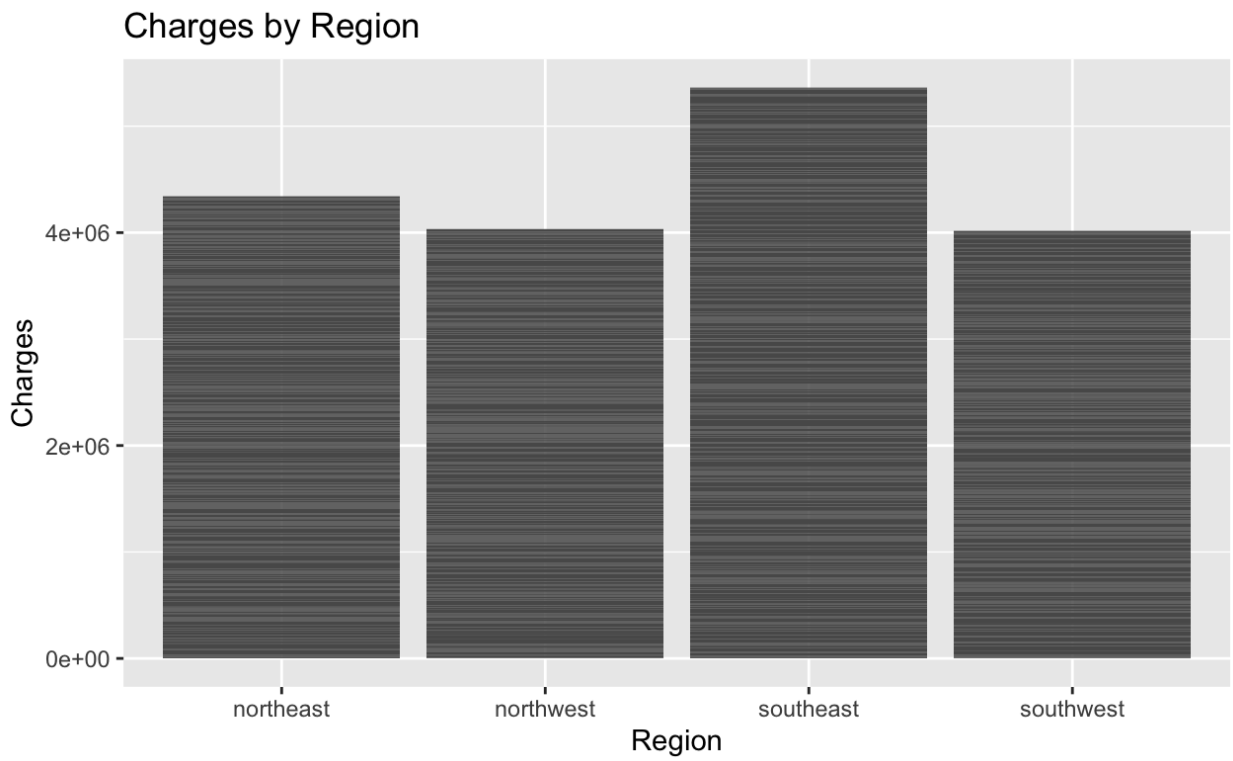

Figure 2.0

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
region	3	1.301e+09	433586560	2.97	0.0309	*
Residuals	1334	1.948e+11	146007093			

Figure 3.0

Now, we check for an interaction between smoking and regions to see if there is an interaction there.

H0: There is no interaction between smoking and region.

HA: There IS an interaction between smoking and region.

The F statistic was 8.958 with a p-value of 1.18-05e. The p-value < 0.05, we reject the null hypothesis, indicating that there is an interaction between smoking and region.

```
                Df    Sum Sq   Mean Sq  F value   Pr(>F)
region           3 1.301e+09 4.336e+08    7.896 3.19e-05 ***
smoker           1 1.203e+11 1.203e+11 2191.337  < 2e-16 ***
region:smoker    3 1.416e+09 4.721e+08    8.598 1.18e-05 ***
Residuals     1330 7.303e+10 5.491e+07
```

Figure 4.0

So we know there's an interaction between smoking and nonsmoking! But what if we wanted to check for just nonsmokers and see if there's significance there? Another test.

HO: μNorthwest&Nonsmoking = μNortheast&Nonsmoking = μSouthwest&Nonsmoking = μSoutheast&Nonsmoking

HA: At least two means are different.

```
             Df    Sum Sq  Mean Sq F value Pr(>F)
region        3 2.315e+08 77175436   2.155 0.0917 .
Residuals  1060 3.796e+10 35808675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = charges ~ region, data = insuranceNonSmoker)

$region
                         diff       lwr       upr     p adj
northwest-northeast  -609.0680 -1954.606  736.4705 0.6492783
southeast-northeast -1133.3154 -2471.582  204.9515 0.1297215
southwest-northeast -1146.2472 -2491.786  199.2913 0.1260743
southeast-northwest  -524.2474 -1849.543  801.0477 0.7390256
southwest-northwest  -537.1792 -1869.817  795.4583 0.7275994
southwest-southeast   -12.9318 -1338.227 1312.3633 0.9999943
```

Figure 5.0

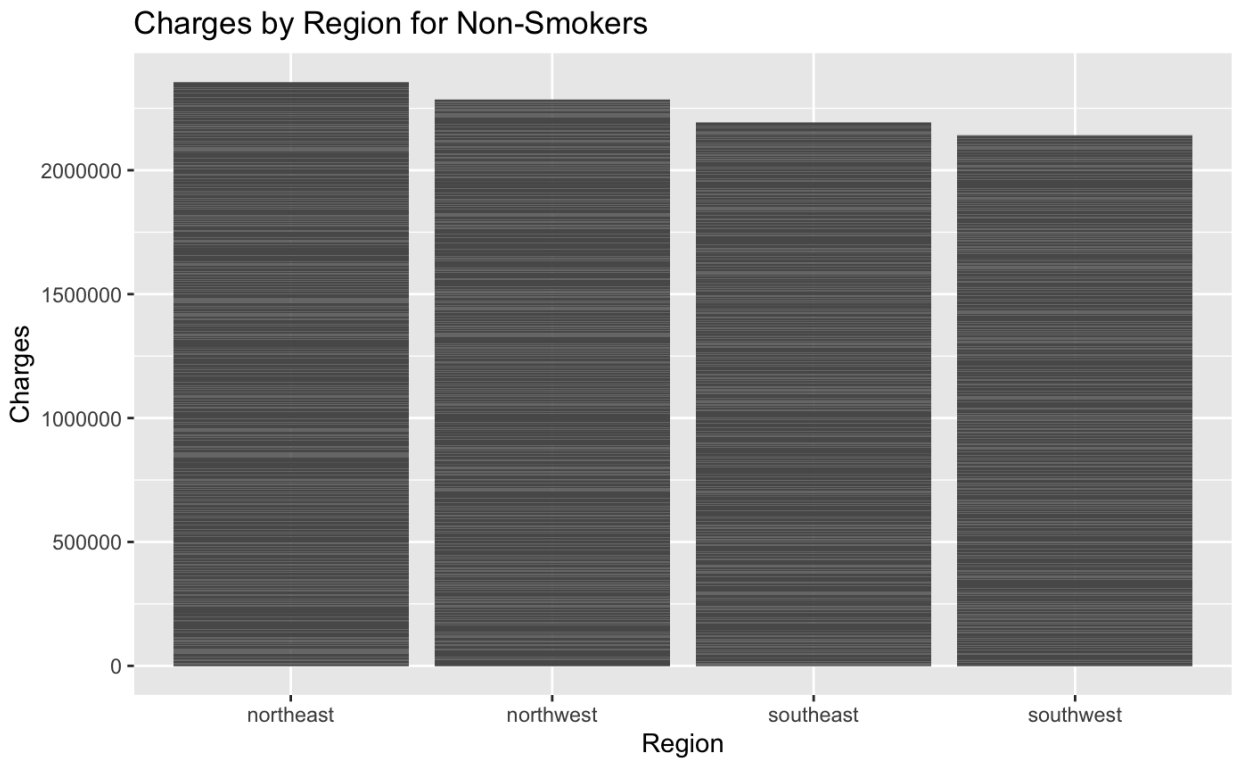

Figure 6.0

We fail to reject the null hypothesis at a p value of 0.09 and F-stat of 2.155. There is no significant difference between the four regions between nonsmokers.

So, we know there is an interaction between nonsmokers and smokers with region, and we know nonsmokers have no significant differences in the regions.

Let's perform our final test to answer our question for interaction between regions and smoking:

To reiterate:

H0: μNorthwest&Smoker = μNortheast&Smoker = μSouthwest&Smoker = μSoutheast&Smoker

Ha: At least two means are significantly different

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
region	3	1.292e+09	430762869	3.316	0.0205	*
Residuals	270	3.507e+10	129901138			

Figure 7.0

With a p-value of 0.0205 and F-stat at 3.316, we can conclude there is a difference between smoking and regions. Splitting up by regions, we can see that the difference is between southeast and northeast.

```
$region
                          diff        lwr      upr     p adj
northwest-northeast   518.4667 -4765.8367 5802.770 0.9942644
southeast-northeast  5171.4604   428.4444 9914.476 0.0264693
southwest-northeast  2595.5270 -2688.7764 7879.830 0.5830080
southeast-northwest  4652.9936  -297.4324 9603.420 0.0739588
southwest-northwest  2077.0603 -3394.1718 7548.292 0.7601837
southwest-southeast -2575.9333 -7526.3594 2374.493 0.5350117
```

Figure 8.0

We can conclude at 0.05 significance that there is a difference between charges of smoking and nonsmoking by region. Specifically, between southeast and northeast. Although these charges do not appear ‘extremely’ significant as if we tested 99%, this test would have failed.

Thus we can conclude for our first research question that there is no significant differences for nonsmokers throughout the US, but there IS a significant difference between smokers throughout the US as far as charges and region go.

For research question 2 examining differences in smoker and non-smokers, we used a simple t-test. (Frankly, that was a lot of testing we did with question 1!) with a t-test of –32.752 with a p-value that is less than 2e-16.

The p-value < 0.05, we reject the null hypothesis, confirming that those who smoke pay significantly higher rates.

```
        Welch Two Sample t-test

data:  charges by smoker
t = -32.752, df = 311.85, p-value < 2.2e-16
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 -25034.71 -22197.21
sample estimates:
 mean in group no mean in group yes
         8434.268         32050.232
```

Figure 9.0

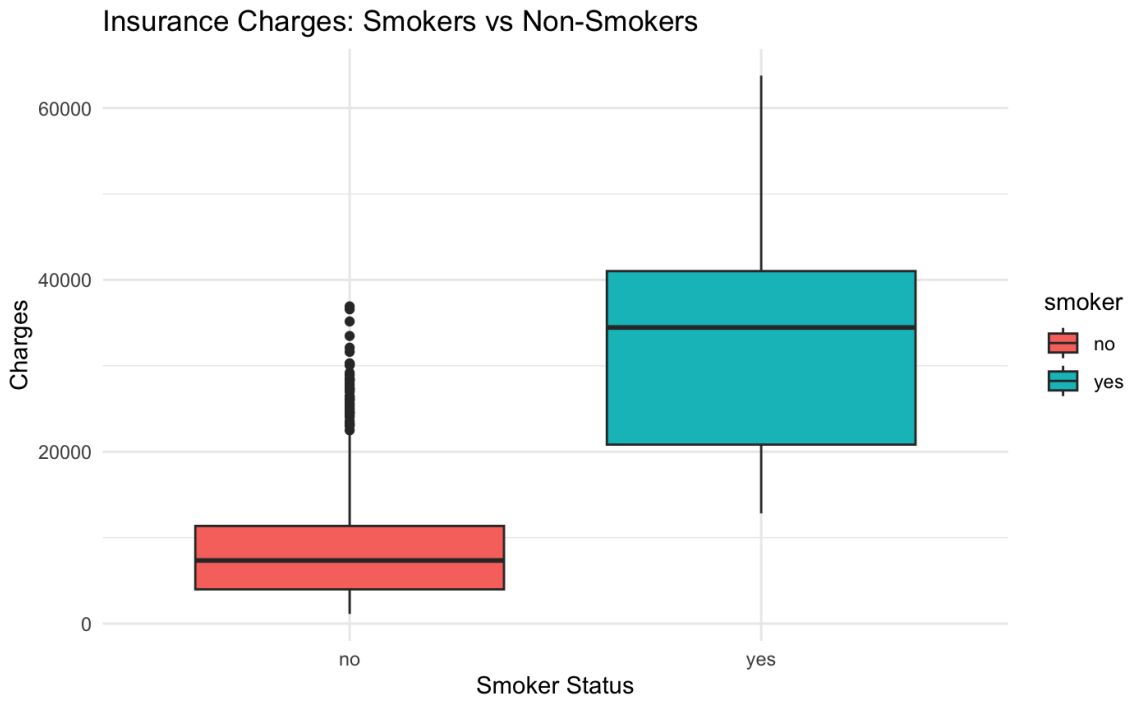

Figure 10.0

We can see there are several outliers for nonsmokers but, in general, smokers pay more for health insurance than nonsmokers.

Conclusion:

For our first question: Between all regions, there are significant differences and that based on our results, the southeast region seems to be the heaviest impacted, having higher charges for smokers. For nonsmoking, there appears to be far less differences. For our second question: Smokers experience higher charges than non-smokers.

Interests:

It's not surprising that smoking impacts the costs of health insurance (That it's higher. Many insurances will literally ask this question when applying) But the different regions are unusual. It's also important to note how different, because it while our data proved it was significant, this was with clear outliers and other forms of data that we could have checked and many other interactions, but this would have complicated things.

Limitations:

Other factors affecting the price such as other medical conditions. The data does not include any information regarding insurance deductibles, which could impact the prices of the data. The dataset is divided into different regions, not different states. Different states have different costs of living which can skew data. (New York is to the Northeast for instance. It is unknown where California is from in the data) The year of the dataset is not known.

Resources:

"Insurance Data." Kaggle, https://www.kaggle.com/datasets/mirichoi0218/insurance. Accessed 11 Apr. 2025.

GitHub:

https://github.com/MatchaMyu/SchoolProjects/upload/main/STAT301-FinalProject