MWSUG 2017 - Paper AA17

Tornado Inflicted Damages Pattern

Vasudev Sharma, Oklahoma State University, Stillwater, OK

ABSTRACT

On average, about a thousand tornadoes hit the United States every year. Three out of every four tornadoes in the world, occur in the United States. They damage life and property in their path and they often hit with very little, sometimes no warning. Tornadoes cause approximately 70 fatalities and 1,500 injuries in US every year. The interest of this study is to find a whether the fatalities and injuries caused by the tornadoes based on the weekday, magnitude are significantly different among the different levels. The idea behind this paper is to find patterns in the damages dealt by the tornadoes and find an insight whether the safety measures are applied correctly.

INTRODUCTION

The idea of this paper started as a small question after a tornado warning. "Can we predict the tornado occurrences in advance?" The idea was to collect data from various sources and try to build a model that can help in predicting the tornadoes based on the changes in the past years. This paper starts on that idea and describes the effect of various variables on the damages caused by the tornadoes.

The preliminary data analysis was surprising and looking at the summary statistics it was hypothesized that there might be uncommon patterns involved in the tornado hits which can be used to better counter the damages inflicted by the tornadoes. Some of the more direct observations were easily recognizable. For example, tornadoes being more in number in the group of states known as tornado alley. But there are more insights like if there are more fatalities on a particular weekday or which magnitude of tornado will cause more damage are the idea for this research paper. The term tornado alley refers to a group of states in United States where the occurrences of tornadoes is higher compared to other states. States which constitute the Tornado Alley are Texas, Oklahoma, Kansas, South Dakota, Iowa, Illinois, Missouri, Nebraska, Colorado, North Dakota and Minnesota. The term will be used for the states mentioned as a group.

PROJECT DATA AND CONSIDERATIONS

The patterns are found by the statistical analysis of the data acquired from National Oceanic and Atmospheric Administration's National Weather Service. Their Storm Prediction Center contains tornado data from 1950 to 2016 with 29 variables like day, month and year of the tornado hit, state affected, magnitude, fatalities and injuries etc. The total number of observations are 62,208. There are categorical and continuous variables which are used for two-Sample T-tests and ANOVA to find out if there are patterns observed in the tornado data.

The data was analyzed and cleaned prior to analysis. Some new variables were created. Variable 'weekday' was created from the variable 'date' (yyyy/mm/dd). Two binary variables 'alley_flag' and 'weekday_flag' were created. 'weekday_flag' denoted whether the day was a weekday or a weekend. 0 denotes a weekend and 1 denotes a weekday. Similarly, alley_flag was created to divide the data based on whether the tornado occurred in a state from tornado alley or not. 0 denotes a non-tornado alley state and 1 denotes a tornado alley state. The start latitude and start longitude were used to plot the tornado hits on the map.

Data Description

The variables which were used for the analysis are as follows

Weekday, weekday_flag, alley_flag, fat, inj, mag, slat and slong.

Variable	Description
Weekday	Day of the week (categorical)
Weekday_flag	Whether the day is weekday or not
Alley_flag	Whether the state is in tornado alley
Fat	Fatalities
Inj	Injuries
Mag	Magnitude of the tornado
Slat	Starting latitude of the tornado
Slong	Starting longitude of the tornado

DATA ANALYSIS

First summary of the data showed the general trends of the tornado occurrences. Texas being the state with highest number of the tornado hits. The map was generated using Tableau desktop 10.2 using 'slat' and 'slong' variables. It is to be noted that variable 'mag' has observation of -9 which means that magnitude is unknown for that tornado. The number of such tornadoes is very small (30) so it is decided to keep it in the data since they won't affect the analysis.

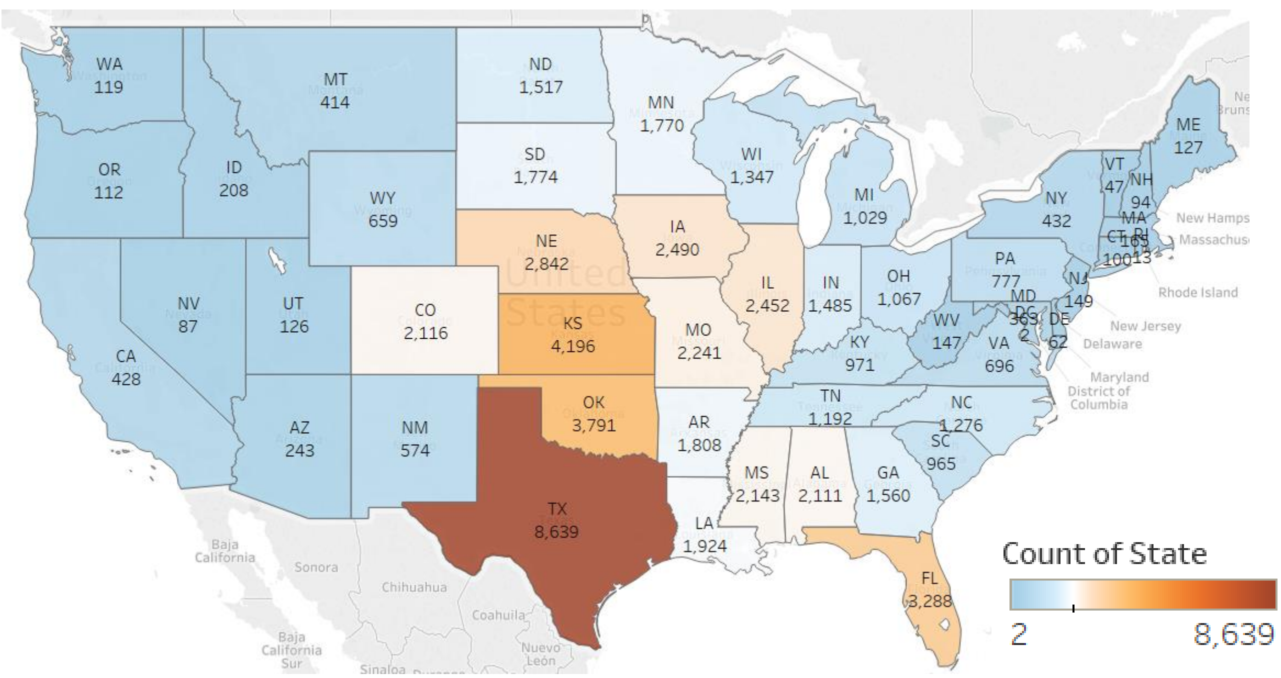

Figure 1. Total number of tornadoes per state in United States

As the map shows, Florida receives a larger number of tornadoes apart from the tornado alley states. Texas has the highest number of tornadoes in all United States.

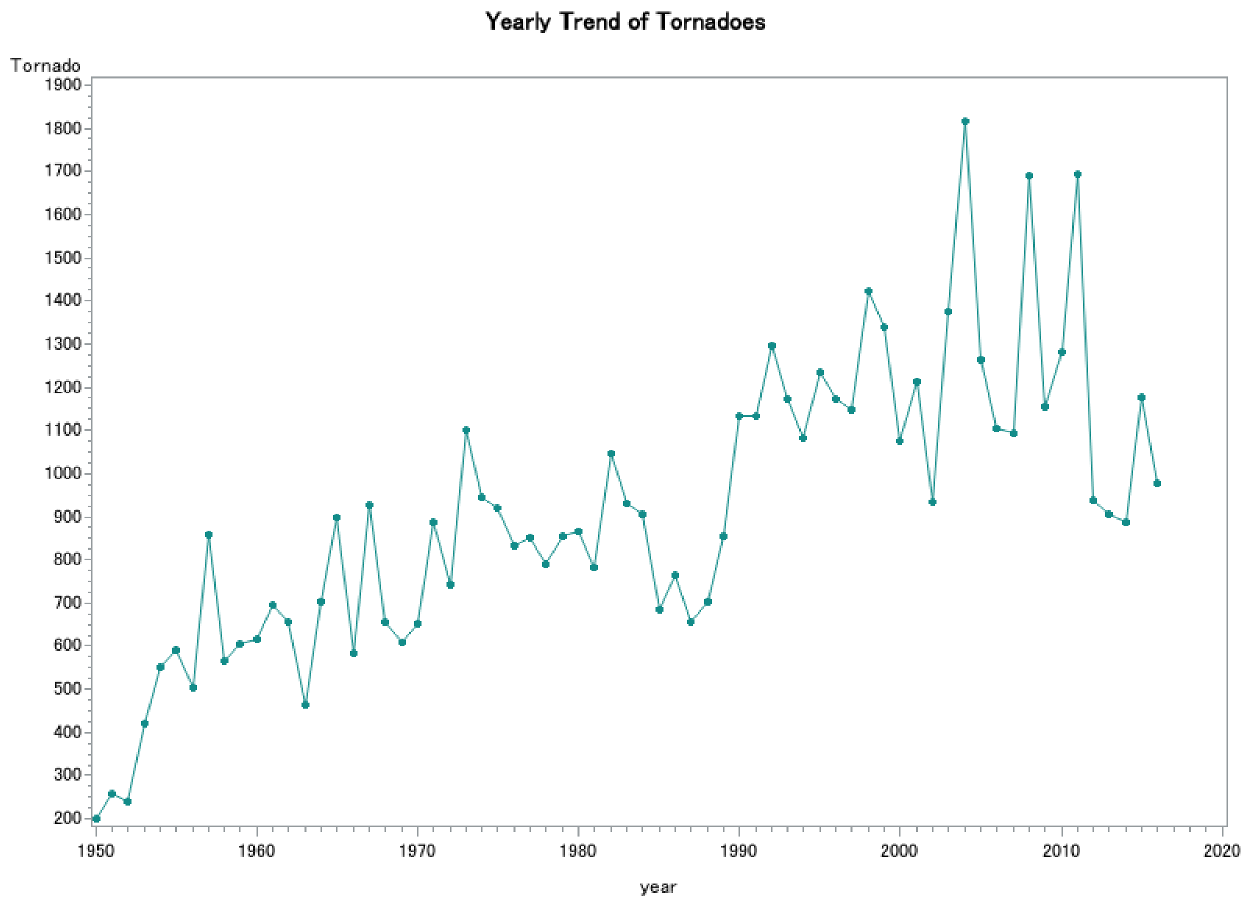

Figure 2. Number of tornadoes per year.

The plot above shows the yearly trend of the tornadoes. It shows the average tornado for every year from 1950 to 2016. It is observed that the overall trend of tornadoes is rising in years.

ANALYSIS OF VARIABLES

The variables were selected and hypothesis was stated for different groups of variables. Analysis of Variance was run for the categorical predictors and the continuous dependent variable. Similarly a two-sample T-test was run for the binary predictor and continuous dependent variable. For all the tests in this paper, normality and independence is checked and data is found to be normal and independent. So, for all the tests done ahead, assumptions of normality and independence are satisfied.

Weekday vs Fatalities

The hypothesis with level of significance, $\alpha = 0.05$ states,

H_0 = mean fatalities on all days of the week is same

H_1 = at least one of the weekdays have different mean of fatalities.

Testing for means fatalities with weekday with PROC GLM

Dependent Variable: fat

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	106.3601	17.7267	6.33	<.0001
Error	62201	174212.1657	2.8008		
Corrected Total	62207	174318.5258			

R-Square	Coeff Var	Root MSE	fat Mean
0.000610	1515.630	1.673557	0.110420

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Weekday	6	106.3601025	17.7266837	6.33	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Weekday	6	106.3601025	17.7266837	6.33	<.0001

Figure 3. Testing fatality means with weekdays with PROC GLM

Since the p-value is less than α, we reject the null hypothesis and the model is significant. So next step is to check assumptions of ANOVA. The normality and independence is checked and data is found to be normal and independent.

H_0 = variances are equal

H_1 = variances are not equal

Level of significance, α = 0.05

ANOVA Diagnostics for testing Assumptions with PROC GLM

Levene's Test for Homogeneity of fat Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Weekday	6	194016	32336.0	1.55	0.1578
Error	62201	1.2986E9	20877.4		

Figure 4. ANOVA Diagnosis for testing Assumptions with PROC GLM

Since the p-value is more than α, we don't reject the null hypothesis. The variances are equal and assumptions are satisfied.

So we run a Tukey test to check which weekdays have a significantly different fatality number.

Multiple Comparisons - All posssible Pairs via Tukey Test

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Weekday	fat LSMEAN	LSMEAN Number
Friday	0.11024931	1
Monday	0.08534473	2
Saturday	0.08149698	3
Sunday	0.15857385	4
Thursday	0.06302886	5
Tuesday	0.08775420	6
Wednesday	0.18247312	7

Figure 5. Multiple comparisons – All possible pairs via Tukey test

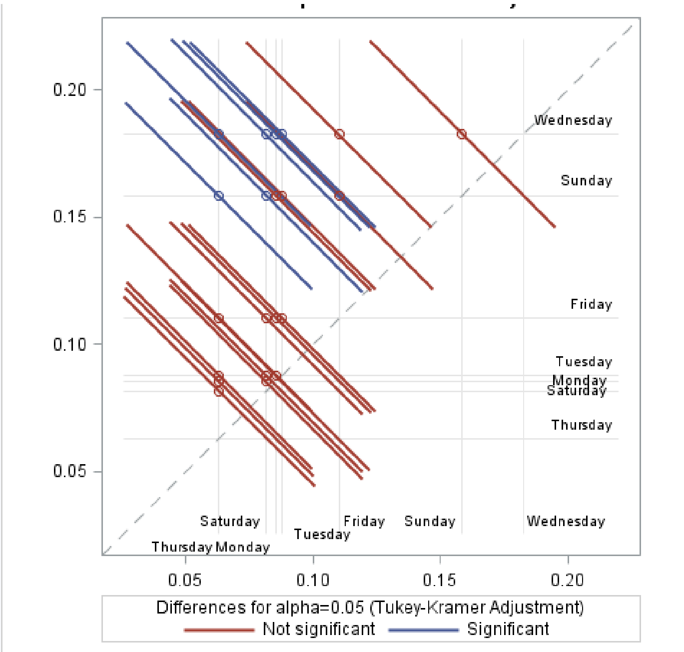

Figure 6. Fatality comparisons for weekdays

Findings show that fatalities on Monday and Wednesday, Tuesday and Wednesday, Wednesday and Thursday, Saturday and Sunday, Saturday and Wednesday, Sunday and Thursday are significantly different from each other.

Weekday vs Injuries

The hypothesis with level of significance, α = 0.05 states

H_0 = mean injuries on all days of the week is same

H_1 = at least one of the weekdays have different mean of injuries.

Dependent Variable: inj

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	12140.60	2023.43	4.53	0.0001
Error	62201	27768672.62	446.43		
Corrected Total	62207	27780813.22			

R-Square	Coeff Var	Root MSE	inj Mean
0.000437	1223.419	21.12900	1.727045

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Weekday	6	12140.59590	2023.43265	4.53	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Weekday	6	12140.59590	2023.43265	4.53	0.0001

Figure 7. Testing injury means with weekdays with PROC GLM

Since the p-value is less than α, we reject the null hypothesis and the model is significant. So next step is to check assumptions of ANOVA.

H_0 = variances are equal

H_1 = variances are not equal

Level of significance, $\alpha = 0.05$

ANOVA Diagnostics for testing Assumptions with PROC GLM

Levene's Test for Homogeneity of inj Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Weekday	6	5.6585E9	9.4308E8	1.82	0.0916
Error	62201	3.229E13	5.1916E8		

ANOVA Diagnostics for testing Assumptions with PROC GLM

Level of Weekday		inj	
	N	Mean	Std Dev
Friday	9025	1.51401662	13.1395119
Monday	8659	1.33930015	17.1918003
Saturday	8123	1.64138865	14.5251286
Sunday	8835	2.24040747	24.3134421
Thursday	9218	1.12898677	11.1951707
Tuesday	9048	1.76226790	31.0460667
Wednesday	9300	2.44043011	27.1228937

Figure 8. ANOVA Diagnosis for testing Assumptions with PROC GLM

Since the p-value is more than α, we don't reject the null hypothesis. The variances are equal and assumptions are satisfied.

So we run a Tukey test to check which weekdays have a significantly different injury count.

Multiple Comparisons - All posssible Pairs via Tukey Test

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Weekday	inj LSMEAN	LSMEAN Number
Friday	1.51401662	1
Monday	1.33930015	2
Saturday	1.64138865	3
Sunday	2.24040747	4
Thursday	1.12898677	5
Tuesday	1.76226790	6
Wednesday	2.44043011	7

Figure 9. Multiple comparisons – All possible pairs via Tukey test

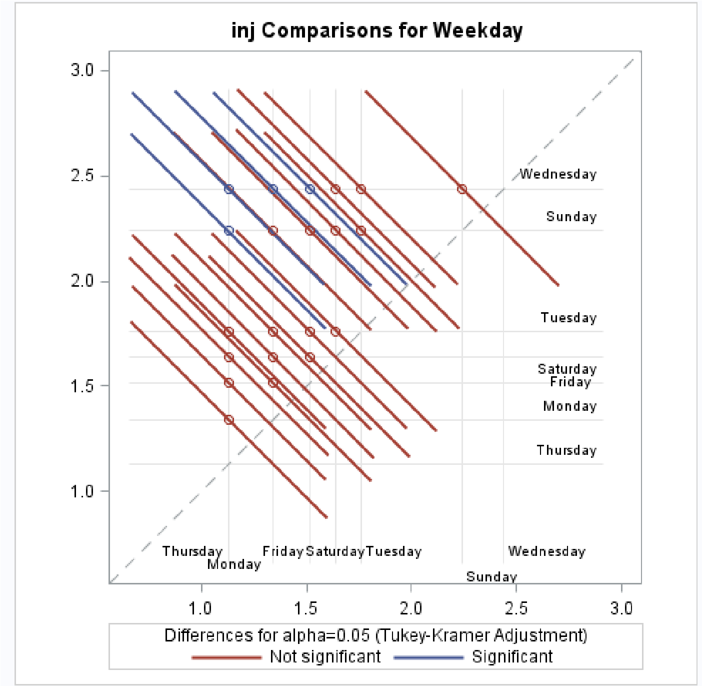

Figure 10. Fatality comparisons for weekdays

Monday and Wednesday, Wednesday and Thursday, Wednesday and Friday, Thursday and Sunday are significantly different from one another in terms of injuries caused by tornadoes.

Magnitude vs Fatalities

Descriptive Statistics of Number of Tornadoes by magnitude

Analysis Variable : year					
N Obs	N	Mean	Std Dev	Minimum	Maximum
62208	62208	1988.445	18.031	1950.000	2016.000

Analysis Variable : year						
mag	N Obs	N	Mean	Std Dev	Minimum	Maximum
-9	30	30	2016.000	0.000	2016.000	2016.000
0	28619	28619	1994.147	15.528	1950.000	2016.000
1	20817	20817	1986.144	18.201	1950.000	2016.000
2	9269	9269	1979.182	18.302	1950.000	2016.000
3	2674	2674	1980.120	18.587	1950.000	2016.000
4	711	711	1978.782	18.660	1950.000	2016.000
5	88	88	1976.045	18.133	1953.000	2013.000

Figure 11. Descriptive statistics of magnitude of tornados

Here we see that most of the tornadoes have magnitude of 0 or 1. Next we run ANOVA

The hypothesis with level of significance, $\alpha = 0.05$ states

H_0 = mean fatalities by all magnitude tornadoes is same

H_1 = at least one of the magnitude tornadoes have different fatalities.

Testing for means fatalities with mag with PROC GLM

Dependent Variable: fat

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	47109.7317	7851.6219	3839.19	<.0001
Error	62201	127208.7941	2.0451		
Corrected Total	62207	174318.5258			

R-Square	Coeff Var	Root MSE	fat Mean
0.270251	1295.128	1.430079	0.110420

Source	DF	Type I SS	Mean Square	F Value	Pr > F
mag	6	47109.73169	7851.62195	3839.19	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
mag	6	47109.73169	7851.62195	3839.19	<.0001

Figure 12. Testing fatality means with magnitude with PROC GLM

Since the p-value is less than α, we reject the null hypothesis and the model is significant. So next step is to check assumptions of ANOVA.

H_0 = variances are equal

H_1 = variances are not equal

Level of significance, α = 0.05

ANOVA Diagnostics for testing Assumptions with PROC GLM

Levene's Test for Homogeneity of fat Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
mag	6	58490906	9748484	966.86	<.0001
Error	62201	6.2715E8	10082.6		

Figure 13. ANOVA Diagnosis for testing Assumptions with PROC GLM

Since p-value is less than 0.05, variances are not equal and assumptions are violated. So Welch Anova is tested against level of significance of 0.05

Welch ANOVA when homogenity of Variance Assumption is violated

Welch's ANOVA for fat			
Source	DF	F Value	Pr > F
mag	5.0000	161.82	<.0001
Error	868.7		

Figure 14. Welch ANOVA testing

Since p-value is less than 0.05, Welch test result is significant. Which means that at least one mean is different than the other.

Multiple Comparisons - All posssible Pairs via Tukey Test

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

mag	fat LSMEAN	LSMEAN Number
0	0.0008386	1
1	0.0111928	2
2	0.0660265	3
3	0.5415108	4
4	3.8888889	5
5	20.3068182	6
-9	-0.0000000	7

Least Squares Means for effect mag
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: fat

i/j	1	2	3	4	5	6	7
1		0.9855	0.0026	<.0001	<.0001	<.0001	1.0000
2	0.9855		0.0348	<.0001	<.0001	<.0001	1.0000
3	0.0026	0.0348		<.0001	<.0001	<.0001	1.0000
4	<.0001	<.0001	<.0001		<.0001	<.0001	0.3753
5	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
6	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
7	1.0000	1.0000	1.0000	0.3753	<.0001	<.0001	

Figure 15. Multiple comparisons – All possible pairs via Tukey test

We see from the results that magnitude of 2 or more on Enhanced Fujita Scale causes much different fatalities than a magnitude of less than 2 on Enhanced Fujita Scale. We can deduce that fatalities are significantly different when the magnitude of the tornado is 0 or 1 than when it is 2 or more.

TWO-SAMPLE T-TESTS FOR BINARY VARIABLES

Two-Sample T-tests are done when the categorical variable is a binary variable. A dummy variable was created for tornado alley states and weekdays. Two-sample T-test code was run in SAS® 9.4 and the following results were generated.

Tornado Alley flag vs Fatalities

The hypothesis with level of significance, $\alpha = 0.05$ states,

H_0 = variances are equal across both groups

H_1 = variances are different across both groups.

Two sample t-test

Variable: fat

alley_flag	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	28380	0.1586	1.8706	0.0111	0	116.0
1	33828	0.0700	1.4880	0.00809	0	158.0
Diff (1-2)		0.0887	1.6734	0.0135		

alley_flag	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		0.1586	0.1369	0.1804	1.8706	1.8554	1.8861
1		0.0700	0.0541	0.0858	1.4880	1.4768	1.4993
Diff (1-2)	Pooled	0.0887	0.0623	0.1151	1.6734	1.6642	1.6828
Diff (1-2)	Satterthwaite	0.0887	0.0617	0.1156			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	62206	6.58	<.0001
Satterthwaite	Unequal	53789	6.45	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	28379	33827	1.58	<.0001

Figure 16. Two-sample t-test for fatalities vs. tornado alley flag

Since p-value for equality of variances is less than α, null is rejected and variances are unequal. For unequal variances, we use Satterthwaite method. The fatalities are found to be 0.0887 more in states which don't fall in tornado alley.

Tornado Alley flag vs Injuries

The hypothesis with level of significance, α = 0.05 states,

H_0 = variances are equal across both groups

H_1 = variances are different across both groups.

Two sample t-test

Variable: inj

alley_flag	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	28380	2.4751	23.7245	0.1408	0	1500.0
1	33828	1.0994	18.6600	0.1015	0	1740.0
Diff (1-2)		1.3757	21.1217	0.1700		

alley_flag	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		2.4751	2.1991	2.7512	23.7245	23.5309	23.9213
1		1.0994	0.9006	1.2983	18.6600	18.5205	18.8017
Diff (1-2)	Pooled	1.3757	1.0424	1.7089	21.1217	21.0049	21.2397
Diff (1-2)	Satterthwaite	1.3757	1.0355	1.7159			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	62206	8.09	<.0001
Satterthwaite	Unequal	53411	7.93	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	28379	33827	1.62	<.0001

Figure 17. Two-sample t-test for injuries vs. tornado alley flag

Since p-value for equality of variances is less than α, null is rejected and variances are unequal. For unequal variances, we use Satterthwaite method. The injuries are found to be 1.3757 more in states which don't fall in tornado alley.

Tornado Alley flag vs Magnitude

Two sample t-test

Variable: mag

alley_flag	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	28380	0.9076	0.9409	0.00559	-9.0000	5.0000
1	33828	0.7282	0.9654	0.00525	-9.0000	5.0000
Diff (1-2)		0.1795	0.9543	0.00768		

alley_flag	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		0.9076	0.8967	0.9186	0.9409	0.9332	0.9487
1		0.7282	0.7179	0.7384	0.9654	0.9581	0.9727
Diff (1-2)	Pooled	0.1795	0.1644	0.1945	0.9543	0.9490	0.9596
Diff (1-2)	Satterthwaite	0.1795	0.1645	0.1945			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	62206	23.37	<.0001
Satterthwaite	Unequal	60833	23.42	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	33827	28379	1.05	<.0001

Figure 17. Two-sample t-test for magnitude vs. tornado alley flag

Since p-value for equality of variances is less than α, null is rejected and variances are unequal. For unequal variances, we use Satterthwaite method. The magnitude is found to be 0.1795 more in states which don't fall in tornado alley.

CONCLUSION

The analyses show that

- The trend is rising for the number of tornadoes that occur every year.
- Weekdays have a significant effect on the fatalities and injuries incurred from the tornadoes.
- Tornadoes having a magnitude greater than or equal to 2 on Enhanced Fujita Scale cause significantly higher number of fatalities even though number of tornadoes with magnitude less than 2 is very high.
- States which do not fall in the tornado alley, have tornadoes with higher magnitudes and cause more fatalities and injuries.

FUTURE WORK

This research generated results which are surprising. The states which have higher number of tornadoes are not the one with higher average of tornado magnitude, fatalities and injuries. Research can be continued to find more insights and inconspicuous results using more variables. The goal is to include variables such as elevation, vegetation and other geographic properties of the various states to find out factors that affect the tornado occurrences. One more factor that may have a significance on fatalities and injuries is early warning systems in a state.

REFERENCES

http://www.spc.noaa.gov/wcm/

ACKNOWLEDGMENTS

I would like to thank Dr. Miriam McGaugh, Clinical Professor, Business Analytics, Oklahoma State University and Koteswara Rao Sriram for their continuous support and guidance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vasudev Sharma
Master's Business Analytics
Oklahoma State University
Phone- 405-762-6021
Email: vasudev.sharma@okstate.edu