Journal of Statistics Education

African Conflict and Climate Data for an Undergraduate Research Project

Darcie A.P. Delzell
Wheaton College

Journal of Statistics Education Volume 20, Number 3 (2012),
www.amstat.org/publications/jse/v20n3/delzell.pdf

Key Words: Class Project; Chi-Square; ANOVA; Bootstrap Analysis.

Abstract

Undergraduate research experiences can be a powerful tool that statistics educators can use to give students an in-depth look at real data analysis as it occurs in multiple professional and academic settings. This article has two goals. The first is to introduce two large and fascinating datasets that are freely available, interesting in content to students, and widely used in current studies. The second is to outline an undergraduate research project that utilized these data. This project was undertaken by four undergraduates over the course of a semester. The phases of the project are discussed as well as example results from the students. There are many possible modifications to the project that can be made at various levels of complexity. Appendices provide relevant R code and descriptions of the merged data available for download.

1. Introduction

There has been much investigation into the benefits of undergraduate research for students (Zydney, Bennett, Shahid, and Bauer 2002; Lopatto 2004; Hunter, Laursen, and Seymour 2007). They are given the unique opportunity to study real problems using real data alongside a professor. These types of projects allow students to perform scientific inquiry using statistical reasoning and methods. This article describes two large and interesting datasets that were used as source files in a semester-long project that was undertaken with a small group of four students with varying levels of statistical knowledge (all four were double majors in mathematics and economics). Two of these students had not completed any statistics courses, but were taking a first course in mathematical statistics concurrently with this project. The other two students had

completed a year-long probability and mathematical statistics sequence. In addition, two merged and simplified datasets created from the sources files are described and provided for download.

The general question of interest, as posed to the students, was to investigate possible relationships between climate and violent political conflict on the African continent. The students were able to experience the real complexities that occur in actual statistical analysis projects. They were introduced to problems related to the merging of data from multiple sources. They quickly discovered that plotting data in multiple ways is one of the most informative steps in any analysis and not always simplistic. As a research group they analyzed the data using various methods and investigated the appropriateness of these methods. This led to multiple strategy changes and the need for the students to learn new methods. They were given the opportunity to communicate their results in both written and verbal form, the latter communication being a presentation they gave to a group of political scientists and economists on campus.

The purpose of this article is to introduce these data sources (as well as the smaller datasets created during the course of the project) and to outline one possible way they can be used in an undergraduate research setting to advance students' statistical and research skills. The phases of the project as implemented are described as well as possible modifications to accommodate different classroom settings or levels of student readiness. In addition, appendices are included with relevant R code and field descriptions for the merged and simplified datasets.

2. Data Sources

The ACLED (Armed Conflict Location and Events Dataset) is a comprehensive source of information regarding the location, dates, and types of political violence on the African continent since 1997 (other continents are included, but this project was limited to Africa). It is a large dataset with thousands of individual records of political violence gathered from multiple sources. These data are meant for public dissemination and can be downloaded (including a codebook) at http://www.acleddata.com/data/ (see Figure 1) (Raleigh, Linke, Hegre, and Karlsen 2010). The available datasets are in Microsoft Excel format. Table 1 gives the names of some of the fields used in this project and their descriptions. Further information regarding the fields can be found in the codebook. Other data such as the primary actors and number of fatalities are also available and invite new questions of interest not considered in this paper.

The conflict data used in this paper can be accessed in a tab-delimited text file at: http://www.amstat.org/publications/jse/v20n3/delzell/conflictdata.txt.

The conflict data used in this paper is also available in an Excel file at: http://www.amstat.org/publications/jse/v20n3/delzell/conflictdata.xlsx.

A documentation file for the data set can be accessed in a text file at: http://www.amstat.org/publications/jse/v20n3/delzell/conflictdoc.txt.

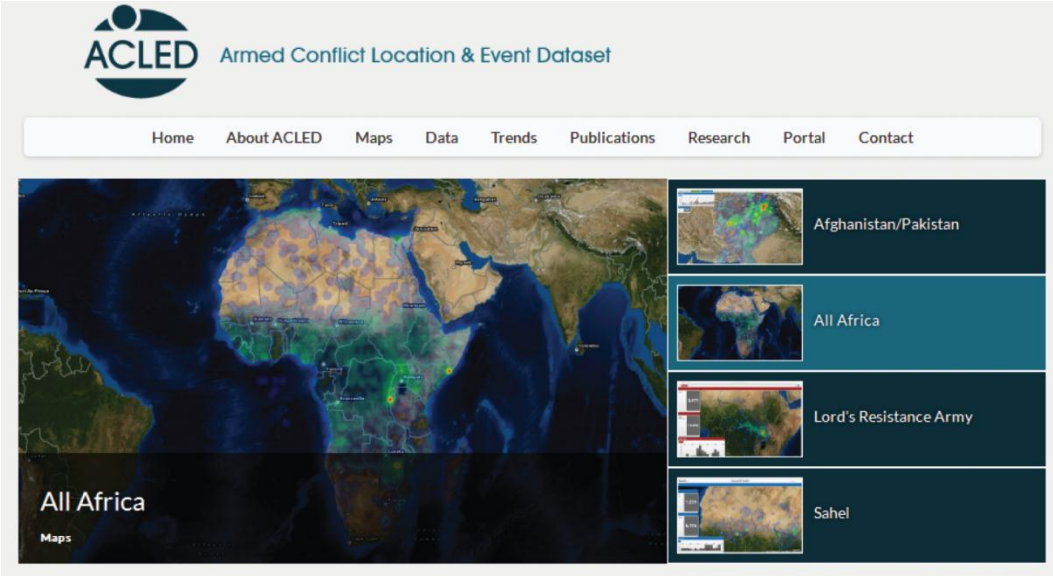

Figure 1: Screenshot of the ACLED homepage (http://www.acleddata.com/)

The National Climatic Data Center (http://www.ncdc.noaa.gov/oa/ncdc.html; see Figure 2) of the U.S. Department of Commerce maintains various datasets housing global climate information through the NOAA (National Oceanic and Atmospheric Administration) Satellite and Information Service. Various weather stations around the world record temperature, precipitation and other climatological information that is publicly available and in various formats. The GSOD (Global Summary of the Day) is one data source maintained by NNDC (NOAA's National Data Centers) that is a compilation of global historical daily temperature, precipitation, and many other meteorological measures.

Table 1: Selected fields from the ACLED dataset

Field Name	Description
EVENT_DATE	date of conflict event (day/month/year)
YEAR	year of conflict event
EVENT_TYPE	Coded as Violence against civilians, Riots, Battles, or Other (nonviolent activity; these records were ignored)
LATITUDE	geographic latitude
LONGITUDE	geographic longitude

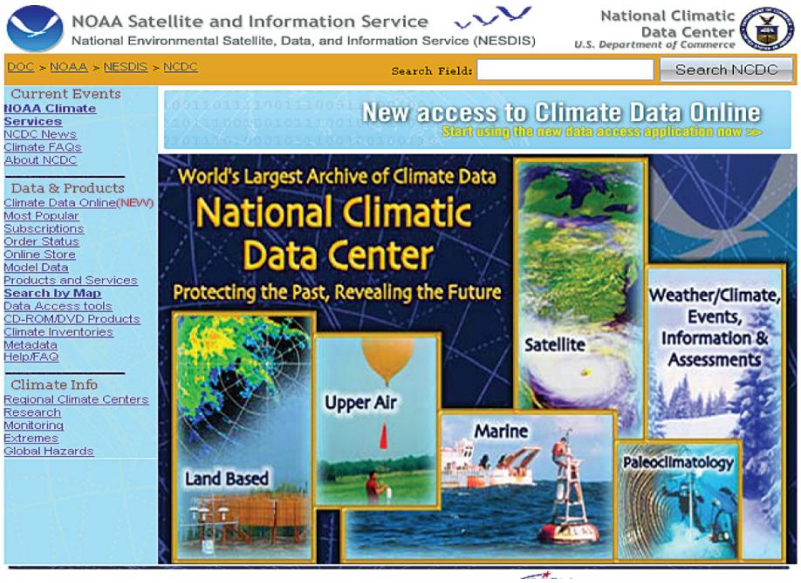

Figure 2: Screenshot of the NOAA Satellite and Information Service

Downloading the correct climate data efficiently can be daunting. The simplest way to attain the GSOD is to navigate to http://www7.ncdc.noaa.gov/CDO/cdo and choose "Geographic Region". It is then possible to download climate data for Africa for the years corresponding to the ACLED data events under consideration (for this project, 1997 – 2010). When the file is available, a station list is also available that provides the latitudes and longitudes for all of the individual weather stations. Table 2 gives a sample of the fields available from the GSOD dataset and their descriptions.

> *Potential Pitfall: When students download the GSOD, they may feel they have used the website incorrectly if they don't realize the large amount of data they are downloading (especially if internet speed is slow). It would be wise to have them practice downloading a single year's worth of data from a single location before attempting to download a large amount of data.*

The GSOD data used in this paper can be accessed in a text file at:
http://www.amstat.org/publications/jse/v20n3/delzell/weeklydata.txt.

The GSOD data used in this paper is also available in an Excel file at:
http://www.amstat.org/publications/jse/v20n3/delzell/weeklydata.xlsx.

A GSOD documentation file for the data set can be accessed in a text file at:
http://www.amstat.org/publications/jse/v20n3/delzell/weeklydoc.txt.

Table 2: Selected fields from the GSOD dataset

Field Name	Description
STN	weather station number
YEAR	Year
MODA	month and day
DEWP	mean dew point for the day in F°
SLP	mean sea level pressure for the day in millibars
STP	mean station pressure for the day in millibars
VISIB	mean visibility for the day in miles
MXSPD	maximum sustained wind speed for the day in knots
MAX	maximum temperature for the day in F°
SNDP	snow depth for the day in inches

A major challenge that was part of the research project was to merge temperature data from the GSOD with armed conflict data from the ACLED and use appropriate statistical methodologies to determine if evidence exists to claim that climate has some effect on the type or "amount" of armed conflict in African countries. The students chose to consider only the temperature measure of climate, but with the varied number of data fields included in these sources, there are many other research questions that could be posed requiring various levels of statistical expertise.

> *Alternative Application: It is not necessary to combine these sources. Separately they provide many teaching opportunities. For example, various questions could be asked using the ACLED alone (such as potential relationships between conflict count/types and actors or regions). Students could also investigate the types of conflicts in certain regions and conduct research to provide possible historical explanations.*

3. Phases of the Project

The project was broken up into four phases. The first phase involved the acquisition of the data from the websites listed in Section 2 and the merging of the data sources into a single text file. The goal of this phase was to allow the students the opportunity to become familiar with the data's structure, which is a necessary step in any research project. This step is often skipped by students who are tempted to simply begin data analysis. In the second phase of the project the students spent time plotting the data in various ways in order to gain a better sense of appropriate statistical tests and, again, to gain familiarity with the data. In the third phase the students began analyzing the data and also assessed the adequacy of the methods used, especially in terms of statistical assumptions. The fourth phase consisted of a paper and a presentation.

3.1. Data Acquisition and Technology – Phase 1

The students downloaded the data files and spent time determining how to merge the two sets of data for further analysis. This primarily consisted of a determination of how to assign a temperature (we chose to use the maximum daily temperature) to a given event in the ACLED.

This was a not a simple decision, as the weather stations were often not in the same location as a conflict event. In addition, climate information was not always available for the day a conflict event occurred. The students had to decide the relative importance of a climate record being close in both time and space to the conflict record and had to devise a defensible way of relating these two distances. These students chose a simple minimization formula that converted delays in time to miles (10 miles = 1 day). In other words, a difference in time of one day was taken to be equivalent to a difference in space of 10 miles. Then, for a given conflict event, the nearest weather station in both time and distance was chosen to contribute the temperature for that day using the latitude and longitude values for both the conflict event and weather station. For example, one ACLED conflict event in Uganda has the following data:

EVENT_DATE(day/month/year)	YEAR	EVENT_TYPE	LATITUDE	LONGITUDE
16/06/2003	2003	Battle	3.250	32.141

Using the minimization algorithm described in the preceding paragraph, the corresponding climate record using the GSOD data and the station information is,

STN	LATITUDE	LONGITUDE	YEAR	MODA(month/day)	MAX(TEMP)
636300	2.750	32.333	2003	06/16	80.6

These data were merged to create the following:

EVENT_DATE	YEAR	EVENT_TYPE	MAX(TEMP)
16/06/2003	2003	Battle	80.6

The actual merging process would be difficult to accomplish in Excel, but is rather simple in R (R Development Core Team 2011) or in SQL, which these students chose to use. However, the merged data is available for download for shorter projects (this phase took approximately 4 weeks) as *conflictdata.txt* or *conflictdata.xlsx*. A description of the fields included in this dataset is included in Appendix A.

> *Helpful Hint: This is a wonderful opportunity for students to propose various solutions to the merging problem and debate amongst themselves as to which solution is the most appropriate. It is a great way for students to see the complexity and choice that sometime arises in real research. Many algorithms could be used to choose the contributing weather station. The learning process for the students is in the debate and the realization that this choice should be defensible to others.*

During the course of this phase the students were encouraged to consider a very important assumption for many statistical methods, the assumption that the observations are independent. In order to have this discussion they were forced to define the population and the sample as it related to the population. For example, do we assume that our data is all African conflict from 1997 to present? Is our population past conflict or future conflict? If future, is our sample a representative random sample? These discussions were very fruitful and led to deep reflection on the importance of assumptions and definitions.

Potential Pitfall: The ACLED data themselves are complex. The information regarding the individual events are compiled from multiple sources and are consistently being updated to reflect new information. The questions listed above do not all have definitive answers, and that can be frustrating and disturbing to students. The discussions are important and they are a useful learning experience for students.

As a result of these discussions, the decision was made by the students to aggregate conflict events by week and record the total number of conflict events per type. The average maximum daily temperature for that given week was kept. This was done as an effort to correct (at least partially) potential dependence of events that occur close together in time. This second aggregated dataset is also available for download as *weeklydata*.txt or *weeklydata.xlsx*. A description of the fields included in this dataset is included in Appendix B.

Potential Pitfall: Weeklydata.txt does not include records for weeks with zero conflict events. Students might not catch this; when analyses are run on these data they are, in essence, conditioning on the fact that TotalCount > 0.

3.2. Visualization – Phase 2

After the data were merged, the students were given a week to explore the data visually. They learned how to create histograms, bar charts, boxplots and scatterplots. The aim was to determine which specific hypotheses the students should investigate with statistical methods that would answer our general question of relationship between conflict and climate. An additional benefit of this phase of the project was that students learned that some visualizations of data are not informative due to the characteristics of the data themselves. Figures 3, 4, & 5 show examples of plots created in R (the code is provided in Appendix C). In Figure 5 a temperature grouping is given. This is discussed in Section 3.3.

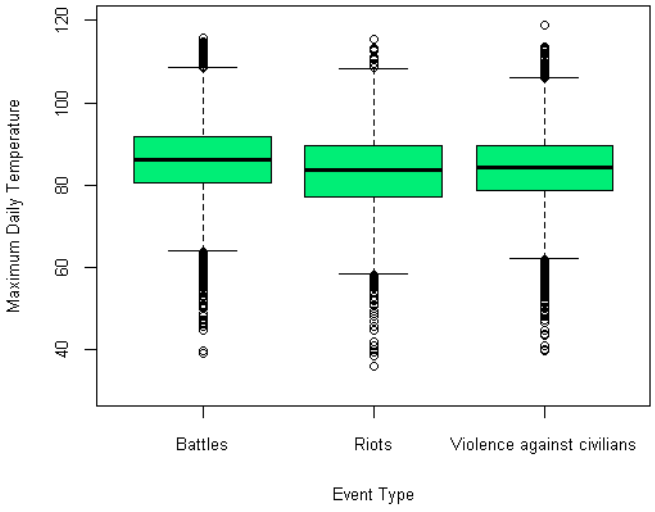

Figure 3: Boxplots of the maximum daily temperature by conflict event type (from *conflictdata.txt*)

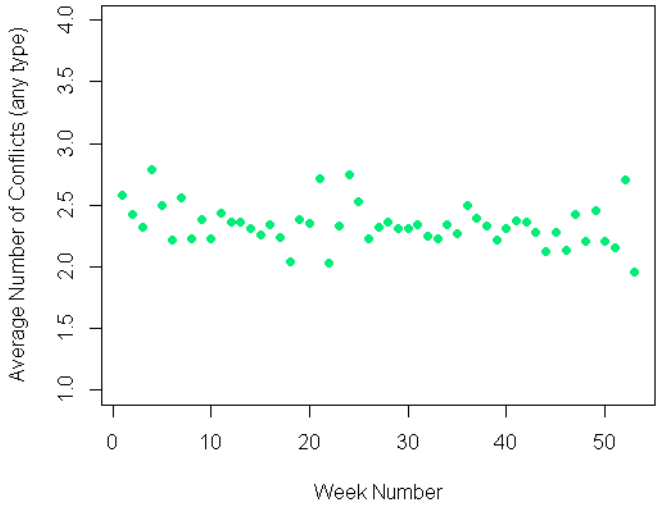

Figure 4: Scatterplot of average number of conflicts over time (from *weeklydata.txt*)

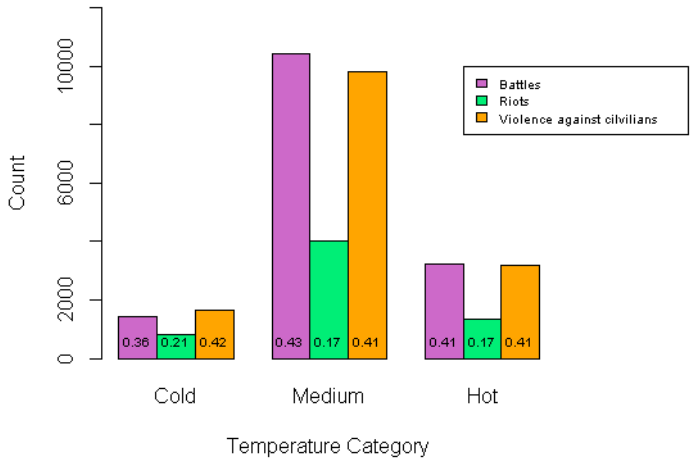

Figure 5: Bar graph of the number of occurrences of various types of conflict by temperature designation (from *weeklydata.txt*)

3.3. Statistical Analyses – Phase 3

A variety of analyses were performed with these data. The two most fruitful are reported here as an example as to the various types of methods that are appropriate for these data. A simple chi-square test is discussed as an example of a simple analysis. In addition, an ANOVA was performed followed by a bootstrapped ANOVA. In all cases reported in this section, the analyses were performed using the merged data that had been aggregated by week (*weeklydata.txt*). The R code for all of these methods is included in Appendix C. The students referred to two textbooks for most of their work on this project (Ramsey and Schafer 2002; Wackerly et al. 2008).

As noted above, each type of violent conflict event was classified as one of three types (Battles, Riots, or Violence against civilians). One investigation focused on the relationship extreme temperature might have on the type of conflict event. The students had noticed that there was a lower proportion of Battles and a higher proportion of Violence against civilians during weeks with colder temperatures and chose to classify events as belonging to a particular temperature category. Temperatures for each event were classified as Cold, Medium, or Hot (C, M, or H) based on the distance a given temperature was from the mean for its weather station. Temperatures below/above one standard deviation below/above the mean were classified as Cold and Hot, respectively. Temperatures within one standard deviation of the mean were classified as Medium. This was, again, an ad hoc classification that the students discussed at length. Figure 5 indicates that there might be evidence of a dependency (while perhaps not strong) between conflict type and temperature category.

> *Potential Pitfall: Students might think that there is evidence of a dependence between temperature category and conflict type simply because the Medium temperature category has higher bars. However, the temperature categories themselves were calculated in such a way that it is expected that approximately 68% of the temperatures would be categorized as Medium. It is the change in proportions of conflict types across temperature categories that gives the visual evidence of a potential dependence.*

However, independence of observations remained a concern. For these reasons, conflicts of the same type in the same week counted as a single observation in this implementation of the chi-square test for independence. The observed counts are given in Table 3. A chi-square test with a null hypothesis of independence between conflict type and temperature category was performed, resulting in a p-value of 6.2×10^{-5} (test statistic = 24.55, d.f. = 4).

Table 3: Observed counts from a chi-square analysis

Conflict Type/Temp Category	Cold	Medium	Hot
Battles	838	4612	1352
Riots	621	2649	765
Violence against civilians	1003	5419	1576

> *Alternative Application: Have the students graph the numbers in Table 3 as a bar graph similar to Figure 5 and discuss the reasons for the difference. This is also a good time to discuss statistical significance versus practical significance. As discussed in Section 3.4, many results were statistically significant, but not necessarily practically significant.*

Another investigation that eventually led to a bootstrapping implementation was the consideration of the average number of conflicts per week. For example, the students considered that when Battles occur, it could be the case that a higher number of Battles tend to occur in weeks with higher temperatures. Analyses were performed for each conflict type using R's oneway.test function. A Welch's ANOVA (which does not assume equal variances) was performed to detect differences of mean conflicts per week (conditioned on conflict having occurred that week) by temperature type. For each ANOVA performed, the null hypothesis was

that there is no difference in the mean number of conflicts per week between the three temperature types. The results are reported in Table 4.

Table 4: Results from ANOVA analyses

Conflict Type	Sample Means			*F*-Statistic (d.f.)	*p*-value
	Cold (s.d.)	Medium (s.d.)	Hot (s.d.)		
Battles	0.672 (1.46)	1.009 (2.18)	1.121 (2.30)	49.993 (2,5090)	2.2×10^{-16}
Riots	0.396 (0.84)	0.387 (0.95)	0.471 (1.24)	5.551 (2,4456)	0.004
Violence against civilians	0.787 (1.91)	0.951 (1.70)	1.112 (2.17)	15.700 (2,4196)	1.6×10^{-7}

Helpful Hint: The socio/political meanings of the group means for this test are subtle. One can present this particular ANOVA to students in a classroom setting and have them discuss the physical meaning of the average number of conflicts per week given the condition that there is at least one conflict that has/will occur. One interpretation is that this is measuring the longevity or severity of conflict when it occurs.

The students were then instructed to check their assumptions. Group histograms were extremely right-skewed and indicated that perhaps the assumption of normality was violated to the extent that a nonparametric approach should be considered. This is also evidenced by the large standard deviations in Table 4.

Helpful Hint: A right-skewed histogram is to be expected with count data that also has a small sample mean.

Alternative Application: The weekly data are conditioned on any type of conflict occurring during a given week. Therefore, for a given conflict type there are many records with no occurrences during a given week (for example, for a particular location, a week contains a Riot, but not a Battle). Students could choose to model these data with a Poisson-like discrete distribution and consider the properties of the distributions and how they might differ by temperature categories and conflict type. This would give students the opportunity to use data in the creation of probability distributions.

The students investigated using a bootstrapped version of ANOVA that doesn't assume an *F* distribution for the test statistic (only the bootstrapped ANOVA for the Battles conflict type is presented for illustration). They spent approximately two weeks learning about bootstrap methods. In essence, a bootstrapped ANOVA translates all samples (three in this example: C, M, H) to have means equal to the global mean, regardless of temperature classification. In this manner the data are forced to agree with the ANOVA null hypothesis of equal group means. Resampling is done with these translated data sets and test statistics are computed. The bootstrapped *p*-value is the proportion of bootstrapped test statistics exceeding the observed test statistic. A classic bootstrap reference is by Efron and Tibshirani (1994). A very accessible and

somewhat short introduction to bootstrapped ANOVA is found in Hayes (2005). Most students should find Hayes' paper very easy to both read and implement.

The students used R to generate 10,000 bootstrapped samples (sampling from the translated samples) and an *F*-statistic generated from each. Figure 6 shows the histogram of test statistics and overlaid in red is the theoretical *F* distribution assumed with the Welch's ANOVA analysis. There is very little difference between the theoretical curve and the shape of the histogram. The bootstrapped *p*-value obtained was 0, as no bootstrapped test statistics (assuming the null hypothesis of equal means) were greater than the original test statistic of 49.993. Students often read that the assumption of normality is often not crucial for ANOVA; the bootstrapping analysis gave them the opportunity to observe, that in this case, normality was not a critical assumption and the ANOVA procedure was robust to the assumption of normality. While the nonparametric bootstrapping procedure is valid, both analysis methods resulted in the same inference.

> *Alternative Application: Students could go a step further with these analyses and compute individual confidence intervals for means or differences in means between the temperature category groups. In so doing, issues involving multiple comparisons could be addressed.*

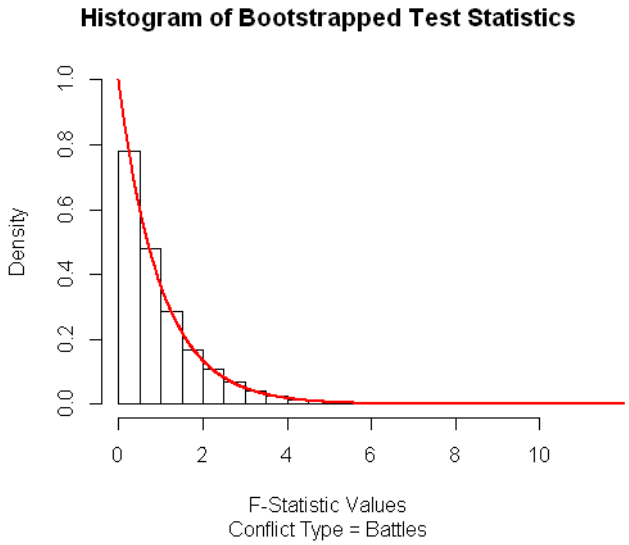

Figure 6: Histogram of bootstrapped *F*-statistics from 10,000 samples; theoretical distribution overlaid in red

3.4. Communication of Results– Phase 4

After the analysis, the students began writing their results in a format suitable to publication in a scientific journal and also prepared to present their findings to professors with expertise in the subject of violent political conflict. In this research project, we had the privilege of presenting to a group of five faculty from the Political Science and International Relations department and one faculty member from the Economics department. Because of this opportunity the students focused much more on the presentation than the paper. They spent many hours developing a clear narrative for their presentation and practiced it many times beforehand. This was an

excellent experience for them and the conversation we had following the presentation with these faculty was extremely influential on the students' view of research. They had pondered many questions about African conflict over the course of the project and were able to pose many of those questions to the professors. These questions ranged from simple inquiries regarding data from particular African countries, to more complex questions regarding the necessary assumption of independence that these data seemed to violate. They commented numerous times that they not only learned much more about these data during the discussion, but also gained new respect for experts in the fields of economics and political science.

In addition, the students had to explain to the faculty that while the results were statistically significant, their results might not indicate a practically significant relationship between conflict and climate. For example, consider Table 4. The students were initially very excited that the *F*-statistics were all significantly large to induce small *p*-values. However, they were encouraged to consider the ANOVA test for the Riots conflict type. While the *p*-value is small, the largest difference between sample means is less than 0.1. This would indicate that, on average, hot temperatures are associated with 0.1 more riots per week than riots occurring during cold temperatures. The students were asked to consider how such a small difference would be detectable, even though there might not be any practically significant difference. By considering the mathematical form of the test statistic itself, they eventually came to the correct conclusion that the very large number of data values allowed for detection of very small differences.

4. Student Gains

After the completion of the project, the students were asked to write a summary of their experience. They all reported that this research experience was extremely valuable in their study of statistics (see Appendix D for examples of written comments). In particular, they were impressed by the complexities that often arise in real statistical analysis outside of a traditional classroom setting.

Specifically, these data sources provided the students with the following experiences:

1. They acquired data and implemented a merging strategy they had to defend.
2. They had to convert a general question of interest into specific statistical hypotheses.
3. They learned a new statistical method that they had not been previously exposed to (bootstrapping).
4. They learned that statistical significance does not always imply practical significance. They saw first-hand that large amounts of data can often lead to low *p*-values.
5. They were able to use both a parametric (ANOVA) and nonparametric (bootstrapped ANOVA) method to test hypotheses of equal means. They were able to see a close association between the theoretical distribution of a test statistic (under the assumption of normality) and an estimated distribution of the same test statistic (not assuming normality).
6. They were given the opportunity to communicate their findings to non-statisticians who did not possess expertise in statistical science, but who did possess expertise in the field of application the students' analyses pertained to. Practicing statisticians are often

expected to explain their results verbally to non-statisticians in an accessible and easily understood way.

5. Conclusion

The two source datasets (the ACLED and the GSOD) provide excellent opportunities for students to do real research in a mentoring setting with a professor. For smaller, more traditional classroom projects the simplified and merged datasets (*conflictdata.txt* and *weeklydata.txt*) allow for an abundance of statistical application experiences. These data also can also be used to stimulate classroom discussion as to the practical meaning of measures used in analyses, the determination of merging strategies, and the use of nonparametric approaches. It might be beneficial to have a meteorologist or political scientist participate in a discussion on the nature and possible particularities of these data. For example, in the discussion following the students' presentation, they were informed by a few of the political science professors that the results in Table 4 for the Battles conflict type made logical sense as many strategic offensives (Battles) are planned for the spring (a.k.a. "the spring offensive") and fewer occur in the colder months. The students involved in this project were very excited to work with 'real' data of such an interesting nature; political conflict and climate stories abound in the news and these data give students an opportunity to use the same data professional scientists around the world are currently using to understand very complex, but important issues.

Appendix A

The dataset designated *conflictdata.txt* is available in tab-delimited format or as an Excel file at:
http://www.amstat.org/publications/jse/v20n3/delzell/conflictdata.txt.
http://www.amstat.org/publications/jse/v20n3/delzell/conflictdata.xlsx.

A documentation file for the data set can be accessed in a text file at:
http://www.amstat.org/publications/jse/v20n3/delzell/conflictdoc.txt.

The data file was created as the result of a merging process of multiple fields from the ACLED and GSOD datasets and has the following fields:

Name	**Description**
Year	year of conflict event as assigned by ACLED
EventType	type of conflict event (choices are Riots, Battles, or Violence against civilians) as assigned by ACLED
Actor1	conflict actor as assigned by ACLED
Actor2	conflict actor as assigned by ACLED
Country	country of conflict event as assigned by ACLED
Region	region of conflict event as assigned by ACLED
Location	location of conflict event as assigned by ACLED
ConflictLat	geographic latitude of the conflict event as assigned by ACLED
ConflictLong	geographic longitude of the conflict event as assigned by ACLED
StationID	weather station identification code as assigned by the GSOD
YrMoDy	further date information for the conflict event in Year/Month/Day format as assigned by ACLED
MaxTemp	maximum temperature for the day assigned to the conflict event in degrees Fahrenheit as obtained from the GSOD data
StationName	name of the weather station as assigned by the GSOD
StationLong	geographic latitude of the weather station as assigned by the GSOD station list
StationLat	geographic longitude of the weather station as assigned by the GSOD station list
TempCat	a temperature category {cold(C), medium(M), or hot(H)} assigned based on temperatures recorded at a given weather station (for given weather station (StationID), temperatures below/above one standard deviation below/above the mean for that station were classified 'cold'/'hot'; other values were classified 'medium')

Appendix B

The dataset designated *weeklydata.txt* is available in tab-delimited format or as an Excel file at:
http://www.amstat.org/publications/jse/v20n3/delzell/weeklydata.txt.
http://www.amstat.org/publications/jse/v20n3/delzell/weeklydata.xlsx.

A documentation file for the data set can be accessed in a text file at:
http://www.amstat.org/publications/jse/v20n3/delzell/weeklydoc.txt.

The data file was created as the result of an aggregation process applied to *conflictdata.txt* and has the following fields:

Name	**Description**
Year	year of conflict event as assigned by ACLED
Week	The week in a given year (1-53). The R function 'strptime' was used to calculate the day of the year from the YrMoDy field in *conflictdata.txt*. This value was divided by 7 and rounded up to the nearest integer to obtain a week number.
StationID	weather station identification code as assigned by the GSOD
BattlesCount	total number of conflicts categorized as Battles during the week (each event in *conflictdata.txt* was assigned a week number and the total events categorized as 'Battles' for the given week, year, and stationID were calculated)
VACCount	total number of conflicts categorized as Violence against civilians during the week (each event in *conflictdata.txt* was assigned a week number and the total events categorized as 'Violence against civilians' for the given week, year, and stationID were calculated)
RiotsCount	total number of conflicts categorized as Riots during the week (each event in *conflictdata.txt* was assigned a week number and the total events categorized as 'Riots' for the given week, year, and stationID were calculated)
TotalCount	total number of conflicts during the week (sum of BattlesCount, RiotsCount, and VACCount for a given year, week, and stationID)
HighMaxTemp	highest daily maximum temperature for the week assigned to the conflict event in degrees Fahrenheit (for the events with the same year, week, and stationID, this is the maximum value of the field MaxTemp in *conflictdata.txt*)
TempCat	the temperature category {cold(C), medium(M), or hot(H)} belonging to the HighMaxTemp value above (for given weather station (stationID), temperatures below/above one standard deviation below/above the mean for that station were classified 'cold'/'hot'; other values were classified 'medium')

Appendix C

Sample R Code

```
conflict = read.delim("conflictdata.txt",sep="\t")
weekly = read.delim("weeklydata.txt",sep="\t")

#Figure 3
par(mar=c(5.1,4.1,2.1,2.1))
boxplot(conflict$MaxTemp ~
conflict$EventType,cex.lab=.8,cex.axis=.8,ylim=c(30,120),col=colors()[612],ylab="Maximum
Daily Temperature",xlab="Event Type")

#Figure 4
weekagg = aggregate(weekly$TotalCount,by=list(weekly$Week),mean)
plot(weekagg[,1],weekagg[,2],xlab="Week Number",ylim=c(1,4),ylab="Average Number of
Conflicts (any type)",pch=16,col=colors()[612])

#Figure 5
attach(weekly)
Ctemp = subset(weekly,TempCat=="C")
Mtemp = subset(weekly,TempCat=="M")
Htemp = subset(weekly,TempCat=="H")
bin11 <- sum(Ctemp$BattlesCount)
bin12 <- sum(Mtemp$BattlesCount)
bin13 <- sum(Htemp$BattlesCount)
bin21 <- sum(Ctemp$RiotsCount)
bin22 <- sum(Mtemp$RiotsCount)
bin23 <- sum(Htemp$RiotsCount)
bin31 <- sum(Ctemp$VACCount)
bin32 <- sum(Mtemp$VACCount)
bin33 <- sum(Htemp$VACCount)
M <- as.table(rbind(c(bin11, bin12, bin13), c(bin21, bin22, bin23), c(bin31, bin32, bin33)))
dimnames(M) <- list(type=c("Battles","Riots", "Violence against
cilvilians"),Temp=c("Cold","Medium","Hot"))

prop11 <- bin11/(bin11+bin21+bin31)
prop12 <- bin12/(bin12+bin22+bin32)
prop13 <- bin13/(bin13+bin23+bin33)
prop21 <- bin21/(bin11+bin21+bin31)
prop22 <- bin22/(bin12+bin22+bin32)
prop23 <- bin23/(bin13+bin23+bin33)
prop31 <- bin31/(bin11+bin21+bin31)
prop32 <- bin32/(bin12+bin22+bin32)
prop33 <- bin33/(bin13+bin23+bin33)
```

```
P =  as.table(rbind(c(prop11, prop12, prop13), c(prop21, prop22, prop23), c(prop31, prop32, prop33)))
dimnames(P) <- list(type=c("Battles","Riots", "Violence against cilvilians"),Temp=c("Cold","Medium","Hot"))
M <- as.table(rbind(c(bin11, bin12, bin13), c(bin21, bin22, bin23), c(bin31, bin32, bin33)))
dimnames(M) <- list(type=c("Battles","Riots", "Violence against cilvilians"),Temp=c("Cold","Medium","Hot"))
par(xpd=TRUE,mar=c(5.1 ,4.1, 4.1, 6.1))
xpos = barplot(M,ylab="Count",xlab="Temperature Category", ylim=c(0,12000),col=colors()[c(511,612,498)], beside=TRUE)
legend(x=10, y = 10000, legend=c("Battles","Riots", "Violence against cilvilians"), fill = colors()[c(511,612,498)],cex=.6)
text(xpos, 0, round(P, 2),cex=.6,pos=3)

#Chi-Square
xbin11 <- nrow(subset(weekly,BattlesCount>0 & TempCat=="C"))
xbin12 <- nrow(subset(weekly,BattlesCount>0 & TempCat=='M'))
xbin13 <- nrow(subset(weekly,BattlesCount>0 & TempCat=='H'))
xbin21 <- nrow(subset(weekly,RiotsCount>0 & TempCat=='C'))
xbin22 <- nrow(subset(weekly,RiotsCount>0 & TempCat=='M'))
xbin23 <- nrow(subset(weekly,RiotsCount>0 & TempCat=='H'))
xbin31 <- nrow(subset(weekly,VACCount>0 & TempCat=='C'))
xbin32 <- nrow(subset(weekly,VACCount>0 & TempCat=='M'))
xbin33 <- nrow(subset(weekly,VACCount>0 & TempCat=='H'))
M <- as.table(rbind(c(xbin11, xbin12, xbin13), c(xbin21, xbin22, xbin23), c(xbin31, xbin32, xbin33)))
dimnames(M) <- list(type=c("Battles","Riots", "Violence against cilvilians"),Temp=c("Cold","Medium","Hot"))

Xsq <- chisq.test(M)
print(Xsq) # Prints test summary
Xsq$observed   # observed BattleCounts (same as M)

#ANOVA
oneway.test(BattlesCount ~ TempCat,data=weekly)
oneway.test(RiotsCount ~ TempCat,data=weekly)
oneway.test(VACCount ~ TempCat,data=weekly)

#Bootstrap : BattlesCount only
with(weekly,tapply(BattlesCount,TempCat,mean))
with(weekly,tapply(BattlesCount,TempCat,var))
sizes = with(weekly,tapply(BattlesCount,TempCat,length))

meanstar = with(weekly, tapply(BattlesCount,TempCat,mean))
globalmean = mean(BattlesCount)
grpC = subset(weekly,weekly$TempCat=="C",select= c(BattlesCount,TempCat))
```

```
grpC[,1] = grpC[,1] - meanstar[1] + globalmean
grpM = subset(weekly,weekly$TempCat=="M",select= c(BattlesCount,TempCat))
grpM[,1] = grpM[,1] - meanstar[3] + globalmean
grpH = subset(weekly,weekly$TempCat=="H",select= c(BattlesCount,TempCat))
grpH[,1] = grpH[,1] - meanstar[2] + globalmean
R = 10000
Fstar = numeric(R)
for (i in 1:R) {
  groupC = grpC[sample(1:sizes[1],size=sizes[1],replace=T),]
  groupM = grpM[sample(1:sizes[3],size=sizes[3],replace=T),]
  groupH = grpH[sample(1:sizes[2],size=sizes[2],replace=T),]
  bootdata = rbind(groupC,groupM, groupH)
  Fstar[i] = oneway.test(BattlesCount~TempCat, data=bootdata)$statistic
}
hist(Fstar, breaks=20,prob=T,ylim=c(0,1),xlab="F-Statistic Values",main="Histogram of
Bootstrapped Test Statistics",sub="Conflict Type = Battles")
y <- seq(0,12,by=.01)
points(y,df(y,df1=2,df2=5089.974),type='l',col='red',lwd=2)
pvalue = length(Fstar[Fstar>oneway.test(BattlesCount ~
TempCat,data=weekly)$statistic])/10000
pvalue
```

Appendix D

Some student comments…

Student A: "My experience this semester in [the mentored research course] was an incredibly valuable and rich experience that taught me a ton not only about statistics or research, but about leadership, teamwork and organization."

Student B: "I am glad that I took this mentored research seminar because I had a great time, and because I learned a lot about the research process. I now have had a taste of research and I know the experience will help me in my future endeavors." "…I see that research can be a hard process, and that making sure that results are clear and defendable is a difficult procedure to go through."

Student C: "My knowledge of R increased, and I now feel comfortable running basic code. I also really enjoyed working with SPSS, and hope to increase my knowledge of that program."

Student D: "I learned quite a bit this semester about research and about statistics, though it was in a different way than I had learned before." "…working with real data can be very fuzzy, and involves making a lot of judgment calls even before analysis is actually done."

References

Efron, B. and Tibshirani, R. (1994), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Hayes, A. F. (2005), In B. Everitt & D. Howell (Eds.) "One Way Designs: Nonparametric and Resampling Approaches" In *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1468-1474), Chichester, UK: Wiley & Sons.

Hunter, A. B., Laursen, S. L. and Seymour, E. (2007), "Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development," *Science Education*, 91, 36-74.

Lopatto, D. (2004), "Survey of undergraduate research experiences (SURE): first findings," *Cell Biology Education*, 3, 270-277.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (08/13/2012).

Raleigh, C., Linke A., Hegre H., and Karlsen J. (2010), "Introducing ACLED-Armed Conflict Location and Event Data," *Journal of Peace Research*, 47, 1-10.

Ramsey, F. L. and Schafer, D. W. (2002), *The Statistical Sleuth : A Course in Methods of Data Analysis*, Pacific Grove, CA: Duxbury/Thomson Learning.

Wackerly, D. D., Mendenhall, W. and Scheaffer, R. L. (2008), *Mathematical Statistics with Applications*, Belmont, CA: Thomson Brooks/Cole.

Zydney, A. L., Bennett, J. S., Shahid, A. and Bauer, K. W. (2002), "Impact of undergraduate research experience in engineering," *Journal of Engineering Education*, 91, 151-158.

Darcie Delzell
Wheaton College
Mathematics Department
501 College Ave.
Wheaton, IL 60187
darcie.delzell@wheaton.edu
Phone: 630-752-5872
Fax: 630-752-5996
