Most Lethal and Deadly Storm Types in 2011

Jim Callahan

September 26, 2015

Synopsis: This paper is in response to **two questions:**

1. **Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?** According to the 2011 data: Tornadoes.
2. **Across the United States, which types of events have the greatest economic consequences?** According to the 2011 data: Floods.

These questions are addressed using the **"R"** statistical language and data from the **National Oceanic and Atmospheric Administration (NOAA) "Storm Data" database**.

Data Processing The **R** language is an open source version of the the **S** language developed at Bell Labs during the "golden age" that also produced the **Unix** operating system and the **C** programming language. See: https://www.R-project.org/ for the **R** language and for the history of S and R languages, see: https://www.r-project.org/nosvn/conferences/useR-2006/Slides/Chambers.pdf ,

http://blog.revolutionanalytics.com/2014/01/john-chambers-recounts-the-history-of-s-and-r.html

and https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf

Unlike **C**, **R** is an interpretive command language where the user types commands at the command line and gets an immediate response:

```r
2+2
```

```
## [1] 4
```

```r
sqrt(25)
```

```
## [1] 5
```

```r
GaussDidThisInHisHeadInElemetarySchool <- sum(1:100)
print(GaussDidThisInHisHeadInElemetarySchool)
```

```
## [1] 5050
```

The last of these three examples is a problem solved by famous mathematician **Carl Friedrich Gauss (1777-1855)** while he was an eight year old child math prodigy in elementary school. He solved it in his head, amazing his teacher. For more of the story see: "Clever Carl" http://nrich.maths.org/2478/index?nomenu=1'

For those of us who are not (child or adult) math prodigies we can solve the problem with **R** by typing **sum(1:100)** at the command line. The **"<-"** assigns the result of the function to the variable name on the left.

The command line commands can be combined in simple text files **"scripts"** or combined with compliled programs (compiled in **FORTRAN**, **C** or **C++**), data and documentation to form complete **"packages"**.

Many open source statisical **"packages"** have been written in **R** making the complete system, the base language plus the optional downloadable statistical packages competitive with traditional statistical systems such as **SAS** or **SPSS**.

When we load the **NOAA "Storm Data" data file** into the **R** statistical system, we will also be using the "**<-**" to assign the result to the variable name on the left.

```
filename <- "~/GitHub/RepData_PeerAssessment2/data/repdata%2Fdata%2FStormData.csv.bz2"
NOAA <- read.csv(filename,
                 stringsAsFactors = FALSE )
```

```
# Select Columns of interest:
#        Primary key:    REFNUM
#        Date and Time:  BGN_DATE, BGN_TIME, TIME_ZONE,
#        Location:       STATE, COUNTY, COUNTY_END, LATITUDE, LONGITUDE,
#        Type of Storm:  EVTYPE,
#        Damage:         PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP,
#        Casualties:     FATALITIES, INJURIES

ColumnSubset <- c("REFNUM", "BGN_DATE",
                  "STATE", "COUNTY", "COUNTY_END", "LATITUDE", "LONGITUDE",
                  "EVTYPE", "PROPDMG", "PROPDMGEXP", "CROPDMG", "CROPDMGEXP",
                  "FATALITIES", "INJURIES")

storms <- NOAA[, ColumnSubset]
str(storms)
```

```
## 'data.frame':    902297 obs. of  14 variables:
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
```

```
# Select Rows of interest
# Select years of interest (Since 2001 "21st Century Storms" or "last 25 years")

# Step 1: Create a year variable from the date;
datetime = as.POSIXct(storms$BGN_DATE, "%m/%d/%Y %H:%M:%S", tz = "")
storms$year <- format(datetime, "%Y")

# Step 2: Filter by year: "21st Century Storms"
storms <- storms[storms$year >= "2001", ]
```

```r
# Garbage Collection: Remove the original NOAA database from memory
rm(NOAA)
```

We can't simply sum the property damage variable, because some values are in thousands ("K") and others are in millions ("M"). So, we need to rescale the variables using the appropriate multipliers. The same goes for the crop damage variable.

```r
# Scale Property Damage and Crop Damage by thouands and millions
storms$PropertyDamage <- storms$PROPDMG   # Copy data; retain orginal intact

storms$PropertyDamage <- ifelse(storms$PROPDMGEXP == "K",
                                storms$PropertyDamage * 1000,
                                storms$PropertyDamage)

storms$PropertyDamage <- ifelse(storms$PROPDMGEXP == "M",
                                storms$PropertyDamage * 1000*1000,
                                storms$PropertyDamage)


storms$CropDamage     <- storms$CROPDMG   # Copy data; retain orginal intact

storms$CropDamage     <- ifelse(storms$CROPDMGEXP == "K",
                                storms$CropDamage * 1000,
                                storms$CropDamage)

storms$CropDamage     <- ifelse(storms$CROPDMGEXP == "M",
                                storms$CropDamage * 1000*1000,
                                storms$CropDamage)
```

Fatalities and Damage by Storm Type and Year NOAA's "Storm Data" database is a comprehensive listing of major storm events. Each tornado is a seperate event and tornados that cross state lines may count as two tornado events. So, to compare the annual impact of "tornadoes" to the annual impact of floods, we have to add up all of the data from the tornado events in a given year. It is helpful to sum by year, because many managers are used to annual summaries. Moreover, while human lives are comparable, it is more problematic to add together damage estimates from the 1950s and 1960s when houses might cost in the tens of thousands of dollars to damage estimates from the current century when the cost of houses are measured in the hundreds of thousands of dollars. While it is simple to multiply by a Consumer Price Index (CPI) or a more specialized housing price index, given the magnitude of the price changes the choice of index could by itself could distort the analysis.

```r
# Tabulate Fatalities and Damage by Storm Type and Year
# Userful R commands include table(), xtabs(), ftable()  or aggregate()
# This use of aggregate() is based on Jared Lander's "R for Everyone" page 123
# where he uses aggregate() on the diamonds data set from the ggplot2 package.
# Template: aggregate(formula, data, FUN, ..., subset, na.action = na.omit)
StormTot <- aggregate(
    formula = cbind(FATALITIES, INJURIES, PropertyDamage, CropDamage) ~ EVTYPE + year,
    data    = storms,
    FUN     = sum)

# Round off total to nearest dollar
# because the estimates are not accurate to nearest penny.
```

```r
StormTot$PropertyDamage <- round(StormTot$PropertyDamage, digits = 0)
StormTot$CropDamage     <- round(StormTot$CropDamage, digits = 0)

#### Rename and put variables in logical order.
#### Order of columns:
StormTot <- StormTot[ , c("year", "EVTYPE", "FATALITIES", "INJURIES",
                          "PropertyDamage", "CropDamage") ]

#### The variables in "stormtot" have been aggregated by type of storm and year
#### and thus the NOAA supplied names reflect the origin of the variable
#### but not its current content, so it is appropriate to rename the variables
#### for display.
ColumnNames <- c("Year", "StormType", "Fatalities", "Injuries", "PropertyDamage", "CropDamage")
colnames(StormTot) <- ColumnNames
```

Results

Health Impact This is in answer to the question:

> **"Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?"**

To prepare a ranking we sort the data and number the rows.

```r
# What type of storms caused the most fatalities in the most recent year (2011)?
MostRecentYear <- max(StormTot$Year)
StormsRankYear <- StormTot[StormTot$Year == MostRecentYear, ]

# This is the sort -- rank by fatalities in the stormrankyear
DeadlyStorms <- StormsRankYear[order(-StormsRankYear$Fatalities, na.last = NA), ]

# Renumber the rows
RowNames <- as.character(1:nrow(DeadlyStorms))
rownames(DeadlyStorms) <- RowNames
```

```r
print( DeadlyStorms[1:25,
         c("StormType", "Fatalities", "Injuries", "PropertyDamage", "CropDamage")]
      )
```

Health Impact in 2011 alone

```
##                    StormType Fatalities Injuries PropertyDamage CropDamage
## 1                    TORNADO        587     6163     4519600705   31361000
## 2                FLASH FLOOD         68       30     1384044700   88447000
## 3                       HEAT         63      611              0          0
## 4                      FLOOD         58       10     4717677453  154872000
## 5          THUNDERSTORM WIND         54      373      381891410  139832000
## 6             EXCESSIVE HEAT         36      138        1143200          0
```

```
## 7                RIP CURRENT          29        27               0          0
## 8                  LIGHTNING          26       194        46978920     112000
## 9             COLD/WIND CHILL          21         1           70000          0
## 10                 HIGH SURF          11        11          222000          0
## 11               STRONG WIND          10        33        16545130   15059000
## 12                 AVALANCHE           9         8           55000          0
## 13                  WILDFIRE           6       116       648318400    9797000
## 14                 HIGH WIND           4        11        41951000   44293000
## 15            TROPICAL STORM           4         1       138742200   24501000
## 16 MARINE THUNDERSTORM WIND           3        14          108800      50000
## 17                  BLIZZARD           2         0         2742000          0
## 18  EXTREME COLD/WIND CHILL           2         1         7035000          0
## 19        MARINE STRONG WIND           2         5          351600          0
## 20            WINTER WEATHER           2         0         1895000          0
## 21             COASTAL FLOOD           1         1        27274000          0
## 22                HEAVY RAIN           1         1        11791000   20713000
## 23                 LANDSLIDE           1         0        21136000      17000
## 24                   TSUNAMI           1         0        53554000          0
## 25              WINTER STORM           1         0        18157000      70000
```

Property Damage This is in answer to the question:

> **Across the United States, which types of events have the greatest economic consequences?**

Again, to prepare a ranking we sort the data and number the rows.

```r
# What type of storms caused the most property damage in the most recent year (2011)?
DamageStorms <- StormsRankYear[order(-StormsRankYear$PropertyDamage, na.last = NA), ]

# Renumber the rows
RowNames <- as.character(1:nrow(DamageStorms))
rownames(DamageStorms) <- RowNames
```

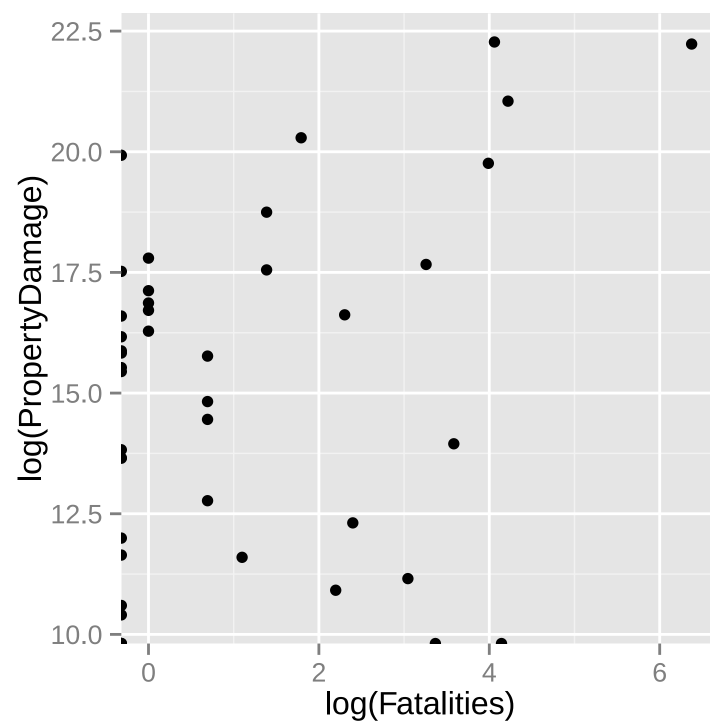
```r
print( DamageStorms[1:25,
         c("StormType", "PropertyDamage", "CropDamage", "Fatalities", "Injuries")]
      )
```

Property Damage in 2011 alone

```
##                  StormType PropertyDamage CropDamage Fatalities Injuries
## 1                    FLOOD     4717677453  154872000         58       10
## 2                  TORNADO     4519600705   31361000        587     6163
## 3              FLASH FLOOD     1384044700   88447000         68       30
## 4                 WILDFIRE      648318400    9797000          6      116
## 5                     HAIL      451329550   82334000          0       31
## 6        THUNDERSTORM WIND      381891410  139832000         54      373
## 7           TROPICAL STORM      138742200   24501000          4        1
## 8                  TSUNAMI       53554000          0          1        0
```

```
## 9                LIGHTNING       46978920     112000          26       194
## 10               HIGH WIND       41951000   44293000           4        11
## 11        STORM SURGE/TIDE       40695000          0           0         0
## 12           COASTAL FLOOD       27274000          0           1         1
## 13               LANDSLIDE       21136000      17000           1         0
## 14            WINTER STORM       18157000      70000           1         0
## 15             STRONG WIND       16545130   15059000          10        33
## 16              HEAVY SNOW       16125300      20000           0         0
## 17              HEAVY RAIN       11791000   20713000           1         1
## 18               HURRICANE       10500000   10500000           0         0
## 19               ICE STORM        7837500      80000           0         0
## 20         LAKESHORE FLOOD        7500000          0           0         0
## 21 EXTREME COLD/WIND CHILL        7035000          0           2         1
## 22            FROST/FREEZE        5540000   13410000           0         0
## 23              WATERSPOUT        5110000          0           0         0
## 24                BLIZZARD        2742000          0           2         0
## 25          WINTER WEATHER        1895000          0           2         0
```

```r
# See Jared Lander's "R for Everyone: Advanced Analytics and Graphics", page 223
library(ggplot2)
ggplot(StormsRankYear,
       aes(x=log(Fatalities), y=log(PropertyDamage))) + geom_point()
```



Do fatalities and storm damage always go together?

Based on the graph there seem to be many events with high property damage (the dots are clustered to the left of the graph, rather than along the diagonal or uniformly spread out) and low fatalities this may be due to evacuations which save human lives when property damage is unavoidable.