The Polar Refresh; Updating Emanuel Parzen's Buffalo NY Snowfall Dataset

Written by Jeffrey C. Miecznikowski

Department of Biostatistics, SUNY University at Buffalo

723 Kimball Tower, 3435 Main St., Buffalo, NY, 14214

Over the last 40 years the Buffalo, New York snowfall dataset has been used as the fundamental example in seminal works demonstrating important statistical developments. This dataset consisting of annual snowfall totals from 1910 to 1972 in Buffalo, NY was originally assembled by Dr. Emanuel Parzen and has been used to establish density estimation, non parametric time series modeling, and violin plots. What makes this dataset so popular? What does an updated version of this dataset reveal? Can Buffalo's snowfall predictions trace its roots to the Civil War era and, ultimately, the formation of the National Weather Service?

The city of Buffalo, New York originated around 1789 and grew quickly after the opening of the Erie Canal in 1825. Buffalo became home to The University of Buffalo (UB), a university that was founded in 1846 and later became part of the State University of New York (SUNY) system. In 1965 the Department of Mathematical Statistics was formed at UB and in 1971 Emanuel Parzen became the Chair of the department. To demonstrate many of his techniques, Dr. Parzen developed the Buffalo snowfall dataset. This dataset consists of 63 observations measuring the annual snowfall amounts in inches (to the nearest tenth) as observed in Buffalo, NY in winters from 1909/10 to

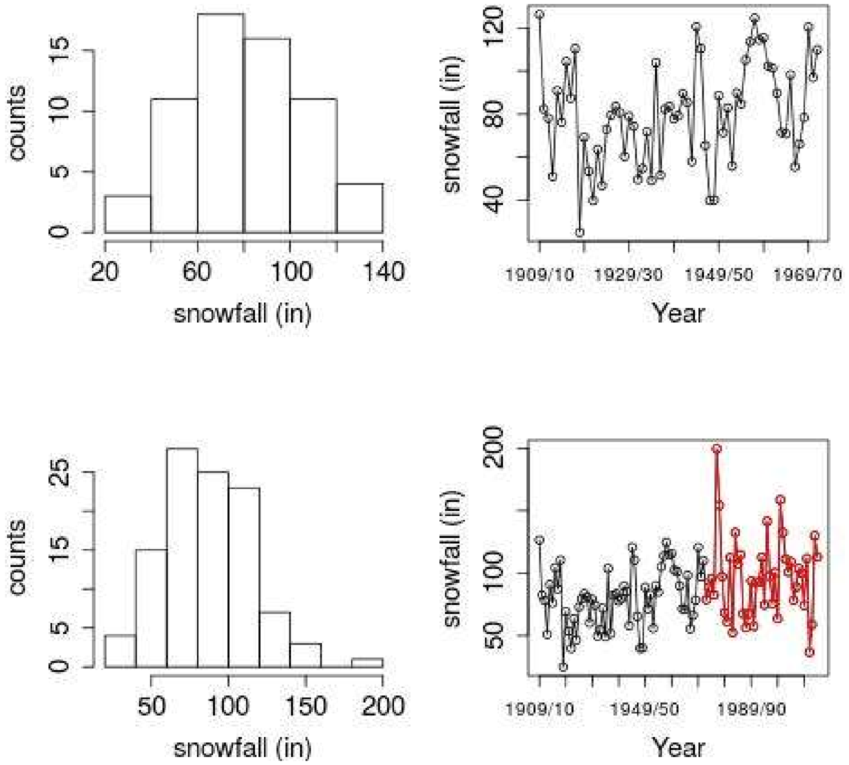

Figure 1 (Top left) Histogram of snowfall data measured in inches in Buffalo, NY from 1909/10 to 1971/72. (Top right) The snowfall data displayed as a time series. (Bottom left) Histogram of the updated snowfall data measured in inches in Buffalo, NY from 1909/10 to 2014/15. The outlying large value refers to snowfall in 1976/77 and is discussed later in the article. (Bottom right) Snowfall data displayed as a time series showing the additional datapoints in red with the original dataset in black.

1971/72. The data are shown in Figure 1 and given in Table 1.

Table 1: Annual Snowfall in inches in Buffalo, NY from 1909/10 to 2014/15. Additional points in red update Parzen's original data and represent 1972/73-2014/15 snowfall amounts.

126.4	82.4	78.1	51.1	90.9	76.2	104.5	87.4	110.5
25.0	69.3	53.5	39.8	63.6	46.7	72.9	79.6	83.6
80.7	60.3	79.0	74.4	49.6	54.7	71.8	49.1	103.9
51.6	82.4	83.6	77.8	79.3	89.6	85.5	58.0	120.7
110.5	65.4	39.9	40.1	88.7	71.4	83.0	55.9	89.9
84.8	105.2	113.7	124.7	114.5	115.6	102.4	101.4	89.8
71.5	70.9	98.3	55.5	66.1	78.4	120.5	97.0	110.0
78.8	88.7	95.6	82.5	199.4	154.3	97.3	68.4	60.9
112.4	52.4	132.5	107.2	114.7	67.5	56.4	67.4	93.7
57.5	92.8	93.2	112.7	74.6	141.4	97.6	75.6	100.5
63.6	158.7	132.4	111.3	100.9	109.1	78.2	88.9	103.8
100.2	74.1	111.8	36.7	58.8	129.9	112.9		

The snowfall dataset has appeared in at least 20 statistical journal articles and is available within MATLAB and many R libraries (dbEmpLikeGof, gss, and LearnEDA) ensuring its continued popularity. In addition, there are at least 10 textbooks that include this dataset in examples and exercises in various statistical frameworks. The popularity of this dataset may be due to several distinguishing features:

1. It encompasses a relatively large time span and contains no missing data.
2. It demonstrates a nearly normally distributed natural phenomenon.
3. It's a dataset that requires no extensive background in statistics or specialty science to understand.
4. It allows for numerous easily obtained variables of mixed type (continuous, ordinal, and factor) to be integrated into the analysis where basic association tests and graphical techniques can be demonstrated.
5. It is a dataset with possibly significant conclusions, that is, discovering accurate predictors for snowfall amounts would allow civic leaders to promptly and effectively allocate resources thus saving time and money.

In Figure 1 and Table 1 we update the dataset to include data after 1972. The updated data were obtained from the national weather service forecast office and can be found at: www.weather.gov/buf/BuffaloSnow.

Kernel Density Estimation

In the general problem of density estimation we are given $X_1, \ldots, X_n$ independent and identically distributed data points from a distribution with cumulative distribution function (CDF) F or probability density function (PDF) f. The goal is to estimate f (or F) using the estimator $\hat{f}(x)$.

In kernel density estimation (kde), the density estimator is given by,

$$\hat{f}(x;h) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right). \quad (1)$$

In general, K is called the kernel function and h is the bandwidth. The kernel function is usually a smooth function with mean 0 that is non-negative and integrates to 1, e.g. the standard Normal pdf. In essence the kde at x is obtained by averaging the kernel function evaluated at $x - X_i$ over the n data points where the snowfall data points X_i nearest to x have the largest values of $K(x - X_i)$. Figure 3 shows the kernel density for the snowfall data with the data points as short vertical bars. For a general setting, the rule-of-thumb bandwidth is given by 1:06sd £ n$^{i\ 1=5}$ where *sd* is the sample standard deviation and *n* is the number of data points.

With our density estimation for annual snowfall, we can determine the estimated 100-year- annual snowfall, in other words, the snowfall amount that has a 1 percent probability of occurring in a given winter. This is simply the 99th percentile of our kernel density estimator (see Figure 3). This estimate can be a measure to characterize extreme annual snowfall amounts and can be used to estimate necessary resources needed in the winter.

With the updated data, it seems there's at least one winter season (post 1972) with an unusually large snowfall total. One of the simplest methods to confirm this outlier is a Grubbs' test where the test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation. Significance is assessed using a t-distribution. Using Grubbs' test, the p-value for the updated dataset is $<$ 0.002 suggesting that the snowfall amount in 1976/77 is an outlier. During the winter of 1976/77 a notorious storm occurred that was coined the Blizzard of '77. This tremendous blizzard hit the Buffalo region in January of 1977 with snow drifts (not snowfall) from this storm as high as 100 in. in several areas. Figure 2 is a newspaper headline from January 31st, 1977 depicting the damage and destruction caused by the storm. While this blizzard gathered headlines it is not the source of this season as an outlier, as this storm consisted of about only 10 in. of new snowfall with most of the damage caused by strong winds blowing the previously fallen snow creating large snow drifts and accumulations in certain regions.

TV Topics ★ Pause ★ News Magazine ★ Color Comics

BUFFALO EVENING NEWS

Week-End Edition Storm Special

Blizzard Paralyzes WNY, Kills 7, Strands Thousands

Helicopters, Federal Aid Are Sought

Helping Hands Reach Out to the Stranded As Doors, Hearts Open to Victims of Storm

News special

In The Week-End Magazine

Also in today's News

TV Topics including Pause

Index to Today's News

Braving Inferno, Winter's Fury, Firemen Save a Neighborhood

Figure 2 Newspaper cover page during the Blizzard of '77. Permission received from The Buffalo News for reprinting.

With this updated dataset, in addition to seeing an outlier, we can re-examine the annual snowfall density estimation. In density estimation with *n* observations, one can start with the naive empirical estimators such as the empirical density function which puts a mass of $1/n$ at each data point. These empirical estimators constitute the so-called "raw" estimators of the distribution. Parzen, himself, in [1] proposes "that the problem in density estimation is to find a suitable 'smooth' density which is closest to a 'raw' estimator of the density." We show the density estimation for the original and updated datasets as estimated using the kernel density technique in Figure 3 (see sidebar for details of the kernel density method). Ultimately, from Figure 3, we conclude the annual snowfall density has one peak with several years (post 1972) having larger than expected snowfall amounts (skewed right tail).

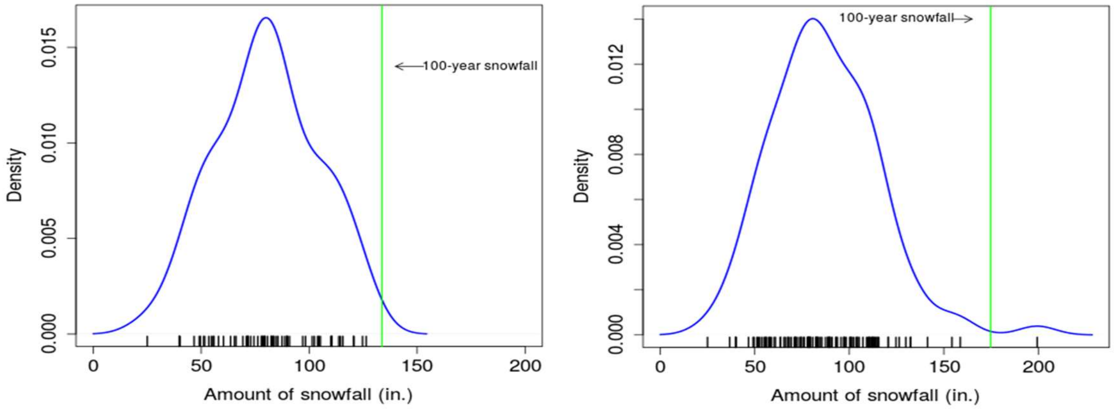

Figure 3 : Left: Kernel density estimate for the original snowfall dataset (1909/10-1971/72). Right: Kernel density estimate for the snowfall data measured in inches in Buffalo, NY from 1909/10 to 2014/15. The data points are shown as short vertical bars. The green line at 175 in. represents the 100-year snowfall (see sidebar for explanation).

To further extend the snowfall study we integrate other variables into the analysis such as the average annual winter temperature as we may expect temperature to be associated with snowfall amounts. We perform a “web-scrape” to find monthly average temperature data for Buffalo, NY. The National Weather Service webpage www.weather.gov/buf/BUFtemp provides average monthly temperatures from 1940 to present time. (Note there are other webpages for temperature data although they charge a fee for the data.) Since we are interested in the average winter temperature, we average the (average) monthly temperature from November through April for each winter season. The average winter temperature data in Fahrenheit are given in Table 2.

Table 2: Annual Winter Temperature (Nov-April) in Fahrenheit (F) from 1940/41-2014/15.

32.42	34.27	29.95	31.67	34.00	34.28	33.52	32.32	36.93
33.22	33.72	34.03	36.47	35.88	34.65	31.72	35.38	33.97
31.27	32.85	30.57	31.97	29.92	33.45	32.03	32.93	32.92
32.47	32.47	30.55	31.42	31.88	34.43	33.07	33.53	35.07
30.22	29.63	31.45	33.63	33.00	30.65	36.23	32.08	34.27
33.60	34.82	34.83	33.65	33.90	37.12	33.47	33.38	31.80
35.13	29.80	33.18	36.32	34.80	35.6	32.42	37.67	30.68
33.72	32.90	36.07	34.47	34.25	32.83	34.95	31.87	39.43
35.13	29.65	29.60						

A simple linear regression model for the data in Figure 4 (left) shows that, on average, for each degree Fahrenheit increase in winter temperature, the snowfall will decrease 7.2 inches (R-squared = 0.27, p<0.001). This analysis demonstrates a natural conclusion, namely that colder temperatures will be associated with more snowfall.

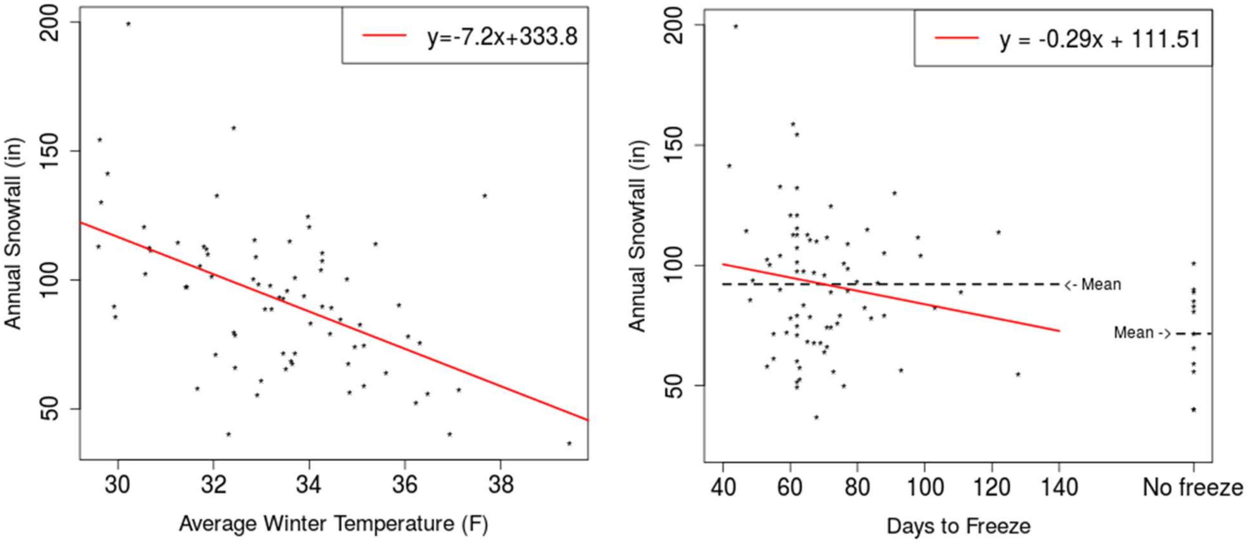

Figure 4 (Left) A plot of the annual winter temperature (November of one year through April of the next year) starting from 1940/41 to 2014/15. The linear regression suggests that for each increase in average winter temperature the snowfall decreases 7.2 inches. (Right) Snowfall as a function of the days required (since Nov 1st) for Lake Erie to freeze.

In addition to air temperature an additional variable to possibly explain the annual snowfall is Lake Erie water temperature. Note lake water temperatures are a factor in producing so called “lake-effect” snow where moisture from the lake is lifted by moving storms and deposited upon landfall as snow. Once the lake freezes over, traditional lake effect snow is no longer possible. Thus, we may expect that an unfrozen Lake Erie is associated with more snowfall. We compute the number of winter days from November 1st each season until the lake freezes, that is, obtains a temperature of 32 degrees. Daily Lake Erie temperature data are available at http://www.erh.noaa.gov/buf/laketemps/laketemps.php or by request to author. Granted November 1st is an arbitrary date, but this variable nevertheless allows us to examine the number of possible lake effect snow days in a winter season. Figure 4 (right) shows annual snowfall as a function of the number of days required for the lake to freeze in a winter season. Note there are 12 winter seasons where the lake did not freeze. We see a significant mean shift (Wilcoxon Rank Sum p=0.02) in the annual snowfall between seasons where the lake freezes and seasons where the lake does not freeze. This shift surprisingly suggests that snowfall amounts are smaller in years where the lake does not freeze. Further in seasons where the lake does freeze over, we see that the snowfall decreases 0.29 in. (p=0.17, R-squared=0.24) for each additional day required for the lake to freeze. Note the 0.29 coefficient is not significantly different in zero. Put in other words, there is no evidence that Lake Erie being unfrozen for a longer time frame is associated with more snowfall in Buffalo, NY. From these expanded analyses on snowfall and winter temperatures we discover that a colder ambient winter temperature is associated with more snowfall and that an increased presence of lake effect snow conditions do not significantly contribute to more snowfall.

In conclusion, Parzen’s Buffalo snowfall dataset has played a prominent role in demonstrating statistical methods. Here we update the dataset and re-examine several of the statistical techniques for inference including density estimation. Ultimately, our findings show that the annual snowfall density is skewed towards larger annual snowfall amounts and larger snowfalls are significantly associated with colder ambient winter temperatures. This dataset can also be viewed in a larger historical perspective (see sidebar) with the history of national weather forecasting originating in Buffalo, NY and will, no doubt, continue to be a testing ground for future generations of statisticians.

The National Weather Service and Buffalo, NY

We have just scratched the surface of studying lake and temperature variables and their possible associations with snowfall. As one of the Great Lakes, the weather effects related to Lake Erie have an important history dating back to the formation of the national weather service. Following Great Lakes storms of 1868 and 1869 which sank or damaged over a 1000 vessels killing over 500 sailors and passengers, the scientist, Increase A. Lapham, sent letters detailing these wrecks and other weather related events such as sea level changes to General Halbert Paine, Congressman for Milwaukee. The extensive research in these letters led Congressman Paine to introduce a congressional resolution that created a meteorological observing network to provide notice of approaching storms on the Great Lakes and other seacoasts. The resolution was passed by congress and signed into law one week later by President Ulysses S. Grant with Lapham appointed as the Assistant to the Chief Signal Officer of the newly created meteorological division of the Signal Service (later the U.S. Weather Bureau and then the National Weather Service). With this appointment, Lapham became known as the “Father of the National Weather Service.”

Interestingly these first meteorological observations were taken at military stations with those duties assigned to General Albert Myer and his military Signal Core. General Myer was born in Buffalo, NY in 1828 and had developed aerial signaling methods for field communications during the Civil War. After the Civil War, he acquired the nickname “Old Probabilities” since his weather forecasts appeared in bulletins and articles under *The Probabilities.* No doubt these early forecasts with primitive forecasting equipment were very limited. General Myer later hired Lapham as a civilian assistant to assist with the daily weather recording! This new office was coined the “U.S. Weather Bureau” and is now known National Weather Service, a component of the National Oceanic and Atmospheric Administration. Ultimately, this department was the source for the updated version of the snowfall data and lake temperature variables.

Figure 5 (Left) General Albert Myer nicknamed "Old Probabilities" due to his early periodicals on weather forecasting. (Right) Increase A. Lapham the "Father of the U.S. Weather Service." Both photos are in the public

References

[1] Parzen, E. 1977. “A Unified Approach Based on Density Estimation and Testing for ‘White Noise’” *Technical Report No. ARO-1.* Statistical Laboratory State University of New York at Buffalo.