[redacted]

MATH 216

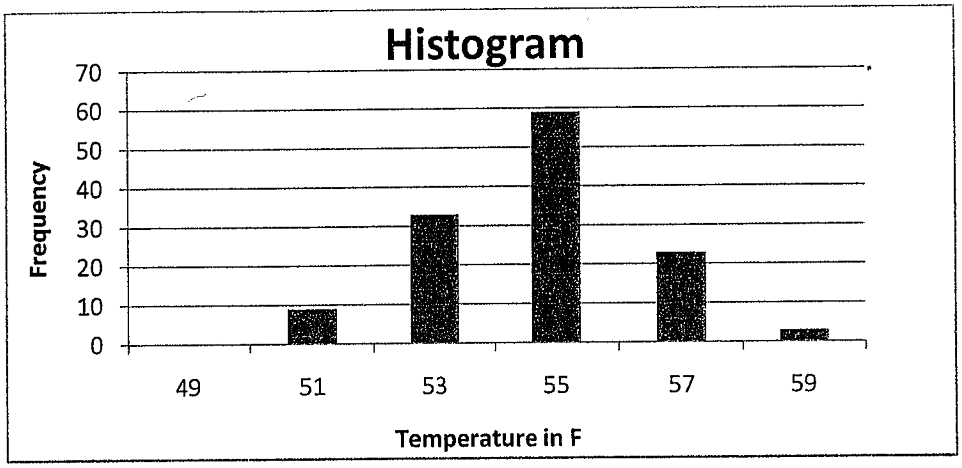
Statistical Analysis of the Mean Annual Temperature in New York City (Central Park)

Global warming is the theory that the earth is warming due to human activities – particularly due to emissions from fossil fuels. This topic has been hotly debated, particularly over the past decade. The ramifications of this theory – if true – are tremendous. If global warming is indeed occurring, and we dismiss it (Type II Error), then our children's children could very well pay the price of our mistake. However, if global warming is incorrectly validated (Type I Error), then we could all pay for years to come in pointless political banter and policy. The following analysis is preformed on data collected from a single location in New York City's Central Park[1]. While this analysis alone is not nearly enough to prove whether or not global warming is occurring, it certainly will give some perspective on the magnitude of temperature change of a location fairly close to home.

In order to determine if there is a basis for Global Warming present in this data is to to construct hypothesis tests to determine if the mean temperature of a number of samples has increased by a significant amount. However, before we can perform this analysis it is vital to make sure that the underlying data follows a normal distribution. To do this, I first averaged the data by year, and then I constructed a histogram. The histogram is shown below in figure 1.

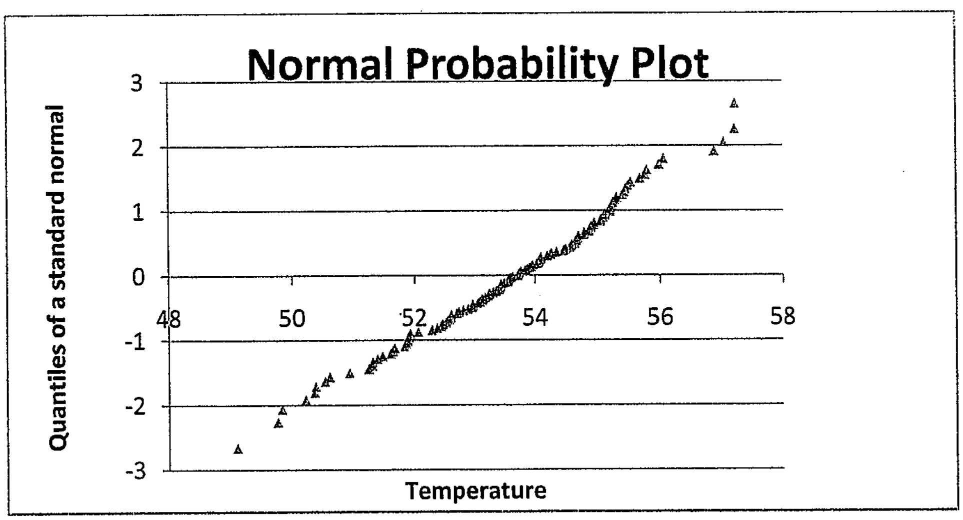
[1] Data set obtained from http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.USHCN/ID+305801+VALUE/.filnet/.mean/.temp/

Figure 1: Histogram



As the histogram indicates, the data is single peaked, relatively symmetrical, and has rapid dying tails. This supports our assumption that the underlying data is normal.

Another method of showing that the data is normal is by constructing a normal probability plot. The plot for this data set is shown below.



From the normal probability plot, we can see that the plot is very straight. The end of it is slightly skewed however. Yet, even with the skew, the plot passes the "Fat pencil" test. Since the sample size (n = 128) is substantially large, it is safe to assume that the data is normal.

After determining that the data is normal, the analysis can proceed. First, I calculated the overall mean and standard deviation of the yearly mean data. These values will be used later on in the hypothesis tests. Next, I arranged the data into mean temperatures by decade. These sets of data will be the samples used to conduct the hypothesis tests. The resulting chart is shown in figure 3.

Figure 3: Mean Temperatures by Decade

From	To	Mean Temperature
1870	1879	51.84933333
1880	1889	51.08999992
1890	1899	52.88249983
1900	1909	52.9741665
1910	1919	52.64499992
1920	1929	53.12583333
1930	1939	54.3375
1940	1949	54.32916667
1950	1959	54.78916667
1960	1969	54.16916667
1970	1979	54.53666667
1980	1989	55.11083333
1990	1996	55.43940476

In order to conduct our hypothesis test, we must determine the Z values for each of the samples. The Z-values can be obtained by using the following equation[2]:

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

The mean, μ, is the mean all of the yearly temperatures in the data set, and equals 53.59. The standard deviation, σ, is the standard deviation of the entire data set, and equals 1.63. As an example, the Z value calculation for the decade of 1870-1879 is shown below:

[2] All equations and Z values obtained from *Statistical Methods for Engineers*

$$Z = \frac{51.849 - 53.59}{1.63/\sqrt{10}} = -3.39$$

For this analysis, I decided to use a significance level of $\alpha = .05$. The Z-Value for this significance level is $Z_\alpha = 1.645$. For my hypothesis, I want to determine if the mean of the each is significantly greater than the mean of the overall data set, or if it is unchanged. The hypotheses are shown below.

$$H_0: \mu = 53.59$$

$$H_{A}: \mu > 53.59$$

For the sample from the years 1870 – 1879, H_0 must be accepted because there is not enough evidence to say otherwise. This is illustrated below:

$$Reject\ H_0\ if\ Z > Z_\alpha$$

$$-3.39 > 1.645?$$

$$False$$

The results from the hypothesis tests for each of the decades are shown below: (note that the last sample is not a full decade and n=7 was used for its calculation)

Figure 4: Results from Hypothesis Tests

Start	End	n	H_0	H_A	Z	α	Z_α	Rejection Region ($Z > Z_\alpha$)			Reject H_0?
1870	1879	10	μ = 53.59	μ > 53.59	-3.4	0.05	1.644854	-3.39	>	1.6449	N
1880	1889	10	μ = 53.59	μ > 53.59	-4.9	0.05	1.644854	-4.86	>	1.6449	N
1890	1899	10	μ = 53.59	μ > 53.59	-1.4	0.05	1.644854	-1.38	>	1.6449	N
1900	1909	10	μ = 53.59	μ > 53.59	-1.2	0.05	1.644854	-1.2	>	1.6449	N
1910	1919	10	μ = 53.59	μ > 53.59	-1.8	0.05	1.644854	-1.84	>	1.6449	N
1920	1929	10	μ = 53.59	μ > 53.59	-0.9	0.05	1.644854	-0.91	>	1.6449	N
1930	1939	10	μ = 53.59	μ > 53.59	1.44	0.05	1.644854	1.442	>	1.6449	N
1940	1949	10	μ = 53.59	μ > 53.59	1.43	0.05	1.644854	1.426	>	1.6449	N
1950	1959	10	μ = 53.59	μ > 53.59	2.32	0.05	1.644854	2.318	>	1.6449	Y

1960	1969	10	$\mu = 53.59$	$\mu > 53.59$	1.12	0.05	1.644854	1.116	>	1.6449	N
1970	1979	10	$\mu = 53.59$	$\mu > 53.59$	1.83	0.05	1.644854	1.829	>	1.6449	Y
1980	1989	10	$\mu = 53.59$	$\mu > 53.59$	2.94	0.05	1.644854	2.943	>	1.6449	Y
1990	1996	7	$\mu = 53.59$	$\mu > 53.59$	3	0.05	1.644854	2.995	>	1.6449	Y

From these results, we can see that four of the thirteen decades have mean temperatures high enough above the population mean to be considered mathematical anomalies at the .05 significance level.

Since we have rejected H_0 for a few of the samples, it would be informative to construct a 99% confidence interval for the true mean of these samples. The interval is determined by using the following equation:

$$\bar{y} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

An example calculation is shown below for the years 1950-1959:

$$54.79 \pm 2.57 * \frac{1.63}{\sqrt{10}} = (53.46,56.11)$$

Repeating these calculations for each of the decades that rejected H_0 yields the following table of confidence intervals:

Decade	**From**	**To**
1950-1959	53.46	56.11
1970-1979	53.21	55.85
1980-1989	53.79	56.43
1990-1996	53.84	57.01

The minimum's are well within one standard deviation of the mean. However, the maximum's for the years 1990-1996 is substantially above two standard deviations from the mean. Another interesting feature that can be gleaned from the hypothesis tests is that the decades which rejected H_0 are all relatively recent, and the latest three from the data set also failed. In fact, by looking at the confidence intervals for the failed decades, we can see that they progressively get higher and higher, with the highest being from 1990-1996. This gives credit to the theory of global warming. One would expect that since global warming is a result of carbon emissions, and since our nation has used more and more fossil fuels as time has progressed, the problem would become exacerbated.

In order to ascertain how good these hypothesis tests are at determining change in the mean annual temperature, a power analysis is needed. Power is the probability that we reject H_0 when H_A is correct. Power is calculated using the following equation:

$$power = P(Z > z_\alpha - \gamma), where\, \gamma = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$$

The following calculation shows the power of the tests to detect a change in the mean temperature by one degree Fahrenheit (54.59°).

$$power = P\left(Z > 1.645 - \frac{54.59 - 53.59}{1.63/\sqrt{10}}\right) = .6141$$

This means that the null hypothesis will be rejected 61.41% of the time when the true mean temperature is 54.59°. Naturally, as the desired temperature increases, so does the power of these tests.

From this analysis, it can be concluded that the mean temperature has definitely risen since the 1950's. The data shows that the temperature was fairly close to the mean, even below for some decades, up until this point. However, from 1950 on, the data showed significant increases in the mean temperature. This is reflected by the results of the hypothesis tests. However, it is crucial to remember that correlation does not necessarily mean causality. Just because the temperature has risen does not necessarily mean that it is the result of human-caused global warming. It is entirely possible that another phenomenon could explain this rise in temperature. It is not this papers intent to either prove or disprove global warming. However, this data used in conjunction with other similar studies (along with someone who knows significantly more about the climate than I do), would surely aid in this debate.

Works Cited

None. *IRI/LDEO Climate Data Library.* September 25, 2006. http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.USHCN/ID+305801+VALUE/.filnet/.mean/.temp/ (accessed April 23, 2007).

Vining, Geoffrey, and Scott M. Kowalski. *Statistical Methods for Engineers - Second Edition.* Belmont, CA: Thomson Higher Education, 2006.