# ORF 245 Fundamentals of Statistics
# Chapter 10
# Summarizing Data

Robert Vanderbei

Fall 2015

Slides last edited on December 14, 2015

# Median and Mode

Let $X$ be a random variable with cdf $F(x)$ and pdf $f(x)$. Recall that the *mean*, $\mu$, of $X$ is defined as the expected value of X:

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Another "measure of location" is the *median*, $\eta$. It is that place on the real line where the probability that $X$ comes out bigger than $\eta$ (or smaller than $\eta$) is exactly $1/2$. In terms of the cdf, we can write that

$$F(\eta) = 1/2 \qquad \text{or equivalently} \qquad \eta = F^{-1}(1/2)$$

Let $X_1, X_2, \ldots, X_n$ be a finite collection of independent identically distributed random variables. In Chapter 7, we derived a formula for the confidence interval associated with estimating the mean $\mu$.
Our aim a few slides hence is to derive a confidence interval for the median $\eta$.

A third measure of location is the *mode*. It is the value of $x$ that has the greatest "likelihood" as measured by the pdf:

$$\text{mode} = \text{argmax}_x f(x).$$

# Log-Normal Distribution

The best example of random variables for which the mean and median differ are those with the log-normal distribution. A random variable $Y$ is said to have a *log-normal* distribution if the log of $Y$ has a normal distribution. So, suppose that

$$Y = e^{\mu+\sigma X}$$

where $X$ is a standard normal random variable.

Let's compute the mean value of $Y$:

$$\begin{aligned}\mathbb{E}(Y) = \mathbb{E}e^{\mu+\sigma X} &= \int_{-\infty}^{\infty} e^{\mu+\sigma x}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = e^{\mu}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}+\sigma x}dx \\ &= e^{\mu}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\sigma)^2}{2}+\frac{\sigma^2}{2}}dx = e^{\mu+\frac{\sigma^2}{2}}\end{aligned}$$

We compute the median as follows:

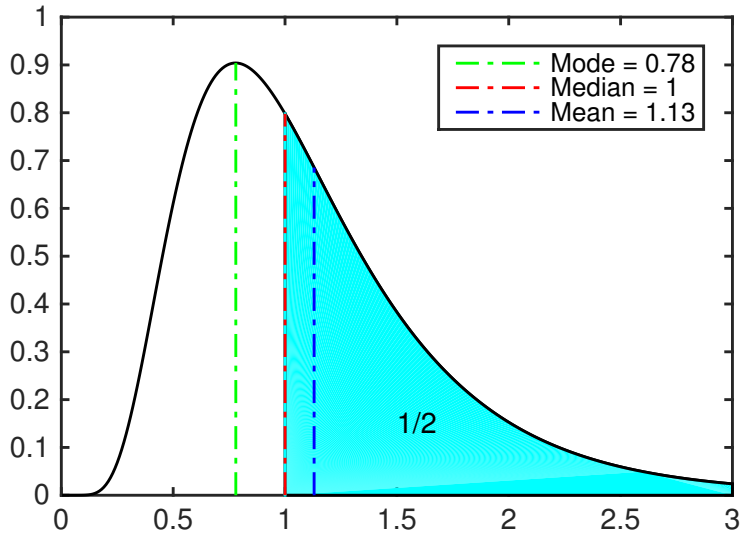
$$\frac{1}{2} = P(Y \le y) = P(e^{\mu+\sigma X} \le y) = P(\mu+\sigma X \le \log(y)) = P(\sigma X \le \log(y)-\mu)$$

By the symmetry of the standard normal distribution, we see that $\log(y)-\mu = 0$ which translates to

$$\text{Median}(Y) = e^{\mu}$$

The mode is also "easy" to compute: $\text{Mode}(Y) = e^{\mu-\sigma^2}$

# Mode vs. Median vs. Mean




```
x = 0.01:0.01:3;
y = pdf('logn',x,0,1/2);
plot(x,y,'k-');
hold on;
fill([x(100:end) x(end) x(100)], [y(100:end) 0 0], 'c');
l1=plot([0.78 0.78], [0 y(78)], 'g-.');
l2=plot([1.00 1.00], [0 y(100)], 'r-.');
l3=plot([1.13 1.13], [0 y(113)], 'b-.');
legend([l1,l2,l3], 'Mode = 0.78', 'Median = 1', 'Mean = 1.13');
```

# Order Statistics — Sample Median

As before, let $X_1, X_2, \ldots, X_n$ be a finite collection of independent identically distributed random variables.

Rearrange these $n$ random numbers into increasing order:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

In other words, $X_{(1)}$ is the smallest of the $n$ random numbers and $X_{(n)}$ is the largest of the set. Written like this in increasing order, these derived random variables are called the *order statistics*.

If $n$ is odd, the middle point, $X_{((n+1)/2)}$, is called the *sample median*.

The order statistics allow us to make a confidence interval for the median. To this end, for any fixed integer $k$, we start by computing

$$P(X_{(j)} \leq \eta \leq X_{(j+1)}) = P(\text{exactly } j \text{ of the } X_i\text{'s are less than } \eta) = \binom{n}{j} \frac{1}{2^n}$$

the second equality follows from the fact that each $X_j$ is less than $\eta$ with probability $1/2$ and these are independent events. Hence, the number of them that are less than $\eta$ is a binomial random variable with parameter $1/2$.

# Confidence Interval for the Median

From

$$P(X_{(j)} \leq \eta \leq X_{(j+1)}) = \binom{n}{j} \frac{1}{2^n}$$

it follows that

$$P(X_{(k)} \leq \eta \leq X_{(n-k+1)}) = \sum_{j=k}^{n-k} \binom{n}{j} \frac{1}{2^n}$$

If we pick $k$ such that the right-hand side is close to $0.95$, then we get that

$$P(X_{(k)} \leq \eta \leq X_{(n-k+1)}) \approx 0.95$$

In other words, for such a choice of $k$,

$$X_{(k)} \leq \eta \leq X_{(n-k+1)}$$

is a $95\%$ confidence interval for the median $\eta$.

# Finding $k$

For $n$ small, $k$ can be found by explicitly computing

$$\sum_{j=0}^{k-1} \binom{n}{j} \frac{1}{2^n}$$

for small values of $k$ and stopping when this sum is approximately $0.025$.

For large values of $n$, we can apply to the Central Limit Theorem.

Specifically, let $Y$ be a binomial random variable with mean $p = 1/2$ and $n$ equal to the number of $X_i$'s. We know that

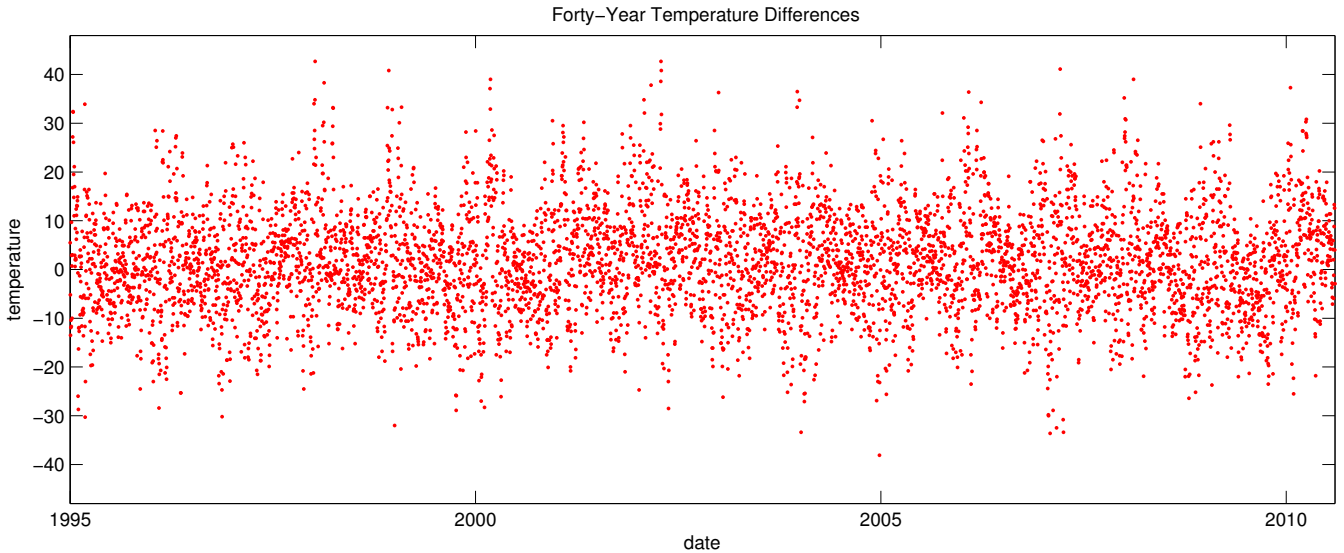
$$\mathbb{E}(Y) = np = \frac{n}{2} \qquad \text{and} \qquad \text{Var}(Y) = npq = \frac{n}{4}$$

If $n$ is large (say, greater than $20$), $Y$ is approximately normally distributed and therefore $95\%$ of the probability falls with two standard deviations of the mean. Hence,

$$k = \frac{n}{2} - \sqrt{n}$$

rounded off to the nearest integer.

# Local Climate Data – Forty Year Differences




$$n = 5699$$

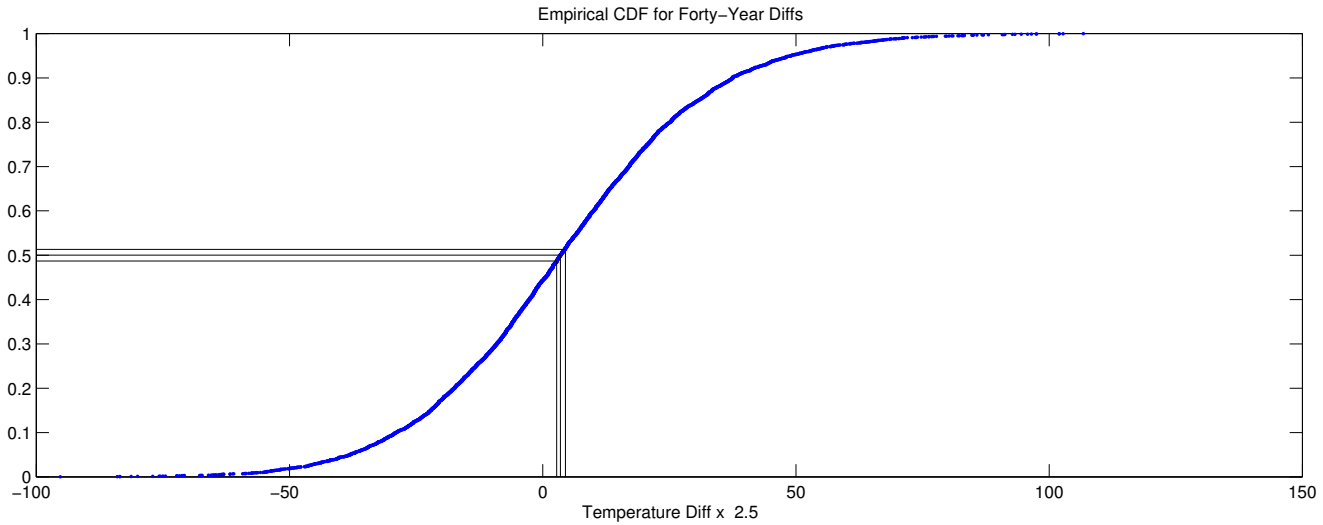
On a degrees-Fahrenheit-per-century basis...

$$\bar{X} = 4.25, \qquad S = 26.5, \qquad S/\sqrt{n} = 0.35$$

Confidence interval...

$$\mu = 4.25 \pm 0.69$$

# Local Climate Data – Forty Year Differences




$$n = 5699 \qquad \sqrt{n} \approx 75$$

$$k = 2775 \qquad (n+1)/2 = 2850 \qquad n - k + 1 = 2925$$

On a degrees-Fahrenheit-per-century basis...

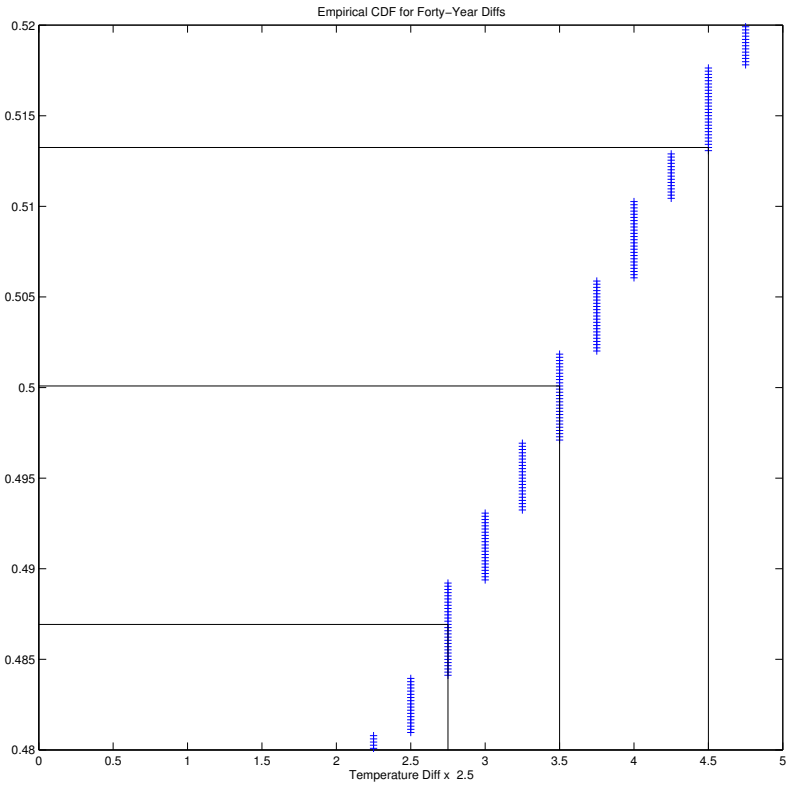
$$X_{(2775)} = 2.75, \qquad X_{(2850)} = 3.50, \qquad X_{(2925)} = 4.50$$

Confidence interval...

$$\eta \in [2.75, 4.50]$$

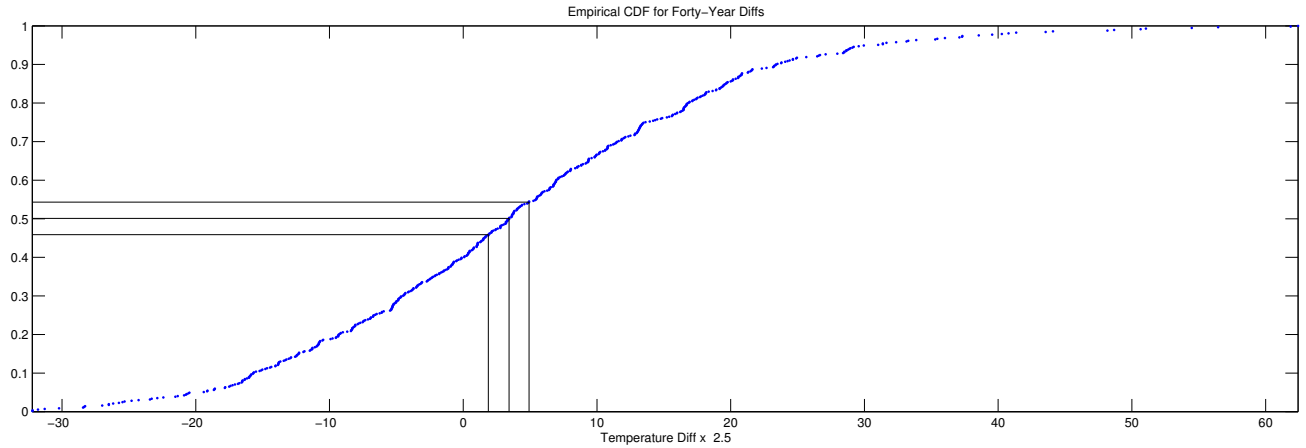
# Local Climate Data – Forty Year Differences

Close up view




NOTE: Precision in original data is only to one decimal place.

# Ten-Day Averages of Forty Year Differences




$$n = 569 \qquad \sqrt{n} \approx 24$$

$$k = 261 \qquad (n+1)/2 = 285 \qquad n - k + 1 = 309$$

On a degrees-Fahrenheit-per-century basis...

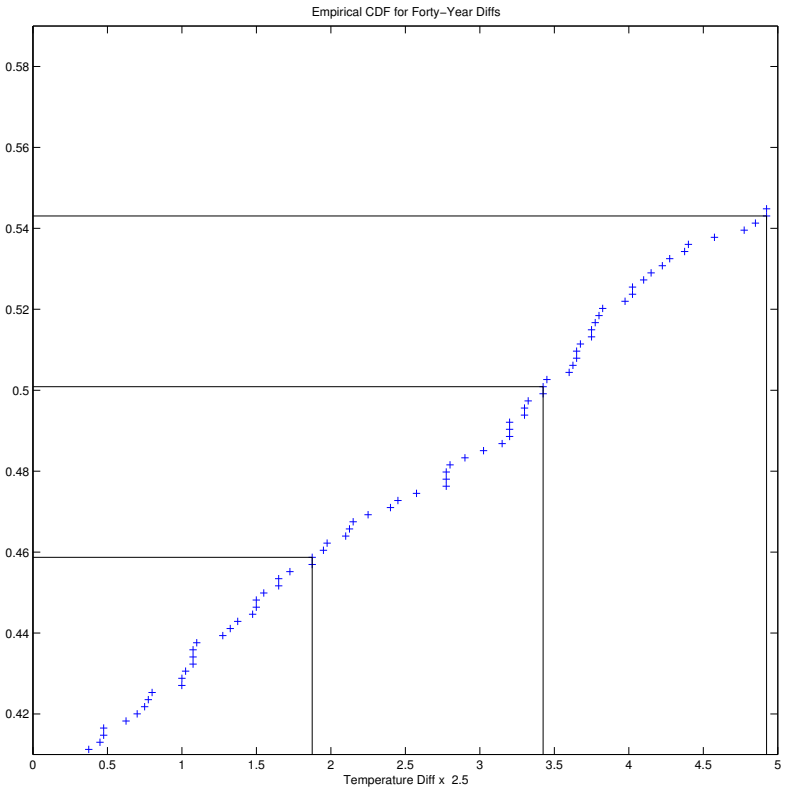
$$X_{(2775)} = 1.875, \qquad X_{(2850)} = 3.425, \qquad X_{(2925)} = 4.925$$

Confidence interval...

$$\eta \in [1.8755, 4.925]$$

# Ten-Day Averages of Forty Year Differences

Close up view




NOTE: Precision is improved but confidence interval has widened.

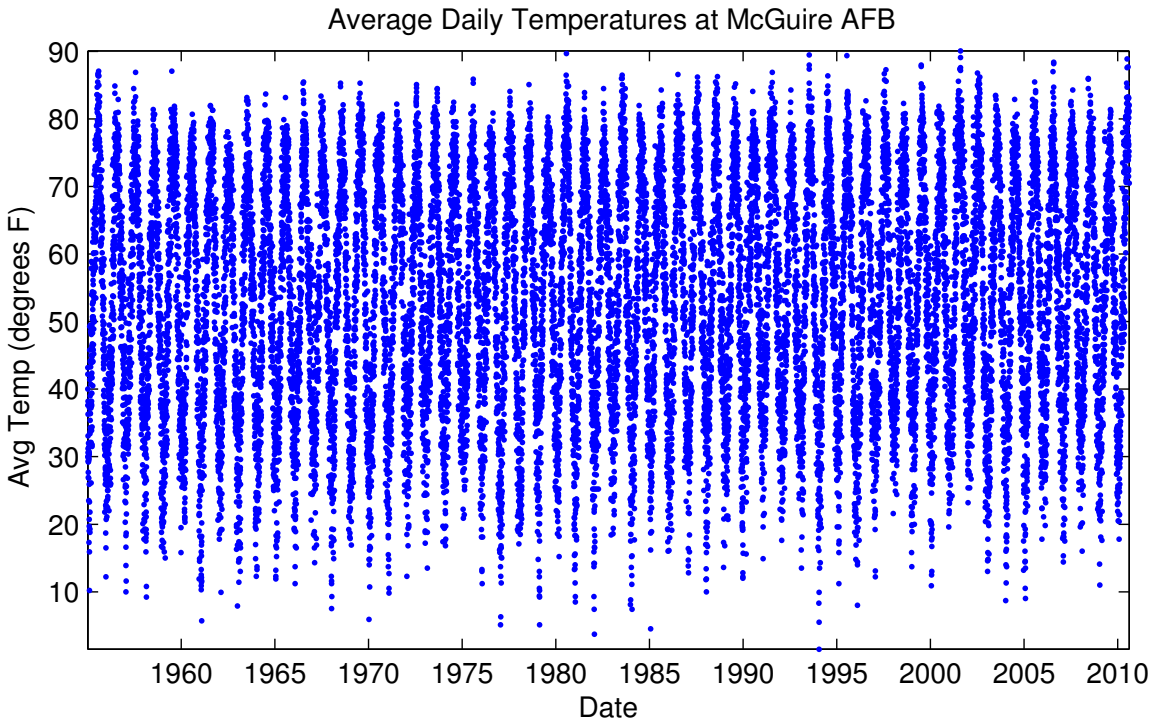
# Matlab Code

```
load -ascii '/Users/rvdb/ampl/nlmodels/LocalWarming/McGuireAFB/data/McGuireAFB.dat';
temp = McGuireAFB(:,2);

diffs = temp(1+40*365.25:end) - temp(1:end-40*365.25);
diffs = 2.5*diffs; % convert to 'per century'
[n m] = size(diffs);
xbar = mean(diffs)
stddev = std(diffs)
n
stddev/sqrt(n)
1.96*stddev/sqrt(n)
y = (1:n)/n;
tempdiffssorted = sort(diffs);
figure(6); % FortyYearDiffs2cdf.pdf
plot(tempdiffssorted,y,'b.');
title('Empirical CDF for Forty-Year Diffs');
xlabel('Temperature Diff x  2.5');
hold on;
n2 = round(n/2);
k = n2 - round(sqrt(n));
plot([-100 tempdiffssorted(n2) tempdiffssorted(n2)], [n2/n n2/n 0], 'k-')
plot([-100 tempdiffssorted(k) tempdiffssorted(k)], [k/n k/n 0], 'k-')
plot([-100 tempdiffssorted(n-k+1) tempdiffssorted(n-k+1)], [(n-k+1)/n (n-k+1)/n 0], 'k-')
hold off;
tempdiffssorted([k n2 n-k+1])
[k n2 n-k+1]
[k n2 n-k+1]/n
round(sqrt(n))
```

# Graphical Analyses of Data

Recall the original 55-years of daily average temperature data from McGuire AFB:

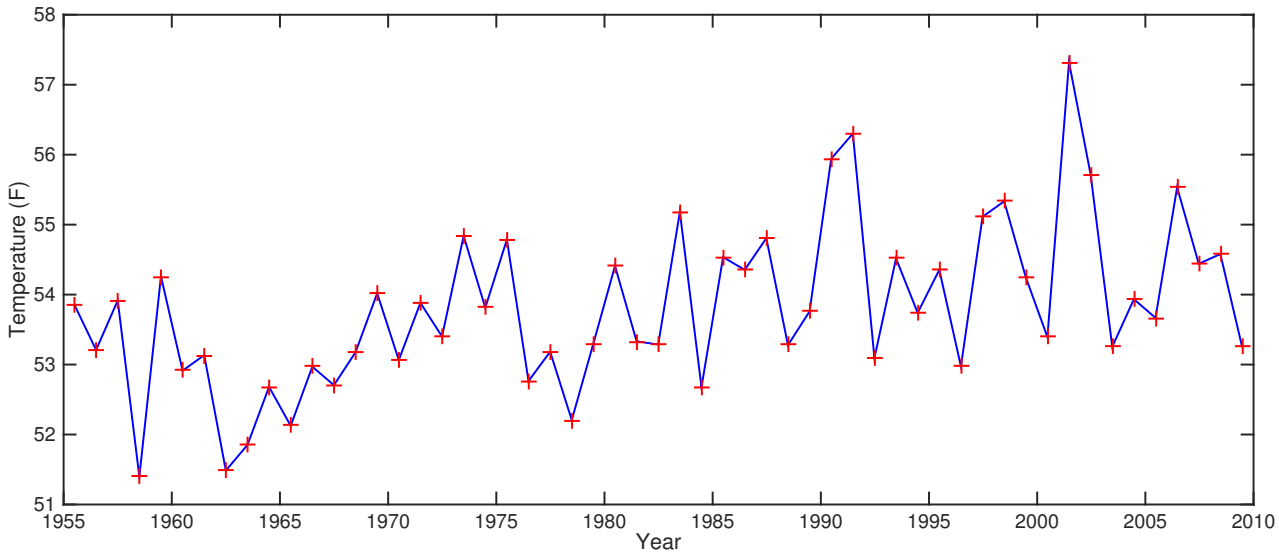



We have shown that 40-year differences in temperatures are, on average, significantly different from zero.

Yet, no trend is evident in this display of the raw data.

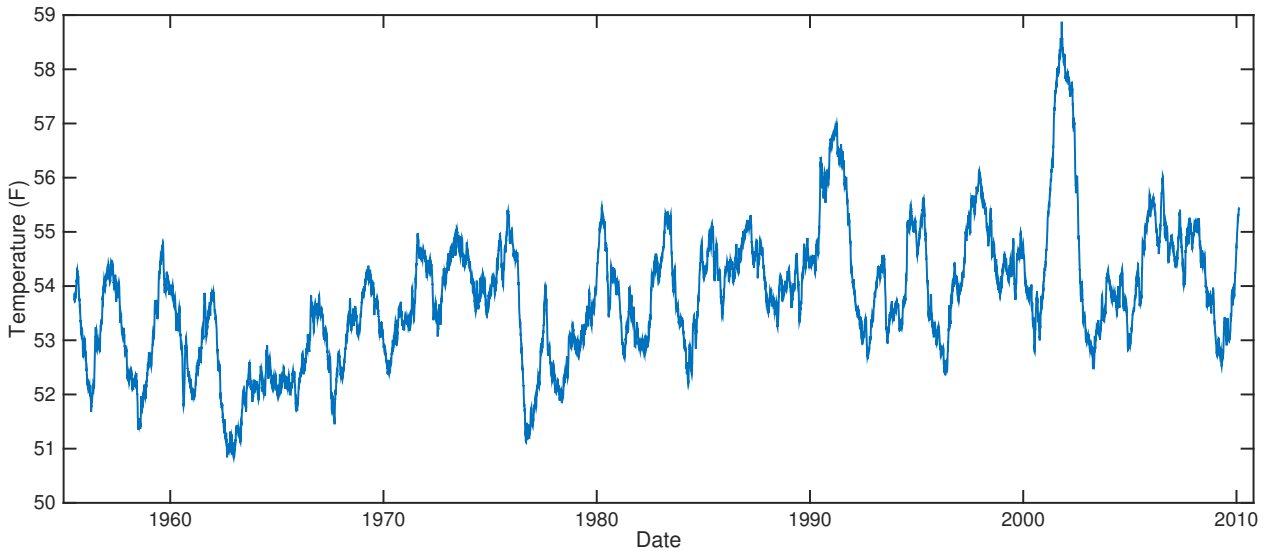
Can we find better ways to summarize the data?

# One-Year Averages




```
T = McGuireAFB(:,2);
figure(6);
window = ones(365,1);
Tw = conv(T,window,'valid');
time = 1955+1/2 + (1:size(Tw))/365.25;
yearlytime = time(1:365:end)';
yearlytemp = Tw(1:365:end)/365;
plot(yearlytime,yearlytemp,'b');
```

# One-Year Averages Day-by-Day




```
T = McGuireAFB(:,2);
figure(8);
window = ones(365,1);
Tw = conv(T,window,'valid');
time = 1955+1/2 + (1:size(Tw))/365.25;
plot(time(1:end),Tw(1:end)/365);
```