ORF 245 Fundamentals of Statistics
Chapter 7
Survey Sampling

Robert Vanderbei

Fall 2014

Slides last edited on December 31, 2014

Chapter Overview

This chapter deals with a special case in which the sample space Ω consists of a large, but finite, number of elements: N.

As usual, we consider n random variables, $X_1, X_2, \ldots, X_n$, defined on Ω.

The only difference with what has been discussed before is that the random variables represent *sampling without replacement* and are, therefore, not independent random variables.

Sampling with replacement produces iid random variables.

Key point: In modern applications, N is generally *very large*.

Also, n is big but generally *much smaller* than N.

In such cases, the differences between sampling with replacement vs without replacement are tiny.

We will ignore them.

Hence, we will skip everything except Section 7.3.3.

Election Polling

(Midterm election is tomorrow!)

Suppose that $p = 48\%$ of voters support candidate A. A polling agency queries $n = 1600$ randomly selected voters. What is the probability that the poll will show that candidate A has the support of a majority of the voters?

Let X_i be a Bernoulli random variable that is 1 if the i-th randomly selected voter supports candidate A and 0 otherwise.

Let

$$S_n = \sum_{i=1}^{n} X_i$$

denote the number of polled voters who support candidate A.

The random variable S_n is Binomial with parameters n and p. Its mean and standard deviation are

$$\mu = \mathbb{E}(S_n) = np = 1600 \times 0.48 = 768$$

and

$$\sigma^2 = \mathsf{Var}(S_n) = npq = 1600 \times 0.48 \times 0.52 = 399 \quad \Longrightarrow \quad \sigma = 20$$

Hence,

$$\begin{aligned} \mathbb{P}(S_n \geq 801) = \mathbb{P}(S_n \geq 800.5) &= \mathbb{P}((S_n - \mu)/\sigma \geq 32.5/20) \\ &\approx \mathbb{P}(Z \geq 1.625) = \mathbb{P}(Z \leq -1.625) = 0.05208 \end{aligned}$$

where Z is a std. normal rv. (Note: cdf('bino', 800.5, 1600, 768/1600) $= 0.94801$)

Election Polling (p unknown)

Suppose that p is not known. A poll of $n = 1600$ voters is taken and finds that $s_n = 831$ say they are in favor of candidate A. We can use s_n as an estimate of the random variable S_n. Let $\hat{P} = S_n/n$. We can estimate p using the sample estimate of $\hat{P}$:

$$\hat{p} = s_n/n = 831/1600$$

Similarly, we can estimate $\sigma^2 = \text{Var}(S_n)$ using $n\hat{p}\hat{q}$ where $\hat{q} = 1 - \hat{p}$. Hence, the variance of $\hat{P}$ is estimated using

$$\hat{p}\hat{q}/n$$

We can compute a probability that another poll will also show candidate A ahead:

$$\begin{aligned}
\mathbb{P}(\hat{P} \geq 0.5) &= \mathbb{P}\left(\frac{\hat{P} - \hat{p}}{\sqrt{\hat{p}\hat{q}/n}} \geq \frac{0.5 - \hat{p}}{\sqrt{\hat{p}\hat{q}/n}}\right) \approx \mathbb{P}\left(Z \geq \frac{0.5 - 831/1600}{\sqrt{\frac{831}{1600}\frac{769}{1600}/1600}}\right) \\
&= \mathbb{P}(Z \geq -1.55) = \mathbb{P}(Z \leq 1.55) \\
&= 0.9394
\end{aligned}$$

Election Polling – Confidence Interval

Using the fact that $\dfrac{\hat{P}-p}{\sqrt{pq/n}}$ is approximately a standard normal rv, we see that

$$\mathbb{P}\left(\left|\frac{\hat{P}-p}{\sqrt{pq/n}}\right| \leq 1.96\right) \approx 0.95$$

or equivalently

$$\mathbb{P}\left(\hat{P}-1.96\sqrt{pq/n} \leq p \leq \hat{P}+1.96\sqrt{pq/n}\right) \approx 0.95$$

Hence, there is a 95% chance that the *random interval*

$$\left[\hat{P}-1.96\sqrt{pq/n},\ \hat{P}+1.96\sqrt{pq/n}\right]$$

covers p. Don't forget that $\hat{P}$ is a random variable.

Unfortunately, we don't know p and q. So, we approximate them by $p \approx \hat{P}$ and $q \approx 1-\hat{P}$.

Confidence Interval – Continued

From the actual poll numbers, we have that

$$n = 1600, \qquad \hat{P} = \frac{831}{1600}, \qquad 1 - \hat{P} = \frac{769}{1600}$$

and so we get a specific interval for p:

$$[\, 0.4949 \;\leq\; p \;\leq\; 0.5439 \,]$$

The number p is fixed. We don't know what it is, but it is not random. Hence, the above statement is either right or wrong. If we repeat the experiment (i.e., the poll) over and over again we will get similar intervals but the endpoints will be a little different every time. The statement that p lies in the interval will be right about 95% of the time.

Confidence Intervals – Picking n

The "half-width" of the confidence interval is $1.96\sqrt{pq/n} \approx 0.0245$.

Suppose we want to preselect the polling sample size n to be big enough that we can guarantee that the confidence interval's half-width is at most 0.01:

$$1.96\sqrt{pq/n} \;\leq\; 0.01$$

How big must n be? Clearly n must satisfy

$$n \;\geq\; pq\,(1.96/0.01)^2$$

Unfortunately, we don't know p before taking the poll. To be safe, we should use the worst case. In other words, we need to find that value of p that maximizes $pq = p(1-p)$. It is easy to see that this quadratic function of p is maximized when $p = 1/2$. With this choice, we get

$$n \;\geq\; \frac{1}{4}\,(1.96/0.01)^2 \;=\; 9604$$

Conclusion: The poll must query about $10{,}000$ people in order to get a confidence interval whose half-width is $\pm 1\%$.

Election Polling – One-Sided Confidence Interval

Again using the fact that $\dfrac{\hat{P}-p}{\sqrt{pq/n}}$ is approximately a standard normal rv, we see that

$$\mathbb{P}\left(\frac{\hat{P}-p}{\sqrt{pq/n}} \leq 1.645\right) \approx 0.95$$

or equivalently

$$\mathbb{P}\left(\ \hat{P} - 1.645\sqrt{pq/n}\ \leq\ p\ \right)\ \approx\ 0.95$$

As before, we approximate $p \approx \hat{P}$ and $q \approx 1-\hat{P}$ on the left side of the inequality. From the actual poll numbers, we have that

$$n = 1600, \qquad \hat{P} = \frac{831}{1600}, \qquad 1-\hat{P} = \frac{769}{1600}$$

and so the interval

$$[\ 0.4949, 1\ ]$$

is a *one-sided* 95% *confidence interval* for p.

A "Better" Confidence Interval

Recall that we started our confidence interval derivation with this formula:

$$\mathbb{P}\left(\left|\frac{\hat{P}-p}{\sqrt{pq/n}}\right| \leq 1.96\right) \approx 0.95$$

We then replaced p and q in the denominator with estimates $\hat{P}$ and $1-\hat{P}$. But, let's not do this replacement. Instead, lets solve the inequality as given. We start by squaring both sides:

$$\frac{\left(\hat{P}-p\right)^2}{pq/n} \leq z^2$$

where z is a shorthand for 1.96. Multiplying by both sides by pq/n and expanding out the square, we get

$$\hat{P}^2 - 2\hat{P}p + p^2 \leq z^2 pq/n$$

Recalling that $q = 1-p$, we can view this as a quadratic inequality in p that defines an interval whose end points are the solutions to the quadratic equation:

$$\left(1+\frac{z^2}{n}\right)p^2 - \left(2\hat{P}+\frac{z^2}{n}\right)p + \hat{P}^2 = 0.$$

A "Better" Confidence Interval — Continued

From the quadratic formula, we get an explicit formula for the endpoints:

$$p = \frac{\left(2\hat{P} + \frac{z^2}{n}\right) \pm \sqrt{\left(2\hat{P} + \frac{z^2}{n}\right)^2 - 4\left(1 + \frac{z^2}{n}\right)\hat{P}^2}}{2\left(1 + \frac{z^2}{n}\right)}$$

$$= \frac{\hat{P} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{P}\hat{Q}}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Hence, we arrive at the following confidence interval:

$$\mathbb{P}\left(\frac{\hat{P} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{P}\hat{Q}}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \leq p \leq \frac{\hat{P} + \frac{z^2}{2n} + z\sqrt{\frac{\hat{P}\hat{Q}}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}\right) \approx 0.95$$

Confidence Intervals

Suppose that $X_1, X_2, \ldots, X_n$ are iid random variables with

- *unknown distribution*
- and *unknown mean*, μ,
- but (strangely) *known variance*, σ^2.

As usual, we denote the sample mean by $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

By the central limit theorem, for n "large", $\bar{X}$ is approximately normally distributed with mean μ and variance σ^2/n.
Hence, by the properties derived earlier for normally distributed random variables,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately normally distributed with mean zero and variance one, and so

$$\mathbb{P}\left(-z \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z) - \Phi(-z) = 1 - 2\Phi(-z)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for $N(0,1)$.

Confidence Intervals – Continued

For $z = 1.96$, we have $\Phi(-z) = 0.025$ and therefore

$$\mathbb{P}\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

The inequalities inside $\mathbb{P}(\cdot)$ can be rearranged to read

$$\mathbb{P}\left(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}\right) = 0.95$$

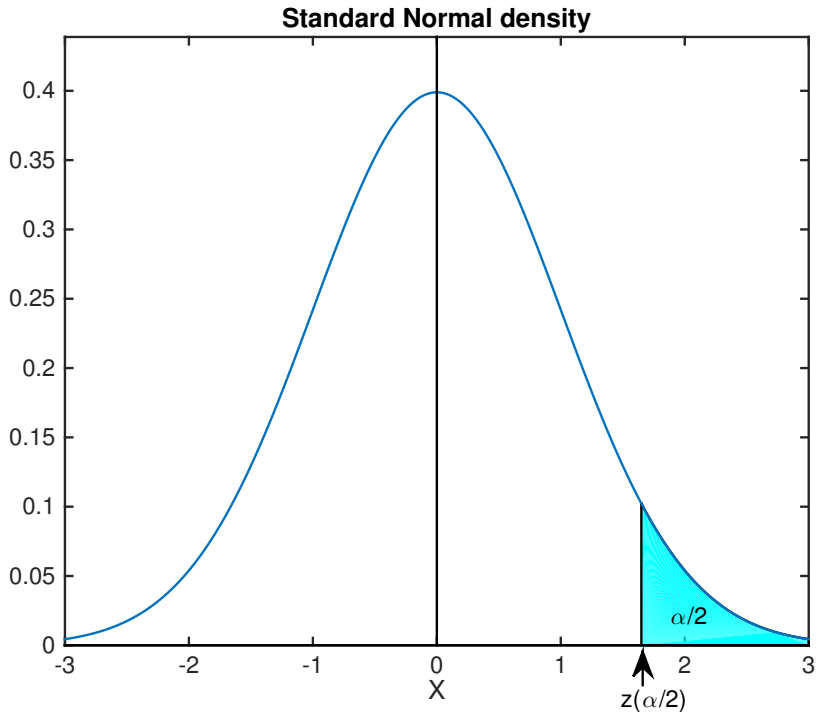
In words, we say that there is a 95% chance that the true mean lies within 1.96 standard deviations of the sample mean.

Since 1.96 is close to 2, it is common practice to report the $\bar{X} \pm 2\sigma/\sqrt{n}$ interval as the 95% confidence interval.

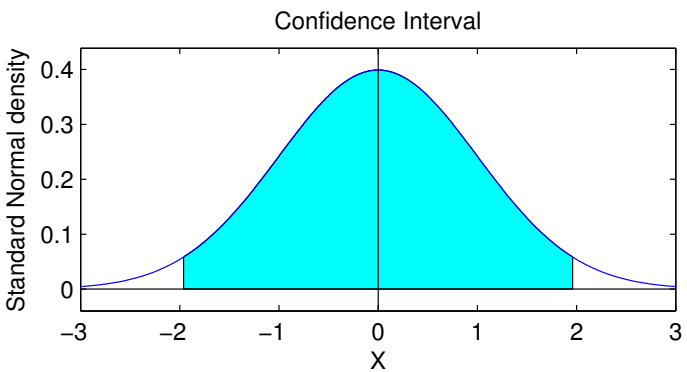
Confidence Intervals – Continued

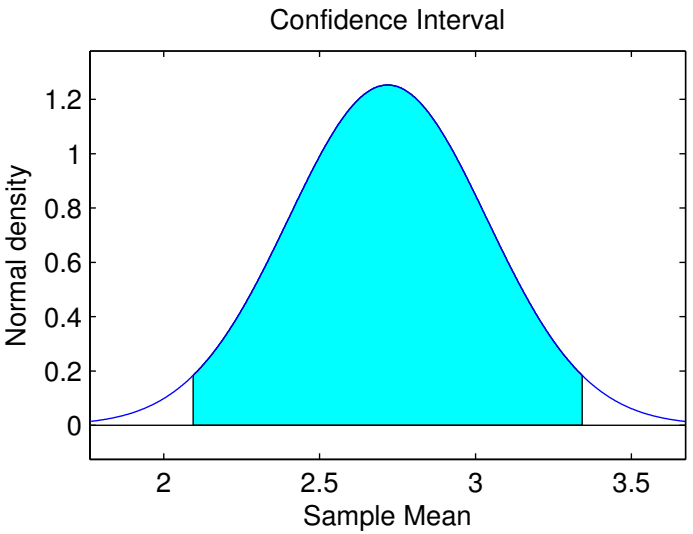
More generally

$$\mathbb{P}\left(\bar{X} - \ z(\alpha/2)\,\sigma/\sqrt{n} \ \leq \ \mu \ \leq \ \bar{X} + \ z(\alpha/2)\,\sigma/\sqrt{n}\right) = 1 - \alpha$$



Confidence Intervals – Continued





Estimating π

40 independent measurements of "circumference over diameter":

3.0599,	3.4868,	3.0575,	3.2583,	3.3841,
2.7767,	2.9833,	3.1937,	3.3270,	3.1986,
2.8534,	3.2194,	3.4213,	2.8607,	2.8056,
3.0890,	3.5169,	3.3082,	3.4082,	3.3041,
3.2140,	3.1693,	3.1045,	3.0943,	2.7239,
2.8365,	3.3146,	3.3173,	3.0096,	3.0476,
3.0557,	3.2060,	2.9506,	2.8569,	3.0850,
3.0136,	3.2344,	3.0893,	3.4888,	3.1426

$$\bar{x} = 3.1367, \qquad \sigma \approx 0.188.$$

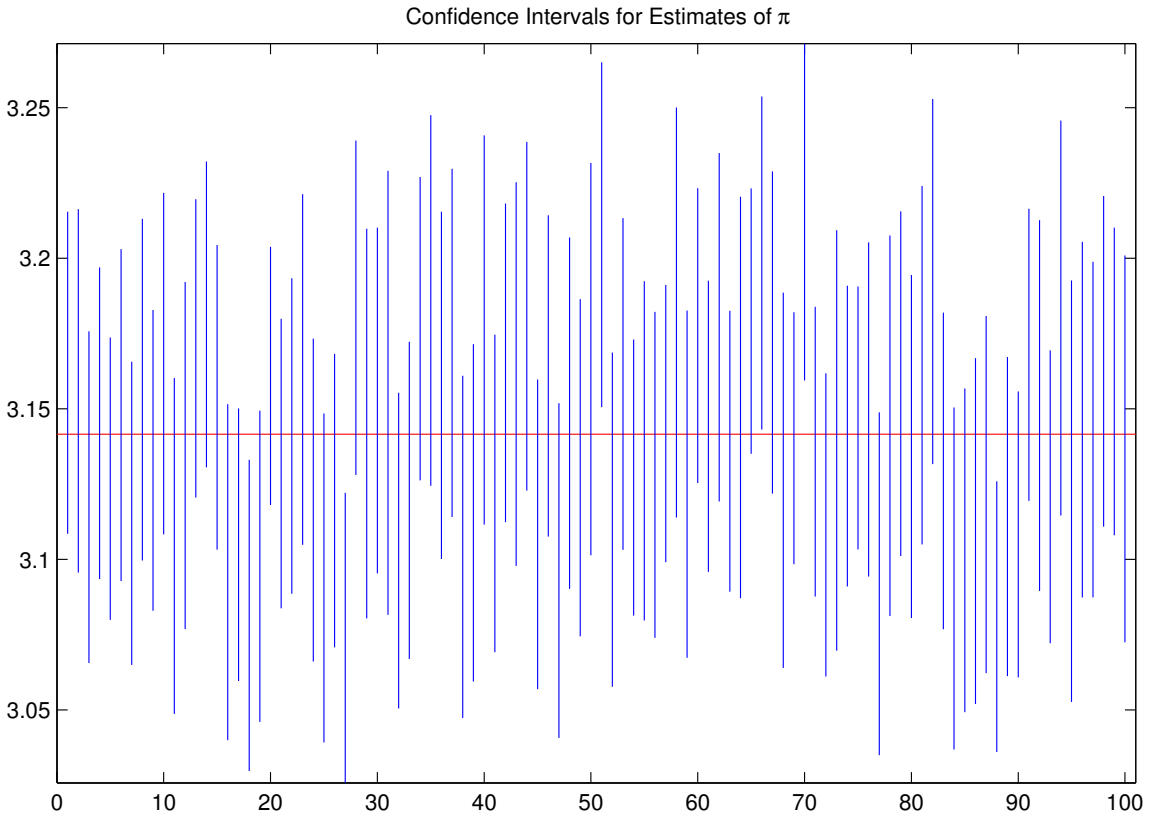
95% Confidence interval:

$$3.1367 - 1.96 \times 0.188/\sqrt{40} \;\leq\; \pi \;\leq\; 3.1367 + 1.96 \times 0.188/\sqrt{40}$$

In other words:

$$3.0793 \;\leq\; \pi \;\leq\; 3.1941$$

Estimating π – One Hundred Repetitions



Using all 4000 measurements, we get that $\pi = 3.1405 \pm 0.0057$

Confidence Intervals – Unknown Variance

Usually the variance, σ^2, is *not known*.
In such cases, we approximate the variance by the sample variance

$$\sigma^2 \approx S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

The random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a *t-distribution* with parameter n.
Hence, if we pick z so that $F(z) = \mathbb{P}(T \leq z) = 0.975$, then

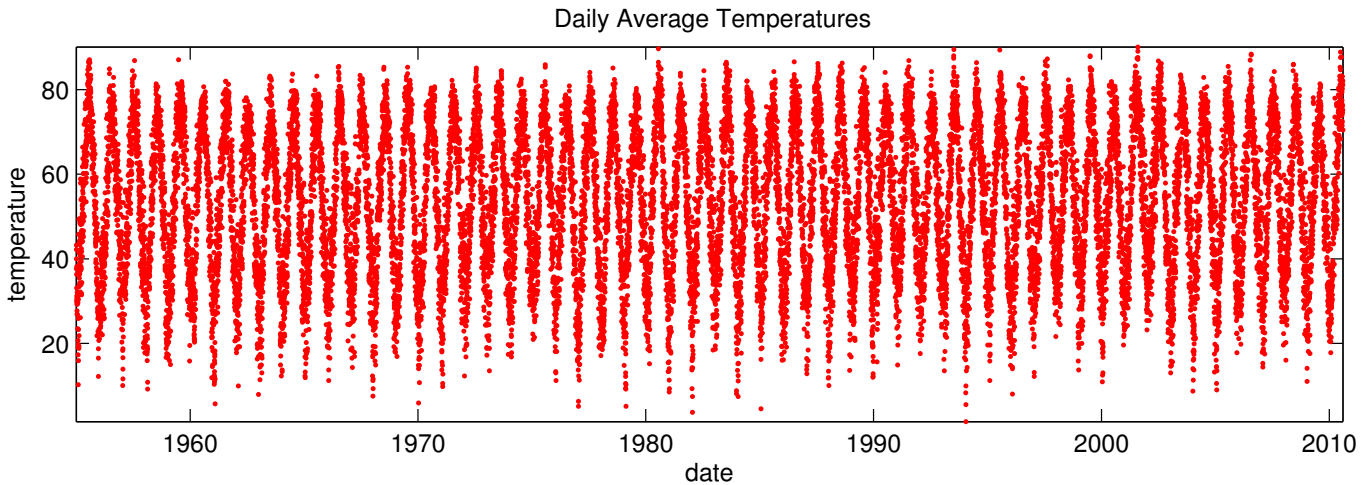
$$\mathbb{P}\left(-z \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z\right) = F(z) - F(-z) = 0.95$$

And, therefore a 95% confidence interval for μ can be written as

$$\mathbb{P}\left(\bar{X} - zS/\sqrt{n} \leq \mu \leq \bar{X} + zS/\sqrt{n}\right) = 0.95$$

Of course, the constant z depends on n. Values can be found in Table 4 of the textbook or using Matlab's `tinv` function. For large values of n, $z \approx 1.96$.

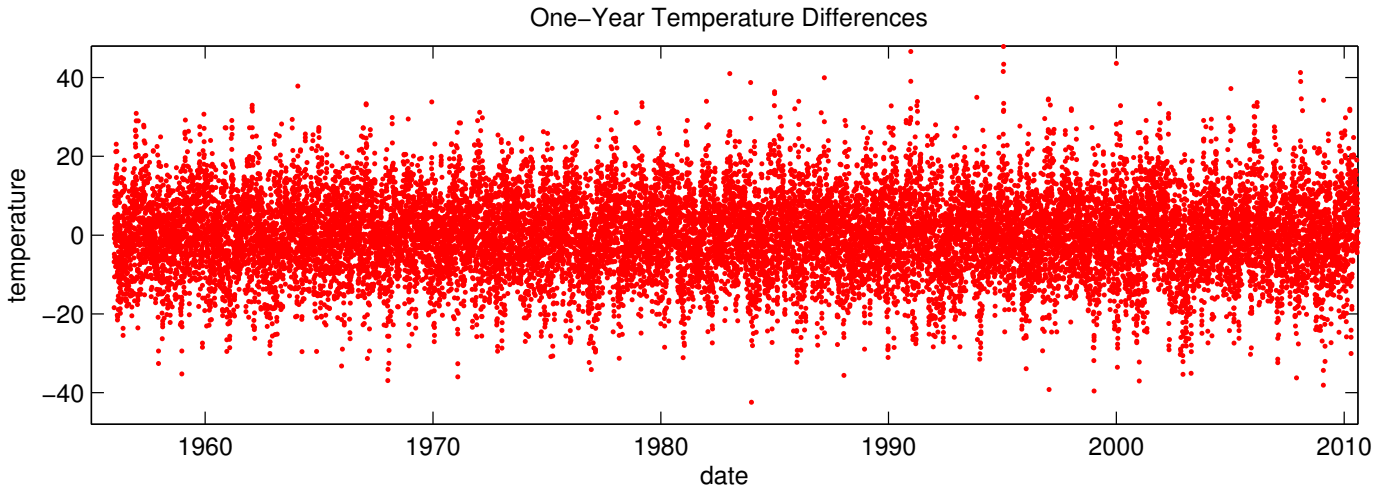
Local Climate Data



The data can be grabbed from here:

http://www.princeton.edu/∼rvdb/tmp/McGuireAFB.dat

Local Climate Data – One Year Differences



$$n = 19943$$

$$\bar{X} = 0.0298°\text{F/yr}, \qquad S = 10.1°\text{F/yr}, \qquad S/\sqrt{n} = 0.0716°\text{F/yr}$$

On a per century basis...

$$\bar{X} = 2.98°\text{F/century}, \qquad S = 1010°\text{F/century}, \qquad S/\sqrt{n} = 7.16°\text{F/century}$$

Confidence interval...

$$\mu = 2.98 \pm 14.03°\text{F/century}$$

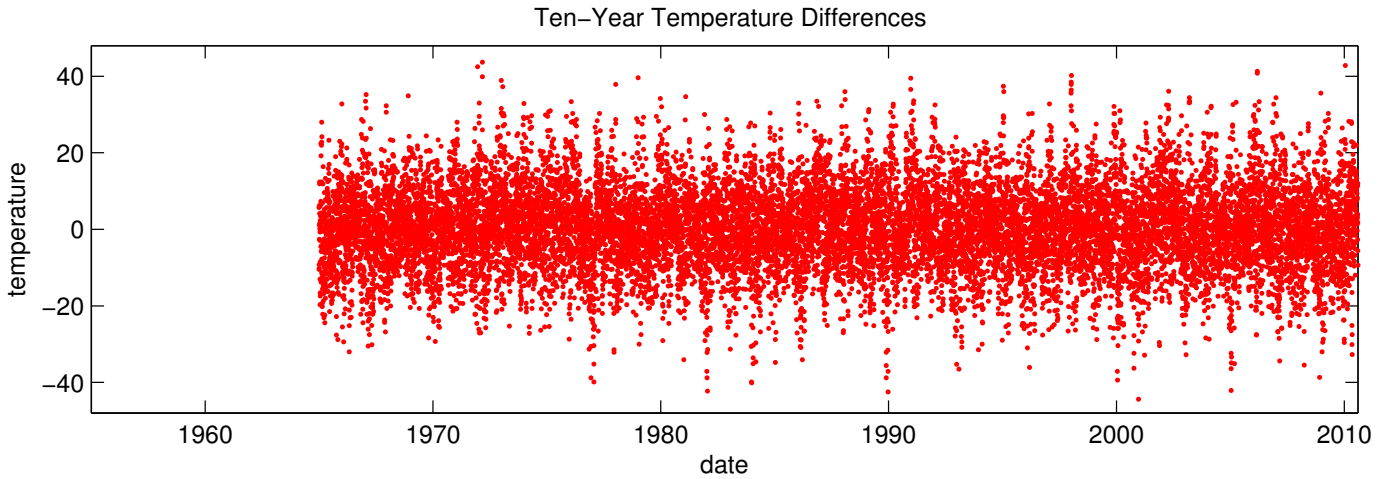
Not *convincing!*

The Matlab Code

```
load -ascii 'McGuireAFB.dat';
[n m] = size(McGuireAFB);
date = McGuireAFB(:,1);
date2 = 1955 + (0:n-1)/365.25;
temp = McGuireAFB(:,2);
plot(date2,temp,'r.');  % DailyAvgTemp.pdf
axis tight;
xlim([date2(1) date2(end)]);
xlabel('date');
ylabel('temperature');
title('Daily Average Temperatures');

'one year diffs'
diffs = temp(1+366:end) - 0.75*temp(2:end-365) - 0.25*temp(1:end-366);
diffs = 100*diffs;
[n m] = size(diffs);
xbar = mean(diffs)
stddev = std(diffs)
stddev/sqrt(n)
1.96*stddev/sqrt(n)
figure(2);
plot(date2(1+366:end),diffs/100,'r.');  % OneYearDiffs.pdf
axis tight;
xlim([date2(1) date2(end)]);
ylim([-48 48]);
xlabel('date');
ylabel('temperature');
title('One-Year Temperature Differences');
```

Local Climate Data – Ten Year Differences



$$n = 16657$$

$$\bar{X} = 0.397°\text{F/10 yrs}, \qquad S = 10.6°\text{F/10 yrs}, \qquad S/\sqrt{n} = 0.082°\text{F/10 yrs}$$

On a per century basis...

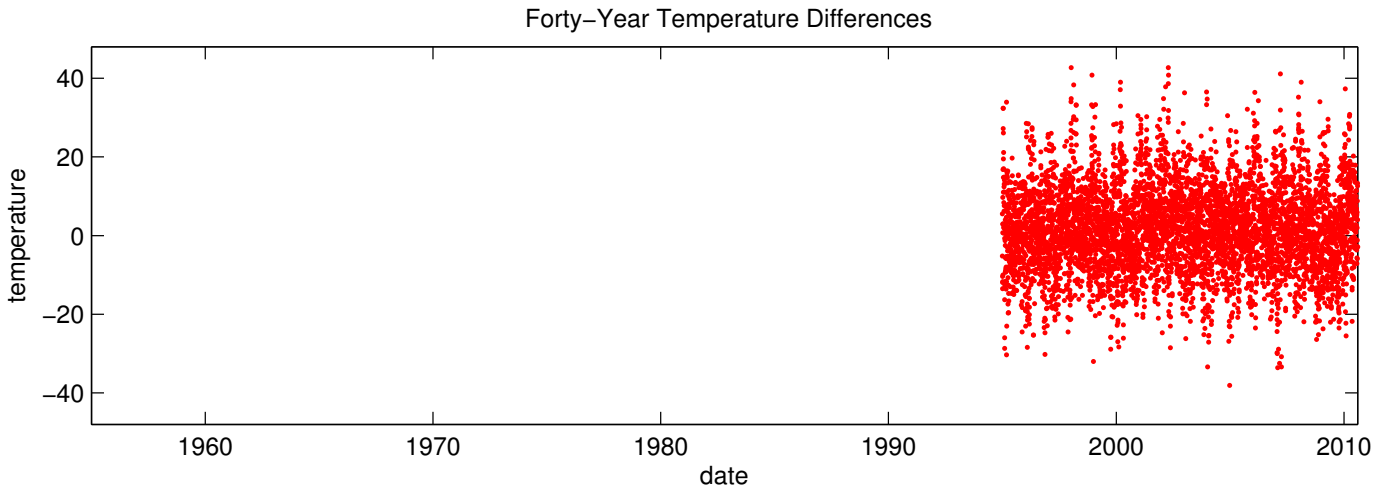
$$\bar{X} = 3.97°\text{F/century}, \qquad S = 105.9°\text{F/century}, \qquad S/\sqrt{n} = 0.82°\text{F/century}$$

Confidence interval...

$$\mu = 3.97 \pm 1.61°\text{F/century}$$

Okay, now I'm convinced.

Local Climate Data – Forty Year Differences



$$n = 5699$$

$$\bar{X} = 1.70°\text{F/40 yrs}, \qquad S = 10.6°\text{F/40 yrs}, \qquad S/\sqrt{n} = 0.140°\text{F/40 yrs}$$

On a per century basis...

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

Confidence interval...

$$\mu = 4.25 \pm 0.69°\text{F/century}$$

Now I'm even *more convinced!*