STA305/1004

Jan. 11, 2016

INTRODUCTION

- Course syllabus
- Course schedule
- Pre-requisites: STA302 or ECO375
- Statistical computing: R
- Discussion forums.

COURSE WEBSITES

- UofT Portal: https://portal.utoronto.ca/
- Piazza discussion forum: http://piazza.com/utoronto.ca/winter2016/sta305h
- Class notes: http://utstat.toronto.edu/~nathan/designscistudynotes.htm

WHY DESIGN?

Why should scientific studies be designed?

WHY DESIGN?

Why should scientific studies be designed?

- Avoid bias
- Variance reduction
- System optimization

BIG DATA

"… big data may be as important to business - and society - as the Internet has become. Why? More data lead to more accurate analyses."

(SAS, http://www.sas.com/en_id/insights/big-data/what-is-big-data.html)

BIG DATA

In 2015 the population of Canada is 35.8 Million people.

To estimate the mean number of hours spent on the Internet is it better to:

(a) take a simple random sample of 100 people (and ask about hours spent on internet) and estimate the mean number of hours spent on the Internet; or

(b) use a large database (e.g., millions of people) that contain hours spent on the Internet for each person?

BIG DATA

variance-bias trade-off

by precision, we mean 'mean square error' = variance + bias

- To have equivalent precision of a random sample of 100 people a database would have to contain over 96% of the population 34.3 Million people.
- This illustrates the power of random sampling and the danger of putting faith in "Big Data" simply because it's big.

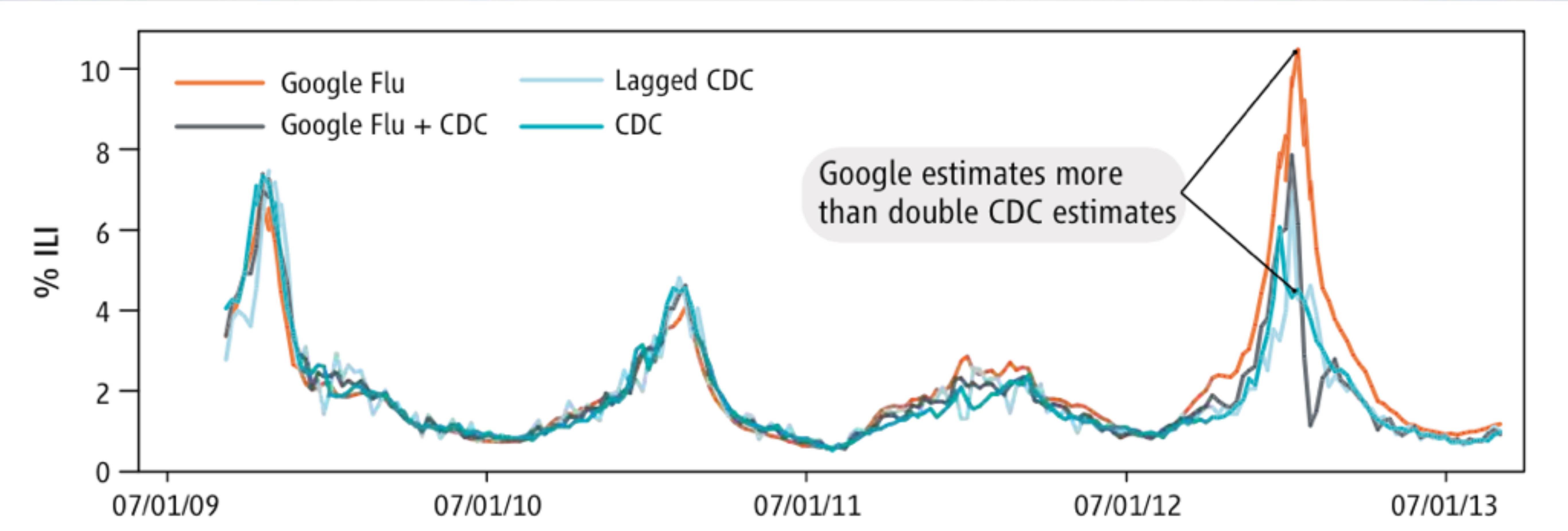
INTRODUCTION

- Data is usually very expensive.
- In a clinical trial the average per patient cost is between \$5500-\$7600.
- Statistics can help unfold what’s going on in the lab or production facility.

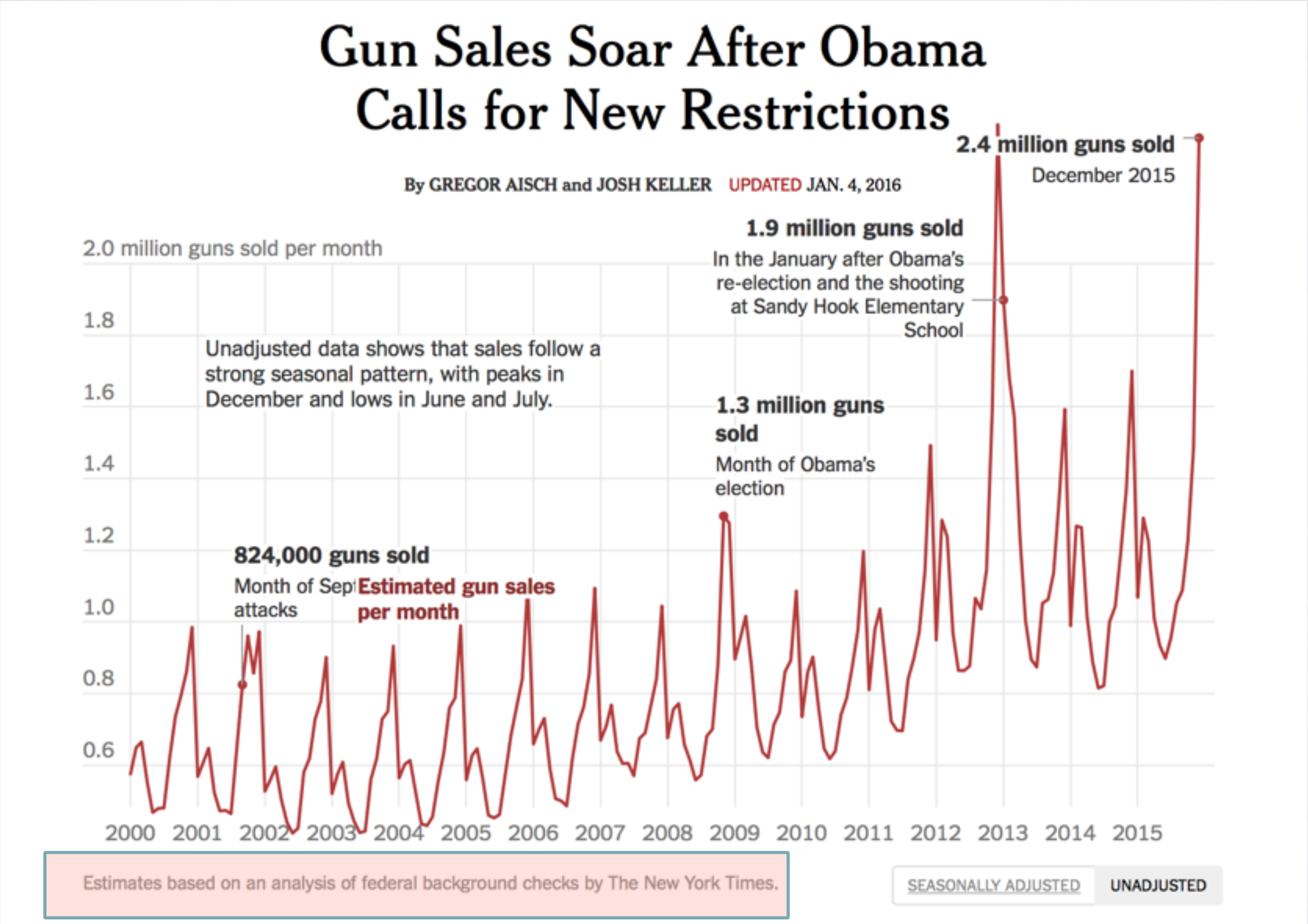
INTRODUCTION

Most "big data" is not obtained from instruments designed to produce valid and reliable data amenable for scientific analysis.

Google Flu (Lazer et al., Science 14 March 2014)



U.S. GUN SALES



Gun Sales Soar After Obama Calls for New Restrictions
By GREGOR AISCH and JOSH KELLER UPDATED JAN. 4, 2016
2.4 million guns sold
December 2015
1.9 million guns sold
In the January after Obama's re-election and the shooting at Sandy Hook Elementary School
2.0 million guns sold per month
1.8
Unadjusted data shows that sales follow a strong seasonal pattern, with peaks in December and lows in June and July.
1.6
1.3 million guns sold
Month of Obama's election
1.4
1.2
824,000 guns sold
Month of Sept
attacks
Estimated gun sales per month
1.0
0.8
0.6
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Estimates based on an analysis of federal background checks by The New York Times.
SEASONALLY ADJUSTED
UNADJUSTED

NICS Firearm Background Checks:
Month/Year

November 30, 1998 - December 31, 2015

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Totals
1998											21,196	871,644	892,840
1999	591,355	696,323	753,083	646,712	576,272	569,493	589,476	703,394	808,627	945,701	1,004,333	1,253,354	9,138,123
2000	639,972	707,070	736,543	617,689	538,648	550,561	542,520	682,501	782,087	845,886	898,598	1,000,962	8,543,037
2001	640,528	675,156	729,532	594,723	543,501	540,491	539,498	707,288	864,038	1,029,691	983,186	1,062,559	8,910,191
2002	665,803	694,668	714,665	627,745	569,247	518,351	535,594	693,139	724,123	849,281	887,647	974,059	8,454,322
2003	653,751	708,281	736,864	622,832	567,436	529,334	533,289	683,517	738,371	856,863	842,932	1,008,118	8,481,588
2004	695,000	723,654	738,298	642,589	542,456	546,847	561,773	666,598	740,260	865,741	890,754	1,073,701	8,687,671
2005	685,811	743,070	768,290	658,954	557,058	555,560	561,358	687,012	791,353	852,478	927,419	1,164,582	8,952,945
2006	775,518	820,679	845,219	700,373	626,270	616,097	631,156	833,070	919,487	970,030	1,045,194	1,253,840	10,036,933
2007	894,608	914,954	975,806	840,271	803,051	792,943	757,884	917,358	944,889	1,025,123	1,079,923	1,230,525	11,177,335
2008	942,556	1,021,130	1,040,863	940,961	886,183	819,891	891,224	956,872	973,003	1,183,279	1,529,635	1,523,426	12,709,023
2009	1,213,885	1,259,078	1,345,096	1,225,980	1,023,102	968,145	966,162	1,074,757	1,093,230	1,233,982	1,223,252	1,407,155	14,033,824
2010	1,119,229	1,243,211	1,300,100	1,233,761	1,016,876	1,005,876	1,069,792	1,089,374	1,145,798	1,368,184	1,296,223	1,521,192	14,409,616
2011	1,323,336	1,473,513	1,449,724	1,351,255	1,230,953	1,168,322	1,157,041	1,310,041	1,253,752	1,340,273	1,534,414	1,862,327	16,454,951
2012	1,377,301	1,749,903	1,727,881	1,427,343	1,316,226	1,302,660	1,300,704	1,526,206	1,459,363	1,614,032	2,006,919	2,783,765	19,592,303
2013	2,495,440	2,309,393	2,209,407	1,714,433	1,435,917	1,281,351	1,283,912	1,419,088	1,401,562	1,687,599	1,813,643	2,041,528	21,093,273
2014	1,660,355	2,086,863	2,488,842	1,742,946	1,485,259	1,382,975	1,402,228	1,546,497	1,456,032	1,603,469	1,803,397	2,309,684	20,968,547
2015	1,772,794	1,859,584	2,012,488	1,711,340	1,580,980	1,529,057	1,600,832	1,745,410	1,795,102	1,976,759	2,243,030	3,314,594	23,141,970

TOTAL 225,678,492

NOTE: **These statistics represent the number of firearm background checks initiated through the NICS. They do not represent the number of firearms sold. Based on varying state laws and purchase scenarios, a one-to-one correlation cannot be made between a firearm background check and a firearm sale.**

Getting gun sales estimates from background checks

To convert background checks into estimated sales, we relied on a method suggested in the Small Arms Survey by Jurgen Brauer, a professor at Georgia Regents University. Each long gun and handgun check was counted as 1.1 sales. Each multiple-gun check was counted as two sales. Permit checks and other types of checks were omitted. The multiplier is an estimate based on Mr. Brauer's interviews with gun shop owners.

US gun control

Gun sales in the US: how many are actually subject to background checks?

Mona Chalabi
Tuesday 5 January 2016 21.49 GMT

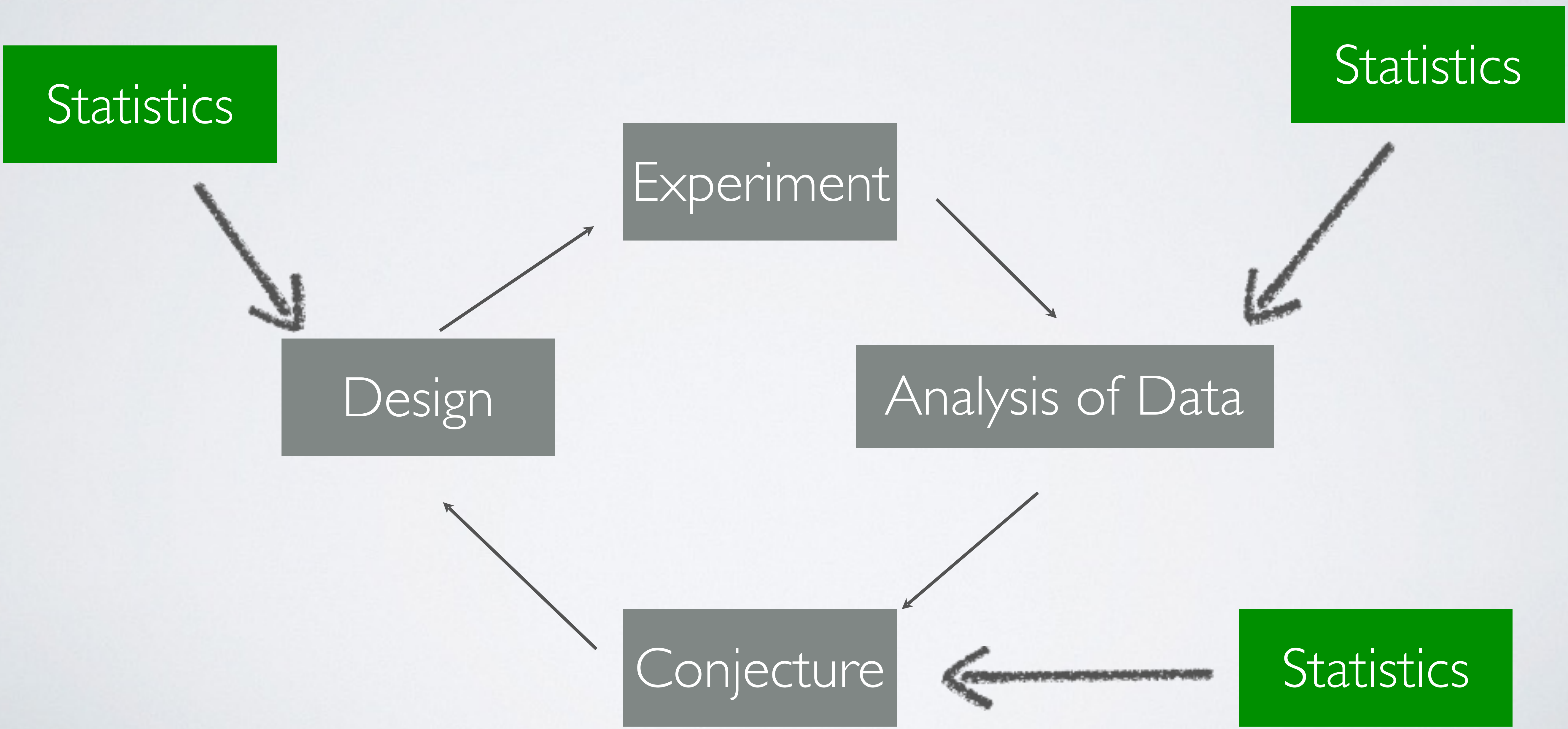
Right now, 40% of firearms are obtained in the US *without* a background check

That number surfaced 22 years ago, when Duke University and the University of Chicago conducted a telephone survey with 251 US adults in 1994. They were trying to find out what percentage of the country's then 44 million gun owners had received background checks to procure their 192 million firearms.

The data collection method has an impact on the quality of conclusions drawn from the data.

INTRODUCTION

Connected to Scientific Method



INTRODUCTION

- When you repeat an experiment you won't get the same response on two different occasions.
- Observation = true response + error
- The observations we get by repeating an experiment differ.

INTRODUCTION

- Good experimental design helps protect real effects from being obscured by experimental error.
- Designed experiments can increase signal-to-noise ratio.
- Statistical analysis provides measures of precision of estimated quantities under study.

INTRODUCTION

- What is the optimal measurement strategy?
- Suppose that we want to measure mass of two apples A and B using an old-fashioned two-pan balance scale.
- Should the apples be weighed one at a time?

INTRODUCTION

- (Hotelling, 1944) Let σ^2 be the variance of individual weighings of two objects.

This apple has weight w_1

This apple has weight w_2

- Weigh two objects together in one pan to obtain the sum of the two weights.

w_1+w_2

- Weigh two objects in opposite pans to obtain the difference between the two weights.

w_1-w_2

INTRODUCTION

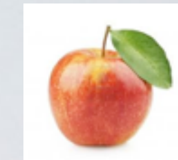
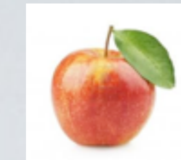
If the objects were weighed one at a time then how many weighings would be required to achieve the same precision (standard error) as the preceding design?

NEEDS 2 MEASUREMENTS

$w_1+w_2=m_1$
$w_1-w_2=m_2$

$\hat{w_1} = (m_1+m_2)/2,$
$\hat{w_2} = (m_1-m_2)/2,$

$Var(\hat{w_1}) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{4}2\sigma^2 = \frac{1}{2}\sigma^2$

INTRODUCTION  

Needs 2 measurements

If the objects were weighed one at a time then how many weighings would be required to achieve the same precision (standard error) as the preceding design?

$$w_1 + w_2 = m_1$$ both in 1 pan

$$w_1 - w_2 = m_2$$ sep pans

$$\hat{w}_1 = \frac{m_1 + m_2}{2}, \quad \hat{w}_2 = \frac{m_1 - m_2}{2}$$ Var " $\hat{w}_2$

$$\mathrm{Var}(\hat{w}_1) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{4} 2\sigma^2 = \sigma^2/2$$

Weighing #	Apple 1	Apple 2
1	a_1	
2	a_2	
3		b_1
4		b_2

Needs 4 Measurements

$$\frac{a_1+a_2}{2} = \text{ave. apple } \#1$$

$$\mathrm{Var}\left(\frac{a_1+a_2}{2}\right) = \frac{1}{4}\left(\sigma^2+\sigma^2\right) = \frac{\sigma^2}{2}$$

$$\mathrm{Var}\left(\frac{b_1+b_2}{2}\right) = \sigma^2/2$$

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \sigma^2$$

X_1 is indep. of X_2

$$\mathrm{Var}\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}\mathrm{Var}(X_1 + X_2)$$

Indep.

$$= \frac{1}{4}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\right)$$

$$= \frac{1}{4}\left(\sigma^2 + \sigma^2\right)$$

$$= \sigma^2 / 2 .$$

INTRODUCTION

- A major issue in experimentation is confusion of correlation with causation.

- Consider the scatterplot of population versus number of storks.

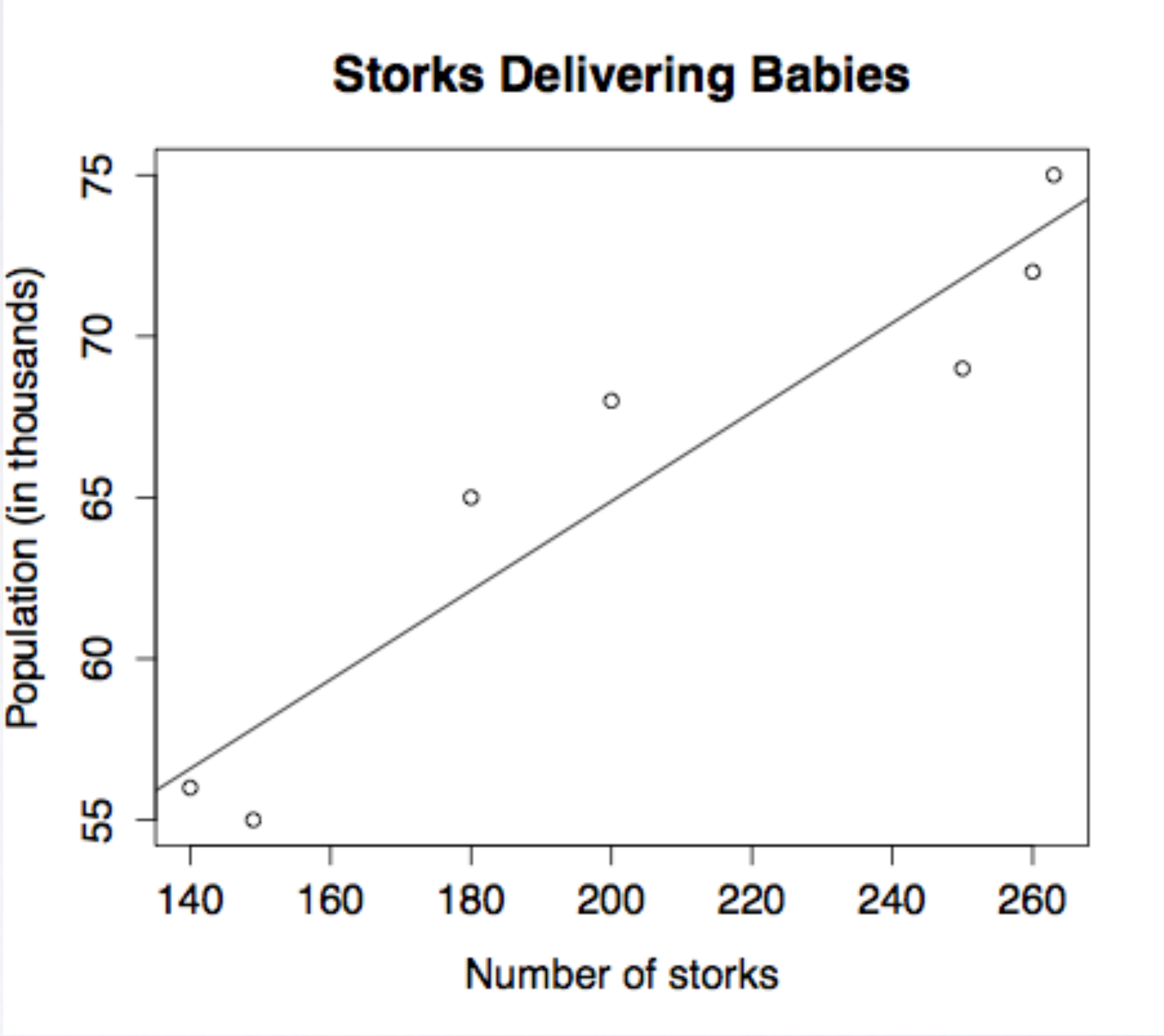
- $R^2 = 0.89$ and p-value of slope is 0.001.

$H_0: \beta_1 = 0$

$y = \beta_0 + \beta_0x + \epsilon$

- Does increase in number of storks *cause* an increase in population?

No, there is correlation, but not necessary causation.



INTRODUCTION

- R.A. Fisher developed many of the methods that we will study in this course.
- In the 1950s large volume of research claimed connection between lung cancer and smoking.
- Fisher spent much of his late life fighting against these conclusions.
- He claimed it was a case of correlation mistaken for causation.

INTRODUCTION

Fisher compared it to a historical correlation between apple imports and marriage rates in England (Marsten, 2008).

Why did he dismiss the theory? (Stolley, 1991)

1. He was a paid consultant of the tobacco companies
2. A lifelong smoker.
3. He disliked anything that smacked of puritanism.