STA305/1004

Jan. 9, 2017

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/winter2017/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

ABRAHAM WALD AND THE MISSING BULLET HOLES

ABRAHAM WALD AND THE MISSING BULLET HOLES

- Abraham Wald born in 1902 in Austria.
- Emigrated to the U.S. and eventually became a professor at Columbia.

ABRAHAM WALD AND THE MISSING BULLET HOLES

- During World War II he spent much of his time in the Statistical Research Group (SRG). A classified program that assembled the best American statisticians to the war effort.

- The SRG was in an apartment building in NYC a few blocks from Columbia U.

ABRAHAM WALD AND THE MISSING BULLET HOLES

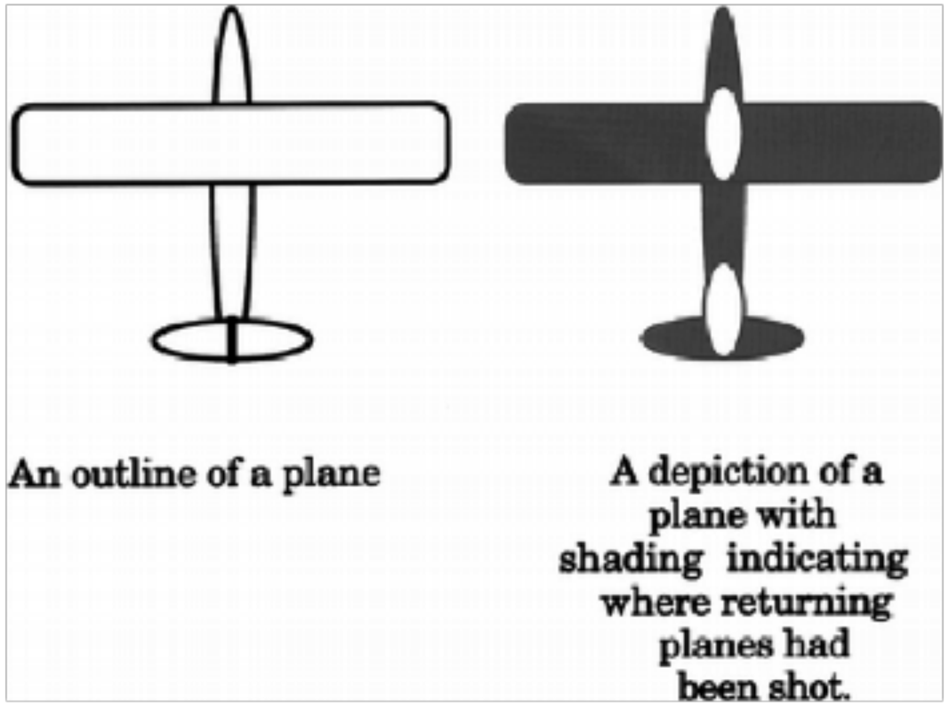
- The SRG was a very influential group and the military frequently listened to their advice.

- Wald at the time was still an “enemy alien”, he was not technically allowed to see the reports he was producing.

ABRAHAM WALD AND THE MISSING BULLET HOLES

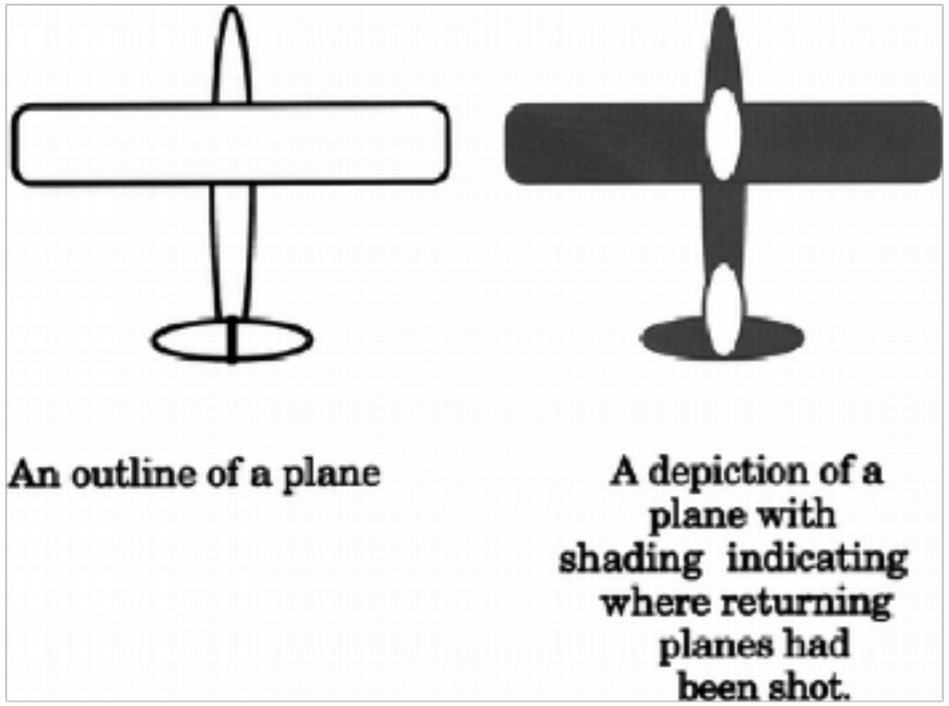
Question: You don't want planes to get shot down by enemy fighters, so you amour them. But armour makes planes heavier, and are less maneuverable and use more fuel. Armouring planes too much is a problem; armouring the planes too little is a problem.

Somewhere in between there's an optimum.


An outline of a plane
A depiction of a plane with shading indicating where returning planes had been shot.

ABRAHAM WALD AND THE MISSING BULLET HOLES

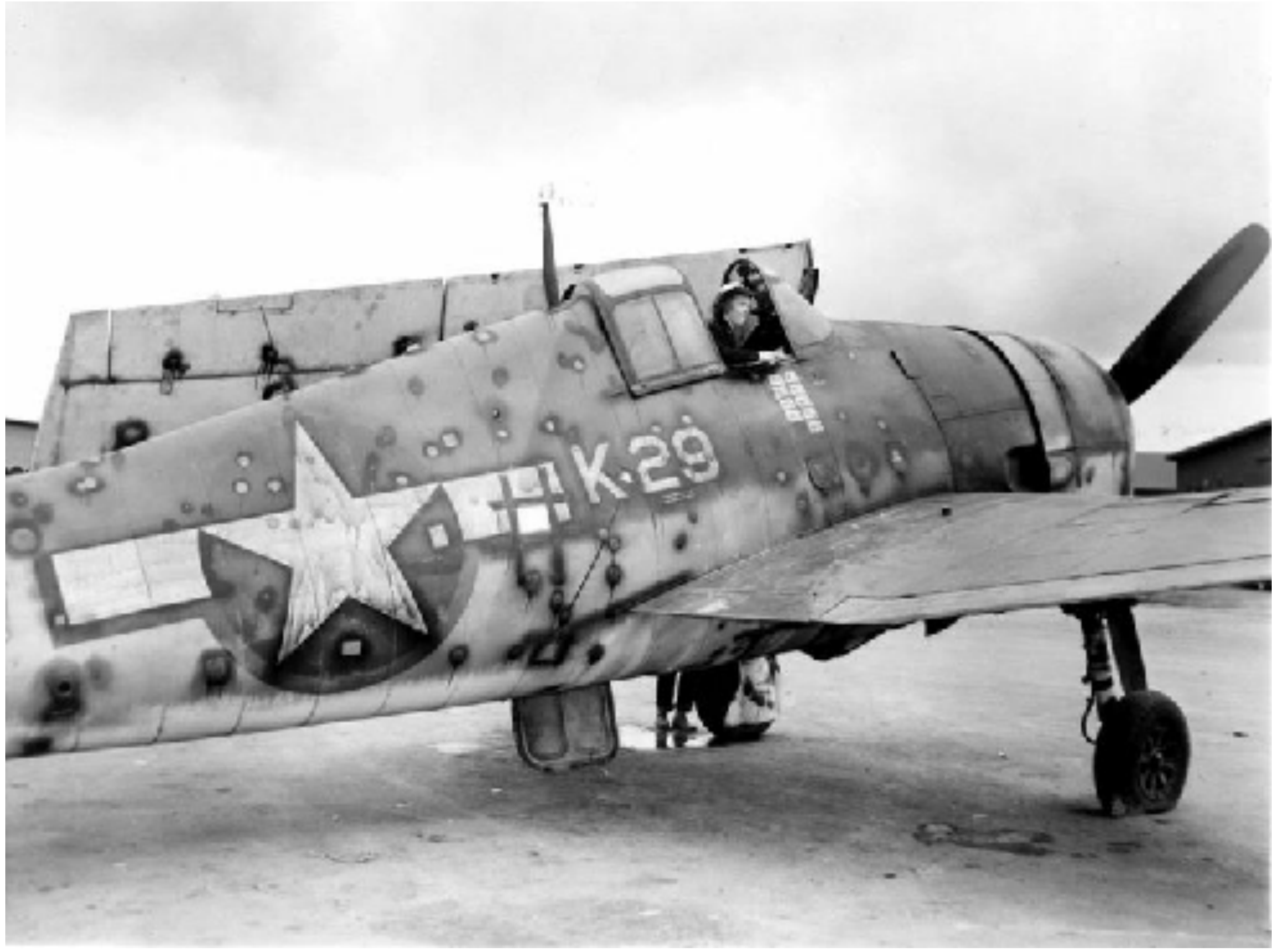
The military supplied the SRG with some data



An outline of a plane
A depiction of a plane with shading indicating where returning planes had been shot.

ABRAHAM WALD AND THE MISSING BULLET HOLES

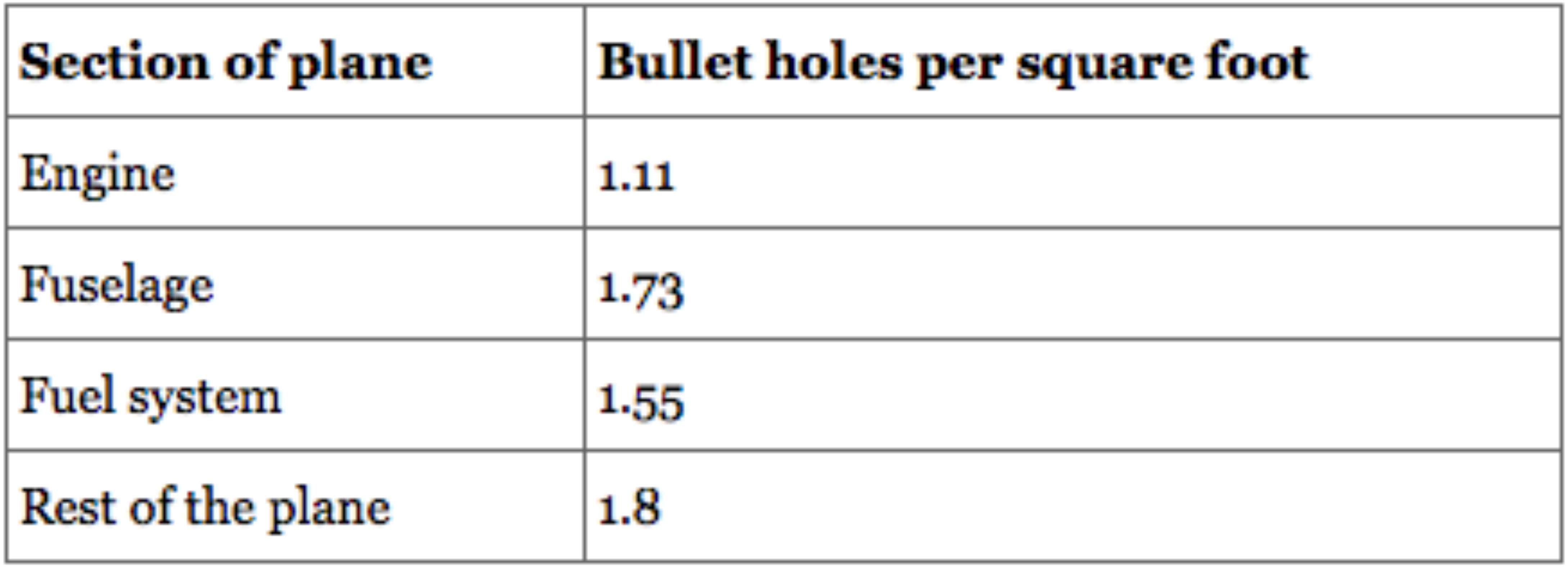
Planes were covered in bullet holes, but the holes weren't uniformly distributed across the aircraft.



ABRAHAM WALD AND THE MISSING BULLET HOLES

Data from American planes that came back from engagements over Europe.

What parts of the plane has the greatest need for armour?



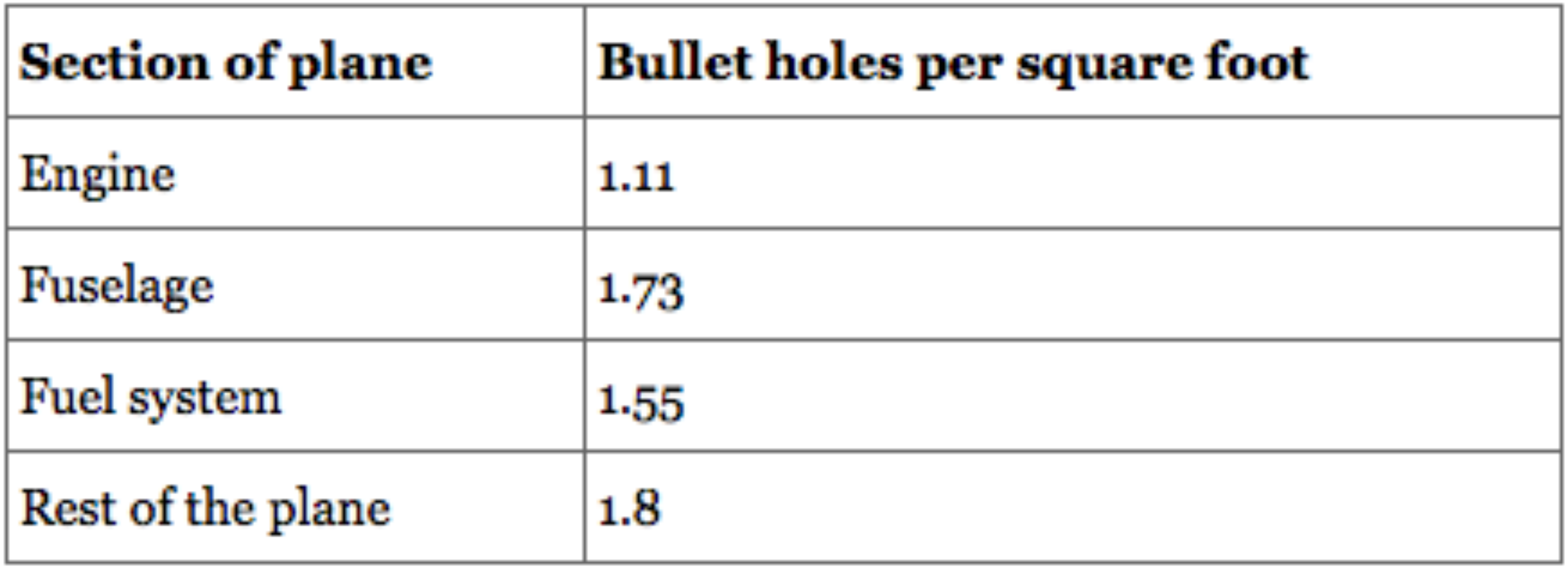
Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

ABRAHAM WALD AND THE MISSING BULLET HOLES

The officers saw an opportunity for efficiency.

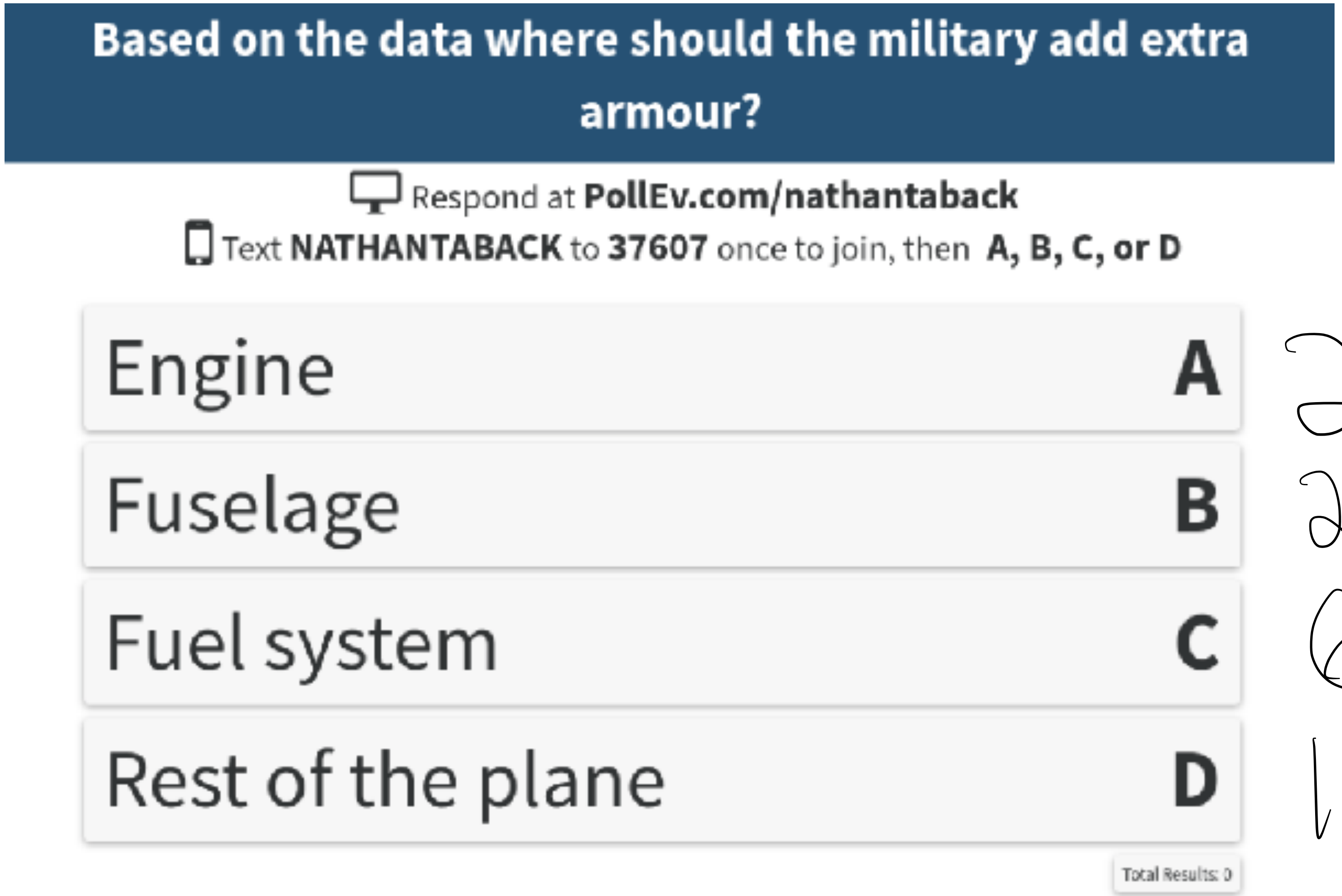
Get the same protection with less armour if you concentrate on places with the greatest need.

They asked Wald how much more armour belonged on those parts of the plane.



Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

ABRAHAM WALD AND THE MISSING BULLET HOLES



Based on the data where should the military add extra armour?

Respond at **PollEv.com/nathantaback**

Text **NATHANTABACK** to **37607** once to join, then **A, B, C, or D**

Engine **A**

Fuselage **B**

Fuel system **C**

Rest of the plane **D**

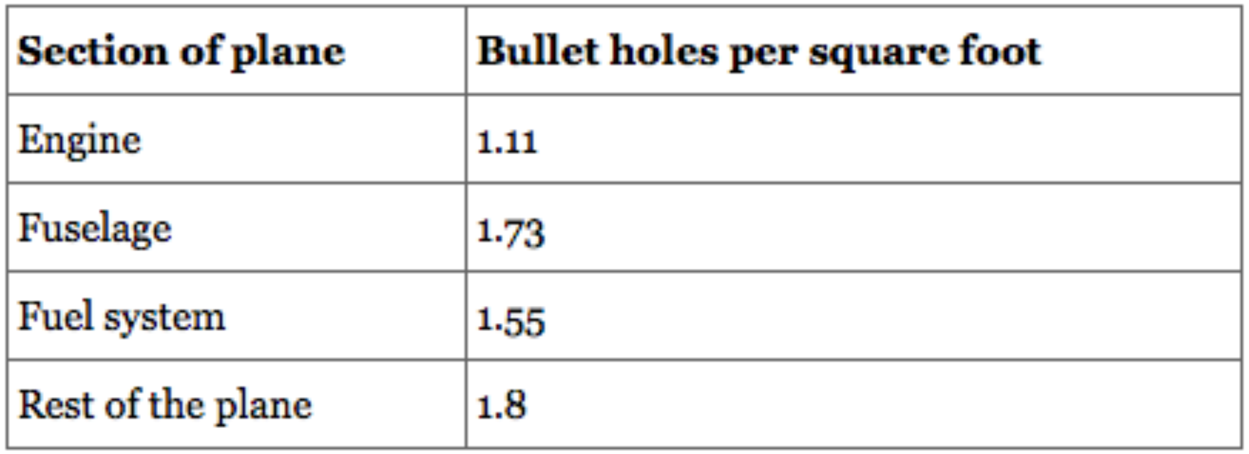
Total Results: 0

20

25

6

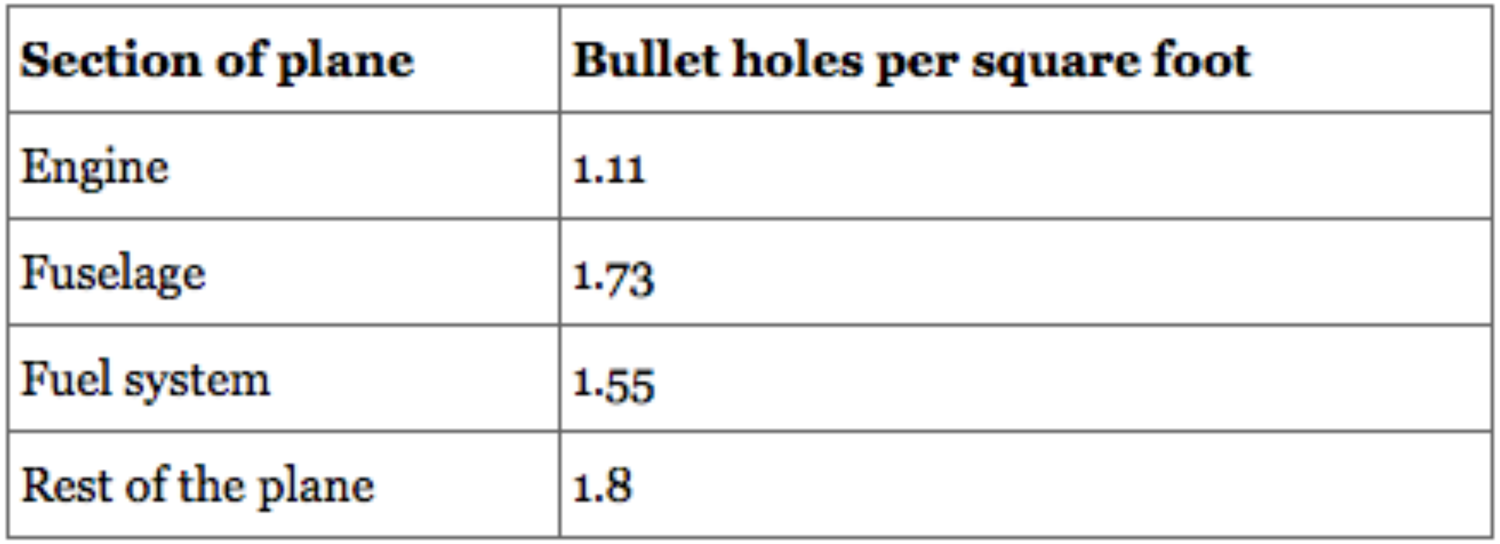
17



Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

ABRAHAM WALD AND THE MISSING BULLET HOLES

Wald said that the armour doesn’t go where the bullet holes are. It goes where the bullet holes aren’t: on the engines.



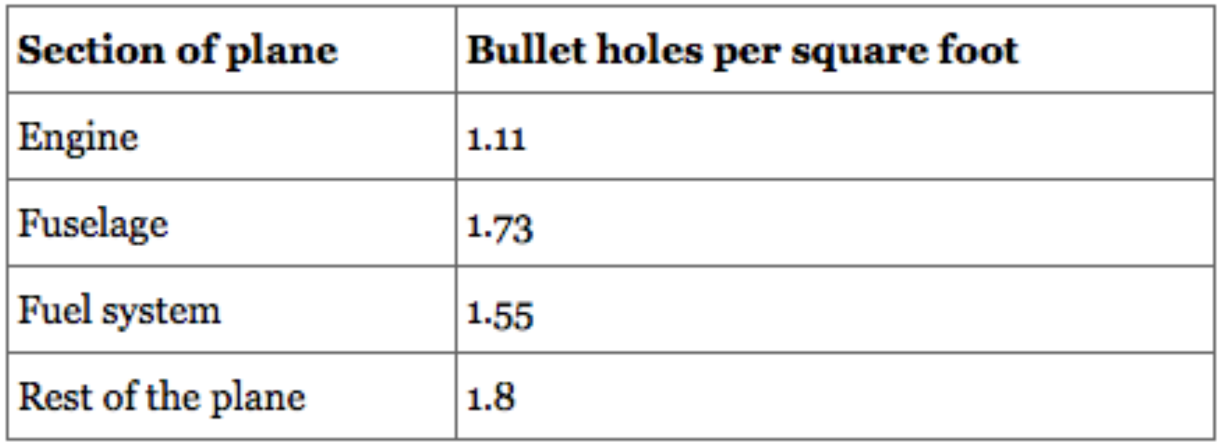
Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

ABRAHAM WALD AND THE MISSING BULLET HOLES

Wald’s insight was to ask: where are the missing holes?

The missing bullet holes were on the missing planes.

The reason planes were coming back with fewer hits to the engine is that planes that got hit in the engine weren’t coming back.



Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

ABRAHAM WALD AND THE MISSING BULLET HOLES

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

ABRAHAM WALD AND THE MISSING BULLET HOLES

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

A Statistician is always asking what assumptions are you making? Are they justified?

ABRAHAM WALD AND THE MISSING BULLET HOLES

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

A Statistician is always asking what assumptions are you making? Are they justified?

The officers were making the assumption that the planes that came back were a random sample of all the planes.

ABRAHAM WALD AND THE MISSING BULLET HOLES

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

A Statistician is always asking what assumptions are you making? Are they justified?

The officers were making the assumption that the planes that came back were a random sample of all the planes.

Once you recognize that you have been making this hypothesis, it takes a moment to realize that it's wrong.

ABRAHAM WALD AND THE MISSING BULLET HOLES

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

A Statistician is always asking what assumptions are you making? Are they justified?

The officers were making the assumption that the planes that came back were a random sample of all the planes.

Once you recognize that you have been making this hypothesis, it takes a moment to realize that it's wrong.

In statistical lingo, the rate of survival and location of bullet holes are correlated.

Survivorship bias

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

- 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.

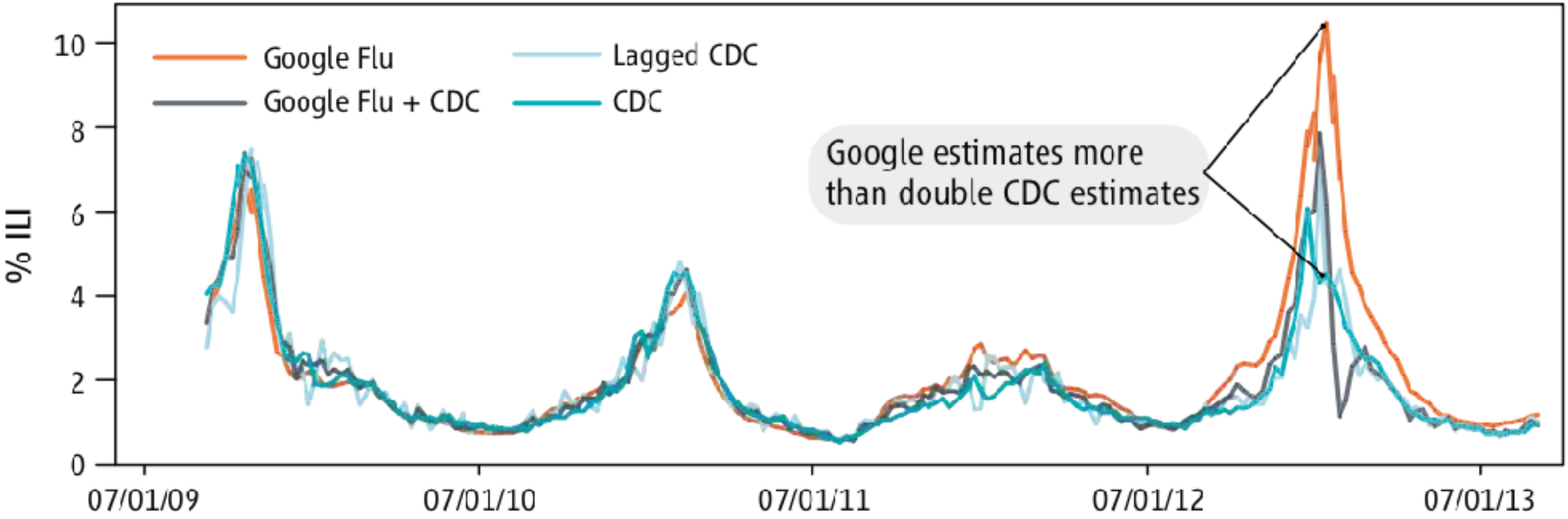
INTRODUCTION

- 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.

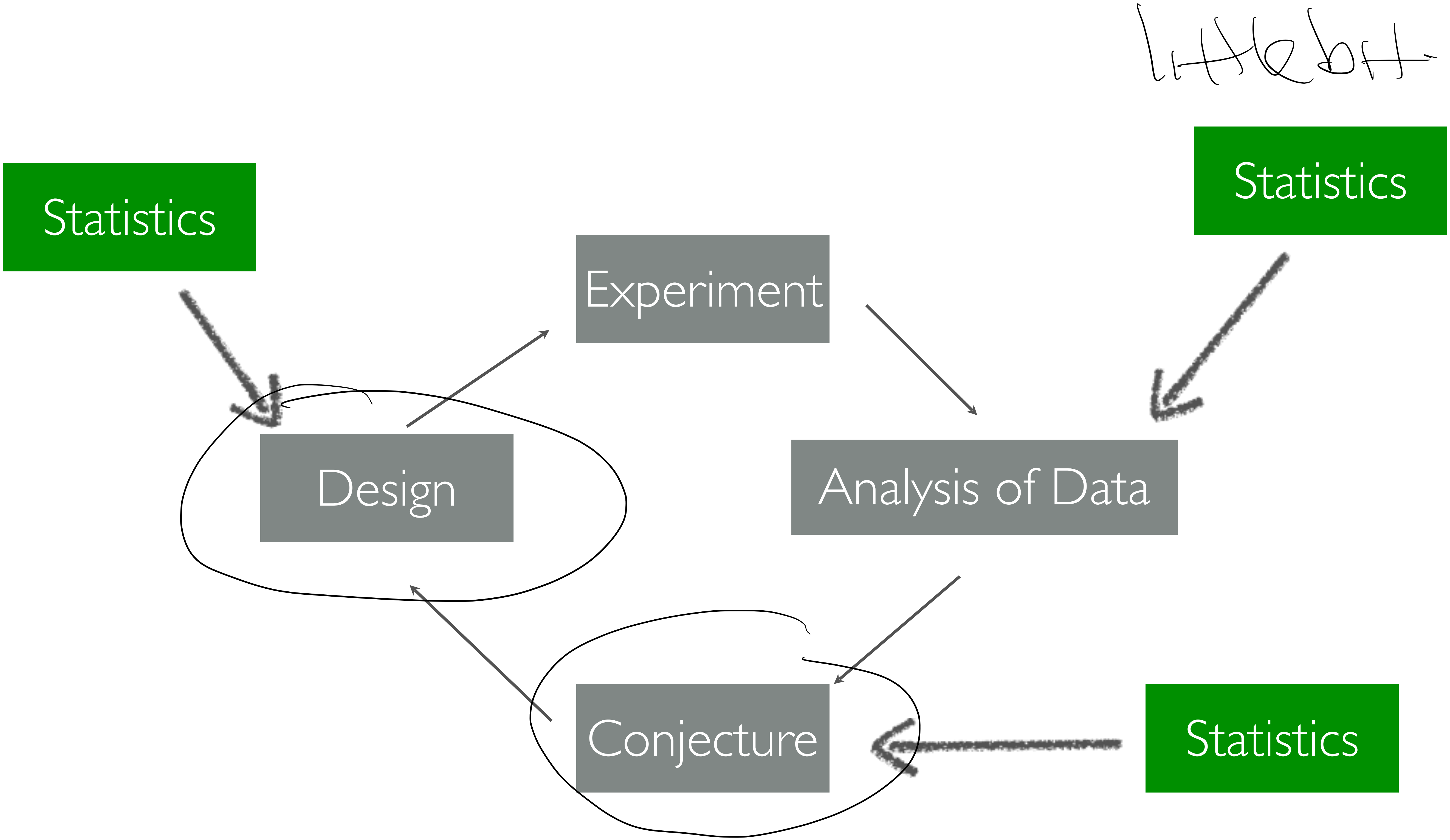
Google Flu (Lazer et al., Science 14 March 2014)



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

INTRODUCTION

Connected to Scientific Method



INTRODUCTION

Nathan Nathan

- 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?

Design#1

$Y_1 = W_1 + W_2$

$Y_2 = W_1 - W_2$

$y_1 + y_2 = 2w_1$

$\Rightarrow w_1 = \frac{y_1 + y_2}{2}$

$y_1 - y_2 = 2w_2$

$w_2 = \frac{y_1 - y_2}{2}$

$$\begin{aligned}
Var(w_1) &= Var\left(\frac{y_1 + y_2}{2}\right) \\
&= \frac{1}{2^2} Var(y_1 + y_2) \\
&= \frac{1}{4}\left(Var(y_1) + Var(y_2)\right) \\
&= \frac{1}{4}\left(\sigma^2 + \sigma^2\right) \\
&= \frac{2\sigma^2}{4} = \sigma^2/2
\end{aligned}$$

$$\mathrm{Var}\left(\frac{y_1 - y_2}{2}\right) = \mathrm{Var}(w_2)$$

$$\frac{1}{2^2}\mathrm{Var}(y_1 - y_2) =$$

$$\frac{1}{2^2}\left(\mathrm{Var}(y_1) + \mathrm{Var}(y_2)\right) =$$

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

$$\frac{1}{2^2}2\sigma^2 =$$

$$\frac{\sigma^2}{2} =$$

$$\mathrm{Var}\left(\frac{w_1^{(1)}+w_2^{(2)}}{2}\right)=\frac{1}{2^2}\left(\mathrm{Var}\left(w_1^{(1)}\right)+\mathrm{Var}\left(w_1^{(2)}\right)\right)$$

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

$$=\frac{2\sigma^2}{2^2}$$

$$=\frac{\sigma^2}{2}$$

$$\mathrm{Var}\left(\frac{w_2^{(1)}+w_2^{(2)}}{2}\right)=\frac{\sigma^2}{2}$$

With individual weighings we need 4 weighings to get the same precision as two weighings using the proposed design.

That is, weighing two apples in one pan then weighing two apples in seperate pans.

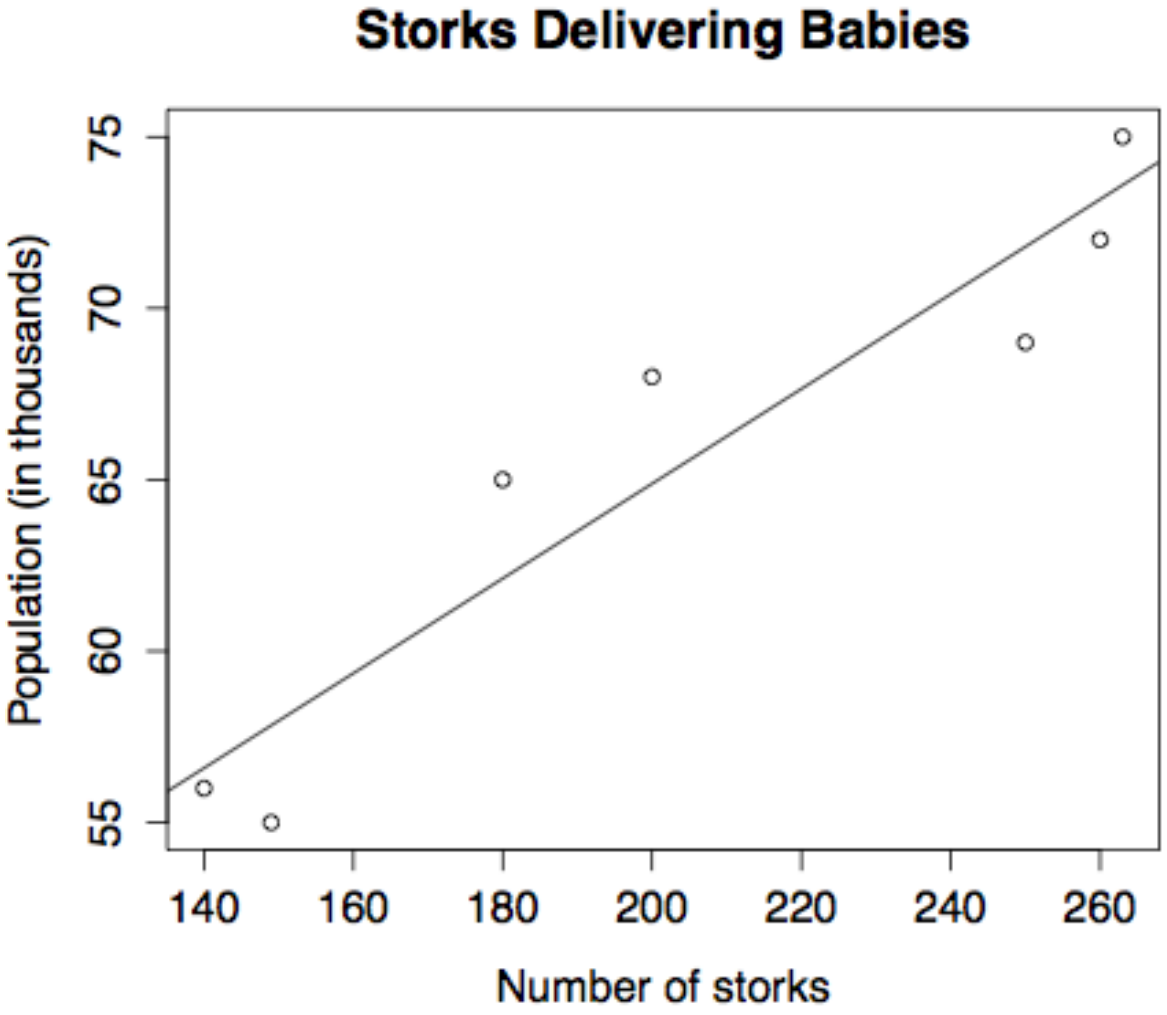
INTRODUCTION

- 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.

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



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.