Statistics Road Tour 2012

Changing Needs in a Changing World

Part I: Chris Wild

Department of Statistics
University of Auckland, New Zealand

THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

The data world …
is getting a whole lot bigger

The data world …
is getting a whole lot bigger

- There is an explosion in the …
  - quantities of data being collected
  - conceptions of what constitutes data
  - settings in which it can arise
  - ways of looking at it

Further, Faster, Better

- Data world exploding
- Green shoots in Software
- Vision
- Future is visual
- Accelerators

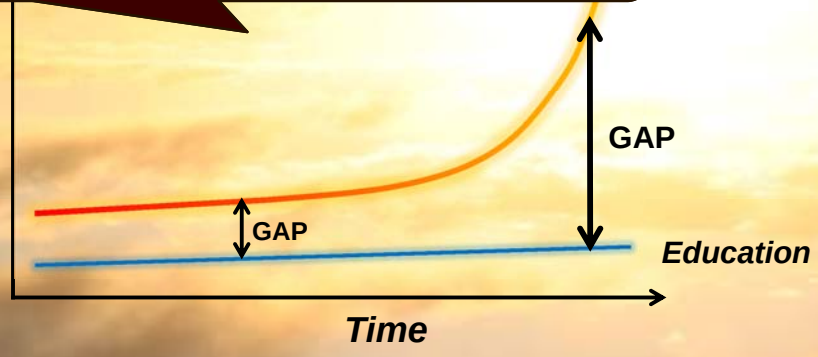
The data world …
is getting a whole lot bigger

Can't just keep illuminating same small patch

- Need to get much …
  - *further*
  - *faster*
  - & with ***better*** *comprehension*

A Growing Gap

Need to **increase the speed** with which we can **open up the world of data**
- But moving slowly enough that teachers can deal with the changes

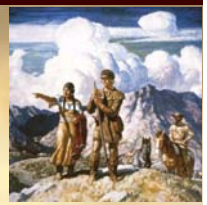


Where did New Curriculum come from?

- Most thrashed out 6 years ago
  in large consultative group organised by NZ Stats Assoc's Ed Committee involving:
  - professional statisticians
  - university lecturers in statistics
  - experienced practicing teachers
  - teacher educators and teacher developers
- volunteers working in their spare time under extremely tight, officially-imposed, time constraints

Where did New Curriculum come from?

It was forward looking



- to the best of our abilities at the time
  - potential to remain in force for a long time
    (Last curriculum sealed in at Year 13 level for ~ 20 years)

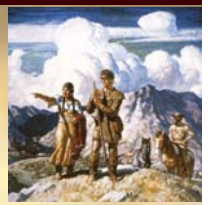
Endeavours to emphasise the fundamentals of the subject

- **Big ideas**
  - which will not change with time

  ahead of the details of the ways things are done
  - which are continually changing

Where did New Curriculum come from?

It was forward looking



- to the best of our abilities at the time
  - potential to remain in force for a long time
    (Last curriculum sealed in at Year 13 level for ~ 20 years)

Endeavours to emphasise the fundamentals of the subject

- **Big ideas**
  - which will not change with time

  ahead of the details of the ways things are done
  - which are continually changing

From **"putting numbers into formulae"** to **"conceptual understandings"**
- help enable them to start "thinking about data like a statistician"

Where did New Curriculum come from?

It was forward looking

Endeavours to emphasise the fundamentals of the subject

If the new curriculum could not …

- **adapt to the rapidly increasing expansion** of the **world of data**
- **reflect the ways in which computers dominate statistics & are continuing to revolutionise it**
  - (Causes some problems temporarily that we collectively have to find ways to work around)

then it would quickly **sink into irrelevance**

Where did New Curriculum come from?

It was forward looking

Endeavours to emphasise the fundamentals of the subject

If the new curriculum could not …

- **adapt to the rapidly increasing expansion** of the **world of data**
- **reflect the ways in which computers dominate statistics & are continuing to revolutionise it**
  - (Causes some problems temporarily that we collectively have to find ways to work around)

then it would quickly **sink into irrelevance**

It helps that "Getting the numbers" moving from the main part of what students have to do to a fairly trivial part

Some of the changes

- All serious statistical analysis uses computers
  - Only real defence for by-hand procedures is to help better understand something done on a computer
    - So educational emphasis shifts from "*how to get the numbers*" to "*discovery through data*"
      - See Handout ***"Confidence Intervals: What matters most?"***
- Statistics is moving much more towards
  - Visualisations via computer graphics
  - Modern computer-intensive inference methods
    - esp. bootstrap and randomisation

Some of the changes

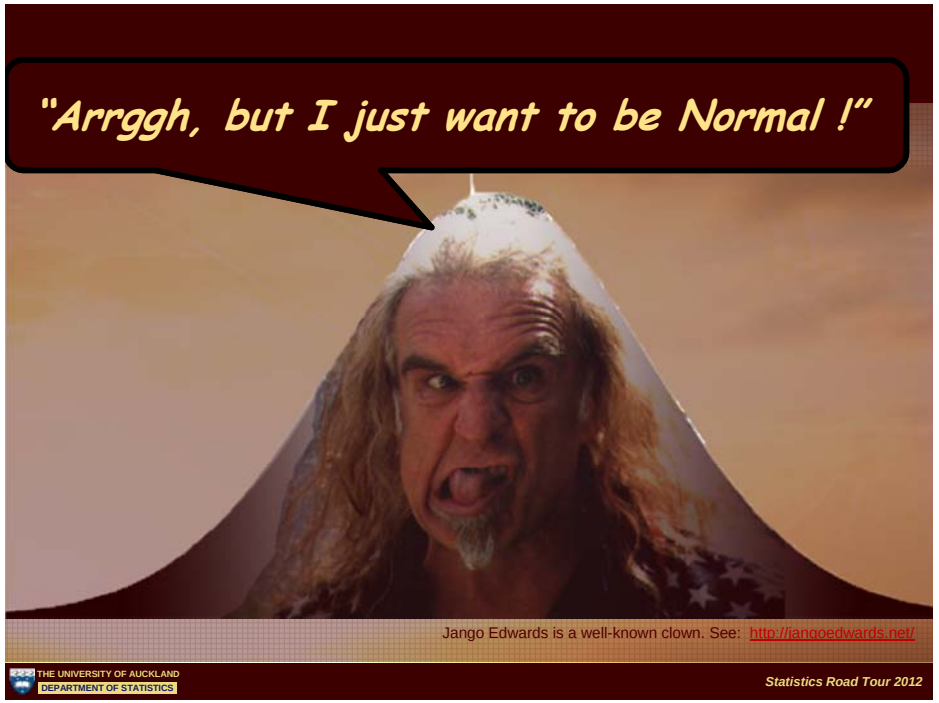
- All serious statistical analysis uses computers
  - Only defence for by-hand procedures is to help better understand some thing done on a computer
    - So educational emphasis shifts from "how to get the numbers" to "discovery through data"
      - See Handout ***"Confidence Intervals: What matters most?"***
- Statistics is moving towards modern computer-intensive inference methods
  - esp. bootstrap and randomisation

Some of the changes

- All serious statistical analysis uses computers
  - Only defence for by-hand procedures is to help better understand some thing done on a computer
    - So educational emphasis shifts from "how to get the numbers" to "discovery through data"

...dout ***"Confidence Intervals: What matters most?"***

Google's Tim Hesterberg (2006) … bootstrapping and randomisation "increasingly pervade statistical practice. They offer ease of use: the same basic procedures can be used in a wide variety of applications, without requiring difficult analytical derivations. This frees statisticians to use a wider range of methods, not just those for which easy formulas for confidence intervals or hypothesis tests are available."



Jango Edwards is a well-known clown. See: http://jangoedwards.net/

Bootstrap

- Brainchild of **Stanford University's Brad Efron**
  - On virtually anyone's list of the 20th century's greatest statisticians and his biggest contribution
- Justified by both high-powered mathematical theory (hundreds of theoretical papers) & extensive computer simulation
- Enables us get further into data world more quickly
  - The most generally applicable method there is of generating confidence intervals
    - Single idea can use for vast majority of quantities of interest
      - e.g. medians, proportions, quartiles, measures of spread (e.g. interquartile ranges), differences in means, medians and proportions, ratios of spreads, regression slopes, correlations and many, many more besides

Bootstrap

- Brainchild of **Stanford University's Brad Efron**
  - On virtually anyone's list of the 20th century's greatest statisticians and his biggest contribution
- Justified by both high-powered mathematical theory (hundreds of theoretical papers) & extensive computer simulation
- Enables us get further into data world more quickly
  - The most generally applicable method there is of generating confidence intervals
    - Single idea can use for vast majority of quantities of interest
      - e.g. medians, proportions, quartiles, measures of spread (e.g. interquartile ranges), differences in means, medians and proportions, ratios of spreads, regression slopes, correlations and many, many more besides

Because the bootstrap is much more general the mathematics is much harder

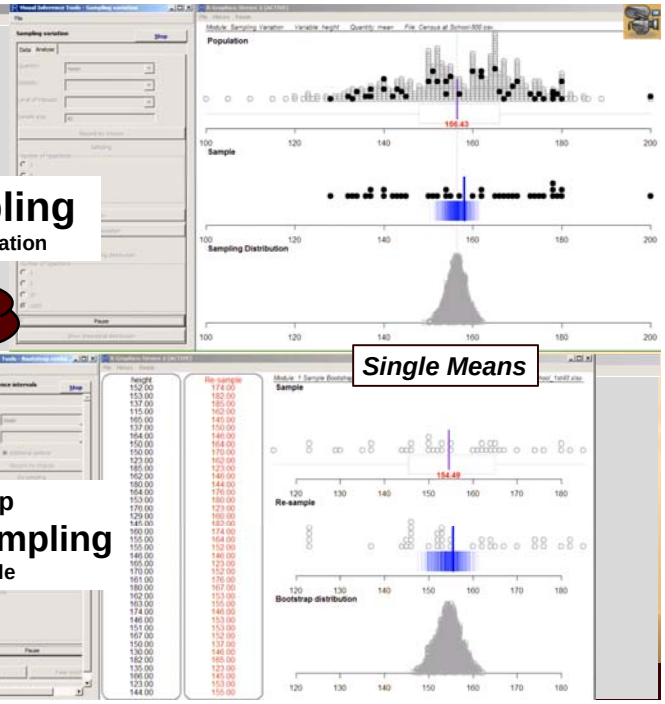
Contrast to methods based on distributional assumptions (e.g. normal) where need to learn a different recipe for each new thing you want to do

Bootstrap
- Sampling with replacement
  - What is it?
  - What does it do?
  - How could we use it?
  - Why might it work?
- Constructing Bootstrap intervals
- Does it work?

Sampling
from population

Looks very similar

Bootstrap Re-Sampling
from sample



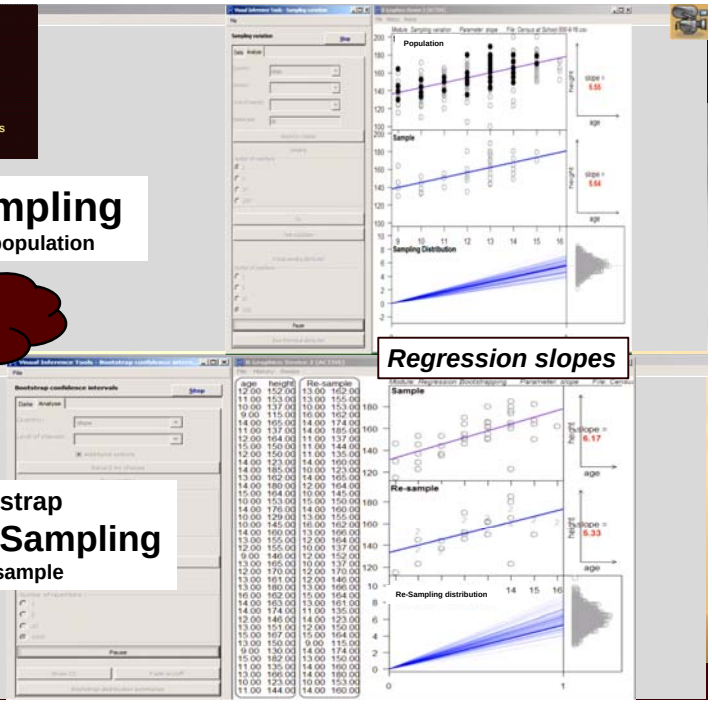
THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

Bootstrap
- Sampling with replacement
  - What is it?
  - What does it do?
  - How could we use it?
  - Why might it work?
- Constructing Bootstrap intervals
- Does it work?

Sampling
from population

Looks very similar

Bootstrap Re-Sampling
from sample



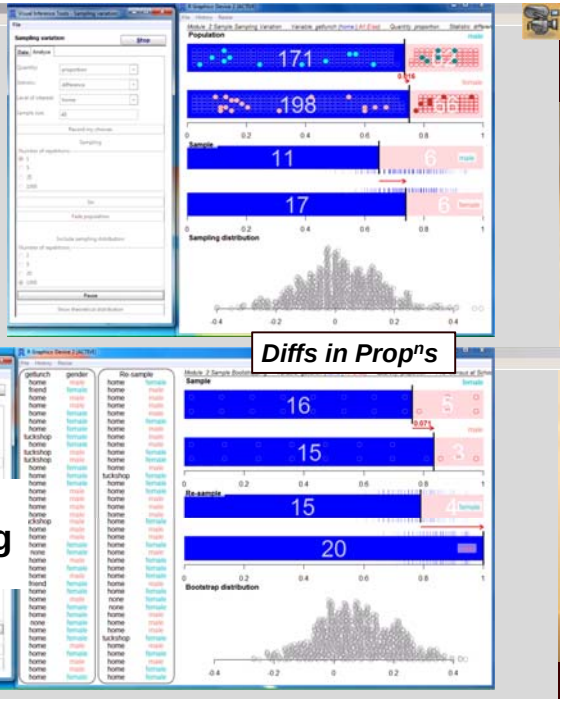
THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

Bootstrap
- Sampling with replacement
  - What is it?
  - What does it do?
  - How could we use it?
  - Why might it work?
- Constructing Bootstrap intervals
- Does it work?

Sampling
from population

Looks very similar

Bootstrap Re-Sampling
from sample



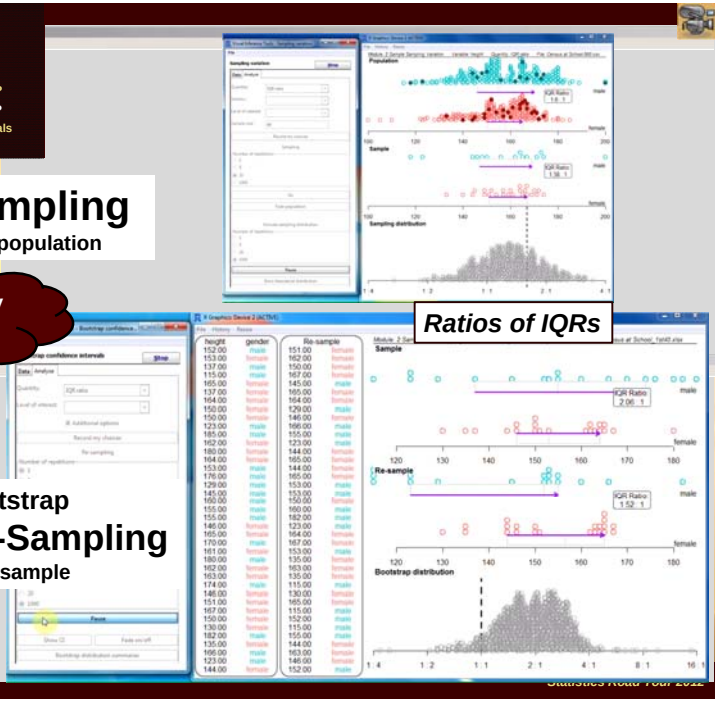
THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

Bootstrap
- Sampling with replacement
  - What is it?
  - What does it do?
  - How could we use it?
  - Why might it work?
- Constructing Bootstrap intervals
- Does it work?

Sampling
from population

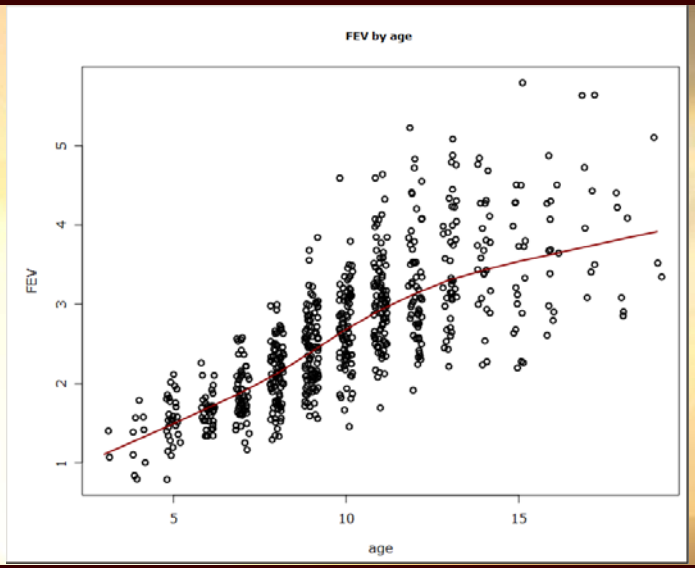
Looks very similar

Bootstrap Re-Sampling
from sample

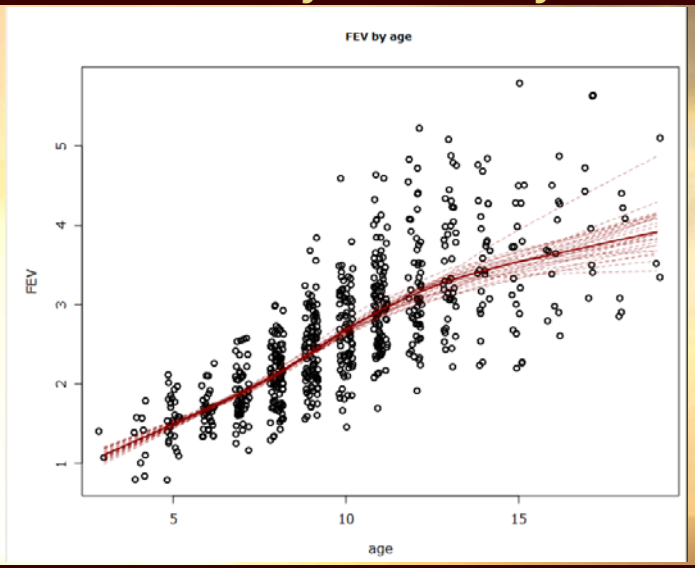


THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

More complex: Scatterplot with smoother



Bootstrap smooths added to convey uncertainty



Bootstrap

- Justified by both high-powered mathematical theory (hundreds of theoretical papers) & extensive computer simulation
- Can get much further into data world more quickly
- Easier to understand
  - More transparent relationship between problem we are trying to solve and method of generating intervals
  - Gives better grounding in “the logic of inference”

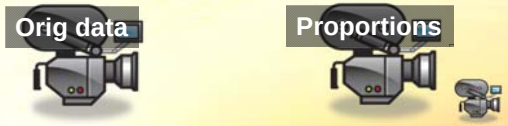
Bootstrap

- Justified by both high-powered mathematical theory (hundreds of theoretical papers) & extensive computer simulation
- Can get much further into data world more quickly
- Easier to understand
- Not reliant on distributional assumptions
- Gives same CIs in the most familiar problems

  (modulo notes on handout)
- Standard tool for very hard problems

Randomisation test visualisations

Can random labelling alone ("*chance alone*") produce differences as large those as I'm seeing?



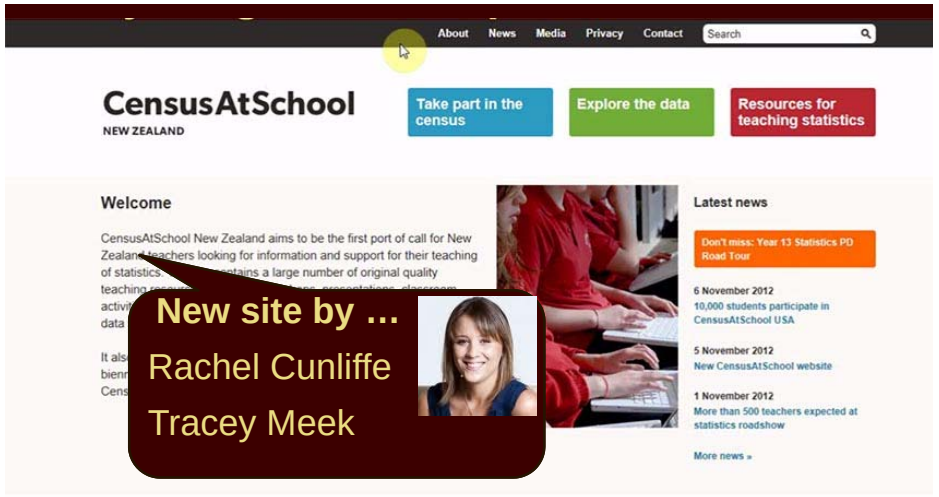
We are in transition

But people in ... are using ...

- Leadership in change comes from centre of the field
  - takes time to percolate out to the peripheries
- Anyone who will do substantial statistical analyses will take some statistics at university
  - Many more do
    - 2 out of every 3 students coming to UoA do at least 1 stats paper
  - UoA is teaching both computer-intensive and traditional and the links between them in 1st year
    - Finding the traditional makes more sense built on the better grounding in the logic of inference from boo & randomisation

This is new for us too

~Everything will end up on Census At School



Thank you