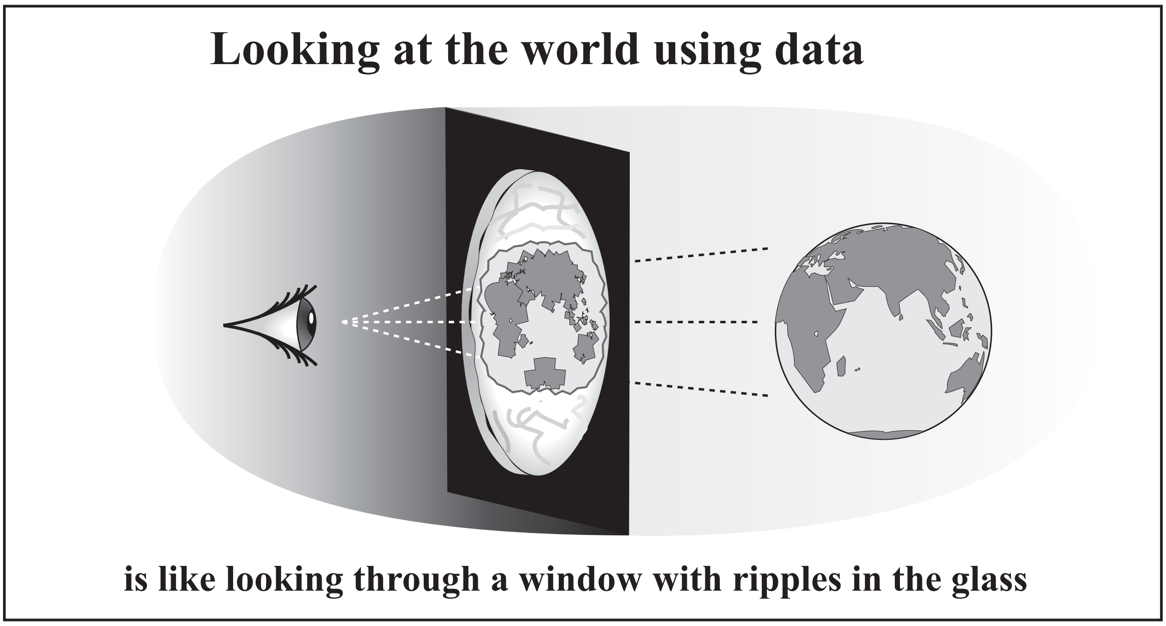
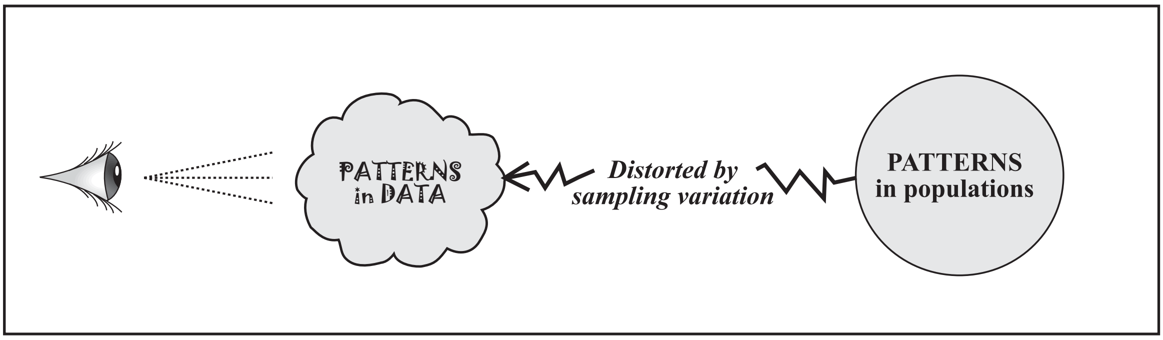
# **Handout 1** for Chris Wild's plenary talk for the
## ***Auckland Mathematics Association Term 2 Meeting***
14 June 2008

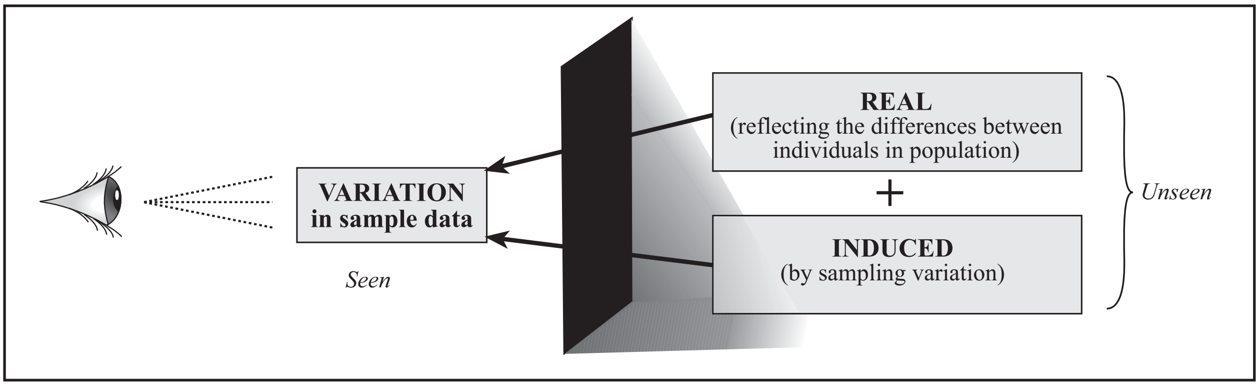
***(Minor Clarifications made to pages 4&5 September 2008)***

*Diagrams by Chris Wild*
*in collaboration with Nicholas Horton, Maxine Pfannkuch & Matt Regan*

## **"What I see may not quite be the way it really is"**

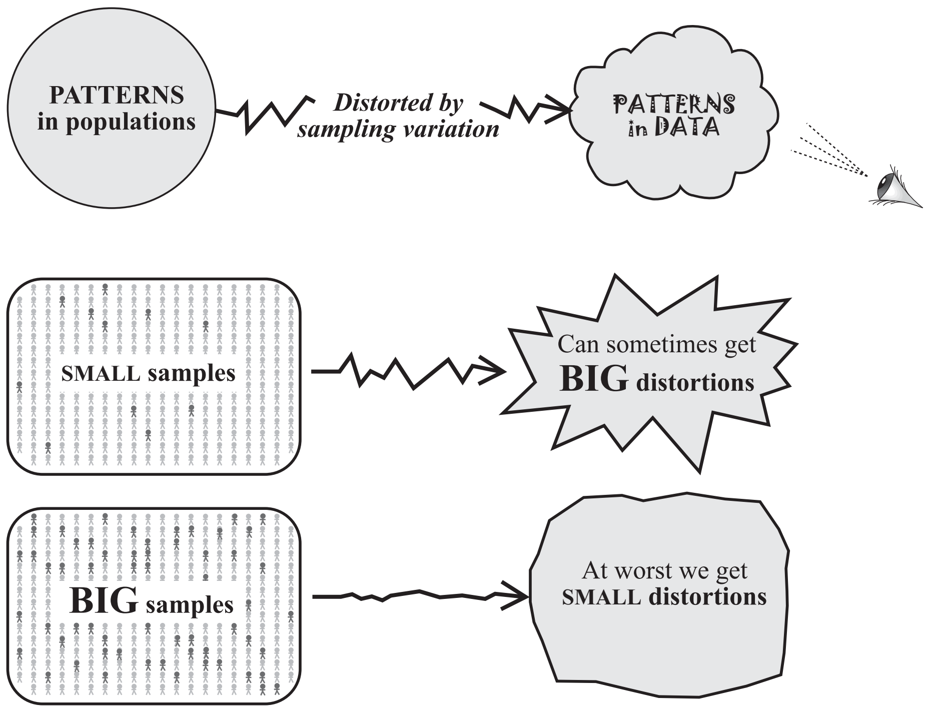


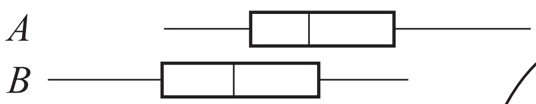

  Improvement suggestions to c.wild@auckland.ac.nz

# "What I see may not quite be the way it really is"




*Bigger sample size* → *More information* → *Allows me to make more precise claims about what is happening back in the population*

*Patterns in data (we have only described the main one)*




**Description:**

- Distribution of *A*-values shifted up scale from that of *B*-values
- *A*-values bigger on average than *B*-values

*Can I claim there is a similar pattern back in the populations?*

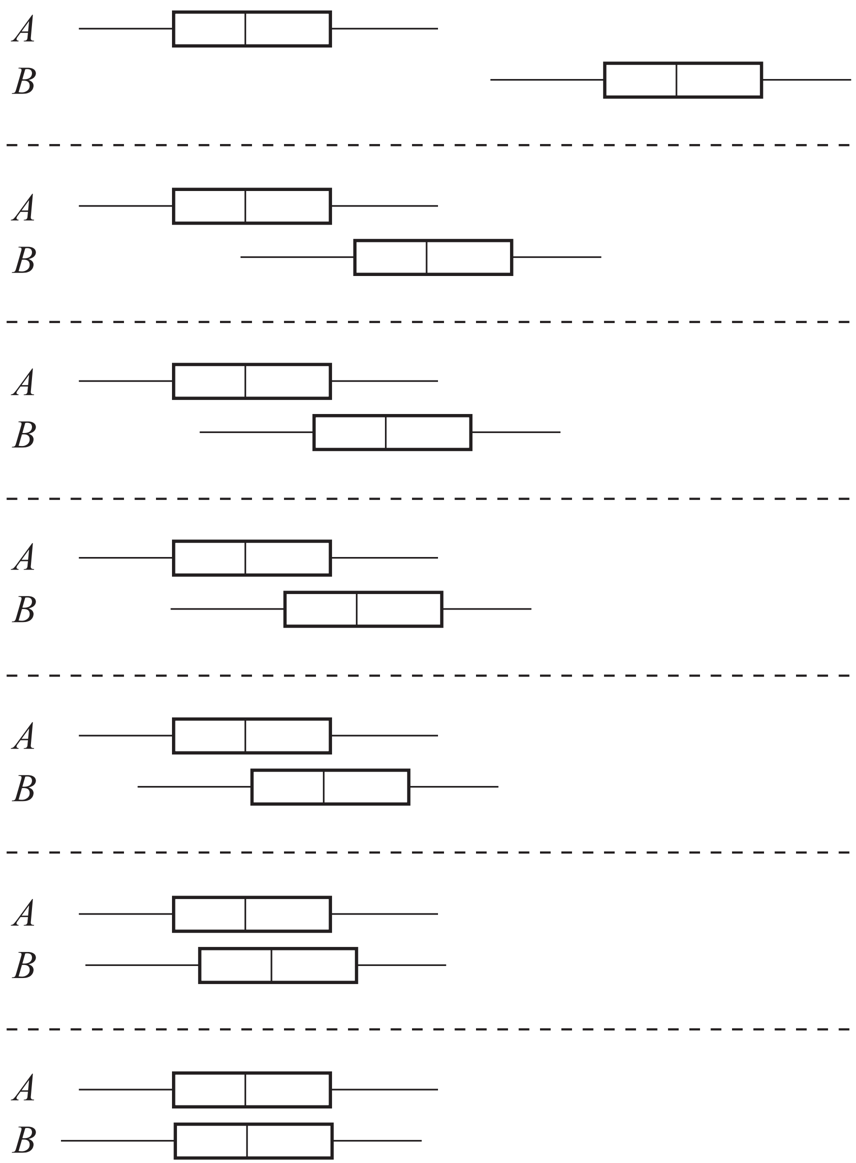
**Assumed Student development at this point:**

- Can describe what they see ***in the observed data***
- Aware of the effects of sampling variation in visual displays
    - Sampling variation alone can produce shifts
    - These shifts are small in very large samples
    - They can be misleadingly large in small samples

**Inference as the next step:**

- Will I claim *A*-values are also bigger on average ***back in the populations***?
    - I will if the shifts are bigger than those produced by sampling variation
    - Otherwise I will not. I cannot tell whether *A*-values are bigger than *B*-values *back in the populations.* It may even be the other way around.

## Observed data:




## Back in the populations:

“Do *B* values tend to be bigger than *A* values?”

***My call is ....***

**B is bigger**

*B* is bigger

*all sample sizes*

*Claim* “B is bigger” *if both sample sizes > 20*

*What’s my call here?*

*What’s my call here?*

*Call* “Cannot tell” *unless both samples are huge*

*Larger random samples have more information about the populations they came from.*

*Thus, with larger random samples, we can make the “B is bigger” call from smaller shifts*

***But how do we decide?***

- *depends on educational level of students*
- *see next page ...*

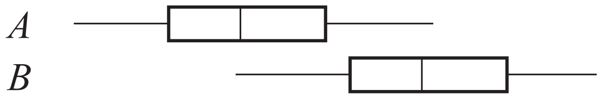
**Cannot tell**

*all sample sizes*

**Warning** to teachers: avoid doing this sample with sizes smaller than about 20 in each group. Small samples quite often give rise to unstable and often very strange boxplots
To echo the previous diagram, we get very large distortions -- see plots for samples of size 10 on page 6

# "How to make the call" by Curriculum Level

## At all levels:

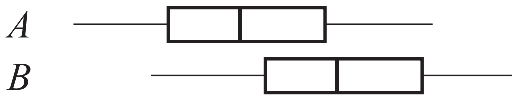



***If there is no overlap of the boxes***, or only a very small overlap
make the claim "***B tends to be bigger than A***" back in the populations

***Apply the following when the boxes do overlap ...***

---

## Curriculum Level 5: *the 3/4-1/2 rule*

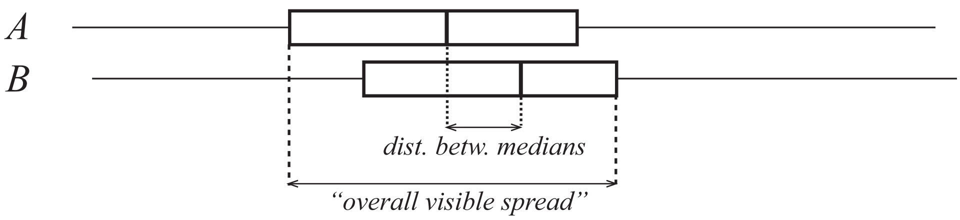



If the median for one of the samples lies outside the box for the other sample
(e.g. "*more than half of the B group are above three quarters of the A group*")
make the claim "***B tends to be bigger than A***" back in the populations

[Restrict to samples sizes of between 20 and 40 in each group]

---

## Curriculum Level 6: *distance between medians as proportion of "overall visible spread"*




Make the claim ***B tends to be bigger than A*** back in the populations
if distance between medians is greater than about ...

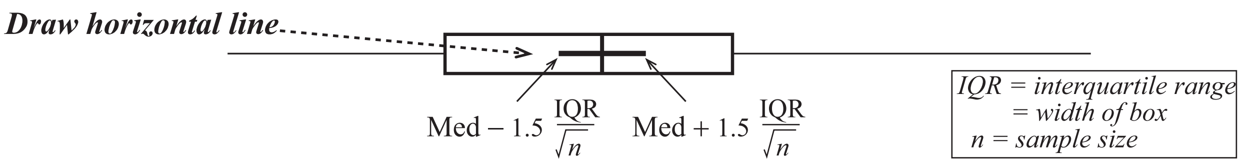
**1/3** of overall visible spread for sample sizes of around **30**

**1/5** of overall visible spread for sample sizes of around **100**

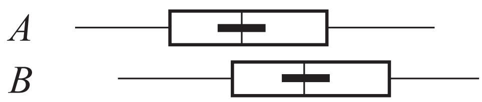
[***Could also use*** 1/10 of overall visible spread for sample sizes of around 1000]

---

## Curriculum Level 7: *based on informal confidence intervals for the population median*




Make the claim ***B tends to be bigger than A*** back in the populations




if these horizontal lines (intervals) do not overlap

---

## Curriculum Level 8: ***on to formal inference***

# Some notes about the rules

## At all levels:

**Emphasize the visual, keep the eyes constantly on the plots**

- What we are doing here is just one small step in interpreting a comparison
  - It is definitely not "what the statistics module is all about"
- While our depictions are in terms of 2 groups do not hesitate to use more groups
  - The stories uncovered in data by comparing several groups are often much more interesting

## Curriculum Level 5: *the 3/4-1/2 rule*

- The intuitive idea here is "the majority of the *B* group is bigger than the 'the great whack' of the *A* group"
- Operate as "the visual shift is big enough to make the call if the median for one of the samples lies outside the box for the other sample" regardless of whether this happens on the lower or upper side of the graphs.
- *Technical aside*: sampling variation alone does not often produce shifts large enough to trigger this rule
  - about 15 times in 100 for samples of size 20 in each group, 7 times in 100 for samples of 30, 3 times in 100 for samples of 40, 1 times in 2,500 for samples of size 100.

## Curriculum Level 6: *distance between medians as proportion of "overall visible spread"*

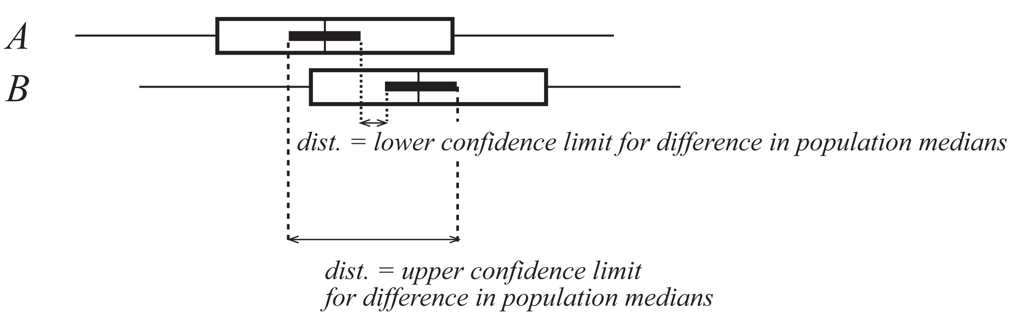
**Students should only be making rough "eye-ball" judgements**

- You are getting the students accustomed to using an idea, not the precise implementation of an algorithm
  - Do ***not*** make this hinge on accuracy of application of the 1/3 and 1/5 rules
- Whether the distance is bigger than 1/3 or 1/5 will often be obvious
  - Otherwise they should do a freehand subdivision of a line into thirds or fifths and then decide
- *Technical aside*: sampling variation alone seldom produces shifts large enough to trigger these rules (about 8 times out of 100 for both rules at the listed sample sizes )

## Curriculum Level 7: *based on informal confidence intervals for the population median*

**About the intervals they are drawing and interpreting**

- They cover the true population Median for approximately 9 out of 10 samples taken (show with simulations)
  - So appeal to "the population median for *A* is probably in here somewhere", similarly for *B*
  - This leads naturally to "*B* bigger than *A*" claim when they do not overlap
    - *Technical aside 1*: sampling variation hardly ever causes shifts big enough to make us mistakenly claim that *B* is bigger than *A* or vice versa using this method (only about once per 40 pairs of samples)
    - *Technical aside 2*: When the intervals do not overlap, a confidence interval for the difference in population medians ranges from the smalller distance between the intervals to the larger

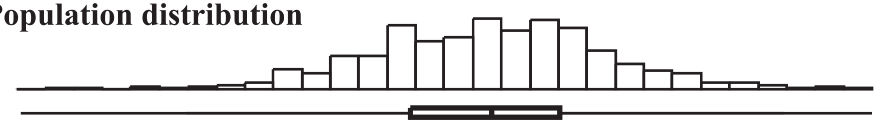


# Examples of shifts caused purely by sampling variation

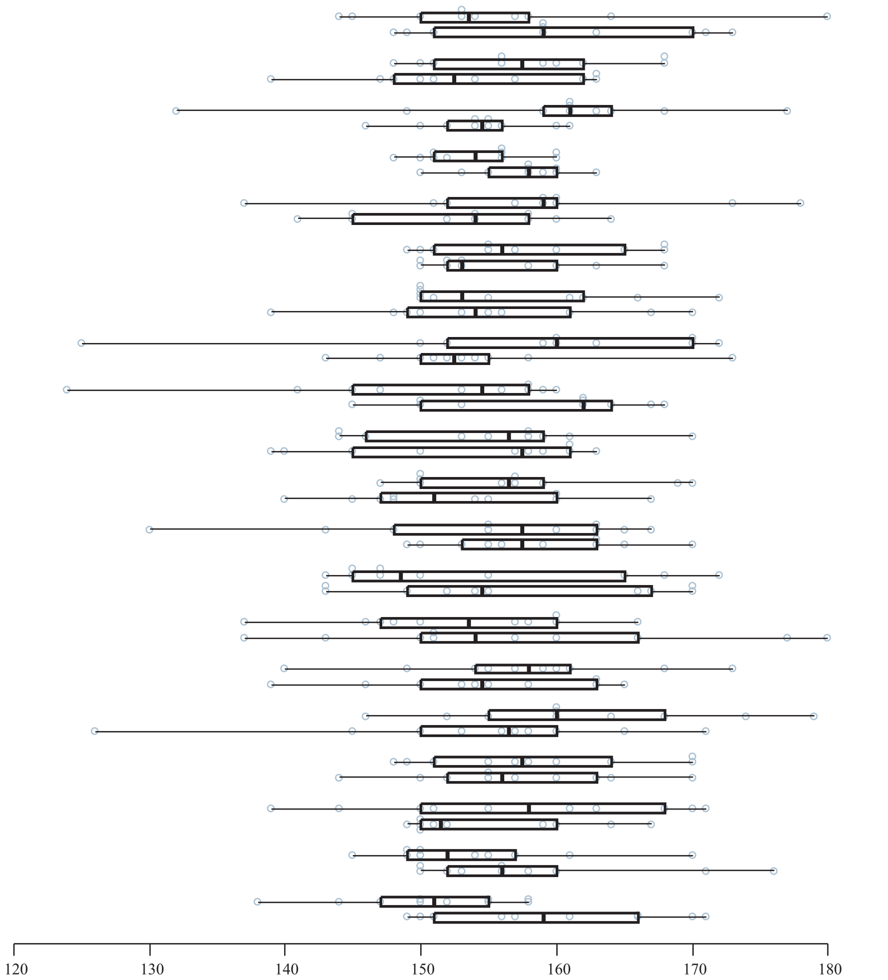
**The population being sampled is the 12 yearolds in NZ CensusAtSchool database**
**the measure used is height**

*This single population is being sampled independently over and over again so any shifts seen are due solely to the sampling*

**Population distribution**

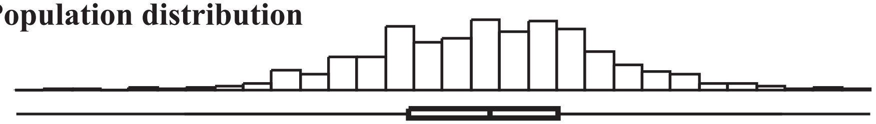


**Samples of size 10**

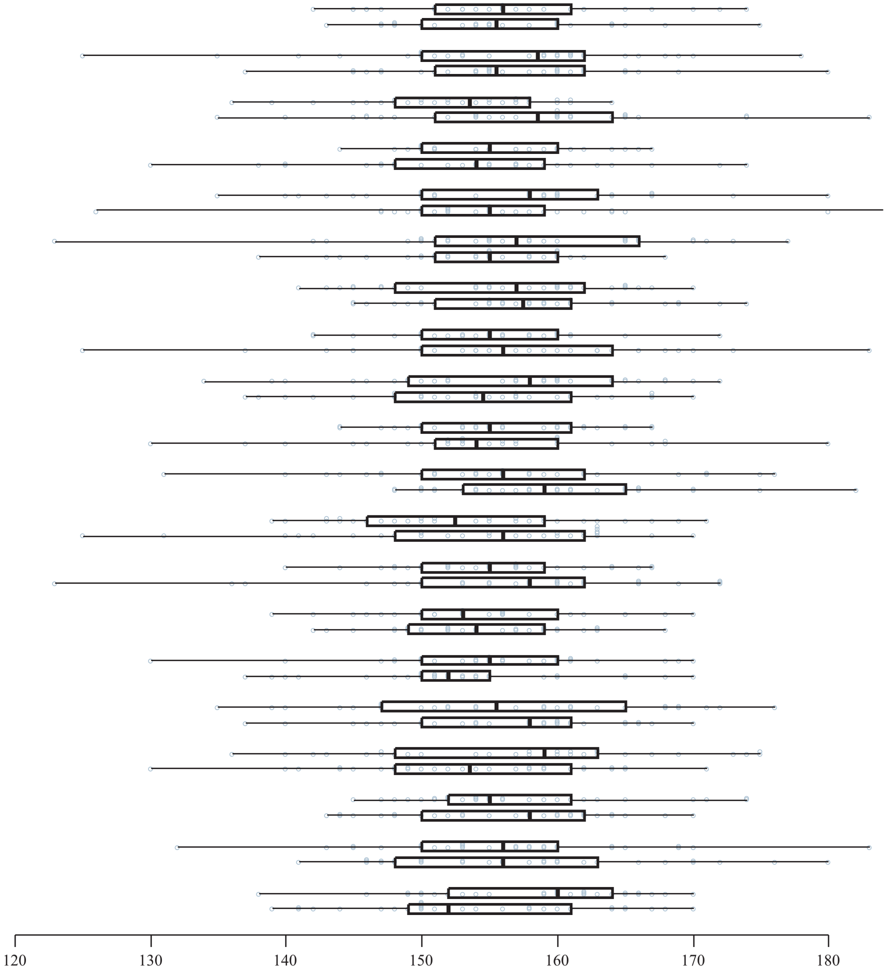



Samples of size 10 shown to demonstrate why we should not be working in this way with such small samples

**Population distribution**



**Samples of size 30**



# Examples of shifts caused purely by sampling variation

**The population being sampled is the 12 yearolds in NZ CensusAtSchool database**
**the measure used is height**

*This single population is being sampled independently over and over again so any shifts seen are due solely to the sampling*