Massey University

Activities for Developing Understanding of Sampling Distributions and Confidence Intervals

Marie Fitch,

IIMS, Massey University, Auckland

30 November 2007

Overview

- Sampling distribution for mean
- Confidence intervals
- Resampling

NZ Curriculum Level 8

- Make inferences from surveys and experiments:
  - Determining estimates and confidence intervals for means, proportions and differences, recognising the relevance of the central limit theorem
  - Using methods such as resampling or randomisation to assess the strength of evidence

Developing the concept of a sampling distribution

- A population of sample means
- A 'class load' of sample means
  - Incomes
  - random numbers
  - blocks
- Thousands of sample means
- http://onlinestatbook.com/stat_sim/sampling_dist/index.html

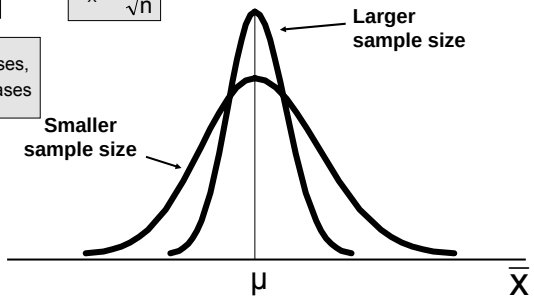
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Key points 1
Sampling Distribution Properties

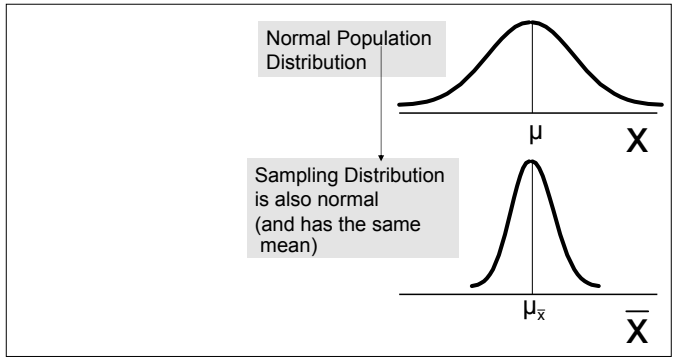
$\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

As n increases, $\sigma_{\bar{x}}$ decreases

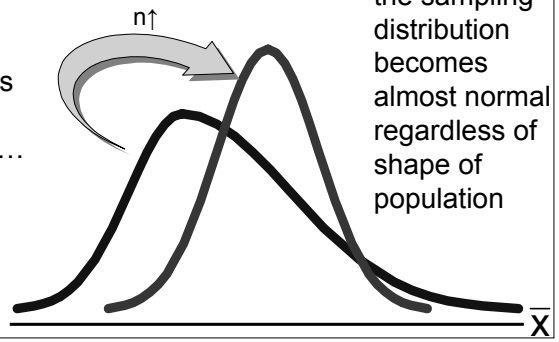


Key points 2
Sampling Distribution Properties



Key points 3
Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

Confidence Intervals

- Using previously collected data
  - –Incomes
  - –random numbers
  - –Blocks

- Simulation

http://onlinestatbook.com/stat_sim/conf_interval/index.html
http://onlinestatbook.com/stat_sim/conf_interval/index.html
http://onlinestatbook.com/stat_sim/normal_approx_conf/index.html
http://onlinestatbook.com/stat_sim/normal_approx_conf/index.html

- Collecting Data - Confident in a Kiss

Interpreting confidence intervals

- Tell us something about the uncertainy associated with a point estimate
- We are …% certain that the … lies within the endpoints of the confidence interval.
  - The endpoints were calculated by a method that gives correct results in …% of all samples.
  - The true … is either in between the endpoints or not (there is no randomenss once a sample has been taken!)
- A range of plausible values

Assumptions

- Confidence interval for mean
  - –Random sample, randomised experiment
  - –Normally distributed or large sample
  - –For difference 2 means – samples must be independent
- Confidence interval for proportion
  - –Random sample, randomised experiment
  - –np>5 and n(1-p) >5
  - –For difference 2 proportions – samples must be independent
- What if these assumptions are not met?

What about other statistics?

- Think about the income data
- Is the median more appropriate?
- Can we construct a 'confidence interval' for a median?
  - Resampling is a possibility
- What if we took a smaller sample – assumption of normality not met?
  - Resampling is a possibility

Resampling

- In resampling, inference is based on repeated sampling within the same sample (the original data)
- Fewer assumptions
  - Independence
  - Random samples
- Especially useful with statistics with unknown distributions (eg median) and for small non-normal samples

The bootstrap Method

•We resample the sample data Suppose I have a sample (S_n) of size *n*

•Take '1000' sample **with replacement** from S_n

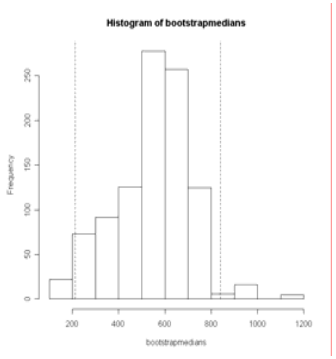
•For each sample calculate the sample statistic (mean, median, standard deviation...) t*

•The 1000 values of t* are the empirical bootstrap distribution

•For a 95% confidence interval find 2.5th and 97.5th percentiles

Example – The income Data

- Sample of 10 incomes
- Median
- For my sample:
  - Median = 569
  - 95% confidence interval for (bootstrap) median (211,839)



Confidence intervals for the difference of the means

- If normality assumptions not met can use randomisation or bootstrap resampling
- NOTE in both cases the interval we obtain is one assuming there is NO difference. So instead of looking to see if 0 is in the interval (it will be) we look to see if the our sample difference is in the interval (no difference) or outside it (a difference)

randomisation

- Can be used with non-random data
- BUT results may not be generalisable to the population

Example

- Researchers wish to know if they can conclude that two poulations of infants differ with respect to the mean age at which they walk alone. The following data (ages in months have been collected…
- Sample A 9.5, 10.5, 9.0, 9.75, 10.0, 13.0, 10.0, 13.5, 10.0, 9.5, 10.0, 9.75
- Sample B 12.5, 9.5, 13.5, 13.75, 12.5, 9.5, 12.0, 13.5, 12.0, 12.0

Confidence interval for the difference of the means

Problem – small samples and not normally distributed

Sample A	Sample B
9 | 55	9 | 05588
10 |	10 | 00005
11 |	11 |
12 | 00055	12 |
13 | 558	13 | 05

The decimal point is at the |

Use randomisation or boostrap resampling

- We begin by combining the two samples into one – i.e. we assume they come from the same (identical) population
- Randomisation
  - Randomly reallocate the data to the two samples
- Bootstrap resampling
  - Randomly choose new samples of correct sizes

Resources

Journal of Statistics Education

http://www.amstat.org/publications/jse/

More geared to university teaching – but 'free' on line

Mathematics in Schools - published by the Mathematical Association (UK)

Some articles are available on line at:
http://www.ma.org.uk/resources/periodicals/online_articles_keyword_index/index.html

Teaching Statistics - to view recent copies need to subscribe – but some articles can be viewed – along with contents of current journals at

http://www.rsscse.org.uk/ts/

mostly geared to teaching in schools