SO5041 Unit 9:

Brendan Halpin, Sociology

Autumn 2019/0

Outline

- In this lecture we will examine
  - Population parameters and sample estimates
  - The Central Limit Theorem: this is why the normal distribution is important
  - How to deal with the imprecision of sample estimates: confidence intervals
  - Reading: Agresti and Finlay, Chapter 4

Sampling error and sampling distributions

- How do we reason about sampling and sampling error?
- Sampling Error = true value – sample value; but true value unknown!
- We can reason about the likely distribution of sampling error
- If the true value is μ, how far is a sample value likely to fall from it?
- $\longrightarrow$ a *distribution* of possible sample values

Referendum sample

- Example: referendum with yes/no answer
- The probability a randomly selected voter goes yes vs no is unknown
- That is equivalent to saying the yes/no proportion is unknown
- What if we want to know overall proportion, and poll 1500 random voters?
- If we get a result of 54/46: how do we know we have any accuracy?

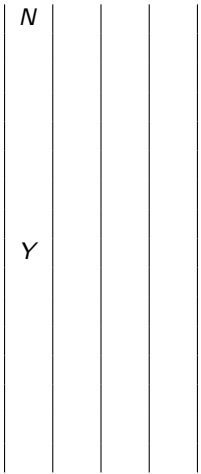
Computer simulation

- Agresti and Finlay do a computer simulation to test
- Computer selects 1500 random numbers: $0 - 0.5 \Rightarrow$ yes $0.5 - 1 \Rightarrow$ no
- First run gets 51.6% yes, next 49.1%
- They run it a million times, and make a histogram of the results: vast majority between 0.46 and 0.54, ie +- 4%

Simulation by hand: $N = 4$

- We can alternatively work through a "toy" example
  - Yes and No equally likely
  - Sample size of 4
  - 16 different possible combinations, each equally likely
- Result is a "binomial" distribution
- Note number of possible combinations is 2^N so this calculation is practical only with tiny samples

Simulating with N=4



Simulating with N=4

N	*N*			
N	*Y*			
Y	*N*			
Y	*Y*			

Simulating with N=4

N	*N*	*N*		
N	*N*	*Y*		
N	*Y*	*N*		
N	*Y*	*Y*		
Y	*N*	*N*		
Y	*N*	*Y*		
Y	*Y*	*N*		
Y	*Y*	*Y*		

Simulating with N=4

N	*N*	*N*	*N*
N	*N*	*N*	*Y*
N	*N*	*Y*	*N*
N	*N*	*Y*	*Y*
N	*Y*	*N*	*N*
N	*Y*	*N*	*Y*
N	*Y*	*Y*	*N*
N	*Y*	*Y*	*Y*
Y	*N*	*N*	*N*
Y	*N*	*N*	*Y*
Y	*N*	*Y*	*N*
Y	*N*	*Y*	*Y*
Y	*Y*	*N*	*N*
Y	*Y*	*N*	*Y*
Y	*Y*	*Y*	*N*
Y	*Y*	*Y*	*Y*

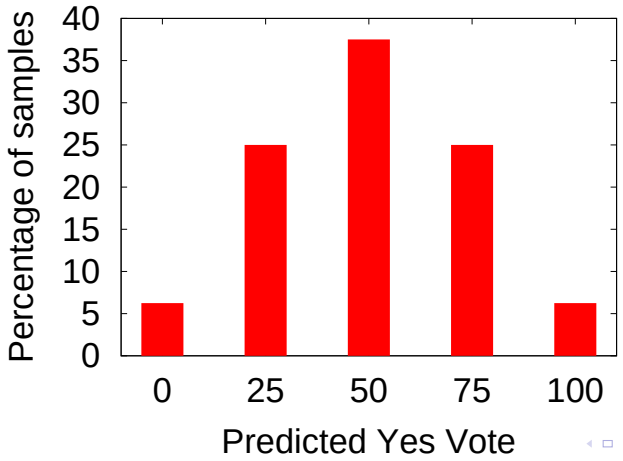
Simulating with N=4

N	*N*	*N*	*N*	0 : 4
N	*N*	*N*	*Y*	1 : 3
N	*N*	*Y*	*N*	1 : 3
N	*N*	*Y*	*Y*	2 : 2
N	*Y*	*N*	*N*	1 : 3
N	*Y*	*N*	*Y*	2 : 2
N	*Y*	*Y*	*N*	2 : 2
N	*Y*	*Y*	*Y*	3 : 1
Y	*N*	*N*	*N*	1 : 3
Y	*N*	*N*	*Y*	2 : 2
Y	*N*	*Y*	*N*	2 : 2
Y	*N*	*Y*	*Y*	3 : 1
Y	*Y*	*N*	*N*	2 : 2
Y	*Y*	*N*	*Y*	3 : 1
Y	*Y*	*Y*	*N*	3 : 1
Y	*Y*	*Y*	*Y*	4 : 0

Simulating with N=4

N	*N*	*N*	*N*	0 : 4
N	*N*	*N*	*Y*	1 : 3
N	*N*	*Y*	*N*	1 : 3
N	*Y*	*N*	*N*	1 : 3
Y	*N*	*N*	*N*	1 : 3
N	*N*	*Y*	*Y*	2 : 2
N	*Y*	*N*	*Y*	2 : 2
N	*Y*	*Y*	*N*	2 : 2
Y	*N*	*N*	*Y*	2 : 2
Y	*N*	*Y*	*N*	2 : 2
Y	*Y*	*N*	*N*	2 : 2
N	*Y*	*Y*	*Y*	3 : 1
Y	*N*	*Y*	*Y*	3 : 1
Y	*Y*	*N*	*Y*	3 : 1
Y	*Y*	*Y*	*N*	3 : 1
Y	*Y*	*Y*	*Y*	4 : 0

$N = 4$



Sampling distribution for N=4

Online Apps:

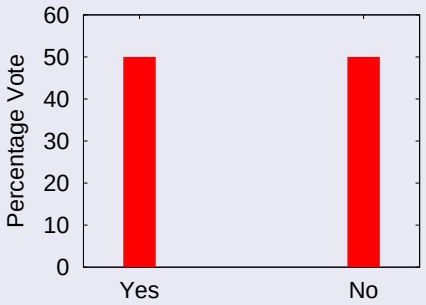
- Heads and Tails
- Simulating binomial sampling

Central Limit Theorem

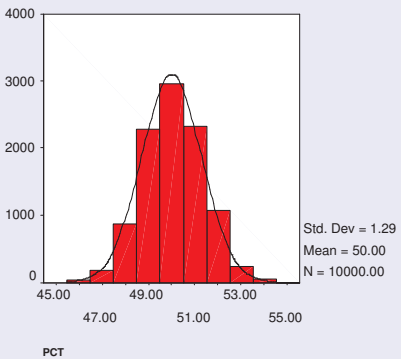
- The Central Limit Theorem: for a sufficiently large sample size, the *sampling distribution* of a statistic such as the sample mean will be approximately normal
- This is true, whatever the distribution of the population of interest
- We have seen this with the simulation of sampling vote where the true population proportions are 50:50

Binomial distribution

Population Distribution



Sampling Distribution



Sampling distribution

- This means that for sufficiently large samples, our sample statistic can be regarded as being drawn from a random distribution with mean μ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- This is a very important theorem because it allows us to reason about the precision of sample estimates
- Note: $\sigma_{\bar{X}}$, the standard deviation of the sample mean, is known as the *standard error*