GV205
Measuring Public Opinion
From Samples to Populations

05 February, 2018

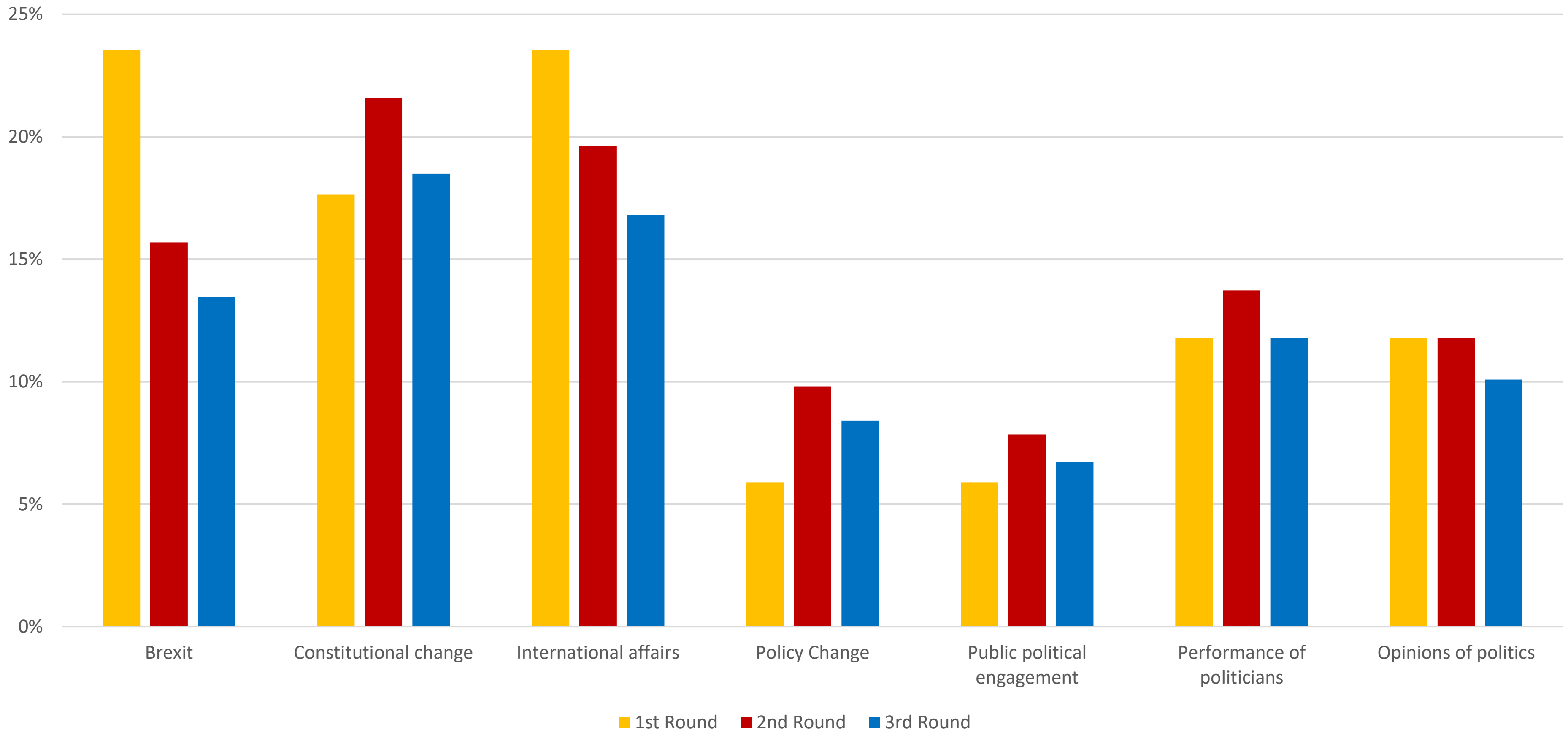
Joe Greenwood
YouGov
@NiceOneCombo

Overview

- The winning topic
- Populations and samples
- The importance of randomness
- The normal distribution
- The central limit theorem
- Confidence intervals
- Statistical significance
- Truly random samples?

- Examples:
  - Approval of Barack Obama
  - Ideology and Income
  - The ‘youthquake’
  - The British Election Study

The Winning Topic



Populations and Samples

- A population is a full group of phenomena that we might be interested in, and might be:
  - The countries of the world
  - The schools in a country
  - The individuals in a local authority
- A sample is a sub-set of phenomena from a population, and might be:
  - A selection of countries from each region of the world
  - A selection of schools from each county of a country
  - A selection of individuals from each postcode in a local authority

Populations: Hard to Reach

- The ideal might be to have data on the whole population, but gathering such data is:
  - Expensive
  - Time consuming
  - Practically challenging

- Samples help us overcome all of the above problems, and can tell us about the whole population.

Types of Sample

- Convenience
- Snowball
- Quota
- Stratified
- Random

The Importance of Randomness

- If a sample is truly random:
  - Every individual in the population has an equal chance of being selected
  - Characteristics that are more or less prevalent in the population will also be more or less prevalent in the sample
  - The sample will naturally approximate the population in terms of the distribution of any variable of interest
- We can be more confident in the estimates emerging from a random sample if it is larger, though there are diminishing returns
- Crucially though, just by chance, a random sample will include random error

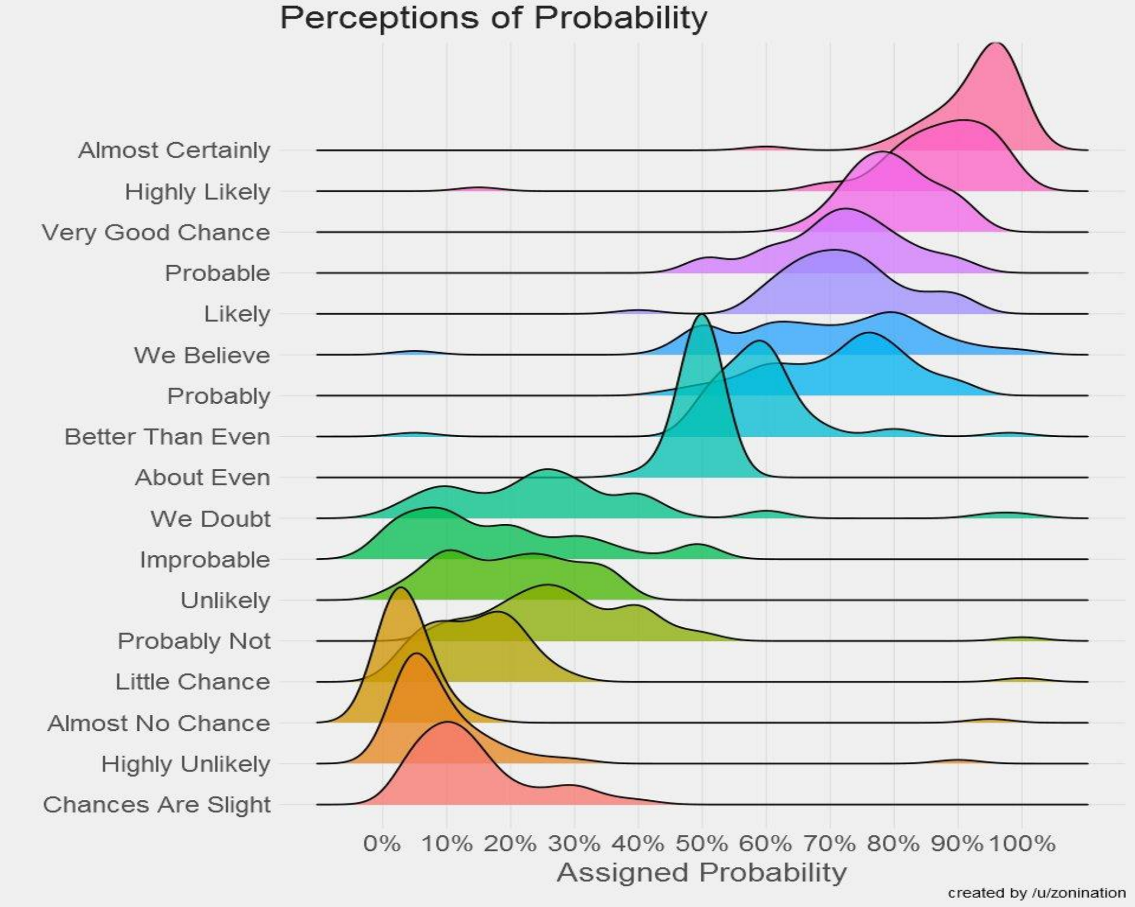
The Normal Distribution

- Has key characteristics:
  - Symmetrical around the mean
  - Equal mean, median, and mode
  - Predictable area under the curve

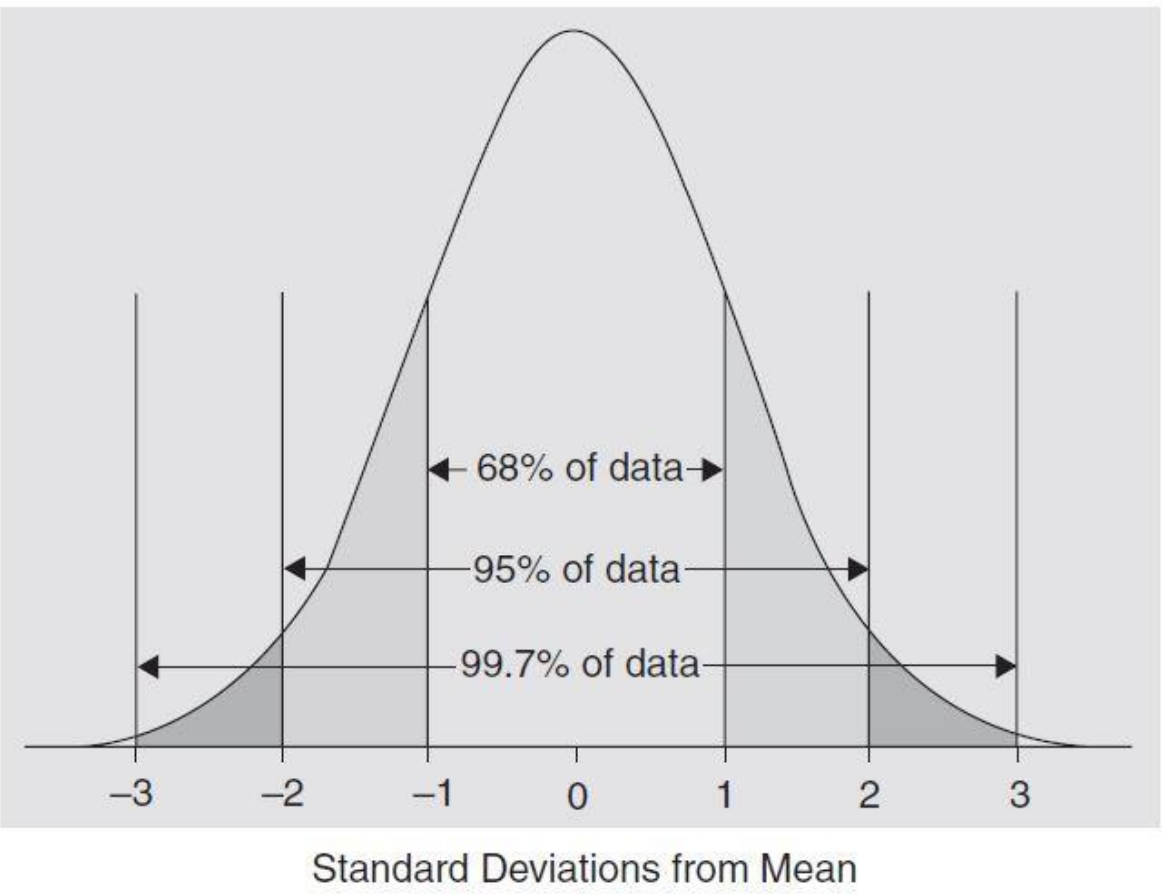
- Lots of variables are (approximately) normally distributed

- Lots of variables are not (approximately) normally distributed

(Non) Normally Distributed Variables



Proportions Under the Normal Curve



Standard Deviations from Mean

Sampling and the Normal Distribution

- The normal distribution is important because random samples produce estimates that are normally distributed around the ‘true’ value in the population
- This can be demonstrated visually in a rather neat fashion:
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
- The idea that estimates provided by random samples are normally distributed around the ‘true’ value in the population is called the central limit theorem

The Central Limit Theorem

- Under general conditions, the distribution of sample means is well approximated by a normal distribution when n is large
- The above is true when the variable is normally distributed in the population, and approximately true even when the variable is not normally distributed in the population
- This requires that the sample is sufficiently large, though what constitutes a large sample varies
- Crucially, the central limit theorem is what allows us to make inferences about the population from samples

Example: Approval of Barack Obama

- NBC and Wall Street Journal 2012 poll of 1,000 Americans:
    - ‘In general, do you approve or disapprove of the job Barack Obama is doing as president?’
- 47% approved of Barack Obama’s performance, 48% disapproved, and 5% were unsure
- To make inferences about the population, we need to calculate:
    - The sample mean
    - The sample standard deviation
    - The standard error of the sample distribution

Approval of Barack Obama: Sample Mean

- The sample mean:
  - $\bar{Y} = \frac{\sum Y_i}{n}$
  - $\bar{Y} = \frac{(470 \times 1) + (530 \times 0)}{1000} = 0.47$

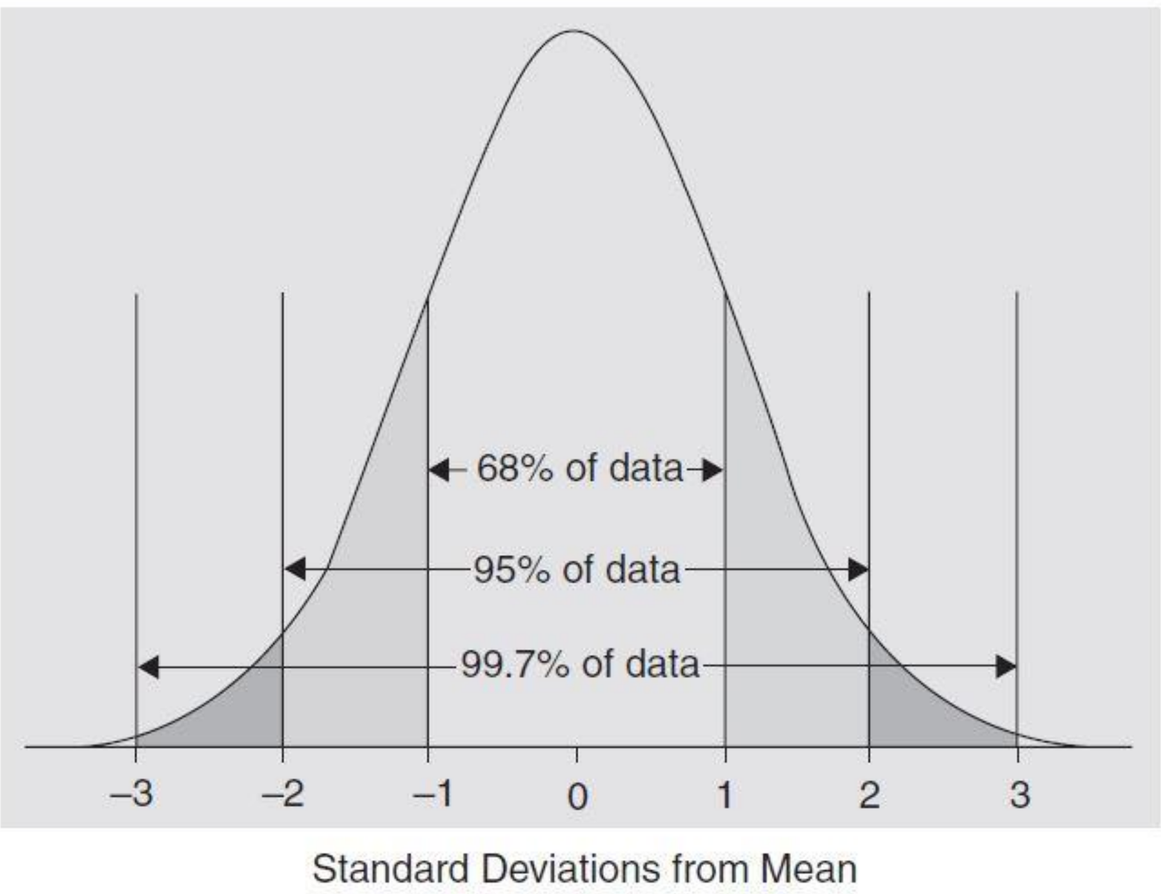
Approval of Barack Obama: Standard Deviation

- The sample standard deviation:
  - $s_Y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n - 1}}$
  - $s_Y = \sqrt{\frac{470(1 - 0.47)^2 + 530(0 - 0.47)^2}{1000 - 1}}$
  - $s_Y = \sqrt{\frac{249.1}{999}} = 0.50$

Approval of Barack Obama: Standard Error

- The standard error of the mean:
  - $\alpha_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$
  - $\alpha_{\bar{Y}} = \frac{0.50}{\sqrt{1000}} = 0.016$

Approval of Barack Obama: The Normal Curve



Approval of Barack Obama: The Normal Curve

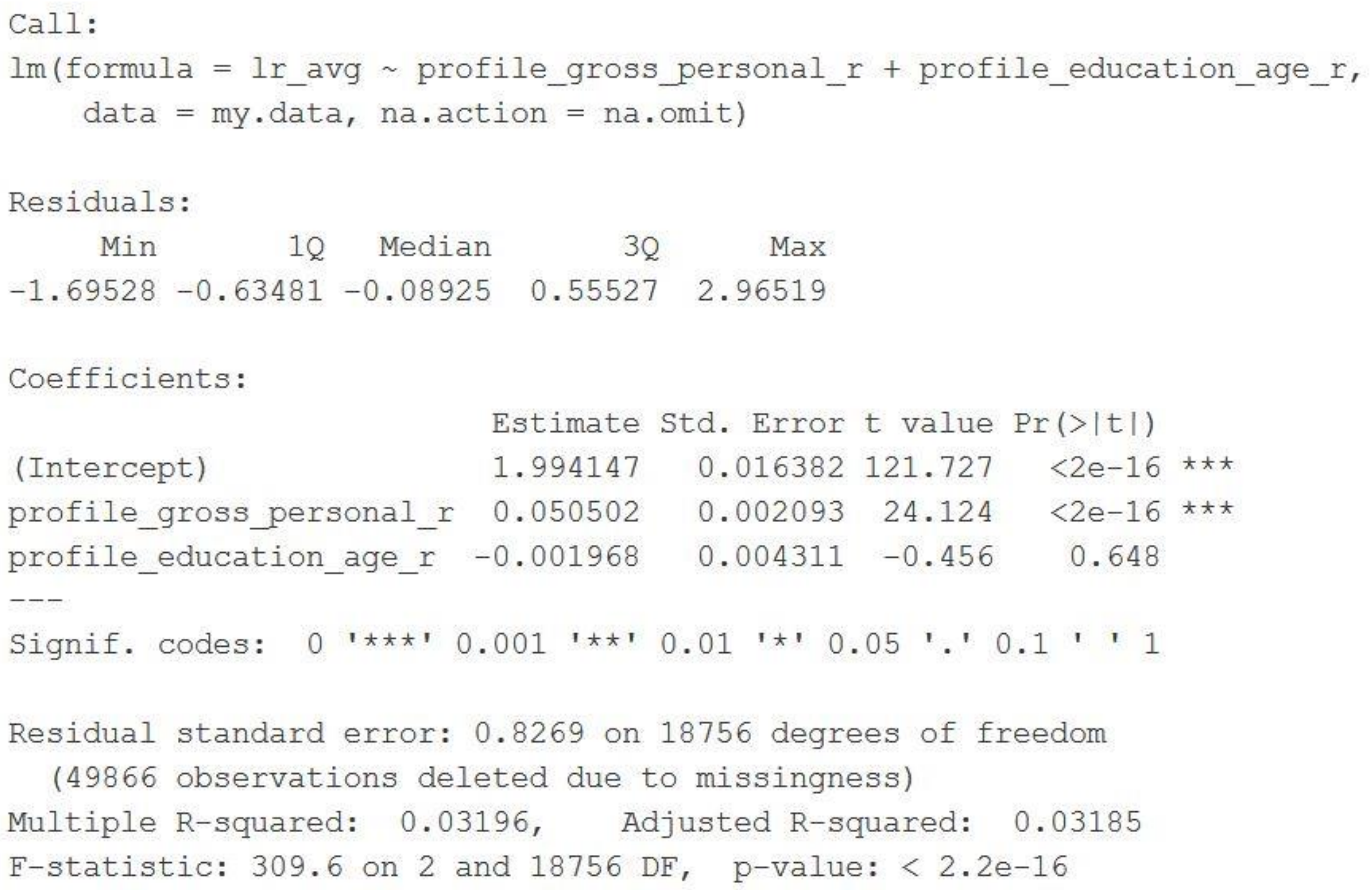
- Our sample mean is 0.47 and our standard error of the mean is 0.016
- We know that, with a random sample, the mean is normally distributed around the 'true' value in the population
- We know that, 95% of the time, the mean in a random sample will be within 2 (or 1.96) standard errors of the 'true' mean in the population
- So, we can be 95% sure that the mean in the population is ± 2 x 0.016 (i.e. ± 0.032) from our sample mean of 0.47
- In other words, we are 95% sure that, with a random sample mean of 0.47, the population mean is between 0.438 and 0.502

Approval of Barack Obama: Confidence Intervals

- Everything on the previous slide was leading us to this conclusion:
  - Based on the mean from the random sample surveyed in 2012, we can be 95% sure that Barack Obama's approval rating was between 43.8% and 50.2%
  - In other words, the 95% confidence intervals around our observed mean 47% approval in the sample are 43.8% (lower bound) and 50.2% (upper bound)
  - We can also call this the margin of error (in this case, $\pm$ 3.2%)

- The above is also important because we can apply the same logic to observed relationships in our samples as to observed means

Example: Ideology and Income



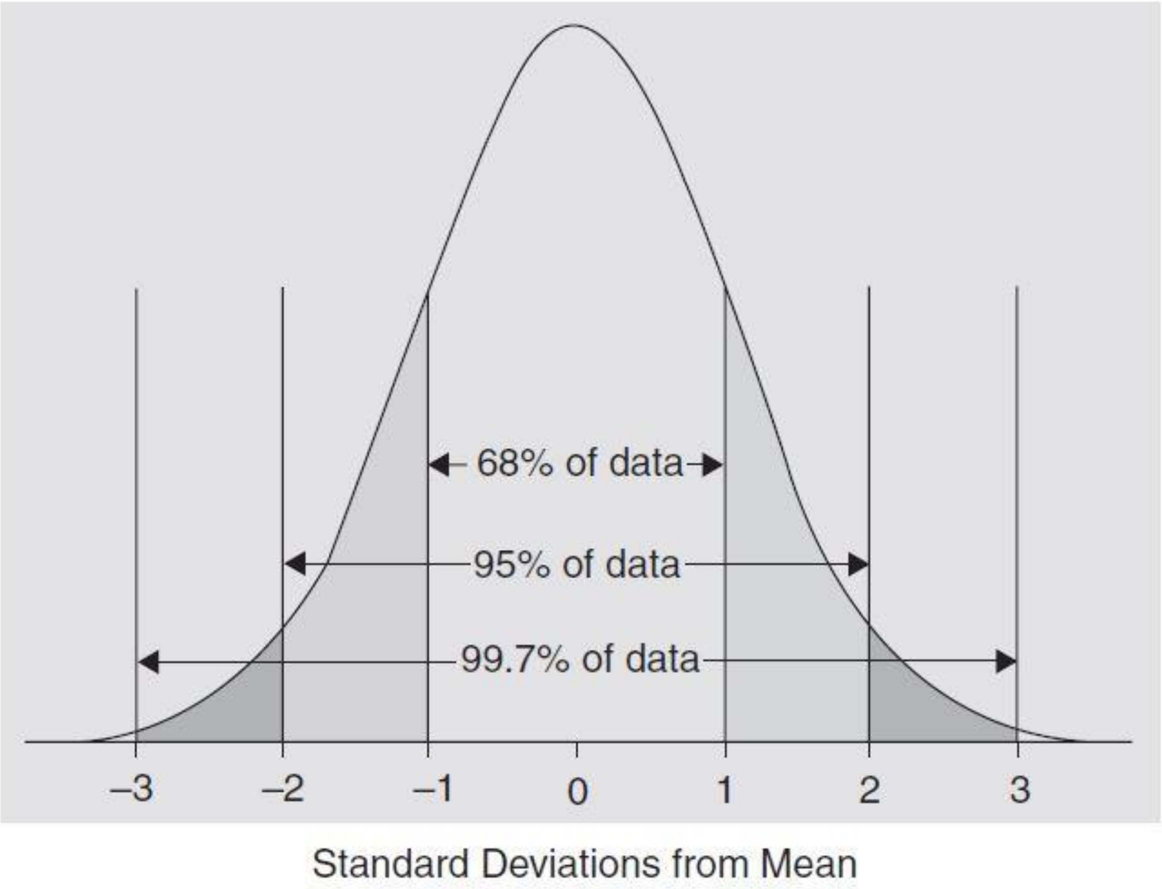
```
Call:
lm(formula = lr_avg ~ profile_gross_personal_r + profile_education_age_r,
    data = my.data, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.69528 -0.63481 -0.08925  0.55527  2.96519 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               1.994147   0.016382 121.727   <2e-16 ***
profile_gross_personal_r  0.050502   0.002093  24.124   <2e-16 ***
profile_education_age_r  -0.001968   0.004311  -0.456    0.648    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8269 on 18756 degrees of freedom
  (49866 observations deleted due to missingness)
Multiple R-squared:  0.03196,	Adjusted R-squared:  0.03185 
F-statistic: 309.6 on 2 and 18756 DF,  p-value: < 2.2e-16
```

Ideology and Income: The Normal Curve



Ideology and Income: Statistical Significance

- If the null hypothesis is correct, there is a 95% chance that we would observe a result in our sample that falls within 2 (or 1.96) standard errors of zero correlation
- 95% sure that, given the strength of the relationship observed in the sample (between ideology and income, controlling for education), such a relationship does not exist just by chance in the population
- This is what we mean when we say that a relationship observed in a sample is statistically significant
- All of this requires the sample to be random

Ideology and Income: Confidence Intervals

- In our sample, we observe a relationship between ideology and income (controlling for education) with an estimated coefficient of 0.05, and a standard error of 0.002
- We are, thus, 95% sure that there is a relationship with a coefficient between 0.046 (i.e. the lower bound is 0.05 – (2 x 0.002)) and 0.054 (i.e. the upper bound is 0.05 + (2 x 0.002)) in the population
- Since the upper and lower bounds (i.e. confidence intervals) do not cross zero we can call the result statistically significant
- Again, all of this requires the sample to be random

Approval of Barack Obama: A Random Sample?

‘If you read the preceding example carefully, you will have noted that the NBC-*Wall Street Journal* poll we described used a *random* sample of 1000 individuals. That means that they used some mechanism (like random-digit telephone dialing) to ensure that all members of the population had an equal probability of being selected for the survey.’

Kellstedt and Whitten, p. 140.

Ideology and Income: A Random Sample?

‘The face-to-face survey is an address-based random probability sample of eligible voters living in 468 wards in 234 Parliamentary Constituencies in England, Scotland, and Wales. 2,194 people completed the face-to-face survey. The fieldwork for the survey was conducted by GfK between June 26th 2017 and October 1st 2017 and achieved an overall response rate of 46.2%.’

British Election Study 2017 Face-to-face survey v1.0: Release note

The ‘Youthquake’: A Random Sample?

‘Opinion polls lent weight to the idea - one polling organisation suggested that turnout among 18 to 24-year-olds went up by as much as 16 percentage points, another suggested an increase of 12 points.’

‘Since 1964, the gold standard measure of electoral behaviour in Britain has been the British Election Study's face-to-face survey.

Newly released results using this data show that there was very little change in turnout by age group between the 2015 and 2017 elections.’

BBC News, ‘The myth of the 2017 “youthquake” election’

The BES: Sample and Population

Sex	Population (ONS)	Sample (BES)	Difference
Male	49.3	45.6	-3.7
Female	50.7	54.4	3.7

The BES: Sample and Population

Age Group	Population (ONS)	Sample (BES)	Difference
18-24	11.2	7.2	-4
25-49	42.2	36.3	-5.9
50-65	23.6	28.2	4.6
65+	22.9	28.3	5.4

The BES: Sample and Population

Education Level	Population (ONS)	Sample (BES)	Difference
High	31.0	41.4	10.4
Medium	39.4	31.3	-8.1
Low	29.6	27.3	-2.3

The BES: Sample and Population

Region	Population (ONS)	Sample (BES)	Difference
South	32.4	30.8	-1.6
London	13.4	9.7	-3.7
Midlands	16.5	17.5	1.0
North	24.1	27.5	3.4
Scotland	8.7	8.7	0.0
Wales	4.9	5.9	1.0

The BES: Sample and Population

Party	Population (BBC)	Sample (BES)	Difference
Conservative	29.8	32.0	2.2
Labour	28.2	34.6	6.5
Liberal Democrat	5.2	5.6	0.4
Green	1.1	1.5	0.4
UKIP	1.3	1.7	0.5
SNP	2.1	2.4	0.3
Plaid Cymru	0.4	1.1	0.8
Other	0.6	0.2	-0.5
Did not vote	31.3	20.9	-10.4

The BES: Sample and Population

Campaign	Population (BBC)	Sample (BES)	Difference
Remain	34.7	41.1	6.4
Leave	37.5	40.7	3.2
Did not vote	27.8	18.2	-9.6

The BES: Random and Quota Samples

- When we have a representative sample obtained by using quotas to ensure that it is a similar to the population as possible on key demographic and political variables, we are asserting that:
  - Our sample is close enough to a random sample that we can apply the same statistical theory and assumptions, including:
    - Estimates from such samples are approximately normally distributed around the 'true' values in the population
    - We can estimate confidence intervals and statistical significant in the same way as if the sample was random
- More on whether this is reasonable next week…