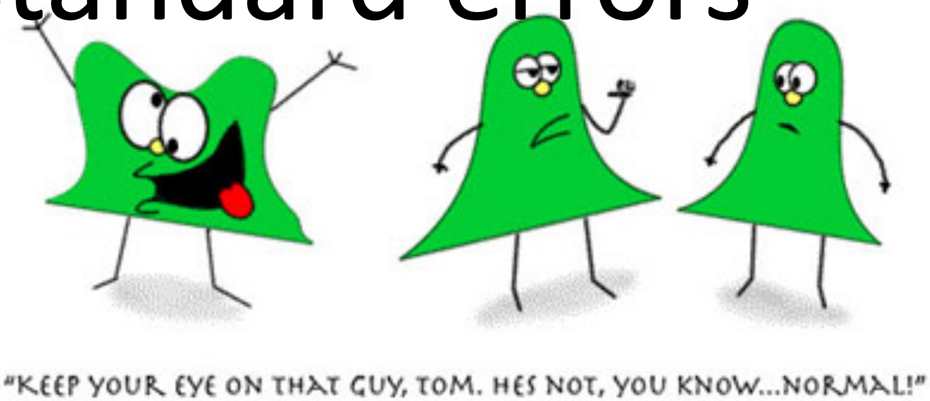
Research Methods for Political Science

MT week 3, lecture 2

Univariate statistics: Measures of dispersion and standard errors



TRINITY COLLEGE DUBLIN
COLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH
THE UNIVERSITY OF DUBLIN

Dr. Thomas Chadefaux

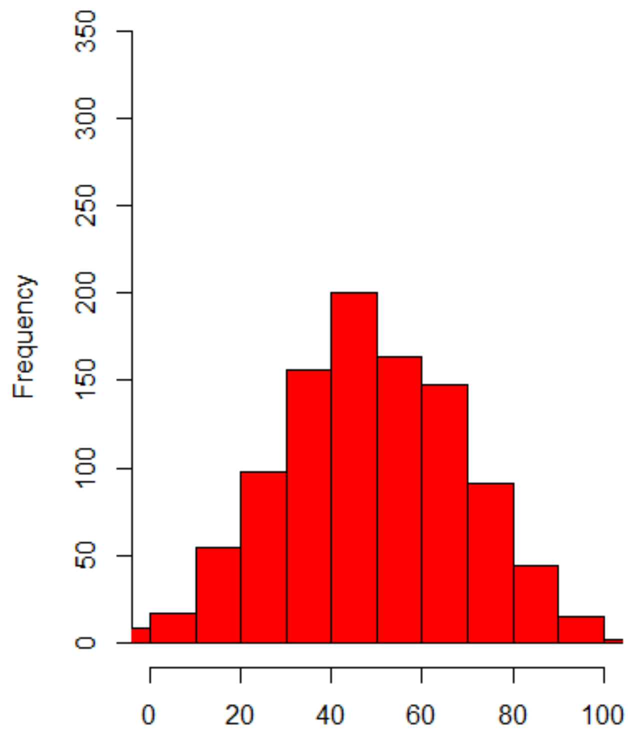
Assistant Professor in Political Science

Thomas.chadefaux@tcd.ie

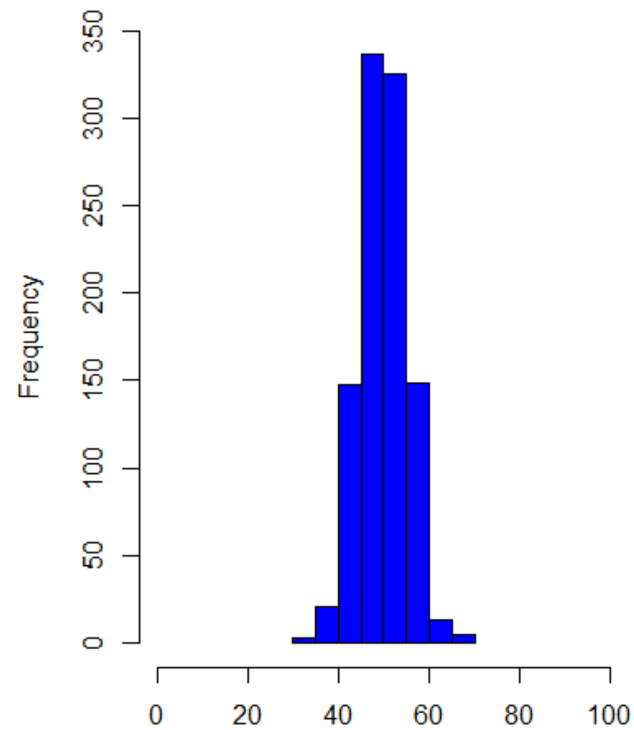
Measuring dispersion

Level	Measure of central tendency	Measure of dispersion
Nominal	Mode	
Ordinal	Median	Range Interquartile range (IQR)
Interval-ratio	Mean	Range, IQR, Variance / Standard deviation

Variance & Standard deviation



M = 50
SD = 20

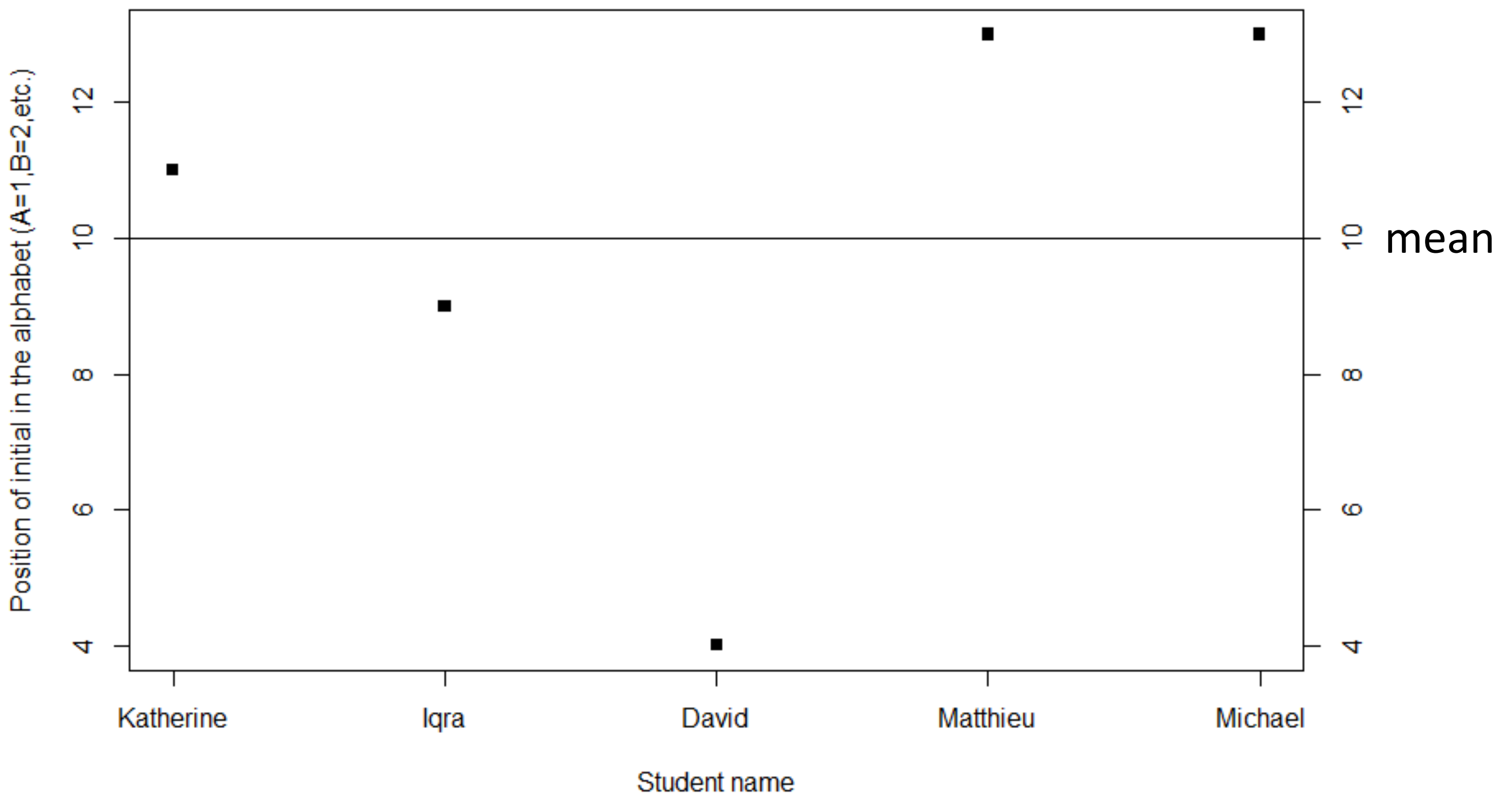


M = 50
SD = 5

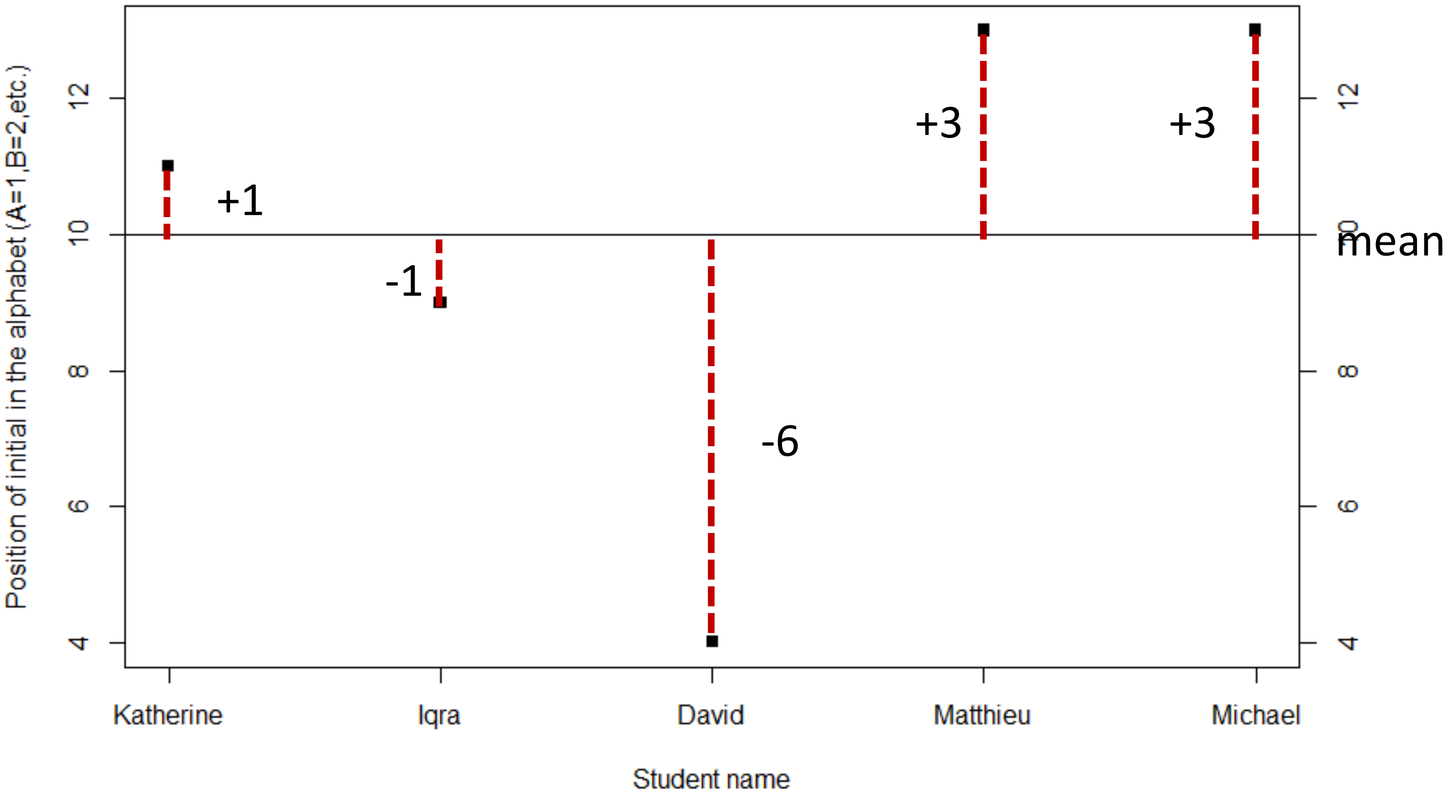
Variance & Standard deviation

- A way to measure the dispersion of a distribution

Example: distribution of initials

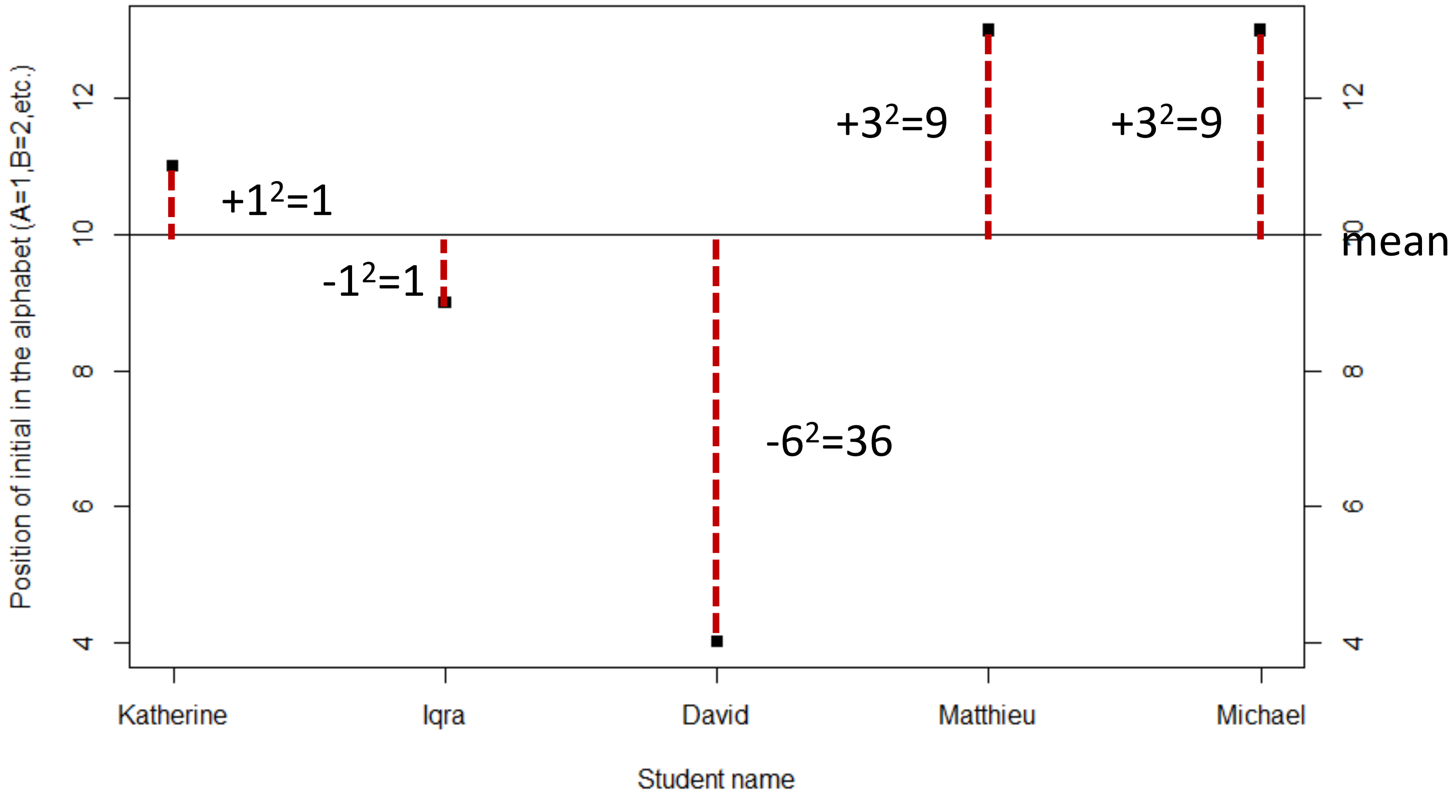


Example: distribution of initials



Sum of deviances = 1 + -1 + -6 + 3 + 3 = 0

Squared errors

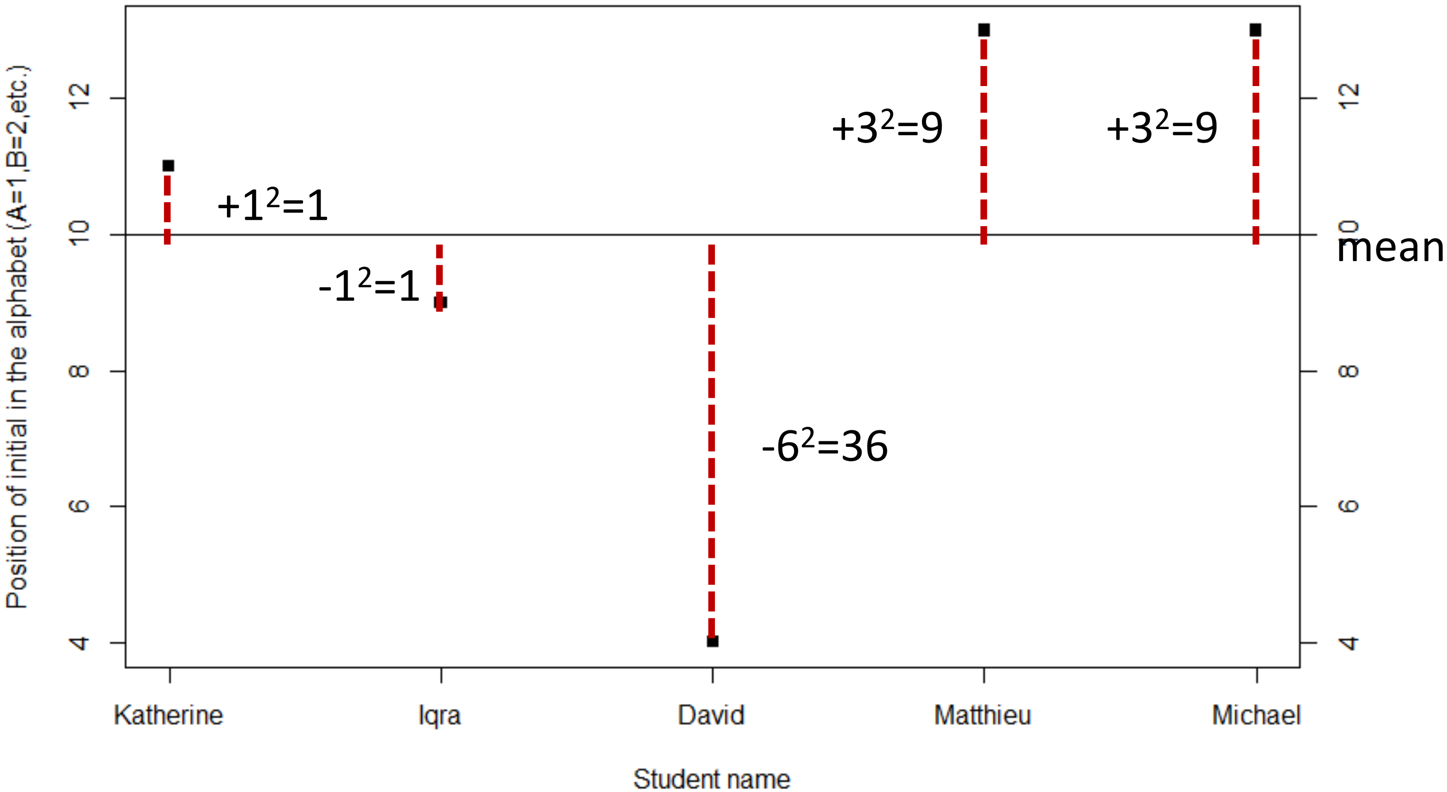


Sum of squared errors = 1 + 1 + 36 + 9 + 9 = 56

Variance

$$Variance = s^2 = SS/(N-1) = \frac{\sum_i (x_i - \bar{x})^2}{N-1}$$

Variance



Sum of squared errors = 1 + 1 + 36 + 9 + 9 = 56

Variance = SS / (N – 1) = 56 / (5 – 1) = 56 / 4 = 14

Variance

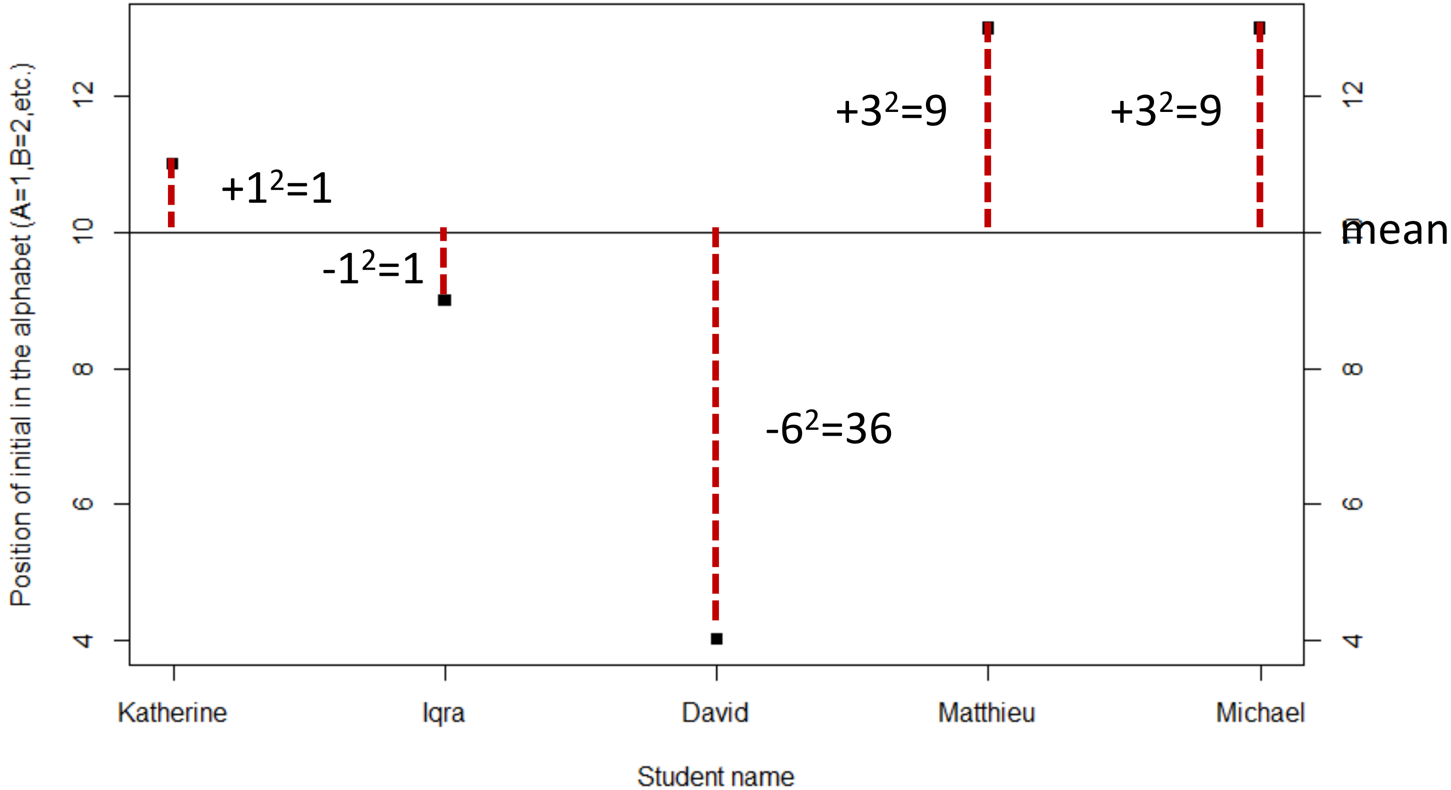
- Downside of variance is that it is expressed in terms of the **squared** distance

- Standard deviation transforms the variance "back" into the original units

$$\text{standard deviation} = s = \sqrt{s^2}$$

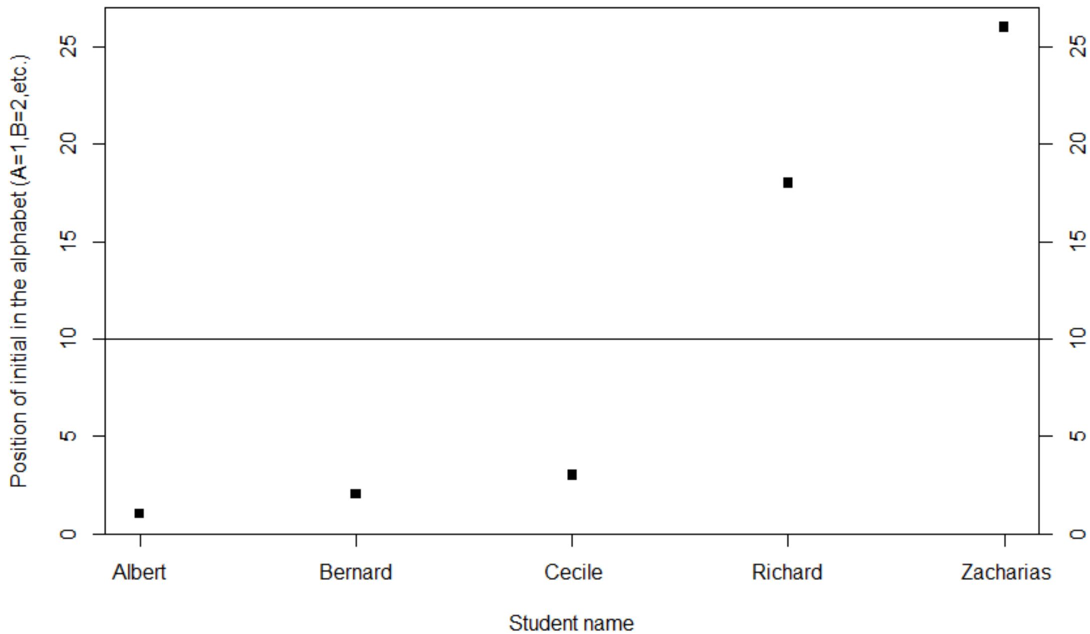
- In case above, SD was $\sqrt{14} \approx 3.74$

Example: standard deviation



Standard deviation = $\sqrt{variance}$ = $\sqrt{14}$ =3.74

Same mean, different SD



Standard deviation = $\sqrt{variance}$ = $\sqrt{128.5}$ =11.33

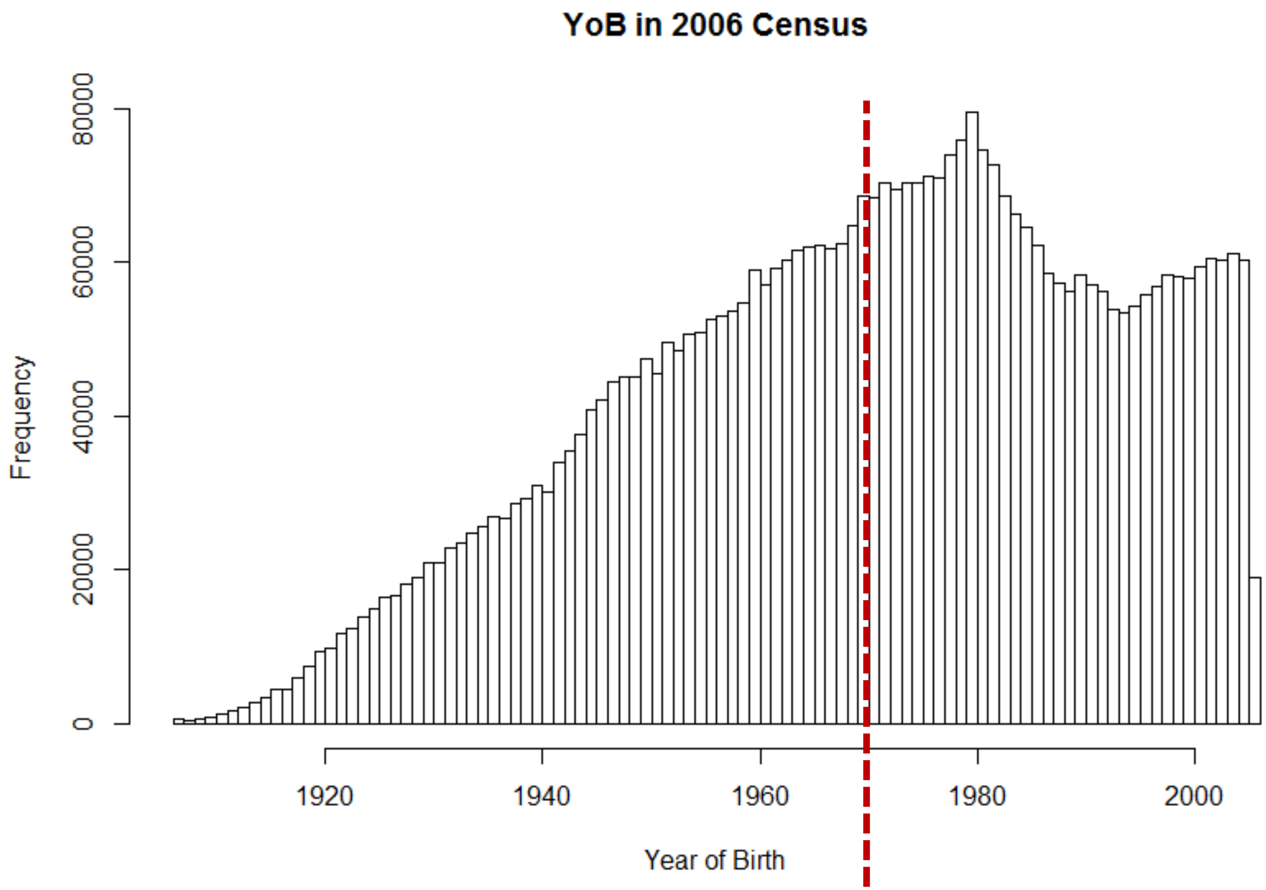
From sample to population

- While you can use the mean and standard deviation to describe the sample, they are even more useful to learn about the population
- Note: our aim is generally not to make claims about 1000 survey respondents or other randomly selected cases, but about the <u>population</u>
- In particular, we'd like to say how confident we are in our estimate of the mean. Surely if we took another sample, we'd get another mean, and so on.

Sampling distribution

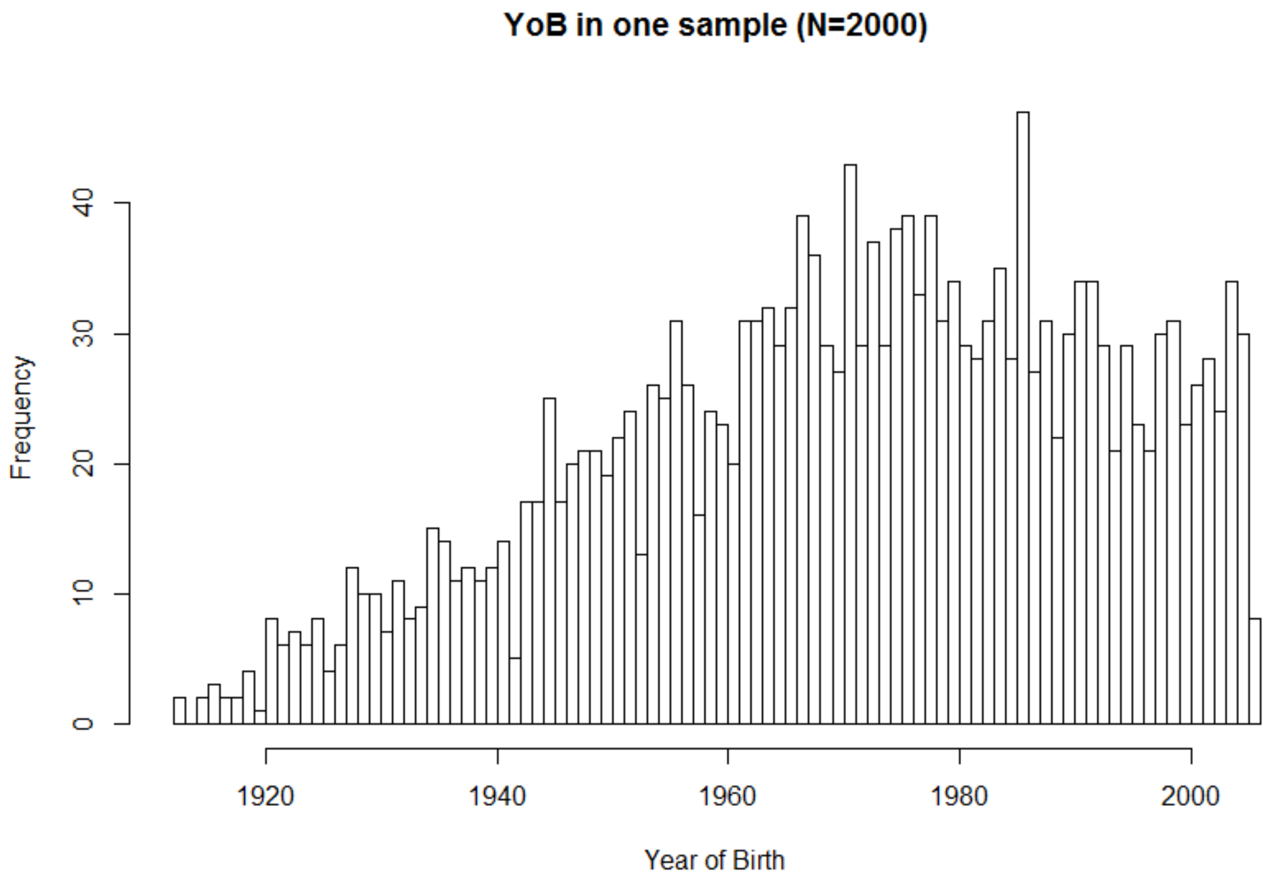
- The distribution of a test statistic (e.g. the mean) if we take many samples from a population
- **Careful: this is almost always confused with the sample distribution. Here we are talking about the distribution of *many sample means*, not the distribution of one sample**
- **I can (almost) guarantee that this will appear on the exam!**

Example: Year of Birth (population distribution)



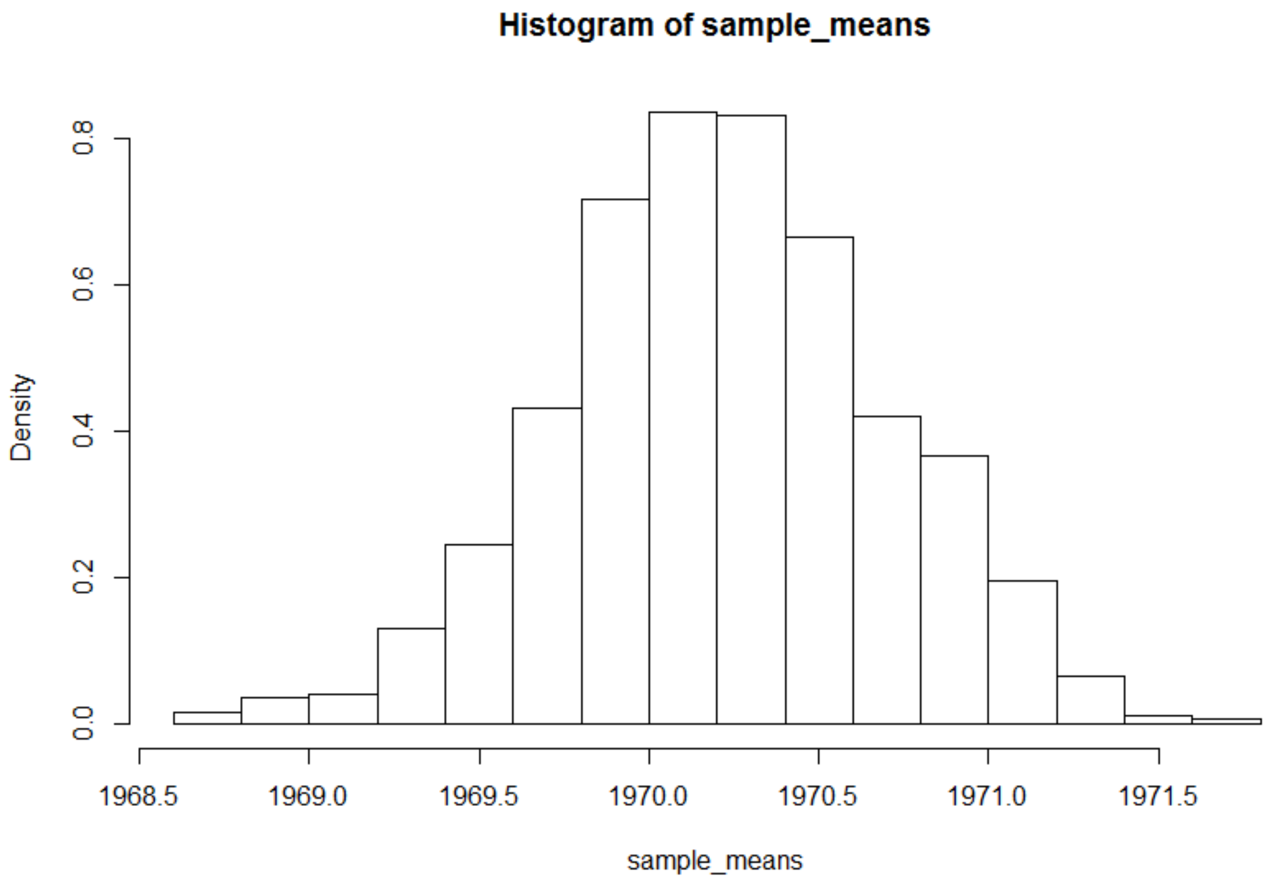
Mean = 1970.2

One sample:

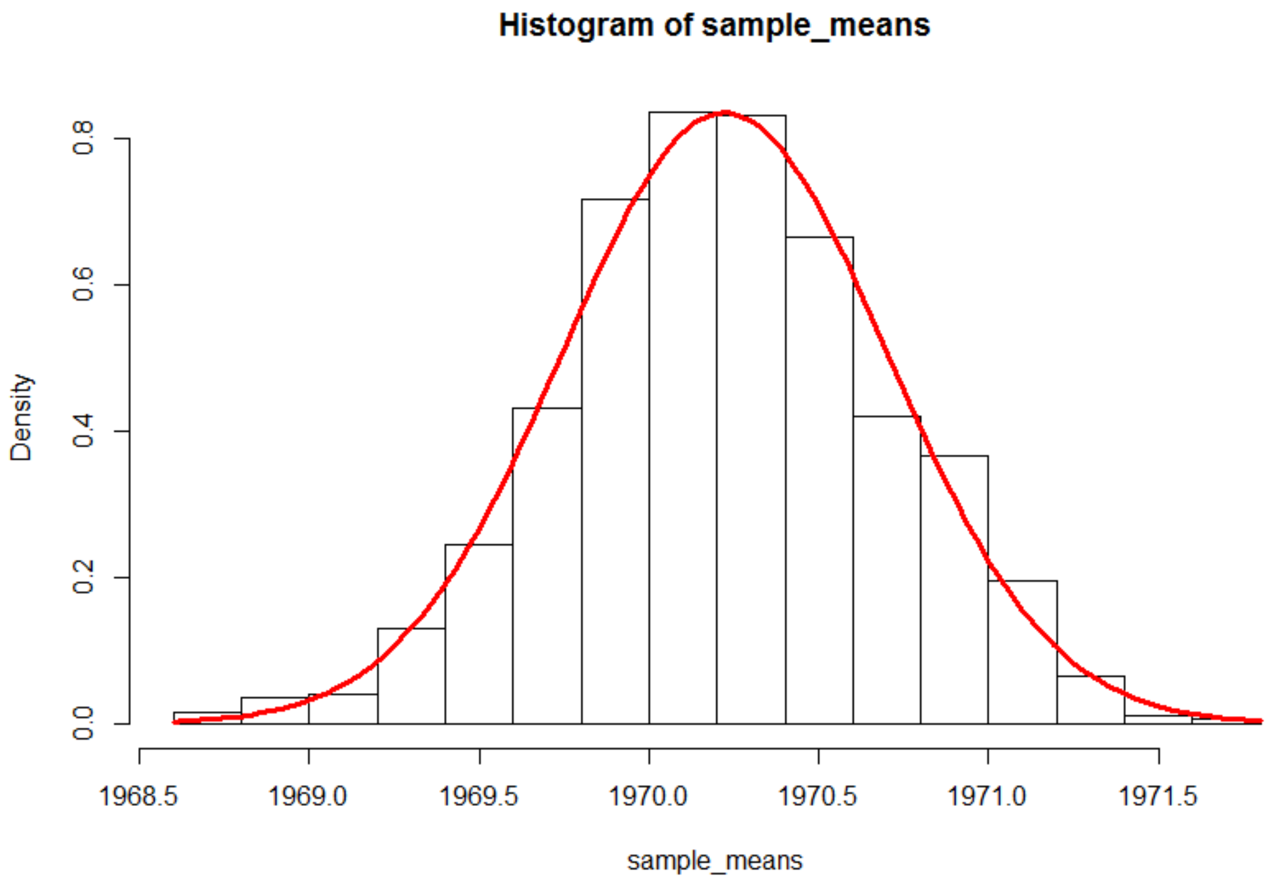


Mean = 1970.9

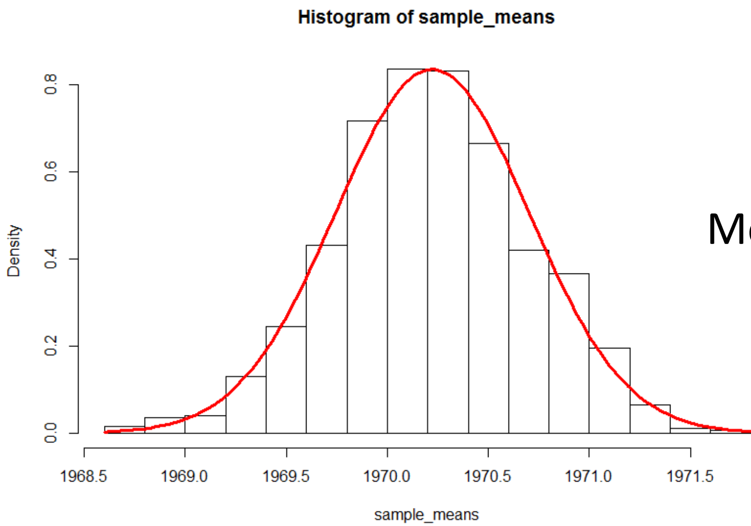
Many samples (means)



Many samples (means)



Many samples (means)



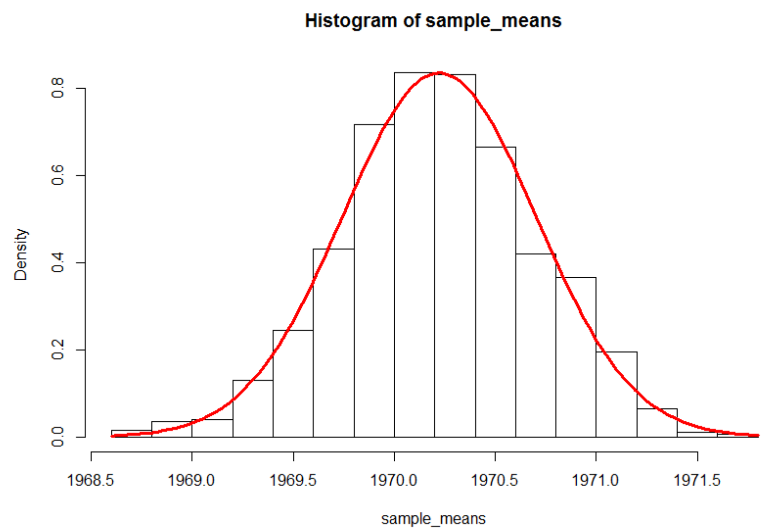
<u>Sampling distribution of the sample mean</u>

Mean of the sampling distribution = Population mean

$$E(\bar{X}) = \mu$$

Usually Greek letters are used for population characteristics
μ = mu = population mean

Many samples (means)



Sigma = population standard deviation

Standard deviation of the sampling distribution =
Standard deviation of the population / square root of the sample size

$$\mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

= standard error

Calculating E($\bar{X}$) and sd($\bar{X}$)

- Usually we don't know the *population* mean and standard deviation, so we estimate them using our *sample*:

This is the standard deviation of the *sample*

$$Standard\ error\ (SE) = \frac{s}{\sqrt{n}}$$

This is just another way to write standard deviation of the sampling distribution of the sample mean

If n gets large, the standard error gets small

$$standard\ error\ (SE) = \frac{s}{\sqrt{n}}$$

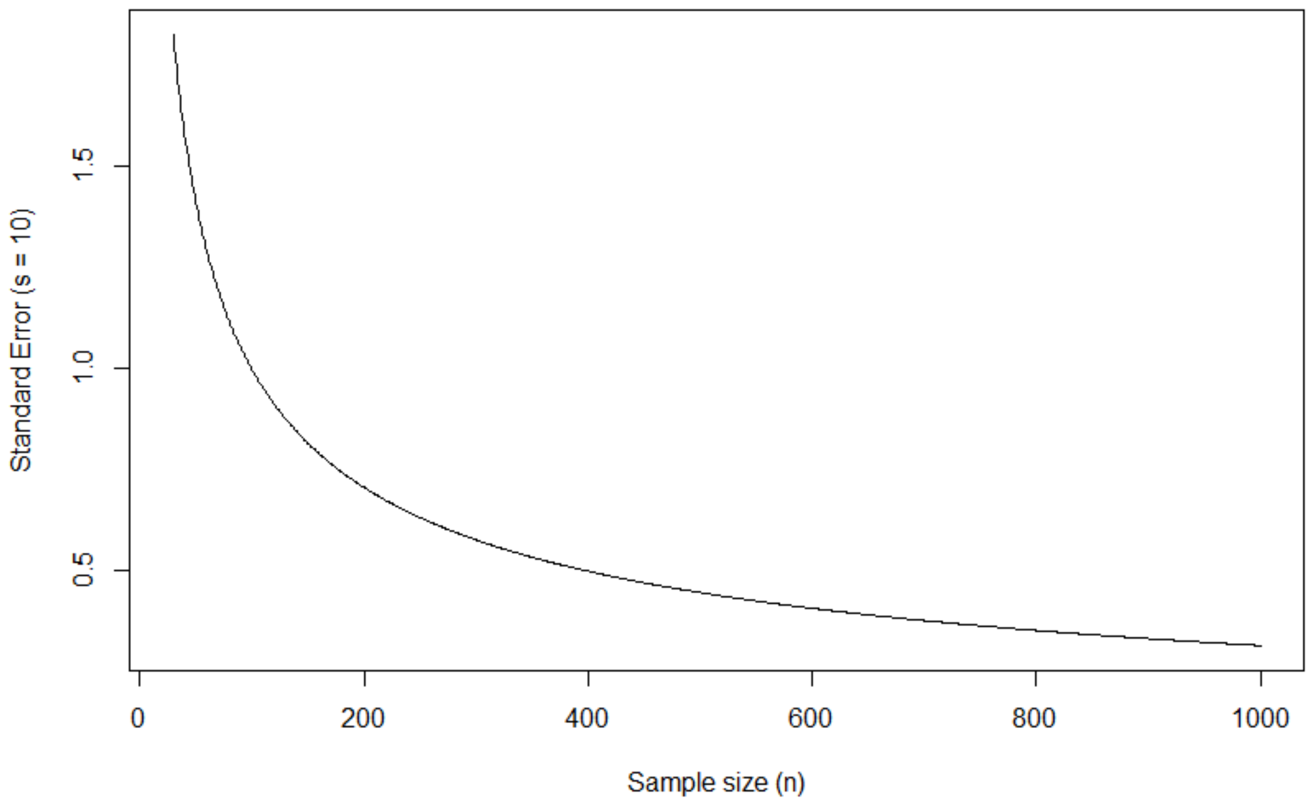
If s = 10 and n = 50, then

$$\mathrm{SE} = \frac{10}{\sqrt{50}} = 1.41$$

If s = 10 and n = 500, then

$$\mathrm{SE} = \frac{10}{\sqrt{500}} = 0.44$$

As n (sample size) increases, SE decreases



Standard error vs. standard deviation

- The standard ERROR is the extent to which your estimate (e.g., of the mean) varies
- The standard ERROR of the mean is the standard DEVIATION of the sample means
- i.e., take a sample, calculate its mean, store it. Repeat this 1000 times, and calculate the SD of all these 1000 means. That is your standard error of the mean

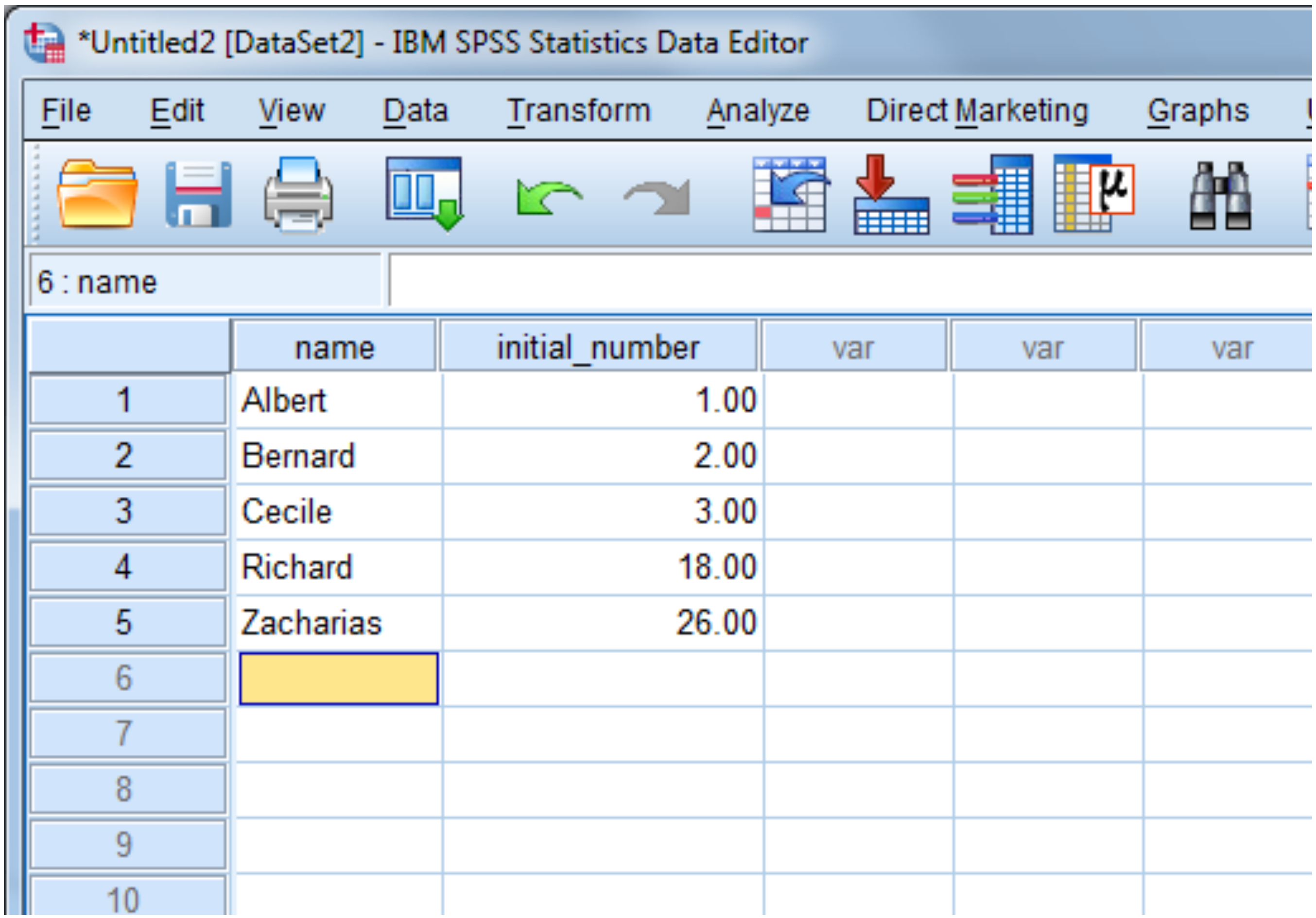
An example

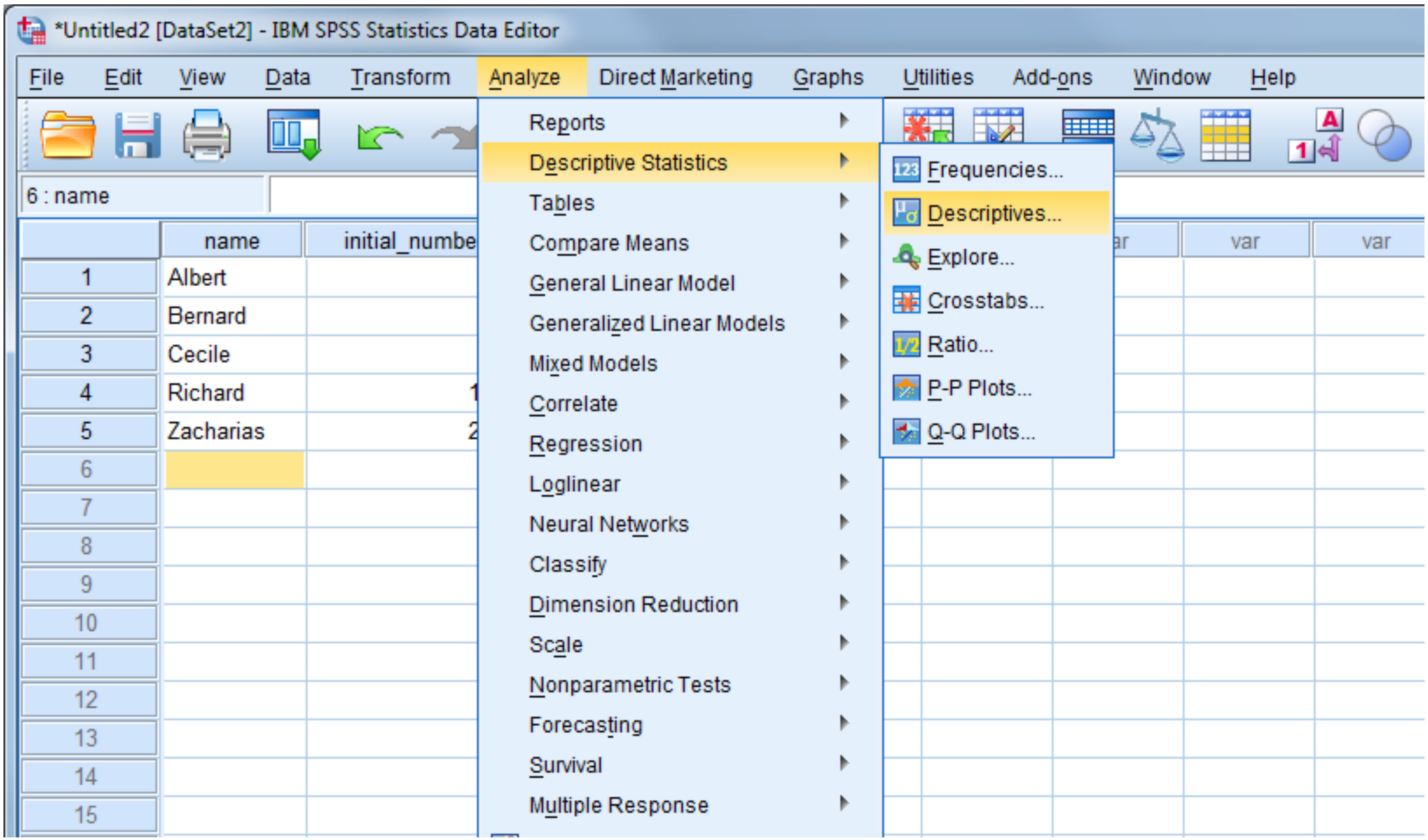
http://onlinestatbook.com/stat_sim/sampling_dist/

Why do we care?

- Calculate confidence intervals
- E.g., confidence interval of the mean:
  - mean+2sd
  - But what is the sd of the mean? It is sd/sqrt(N)=se

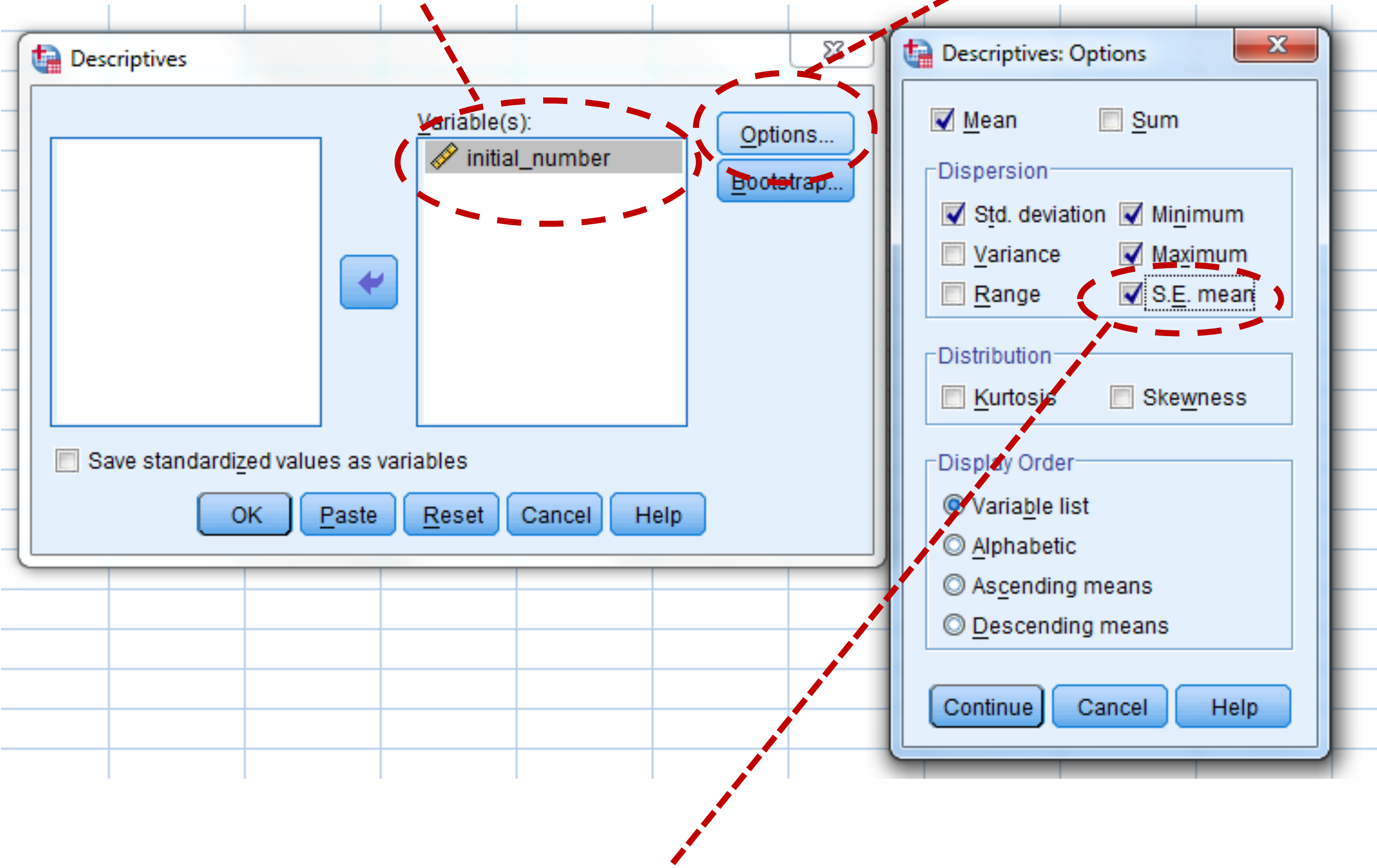
How to calculate the standard error in SPSS




*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
6 : name
name
initial_numbe
var
var
Albert
Bernard
Cecile
Richard
Zacharias
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...

2. Click ‘ Options’

1. Select the variable you wish to analyze



3. Select ‘S.E. mean’

Then click ‘Continue’ and ‘Paste’ / ‘OK’

Descriptives

[DataSet2]

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
initial_number	5	1.00	26.00	10.0000	5.06952	11.33578
Valid N (listwise)	5					

Check that Std. Error of the Mean =

Std. Deviation / Square Root of N =

11.34 / sqrt(5) = 5.07

Central Limit Theorem (important!)

- When samples are large (N > 30)
- The **sampling distribution** will take the form of a normal distribution
- Regardless of the shape of the distribution the sample was drawn from

Things to remember

- Standard deviation: spread of the sample
- Standard error of the mean: spread of the means of many samples. I.e., standard deviation of the sampling distribution
- Central limit theorem: The mean of a large number of random samples will be normally distributed, regardless of the underlying distribution of that variable