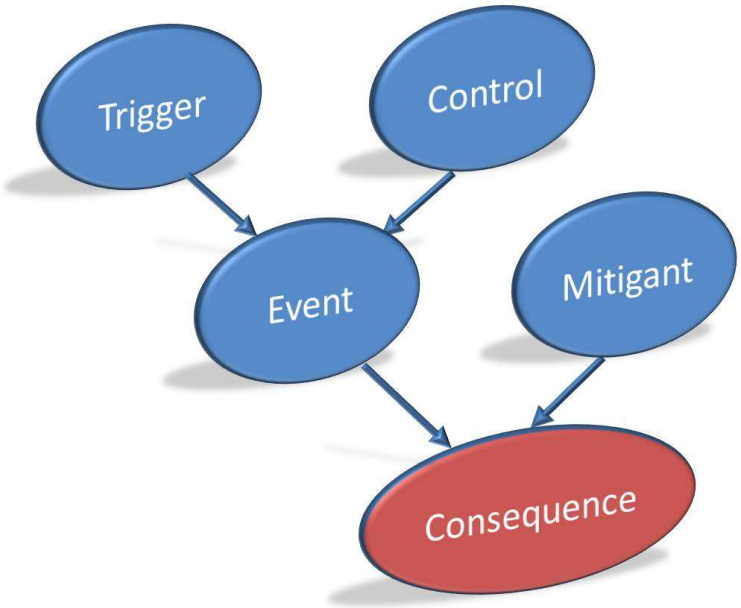


Risk and Decision Making for Data Science and AI

Lesson 3 Classical Statistics for risk assessment

Norman Fenton

@ProfNFenton

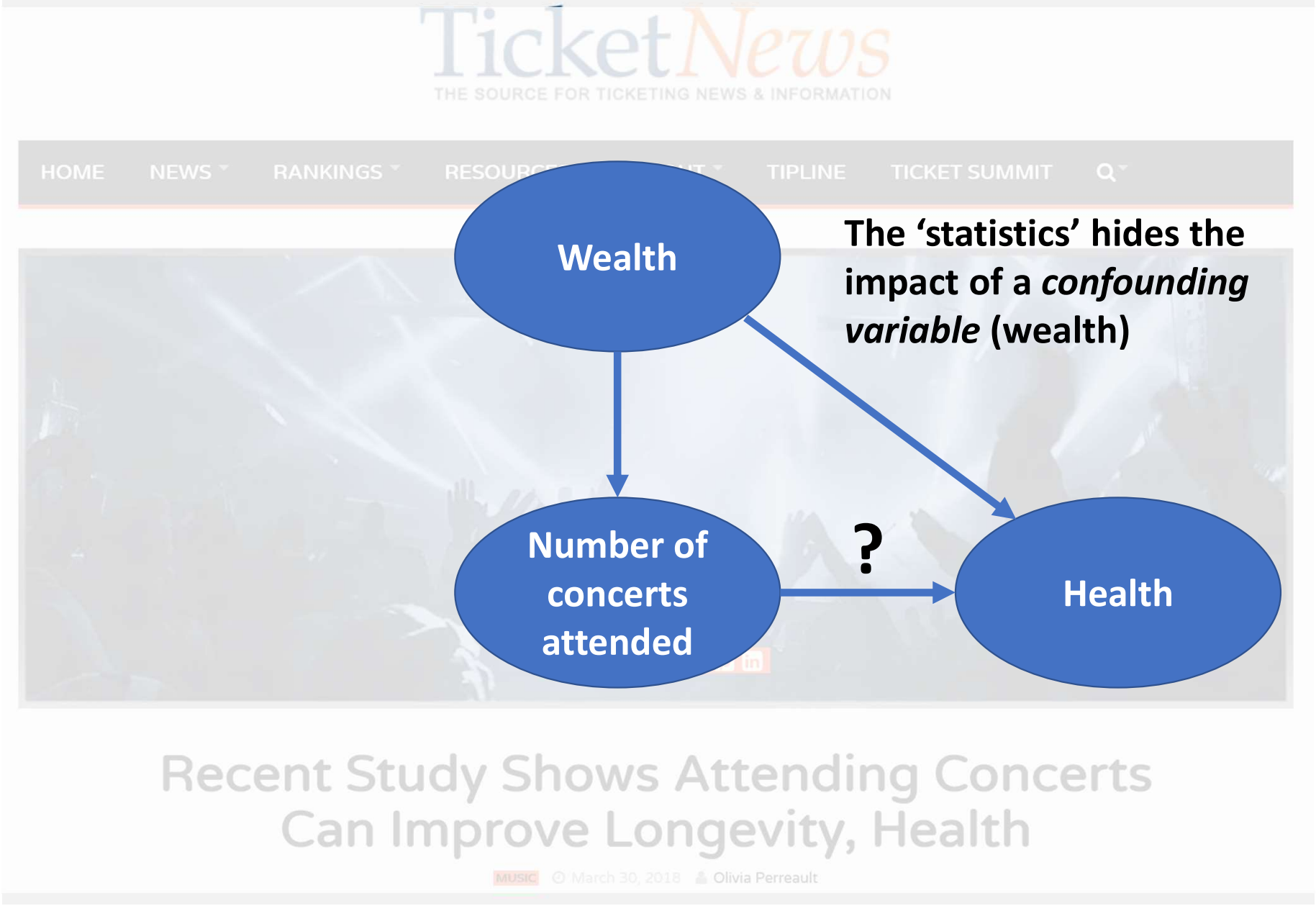
TicketNews

THE SOURCE FOR TICKETING NEWS & INFORMATION

HOME NEWS RANKINGS RESOURCES ABOUT TIPLINE TICKET SUMMIT

Recent Study Shows Attending Concerts Can Improve Longevity, Health

MUSIC March 30, 2018 Olivia Perreault


TicketNews
THE SOURCE FOR TICKETING NEWS & INFORMATION
HOME
NEWS
RANKINGS
TIPLINE
TICKET SUMMIT
Wealth
The 'statistics' hides the impact of a *confounding variable* (wealth)
Number of concerts attended
?
Health
Recent Study Shows Attending Concerts Can Improve Longevity, Health
MUSIC
March 30, 2018
Olivia Perreault

Nine hours of sleep and long naps raise your stroke risk, says study

By **Stephen Beech**

IF you sleep more than nine hours a night or grab a lengthy afternoon nap your risk of suffering a stroke rises by up to a quarter, a study has warned.

Researchers found people who slept for long periods were 23 per cent more likely to have a stroke than people who slept less than eight hours per night.

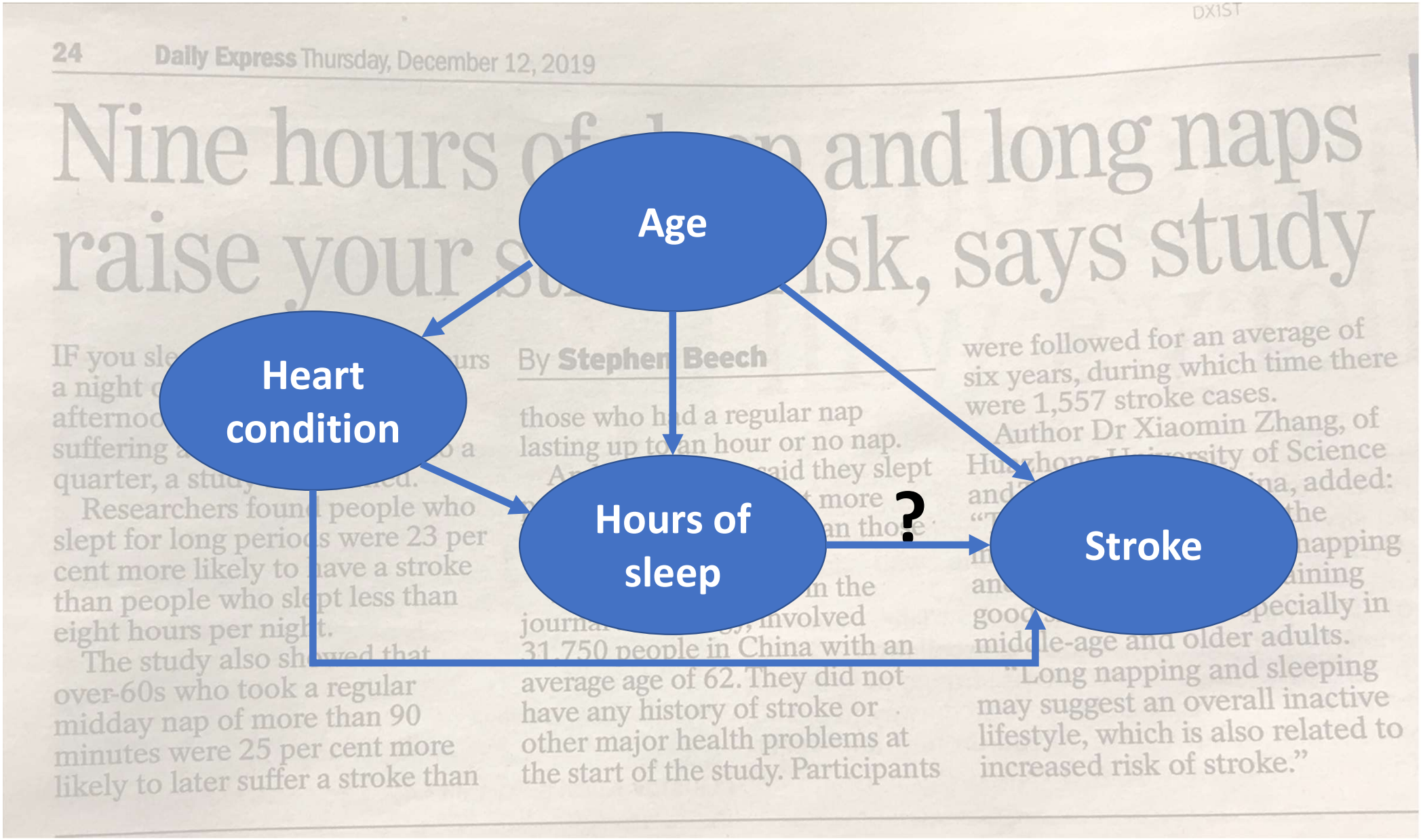
The study also showed that over-60s who took a regular midday nap of more than 90 minutes were 25 per cent more likely to later suffer a stroke than those who had a regular nap lasting up to an hour or no nap.

And people who said they slept poorly were 29 per cent more likely to have a stroke than those who had good sleep.

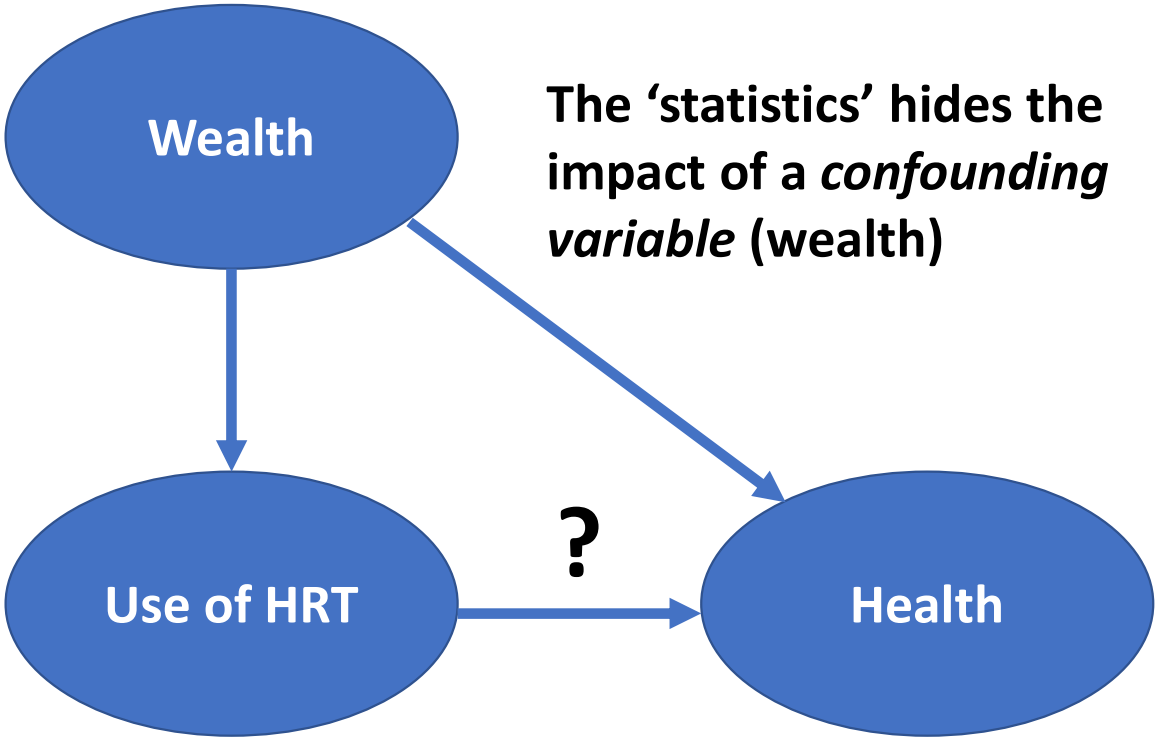
The study, published in the journal Neurology, involved 31,750 people in China with an average age of 62. They did not have any history of stroke or other major health problems at the start of the study. Participants were followed for an average of six years, during which time there were 1,557 stroke cases.

Author Dr Xiaomin Zhang, of Huazhong University of Science and Technology in China, added: "These results highlight the importance of moderate napping and sleeping, and maintaining good sleep quality, especially in middle-age and older adults.

"Long napping and sleeping may suggest an overall inactive lifestyle, which is also related to increased risk of stroke."


Age
Heart condition
Hours of sleep
?
Stroke

Another famous study with 'flawed' conclusions based on observational data (1950's and 1960's): the health benefits of Hormone Replacement Therapy (HRT) treatment for women aged over 50



The 'statistics' hides the impact of a *confounding variable* (wealth)

https://www.teachepi.org/wp-content/uploads/OldTE/documents/courses/bfiles/The%20B%20Files_File1_HRT_Final_Complete.pdf

Does pre-natal care increase chances of infant survival?

Overall data for clinics

		Pre-natal care	
		Yes	No
Survives	Yes	93	90
	No	7	10
	Survival rate	93%	90%

Apparently …yes

Data for Clinic 1

		Pre-natal care	
		Yes	No
Survives	Yes	8	80
	No	2	10
	Survival rate	80%	88%

Data for Clinic 2

		Pre-natal care	
		Yes	No
Survives	Yes	85	10
	No	5	0
	Survival rate	94%	100%

Another example of Simpson's paradox

What is the risk of a person in the UK dying from COVID-19?

What does risk mean here?

- Probability a random person in UK will die from the virus in next X months?
- Probability ***I*** will die from the virus in next X months?

How to factor in:

- Statistics like rate of new cases per day, death rate of people who contract the virus
- Age, health, etc of individuals
- Exposure to people who have the virus
- Recent travel from specific countries
- Potential future interventions like travel bans, new diagnosis and treatments

“Universities just have to do better. It is absolutely unacceptable that 50% of students score below average and so we are going to tackle this“

Higher Education Minister

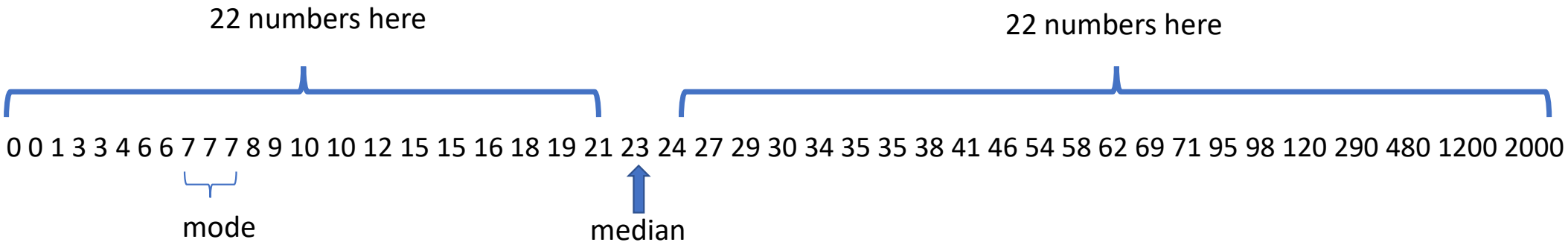
“One in every four children in the UK today lives in poverty. Moreover the number of children living in poverty has been increasing year on year”

Chief Executive of the Children's Society

How much do you agree with statements?

Averages =and their dangers

Here are 45 numbers in increasing order



Which average?

Mean (0+0+1+3+3+….+1200+200)/45 = 115 *in Excel:* *=AVERAGE(A1:A45)*

Median is the number in the middle of the sequence = 23 *in Excel:* *=MEDIAN(A1:A45)*

Mode is the most commonly occurring number = 7 *in Excel:* *=MODE(A1:A45)*

Averages =and their dangers

Here are 45 numbers in increasing order

22 numbers here

22 numbers here

0 0 1 3 3 4 6 6 7 7 7 8 9 10 10 12 15 15 16 18 19 21 23 24 27 29 30 34 35 35 38 41 46 54 58 62 69 71 95 98 120 290 480 1200 2000

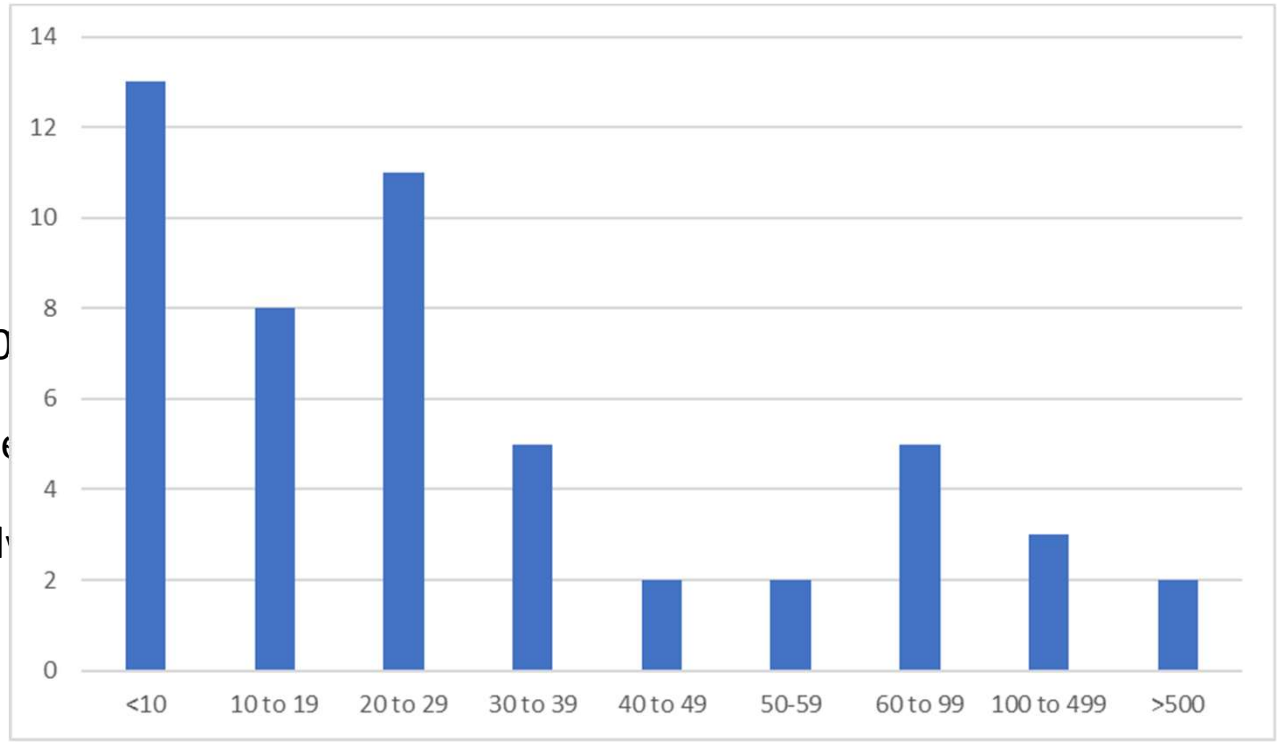
mode

Which average?

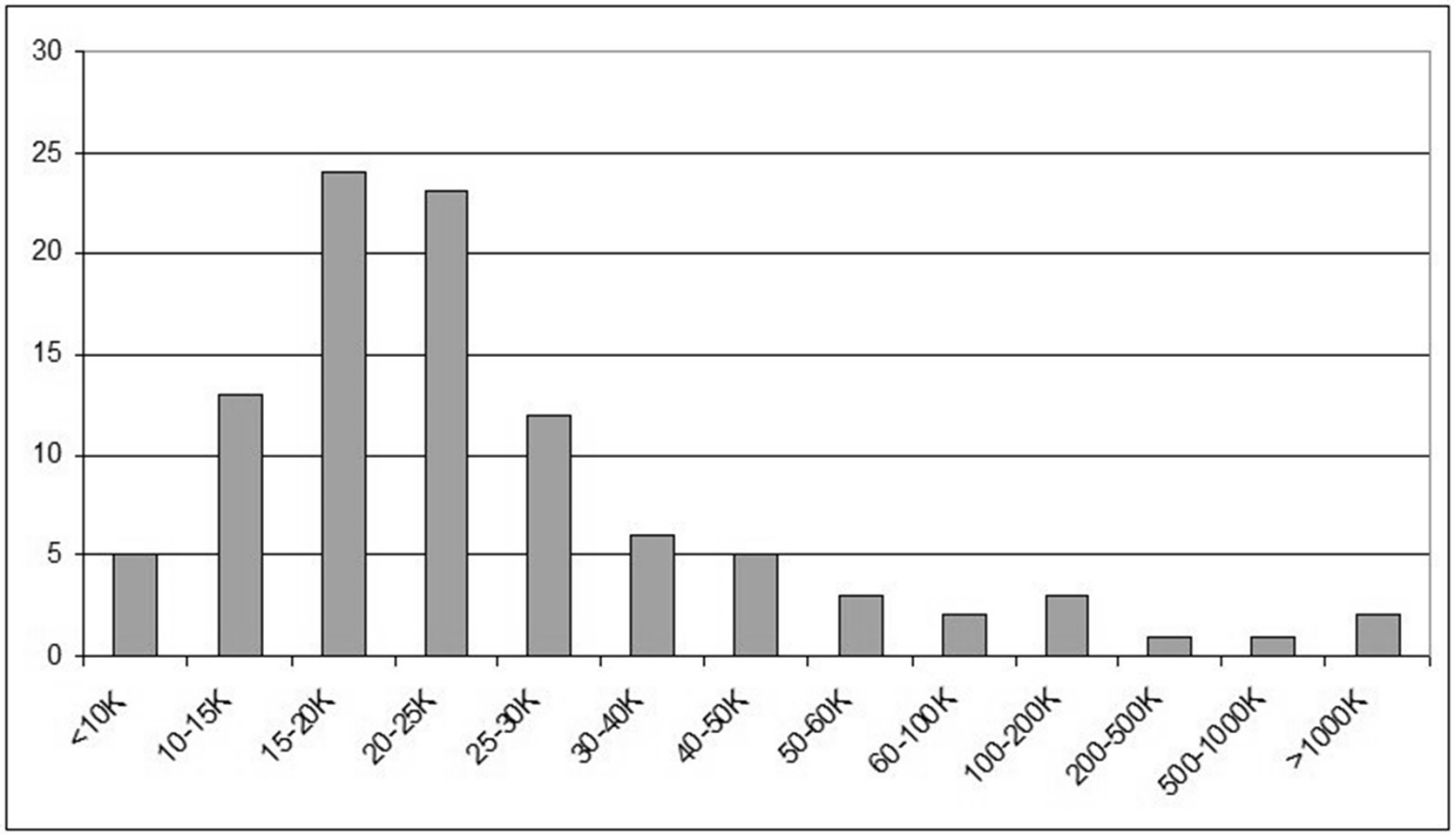
Mean (0+0+1+3+3+….+120

Median is the number in the

Mode is the most commonly



Averages =and their dangers



Distribution of salaries for people in a US city of 100,000 workers

What is the 'average' salary?

Mean salary = $137,000

Median salary = $23,000

By definition half the population earn at least the median salary
But only 5% of the population (5000 people) earn at least the mean salary

Averages =and their dangers

Scenario 1

As the mayor of the City you want to address problem of wealth inequality.
You introduce a modest wealth redistribution package
Every person earning above ‘average’ salary will pay a tax of $100
Every person earning below ‘average’ salary will receive a payment of $100
You are sure this will be popular and tax neutral
But the Finance Office interprets ‘average’ as MEAN
Only 5000 workers pay $100 each, while 95,000 receive $100 each
It’s popular but costs the City $9,000,000 which it does not have ….

Mean salary = $137,000

Median salary = $23,000

By definition half the population earn at least the median salary
But only 5% of the population earn at least the mean salary

Averages =and their dangers

Scenario 2

As the mayor of the City you have to raise $100 million for a new transport project. It's agreed this will be funded by a taking a fixed % of each worker's salary.

Not making same mistake as above this time the Finance offices use the MEDIAN for 'average' salary to calculate the required %.

As the 'average' is $23K they calculate that a tax of 4.35% will raise $100 million (since 4.35% of 23K is $1000)

But this calculation only makes sense if the mean was equal to the median.

Because the mean is $137K, the tax brings in nearly $600 million.

Basing the calculation on the mean salary requires a tax of only 0.73% to raise £100 million.

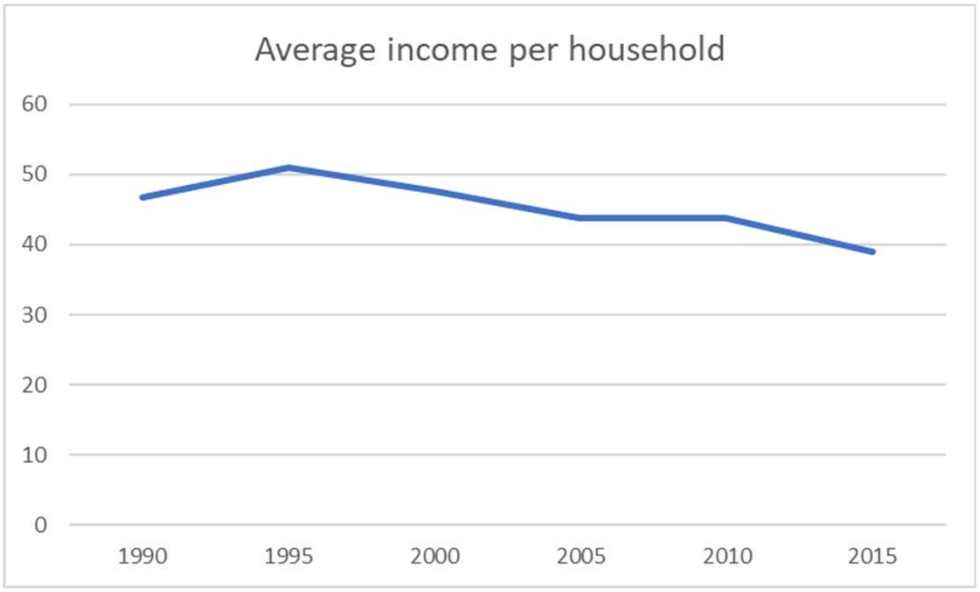
Although the City gets a surplus of $500 million you are voted out of office for what is seen as an unnecessarily savage tax on all workers

Mean salary = $137,000

Median salary = $23,000

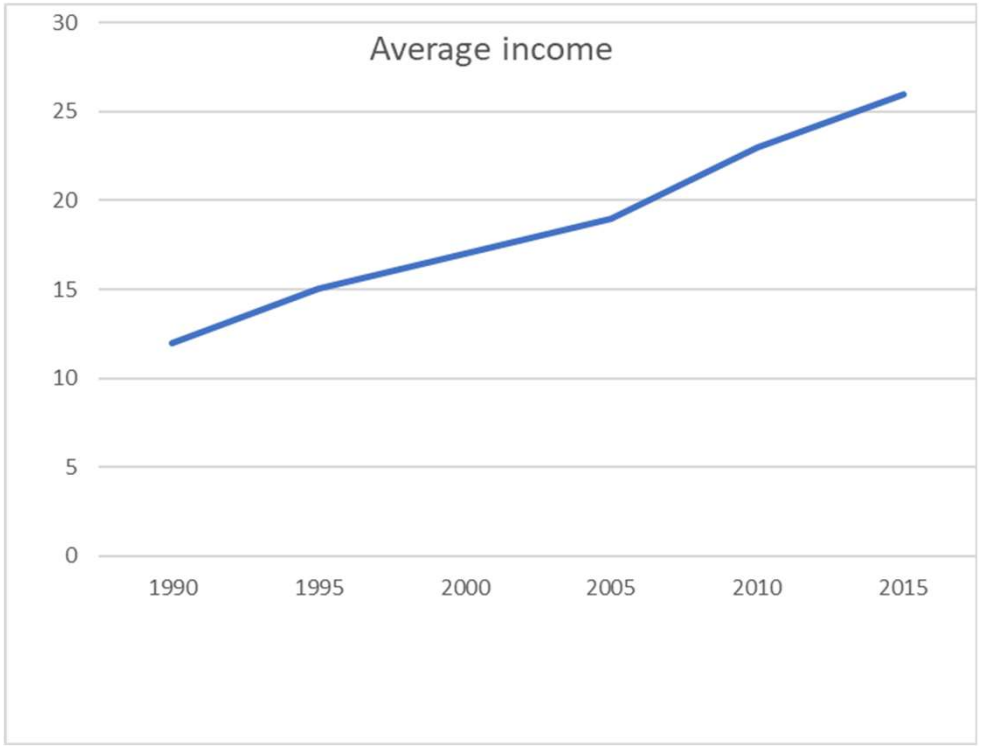
By definition half the population earn at least the median salary
But only 5% of the population earn at least the mean salary

Errors of omission



“Average Household incomes have been declining”

….have they really?



Year	Average income	Income earners per household	Average income per household
1990	12	3.9	46.8
1995	15	3.4	51
2000	17	2.8	47.6
2005	19	2.3	43.7
2010	23	1.9	43.7
2015	26	1.5	39

Omission: decreasing number of people per household

Errors of omission: US Commission of Human Rights, 2000

Racial group	Percentage mortgage applicants rejected	% applicants with credit scores below standard mortgage threshold	% applicants rejected by Black owned banks
Black	**44.3%**	**52%**	**47%**
White	**22.3%**	**16%**	
Asian	**12.4%**	**10%**	
Native Hawaiians	**12.4%**	**10%**	

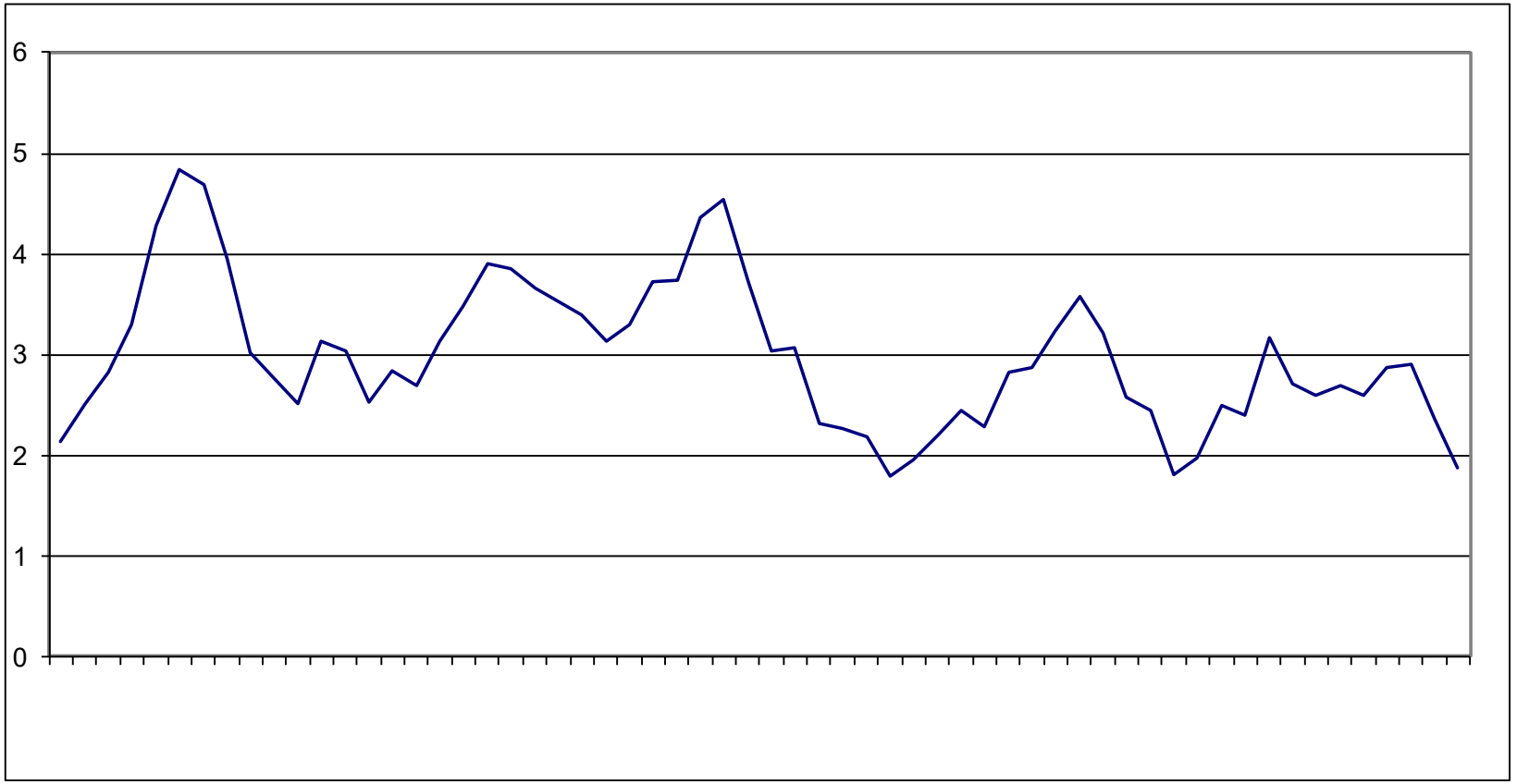
"This provides strong evidence of racial discrimination in mortgage lending institutions"

Does it?

Blacks disproportionately suffer poverty (possibly due to racial discrimination) but no evidence of racial discrimination by banks in mortgage decisions

Sowell, T, "Discrimination and Disparities" (Chapter 4), 2018

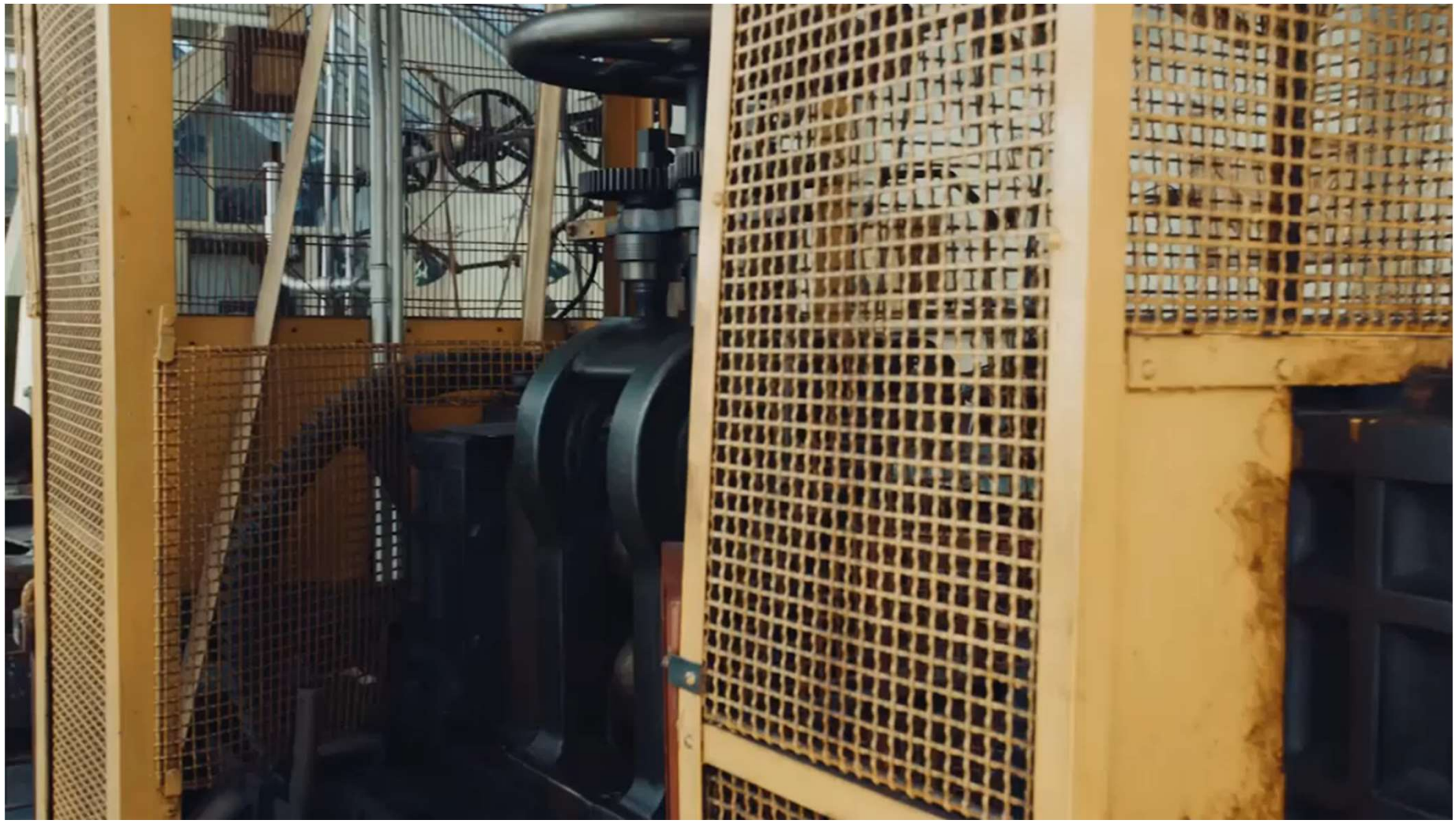
Predicting economic growth



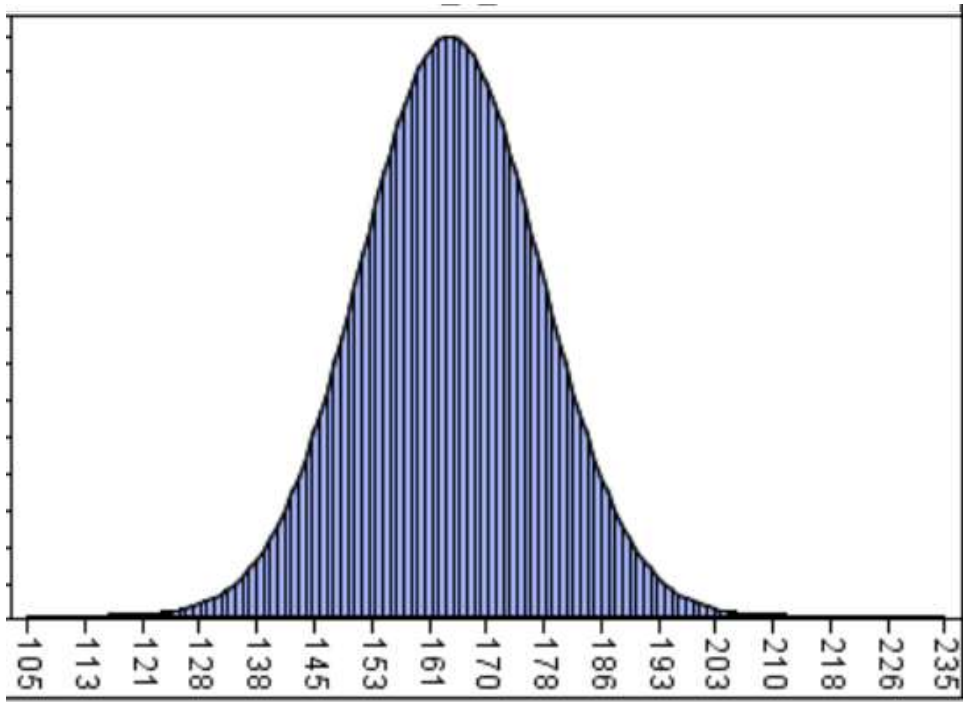
Growth rate in GDP (%) over 16 years (64 quarters)

- What are the chances that the next period growth rate will lie between 1.5% and 3.5%?
- What are the chances that the growth will be less than 1.5%?
- What are the chances that there will be negative growth?

To answer these questions we need a model…….



The Normal Distribution



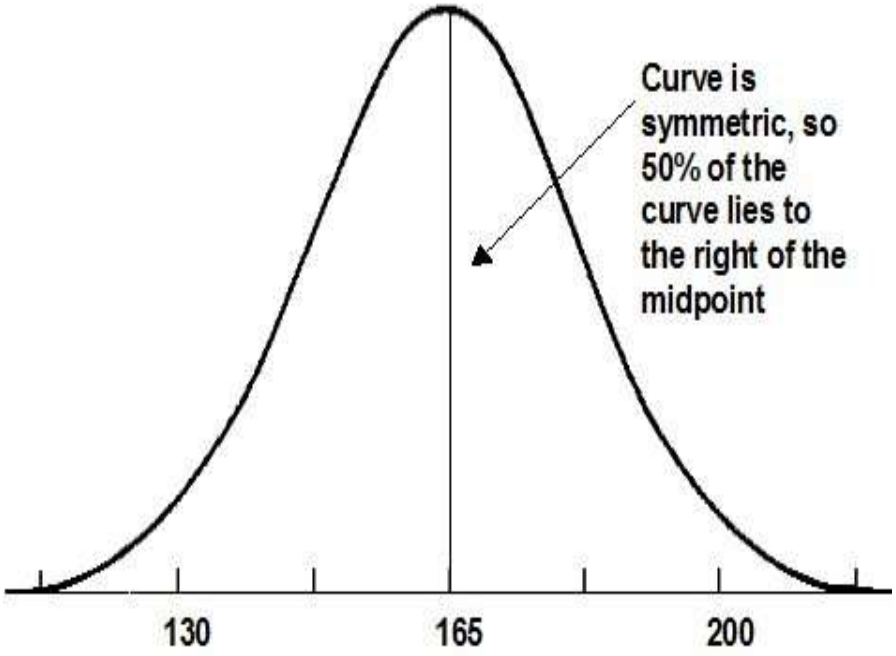
Sample a lot of adults and measure their height in centimetres

First plot the results as a histogram using intervals size 10

Then plot using increasingly small interval sizes

We end up with something approximating a symmetric curve – the Normal distribution

The Normal Distribution



Approximately 20% of this distribution lies between 178 and 190, we can conclude that there is a 20% chance a randomly selected adult will be between 178 and 190 cm tall.

Because the distribution is symmetric the ***mean*** is the distribution midpoint, so is the same as the median (165 in this case)

The 'spread' of the data (which is the extent to which it varies from the mean) is captured by a single number called the ***standard deviation*** (about 13 in this case)

So the Normal distribution is completely characterized by just these 2 parameters.

Tables, calculators, Excel, AgenaRisk etc will then tell you all kinds of 'probability' results

99.7% of any Normal distribution lies beyond the mean and three standard deviations. So in this case there is about 1 in 600 chance of an adult being more than 204 cm tall (6ft 7in) according to the model

BUT……Infinite tails of the normal distribution imply non-zero (albeit very small) chance that an adult can be less than 0 cm tall, and a non-zero chance that an adult can be taller than the Eiffel tower.

Making probabilistic predictions from a distribution like the Normal distribution

If we assume the height of people in the UK is a Normal(165, 13) distribution, what is the probability the next person to enter the room is 210 cm tall?

First note that this is **not a sensible question** because in theory the probability any person is EXACTLY 210 cm is 0.

So, it only makes sense to phrase it as something like:

"what is the probability the next person to enter is between 209.5 and 210.5 cm tall?" (Answer 0.00765%)

Or, more usefully:

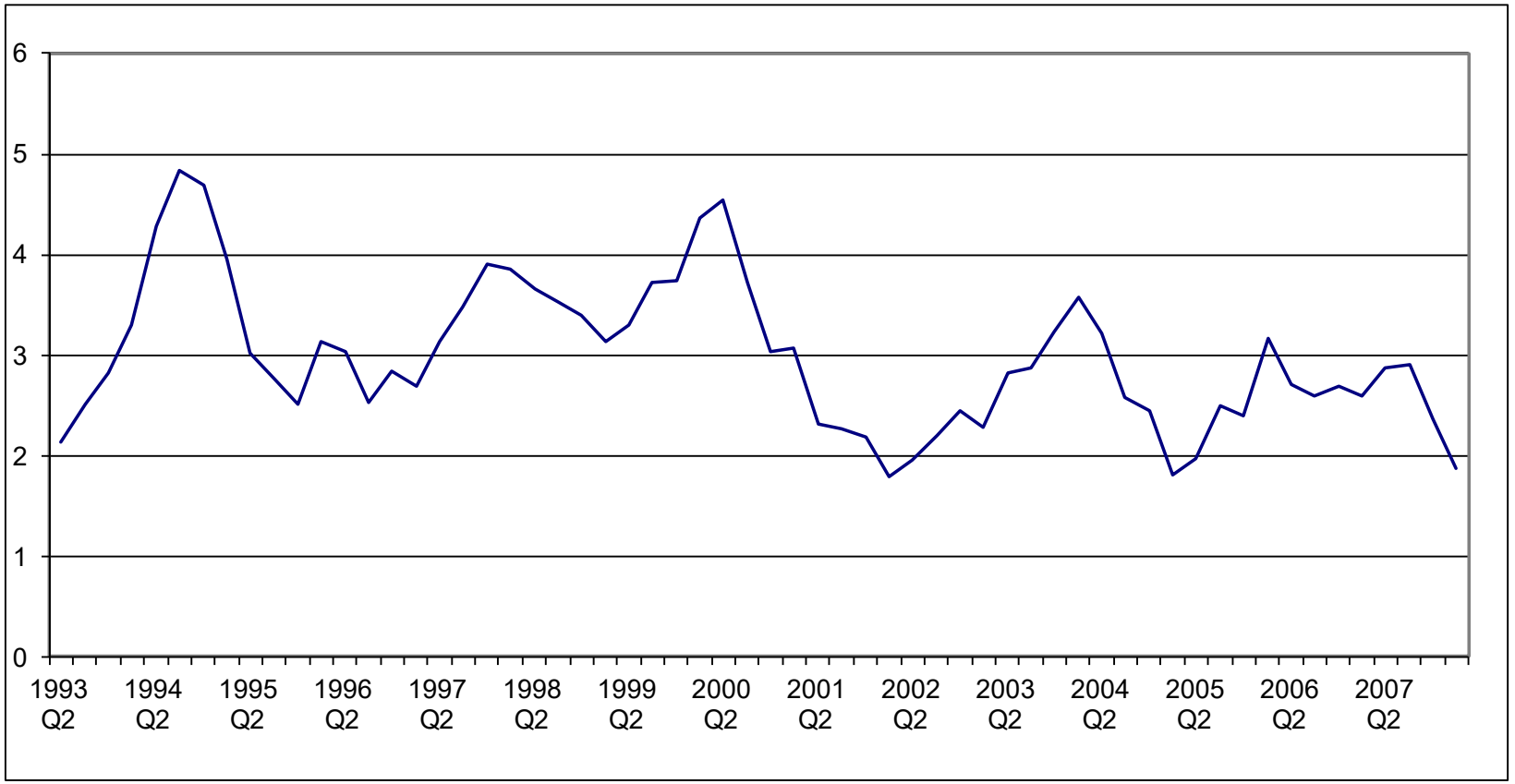
"what is the probability the next person to enter is at least 210 cm tall?" (Answer: 0.027%)

The probability of observing a value greater than *x* is sometimes referred to as the ***p-value*** for *x* (so here, the p-value of 210 is 0.0027%). The p-value gives a feel for how unlikely the observation *x* is; the lower the p-value the less likely you are to observe it.

We will see that p-values are widely (but not necessarily sensibly) used in hypothesis testing.

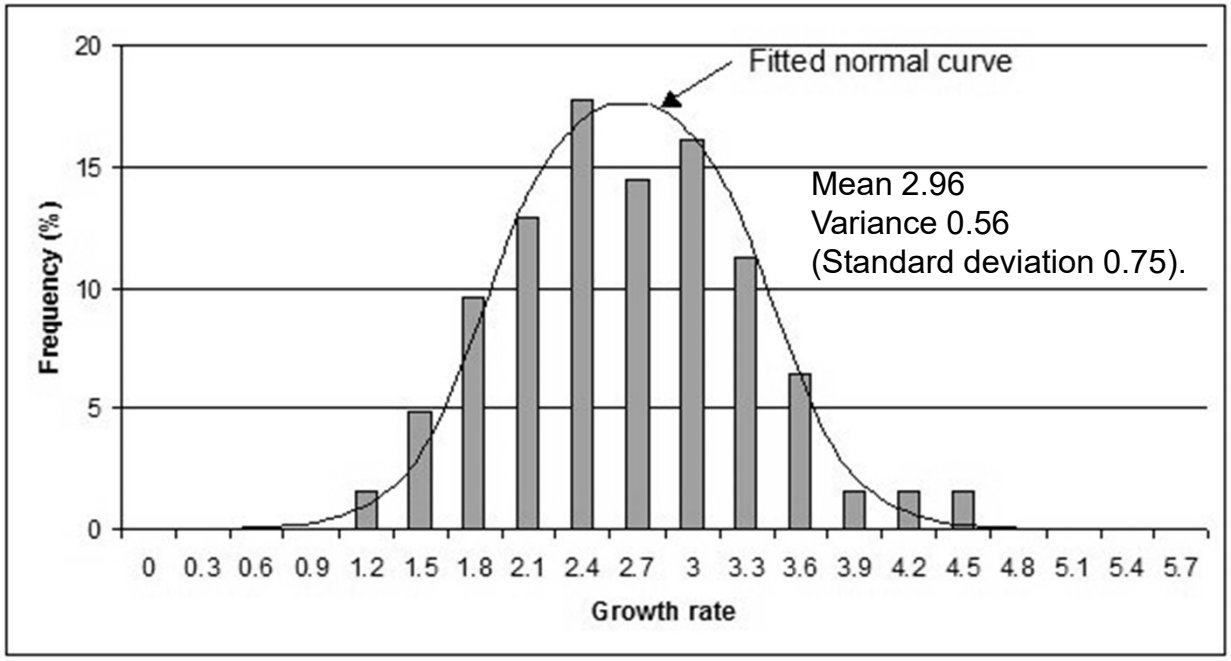
Predicting economic growth

What I showed were the actual UK treasury GDP figures for the period 1993-2008



Growth rate in GDP (%) over time from first quarter 1993 to first quarter 2008

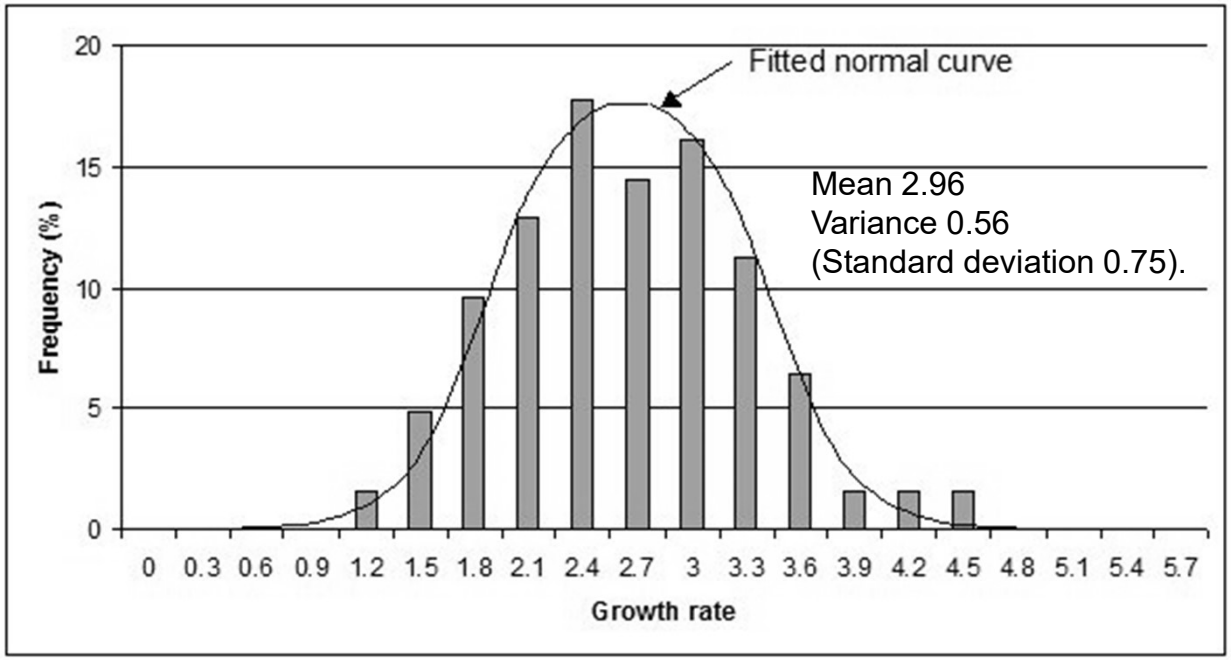
‘Fitting’ a Normal distribution model to the GDP data



Histogram of annualised GDP growth rate from 1993-2008

The data
(see excel file normal_model_fitting)
when plotted as a histogram looks much like a Normal distribution
So it makes sense to ‘fit’ a Normal distribution
Watch the short videos showing you how to do this in MatLab and AgenaRisk
The best ‘fit’ Normal distribution has a mean of 2.96 and standard deviation 0.75
Note that in Excel calculating the mean and standard deviation of the dataset gives slightly different results, because it is not attempting to fit a model.

‘Fitting’ a Normal distribution model to the GDP data



Histogram of annualised GDP growth rate from 1993-2008

What are the chances that next period growth rate will be:

between 1.5% and 3.5%?

- Answer based on the model: approximately 74%.

less than 1.5%?

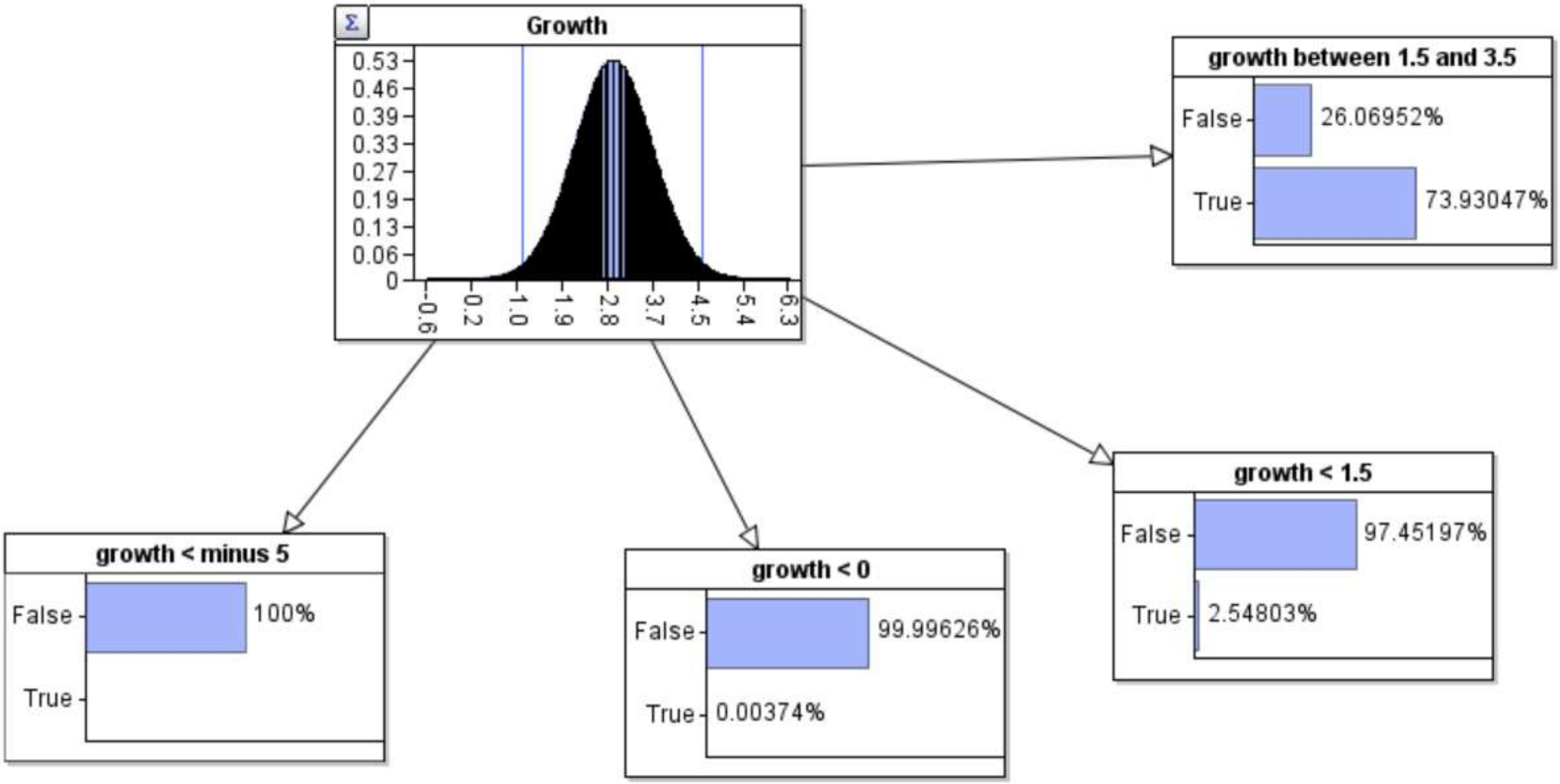
- Answer: about 2.5%

will be negative?

-Answer: less than 0.004%, i.e. about 1 in 25,000

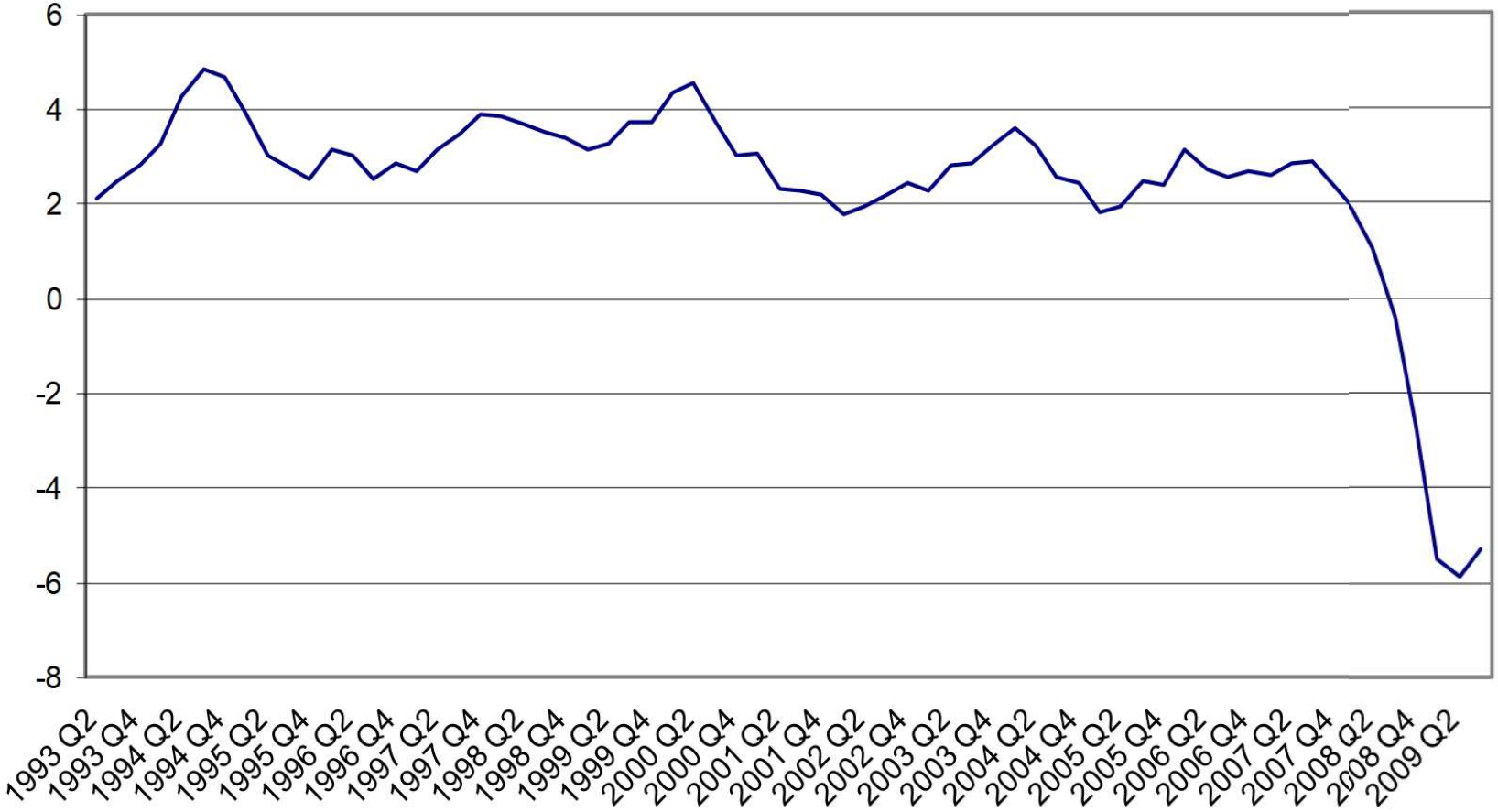
will be less than -5%?

-Answer: essentially zero



MODEL: normal_learning_growth_rate

What happened next?



Growth rate in GDP (%) over time from 1993 to 2010

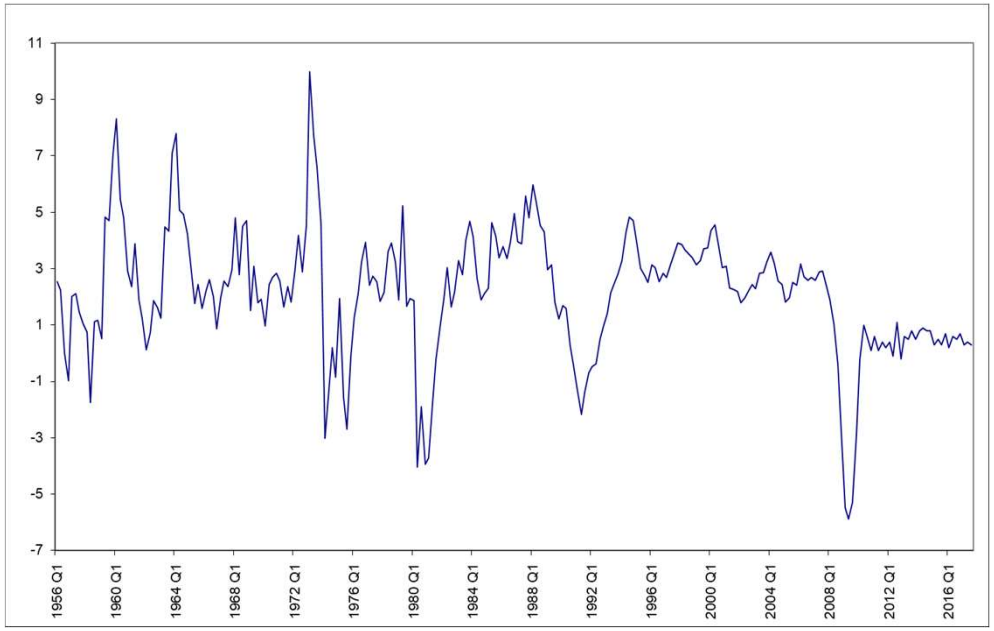
Within less than a year the growth rate was below –5%. According to the model a growth rate below –5% would essentially be impossible.

So what went wrong?

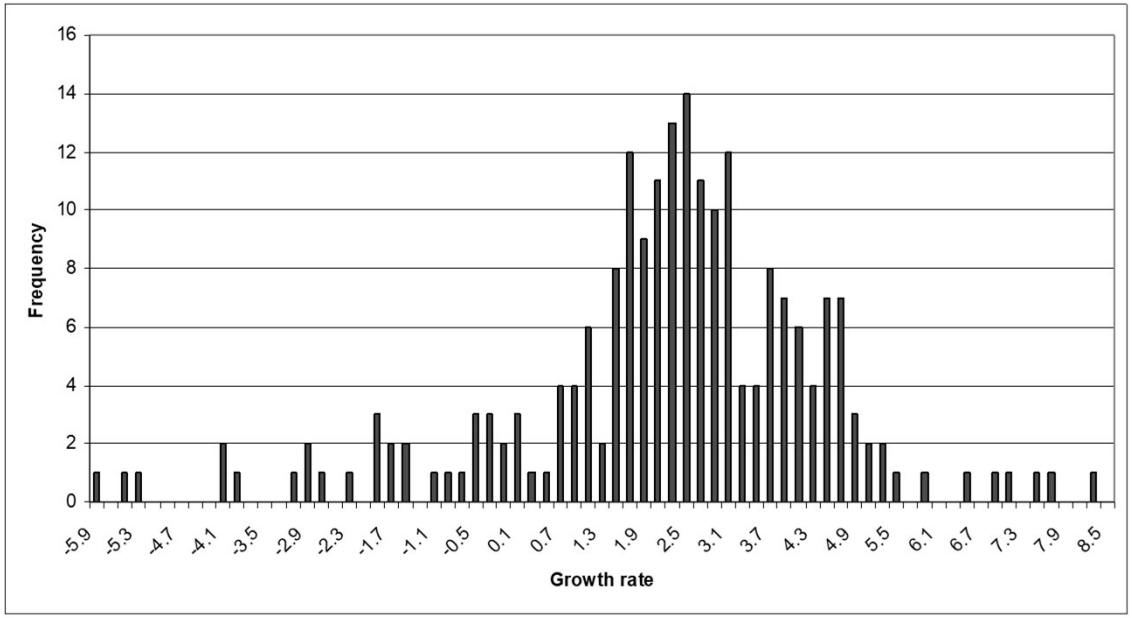
Clearly Normal distribution was a hopelessly inadequate model since it *is inherently incapable of predicting rare events*.

Actual predictions made by financial institutions and regulators in the period running up to the credit crunch were especially optimistic because they based estimates of growth on the so called 'Golden Period' of 1998-2007.

Lessons from history



Growth rate in GDP (%) over time from 1956 to 2017



Histogram of annualised GDP growth rate from 1956-2017

Conditions in 2008 were unlike any that had previously been seen. The standard statistical approaches inevitably fail in such cases.

Not only is the spread of the distribution much wider, but it is clearly not ‘Normal’ because it is not symmetric.

…but it is just as misleading to ignore data that *really is* approximately Normal

- "Ball number 38 is the lucky ball. It has been selected 60% more times than ball number 41"

- "In 2019 the death rate at Hospital X was more than 50% above the national average…"
- "…moreover, it increased significantly since 2019. Action must be taken…."

- "In Dec 2003 there were 18 unexplained deaths in Hospital Y compared to a monthly average of 5. Nurse B was on duty at the time of all these deaths. Hence nurse B must have been maliciously involved.

Sporting Form: quality or luck?

	Ars	Bou	Bur	Che	CP	Ev	Hull	Lei	Liv	MC	MU	Mid	Sou	Sto	Sun	Swa	Tot	Wat	WB	WH
Arsenal		1	2	1	0	1	0	1	1	1	0	0	0	1	0	1	2	1	0	0
Bournemouth			0	2	1	0	0	2	2	1	0	2	1	1	1	0	1	2	0	2
Burnley				2	0	1	0	1	0	0	1	2	2	1	0	1	1	2	2	0
Chelsea					1	0	0	2	2	1	2	2	2	2	0	2	0	2	1	0
Crystal Palace						0	1	1	0	2	1	0	0	2	0	2	2	0	1	2
Everton							1	1	2	2	2	2	2	0	1	2	2	2	0	1
Hull City								2	0	2	2	0	2	1	1	1	2	1	0	2
Leicester									2	1	0	1	0	1	0	0	2	2	0	1
Liverpool										2	2	2	1	0	1	2	0	1	2	2
Man City											2	0	2	0	1	0	2	2	2	0
Man United												0	2	1	1	0	0	1	1	1
Middlesbrough													2	2	1	2	1	2	2	2
Southampton														0	1	0	0	0	0	1
Stoke City															1	1	2	0	0	0
Sunderland																1	0	1	0	2
Swansea City																	1	2	0	0
Tottenham																		1	0	0
Watford																			1	2
West Brom																				2
West Ham																				

A randomly generated set of ‘results’

In Excel use:
RANDBETWEEN(0,2)

2=victory for team on left
1= draw
0=defeat for team on left

so, e.g. Bournemouth lost to Burnley

Sporting Form: quality or luck?

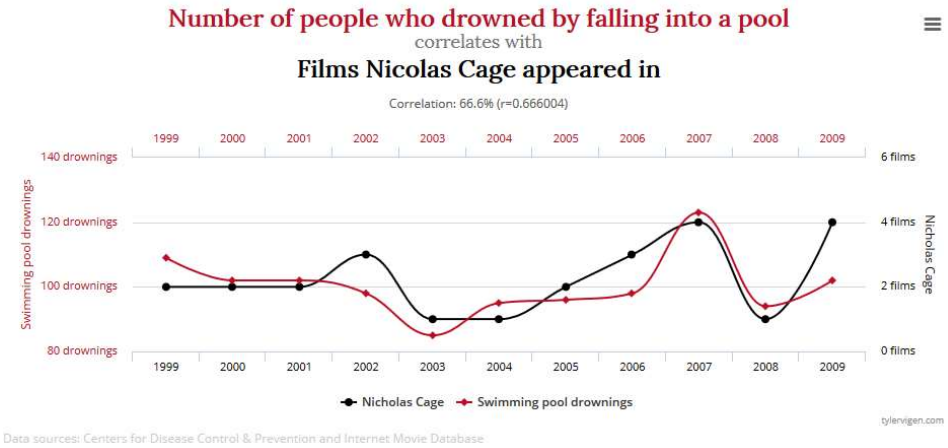
Everton	40
Middlesbrough	39
West Brom	37
Hull City	36
Liverpool	31
Crystal Palace	30
Chelsea	29
Burnley	28
Sunderland	29
Swansea	27
Bournemouth	26
West Ham	25
Man City	24
Stoke City	24
Tottenham	23
Watford	22
Man United	20
Southampton	16
Arsenal	15
Leicester City	15

League “Table” after these 19 games

Chelsea	49
Liverpool	43
Arsenal	40
Tottenham	39
Man City	39
Man United	36
Everton	27
West Brom	26
Bournemouth	24
Southampton	24
Burnley	23
West Ham	22
Watford	22
Stoke City	21
Leicester City	20
Middlesbrough	18
Crystal Palace	16
Sunderland	14
Hull City	13
Swansea	12

Real league table after 19 games Jan 2017

Correlation, regression and ‘significance’

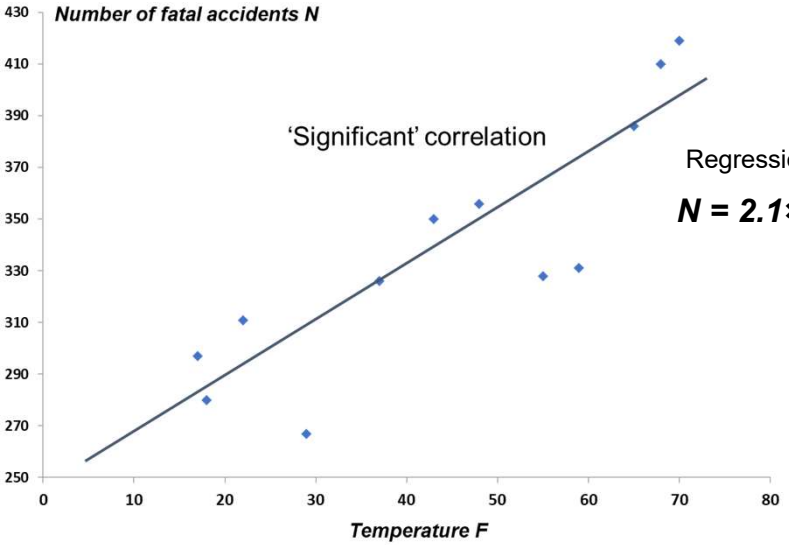


Month	Average temperature	Total fatal crashes
January	17.0	297
February	18.0	280
March	29.0	267
April	43.0	350
May	55.0	328
June	65.0	386
July	70.0	419
August	68.0	410
September	59.0	331
October	48.0	356
November	37.0	326
December	22.0	311

Excel computes:
R=0.8699
Significance = 0.000235
Much less than the 1%, i.e. 0.01 level, so this is a very highly significant relationship

The correlation coefficient R is a number between -1 and 1 that determines whether two paired sets of data are ‘related’.
Excel’s default correlation function is the *Spearman measure* which is the most common
R close to 1 indicates a positive correlation
R close to -1 indicates a negative correlation (e.g. *daily temperature* against *daily hours of darkness*)
R close to 0 indicates no relationship.
However, our ‘confidence’ in whether or not there is a relationship clearly depends on **how much data** we have.
The formal measure of such confidence (statistical significance) is based on a particular notion of p-value (we will explain exactly what this is – and why it is problematic – in the next lecture but it is something like ‘the probability of observing the data if there is no relationship’.
With a lot of data it is possible to have a ‘highly significant’ (e.g. 1% or 0.01 p-value) relationship with values quite close to 0.
Tables or Excel (regression) provide this information. E.g. for 100 pairs of data a correlation is 0.254 has p-value 1%

(Linear) Regression



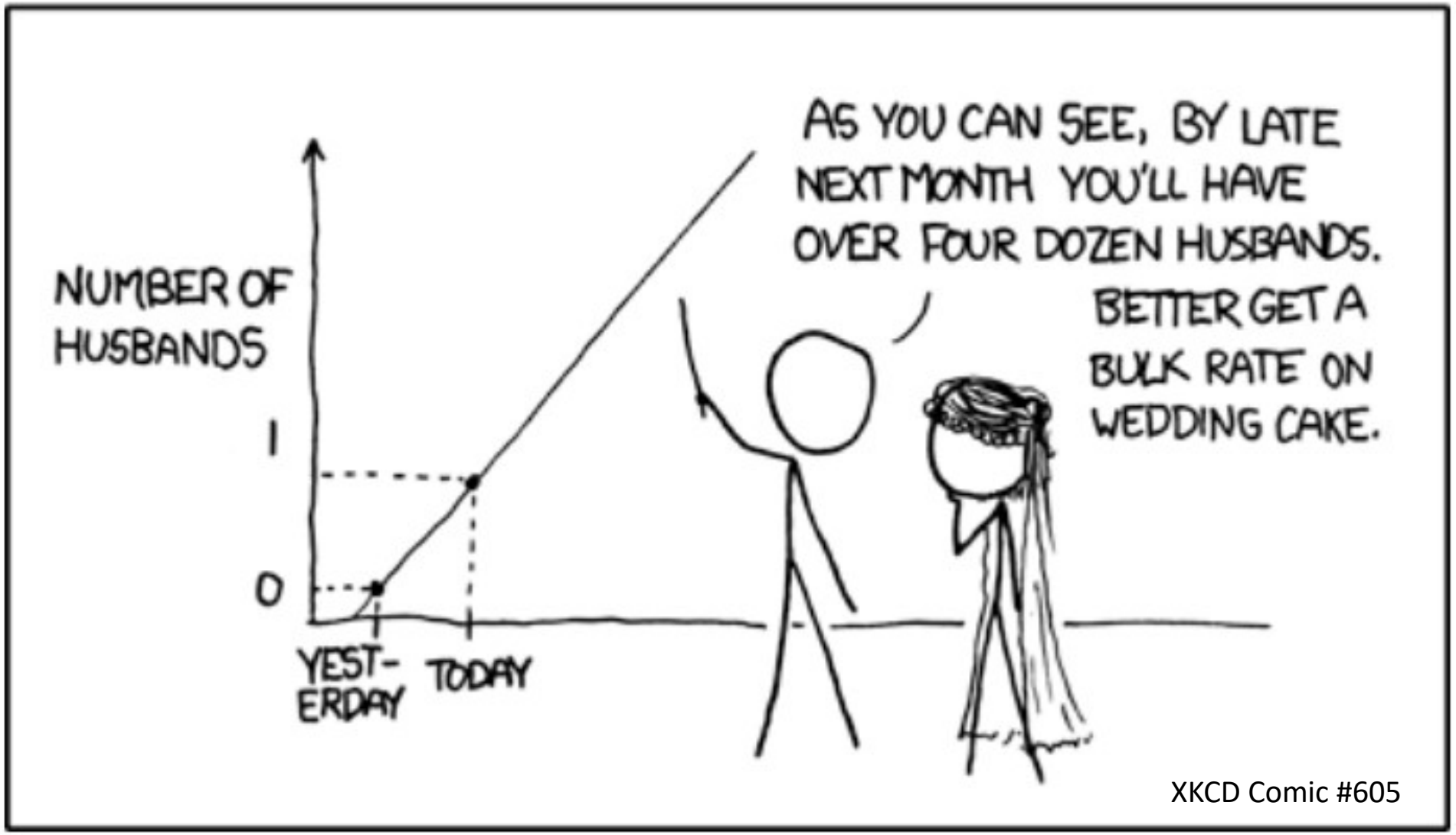
SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.869898				
R Square	0.756722				
Adjusted F	0.732394				
Standard I	24.96697				
Observatic	12				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	19389.42	19389.42	31.1052	0.00023502
Residual	10	6233.498	623.3498		
Total	11	25622.92			
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Uppe*
Intercept	243.5465	18.47424	13.18303	1.2E-07	202.3833618	284
Temp	2.143958	0.384415	5.577203	0.000235	1.287428912	3.0(

See the short videos explaining how to do this in both Excel and MatLab.

The (linear) regression equation is the 'line of best fit'

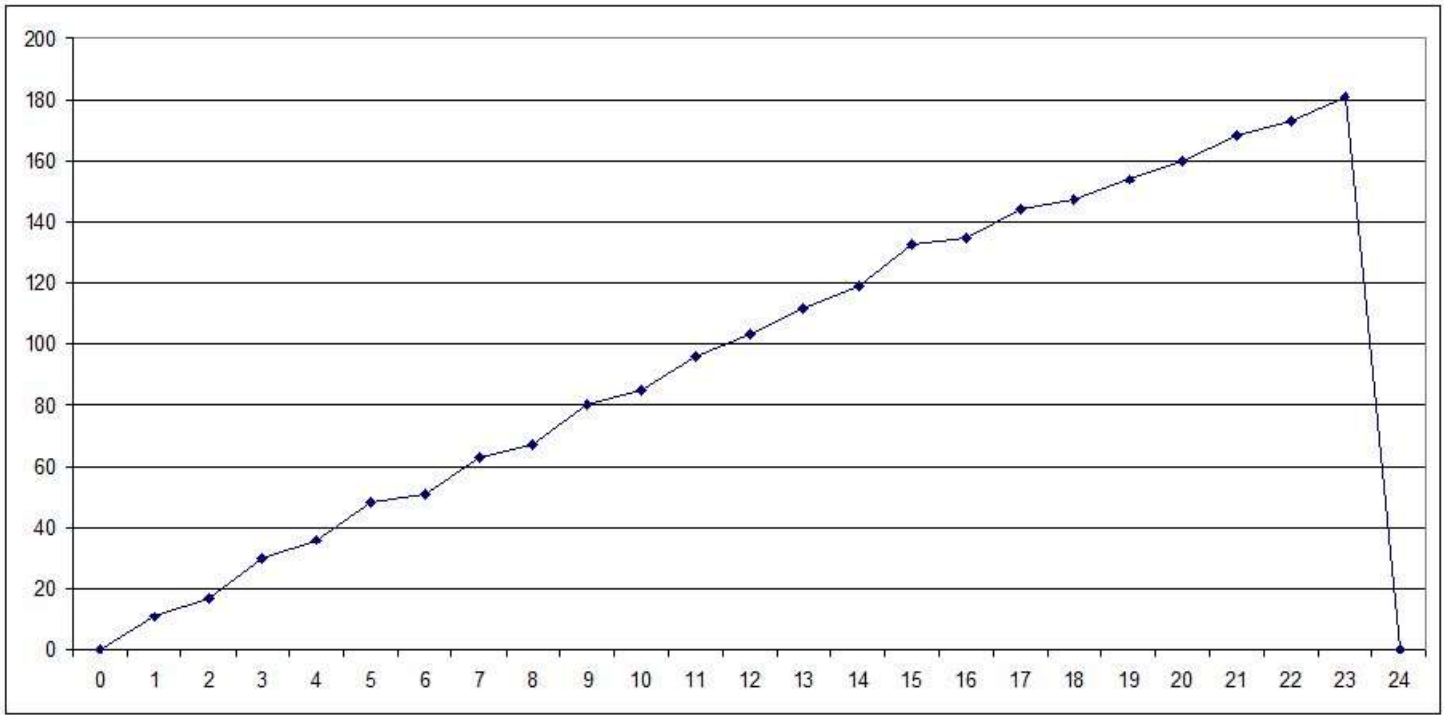
- It minimizes the summed distances from the points

MY HOBBY: EXTRAPOLATING



XKCD Comic #605

The Danger of Regression:
Looking backwards when you need to look forwards?



Suppose that you are blowing up a large balloon. After each puff you measure the surface area and record it.

What will the surface area be on the 24th puff?

Regression models

Spurious correlations and significant relationships are inevitable given much research practice

"Doing/ eating/ drinking.. X increases the risk of getting disease D"

"Working nightshifts leads to increased risk of breast cancer"

"Drinking coffee reduces risk of liver cancer by 50%"
2019 Study, Queens University Belfast

Beware the scattergun approach…..

A typical data-driven study.....

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	S
10	9	5	3	10	4	7	4	3	9	5	4	5	9	4	4	10	5	**5**
10	10	4	2	3	10	6	1	8	5	8	8	8	7	6	3	8	3	**1**
5	7	9	9	3	2	5	2	6	6	7	8	2	9	10	3	8	2	**1**
9	3	6	10	3	1	5	8	2	9	5	8	7	4	8	8	2	7	**7**
2	6	1	1	10	8	8	5	8	7	10	4	7	9	7	4	3	7	**3**
10	3	6	7	1	10	9	9	6	2	8	5	8	3	9	9	2	2	**7**
1	1	1	7	5	1	4	9	1	6	9	8	9	9	4	1	2	7	**5**
3	5	8	4	2	4	6	2	7	9	5	2	2	5	4	3	2	1	**1**
1	8	8	10	6	4	10	7	6	6	5	7	3	7	10	7	4	9	**8**
4	4	8	8	3	1	1	9	1	9	10	9	10	2	8	1	3	4	**10**
9	3	5	3	3	2	4	4	3	10	4	9	8	7	3	10	2	8	**4**
2	3	1	1	6	7	10	5	5	1	4	4	3	10	9	5	7	1	**6**
10	9	1	3	10	6	7	7	8	1	9	4	3	7	3	3	10	3	**7**
4	2	5	10	9	9	2	4	9	8	9	7	5	7	6	6	1	7	**2**
1	7	3	5	5	8	8	10	2	10	7	10	2	10	4	8	5	2	**8**
9	8	8	1	4	2	8	7	10	1	6	8	1	1	9	6	4	1	**2**
4	4	3	2	4	7	5	3	1	3	5	10	2	5	2	6	7	8	**9**
4	7	3	10	5	10	7	3	6	6	6	5	10	8	1	6	2	7	**5**
10	9	6	5	9	7	4	8	10	2	8	6	3	5	9	9	6	3	**2**
7	5	5	3	5	4	8	4	3	4	4	2	1	9	6	4	7	6	**9**

A8 & S
0.56
correlation

0.01
p-value

To generate such a table enter: RANDBETWEEN(1,10) In an Excel cell, then copy & paste to other cells

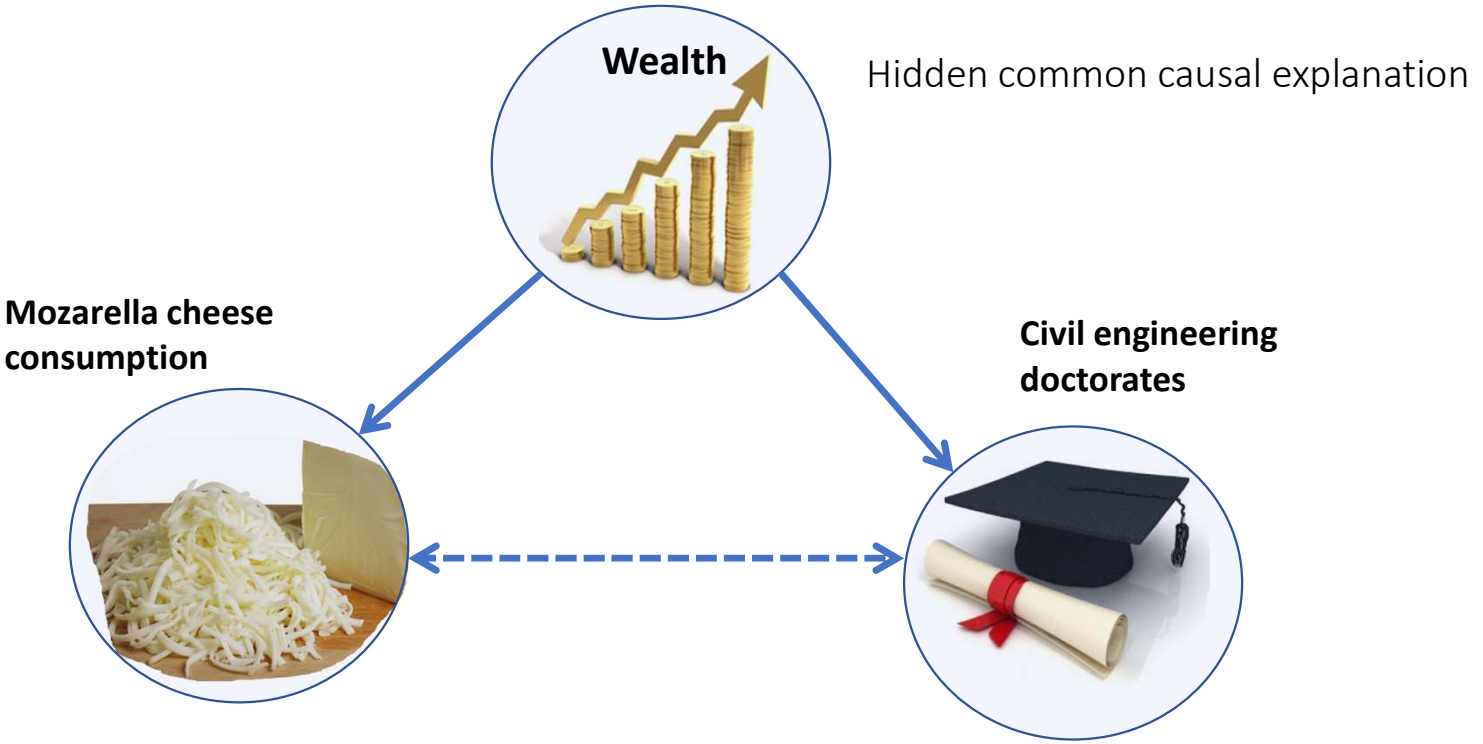
A typical data-driven study…..

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	S
10	9	5	3	10	4	7	4	3	9	5	4	5	9	4	4	10	5	**5**
10	10	4	2	3	10	6	1	8	5	8	8	8	7	6	3	8	3	**1**
5	7	9	9	3	2	5	2	6	6	7	8	2	9	10	3	8	2	**1**
9	3	6	10	3	1	5	8	2	9	5	8	7	4	8	8	2	7	**7**
2	6	1	1	10	8	8	5	8	7	10	4	7	9	7	4	3	7	**3**
10	3	6	7	1	10	9	9	6	2	8	5	8	3	9	9	2	2	**7**
1	1	1	7	5	1	4	9	1	6	9	8	9	9	4	1	2	7	**5**
3	5	8	4	2	4	6	2	7	9	5	2	2	5	4	3	2	1	**1**
1	8	8	10	6	4	10	7	6	6	5	7	3	7	10	7	4	9	**8**
4	4	8	8	3	1	1	9	1	9	10	9	10	2	8	1	3	4	**10**
9	3	5	3	3	2	4	4	3	10	4	9	8	7	3	10	2	8	**4**
2	3	1	1	6	7	10	5	5	1	4	4	3	10	9	5	7	1	**6**
10	9	1	3	10	6	7	7	8	1	9	4	3	7	3	3	10	3	**7**
4	2	5	10	9	9	2	4	9	8	9	7	5	7	6	6	1	7	**2**
1	7	3	5	5	8	8	10	2	10	7	10	2	10	4	8	5	2	**8**
9	8	8	1	4	2	8	7	10	1	6	8	1	1	9	6	4	1	**2**
4	4	3	2	4	7	5	3	1	3	5	10	2	5	2	6	7	8	**9**
4	7	3	10	5	10	7	3	6	6	6	5	10	8	1	6	2	7	**5**
10	9	6	5	9	7	4	8	10	2	8	6	3	5	9	9	6	3	**2**
7	5	5	3	5	4	8	4	3	4	4	2	1	9	6	4	7	6	**9**

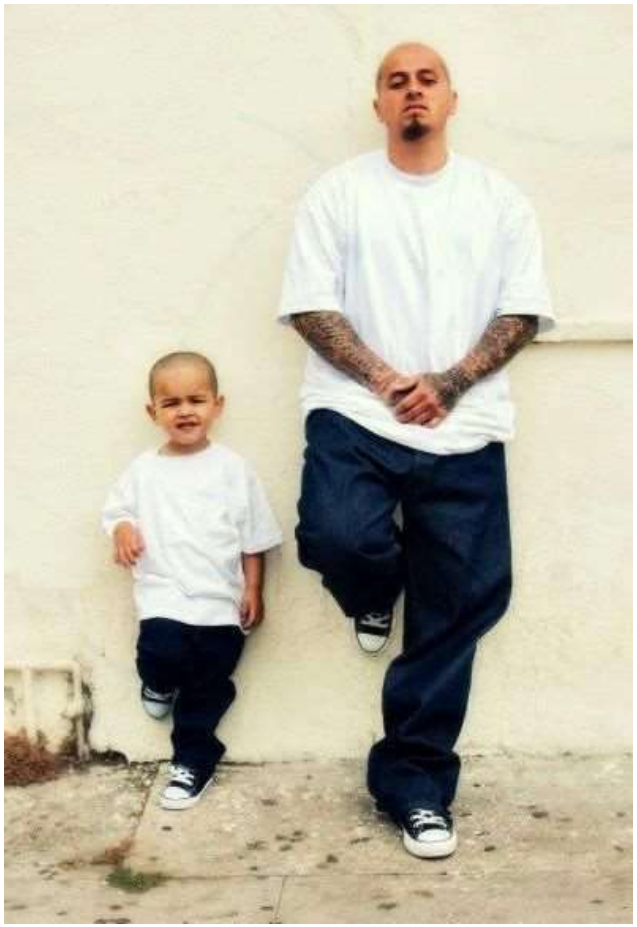
A2 & A17
0.62
correlation

0.0035
p-value

But remember that many relationships that seem ‘spurious’ are not

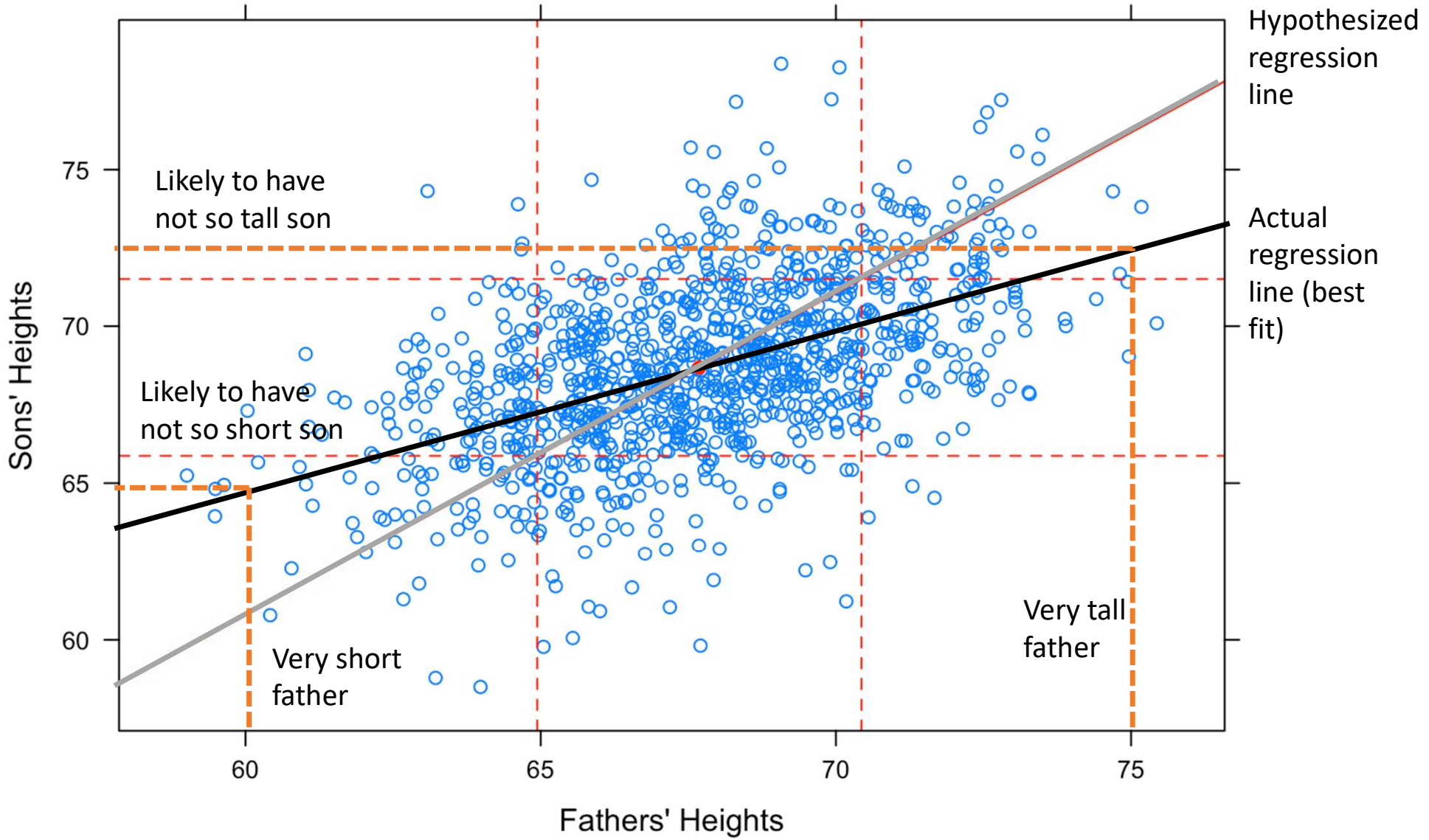


Do tall parents have tall children?

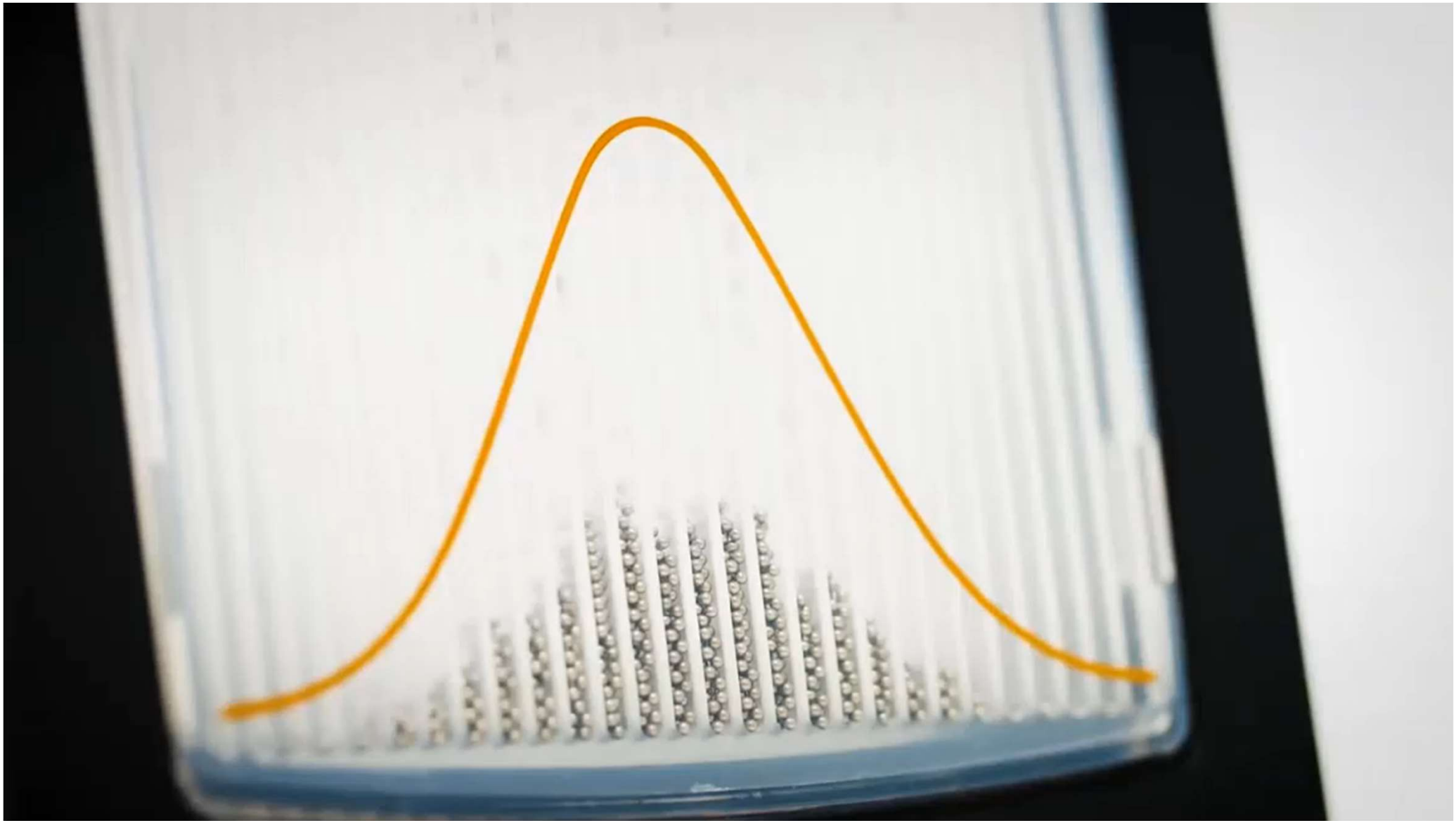


What is the probability that the son of a six foot tall father will be at least six foot tall?

Regression to the mean



Regression to the mean



Regression to the mean

Why do football clubs generally perform better immediately after a manager is sacked?
Why do football clubs generally perform worse immediately after a manager leaves of his own accord (typically for a better job)?

Student test scores: Why do students who perform very badly on Test 1 tend to perform less badly on Test 2? Similarly, why do students who perform very well on Test 1 tend to perform less well on Test 2?

Why does use of a placebo in clinical trials demonstrate both a) the effect of regression to the mean and b) how to potentially overcome wrongly concluding that a new intervention/drug is effective.

Models – workshop session