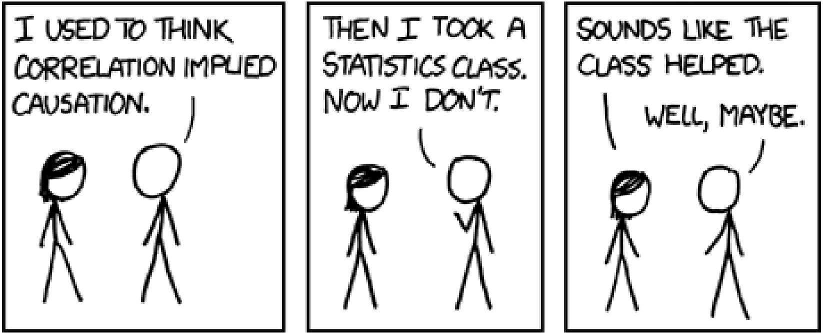
Correlation vs. Causation

Data Analysis for Journalism and Political Communication
(Spring 2024)

Prof. Bell


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

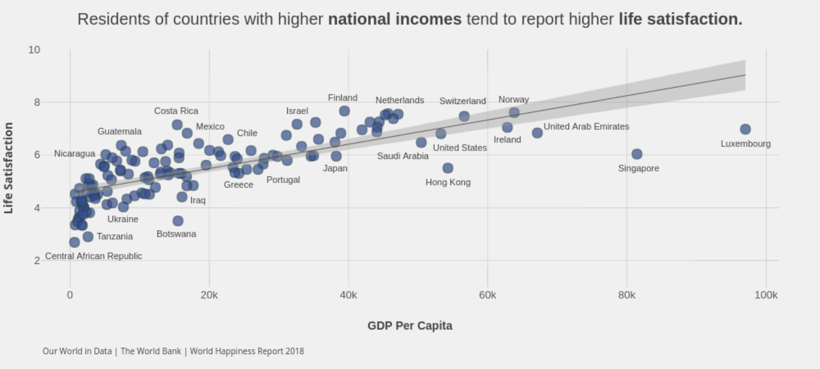
Correlation

Correlation exists when the absolute rate of change in the values of two variables are similar.

Correlation

Correlation exists when the absolute rate of change in the values of two variables are similar.

- **Positive correlation**: As the value of one variable increases (decreases), the value of the other variable increases (decreases) at the same rate


Residents of countries with higher **national incomes** tend to report higher **life satisfaction.**
Life Satisfaction
GDP Per Capita
10
8
6
4
2
0
20k
40k
60k
80k
100k
Finland
Netherlands
Switzerland
Norway
Israel
Costa Rica
Mexico
Chile
Guatemala
Nicaragua
United Arab Emirates
Ireland
United States
Saudi Arabia
Luxembourg
Singapore
Japan
Hong Kong
Portugal
Greece
Iraq
Ukraine
Botswana
Tanzania
Central African Republic
Our World in Data | The World Bank | World Happiness Report 2018

Correlation

Correlation exists when the absolute rate of change in the values of two variables are similar.

- **Positive correlation**: As the value of one variable increases (decreases), the value of the other variable increases (decreases) at the same rate
- **Negative correlation**: As the value of one variable increases (decreases), the value of the other variable decreases (increases) at the same rate

Births per woman by income level, 2013

Our World in Data

Pink bubbles show country averages for income (GDP per capita, PPP adjusted) and for the total fertility rate.
For all other countries the fertility rate is shown for each wealth quintile within the country. It is plotted against the average income per corresponding quintile in the same country.

Data sources: World Bank for all income measures. Fertility rates: national averages from WDI. Fertility by wealth quintile from the DHS (via the WHO) – except for China for which data was added from various research papers. Most data are from 2013 – none of the data refer to a year earlier than 2005.

Causation

Causation exists when a change in the value of one variable would not be observed without a preceding change in the value of another variable.

Causation

Causation exists when a change in the value of one variable would not be observed without a preceding change in the value of another variable.

- Correlation is descriptive, while causation is predictive

Causation

Causation exists when a change in the value of one variable would not be observed without a preceding change in the value of another variable.

- Correlation is descriptive, while causation is predictive
- We say that the **independent** or **explanatory** variable causes the **dependent** or **outcome** variable

Causation

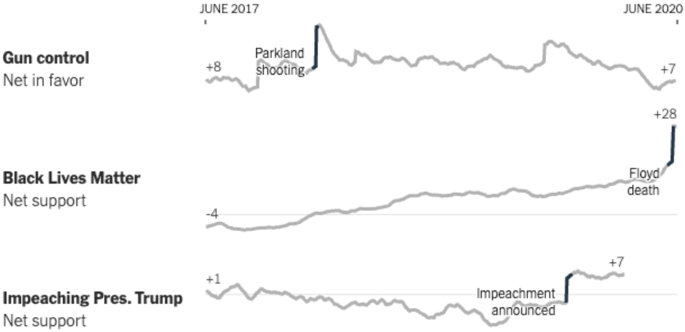
Causation exists when a change in the value of one variable would not be observed without a preceding change in the value of another variable.

- Correlation is descriptive, while causation is predictive
- We say that the **independent** or **explanatory** variable causes the **dependent** or **outcome** variable
- Causation depends on knowing the **counterfactual**: if we did not observe a change in the value of the explanatory variable, we would not observe a change in the value of the outcome variable

THE UPSHOT | How Public Opinion Has Moved on Black Lives Matter

How voters' views on other issues have changed in the last two years

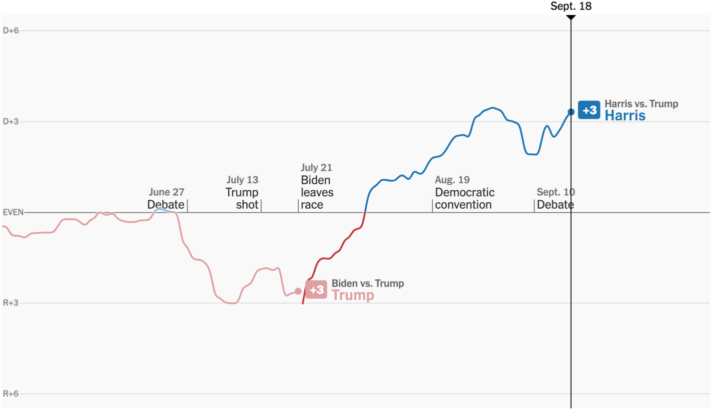
Large swings in public opinion in short periods are not typical. Two-week periods with the biggest shifts in movement are highlighted.



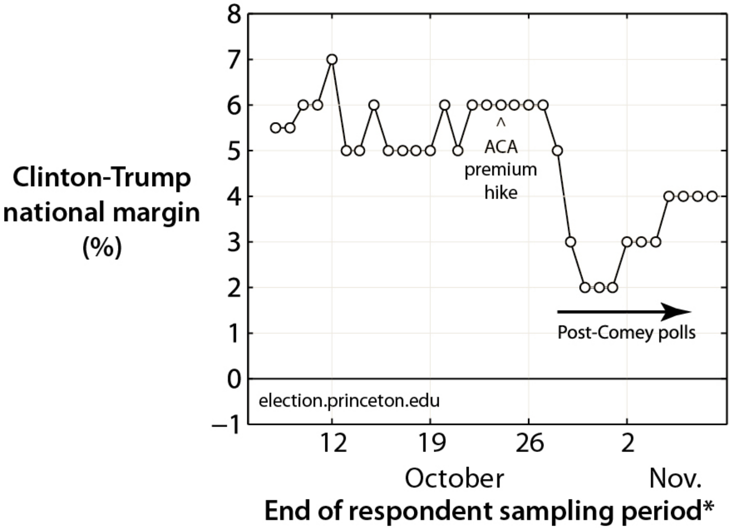
From Biden to Harris

This chart shows how the polling margin has changed over the course of the campaign, first for the Biden vs. Trump matchup, and now for Harris vs. Trump.

National polling average



Note: Head-to-head average shown for the Biden vs. Trump matchup. The Harris vs. Trump average includes polls conducted before Biden dropped out and polls that included Robert F. Kennedy Jr.


Clinton-Trump national margin (%)
8
7
6
5
4
3
2
1
0
–1
^
ACA premium hike
Post-Comey polls
election.princeton.edu
12
19
26
2
October
Nov.
End of respondent sampling period*

The Fundamental Problem of Causal Inference

For any given case, we observe the outcome variable with *either* a change in the independent variable or no change in the independent variable, but not both.

The Fundamental Problem of Causal Inference

For any given case, we observe the outcome variable with *either* a change in the independent variable or no change in the independent variable, but not both.

- In other words, we do not observe the counterfactual

The Fundamental Problem of Causal Inference

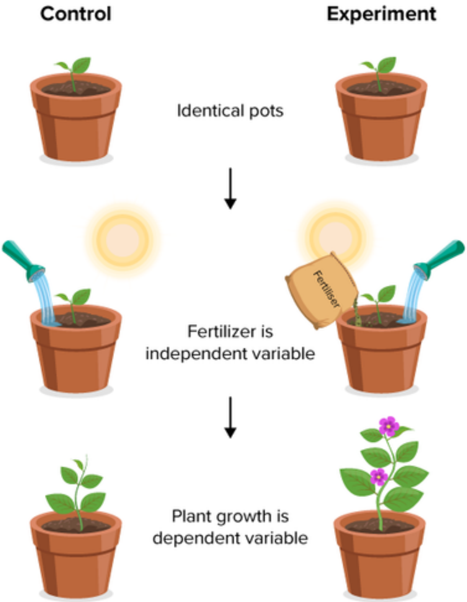
For any given case, we observe the outcome variable with *either* a change in the independent variable or no change in the independent variable, but not both.

- In other words, we do not observe the counterfactual
- But we try out best to observe the counterfactual using **experiments**

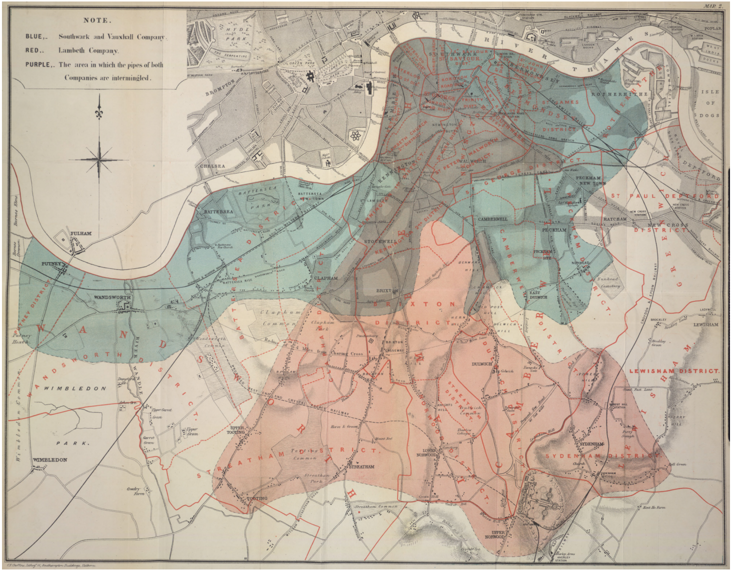
The Fundamental Problem of Causal Inference

For any given case, we observe the outcome variable with *either* a change in the independent variable or no change in the independent variable, but not both.

- In other words, we do not observe the counterfactual
- But we try out best to observe the counterfactual using **experiments**
- Experiments establish a counterfactual by comparing cases that differ only in the explanatory variable that we are interested in


Control
Experiment
Identical pots
Fertiliser
Fertilizer is
independent variable
Plant growth is
dependent variable

John Snow's Cholera Experiment



John Snow's Cholera Experiment

	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423	1,422	59

Experiments

- A critical element of experiments is that the cases are assigned to the **treatment** group (e.g., getting a drug) and **control** group (e.g., getting a placebo) completely at random

Experiments

- A critical element of experiments is that the cases are assigned to the **treatment** group (e.g., getting a drug) and **control** group (e.g., getting a placebo) completely at random
- Recall the definition of a **random sample**: the probability of any given unit being drawn from the population is uniform (the same)

EXPERIMENTS

- A critical element of experiments is that the cases are assigned to the **treatment** group (e.g., getting a drug) and **control** group (e.g., getting a placebo) completely at random
- Recall the definition of a **random sample**: the probability of any given unit being drawn from the population is uniform (the same)
- The intuition for randomness is that there is no **selection bias**: patients aren't getting the drug because they are younger or healthier, for example.

Experiments

- But wait: were households in London assigned to either the Southwark & Vauxhall Company or the Lambeth Company completely at random?

Experiments

- But wait: were households in London assigned to either the Southwark & Vauxhall Company or the Lambeth Company completely at random?
- No. This is called a natural experiment, and it relies on whether we believe being in either group is as-good-as-random.

Experiments

Card and Krueger (1994)

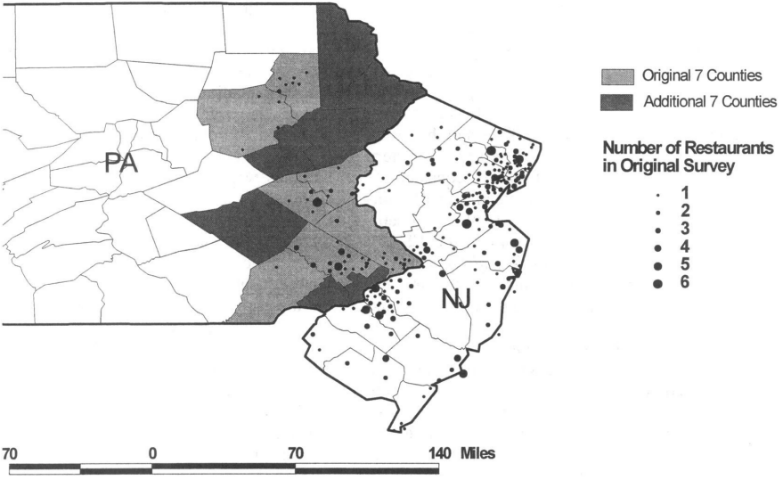


Figure 1. Areas of New Jersey and Pennsylvania Covered by Original Survey and BLS Data

Experiments

- But wait: were households in London assigned to either the Southwark & Vauxhall Company or the Lambeth Company completely at random?
- No. This is called a natural experiment, and it relies on whether we believe being in either group is as-good-as-random.
- Finding natural experiments in the real world is really difficult, so we often design our experiments in controlled settings like laboratories or surveys

Experiments

- But wait: were households in London assigned to either the Southwark & Vauxhall Company or the Lambeth Company completely at random?
- No. This is called a natural experiment, and it relies on whether we believe being in either group is as-good-as-random.
- Finding natural experiments in the real world is really difficult, so we often design our experiments in controlled settings like laboratories or surveys
- We only have a true experiment where the <u>researcher</u> randomly assigns cases to treatment or control.

Experiments

Kam and Zechmeister (2013)

Assessing Causality Without Experiments

- What about when we don't have an experiment, but instead are collecting **observational data** from the world?

Assessing Causality Without Experiments

- What about when we don't have an experiment, but instead are collecting **observational data** from the world?
- Then we face the fundamental problem of causal inference. Correlation does not imply causation.

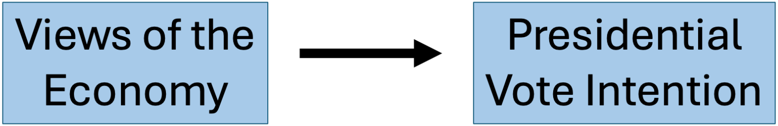
Assessing Causality Without Experiments

- What about when we don't have an experiment, but instead are collecting **observational data** from the world?
- Then we face the fundamental problem of causal inference. Correlation does not imply causation.
- But all hope is not lost - we just have to be much more careful before making causal claims.

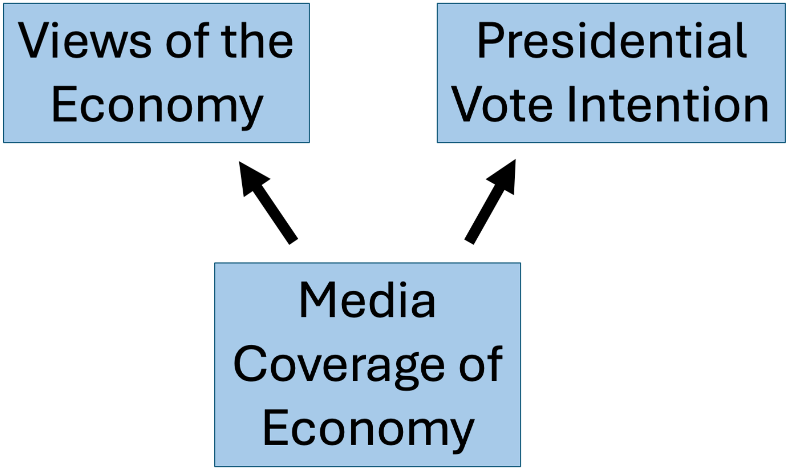
Assessing Causality Without Experiments

- Consider possible **confounders**: other variables that could explain the change in both the explanatory variable and the outcome variable

Assessing Causality Without Experiments



Assessing Causality Without Experiments



Assessing Causality Without Experiments

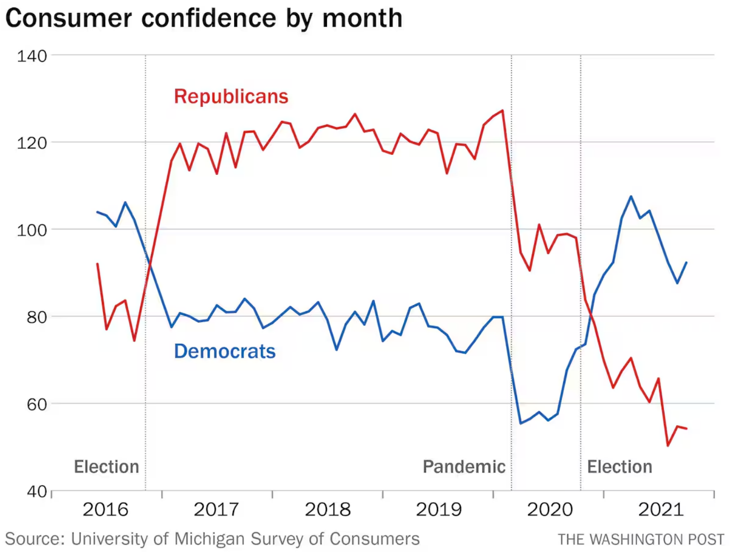
- Consider possible **confounders**: other variables that could explain the change in both the explanatory variable and the outcome variable
- Stronger associations are less susceptible to confounding

Assessing Causality Without Experiments

- Consider possible **confounders**: other variables that could explain the change in both the explanatory variable and the outcome variable
- Stronger associations are less susceptible to confounding
- Consider possible **reverse causation**, especially where the explanatory variable does not clearly precede the outcome variable

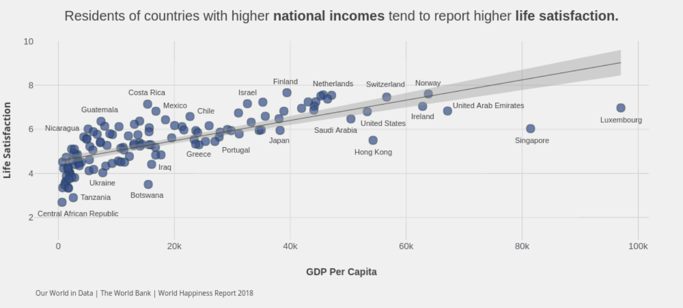
Assessing Causality Without Experiments

Assessing Causality Without Experiments

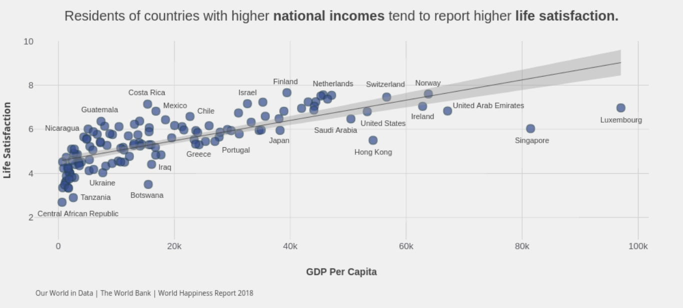


Assessing Causality Without Experiments

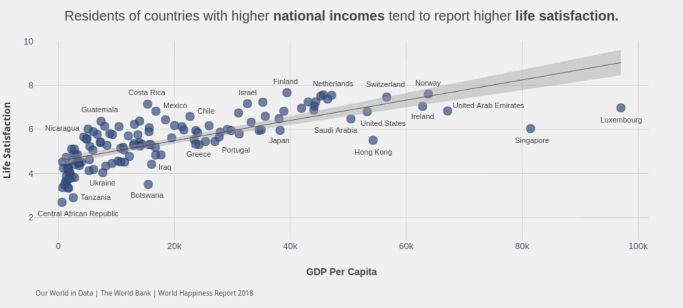
- Consider possible **confounders**: other variables that could explain the change in both the explanatory variable and the outcome variable
- Stronger associations are less susceptible to confounding
- Consider possible **reverse causation**, especially where the explanatory variable does not clearly precede the outcome variable
- Is there a plausible theory for how the explanatory variable causes the outcome variable?



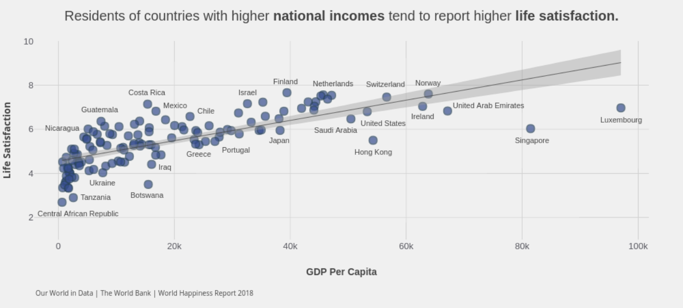
- The correlation between GDP per capita and life satisfaction could be stronger, so check for confounders



- The correlation between GDP per capita and life satisfaction could be stronger, so check for confounders
- Possible confounder: technological advancement increases economic productivity and provides entertainment



- The correlation between GDP per capita and life satisfaction could be stronger, so check for confounders
- Possible confounder: technological advancement increases economic productivity and provides entertainment
- Not clear that GDP per capita precedes happiness, so check for reverse causality



- The correlation between GDP per capita and life satisfaction could be stronger, so check for confounders
- Possible confounder: technological advancement increases economic productivity and provides entertainment
- Not clear that GDP per capita precedes happiness, so check for reverse causality
- Do happy people work harder?

Exercise: Correlation vs. Causation

Conclusion

- Causation depends on knowing a counterfactual that we cannot observe (the fundamental problem of causal inference)
- Experiments overcome this issue by comparing groups that differ only in their assignment to treatment or control (an artificial counterfactual)
- It is harder to make causal claims from observational data due to confounding and reverse causality

"CONVICTION" BY RACHEL AVIV (NEW YORKER)

Operation Hummingbird | Exhibit Ref: CEH/16 | Staff Presence Report

Chart 1: Nurses present on clinical and administrative duties Summary

				Nursing Staff on Clinical and Administrative Duties																																					
Event	Baby	Date Shift Started	Shift Type	Ailsa SIMPSON		Angela McSHANE	Ashleigh HUDSON	Belinda SIMCOCK	Bernadette BUTTERWORTH	Caroline BENNION		Caroline OAKLEY	Cheryl CUTHBERTSON-TAYLOR	Christopher BOOTH	Eirian POWELL	Elizabeth MARSHALL		Janet COX	Jean PEERS	Jennifer JONES-KEY	Joanne WILLIAMS	Kathryn WARD		Laura EAGLES	Lisa WALKER	Lucy LETBY	Mary GRIFFITH	Melanie TAYLOR	Minna LAPPALAINEN	Nicola DENNISON	Patricia STEELE	Rebecca MORGAN	Samantha O'BRIEN	Shelley TOMLINS	Sophie ELLIS		Valerie PARKES	Valerie THOMAS	Vicky BLAMIRE	Yvonne FARMER	Yvonne GRIFFITHS
1	Child A	08/06/2015	NIGHT							X						X									X	X	X	X								X					
2	Child B	09/06/2015	NIGHT										X									X			X	X	X									X					
3	Child C	13/06/2015	NIGHT													X	X									X		X		X					X						
4	Child D	21/06/2015	NIGHT								X	X				X						X				X															
5	Child E	03/08/2015	NIGHT					X				X													X	X												X			
6	Child F	04/08/2015	NIGHT					X					X													X									X	X		X			
7	Child G	06/09/2015	NIGHT	X										X									X		X	X															
8	Child G	21/09/2015	DAY							X		X			X		X	X								X					X			X			X			X	X
9	Child H	25/09/2015	NIGHT													X										X										X		X			X
10	Child H	26/09/2015	NIGHT											X		X	X									X								X							
11	Child I	30/09/2015	DAY				X		X						X		X		X						X	X		X			X						X			X	
12	Child I	12/10/2015	NIGHT				X					X						X						X		X															
13	Child I	13/10/2015	NIGHT									X						X			X			X		X								X						X	
14	Child I	22/10/2015	NIGHT				X							X												X		X										X			
15	Child J	26/11/2015	NIGHT									X												X		X	X			X									X		
16	Child J	17/12/2015	DAY												X				X		X			X		X			X		X								X		X
17	Child K	16/02/2016	NIGHT									X									X					X									X			X			
18	Child M	09/04/2016	DAY			X	X	X									X									X	X														
19	Child L	09/04/2016	DAY			X	X	X									X									X	X				X										
20	Child N	02/06/2016	NIGHT											X												X		X							X			X			
21	Child N	14/06/2016	NIGHT					X	X											X		X	X		X	X															
22	Child O	23/06/2016	DAY								X			X				X								X		X			X	X	X				X			X	X
23	Child P	23/06/2016	DAY								X			X				X								X		X			X	X	X				X			X	X
24	Child P	24/06/2016	DAY		X									X				X								X	X					X					X			X	
25	Child Q	25/06/2016	DAY														X									X	X		X	X											
			Total	1	1	2	5	5	2	2	3	7	2	7	3	5	7	6	2	1	3	3	2	4	6	25	7	7	2	3	6	3	2	3	4	4	5	6	2	6	5

*('X' indicates 'on duty' presence on the shift, where a suspicious event has been identified)