PDU Prevalence Estimation Methods

Dr Gordon Hay
Centre for Public Health
Liverpool John Moores University

g.hay@ljmu.ac.uk

Ice Breaking Exercise

Alcohol

Exercise - alcohol use

- How many people in Croatia drink?
- Street survey - ask 50 people
  - 30 people say they drink (60%)
- What would happen if 500 people were asked?
- Survey carried out at night in the centre of Zagreb - does that matter?
- What does ‘drink alcohol’ mean?

Exercise - alcohol use

- Case definition
  - Drink, drink alcohol
  - Lifetime, last year, last month
  - Recommended units, binge drinking
  - Frequency
  - Under-age drinking

Exercise - alcohol use

- Representative sample
  - Zagreb, Croatia
  - Age, gender, ethnic group
  - Employed / unemployed
  - ‘Hard to reach groups’
    - Prisoners, homeless

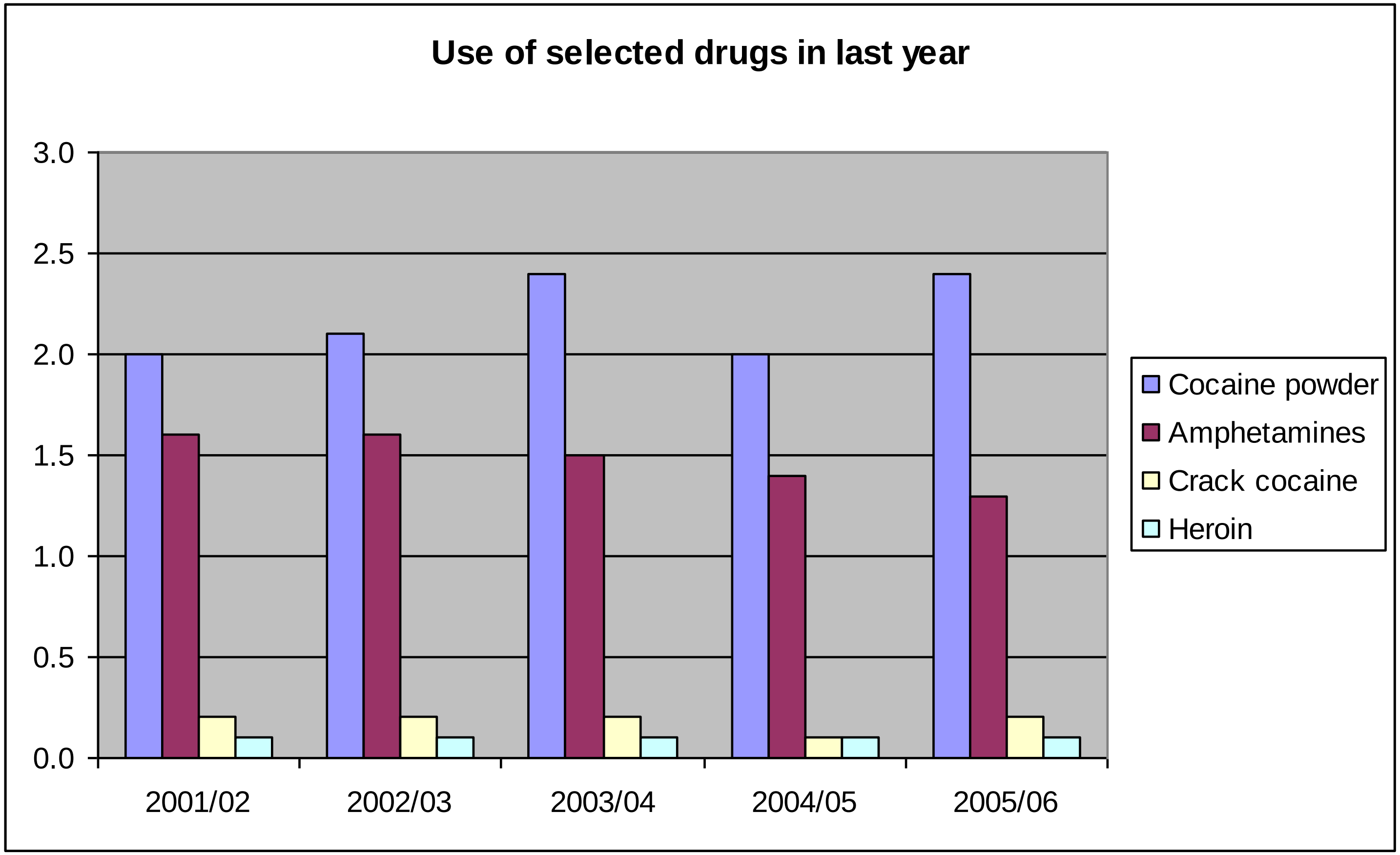
Exercise - alcohol use

- Sample size
  - Should not affect the prevalence rate
  - Can improve the reliability of the estimate

General principle

- Drug use is largely a hidden activity
- Information can be obtained from a sample of the population
- This information can be extrapolated to provide information on the entire population

British Crime Survey



British Crime Survey

- 0.1% of the population used heroin in last year (aged 15 to 59)
- Population of England
  - 31,000,000 (aged 15 to 59)
- 31,000 people in England have used heroin in previous year

British Crime Survey

- Was the sample representative?
- Were respondents 'honest'
- What would the confidence interval be?

Indirect Methods

- Multiplier methods
- Capture-recapture methods
- Multiple indicator methods
- Truncated Poisson

Multiplier Methods

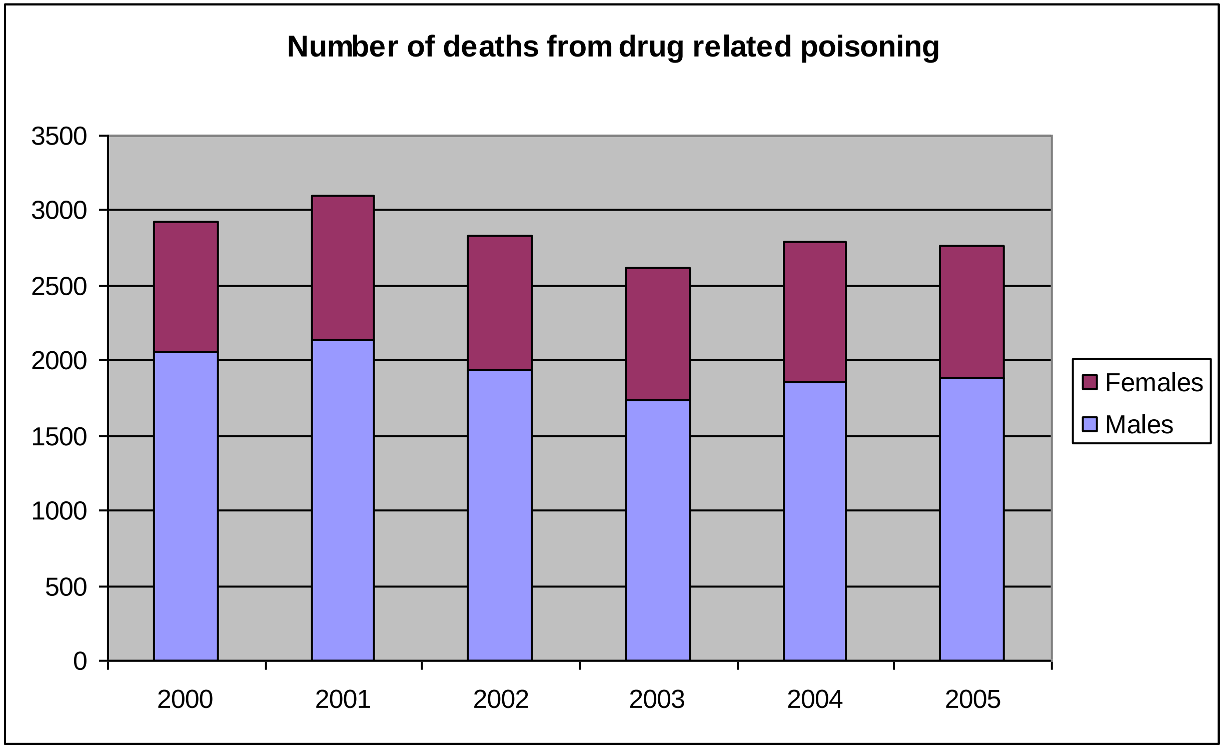
- Information can be obtained from a sample of drug users
  - Contact with treatment services
  - Mortality
- This information can be extrapolated to provide information on all drug users

Multiplier Methods (2)

- Benchmark
  - Number of drug users in treatment
  - Number of drug-related deaths
    - Published mortality statistics
- Multiplier
  - In-treatment rate
  - Mortality rate
    - Anecdotal evidence (between 1% and 2%)
    - Specific studies

Drug related death data

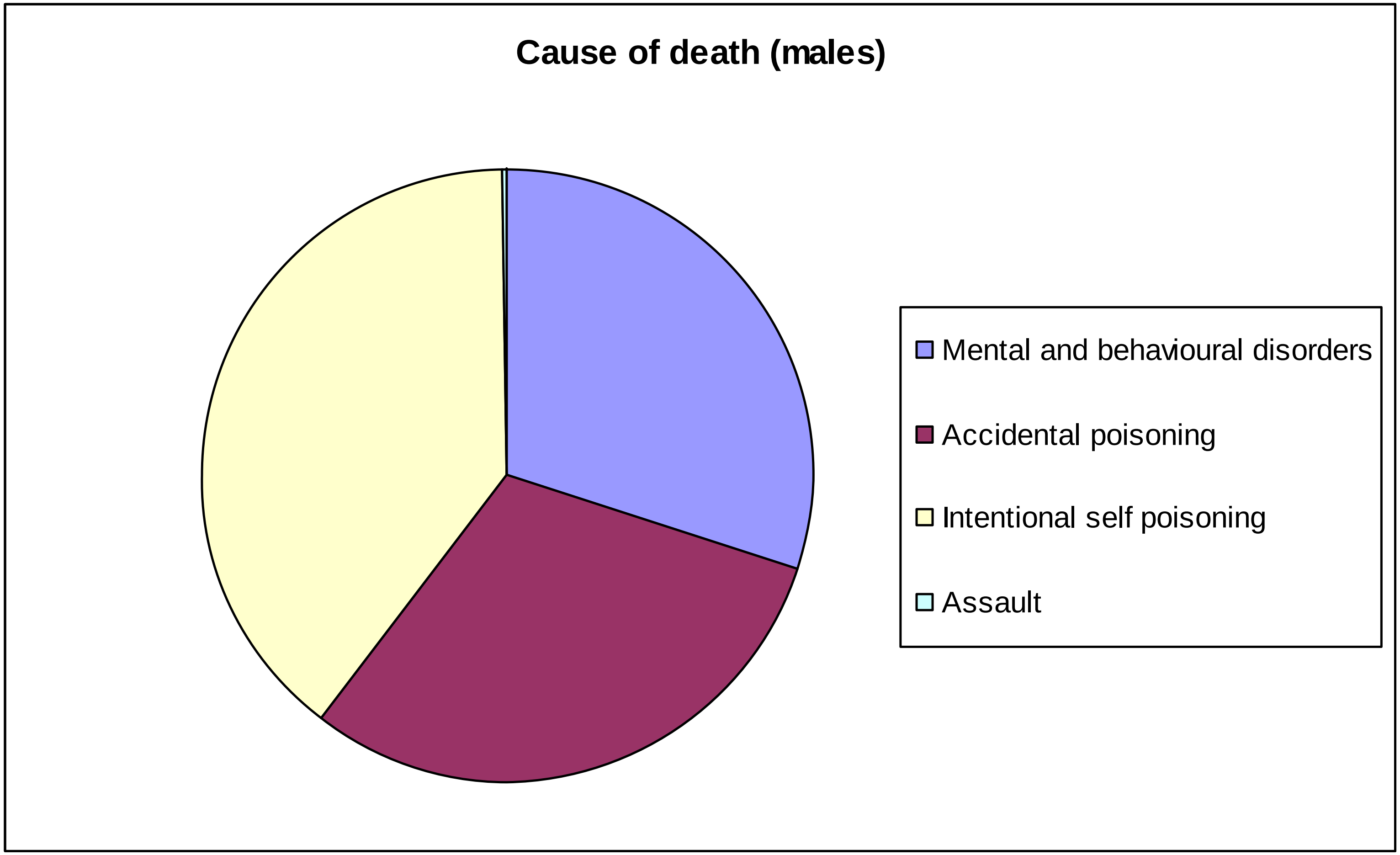
Source: Health Statistics Quarterly



2,762 deaths in 2005

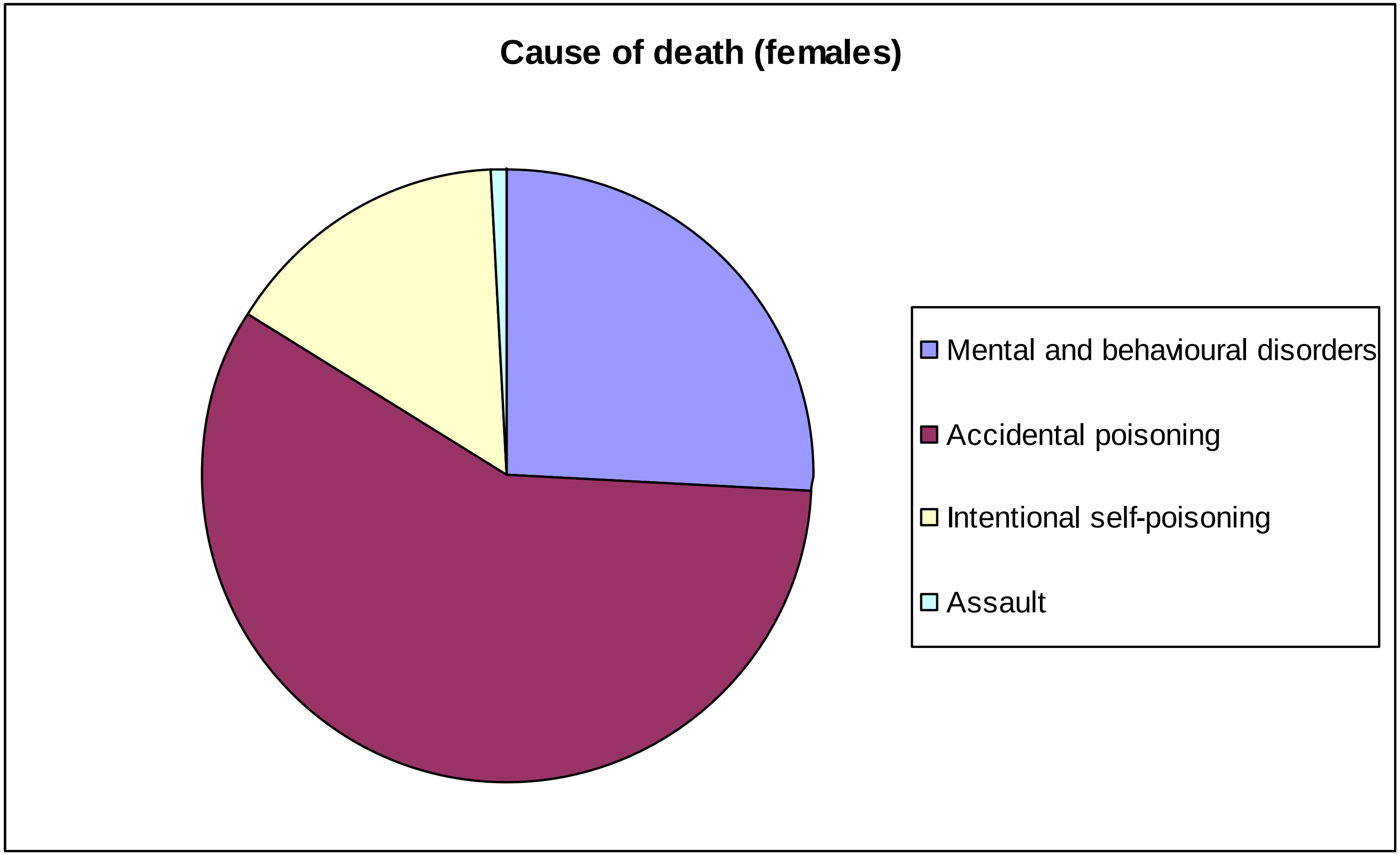
Drug related death data

Source: Health Statistics Quarterly



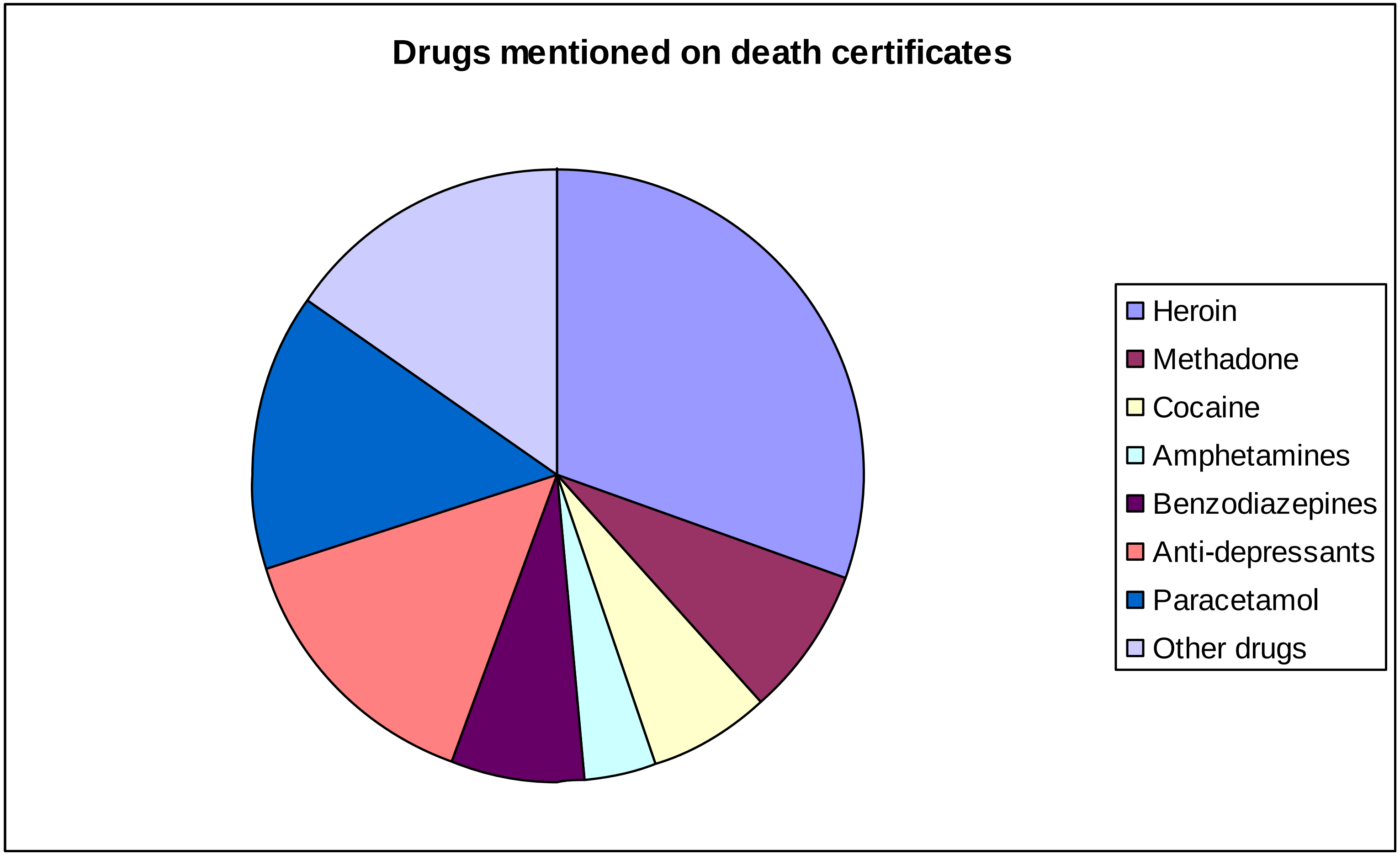
Drug related death data

Source: Health Statistics Quarterly



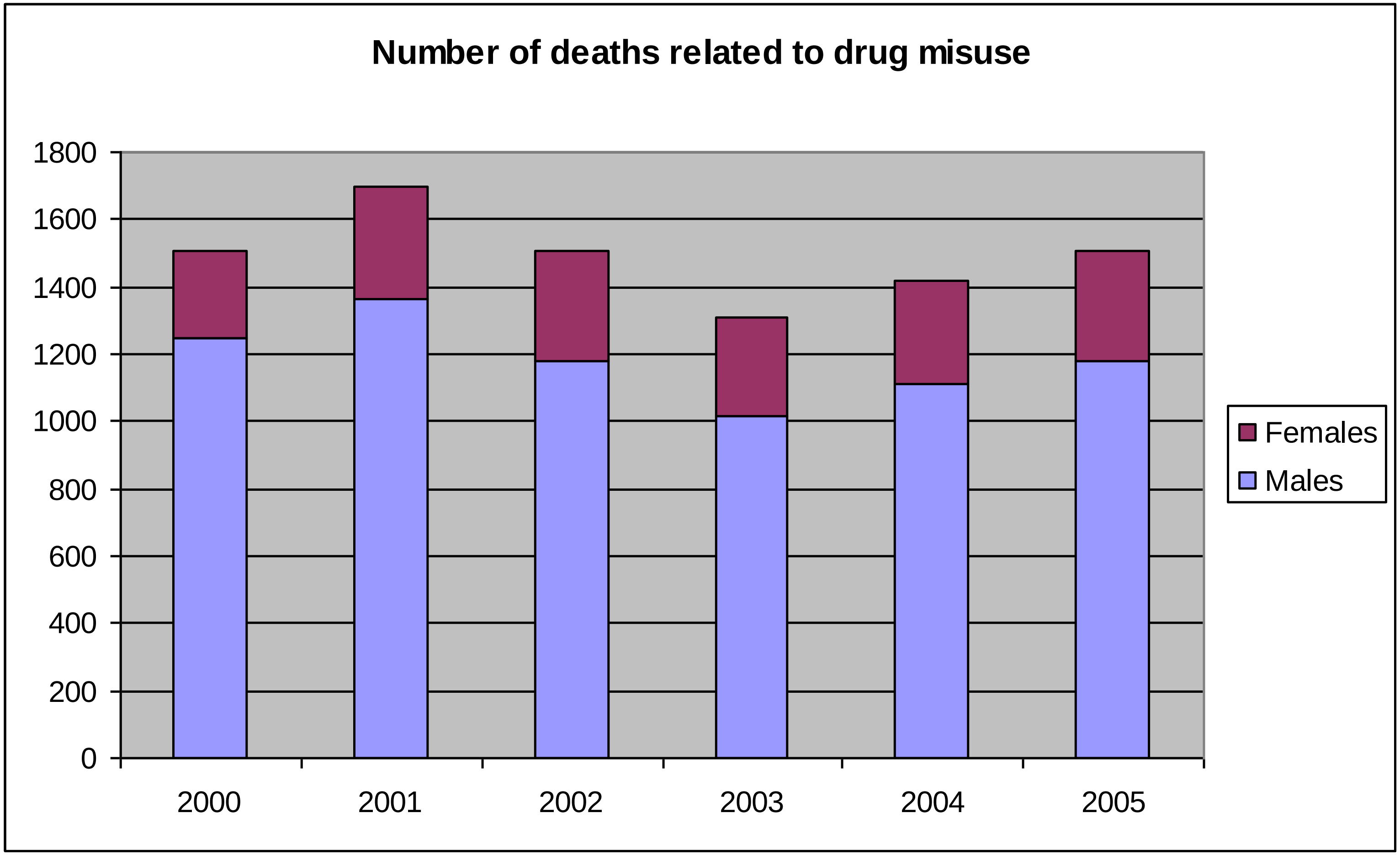
Drug related death data

Source: Health Statistics Quarterly



Drug related death data

Source: Health Statistics Quarterly



1,506 deaths in 2005

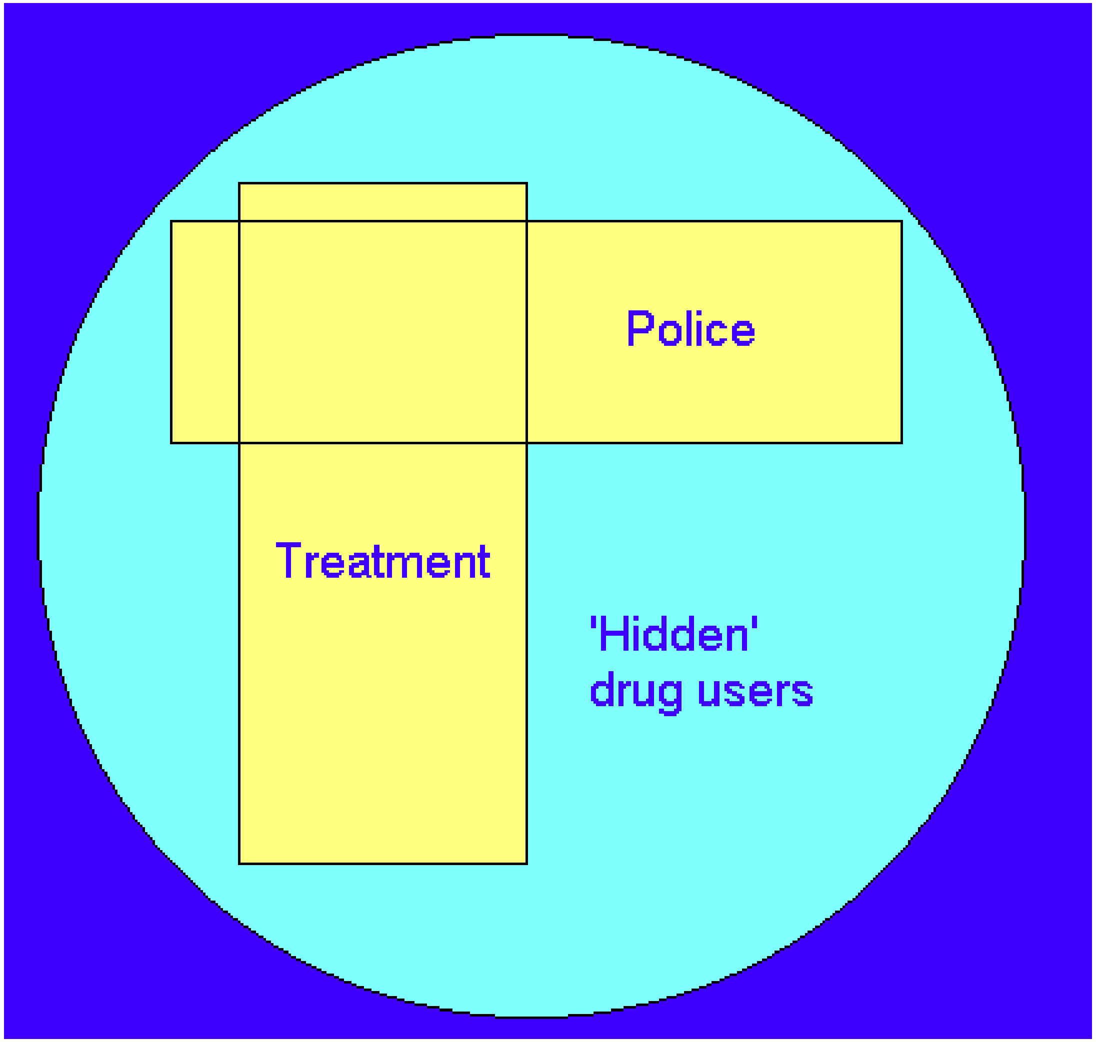
Exercise

Mortality Multipliers

Two sample capture-recapture method

- Simple concept:
  - Only a certain proportion of drug users are in contact with treatment agencies
- Examine the overlap between those in treatment and a second sample (e.g. Police)
- Find the proportion in treatment
- Thus estimate the total number of drug users

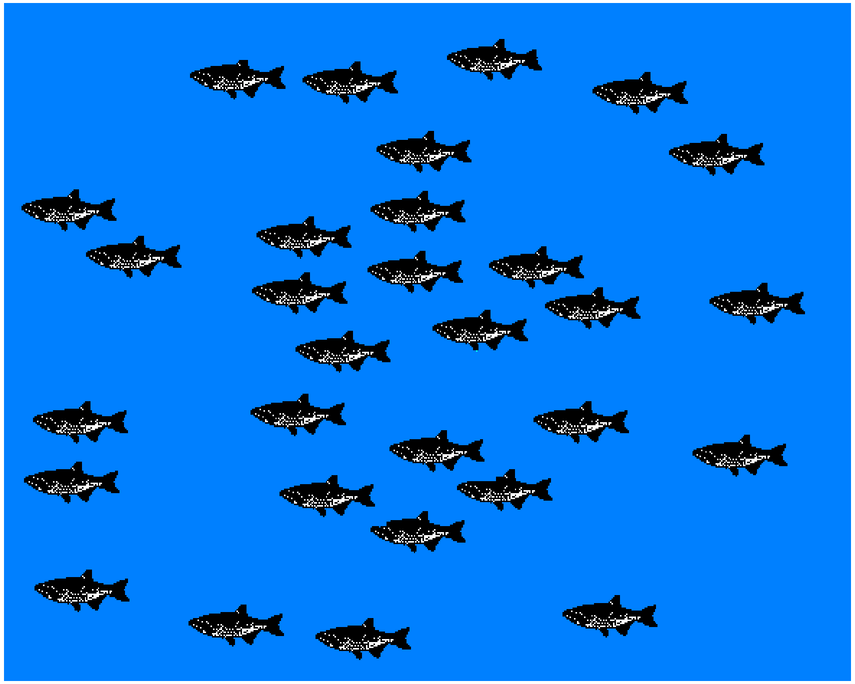
Two sample capture-recapture method



Two sample capture-recapture method

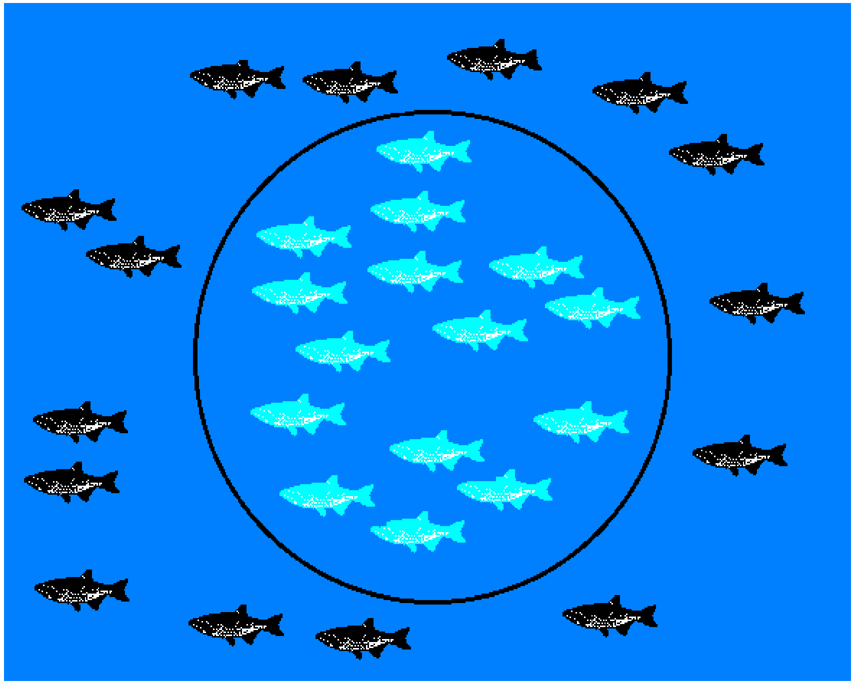
- Animal populations
  - Capture a sample of fish
  - Mark them
  - Release them
  - Recapture a sample at a later date
  - Look for marks
  - Estimate population size

Example - fish



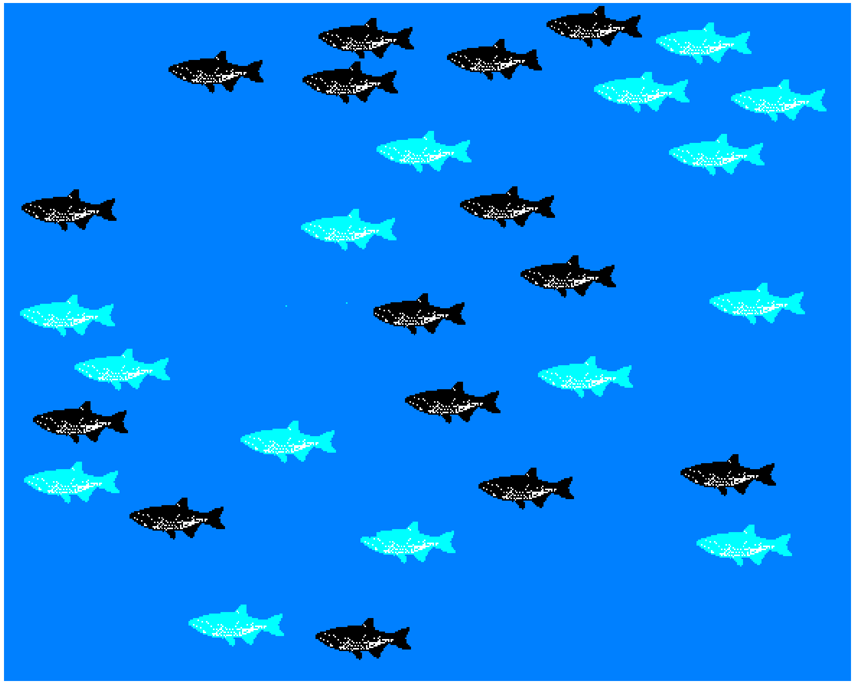
- Unknown number of fish in a lake

Example - fish



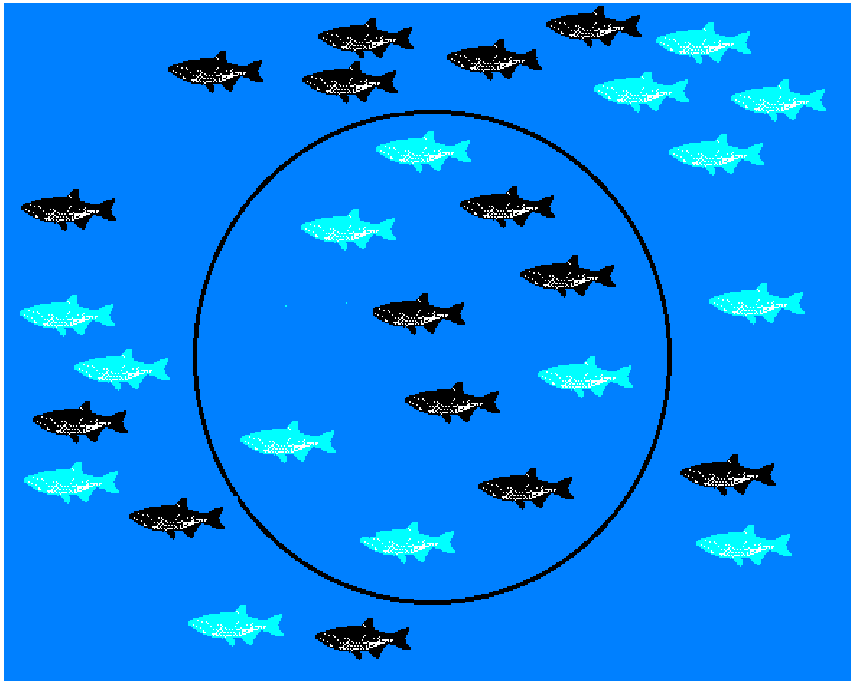
- Unknown number of fish in a lake
- Catch a sample and mark them

Example - fish



- Unknown number of fish in a lake
- Catch a sample and mark them
- Let them loose

Example - fish



- Unknown number of fish in a lake
- Catch a sample and mark them
- Let them loose
- Recapture a sample and look for marks

Estimate population size

n_1 = number in first sample 15

n_2 = number in second sample 10

n_{12} = number in both samples 5

N = total population size

assume that

$n_1/N = n_{12}/n_2$ therefore $15/N = 5/10$

$N = (10 \times 15) / 5 = 30$

Two sample capture-recapture (drug use)

- Drug users
  - Identify two samples
    - Treatment agencies
    - Police
  - Find overlap
  - Estimate population size

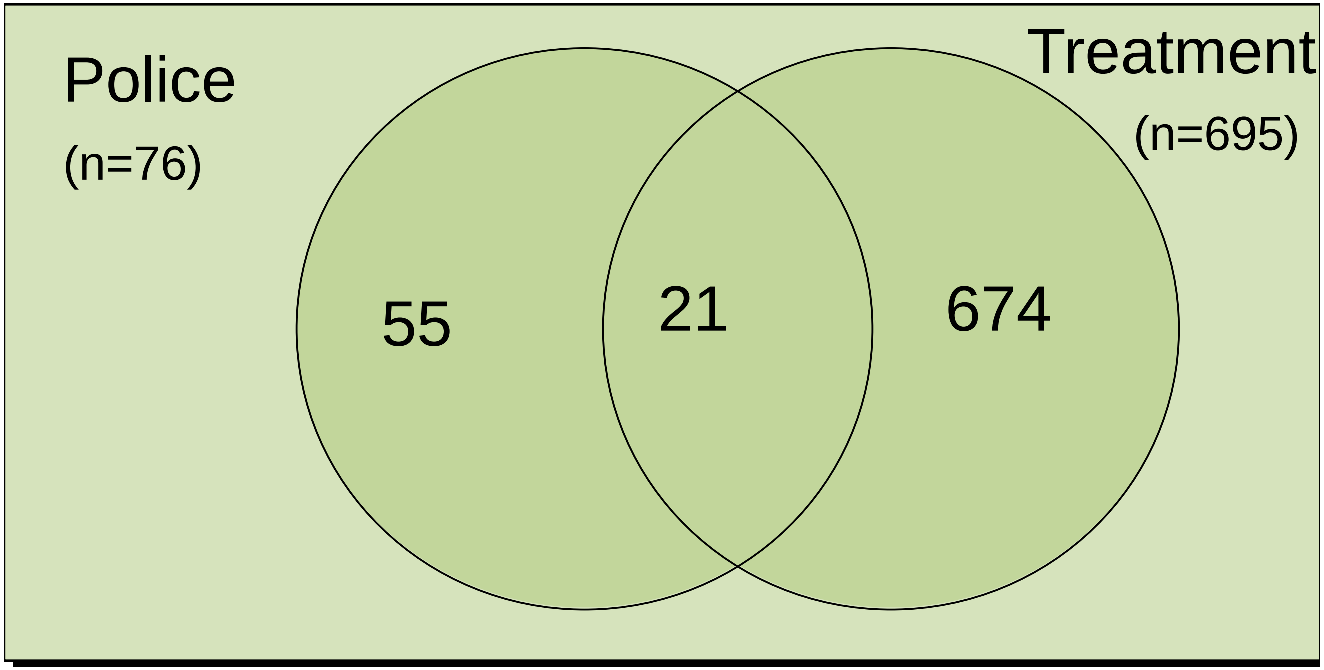
Drug use example

	N
Treatment	695
GPs (family doctors)	148
HIV Test Data	46
Police	76

Computer-based exercise

Find overlap between Treatment and Police Samples

Overlap between Police and treatment



Main assumption

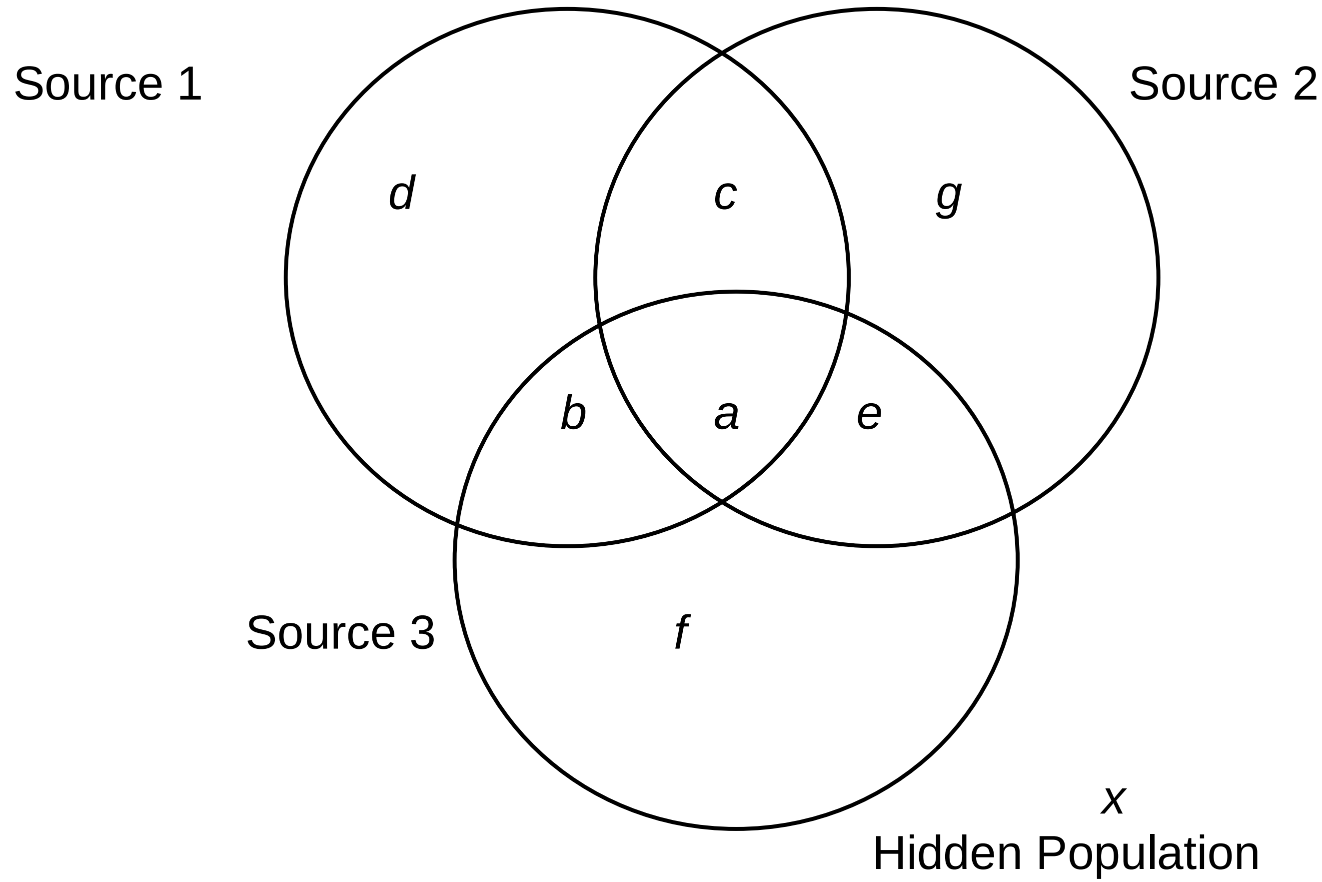
- Samples are independent
  - Police do not stand outside agency arresting people
  - Participation in treatment does not reduce the need to commit crimes

- Samples are often not independent
- Can use a third samples to correct for lack of independence or account for any relationships

Three-sample method

- Statistical analysis
  - Computer package (e.g. SPSS)
  - Log-linear models
  - Explain relationship between sources
- Estimate the size of the hidden population
- Estimate the total population size

Three-sample overlaps
Venn Diagram



Three-sample overlaps
Contingency table

		Source 1			
		Present		Absent	
		Source 2			
		Present	Absent	Present	Absent
Source 3	Present	*a*	*b*	*e*	*f*
	Absent	*c*	*d*	*g*	*x*

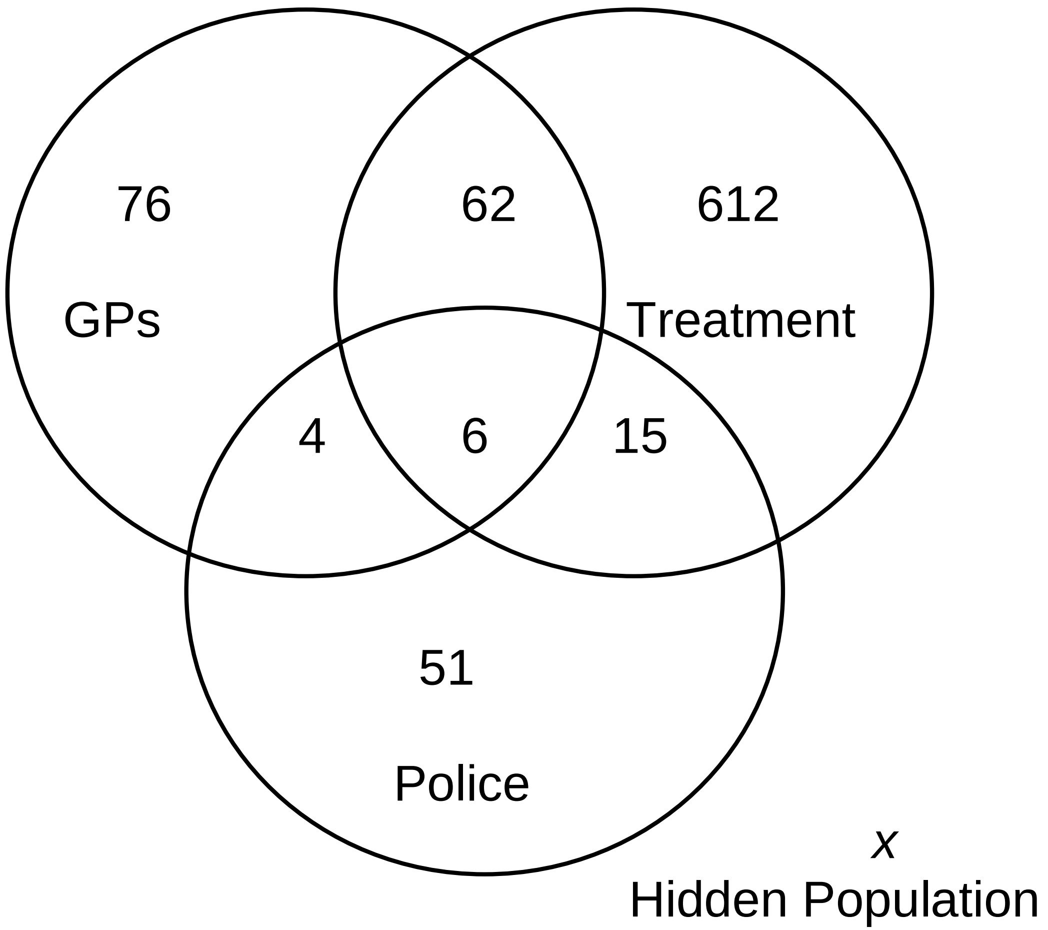
Three-sample overlaps
Data table

Source 1	Source 2	Source 3	Count
1	1	1	*a*
1	1	0	*c*
1	0	1	*b*
1	0	0	*d*
0	1	1	*e*
0	1	0	*g*
0	0	1	*f*
0	0	0	*x*

Computer-based exercise

Find overlap pattern between Treatment, Police and GP data sources

Overlap between treatment, GP and Police data



Contingency table

		Treatment			
		Present		Absent	
		GPs			
		Present	Absent	Present	Absent
Police	Present	6	15	4	51
	Absent	62	612	76	-

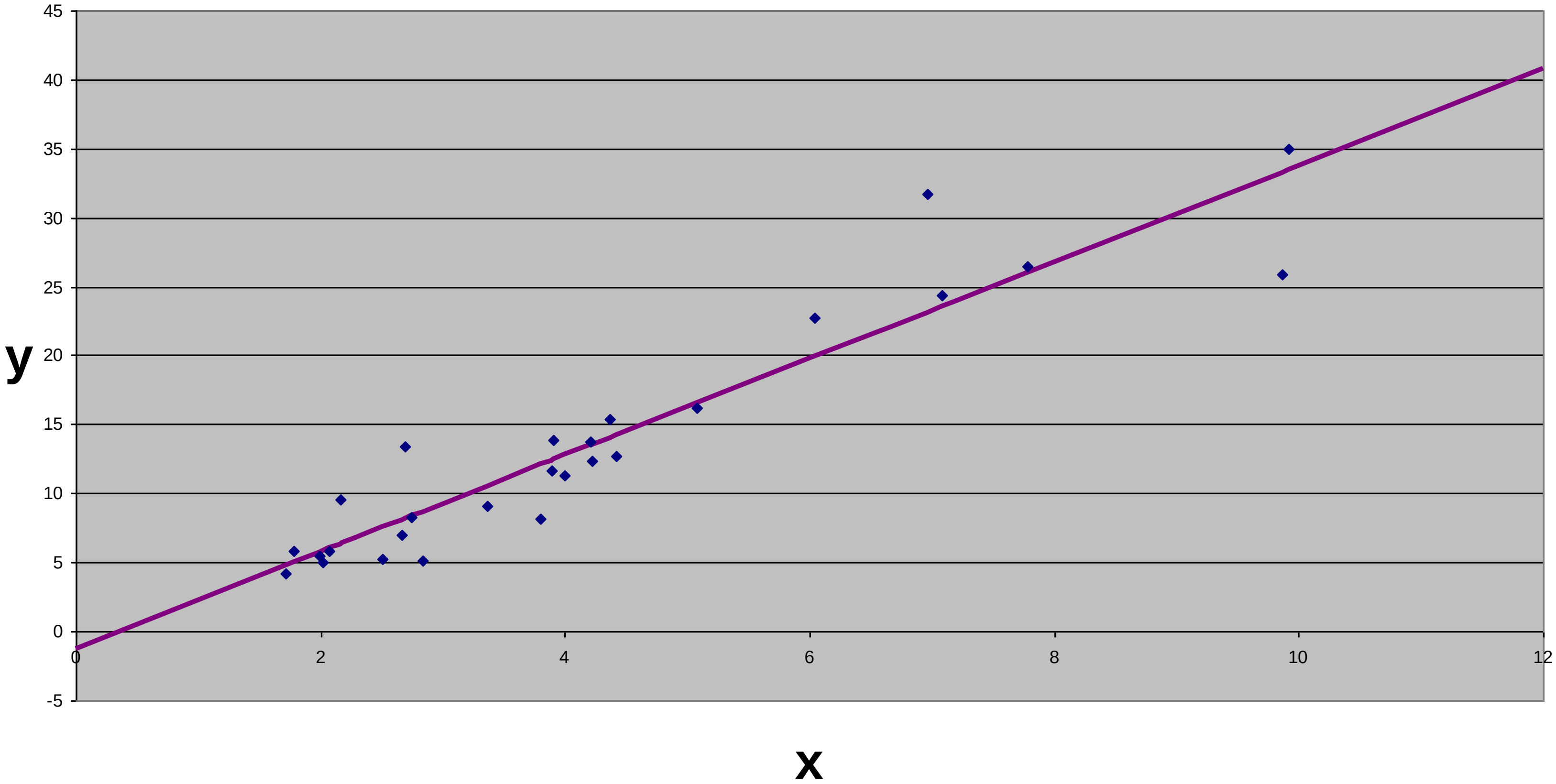
Data table

Treatment	GPs	Police	Count
1	1	1	6
1	1	0	62
1	0	1	15
1	0	0	612
0	1	1	4
0	1	0	76
0	0	1	51
0	0	0	-

Linear Regression (review)

- What is regression?
- What is a dependent variable?
- What are explanatory variables?

Prevalence v Treatment



$$y = ax + c$$

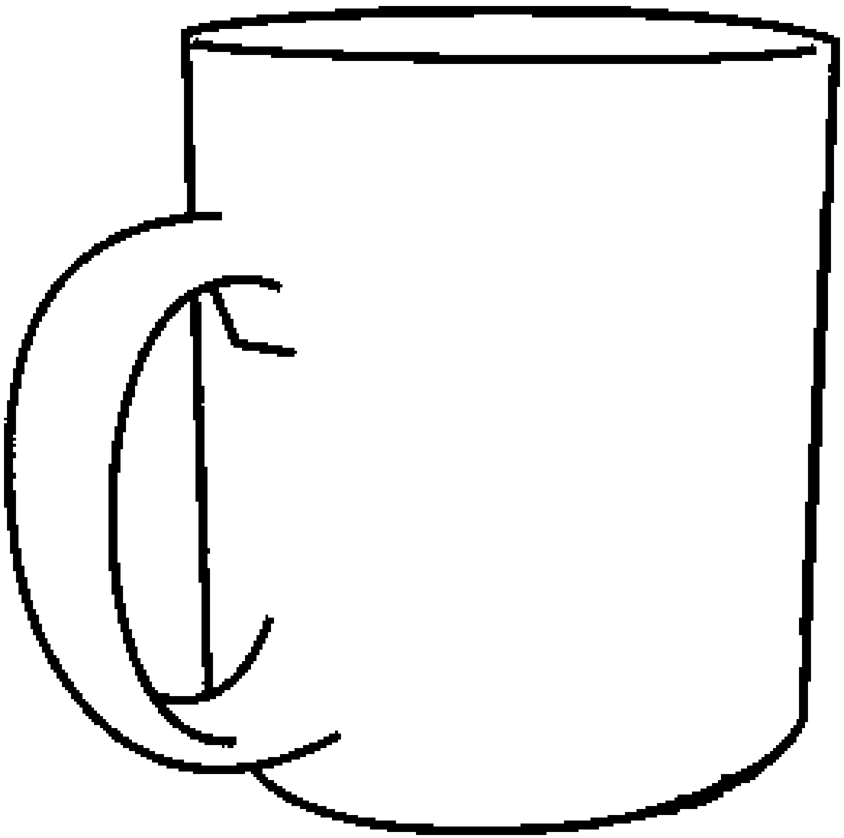
Linear Regression (examples)

$$y = ax + c$$

$$y = a_1x_1 + a_2x_2 + c$$

$$y = a_1x_1 + a_2x_2 + a_3x_3 + c$$

Linear Regression (example)



- Sales of mugs in 120 areas
- How much does advertising costs, number of shops and number of vouchers account for the variation in sales?

Worked Example

Linear regression

Sales of mugs

Linear Regression (issues)

- Model Fitting
- Goodness of fit
- Predicted value
- Confidence interval

Linear Regression

Model	Sales = -29.43 + 0.42(shops) + 2.54(vouch) + 1.02(ads)
R Square	0.626
Predicted Value	15
Confidence Interval	-4 to 33

Log-linear Regression

- Equation for linear regression

$$y = a_1x_1 + a_2x_2 + a_3x_3 + c$$

- Equation for log-linear regression (independence model)

$$\log(y) = \log(x_1) + \log(x_2) + \log(x_3) + \log(c)$$

Computer-based exercise

Fit the independence model to the three sample data

Log-linear Regression

- How realistic is it to assume all sources are independent?

- Possible interactions

- How many interactions are there when there are three sources?
  - FLIPCHART

Log-linear Regression

Models:

- constant+p1+p2+p3
- constant+p1+p2+p3+p1*p2
- constant+p1+p2+p3+p1*p3
- constant+p1+p2+p3+p2*p3
- constant+p1+p2+p3+p1*p2+p1*p3
- constant+p1+p2+p3+p1*p2+p2*p3
- constant+p1+p2+p3+p1*p3+p2*p3
- constant+p1+p2+p3+p1*p2+p1*p3+p2*p3

Computer-based exercise

Fit the other models to the three sample data

3-sample capture-recapture results

Model	Est	Lower	Upper	Deviance	df
Const+p1+p2+p3	921	699	1214	13.78	3
Const+p1+p2+p3+p1*p2	1530	943	2482	6.91	2
Const+p1+p2+p3+p1*p3	716	514	996	6.52	2
Const+p1+p2+p3+p2*p3	966	726	1286	11.72	2
Const+p1+p2+p3+p1*p2+p1*p3	969	342	2748	6.12	1
Const+p1+p2+p3+p1*p2+p2*p3	2081	1164	3721	0.85	1
Const+p1+p2+p3+p1*p3+p2*p3	750	531	1059	5.39	1
Const+p1+p2+p3+p1*p2+p1*p3 +p2*p3	3598	912	14201	0.00	0

Log-linear Regression

- What's the best estimate?
  - Deviance / likelihood ratio
    - degrees of freedom
  - Confidence intervals
  - Credibility

Key assumptions

- Population is closed
- Perfect matching
- Data sources should be representative
- Everyone has the same chance of appearing in any individual data source
- Presence in one source does not influence presence in another
  - Can be relaxed with log-linear models

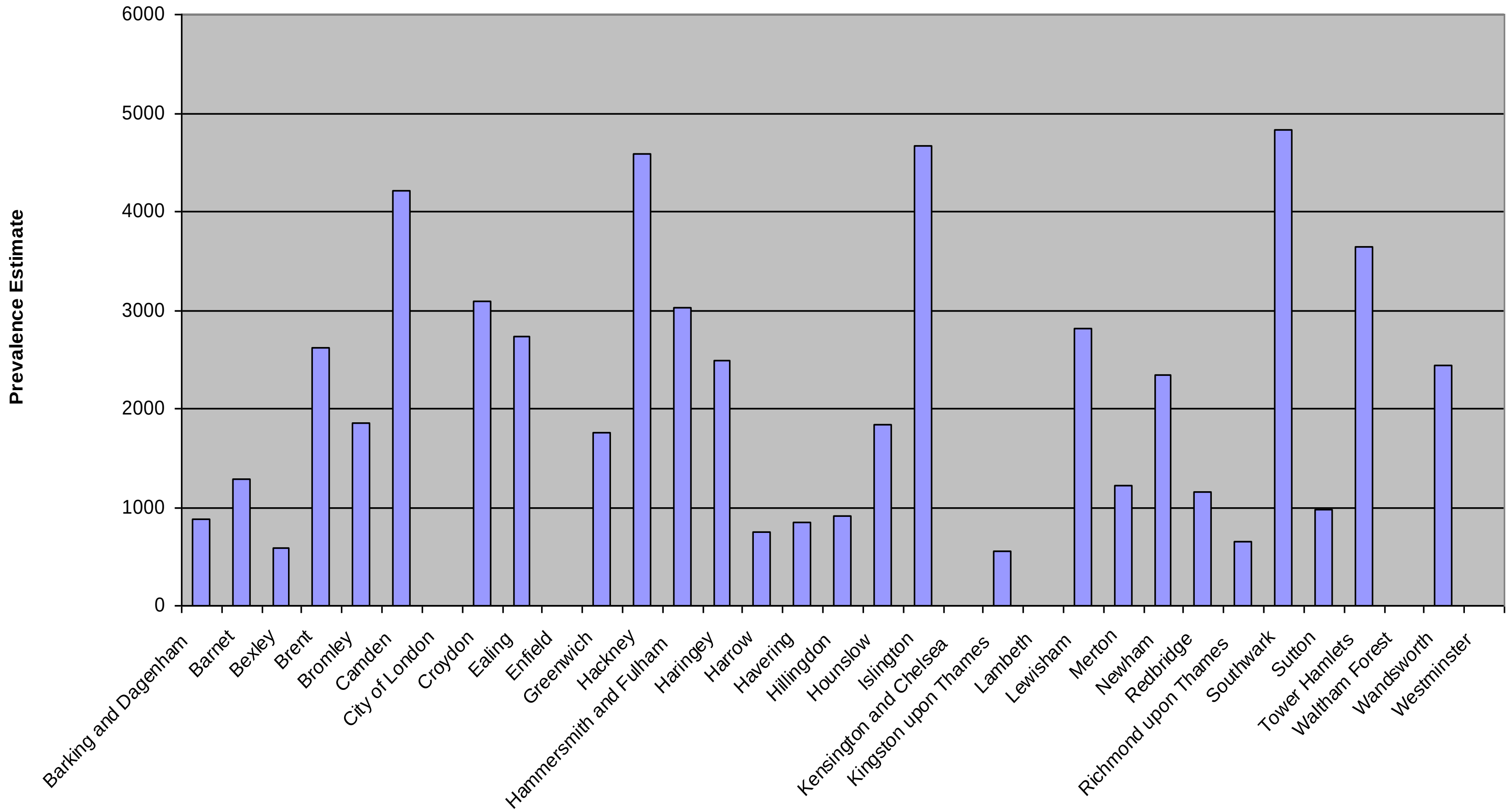
MIM

- Multivariate Indicator Method / Multiple Indicator Method
- Regression
  - Linear regression
  - Model selection

London

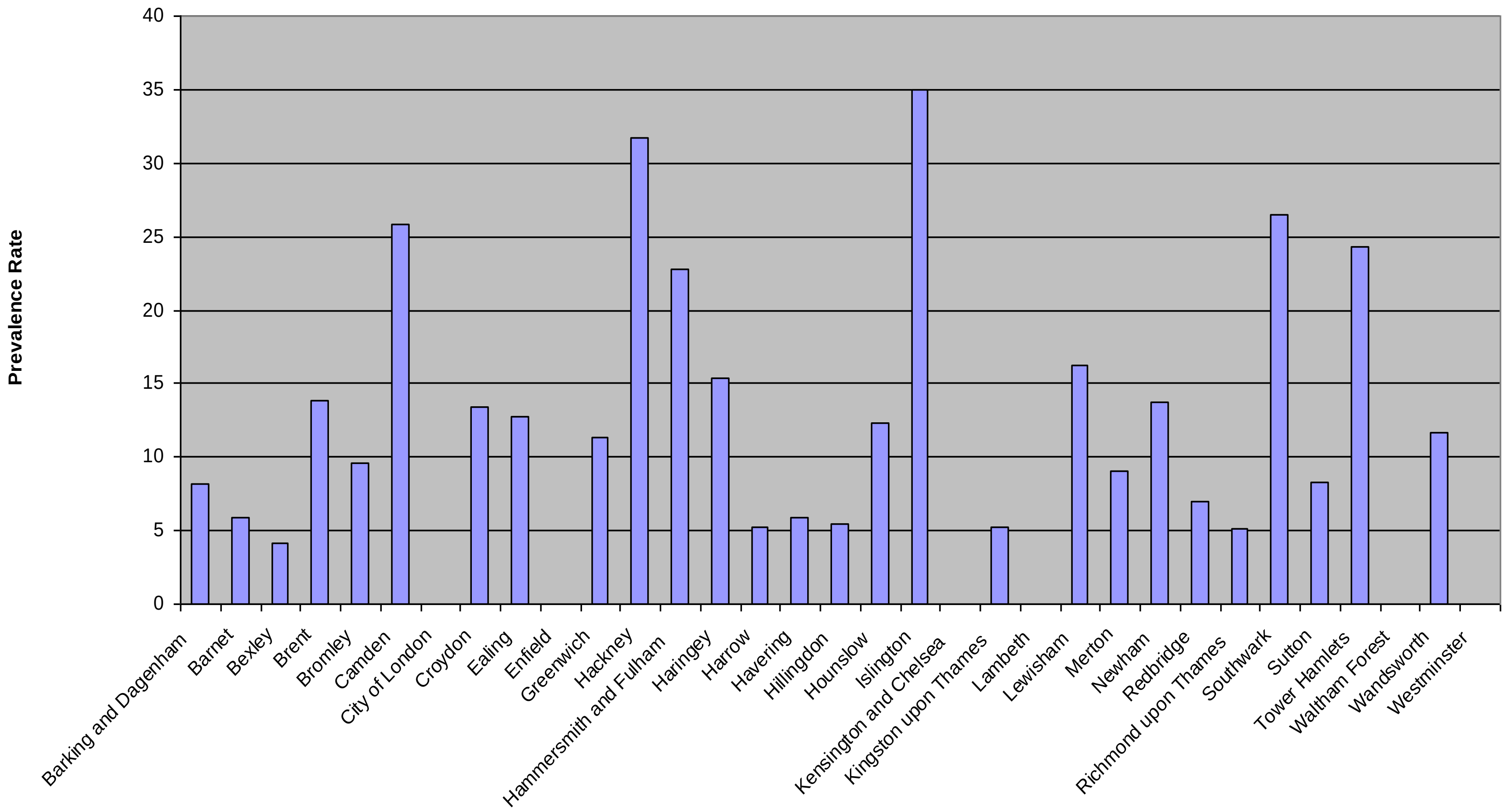
- 33 Drug Action Team (DAT) areas
- 2004/05 data
  - 27 capture-recapture estimates
  - 6 DAT areas where the capture-recapture analysis was not ‘good enough’
  - Need to ‘extrapolate’ to get estimates for those areas

London CRC Estimates

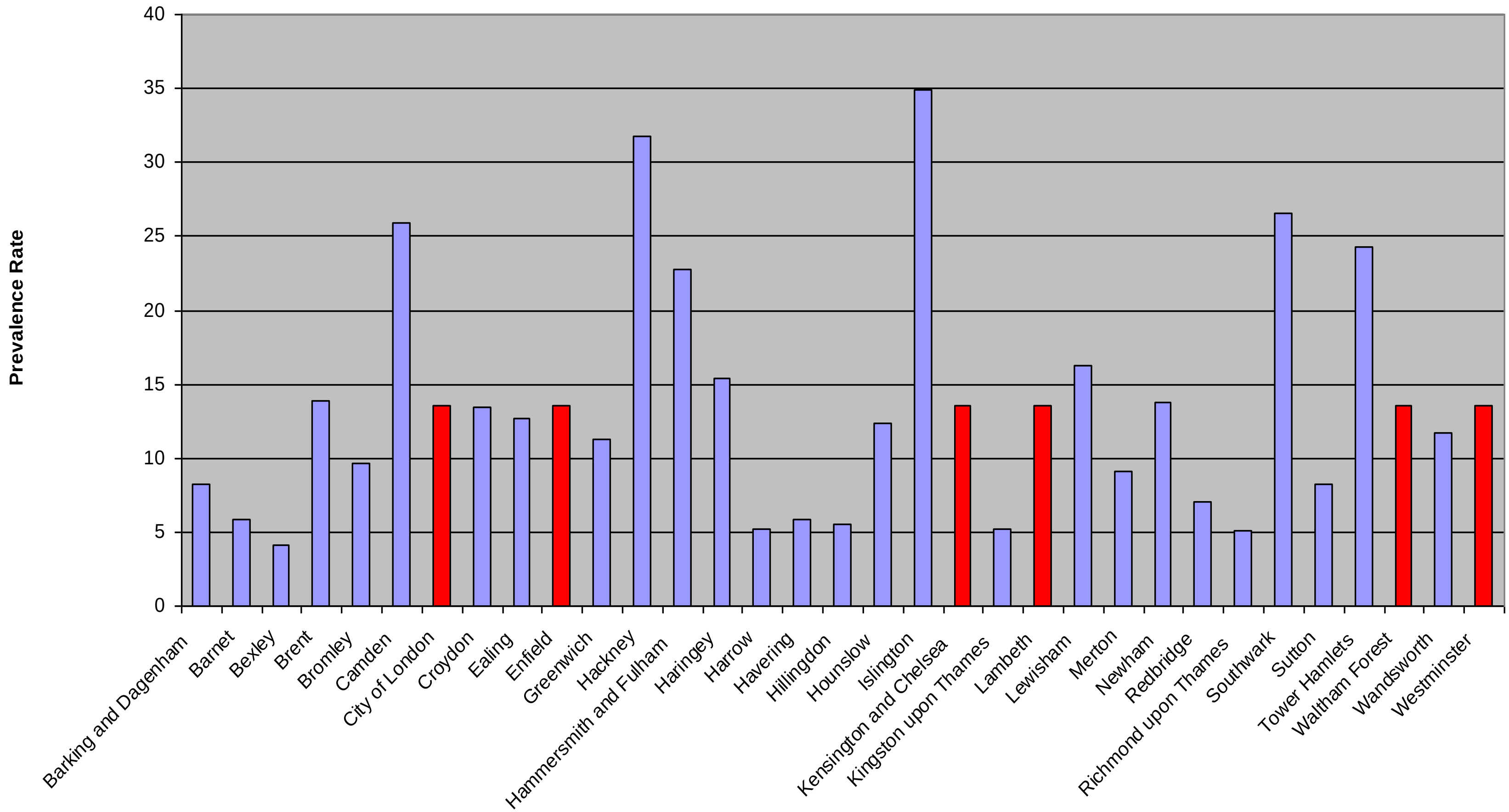


London DAT estimates

rates per 1,000



London DAT estimates

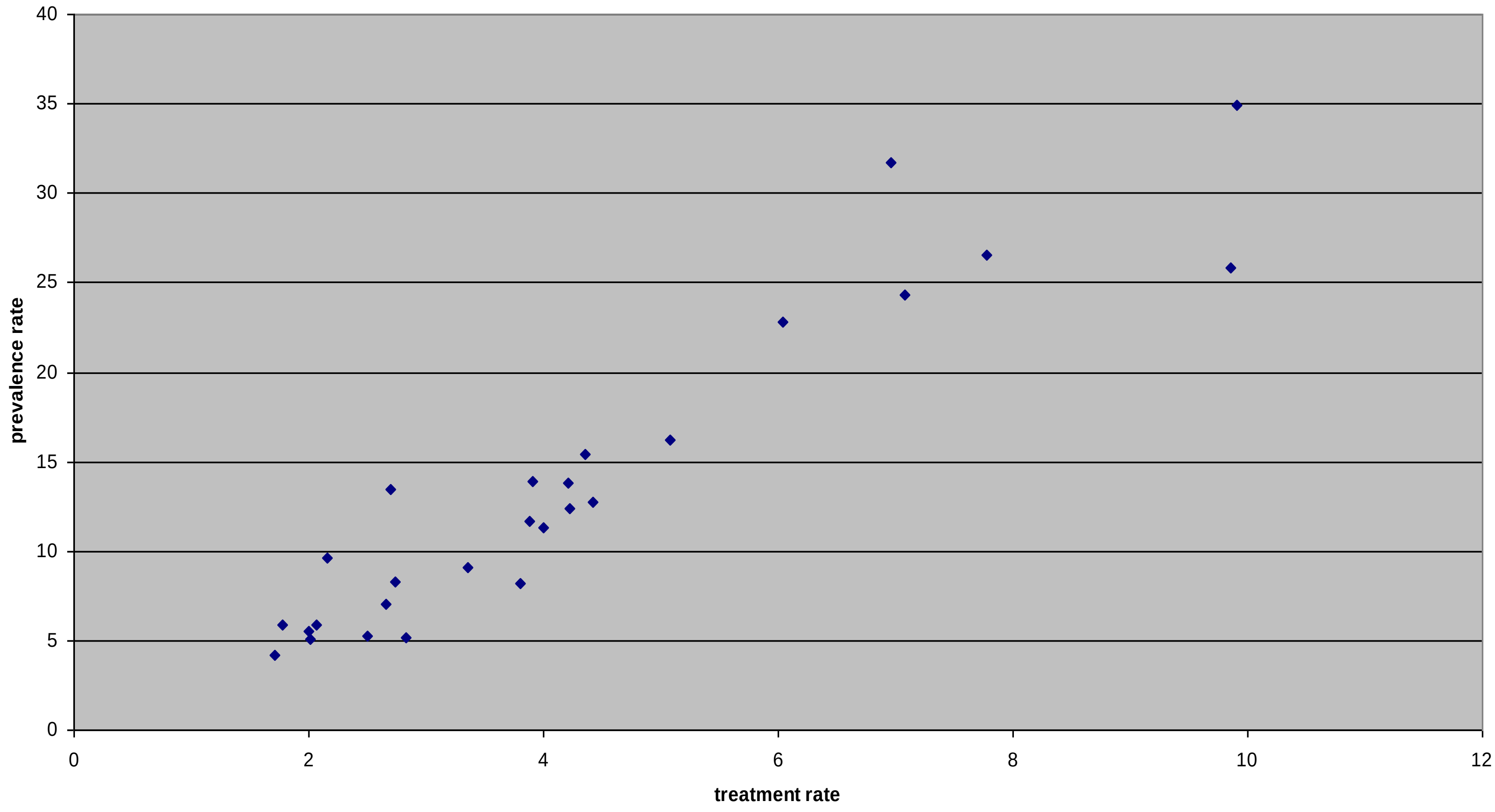


Extrapolation (regression)

$$prevalence = constant$$

$$y = c$$

Prevalence v Treatment



Extrapolation (regression)

prevalence = *constant*

$$y = c$$

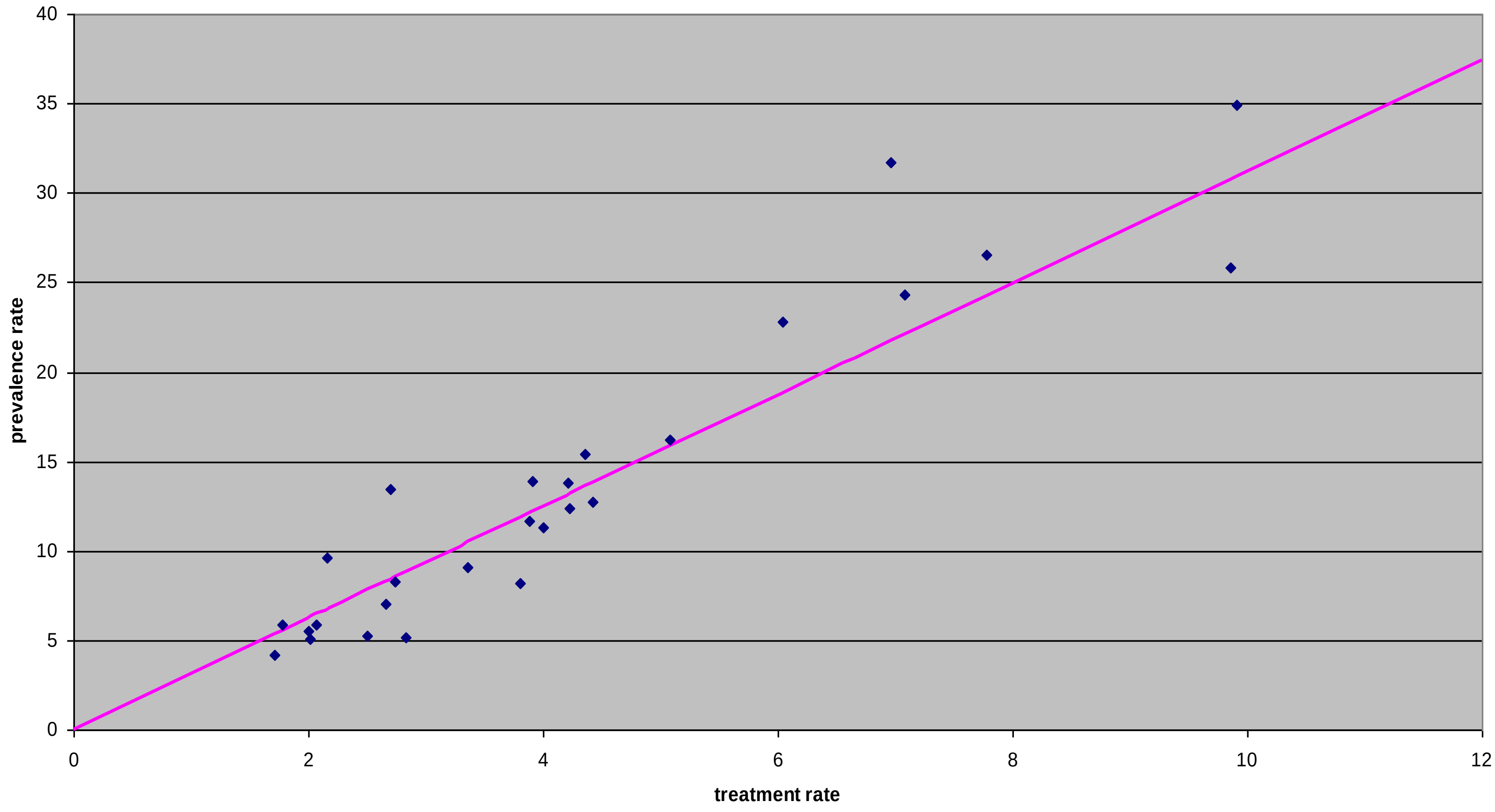
prevalence = *constant* x *treatment*

$$y = ax$$

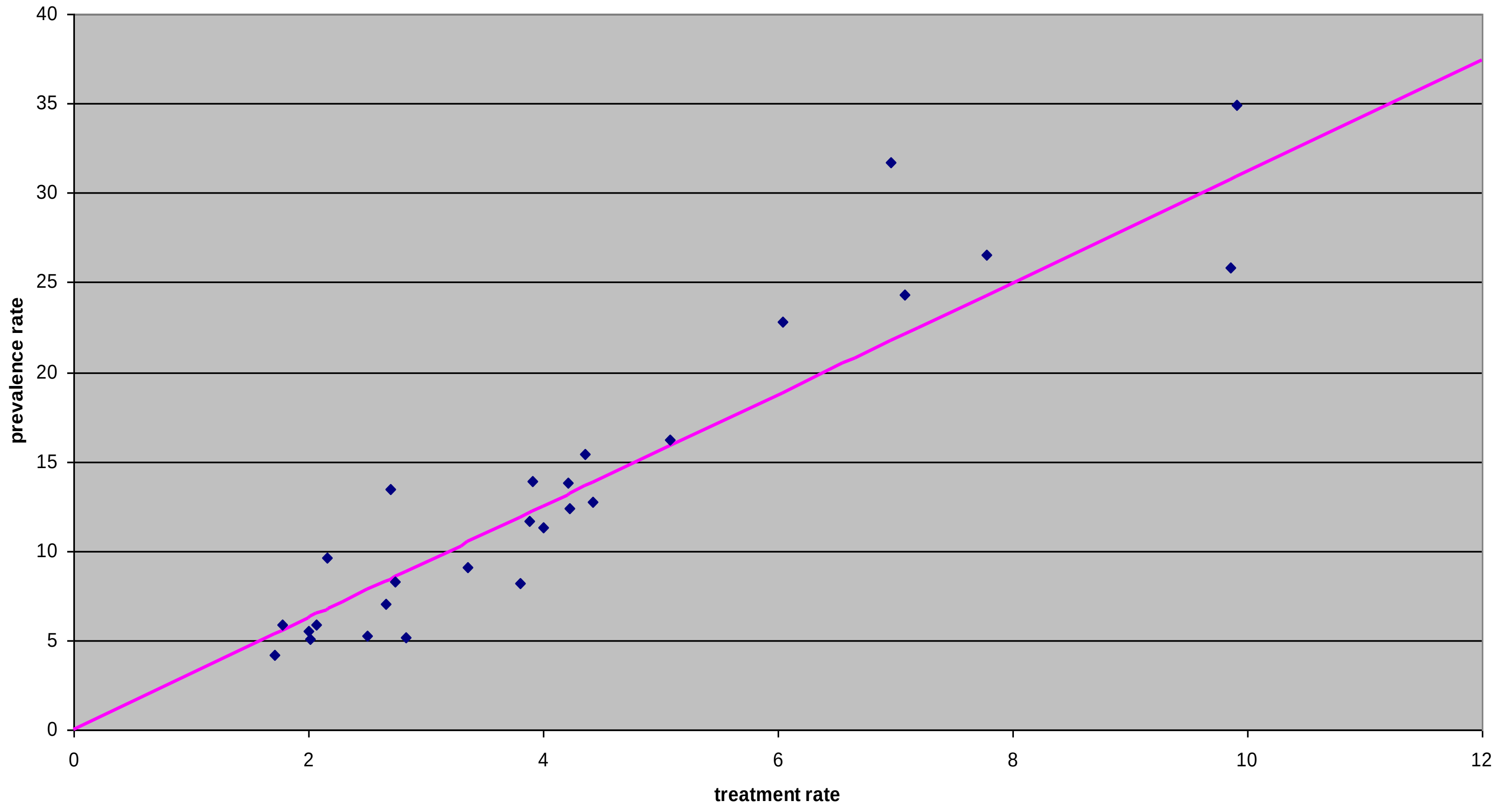
Computer-based exercise

What would be a ‘treatment’ multiplier for London

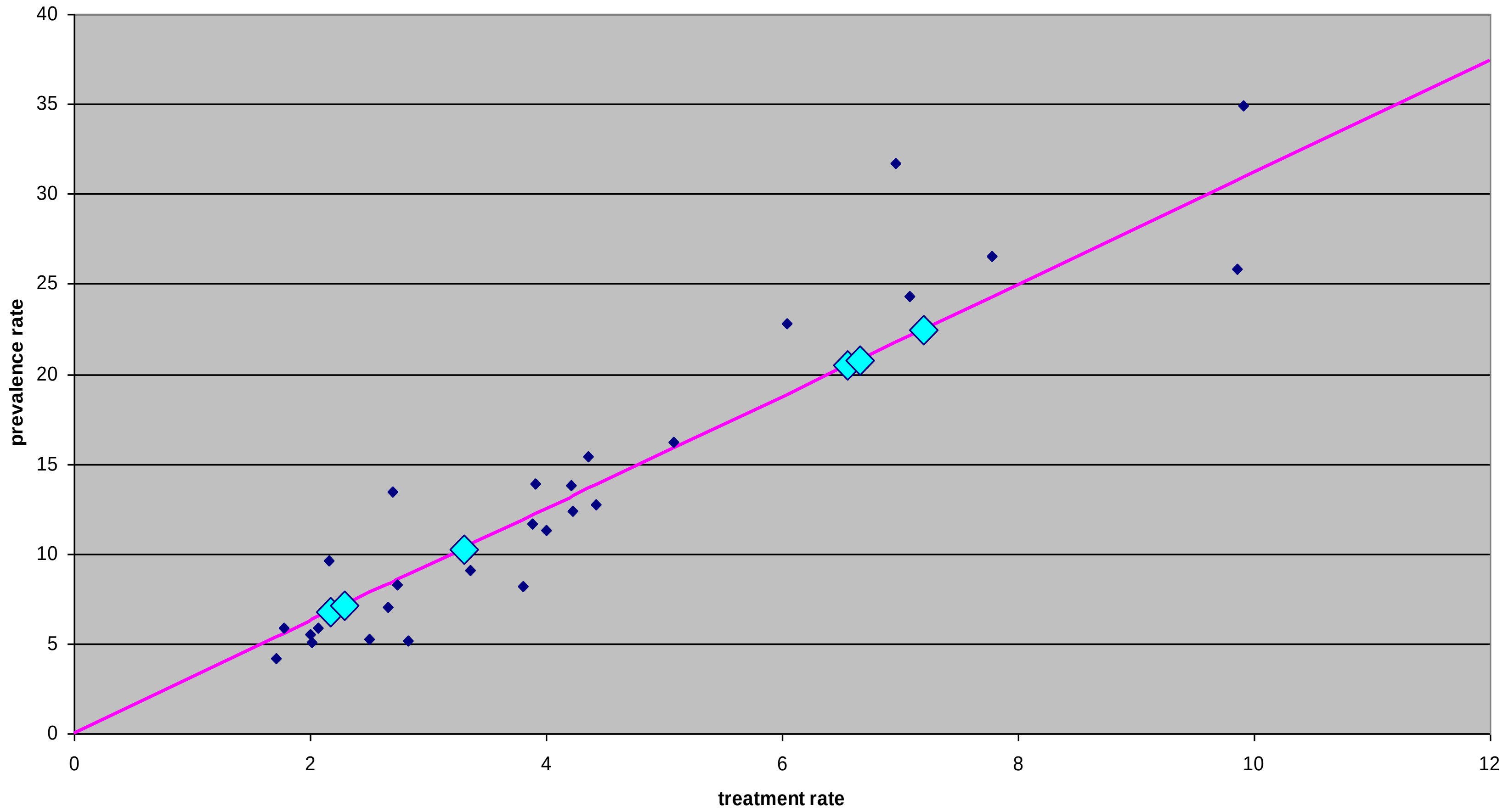
Prevalence v Treatment



Prevalence v Treatment



Prevalence v Treatment



Computer-based exercise

What would be a simple regression model for London

Extrapolation (regression)

$$prevalence = constant$$

$$y = c$$

$$prevalence = constant \times treatment$$

$$y = ax$$ ← *treatment multiplier*

MIM →

$$y = ax + b$$

$$y = a_1x_1 + a_2x_2 + b$$

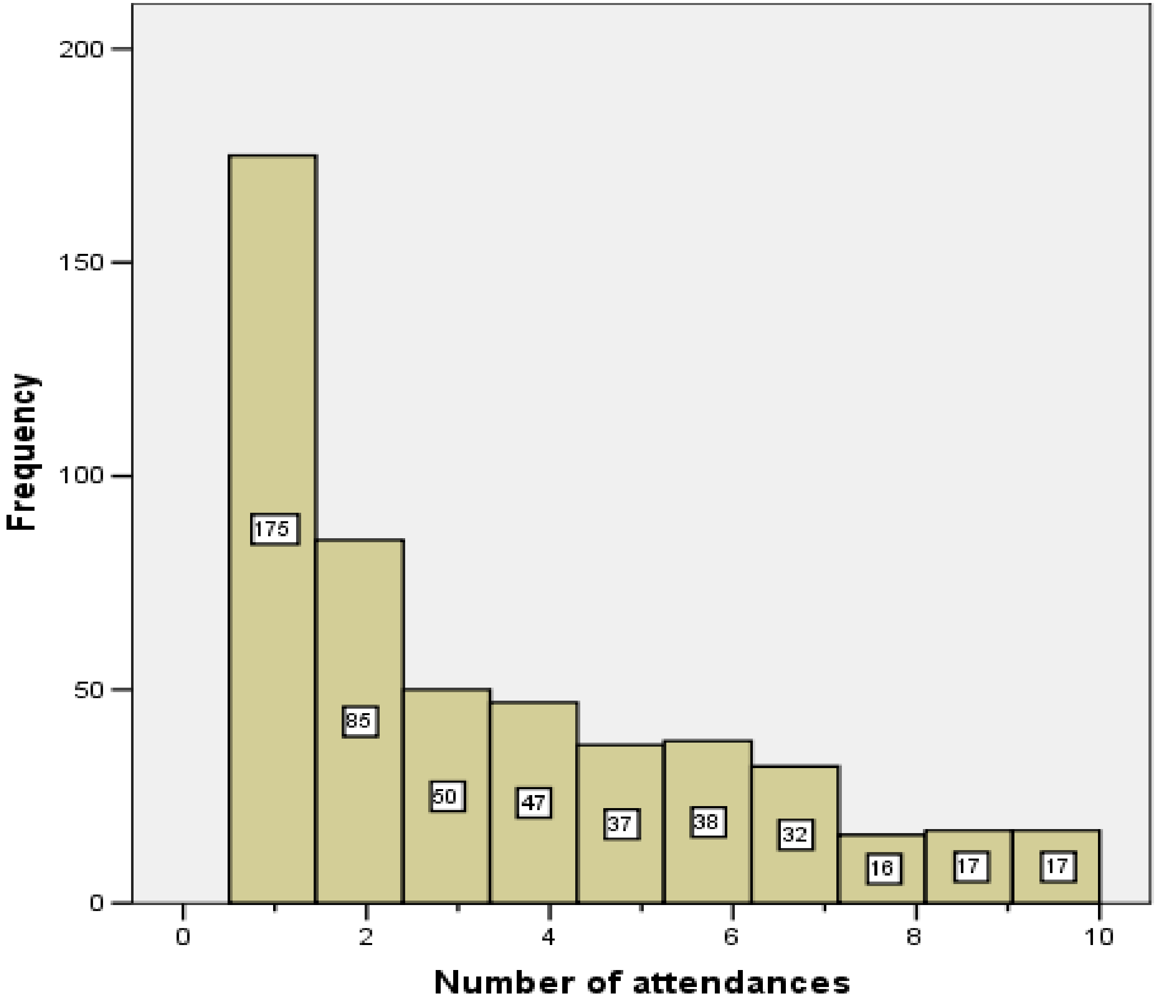
$$y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$$

MIM

- How many indicators to put into model
  - All of them?
  - All that initially seem sensible?
  - Only those that are statistically significant?
- Data reduction
  - Small number of anchor points
  - Principal component analysis
  - Reduces many indicators into 1 or 2 factors

Truncated Poisson

- Can be used with data from only one source
  - Needle exchange visits
- Count how many people have visited
  - Once
  - Twice
- Count the total number of people
- Can estimate the number of people who have visited zero time = hidden population



Frequency
200
150
100
50
0
175
85
50
47
37
38
32
16
17
17
0
2
4
6
8
10
Number of attendances

Truncated Poisson

$$est(n) = S / [\, 1 - \exp(-2\, f_2 / f_1) \,]$$

Where

f_1 = number of people attending only once

f_2 = number of people attending twice

S = total number of people attending

Review

- Introduction
- Two sample capture-recapture analysis
- Using Excel to find overlap patterns
- Three sample capture-recapture analysis
- Multiple Indicator Method
- Truncated Poisson method

Comments?
Questions?

g.hay@ljmu.ac.uk