Friday Morning

An Introduction to longitudinal mixture modelling

What are mixtures?

- Latent (unmeasured) subgroups in the population
- Also known as latent classes

- Mixture modelling is usually a data driven technique
- Can be used to
  - Explain relationships between a set of binary or continuous measures
  - Explain skewness/bimodality in a single cts measure

Mixture models for today

- Single continuous measure
  - GHQ

- Multiple continuous measures
  - Extending Latent Growth Models to GMM
    - Bodyweight example from yesterday
    - Repeated measures of SDQ

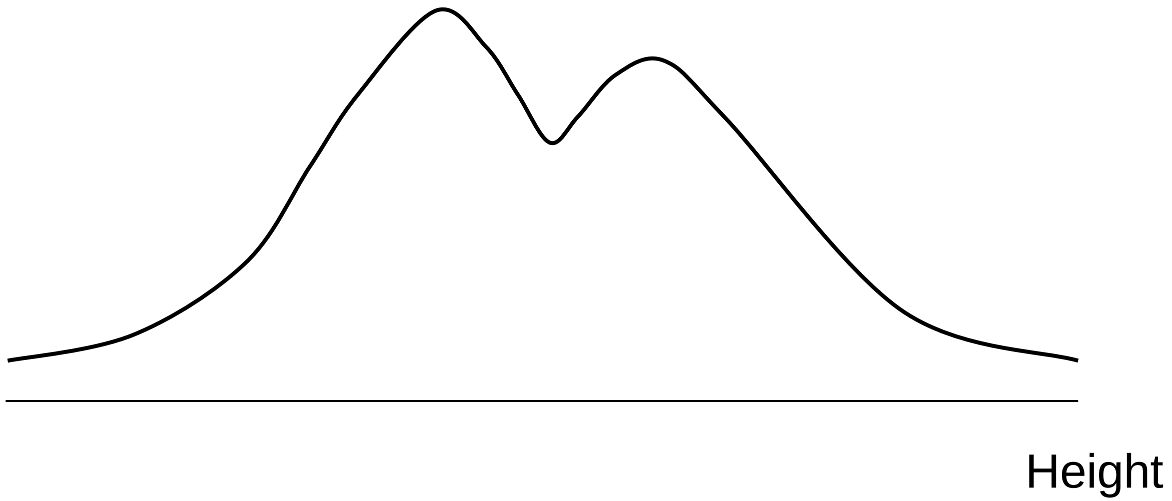
- Repeated binary measures
  - Maternal smoking using LCGA / LLCA

Single continuous variable

Extracting subgroups from a continuous GHQ sum-score

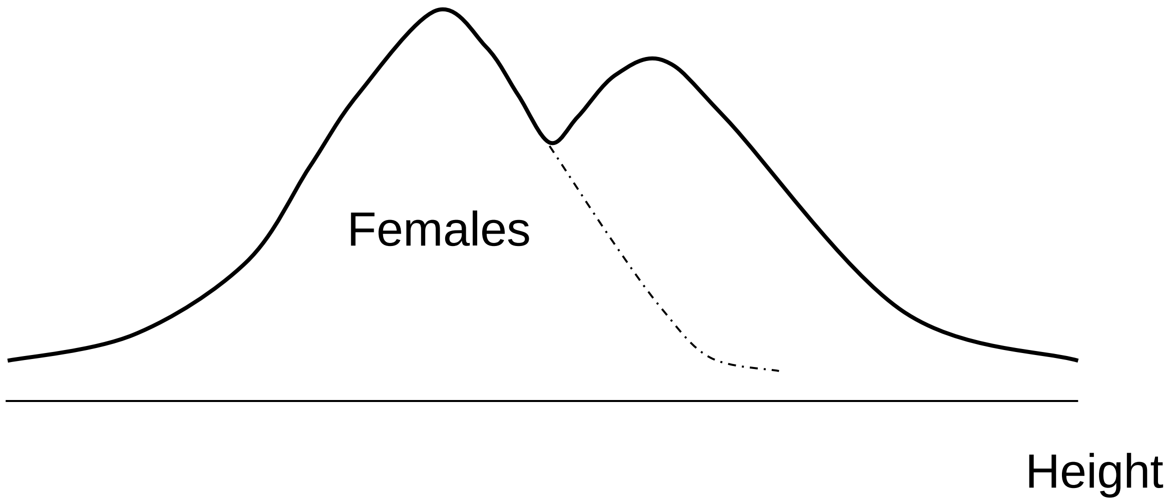
Single continuous variable

- An underlying latent grouping might present itself as a multi-modal distribution for the continuous variable



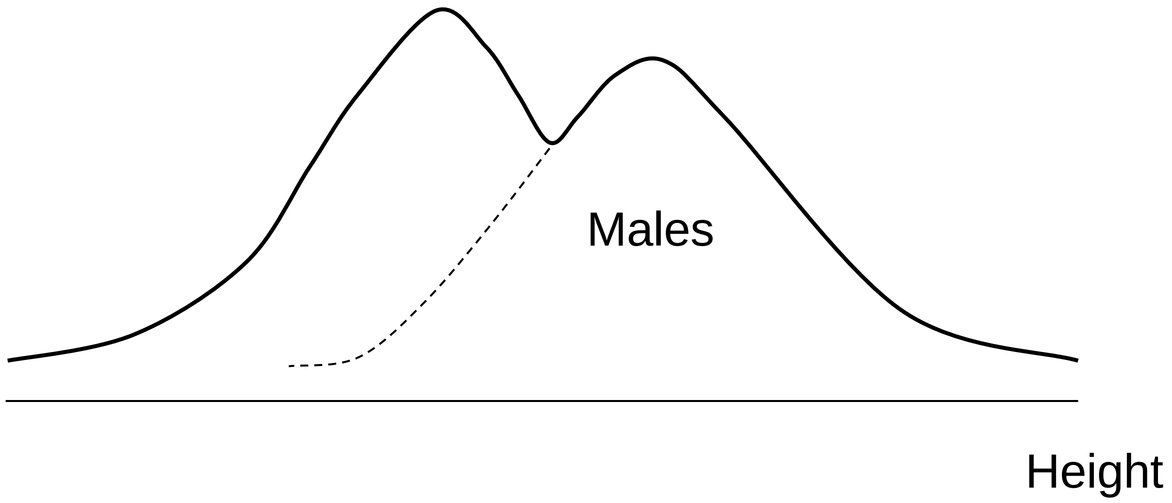
Single continuous variable

- An underlying latent grouping might present itself as a multi-modal distribution for the continuous variable



Single continuous variable

- An underlying latent grouping might present itself as a multi-modal distribution for the continuous variable



Single continuous variable

- We don’t really need a model to estimate gender

- Some groupings may be a bit harder to measure directly
- We can estimate a grouping using the data

- However, distance between modes may be small or even non-existent
- Depends on the variation in the item being measured and also the sample in which the measurement is taken (e.g. clinical or general population)

Single continuous variable

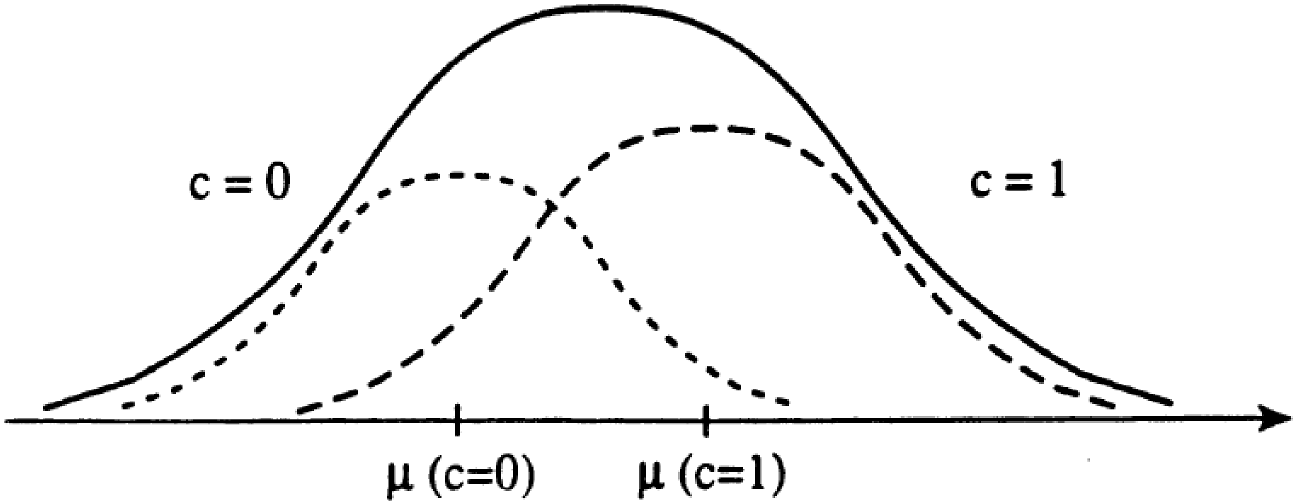
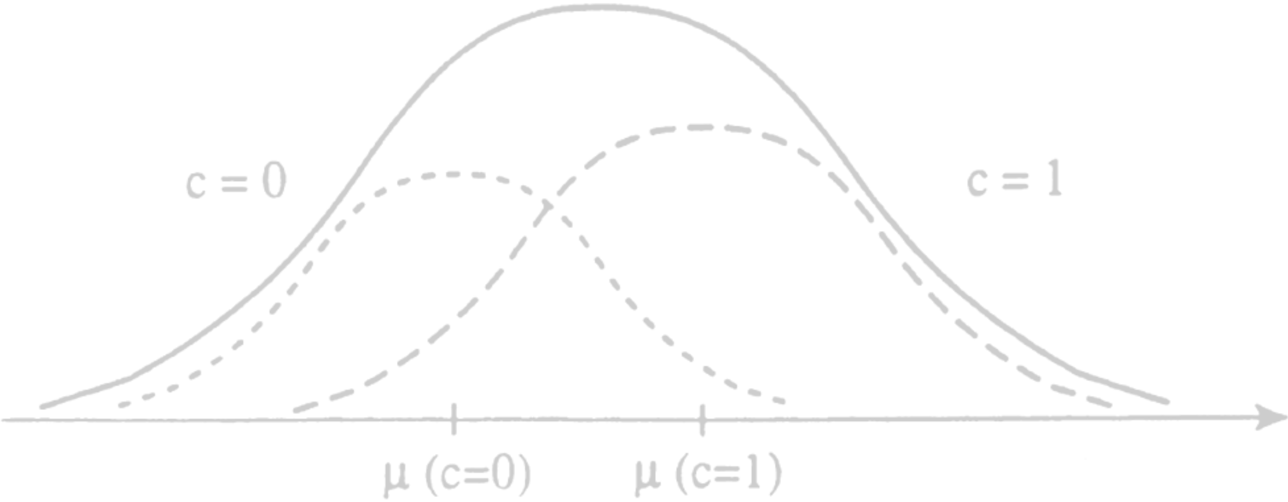


Figure taken from: Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (eds.), New Developments and Techniques in Structural Equation Modeling (pp. 1-33). Lawrence Erlbaum Associates.

Single continuous variable



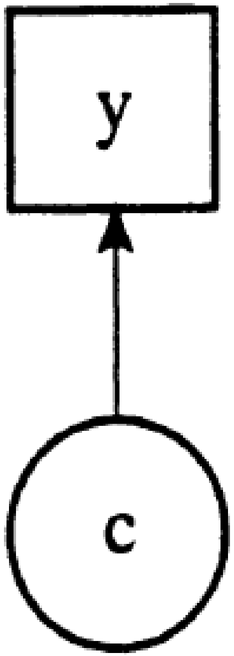


Figure taken from: Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (eds.), New Developments and Techniques in Structural Equation Modeling (pp. 1-33). Lawrence Erlbaum Associates.

Single continuous variable

- We assume that the manifest variable is normally distributed within each latent class
- It’s distribution can then be described by mean/variance

- Can allow means and/or variances to vary between classes

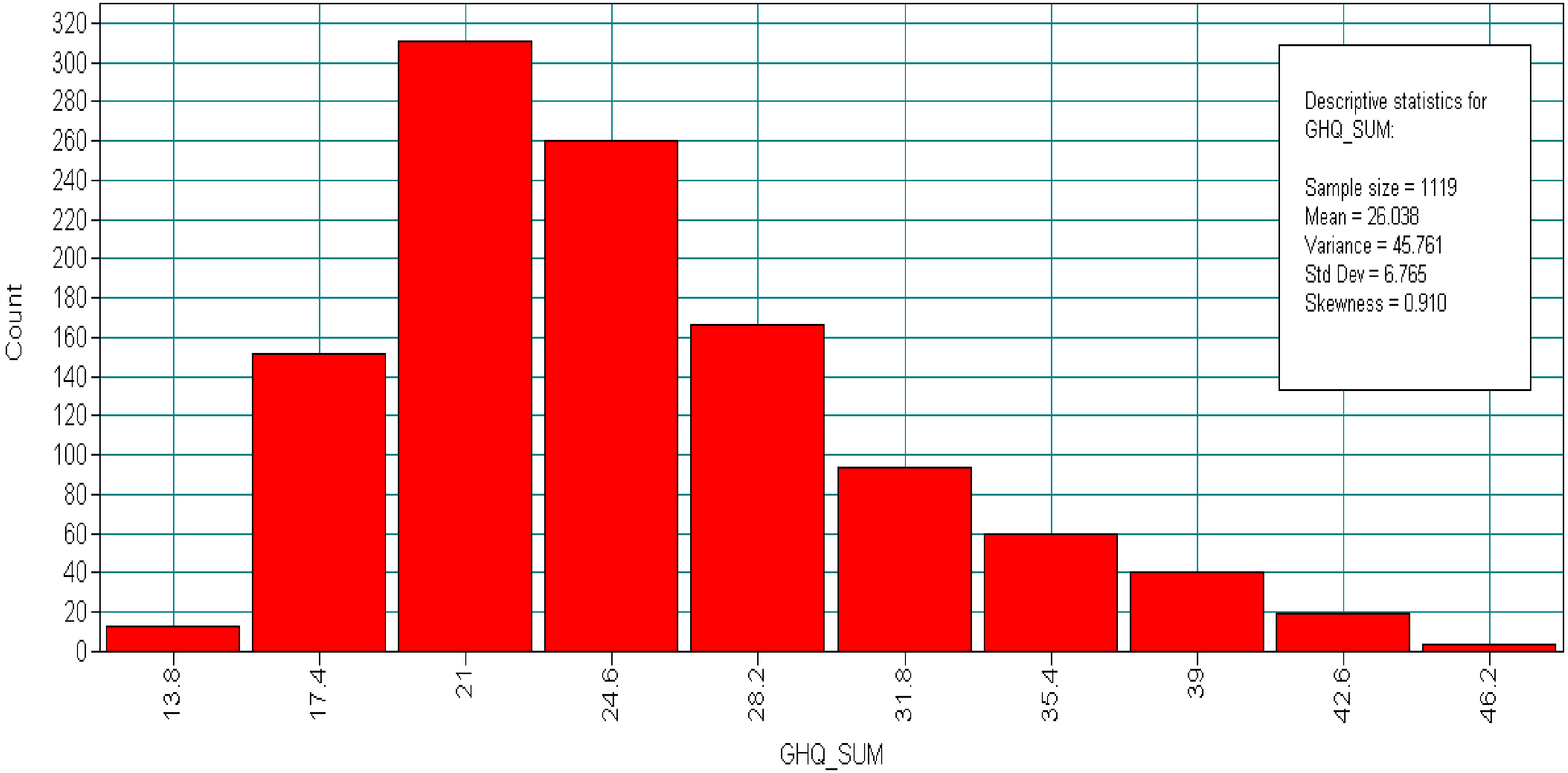
GHQ Example

```
Data:
  File is "ego_ghq12_id.dta.dat" ;

Define:
    sumodd  = ghq01 + ghq03 + ghq05 + ghq07 + ghq09 + ghq11;
    sumeven = ghq02 + ghq04 + ghq06 + ghq08 + ghq10 + ghq12;
    ghq_sum = sumodd + sumeven;

Variable:
  Names are
     ghq01 ghq02 ghq03 ghq04 ghq05 ghq06
     ghq07 ghq08 ghq09 ghq10 ghq11 ghq12
     f1 id;
  Missing are all (-9999) ;
  usevariables = ghq_sum;
```

Here we derive a single sum-score from the 12 ordinal GHQ items

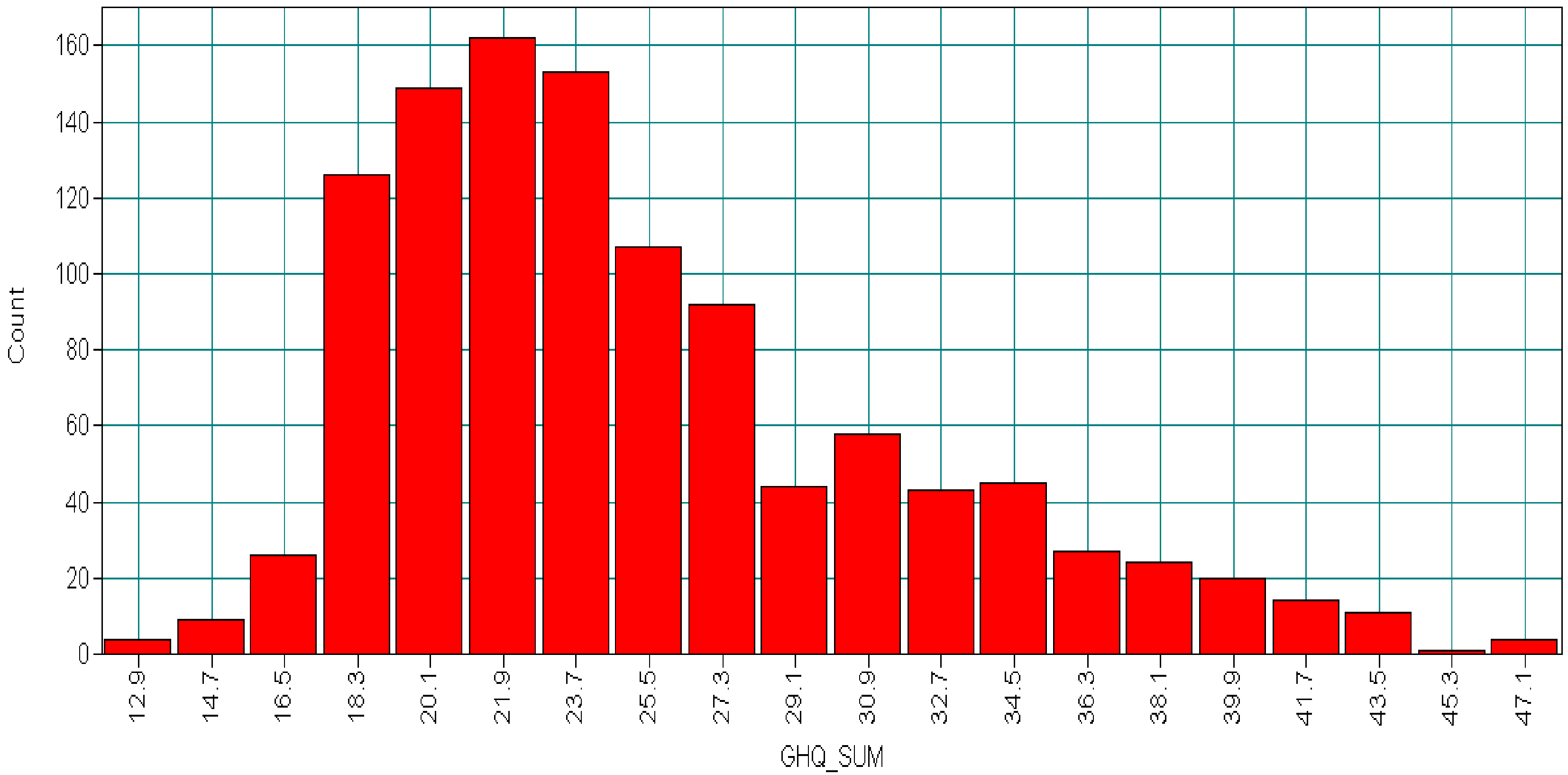
The syntax shows that variables can be created in the define statement which are not then used in the final model

Examine the distribution of the scale



Scale appears unimodal, although there is a long upper-tail

Examine the distribution of the scale



By changing from the default number of bins we see secondary modes appearing

Fit a 2-class mixture

```
Variable:
    <snip>
    classes = c(2);

Analysis:
    type = mixture ;
    proc = 2 (starts);
    starts = 100 20;
    stiterations = 20;
    stscale = 15;

model:
    %overall%

    %c#1%
    [ghq_sum];
    ghq_sum         (equal_var);

    %c#2%
    [ghq_sum];
    ghq_sum         (equal_var);
```

We are extracting two classes

This funny set of symbols refers to the first class

Means are referred to using square brackets.

Variances are bracket-less.

Here we have constrained the variances to be equal between classes

Means will be freely estimated.

Model results

- A smaller class of 17.9% has emerged, consistent with the expected behaviour of the GHQ in this sample from primary care

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THE ESTIMATED MODEL

    Latent classes

       1         200.36980          0.17906
       2         918.63020          0.82094
```

- Cue Tim to say a bit more about this population….

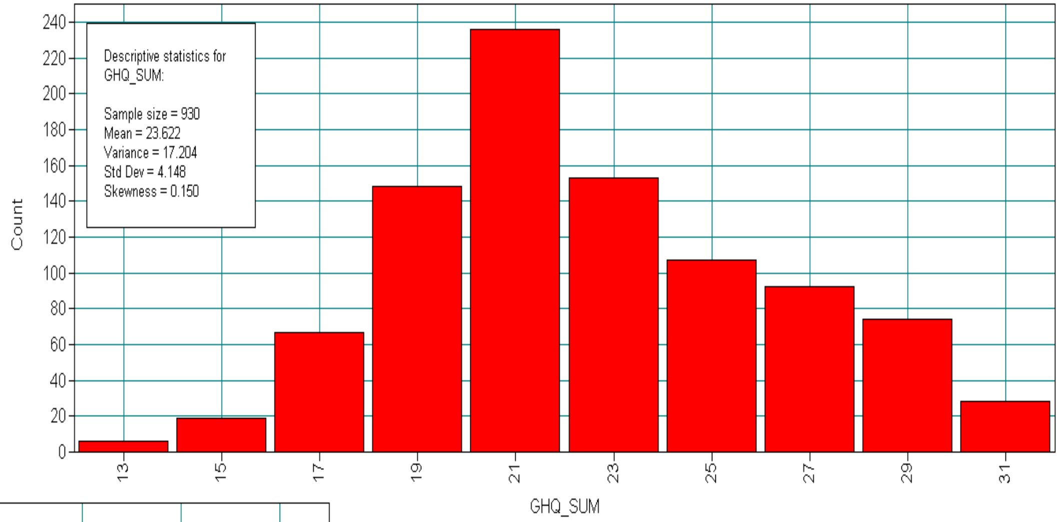
More model results

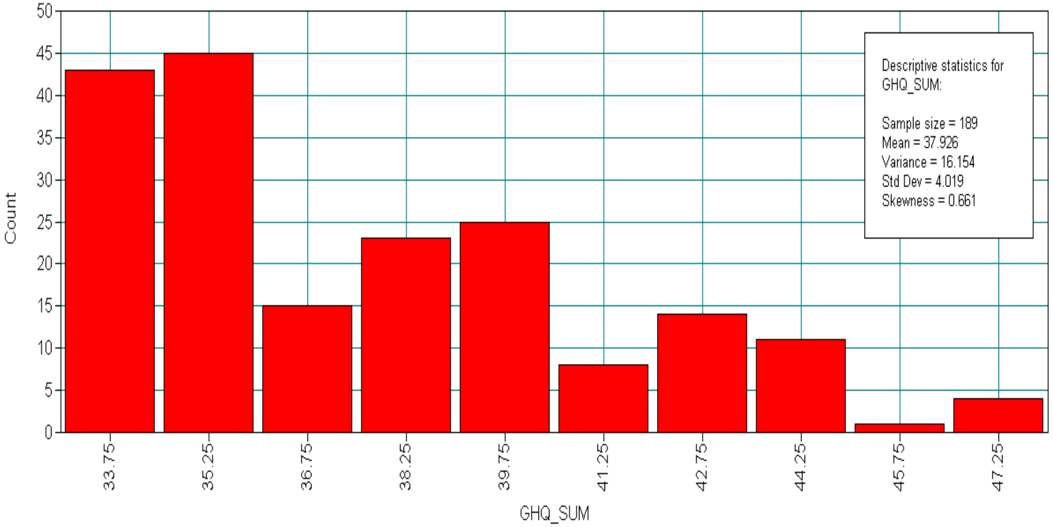
```
                                                  Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Latent Class 1

 Means  GHQ_SUM       37.131      0.574     64.737      0.000
 Vars   GHQ_SUM       18.876      1.016     18.581      0.000

Latent Class 2

 Means  GHQ_SUM       23.618      0.202    117.046      0.000
 Vars   GHQ_SUM       18.876      1.016     18.581      0.000

Categorical Latent Variables

 Means
    C#1               -1.523      0.118    -12.947      0.000
```

Huge separation in means since SD = 4.3 (i.e. sqrt(18.88))

Examine within-class distributions of GHQ score

- **Class 1**





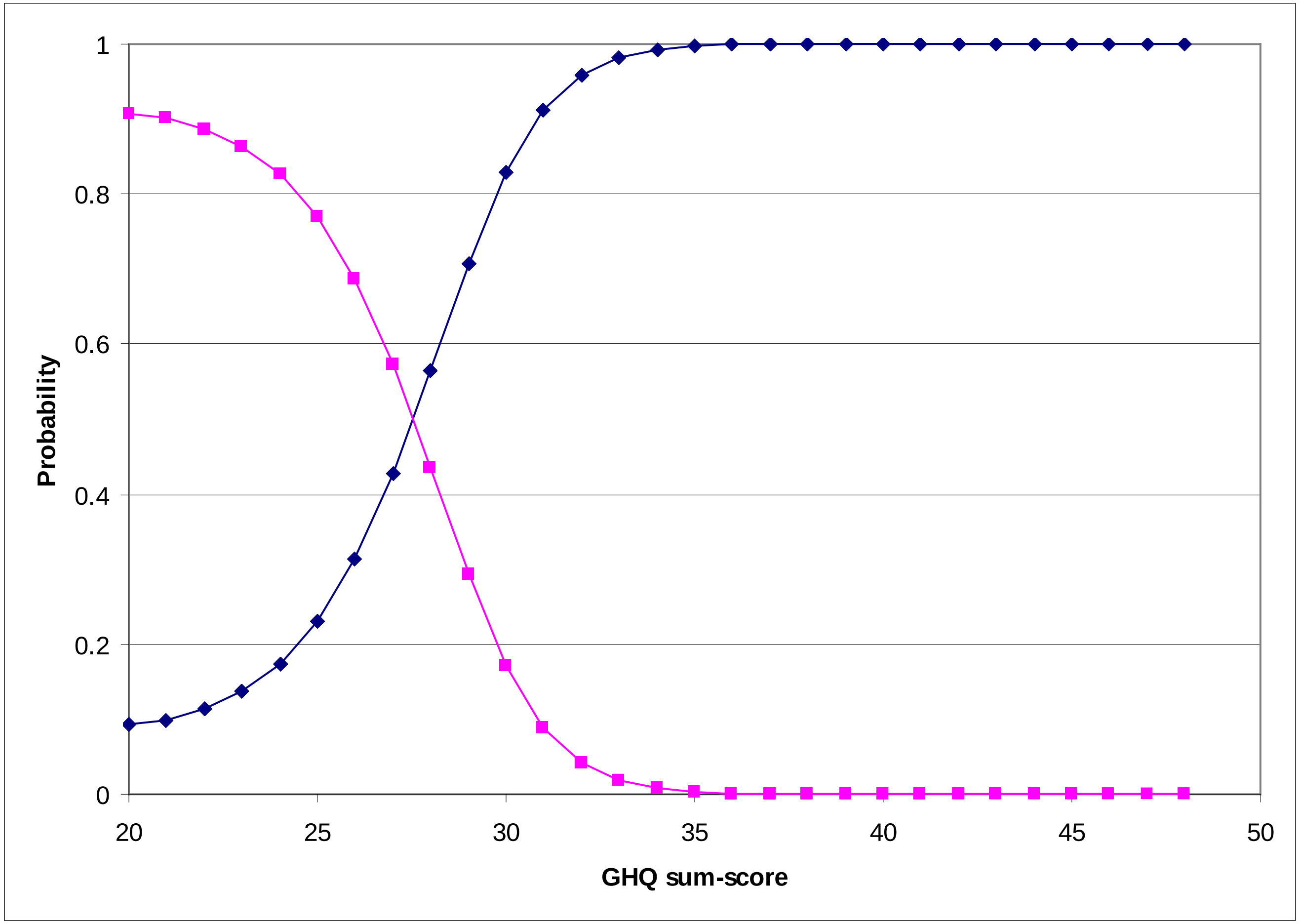
- **Class 2**

What have we done?

- We have effectively done a t-test backwards.
- Rather than obtaining a manifest binary variable, assuming equality of variances and testing for equality of means
- We have derived a latent binary variable based on the assumption of a difference in means (still with equal variance)

- This latent classification might be useful for looking at risk factors for high GHQ score or looking at it’s effect on later outcomes
- We are not actually assigning people

Class-Assignments are probabilistic



There is uncertainty - particularly in the region 25-30

What next?

- The bulk of the sample now falls into a class with a GHQ distribution which is more symmetric than the sample as a whole
- There appear to be additional modes within the smaller class

- The ‘optimal’ number of classes can be assessed using fit statistics and face validity.
- In the univariate case, residual correlations are not an issue, but when moving to a multivariate example, these too will need to be assessed (conditional dependence – come back to this).

Hold on, what was all that about???

- We made some assumptions about what the data should look like within each latent class

- A 2-class mixture was extracted which consisted of
  - [1] Estimated size of each class (class distribution)
  - [2] Within-class characteristics (means/co-variances)
  - [3] Probabilities of class membership for each respondent

- Expected improvement in model fit compared whole sample
  - Some variability explained – lower residuals
  - Members of each class are more homogeneous than whole population

Multiple continuous measures

Extending growth modelling to growth mixture modelling

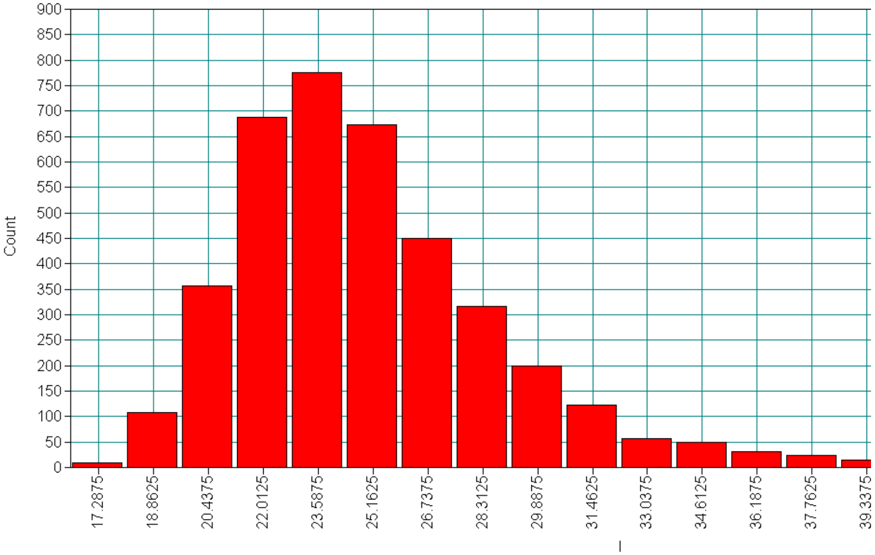
Multiple continuous measures

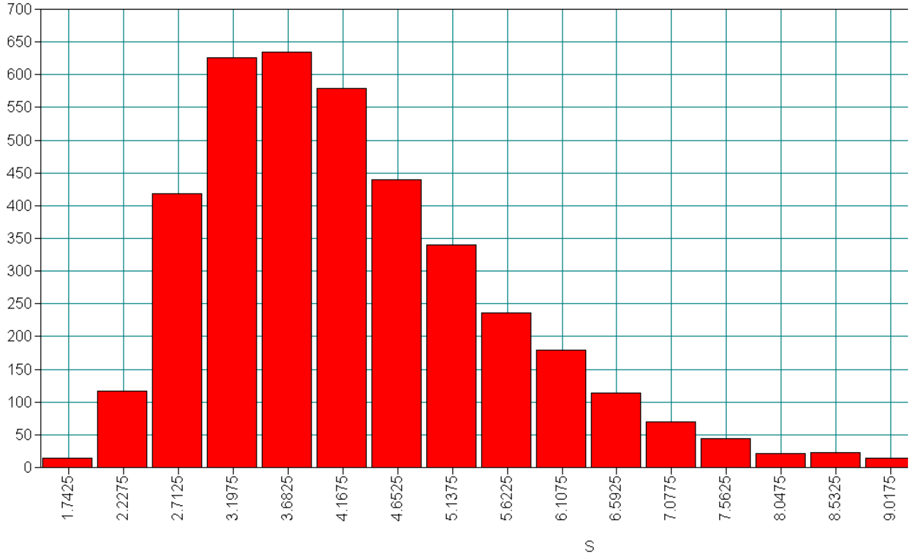
- We can fit a mixture model to multiple continuous measures
- Manifest data -> Latent Profile Analysis
- Can vary means / variances / covariances within each class

- Can also do this for continuous LATENT variables

- E.g. growth factors -> Growth Mixture Modelling

Recall the growth factors from yesterday:-

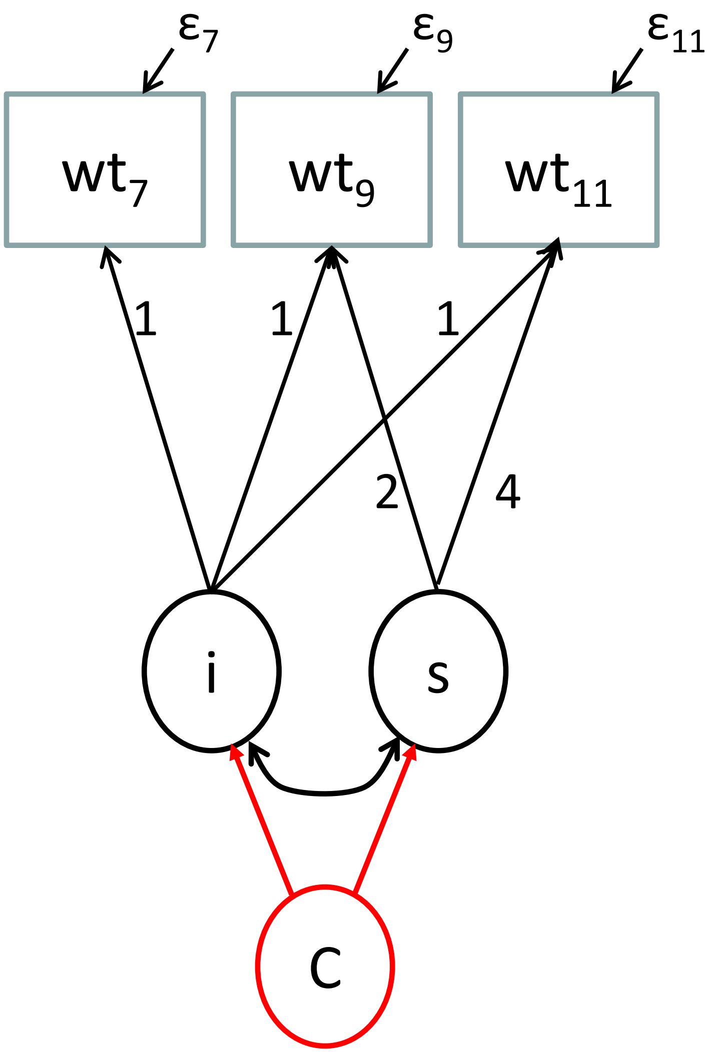




Measures were slightly skewed, this might be because there are underlying sub-populations

Such classes might be of clinical importance (as clinicians do love their groups)

2-class GMM



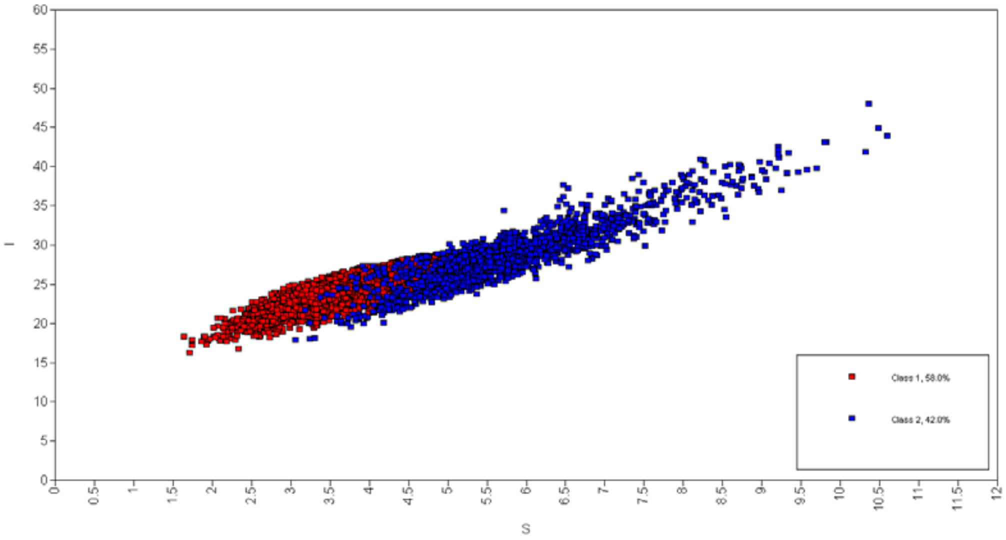
```
Variable:
   Names are  <snip>;
   Missing are all (-9999) ;
   usevariables = wt7 wt9 wt11;
   classes = c(2);

 Analysis:
     processors = 2 (starts);
     type = mixture;
     starts = 500 50;
     stiterations = 25;
     stscale = 25;

 Model:
     %overall%
     i s | wt7@0 wt9@2 wt11@4;
     wt7 wt9 wt11 (1);
```

Large number of possible 2-class GMM’s

Least constrained example

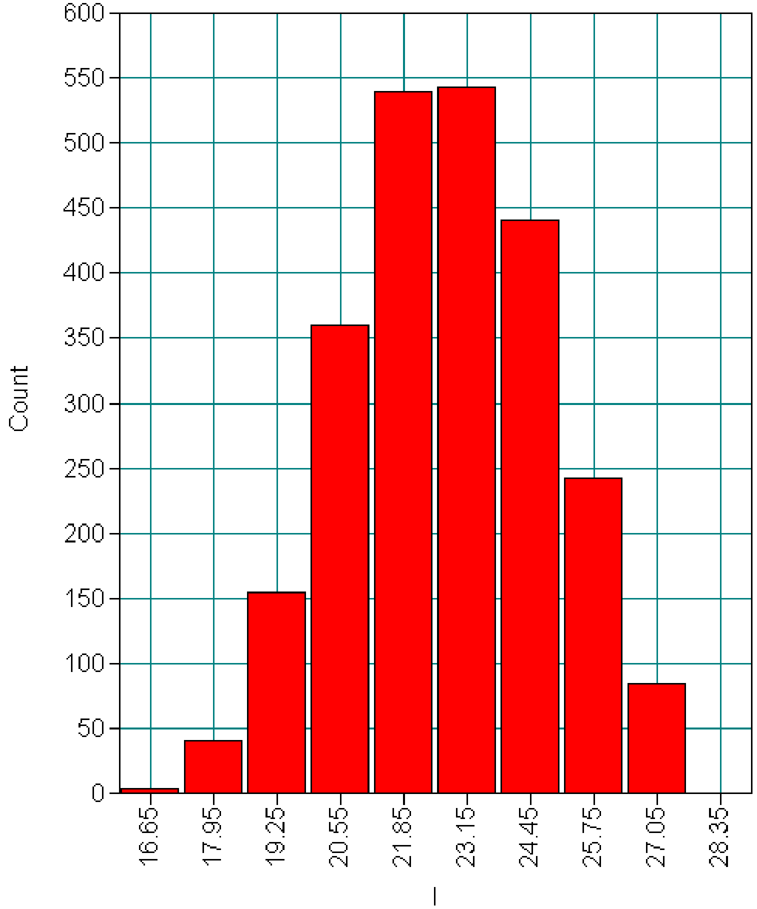
Means more syntax:

```
%overall%
  i s | wt7@0 wt9@2 wt11@4;

%c#1%
  i with s;
  i s;
  [i s];
  wt7 wt9 wt11 (1);

%c#2%
  i with s;
  i s;
  [i s];
  wt7 wt9 wt11 (2);
```

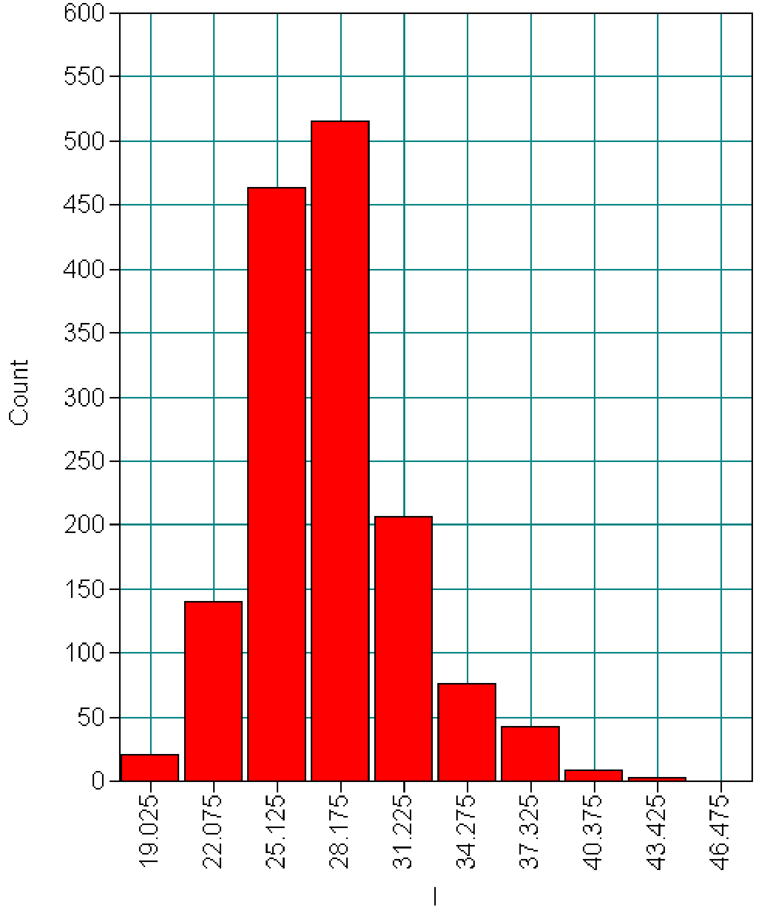


```
                Class 1 (58%)       Class 2 (42%)

S WITH I        1.288     0.071     4.557     0.265

Means
   I           23.637     0.083    28.029     0.195
   S            3.526     0.039     5.657     0.065
Variances
   I            6.215     0.279    20.008     1.117
   S            0.542     0.035     1.670     0.105
Res Vars
   WT7          0.989     0.057     5.943     0.312
   WT9          0.989     0.057     5.943     0.312
   WT11         0.989     0.057     5.943     0.312
```

Intercept factors more normally distributed

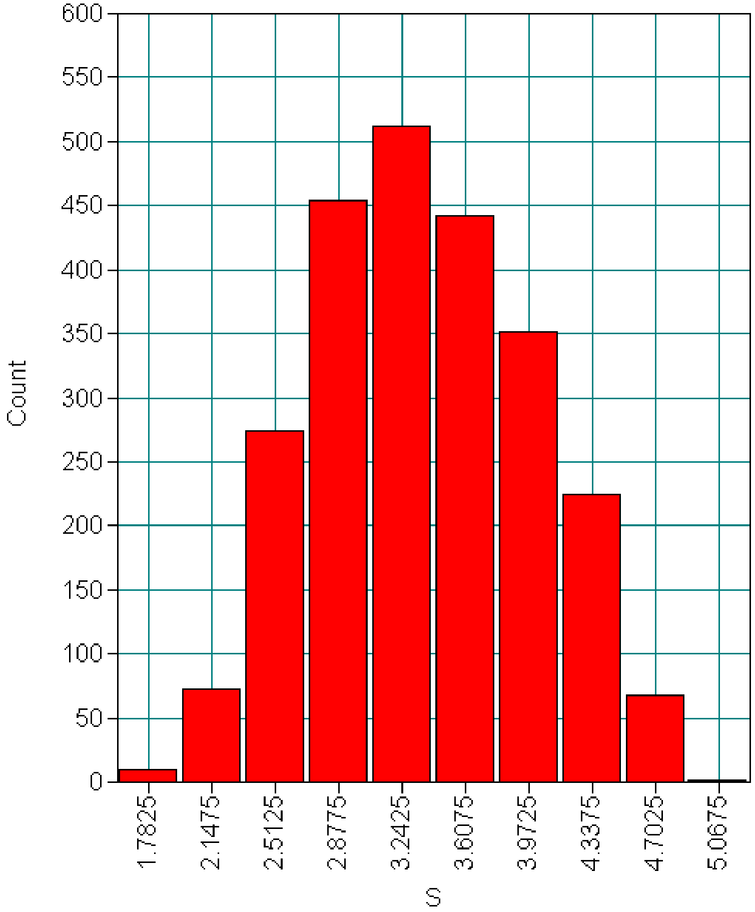


C1

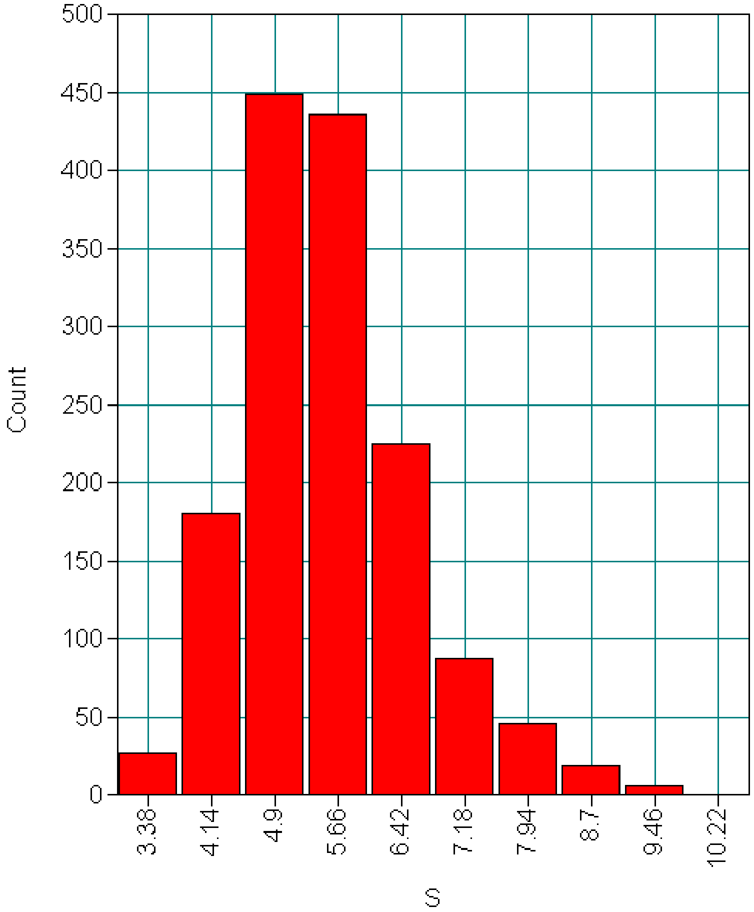


C2

Slope factors are too (ish)

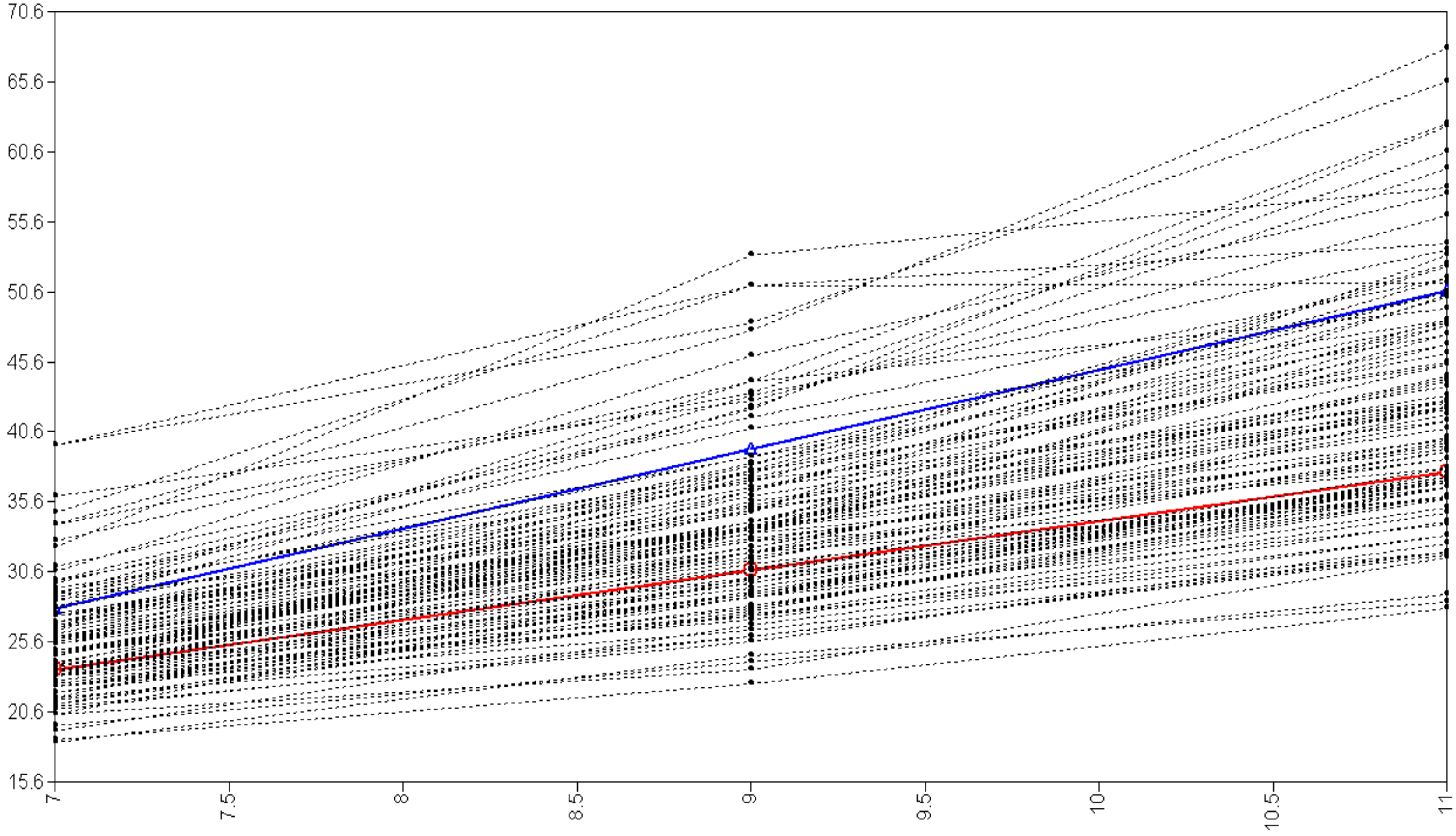


C1



C2

Fit was better, but are mixtures always useful?



Mixture modelling with cts data

- Mixture modelling can improve model fit by deriving discrete classes to represent an unmeasured source of variability (hence reducing unexplained variability ~ residuals)

- Whether these classes have more than merely statistical value will depend on the topic area / population studied

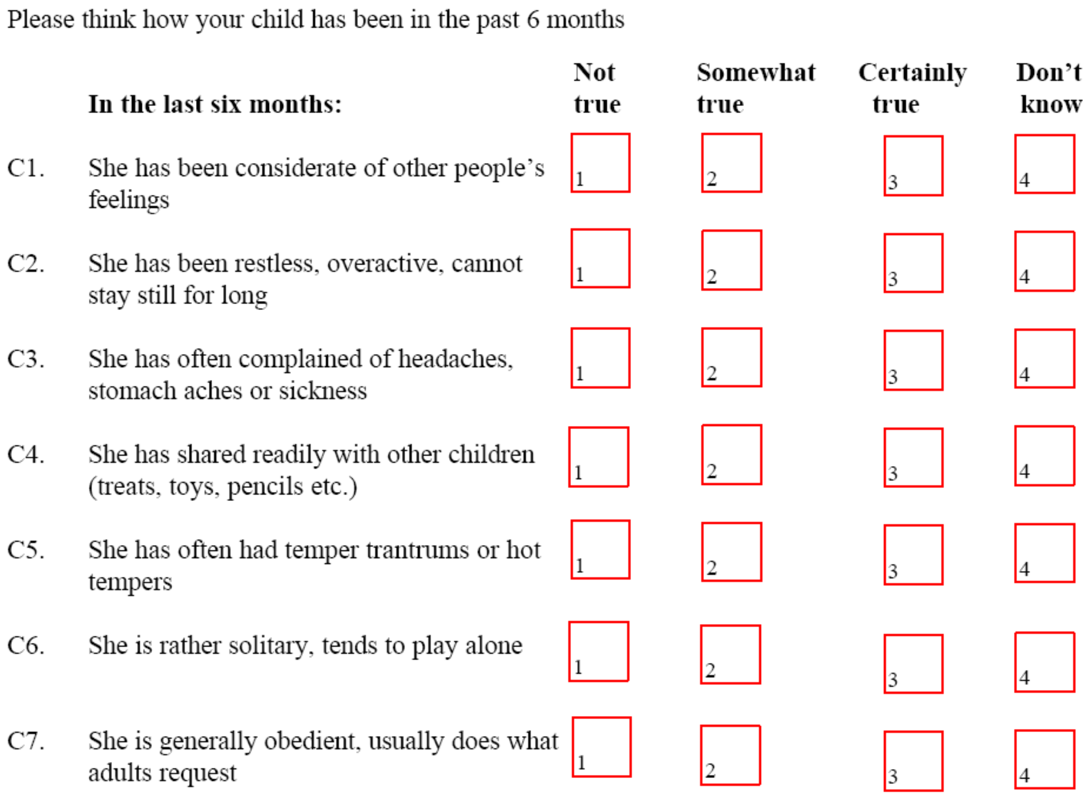
- In situations where most subjects track in a reasonably parallel way, the mixture approach is unlikely to add much to our understanding over+above the growth factors

GMM
Continuous data example 2

Where thinking in terms of classes may be more rewarding

SDQ – Total behavioural difficulties

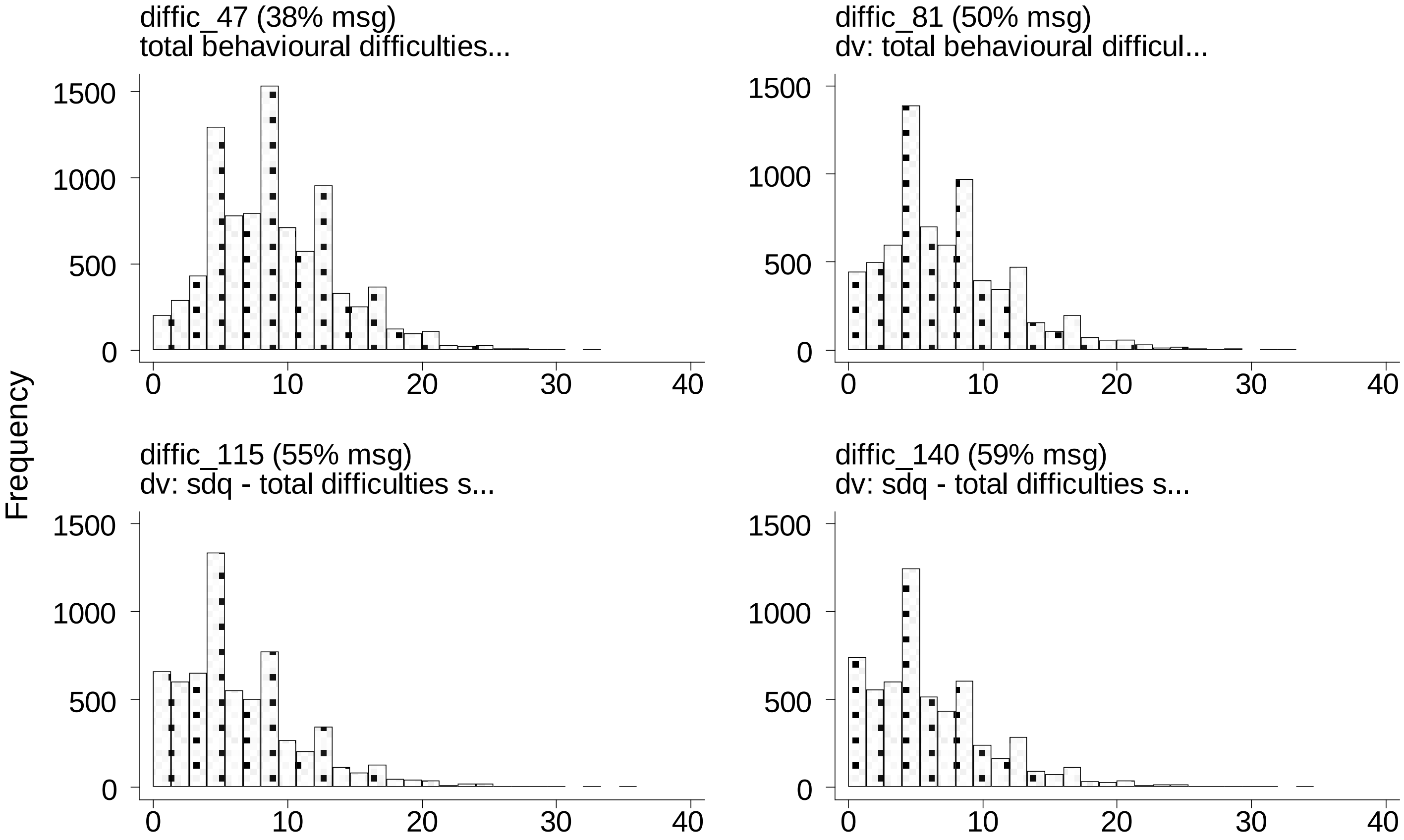
- Mother reported SDQ (strengths and difficulties)



Please think how your child has been in the past 6 months

	In the last six months:	Not true	Somewhat true	Certainly true	Don't know
C1.	She has been considerate of other people's feelings	□ 1	□ 2	□ 3	□ 4
C2.	She has been restless, overactive, cannot stay still for long	□ 1	□ 2	□ 3	□ 4
C3.	She has often complained of headaches, stomach aches or sickness	□ 1	□ 2	□ 3	□ 4
C4.	She has shared readily with other children (treats, toys, pencils etc.)	□ 1	□ 2	□ 3	□ 4
C5.	She has often had temper trantrums or hot tempers	□ 1	□ 2	□ 3	□ 4
C6.	She is rather solitary, tends to play alone	□ 1	□ 2	□ 3	□ 4
C7.	She is generally obedient, usually does what adults request	□ 1	□ 2	□ 3	□ 4

E.g. 2 – Behavioural difficulties: 47/81/115/140 mns



Data four SDQ total difficulties measures.dta (n=14273)

Syntax for linear LGM

```
data:
   File is sdq.dta.dat ;
   listwise is on;

variable:
   Names are sex sdq_47 sdq_81 sdq_115 sdq_140 id;
   Missing are all (-9999) ;
   usevariables = sdq_47 sdq_81 sdq_115 sdq_140;

model:
   i s | sdq_47@0 sdq_81@2.83 sdq_115@5.67 sdq_140@7.75;
   sdq_47 sdq_81 sdq_115 sdq_140 (1);

output:
   tech4 sampstat residual;

plot:
   type is plot3;
   series is sdq_47 (0) sdq_81 (2.83) sdq_115 (5.67) sdq_140 (7.75);
```

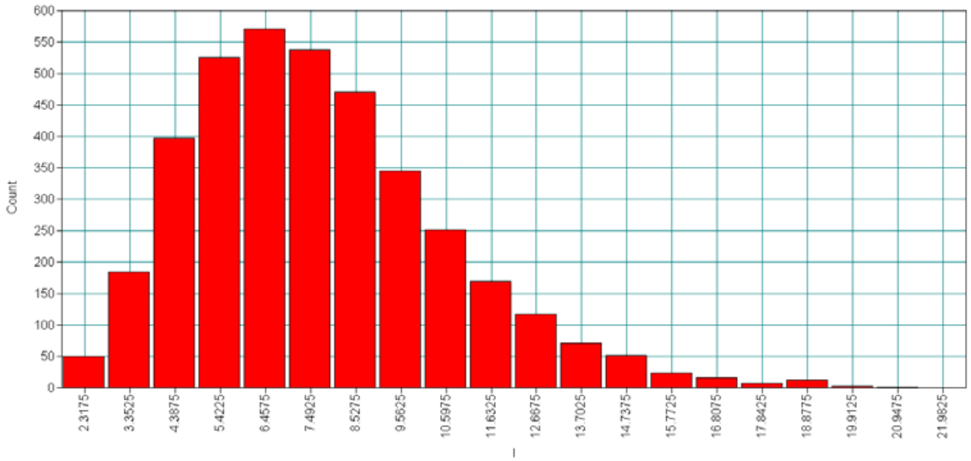
Listwise is on i.e. a complete-case analysis

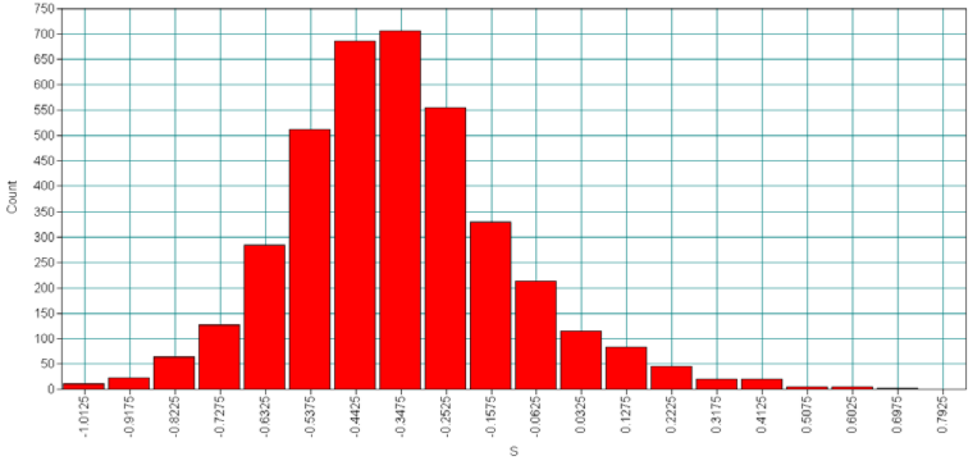
Loadings are the number of years since t1

Residual variance equal through time

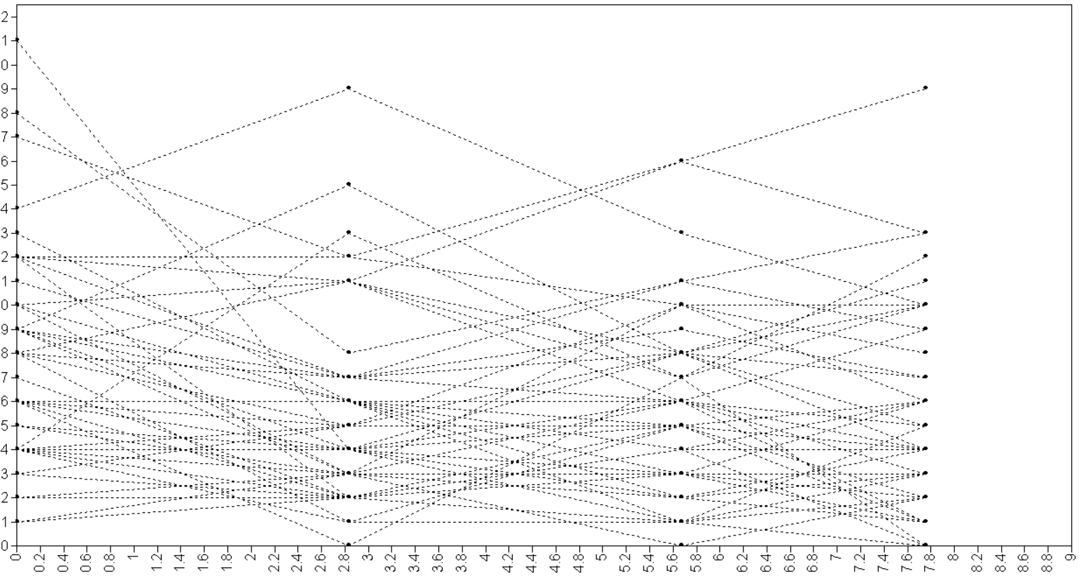
Linear LGM results (n = 3,805)

```
MODEL RESULTS
                       Estimate    S.E.

 S WITH I                -0.311   (0.043)

 Means
    I                     7.995   (0.066)
    S                    -0.315   (0.009)

 Intercepts
    SDQ_47                0.000
    SDQ_81                0.000
    SDQ_115               0.000
    SDQ_140               0.000

 Variances
    I                    12.035   (0.391)
    S                     0.137   (0.008)

 Residual Variances
    SDQ_47                6.414   (0.104)
    SDQ_81                6.414   (0.104)
    SDQ_115               6.414   (0.104)
    SDQ_140               6.414   (0.104)
```



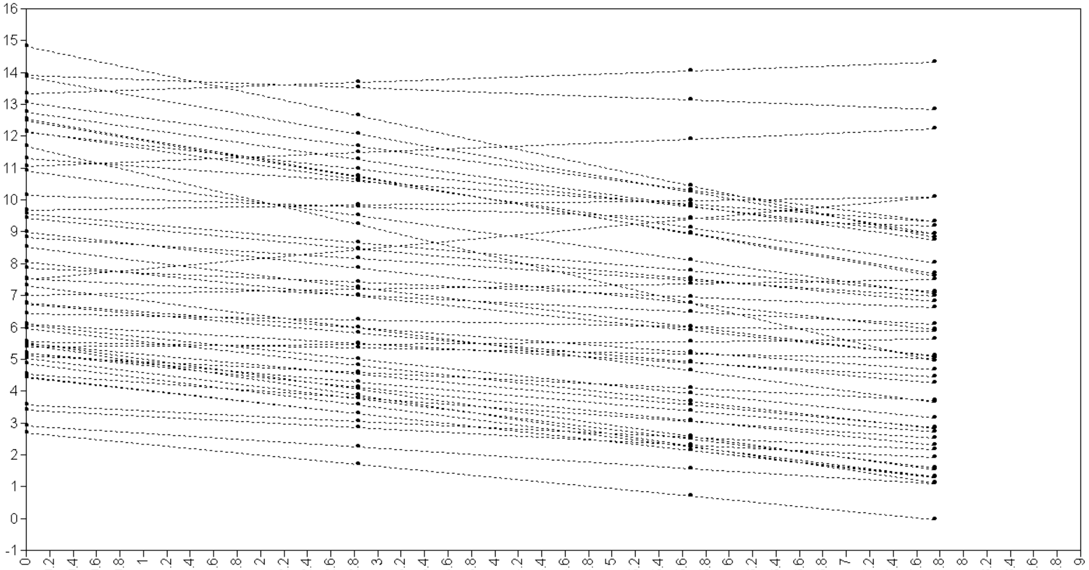


Observed data
(cases 1-50)



- **Not great!!**

Linear LGM model
estimated data
(cases 1-50)



Linear LGM residuals suggest non-linearity

```
RESIDUAL OUTPUT


     ESTIMATED MODEL AND RESIDUALS (OBSERVED - ESTIMATED)


           Model Estimated Means/Intercepts/Thresholds
              SDQ_47        SDQ_81        SDQ_115       SDQ_140
              ________      ________      ________      ________
 1              7.995         7.103         6.208         5.552


           Residuals for Means/Intercepts/Thresholds
              SDQ_47        SDQ_81        SDQ_115       SDQ_140
              ________      ________      ________      ________
 1              0.215        -0.256        -0.196         0.237


           Standardized Residuals (z-scores) for Means/Intercepts/Thresholds
              SDQ_47        SDQ_81        SDQ_115       SDQ_140
              ________      ________      ________      ________
 1             10.737        -6.299        -5.695        25.460
```

Syntax for quadratic LGM – tricky!!

```
data:
   File is sdq.dta.dat ;
   listwise is on;

variable:
   Names are sex sdq_47 sdq_81 sdq_115 sdq_140 id;
   Missing are all (-9999) ;
   usevariables = sdq_47 sdq_81 sdq_115 sdq_140;

model:
   i s q | sdq_47@0 sdq_81@2.83 sdq_115@5.67 sdq_140@7.75;
   sdq_47 sdq_81 sdq_115 sdq_140 (1);

output:
   tech4 sampstat residual;

plot:
   type is plot3;
   series is sdq_47 (0) sdq_81 (2.83) sdq_115 (5.67) sdq_140 (7.75);
```

Just to remind you – without the shorthand

```
data:
   File is sdq.dta.dat ;
   listwise is on;

variable:
   Names are sex sdq_47 sdq_81 sdq_115 sdq_140 id;
   Missing are all (-9999) ;
   usevariables = sdq_47 sdq_81 sdq_115 sdq_140;

model:
   i  by  sdq_47@1  sdq_81@1     sdq_115@1      sdq_140@1;
   s  by  sdq_47@0  sdq_81@2.83  sdq_115@5.67   sdq_140@7.75;
   q  by  sdq_47@0  sdq_81@8.01  sdq_115@32.15  sdq_140@60.06;

   [sdq_47@0 sdq_81@0 sdq_115@0 sdq_140@0];
   [i s q];
   sdq_47 sdq_81 sdq_115 sdq_140 (1);
```

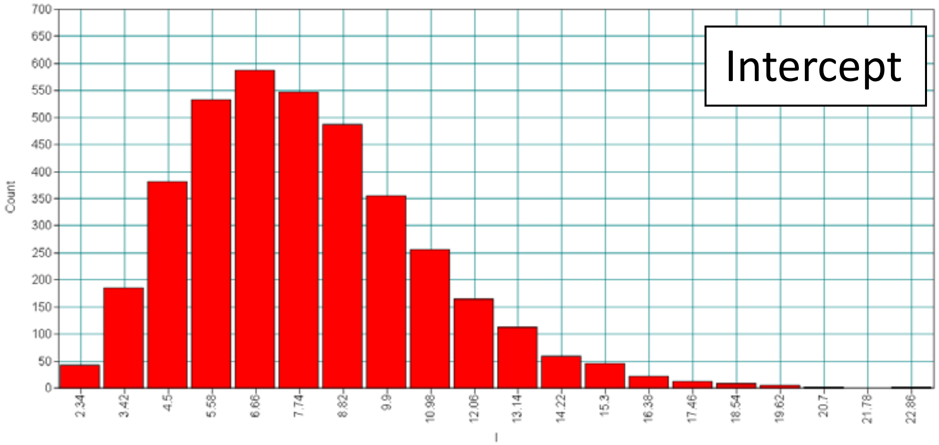
- **Warning – those loadings are getting a bit large!!**

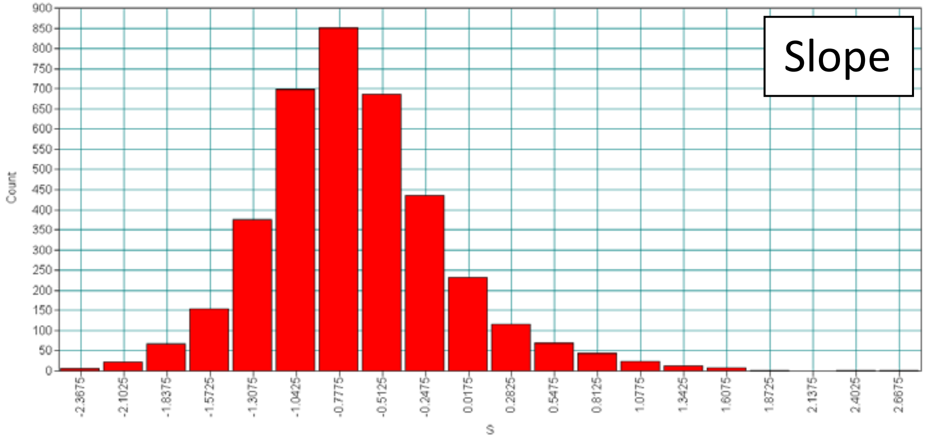
Improved residuals for quadratic model

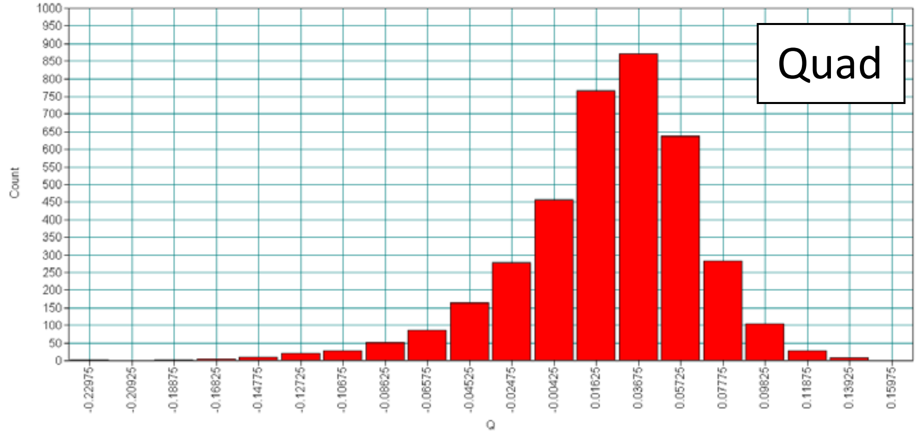
```
RESIDUAL OUTPUT


     ESTIMATED MODEL AND RESIDUALS (OBSERVED - ESTIMATED)


           Model Estimated Means/Intercepts/Thresholds
              SDQ_47        SDQ_81        SDQ_115       SDQ_140
              ________      ________      ________      ________
      1         8.212         6.838         6.023         5.784


           Residuals for Means/Intercepts/Thresholds
              SDQ_47        SDQ_81        SDQ_115       SDQ_140
              ________      ________      ________      ________
      1        -0.003         0.009        -0.011         0.004


           Standardized Residuals (z-scores) for Means/Intercepts/Thresholds
              SDQ_47        SDQ_81        SDQ_115       SDQ_140
              ________      ________      ________      ________
      1        -0.326         0.331        -0.456         0.354
```

Quadratic LGM

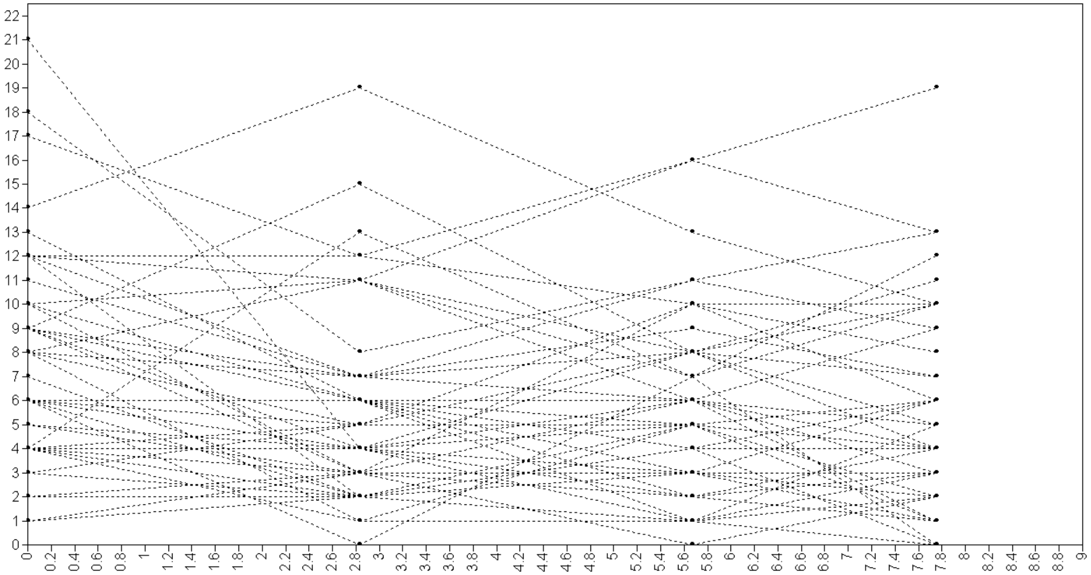
```
MODEL RESULTS
                  Estimate      S.E.

 S WITH I          -0.606      0.139
 Q WITH I           0.014      0.015
 S WITH Q          -0.070      0.010

 Means
  I                 8.212      0.069
  S                -0.585      0.028
  Q                 0.035      0.003

Variances
  I                12.446      0.431
  S                 0.822      0.086
  Q                 0.007      0.001

Residual Variances
  SDQ_47            5.733      0.131
  SDQ_81            5.733      0.131
  SDQ_115           5.733      0.131
  SDQ_140           5.733      0.131
```



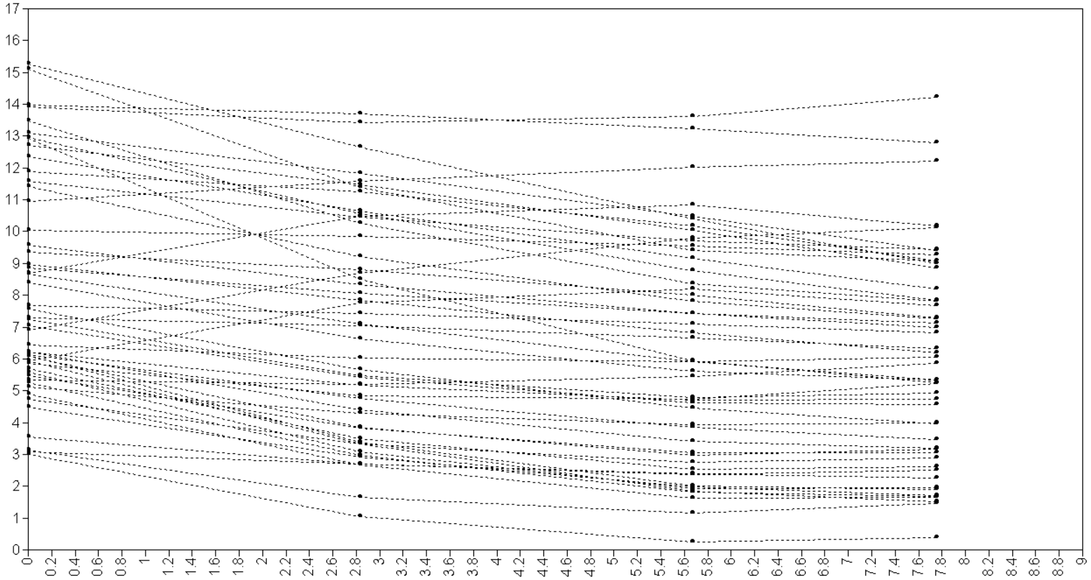




Observed data
(cases 1-50)



Quadractic LGM model
estimated data
(cases 1-50)



Syntax for 2-class quadratic GMM

```
variable:
   <snip>
   classes = c(2);

analysis:
      type = mixture;
      starts = 500 50;

model:
      %overall%
      i s q | sdq_47@0 sdq_81@2.83 sdq_115@5.67 sdq_140@7.75;

      %c#1%
      i with s q;
      s with q;
      i s q;
      sdq_47 sdq_81 sdq_115 sdq_140 (1);

      %c#2%
      i with s q;
      s with q;
      i s q;
      sdq_47 sdq_81 sdq_115 sdq_140 (2);
```

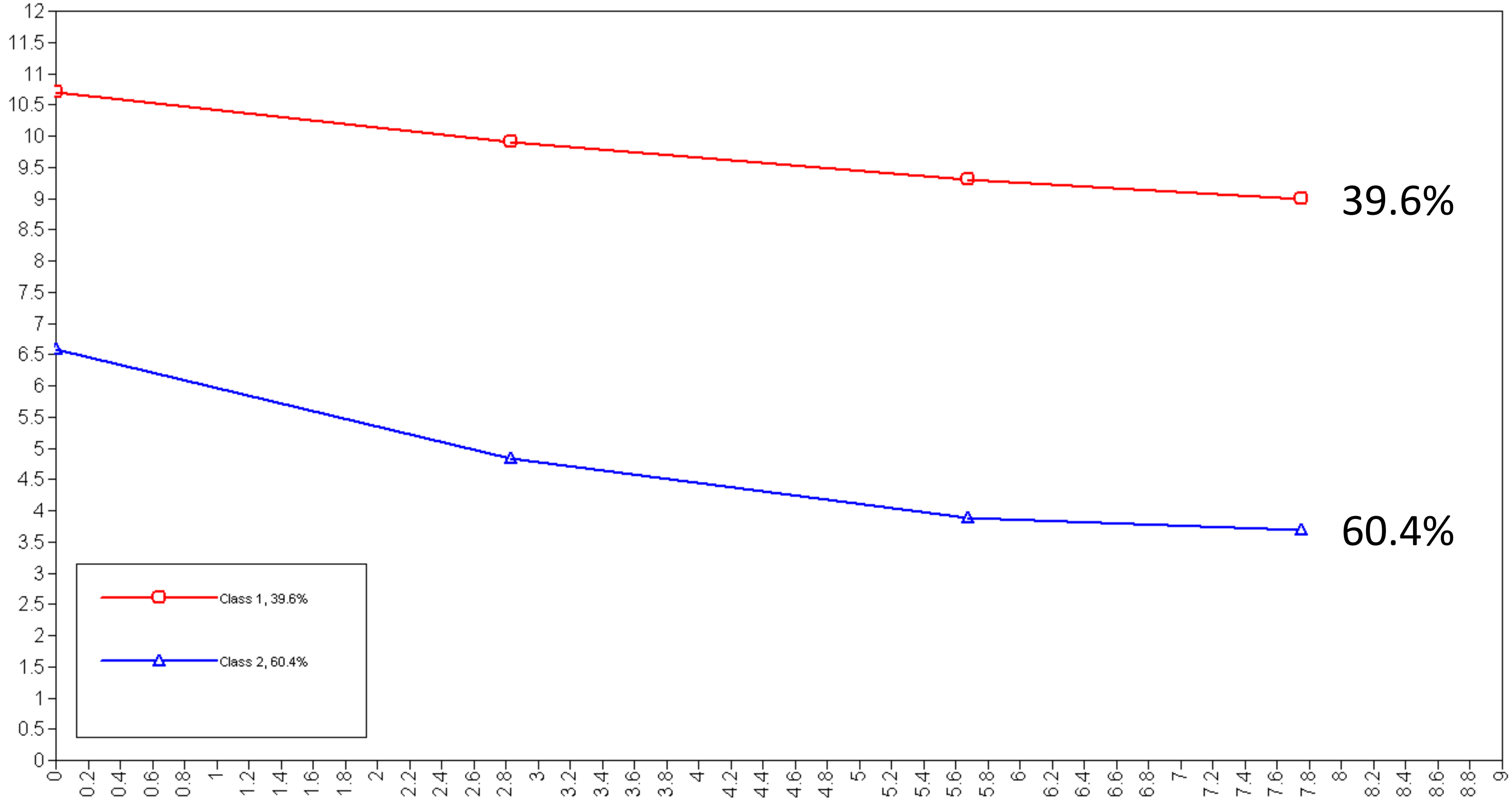
We specify the bits that we'd like to vary between classes

Results for a quadratic GMM[2]

```
         Latent Class 1 (39.6%)


S WITH I                  -0.630      0.373

Q WITH
   I                      -0.014      0.041
   S                      -0.094      0.030

Means
   I                      10.704      0.160
   S                      -0.324      0.072
   Q                       0.013      0.008

Variances
   I                       9.915      0.884
   S                       1.125      0.266
   Q                       0.010      0.004

Residual Variances
   SDQ_47                 10.602      0.560
   SDQ_81                 10.602      0.560
   SDQ_115                10.602      0.560
   SDQ_140                10.602      0.560
```

```
         Latent Class 2 (60.4%)


S WITH I                  -1.295      0.151

Q WITH
   I                       0.091      0.015
   S                      -0.048      0.008

Means
   I                       6.580      0.136
   S                      -0.756      0.037
   Q                       0.049      0.004

Variances
   I                       7.375      0.496
   S                       0.549      0.071
   Q                       0.005      0.001

Residual Variances
   SDQ_47                  2.543      0.139
   SDQ_81                  2.543      0.139
   SDQ_115                 2.543      0.139
   SDQ_140                 2.543      0.139
```

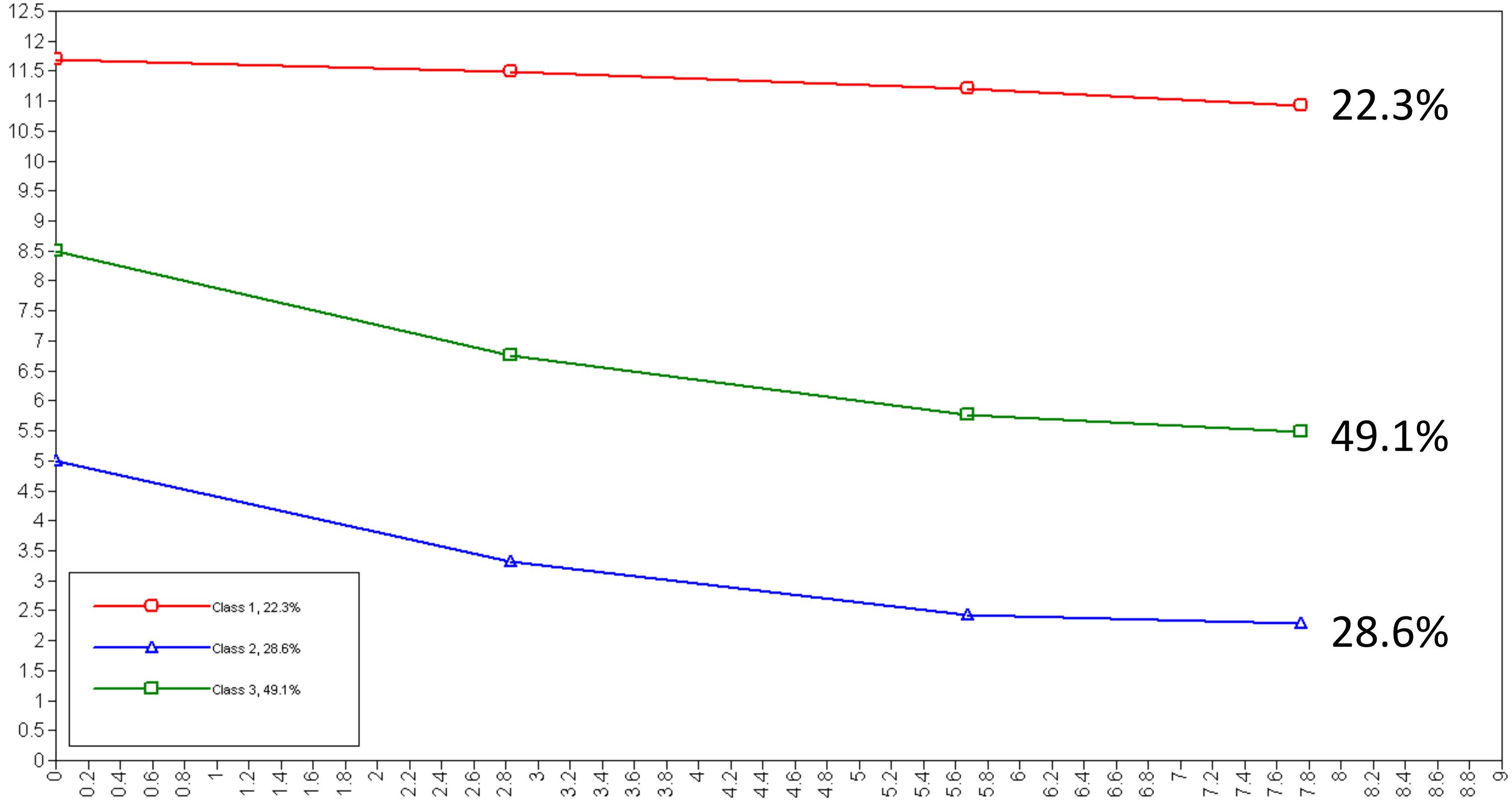
Trajectories for a 2-class quadratic GMM



39.6%

60.4%

Trajectories for a 3-class quadratic GMM



Summary

- The growth mixture model for bodyweight was not particularly useful
- People all grow but at different rates
- Their rank ordering is relatively stable

- When there is more longitudinal variability GMM may prove useful
- This example sort-of demonstrated that!!
- We found a subgroup who's behaviour did not improve like the rest of the population
- They might go on to be really naughty in later life

Coffee time?

Before we move on to binary data

Mixture models for today

- Single continuous measure
  - GHQ
- Multiple continuous measures
  - Extending Latent Growth Models to GMM
    - Bodyweight example from yesterday
    - Repeated measures of SDQ
- Repeated binary measures
  - Maternal smoking using LCGA / LLCA

Modelling with binary data

Trajectories of maternal smoking following childbirth

Maternal smoking following the ALSPAC birth

- Repeated measures of maternal smoking
  - Asked at 6 time points: 2/8/21/33/47/61 months
  - Collected as: None / 1-10 per day / 11-20 per day / over 20 per day
  - Collapsed into Yes/No
- Aims
  - To attempt to model maternal exposure across time period
  - Relate smoking behaviour to
    - earlier risk factors,
    - later outcome in child
  - Totally gloss over the effect of later pregnancies on smoking

The data

```
   msmk2 |      Freq.     Percent        Cum.
---------+-----------------------------------
      no |      9,017       77.09       77.09
     yes |      2,680       22.91      100.00
---------+-----------------------------------
   Total |     11,697      100.00
```

```
   msmk8 |      Freq.     Percent        Cum.
---------+-----------------------------------
      no |      8,402       75.80       75.80
     yes |      2,683       24.20      100.00
---------+-----------------------------------
   Total |     11,085      100.00
```

```
  msmk21 |      Freq.     Percent        Cum.
---------+-----------------------------------
      no |      7,853       77.34       77.34
     yes |      2,301       22.66      100.00
---------+-----------------------------------
   Total |     10,154      100.00
```

```
  msmk33 |      Freq.     Percent        Cum.
---------+-----------------------------------
      no |      7,250       77.40       77.40
     yes |      2,117       22.60      100.00
---------+-----------------------------------
   Total |      9,367      100.00
```

```
  msmk47 |      Freq.     Percent        Cum.
---------+-----------------------------------
      no |      7,389       77.77       77.77
     yes |      2,112       22.23      100.00
---------+-----------------------------------
   Total |      9,501      100.00
```

```
  msmk61 |      Freq.     Percent        Cum.
---------+-----------------------------------
      no |      6,729       76.20       76.20
     yes |      2,102       23.80      100.00
---------+-----------------------------------
   Total |      8,831      100.00
```

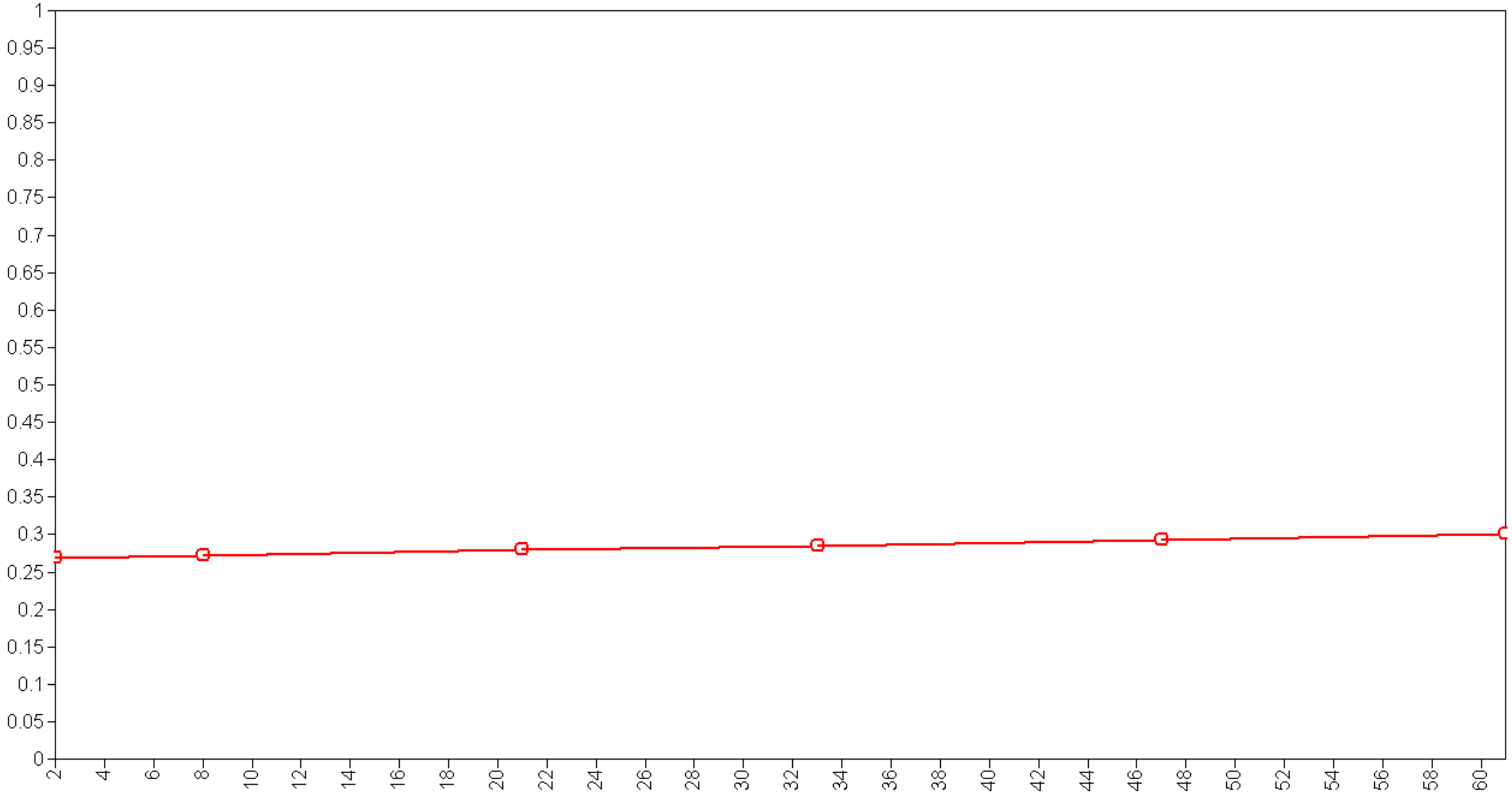
First attempt

- Fit a linear growth model to these data

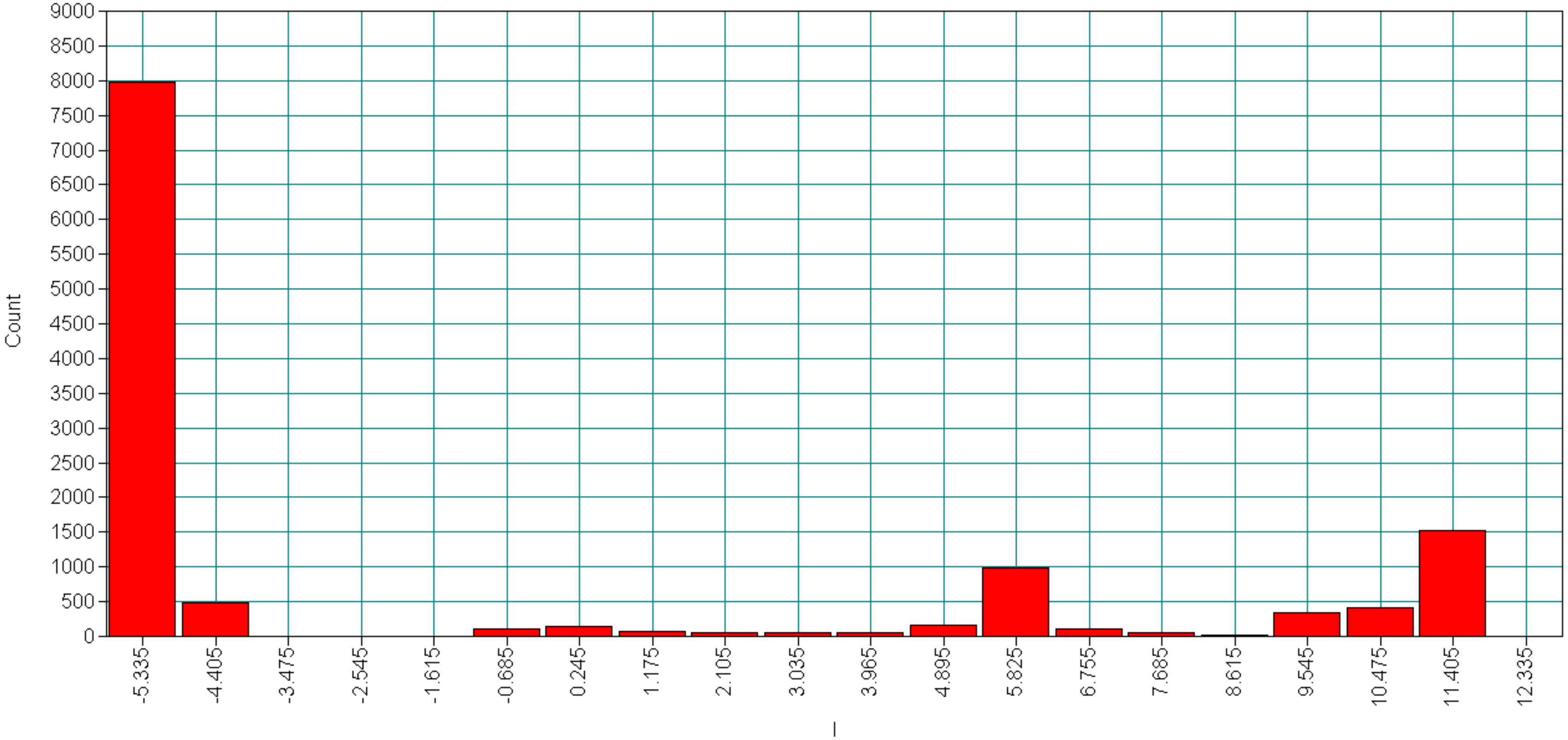
- Similar to yesterday
  - Intercept / slope means
  - Intercept / slope variances + covariance

- Different to yesterday
  - Logit rather than identity link function as measures are binary

LGM: Population average behaviour

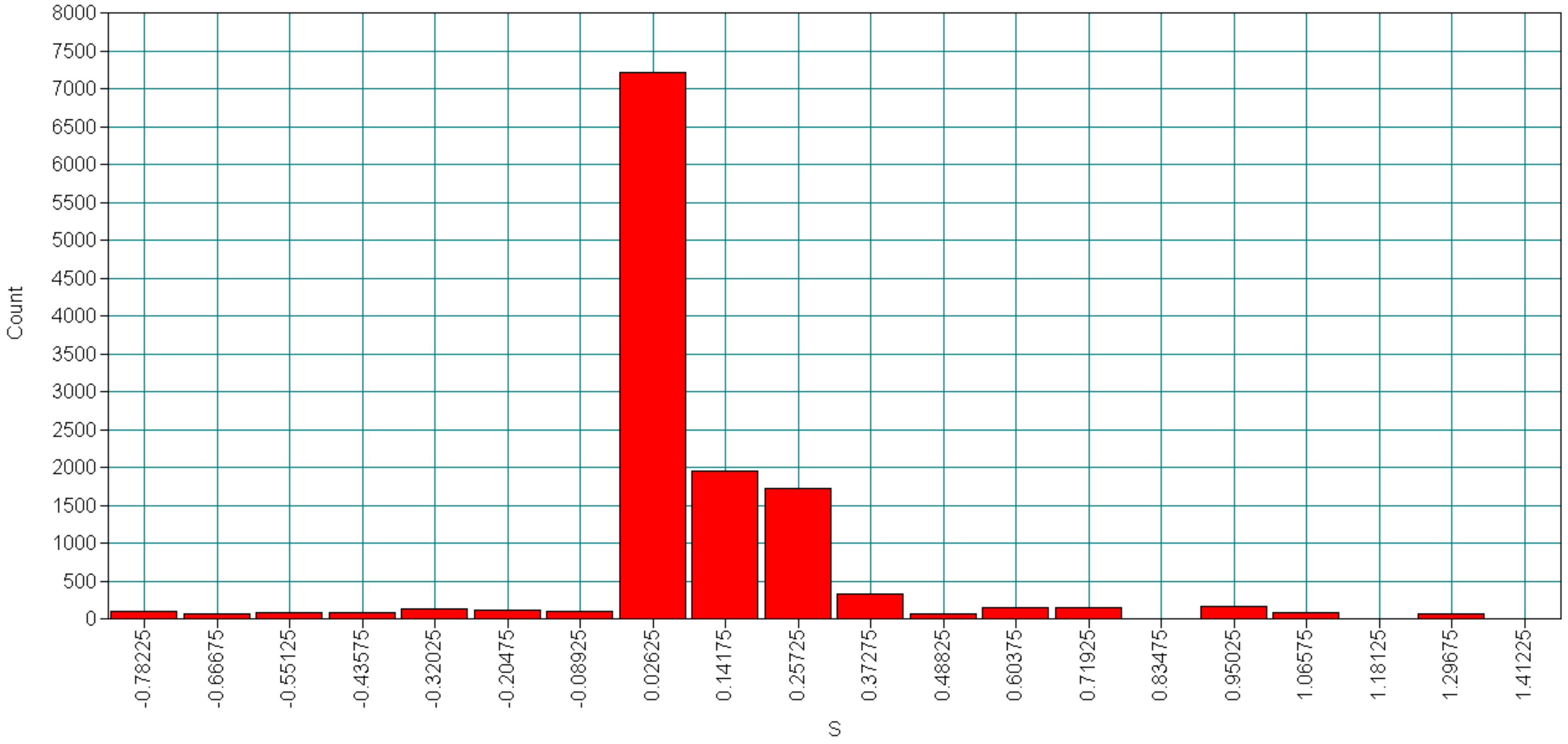


Histogram for intercept factor



Intercept (on logit scale)

Histogram for slope factor



Slope (on logit scale)

Growth modelling and binary data

- It is unusual to find a situation where an LGM can be fitted to binary data
- Intercept/slope distributions severely non-normal
- Additional classes unlikely to rectify this problem

- Favour an approach which uses additional classes rather than continuous factors to capture response heterogeneity
  - LCGA (Latent Class Growth Analysis)
  - LLCA (Longitudinal Latent Class Analysis)

LCGA and LLCA

- Latent Class Growth Analysis (LCGA)
  - Special case of GMM
  - All growth factor variances/covariances constrained to zero
  - Subjects follow polynomial trajectories through time
  - All variability about class-specific trajectory -> error
  - LCGA gives poorer fit (but less assumptions + easier to estimate)

- Longitudinal Latent Class Analysis (LLCA)
  - Trajectories represented as a set of probabilities describing a positive response at each time point

- For binary data, LLCA is actually a special case of LCGA
  - So we'll stick with LCGA for today

Why are we bothering at all?

Person	wt_07	wt_09	wt_11	wt_13	wt_15
1	24.8	29.6	35.2	41.3	55.5
2	23.2	33.4	38	41.6	48.8
3	23.4	31.2	36.6	45.9	60.7
4	28	37.6	50.6	61	64.8
5	26.6	33.6	40.8	45.9	52.3
6	25.8	33.8	40	46.1	52.9
7	27.2	35.4	45.4	48.2	70.4
8	24.8	30.8	36.6	42.3	51.9
9	22.6	28.8	36	41.3	56
10	27.2	48.4	63	71.4	68.9
11	22.8	30.8	37	38.3	53.5
12	31.6	45.2	55.4	64.3	74.8
13	24.8	30.8	39.2	39.5	50.3
14	37.4	48.8	60.2	69.1	68.5
15	20.4	23.6	31.6	39.8	53.2

Why are we bothering at all?

- Repeated continuous are measures complex
- As many response patterns as respondents

- Repeated binary data is relatively simple
- Multiple respondents with any particular pattern

Binary data -> Response patterns

111111 = Yes at all six time points

000000 = No at all six time points

110000 = Yes early on, followed by no

101010 = Alternating pattern

How many are there?

Frequency of response patterns

- The complete-case dataset is dominated by a couple of response patterns
- If there aren’t many patterns then this is all a bit pointless – just use the patterns themselves as a variable e.g.
  - 00 (none)
  - 01 (late)
  - 10 (early)
  - 11 (persistent)

```
+-----------------+
| pattern       z |
|-----------------|
|  000000    5033 |
|  111111     782 |
|  000001     102 |
|  011111      74 |
|  111101      66 |
|-----------------|
|  000011      54 |
|  001111      50 |
|  111110      49 |
|  110111      39 |
|  000111      36 |
|-----------------|
|  010000      35 |
|  110000      34 |
|  111011      34 |
|  111000      32 |
|  101111      29 |
|-----------------|
|  100000      26 |
|  000010      22 |
+-----------------+
```

Quick exercise

- If your data consisted of the 17 complete-case patterns with 20+ observations, how might you group these women yourself?

Quick exercise

- If your data consisted of the 17 complete-case patterns with 20+ observations, how might you group these women yourself?
  - It would probably depend on your hypothesis
  - If you weren't interested in timing you might just add up the number of YES's
  - May be forced to discard some unusual patterns

Quick exercise

- If your data consisted of the 17 complete-case patterns with 20+ observations, how might you group these women yourself?
  - It would probably depend on your hypothesis
  - If you weren’t interested in timing you might just add up the number of YES’s
  - May be forced to discard some unusual patterns

- A mixture modelling approach allows you to extract the strongest signals in the data
- Let’s you work probabilistically to reflect class assignment uncertainty

What about patterns with 1 missing value

```
     +----------------+
     | pattern      z |
     |----------------|
  4. |  00000.    469 |
  7. |  000.00    278 |
 12. |  0000.0    173 |
 13. |  00.000    164 |
 14. |  11111.    126 |
     |----------------|
 18. |  0.0000     89 |
 25. |  11.111     56 |
 27. |  111.11     55 |
 28. |  1111.1     55 |
 34. |  .00000     47 |
     |----------------|
 45. |  .11111     32 |
 47. |  1.1111     29 |
     +----------------+
```

You could have a bash at guessing the missing value for these women

They appear to be either persistent non-smokers or persistent smokers.

Ideally want to account for the uncertainty since we can’t be sure

Particularly when more missing data added

2 missing

```
     +---------------+
     | pattern     z |
     |---------------|
  9. |  0000..   203 |
 15. |  000.0.   113 |
 21. |  000..0    70 |
 24. |  1111..    61 |
 26. |  00..00    55 |
     |---------------|
 30. |  00.00.    50 |
 35. |  111.1.    42 |
 42. |  ..0000    34 |
 51. |  111..1    27 |
 54. |  0..000    24 |
     |---------------|
 56. |  0.000.    23 |
 59. |  00.0.0    22 |
 62. |  11.11.    21 |
     +---------------+
```

3 missing

```
     +---------------+
     | pattern     z |
     |---------------|
  8. |  000...   277 |
 16. |  111...   105 |
 36. |  00..0.    40 |
 38. |  00...0    38 |
 39. |  00.0..    38 |
     |---------------|
 55. |  0...00    23 |
 57. |  11..1.    23 |
 58. |  ...000    22 |
 65. |  0.00..    20 |
 67. |  11.1..    20 |
     +---------------+
```

4 missing

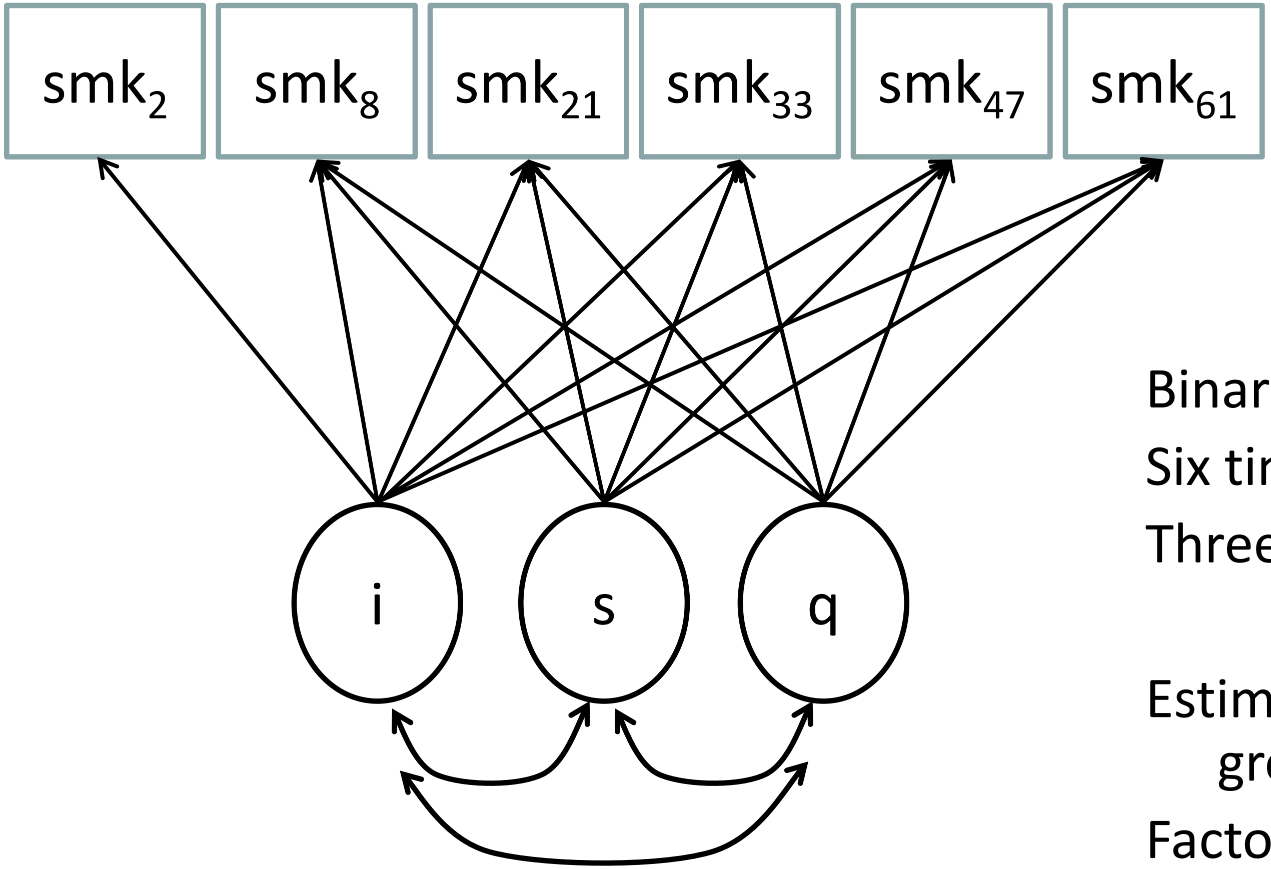
```
     +---------------+
     | pattern     z |
     |---------------|
  5. |  00....   349 |
 11. |  11....   189 |
 33. |  0.0...    48 |
 49. |  .00...    28 |
 50. |  1.1...    27 |
     |---------------|
 64. |  0..0..    20 |
 66. |  01....    20 |
     +---------------+
```

It would be a shame to have to throw all this data away

So, why are we bothering?

- Mixture modelling for binary data
  - is exploratory with the aim of simplifying a complex set of measures
  - assumes data is due to a number of unmeasured subpopulations
  - Can deal with partial non-response
  - Robust to the odd bit of mis-response (unless everyone is lying about their consumption)
  - Gives us nice pretty pictures (see later)
- LCA will always extract groups even if no such subpopulations exist in reality

So, back to the LGM…



Binary data

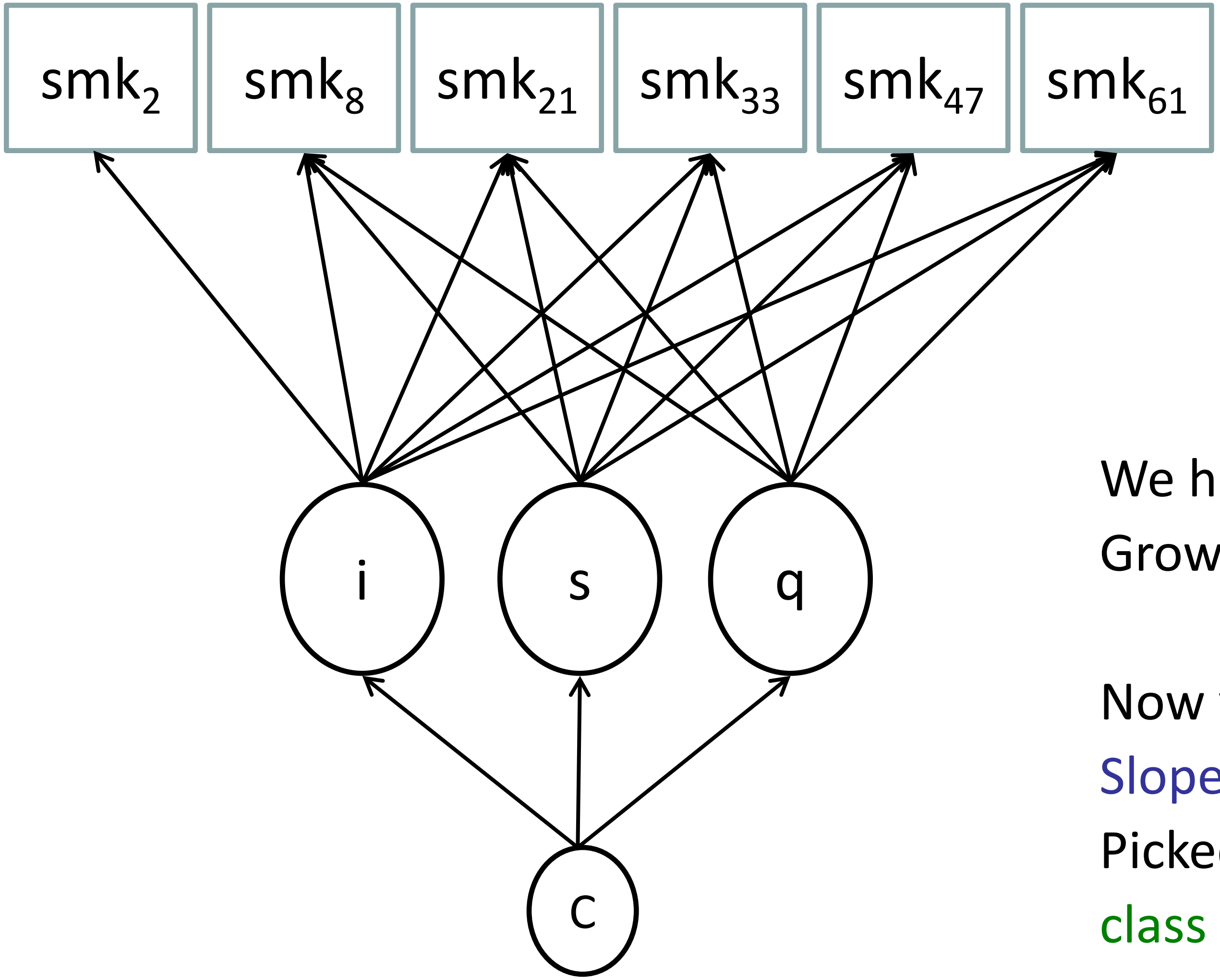
Six time points

Three growth factors (i/s/q)

Estimating co/variance for growth factors

Factors far from normal

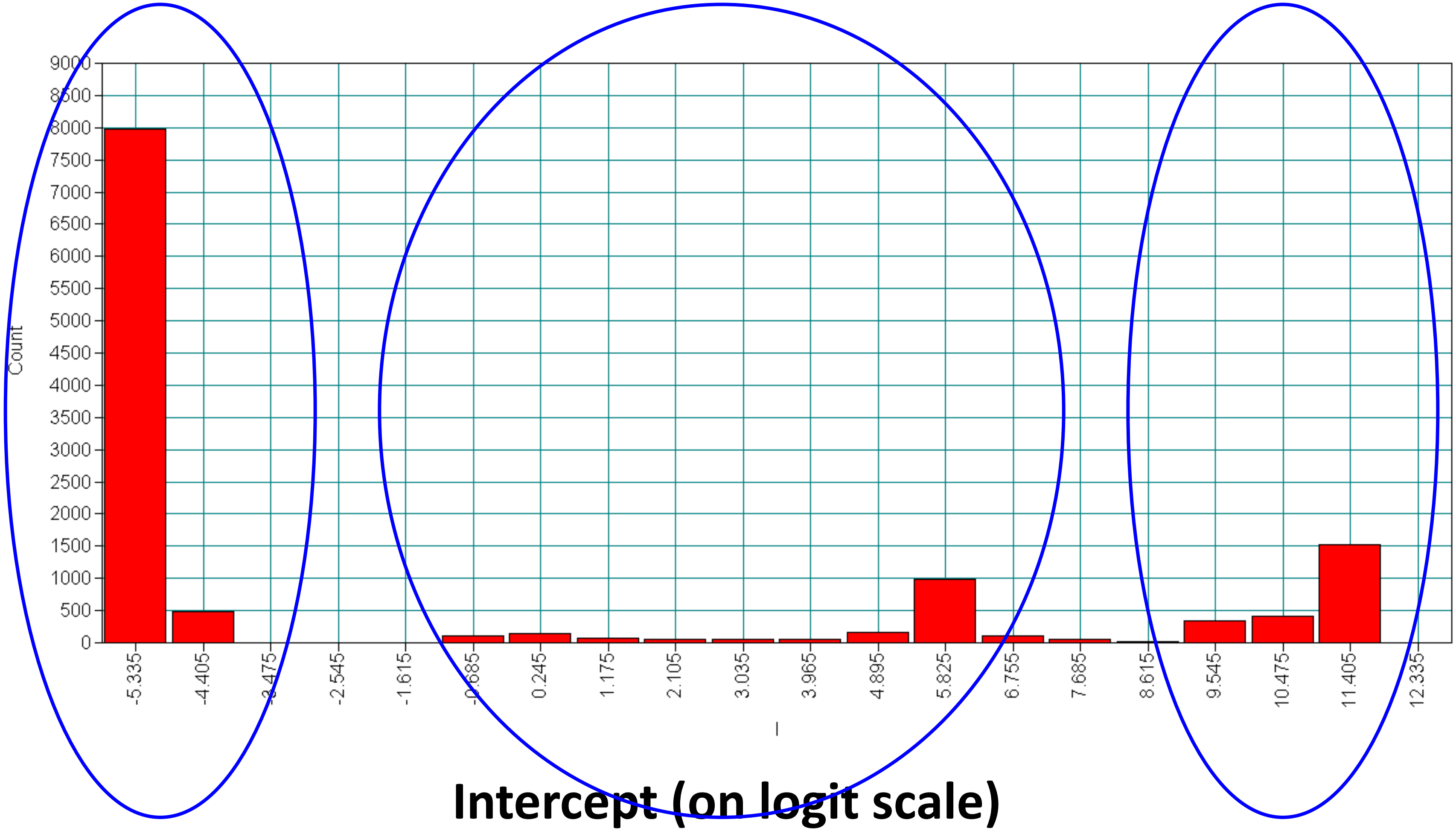
Now add a mixture



We have removed the Growth factor co/variances

Now variability in Intercept, Slope and Quadratic will be Picked up by the latent class variable C

Recall the intercept from earlier



Intercept (on logit scale)

How do we do this?

```
Variable:
  Names are <snip>;
  Missing are all (-9999) ;
  classes = c(3);
  usevariables = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;
  categorical  = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;

Analysis:
  type=mixture ;

Model:
  %overall%
  i s q | msmk2@0.17 msmk8@0.67 msmk21@1.8 msmk33@2.8 msmk47@3.9 msmk61@5.1;

  [msmk2$1@0 msmk8$1@0 msmk21$1@0 msmk33$1@0 msmk47$1@0 msmk61$1@0];

  %c#1%
  [i s q];

  %c#2%
  [i s q];

  %c#3%
  [i s q];
```

How do we do this?

```
Variable:
  Names are <snip>;
  Missing are all (-9999) ;
  classes = c(3);
  usevariables = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;
  categorical  = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;

Analysis:
  type=mixture ;

Model:
  %overall%
  i s q | msmk2@0.17 msmk8@0.67 msmk21@1.8 msmk33@2.8 msmk47@3.9 msmk61@5.1;

  [msmk2$1@0 msmk8$1@0 msmk21$1@0 msmk33$1@0 msmk47$1@0 msmk61$1@0];

  %c#1%
  [i s q];

  %c#2%
  [i s q];

  %c#3%
  [i s q];
```

Loadings are in years

Growth factor means varying across classes

Ignore this bit!

Model results for 3-class LCGA

```
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Latent Class 1 (15.8%)

Means
    I                  2.616      0.233     11.226      0.000
    S                  0.399      0.186      2.145      0.032
    Q                 -0.081      0.030     -2.750      0.006


Latent Class 2 (8.9%)

Means
    I                 -0.427      0.129     -3.313      0.001
    S                  0.070      0.107      0.660      0.510
    Q                  0.009      0.017      0.510      0.610


Latent Class 3 (75.4%)

Means
    I                 -5.130      0.265    -19.332      0.000
    S                 -1.499      0.325     -4.607      0.000
    Q                  0.334      0.063      5.345      0.000
```

Model results for 3-class LCGA

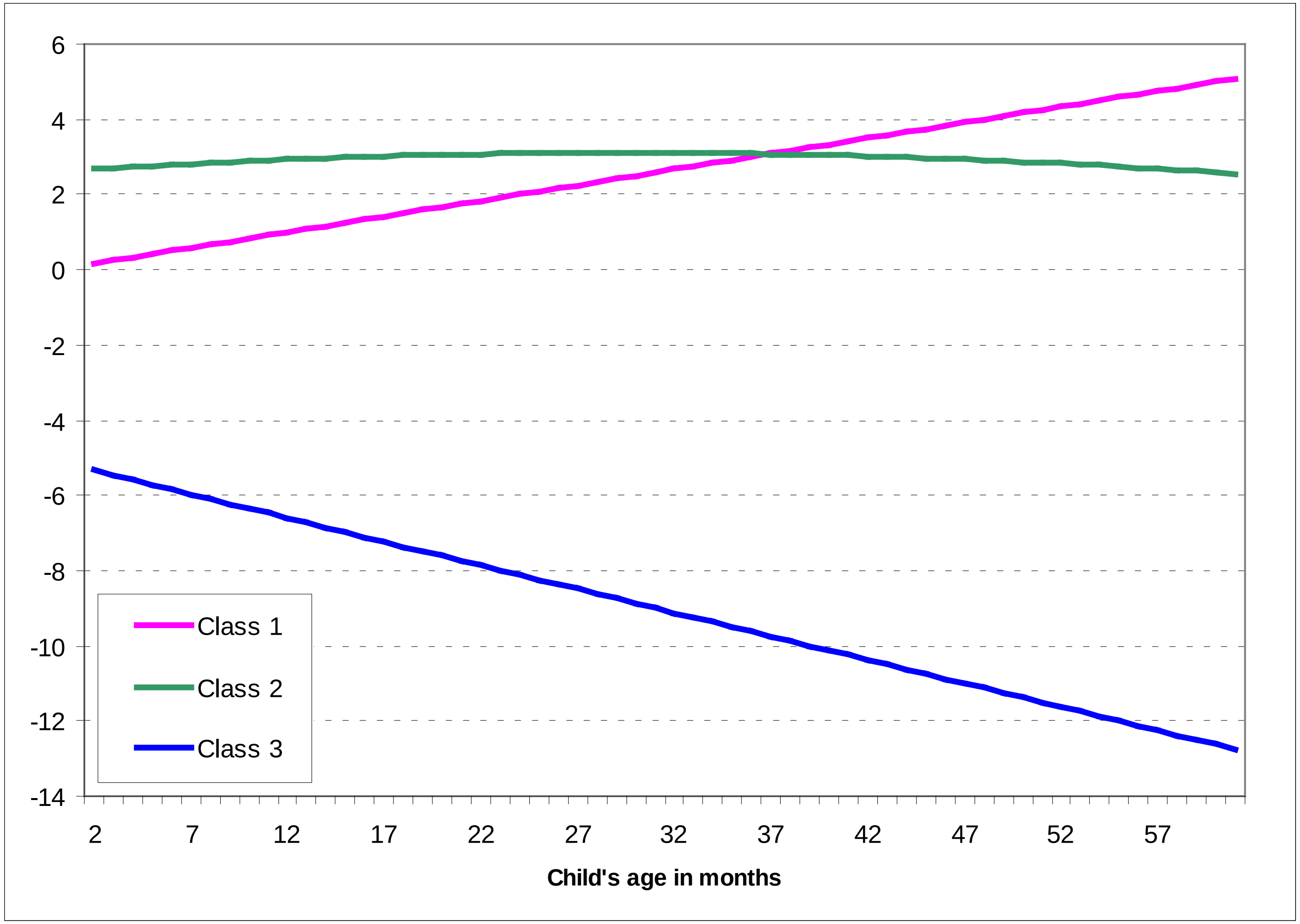
```
                                   Estimate

Latent Class 1 (15.8%)

Means
    I                                2.616
    S                                0.399
    Q                               -0.081


Latent Class 2 (8.9%)

Means
    I                               -0.427
    S                                0.070
    Q                                0.009


Latent Class 3 (75.4%)

Means
    I                               -5.130
    S                               -1.499
    Q                                0.334
```

Instead of continuous growth factors with funny distributions

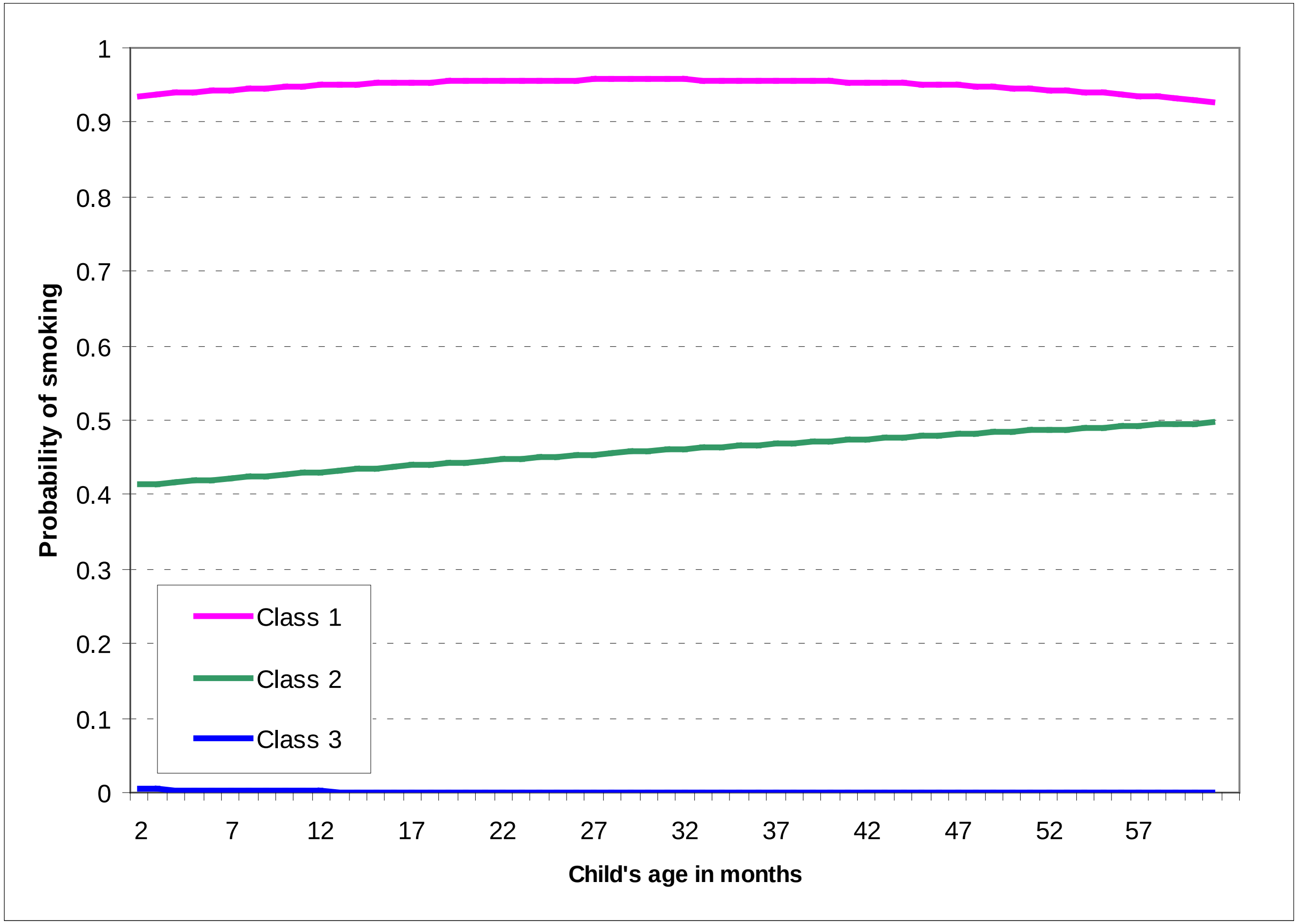
We have three discrete growth factors each described by a mass at three distinct points

Before we come on to issues of whether this model fits, what kind of “growth” do these three classes of individuals exhibit?

Three quadratic growth trajectories



These make a lot more sense in probability space



Persistent smokers?

Dabblers?

Non smokers?

Pardon?

- With a binary outcome we are fitting polynomials in logit space i.e. ln(p/1-p)

- Much easier to interpret in probability space

- Trajectories may no longer appear polynomial

Practical example

1. Use the maternal smoking data to fit a 4-class quadratic LCGA model

2. Look at the estimated means for I/S/Q

3. Plot a trajectory model in probability space to help interpret the classes

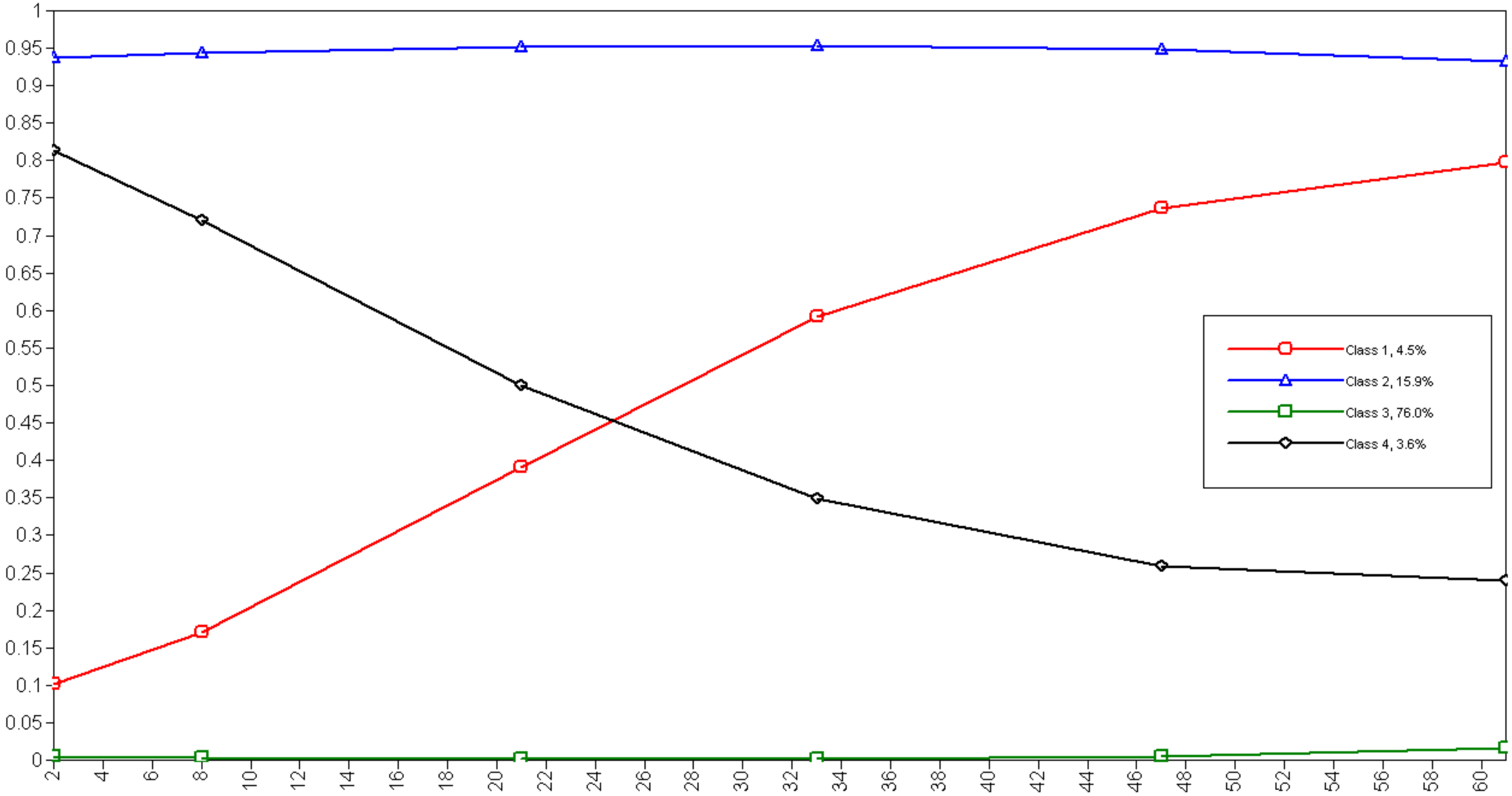
4. We are concentrating on the complete-case dataset for today

Summary of findings from practical

- It appears that adding extra classes is leading to a dissection of the mothers who report smoking at some time

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT
  CLASSES BASED ON THE ESTIMATED MODEL

    Latent classes

       1         307.21670           4.5%
       2        1090.42003          15.9%
       3        5204.78942          76.0%
       4         248.57385           3.6%
```

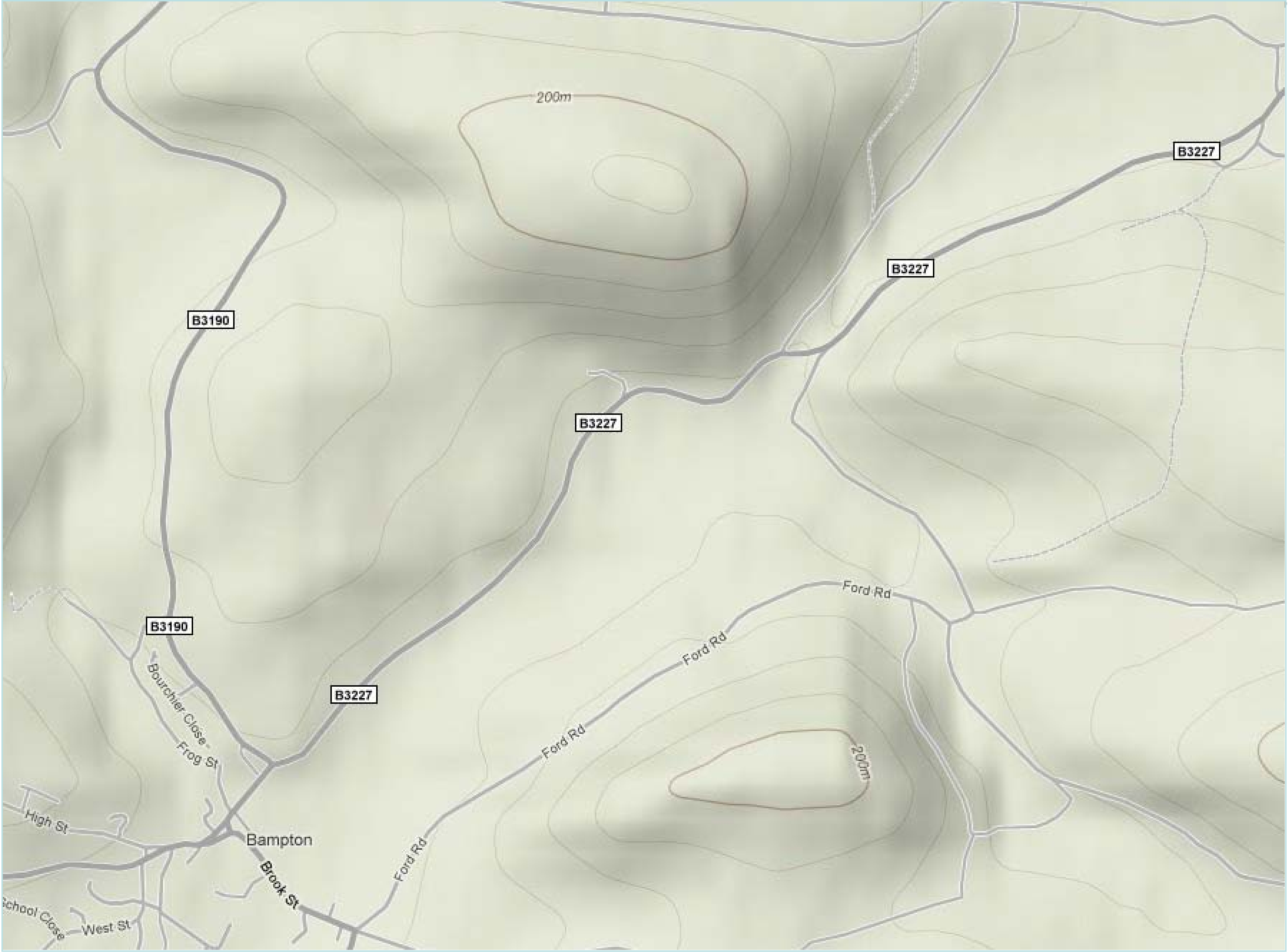
Resulting 4 trajectories

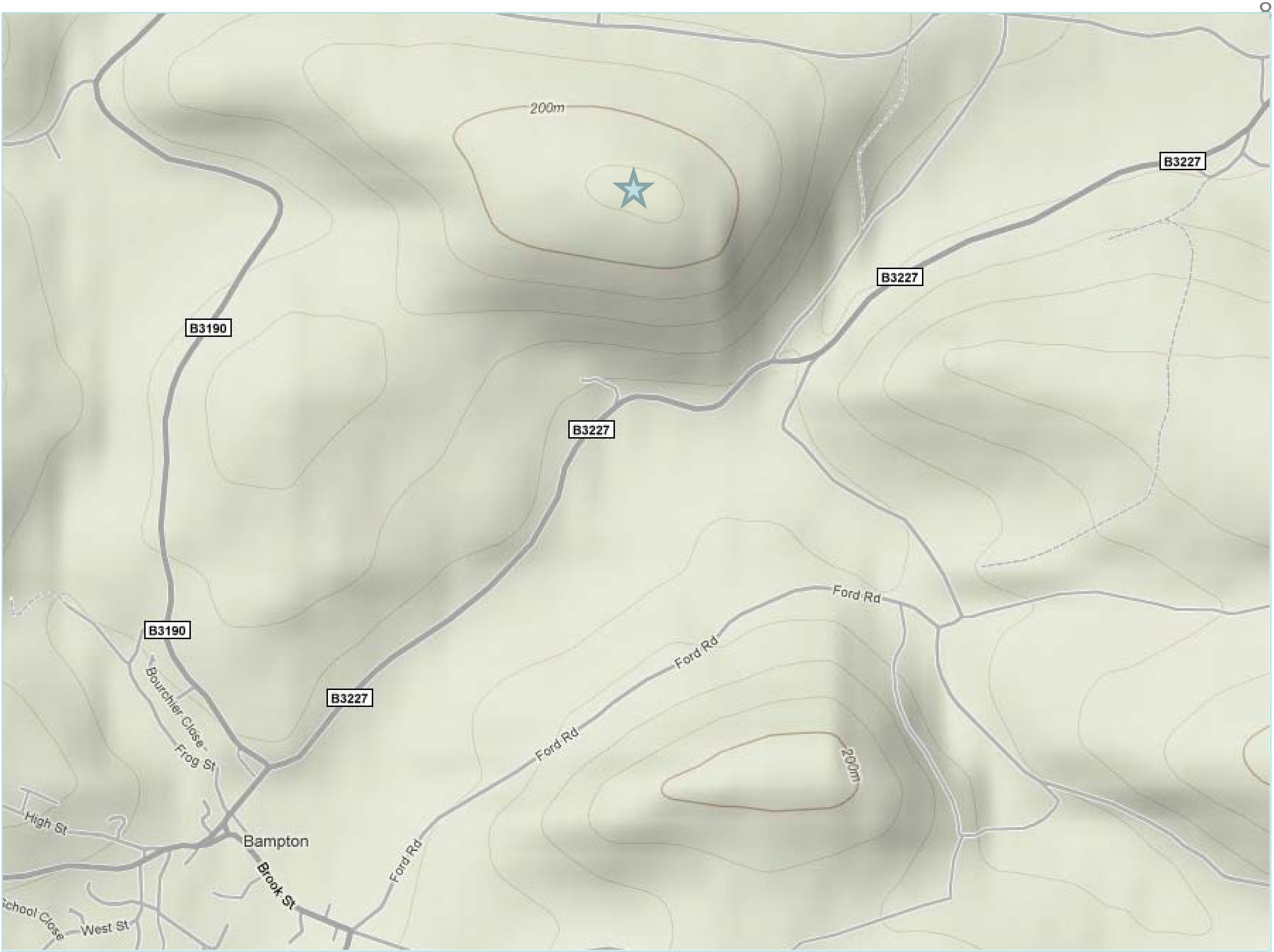


- **Dabblers have split into an offset and onset group**

Parameter Starting Values

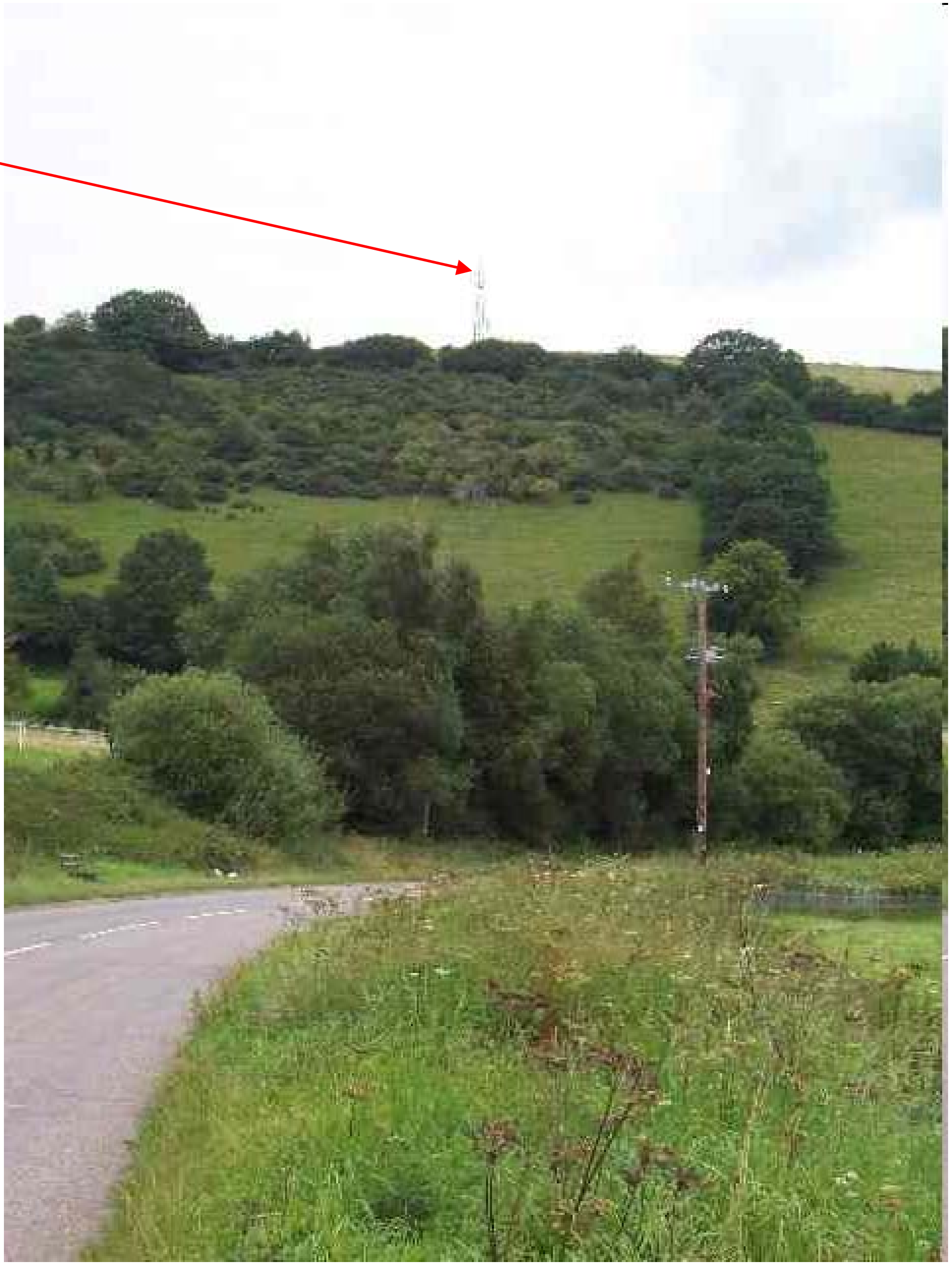
and parachuting over Devon


200m
B3227
B3227
B3190
B3227
B3190
B3227
Ford Rd
Ford Rd
Ford Rd
Ford Rd
200m
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St

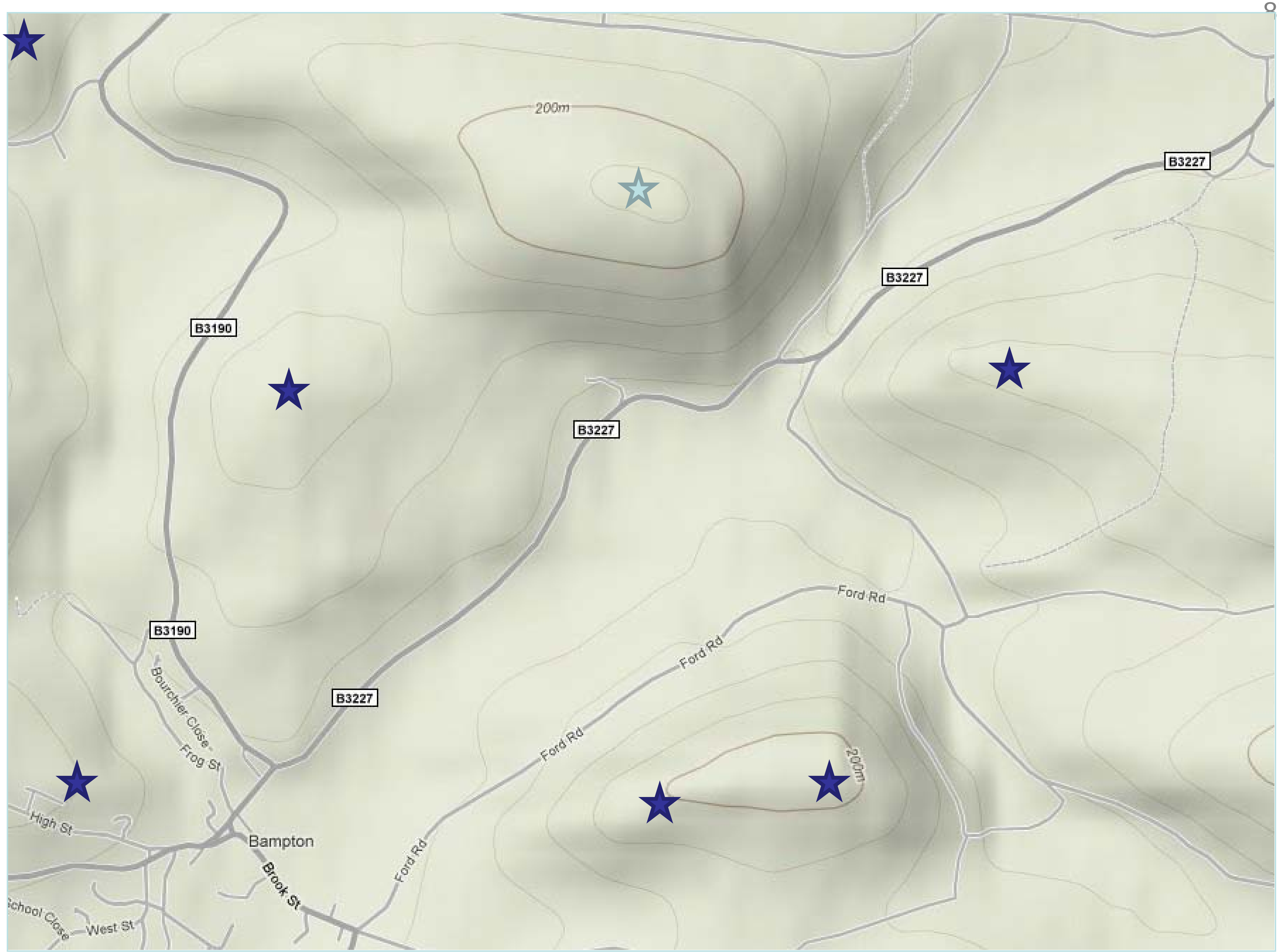


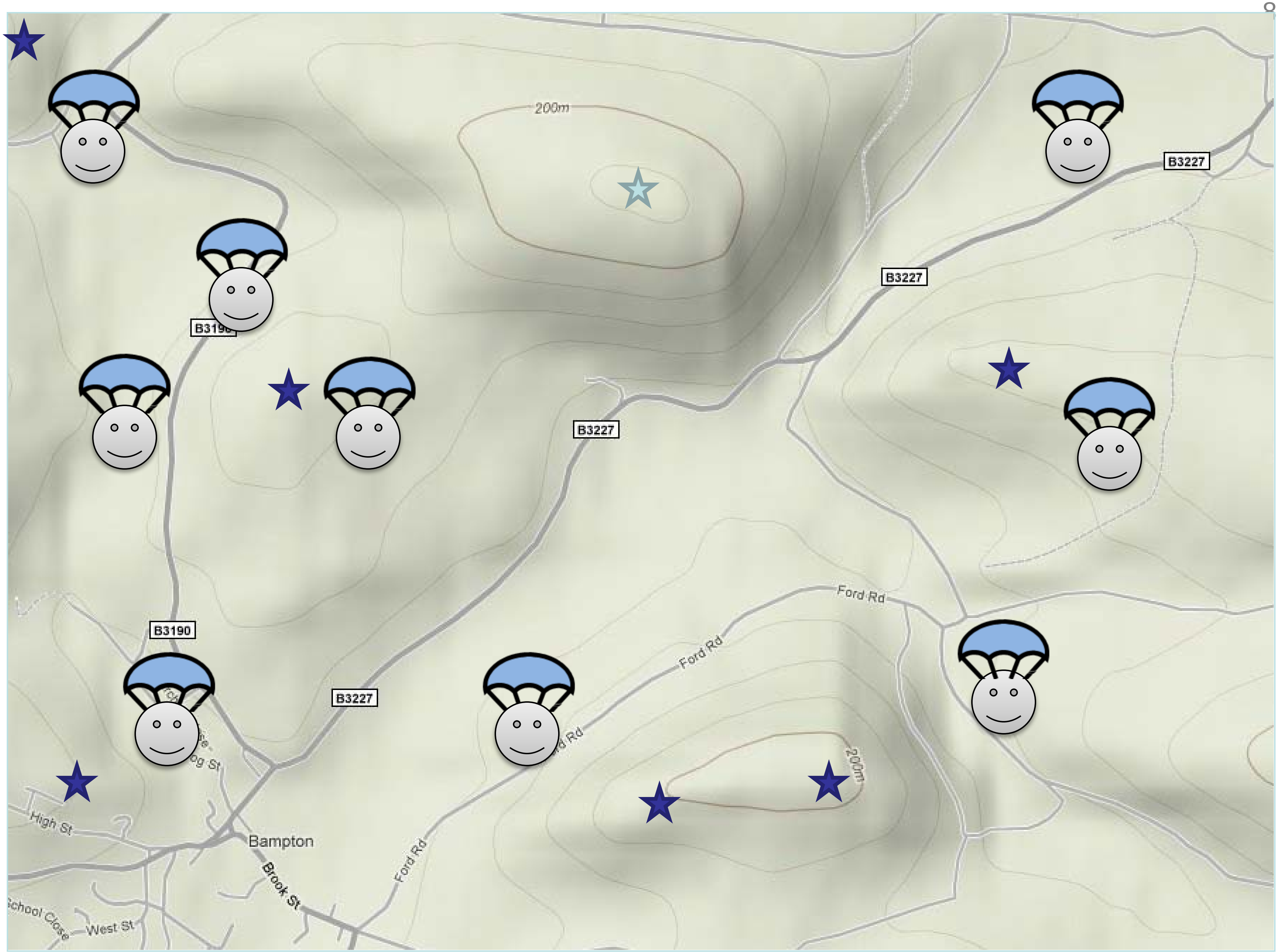
200m
B3227
B3227
B3190
B3227
B3190
B3227
Ford Rd
Ford Rd
Ford Rd
Ford Rd
200m
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St

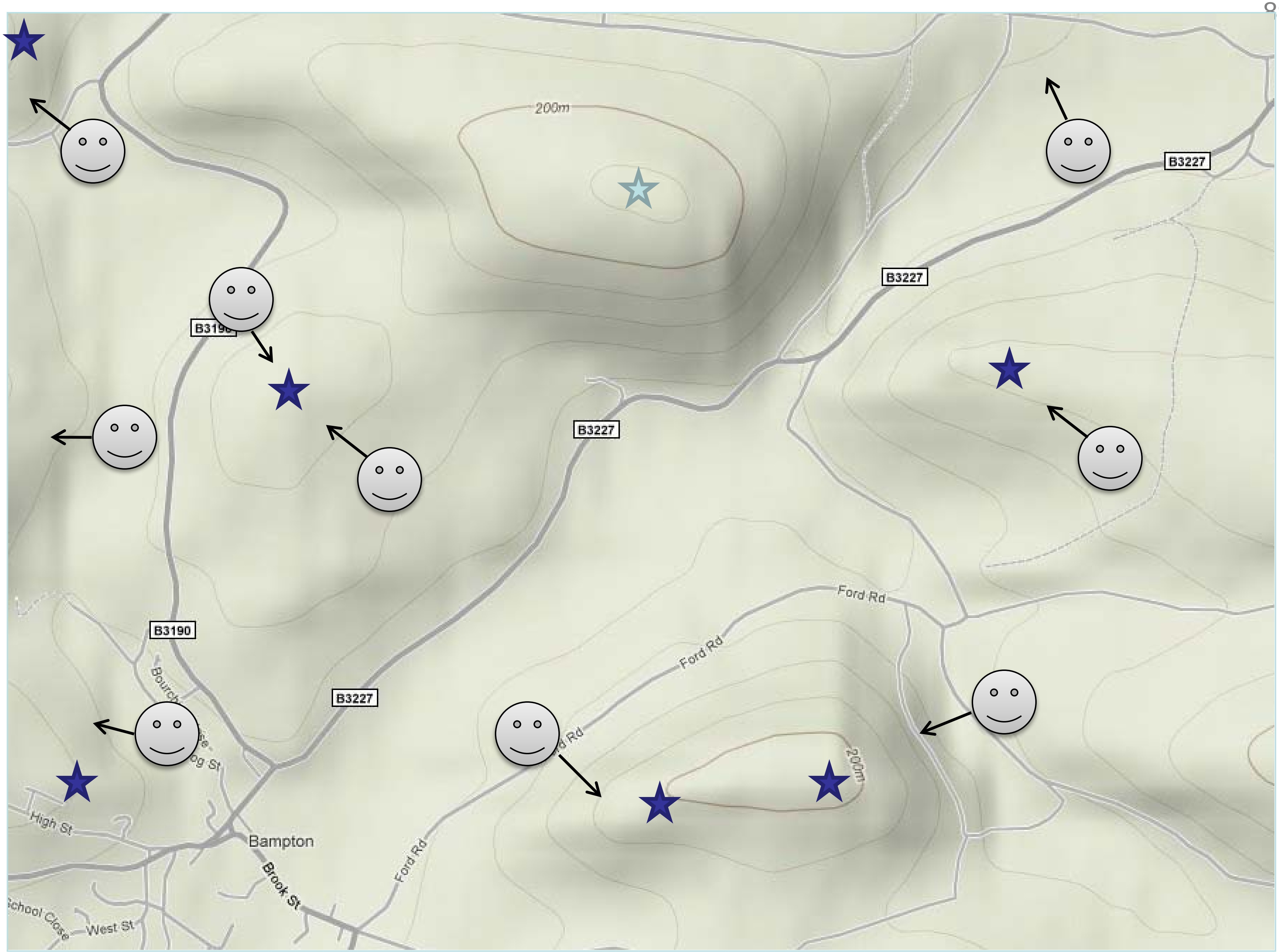
Telecom mast marks the spot

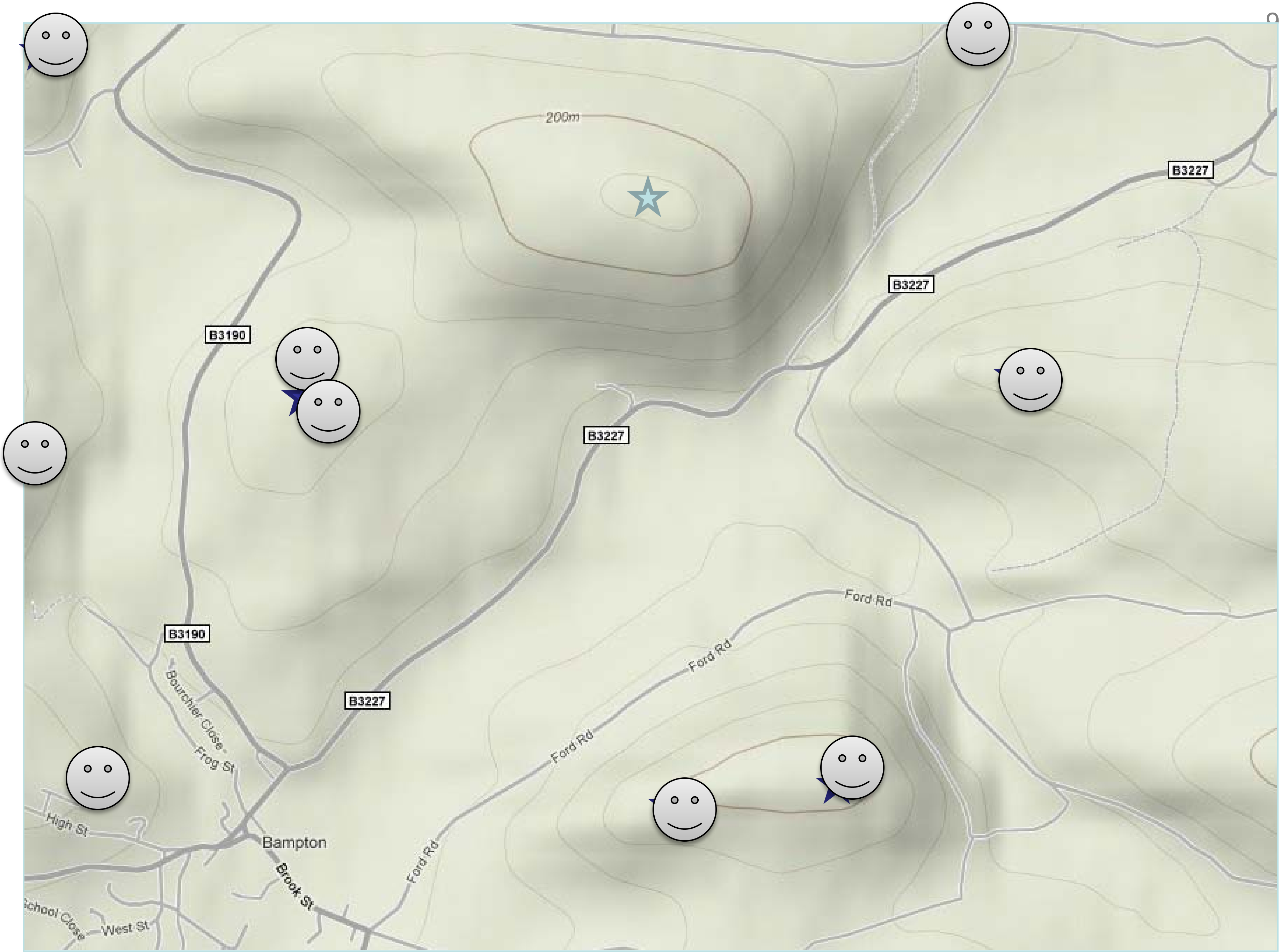


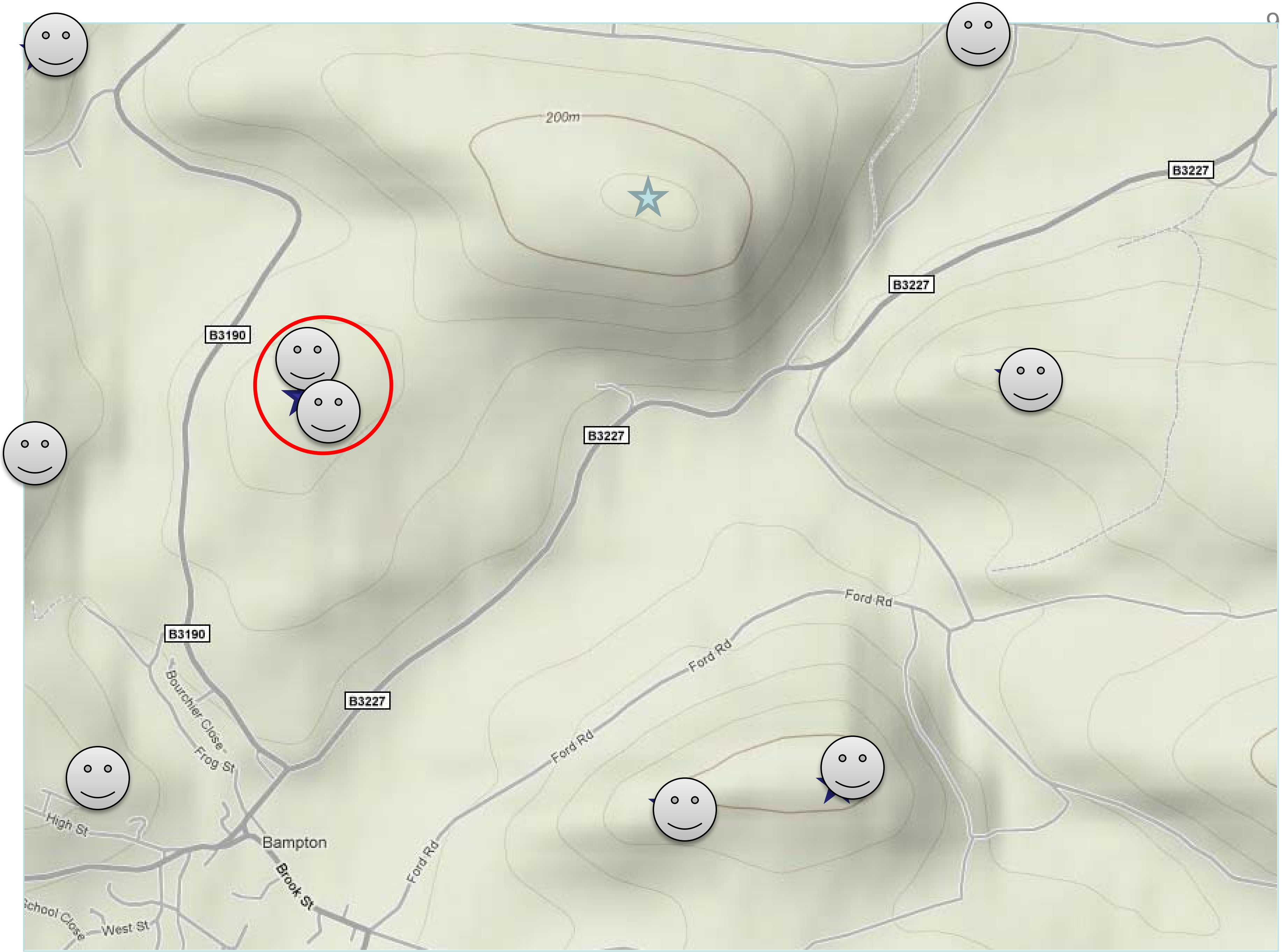
Hukeley Knap picture courtesy of t’internet

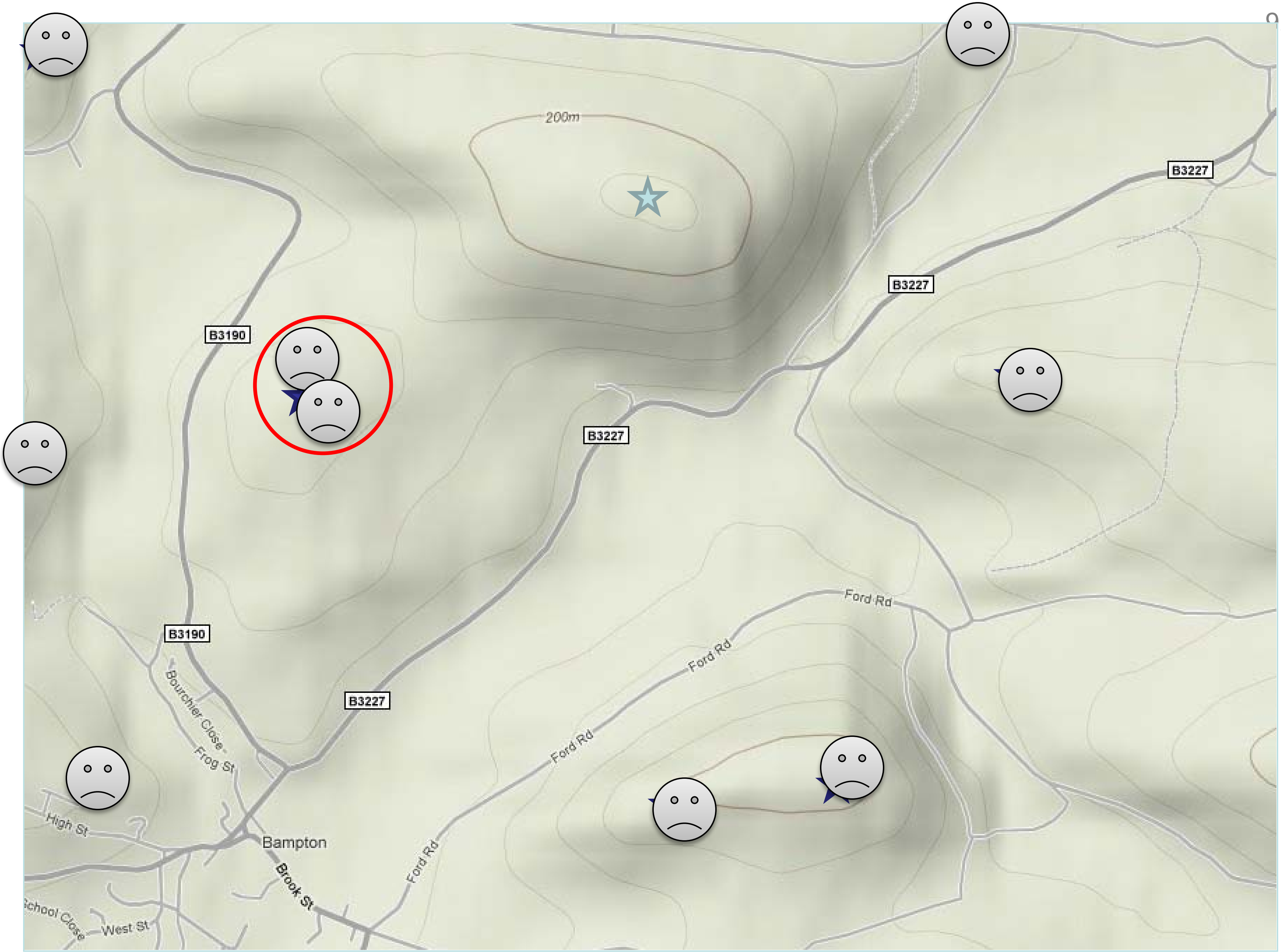

200m
B3227
B3227
B3190
B3227
B3190
B3227
Bourchier Close
Frog St
Ford Rd
Ford Rd
Ford Rd
Ford Rd
200m
High St
Bampton
Brook St
School Close
West St

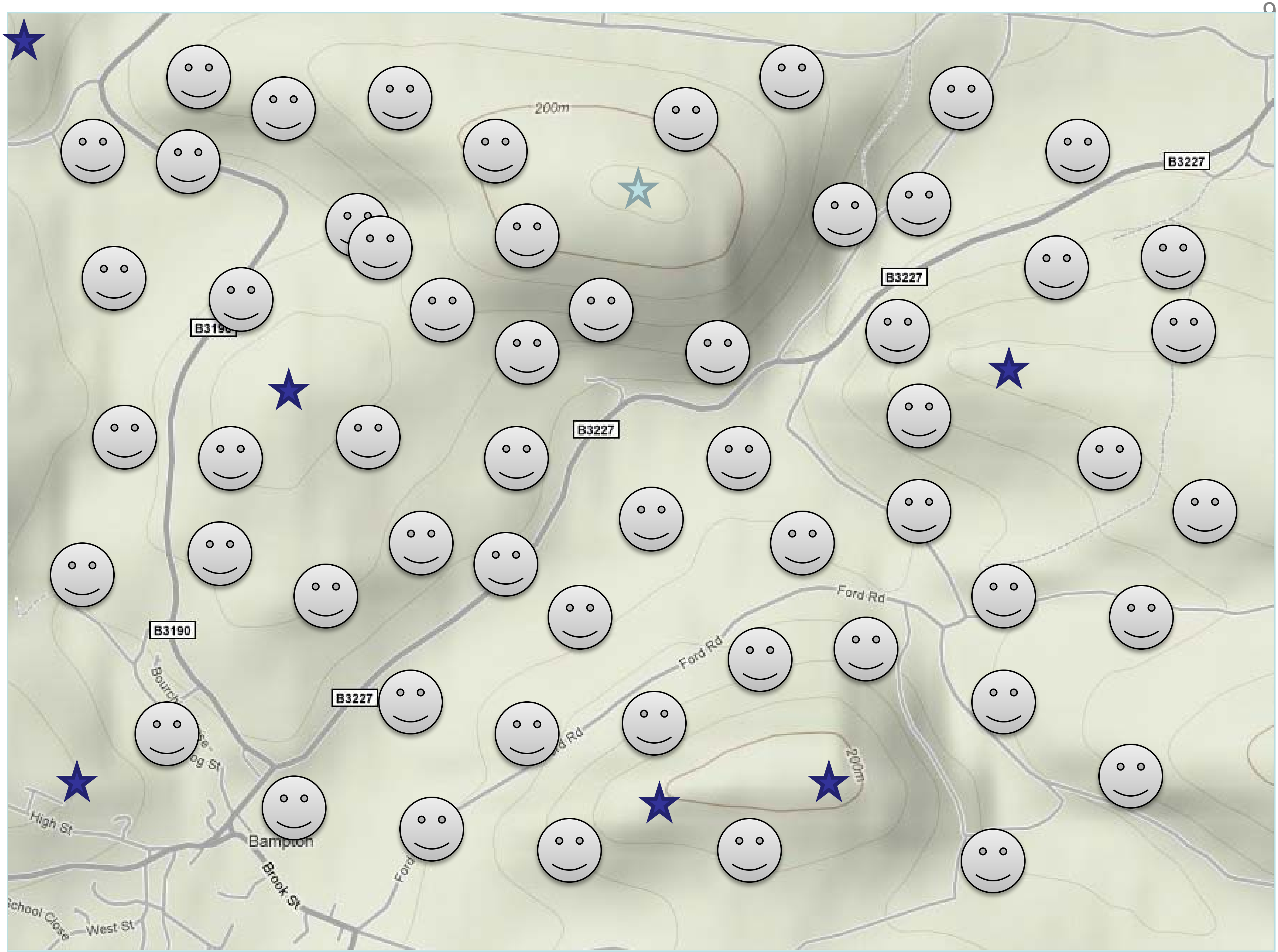

200m
B3227
B3227
B3190
B3227
B3190
B3227
Ford Rd
Ford Rd
Rd
200m
Bampton
Brook St
Ford Rd
High St
School Close
West St
Fore St
Frog St

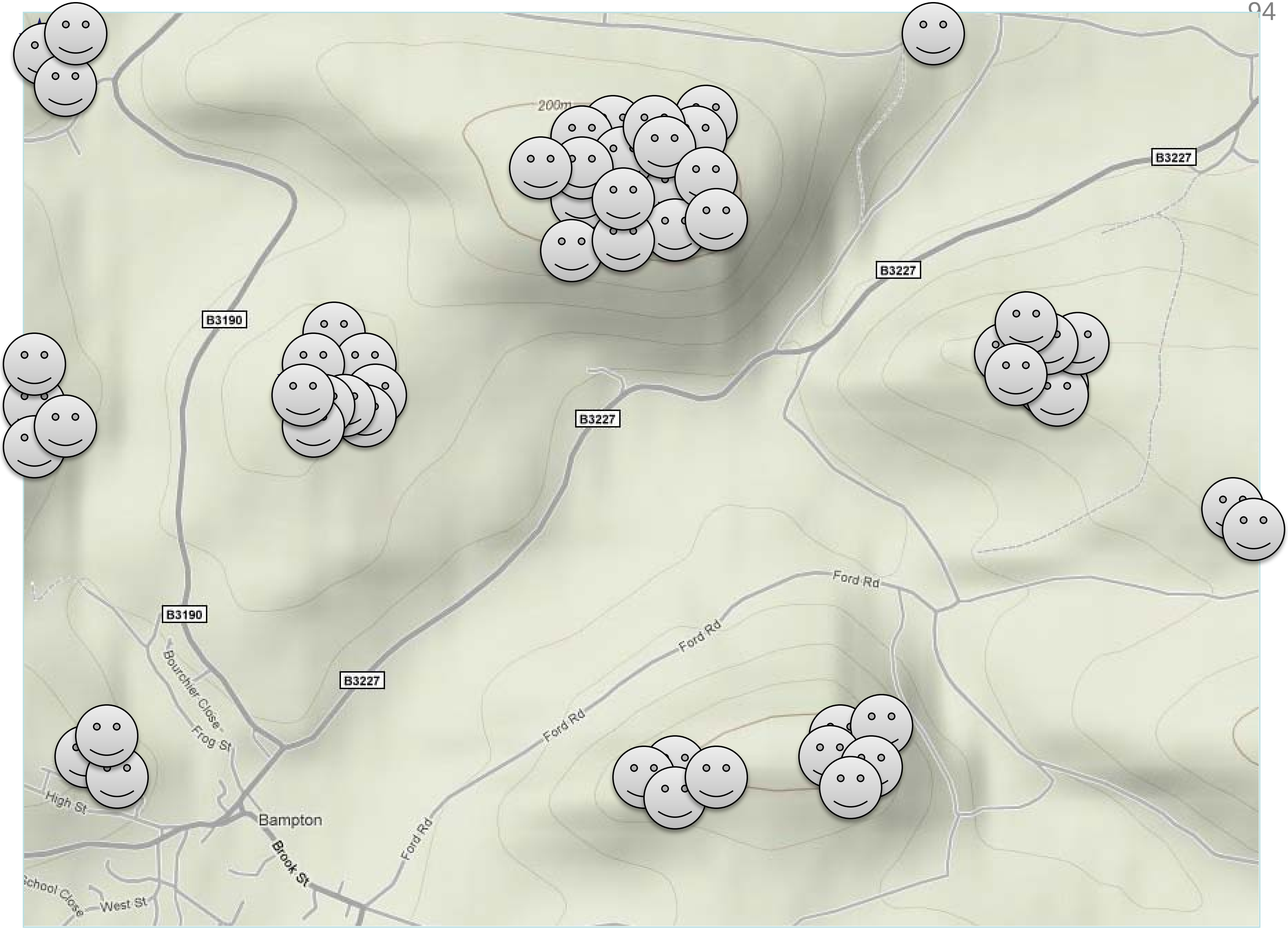

200m
B3227
B3227
B3190
B3227
B3190
B3227
Bourchier Close
Frog St
Bampton
High St
Brook St
School Close
West St
Ford Rd
Ford Rd
Ford Rd
Ford Rd
200m

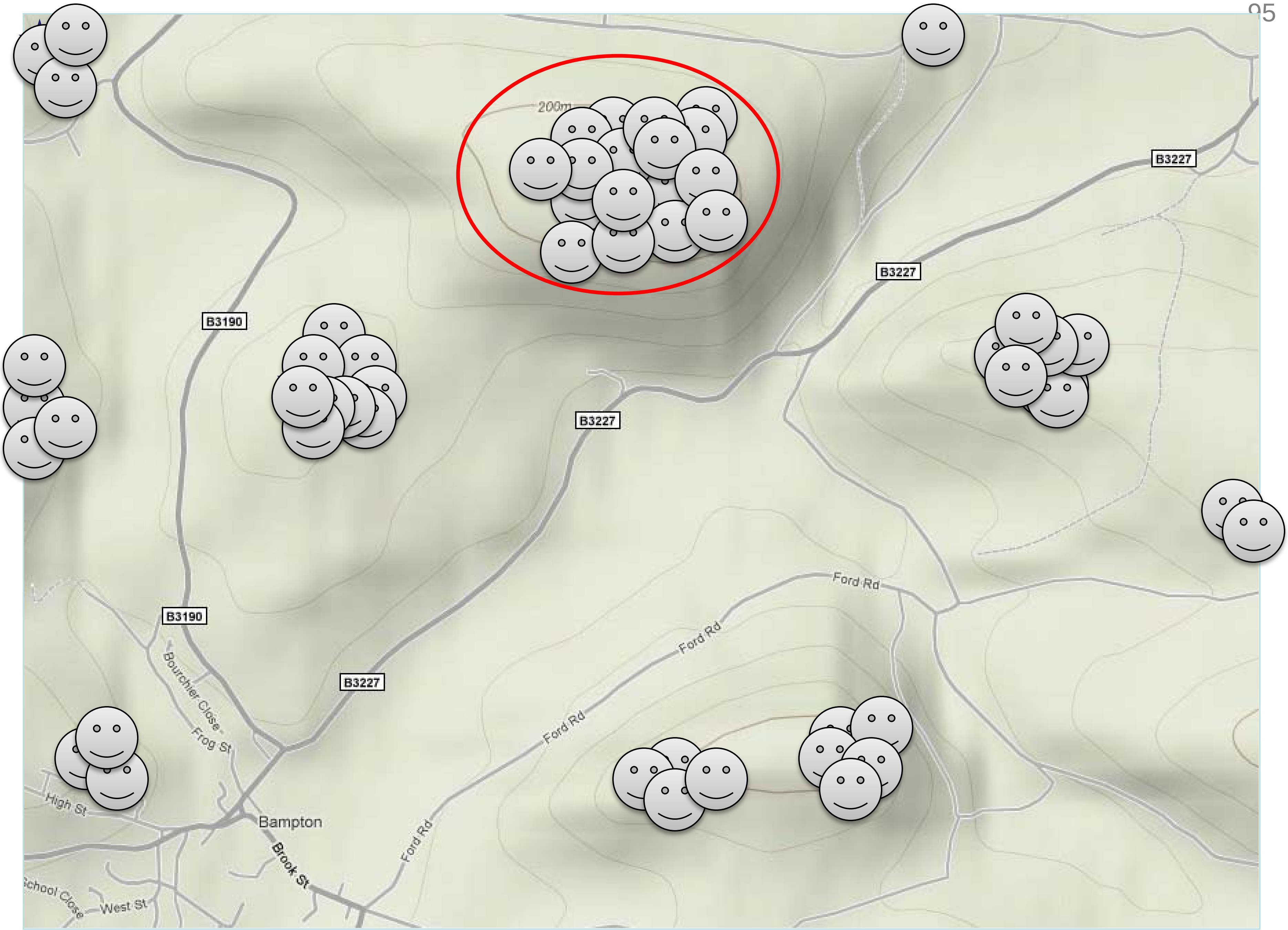

200m
B3227
B3227
B3227
B3227
B3190
B3190
Ford Rd
Ford Rd
Ford Rd
Ford Rd
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St

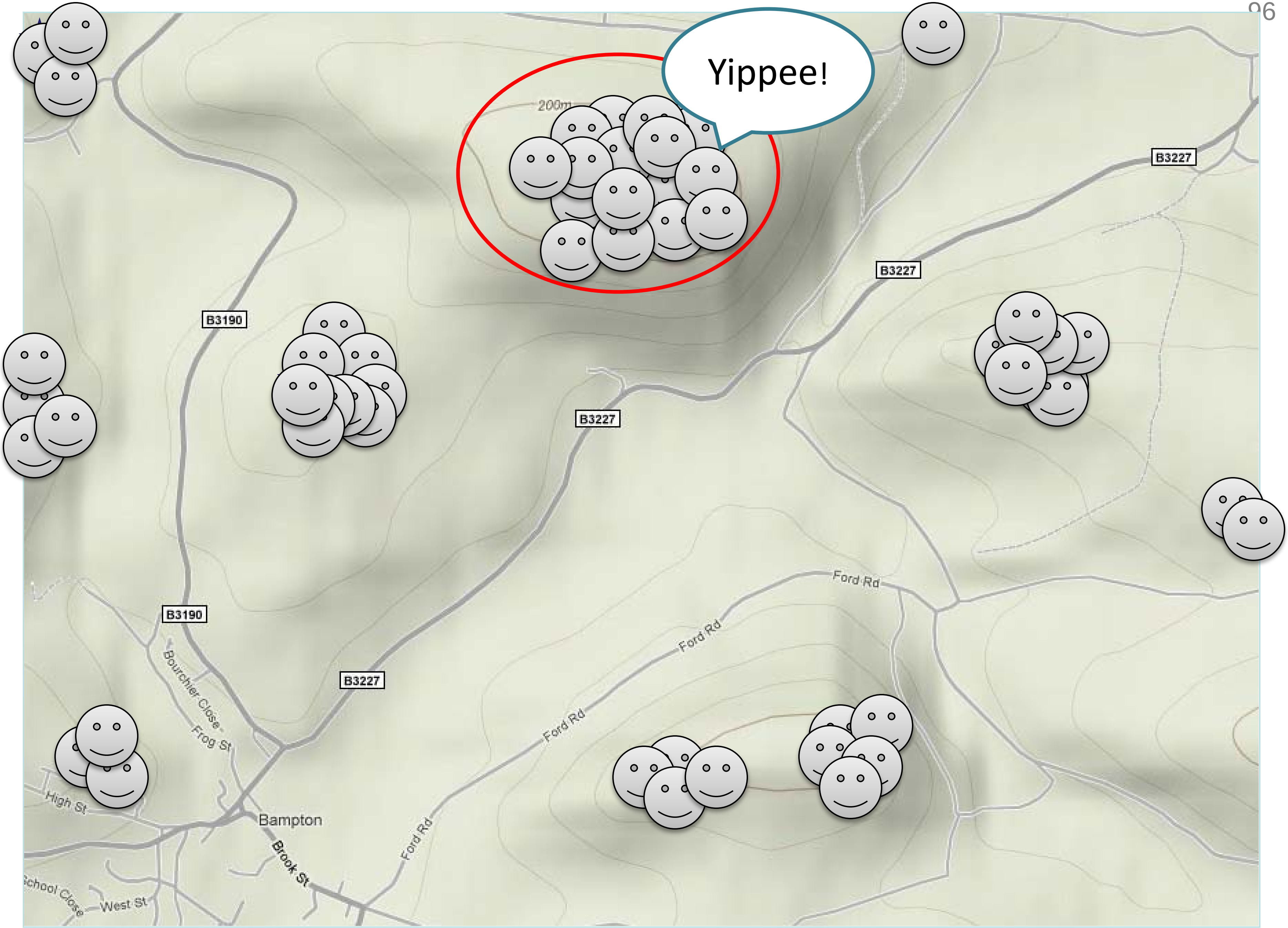

200m
B3227
B3227
B3227
B3227
B3190
B3190
Ford Rd
Ford Rd
Ford Rd
Ford Rd
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St


200m
B3227
B3227
B3190
B3227
B3190
B3227
Ford Rd
Ford Rd
Ford Rd
Ford Rd
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St


200m
B3227
B3227
B3190
B3227
Ford Rd
Ford Rd
B3190
Bourch
B3227
Rd
200m
High St
Bampton
Brook St
Ford
School Close
West St


200m
B3227
B3227
B3190
B3227
B3190
B3227
Ford Rd
Ford Rd
Ford Rd
Ford Rd
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St


200m
B3190
B3190
B3227
B3227
B3227
B3227
B3227
Ford Rd
Ford Rd
Ford Rd
Ford Rd
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St



Yippee!
200m
B3227
B3227
B3227
B3227
B3190
B3190
Ford Rd
Ford Rd
Ford Rd
Ford Rd
Bourchier Close
Frog St
High St
Bampton
Brook St
School Close
West St

Success

```
Loglikelihood values at local maxima, seeds, and
    initial stage start numbers:

    -10148.718  987174          1689
    -10148.718  777300          2522
    -10148.718  406118          3827
    -10148.718  51296           3485
    -10148.718  997836          1208
    -10148.718  119680          4434
    -10148.718  338892          1432
    -10148.718  765744          4617
    -10148.718  636396          168
    -10148.718  189568          3651
    -10148.718  469158          1145
    -10148.718  90078           4008
    -10148.718  373592          4396
    -10148.718  73484           4058
    -10148.718  154192          3972
    -10148.718  203018          3813
    -10148.718  785278          1603
    -10148.718  235356          2878
    -10148.718  681680          3557
    -10148.718  92764           2064
```

Not there yet

```
Loglikelihood values at local maxima, seeds, and
    initial stage start numbers

    -10153.627  23688           4596
    -10153.678  150818          1050
    -10154.388  584226          4481
    -10155.122  735928          916
    -10155.373  309852          2802
    -10155.437  925994          1386
    -10155.482  370560          3292
    -10155.482  662718          460
    -10155.630  320864          2078
    -10155.833  873488          2965
    -10156.017  212934          568
    -10156.231  98352           3636
    -10156.339  12814           4104
    -10156.497  557806          4321
    -10156.644  134830          780
    -10156.741  80226           3041
    -10156.793  276392          2927
    -10156.819  304762          4712
    -10156.950  468300          4176
    -10157.011  83306           2432
```

Success

```
Loglikelihood values at local maxima, seeds, and
    initial stage start numbers:

    -10148.718  987174           1689
    -10148.718  777300           2522
    -10148.718  406118           3827
    -10148.718  51296            3485
    -10148.718  997836           1208
    -10148.718  119680           4434
    -10148.718  338892           1432
    -10148.718  765744           4617
    -10148.718  636396           168
    -10148.718  189568           3651
    -10148.718  469158           1145
    -10148.718  90078            4008
    -10148.718  373592           4396
    -10148.718  73484            4058
    -10148.718  154192           3972
    -10148.718  203018           3813
    -10148.718  785278           1603
    -10148.718  235356           2878
    -10148.718  681680           3557
    -10148.718  92764            2064
```

Not there yet

```
Loglikelihood values at local maxima, seeds, and
    initial stage start numbers

    -10153.627  23688            4596
    -10153.678  150818           1050
    -10154.388  584226           4481
    -10155.122  735928           916
    -10155.373  309852           2802
    -10155.437  925994           1386
    -10155.482  370560           3292
    -10155.482  662718           460
    -10155.630  320864           2078
    -10155.833  873488           2965
    -10156.017  212934           568
    -10156.231  98352            3636
    -10156.339  12814            4104
    -10156.497  557806           4321
    -10156.644  134830           780
    -10156.741  80226            3041
    -10156.793  276392           2927
    -10156.819  304762           4712
    -10156.950  468300           4176
    -10157.011  83306            2432
```

- **These are OPTSEEDS – they mean you can recreate a model without running all the random starts again**

How many random starts?

- Depends on
  - Sample size
  - Complexity of model
    - Number of manifest variables
    - Number of classes

- Aim to find good support for the model with the lowest likelihood, within each run

Time to think about model fit

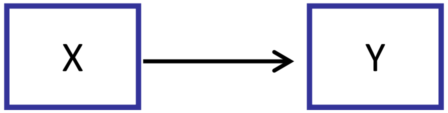
Model fit

- Aim is to add classes until some degree of fit is achieved
- Not universal agreement on what type of fit most important
- Some would say face validity more important

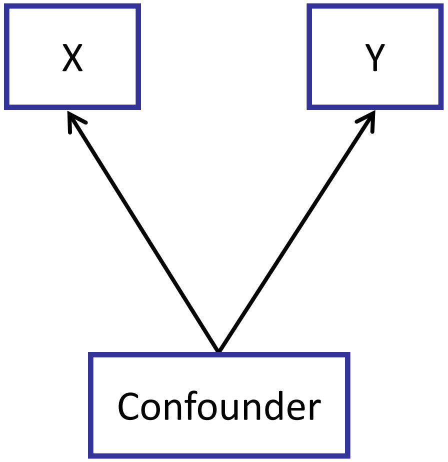
- Commonly used stopping criteria
  - Conditional Independence (Bivariate residuals)
  - Bayesian Information Criterion (BIC)
  - Entropy
  - Change in likelihood (Bootstrapping – for another day!)
  - Running out of degrees of freedom
  - Boredom

Confounding

Before



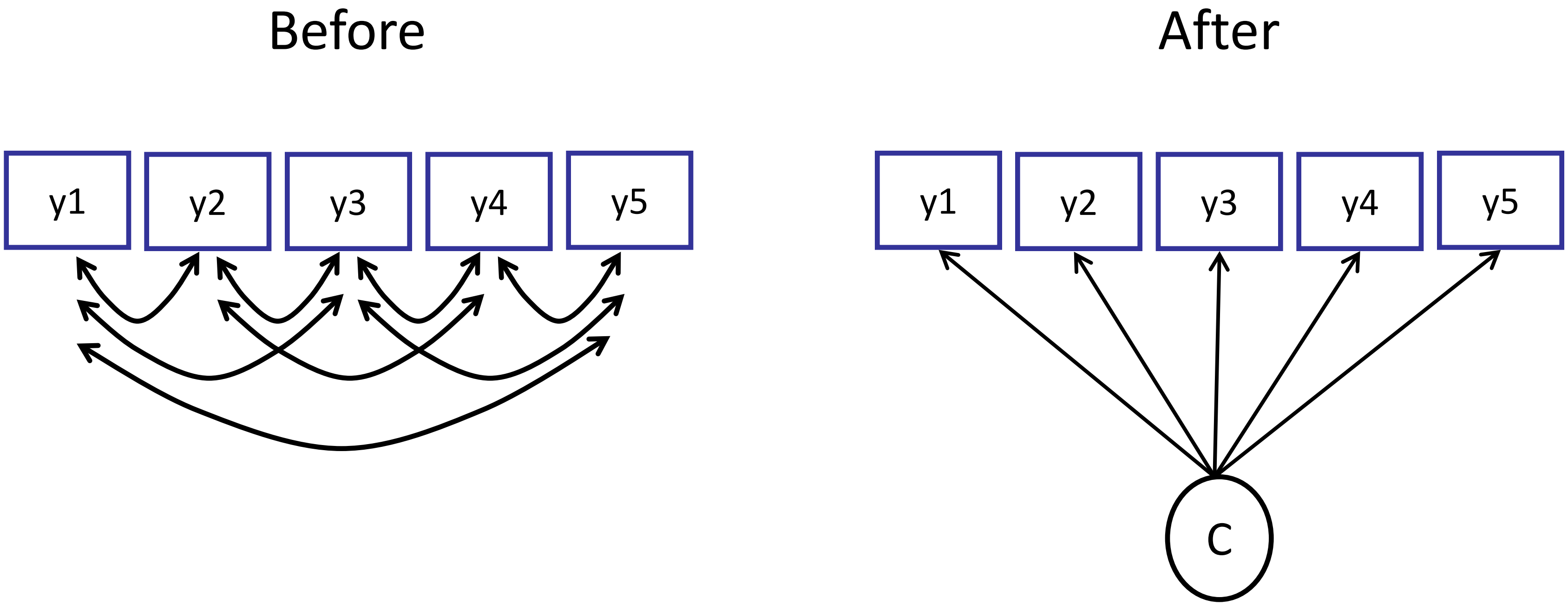
After



The confounding variable explains the association between X and Y

Conditional on C, the items are independent

Conditional independence



The latent class variable explains all the associations between items

Conditional on C, the items should be independent

Here C is latent and determined by the data itself

Bivariate residuals – come from “tech10”

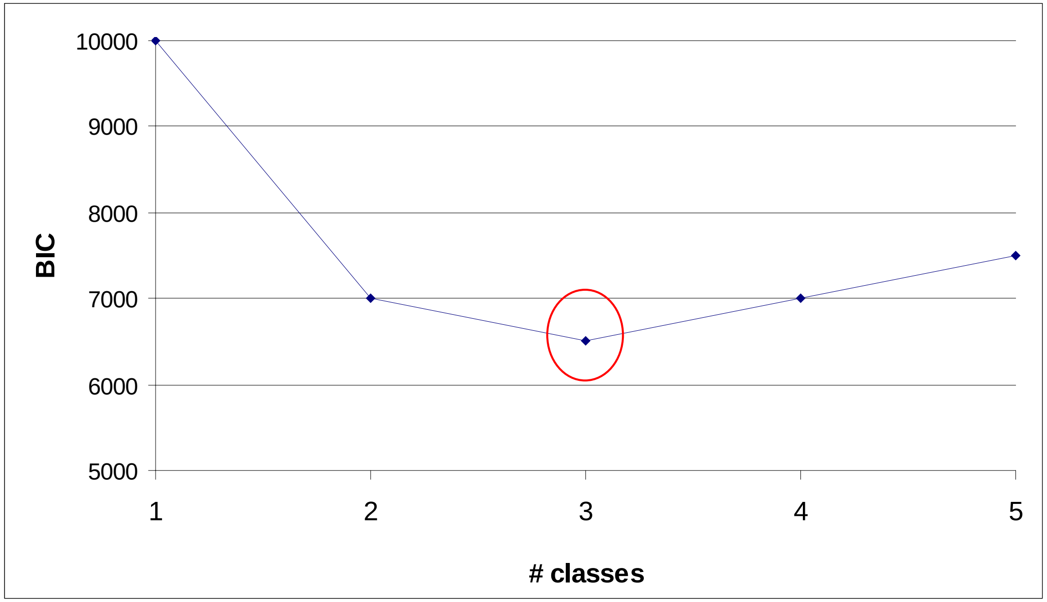
```
BIVARIATE MODEL FIT INFORMATION
                                         Estimated Probabilities
                                                            Standardized
    Variable       Variable                H1           H0          Residual
                                                                    (z-score)
    MSMK2          MSMK8
      Category 1     Category 1          0.789        0.779            1.966
      Category 1     Category 2          0.032        0.042           -4.066
      Category 2     Category 1          0.014        0.024           -5.310
      Category 2     Category 2          0.164        0.155            2.255
      Bivariate Pearson Chi-Square                                    48.502
      Bivariate Log-Likelihood Chi-Square                             54.579
    MSMK2          MSMK21
      Category 1     Category 1          0.783        0.783            0.148
      Category 1     Category 2          0.038        0.039           -0.316
      Category 2     Category 1          0.025        0.026           -0.386
      Category 2     Category 2          0.154        0.153            0.169
      Bivariate Pearson Chi-Square                                     0.270
      Bivariate Log-Likelihood Chi-Square                              0.272

    <snip>

      Overall Bivariate Pearson Chi-Square                           165.254
      Overall Bivariate Log-Likelihood Chi-Square                    172.326
```

Model fit stats - BIC

- Bayesian Information Criterion
  - = -2*Log-likelihood + (# params)*ln(sample size)

- Function of likelihood which rewards a more parsimonious model

- Decrease followed by an increase as extra classes are added



Entropy (for the 4-class LCGA model)

```
CLASSIFICATION QUALITY

     Entropy                         0.948

Average Latent Class Probabilities for Most Likely Latent Class
   Membership (Row) by Latent Class (Column)

           1        2        3        4

    1   0.878    0.053    0.006    0.063
    2   0.018    0.962    0.000    0.020
    3   0.006    0.000    0.990    0.005
    4   0.070    0.089    0.006    0.836
```

- Summarizes the class assignment probabilities
- Higher values are better, 0.6 is “fuzzy”

Remind me what class-assignment probabilities are

- Each respondent is assigned a set of probabilities describing the likely latent class into which they fall
- Respondents with same response pattern have same probabilities
- Same as GHQ example from earlier

- Probabilities below are for subjects with low levels of smoking

```
+-------------------------------------------------------------------------+
| msmk0  msmk8  msmk21  msmk33  msmk47  msmk61   num      p1    p2   p3     p4 |
|-------------------------------------------------------------------------|
|     0      0       0       0       0       0  5033    .001  .999    0      0 |
|     0      0       0       0       0       1   102     .11   .88    0   .009 |
|     0      1       0       0       0       0    35    .017  .608    0   .376 |
|     1      0       0       0       0       0    26    .009  .818    0   .173 |
|     0      0       0       0       1       0    22    .438  .537    0   .025 |
|-------------------------------------------------------------------------|
|     0      0       1       0       0       0    19    .106  .754    0   .141 |
|     0      0       0       1       0       0    18    .218  .733    0   .049 |
+-------------------------------------------------------------------------+
```

Red = less certainty, i.e. more fuzzyness, however these patterns are rare

LCGA model fit summary (n = 6,851)

Classes	# params	BIC	Entropy	Lowest Class-spec entropy	Tech10 (global)
1	3	40347.8	-	-	63542.7
2	7	19335.5	0.975	0.989	491.4
3	11	18425.7	0.934	0.873	298.6
4	15	18187.5	0.948	0.836	180.9
5	19	18169.9	0.933	0.812	119.4
6	23	18153.3	0.926	0.772	34.9
7	Problems				

So now what??

- Focussing solely on model fits stats is sometimes unrewarding
- Not clear-cut picture in this example
  - BIC does not hit a low point
  - Tech10 suggests 6 classes
  - However, entropy always good
- Experiment with more parsimonious models? (remove q terms)
- Importance of non-statistical criteria
  - results look right?
  - Do key covariates distinguish between classes

Summary so far

- Latent Class Growth Analysis is a good way of capturing non-normal variability in growth factors when modelling binary data

- I/S/Q take a different value within each latent class
- No variability is modelled

- Trajectories resemble polynomials, but in LOGIT space

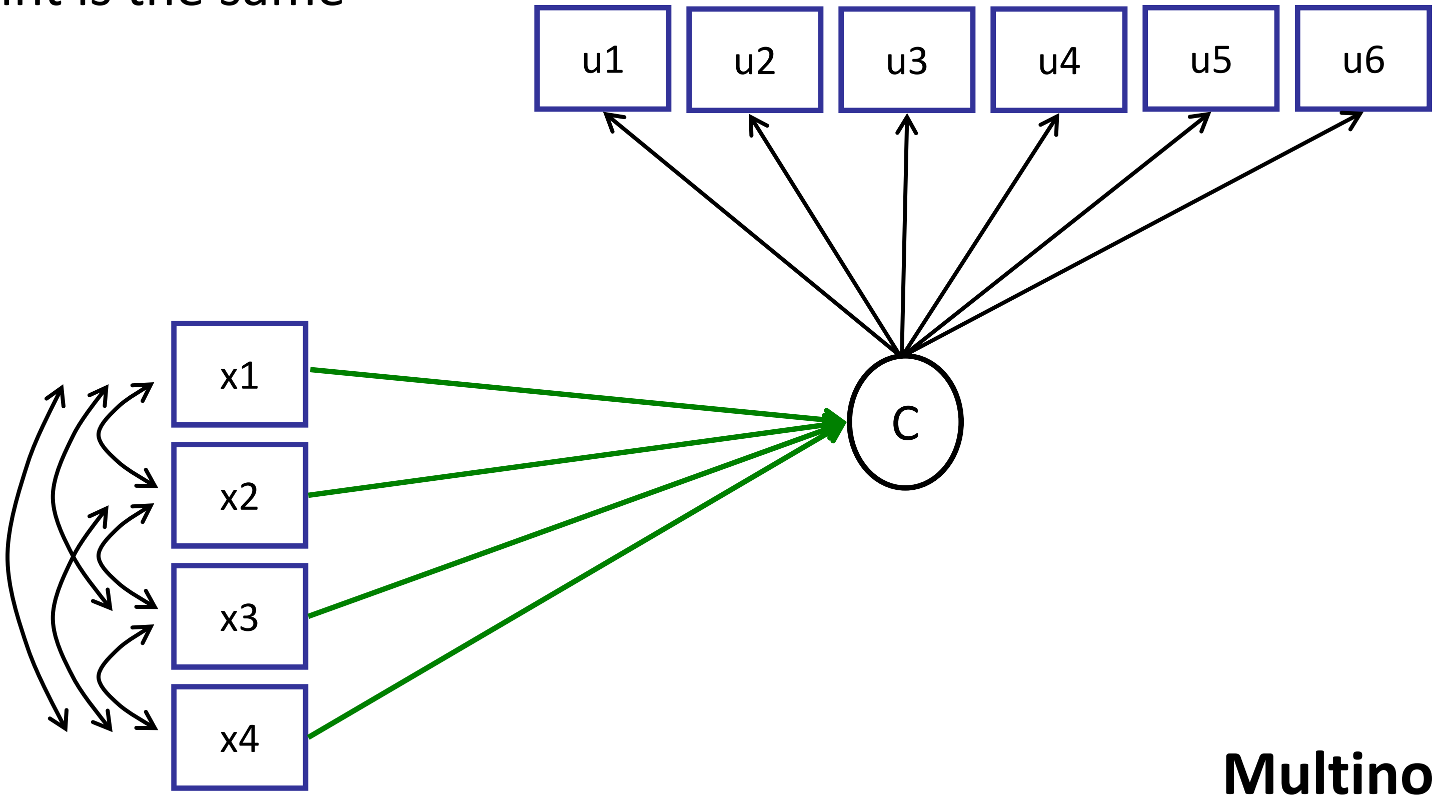
- Various rules for deciding how many trajectories is enough
  - Don’t always work perfectly, even in teaching examples ☹

Incorporating additional variables

Covariates and distal outcomes

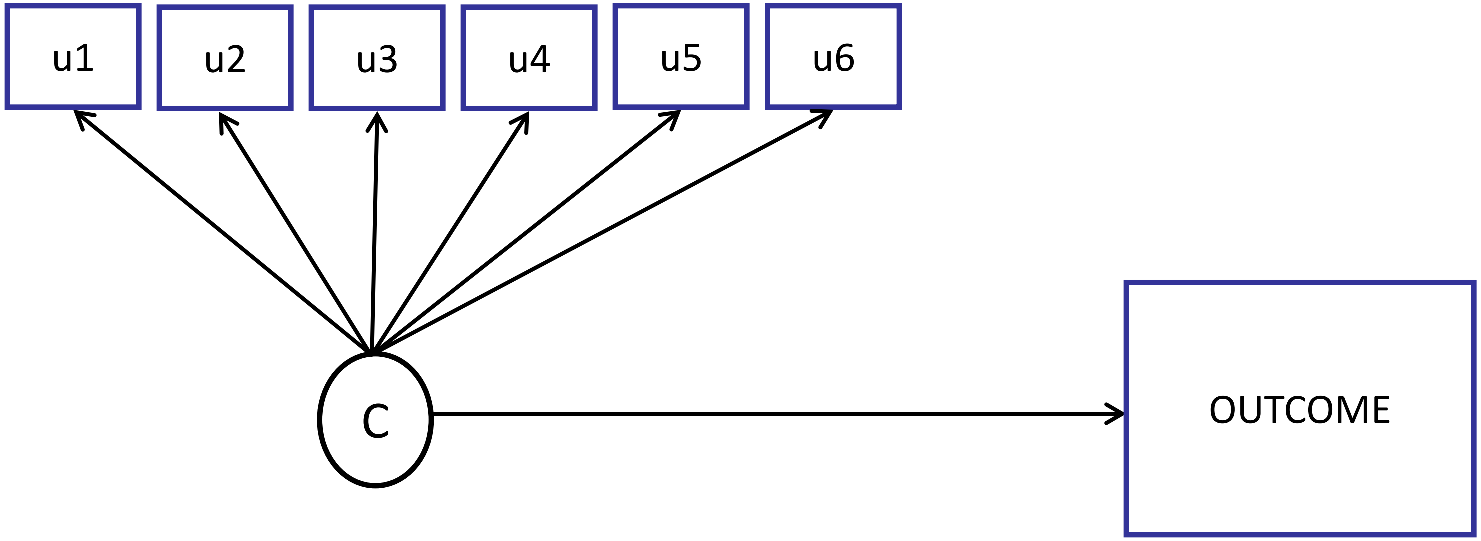
Risk factors for Latent Class membership

Figure is for LLCA but the point is the same



Multinomial logistic model

Latent Classes and a distal outcome



- **Standard linear regression if outcome is continuous**

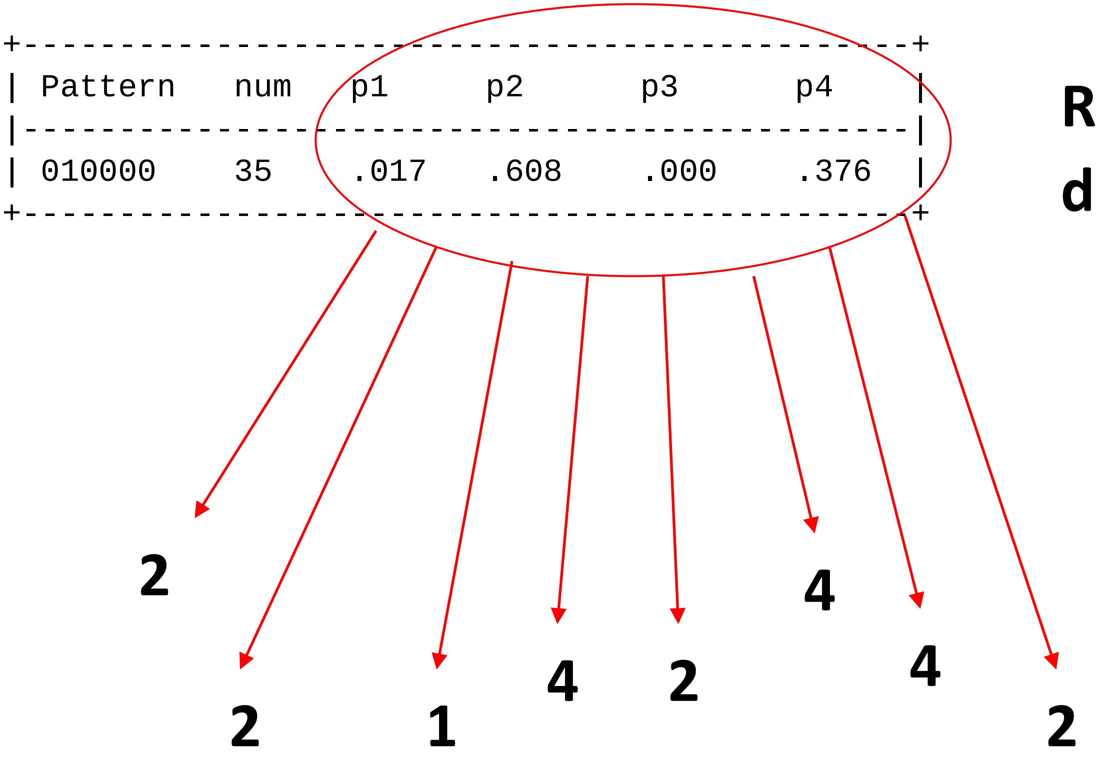
Pseudoclass draws in Mplus

```
+------------------------------------------------+
| Pattern   num   p1     p2      p3      p4     |
|------------------------------------------------|
| 010000    35    .017   .608    .000    .376  |
+------------------------------------------------+
```

A response pattern for smoking

along with it’s class assignment probabilities

Pseudoclass draws in Mplus



```
+---------------------------------------------------------------+
| Pattern    num    p1      p2        p3        p4        |
|---------------------------------------------------------------|
| 010000     35    .017    .608      .000      .376      |
+---------------------------------------------------------------+
```

Random draws from this distribution

Multiple datasets

Analyse like with missing data (Rubin’s rules)

Predictors of class membership

```
DATA:
    FILE = maternal_smoking.dat ;
    listwise is OFF;

VARIABLE:
    Names are
         verbal perf iq
         msmk2 msmk8 msmk21 msmk33 msmk47 msmk61
         tenure crowding smkpreg parity mumed
         tenure2 tenure3 smkpreg1 smkpreg2
         mumed1 mumed2 parity1 parity2 parity3;
    Missing are all (-9999) ;
    classes = c(4);
    usevariables = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;
    categorical = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;
    auxiliary = (r) mumed1 mumed2;

ANALYSIS:
  proc = 2(starts) ;
  type=mixture ;
  starts=3000 300 ;
  stiterations=25 ;
  stscale=10 ;
```

Class membership versus distal outcome

```
DATA:
    FILE = maternal_smoking.dat ;
    listwise is OFF;

VARIABLE:
    Names are
         verbal perf iq
         msmk2 msmk8 msmk21 msmk33 msmk47 msmk61
         tenure crowding smkpreg parity mumed
         tenure2 tenure3 smkpreg1 smkpreg2
         mumed1 mumed2 parity1 parity2 parity3;
    Missing are all (-9999) ;
    classes = c(4);
    usevariables = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;
    categorical = msmk2 msmk8 msmk21 msmk33 msmk47 msmk61;
    auxiliary = verbal (e) perf (e);

ANALYSIS:
  proc = 2(starts) ;
  type=mixture ;
  starts=3000 300 ;
  stiterations=25 ;
  stscale=10 ;
```

Quick exercise (depending on time)

- Return to your 4-class LCGA model

- Add some predictors and try to (i) locate, (ii) interpret the results
- Repeat with one or more IQ outcomes

- Notes
  - Try to use “OPTSEED = XXXX;” in analysis section for speed
  - You will need to use the dummy-vars as predictors are categorical
  - The parameter estimates for the predictors are log-odds as this is a multinomial model
  - Nowhere does it warn you that the auxiliary variables are incomplete!

Results for maternal education

```
                                                     Two-Tailed
                     Estimate       S.E.  Est./S.E.    P-Value

Parameterization using Reference Class 3
                                                                Odds ratios
 C#1 (offset smk) ON
    MUMED1             0.812      0.197      4.125      0.000      2.252
    MUMED2             0.484      0.172      2.807      0.005      1.623

 C#2 (onset smk) ON
    MUMED1             0.542      0.176      3.080      0.002      1.719
    MUMED2             0.264      0.153      1.722      0.085      1.302

 C#4 (persist smk) ON
    MUMED1             1.509      0.091     16.515      0.000      4.522
    MUMED2             0.759      0.089      8.563      0.000      2.136

 Intercepts
    C#1               -3.395      0.131    -25.885      0.000
    C#2               -3.051      0.108    -28.175      0.000
    C#4               -2.279      0.068    -33.272      0.000
```

Smoking classes versus IQ outcome

```
EQUALITY TESTS OF MEANS ACROSS CLASSES
USING POSTERIOR PROBABILITY-BASED MULTIPLE IMPUTATIONS
```

```
              (Verbal IQ)

                    Mean          S.E.

 Non-smoker       109.549         0.265
 Onset smk        107.243         1.357
 Offset smk       106.408         1.328
 Persistent       104.535         0.649


               Chi-Square       P-Value

 Overall test     39.606        0.000
```

```
          (Performance IQ)

                    Mean          S.E.

 Non-smoker       101.755         0.270
 Onset smk        101.454         1.505
 Offset smk       102.229         1.278
 Persistent        97.226         0.678

               Chi-Square       P-Value

 Overall test      27.217         0.000
```

Summary

- We can use the auxiliary command to relate other variables to our latent classes
- Classes strongly socially-patterned
- Potential differences across classes
- Smoking class related to later child IQ but likely to be heavily confounded

- The auxiliary approach is somewhat limited – more can be done by exporting the probabilities to another package e.g. Stata

Global Summary

- Longitudinal variability is interesting
- It can be useful to distinguish between subjects based on the different ways their data moves through time (what like Dr Who?)

- We can employ latent variable models to understand between and within person variability
- Depending on the type of data and the type of application, an approach incorporating continuous and/or discrete latent variables can prove fruitful.