Failing To Measure Any Effect Of Increased Lighting On Crime:

A reply to Profs Farrington and Welsh

By Paul R. Marchant

AG03, Faculty of Information and Technology,

Leeds Metropolitan University,

Calverley Street,

Leeds LS1 3HE

England, UK p.marchant@leedsmet.ac.uk

Abstract

This paper shows that the reply by Profs Farrington and Welsh to my work "A demonstration that the claim that brighter lighting reduces crime is unfounded" still contains flaws and misunderstandings. A key element which is not accounted for is the spatial correlation of crime events in any one period of observation. This effect explains the high observed variability of the crime counts; much higher than is conceded in their reply to me; an order of magnitude greater than Poisson. The effect of 'regression towards the mean' is also one of their problems not properly dealt with. (The Protocol for a Campbell Collaboration systematic review is similarly flawed.) Any study using their methods will be subject to such errors. My work here will explain, in a relatively non-technical fashion, what the problems are. The problems uncovered are general, going beyond any specific evaluation, and can be readily seen with more extensive data from 124 areas.

Introduction

This paper rejoins a debate with Farrington and Welsh (2004) on the merits of my critique (Marchant 2004) of their systematic review which assessed the effects of increased lighting on crime reduction. My thanks go to them for their detailed and painstaking discussion and for seeking to champion the rigour of the statistical foundations of the meta-analysis. In this reply I will show that their original and revised calculations of lighting's efficacy continue to contain flaws and misunderstandings, such that no sound conclusion of the effect of increased lighting can be drawn.

In Marchant (2004), I drew attention to problems in a systematic review of lighting and crime "Home Office Research Study 251" (HORS251), Farrington and Welsh (2002a). (Similar work was published as Welsh and Farrington (2002b)). The problems I uncovered showed that the conclusion that increased lighting reduces crime is unfounded. Farrington and Welsh (2004) have replied and have accepted that the variability is higher than that required by the method they originally used, but they still say that increased lighting reduces crime. However their revised method fails to take proper account of the variability and indeed I will show that it is much larger than they acknowledge. My conclusion from the evidence that I shall present is that the effectiveness of street lighting against crime is unknown; street lighting might reduce crime or it might increase it. Furthermore, I will show that because of this and other problems, the efforts of Farrington and Welsh (2003a and 2003b) in their work with the Campbell Collaboration are bound to fail. My conclusions are likely to apply to other work using the approach of Farrington and Welsh which also attempts to evaluate the effect of similar area-based

crime reduction interventions, because of the fact that correlations exist between the commission of crime events in any one period, compounded by unequal starting conditions in the test areas.

Background

Information about problems was originally sent to the authors of the review and the UK Government's Home Office, its sponsor, in February 2003. Since they were alerted, an addendum has been added to HORS251 (Farrington and Welsh 2003) on the Home Office website, in September 2003, worded in such a way that it may give the false impression that I agree with the statement. Also a revised protocol, for a systematic review on street lighting and crime to be done under the auspices of the Campbell Collaboration,

Welsh and Farrington (2003), has appeared on the Campbell Collaboration's Crime and Justice website, dated November 2003.

Terminology

There are instances in which I employ different terminology to Farrington and Welsh.

<u>Increased not Improved</u>

All the 13 studies included in HORS251 were examining the impact of increased lighting on crime. In fact the increase was between 2 and 7 times in the studies, where the increase was known. Therefore increased is the most appropriate term. Farrington and Welsh use the term 'improved lighting'. Lighting might be improved in many ways, e.g.

lighting which produces less glare. ‘Improved’ is unspecific and unscientific. Improved is more a term associated with advertising.

Comparison not Control

The area not receiving the treatment between the periods should properly be called 'comparison' rather than 'control'. This is so that the distinction from a proper control is clearly maintained. In a randomised controlled trial (RCT), the control group is statistically equivalent to the treatment group, in regard to relevant factors both known and unknown, via the randomisation. However in the situation for the crime studies described, the comparison is just another area with some similarities, yet the term 'control' is used by Farrington and Welsh, which wrongly suggests equivalence. Invariably the comparison areas have less crime, so cannot be considered to be equivalent.

The original Farrington and Welsh meta-analysis.

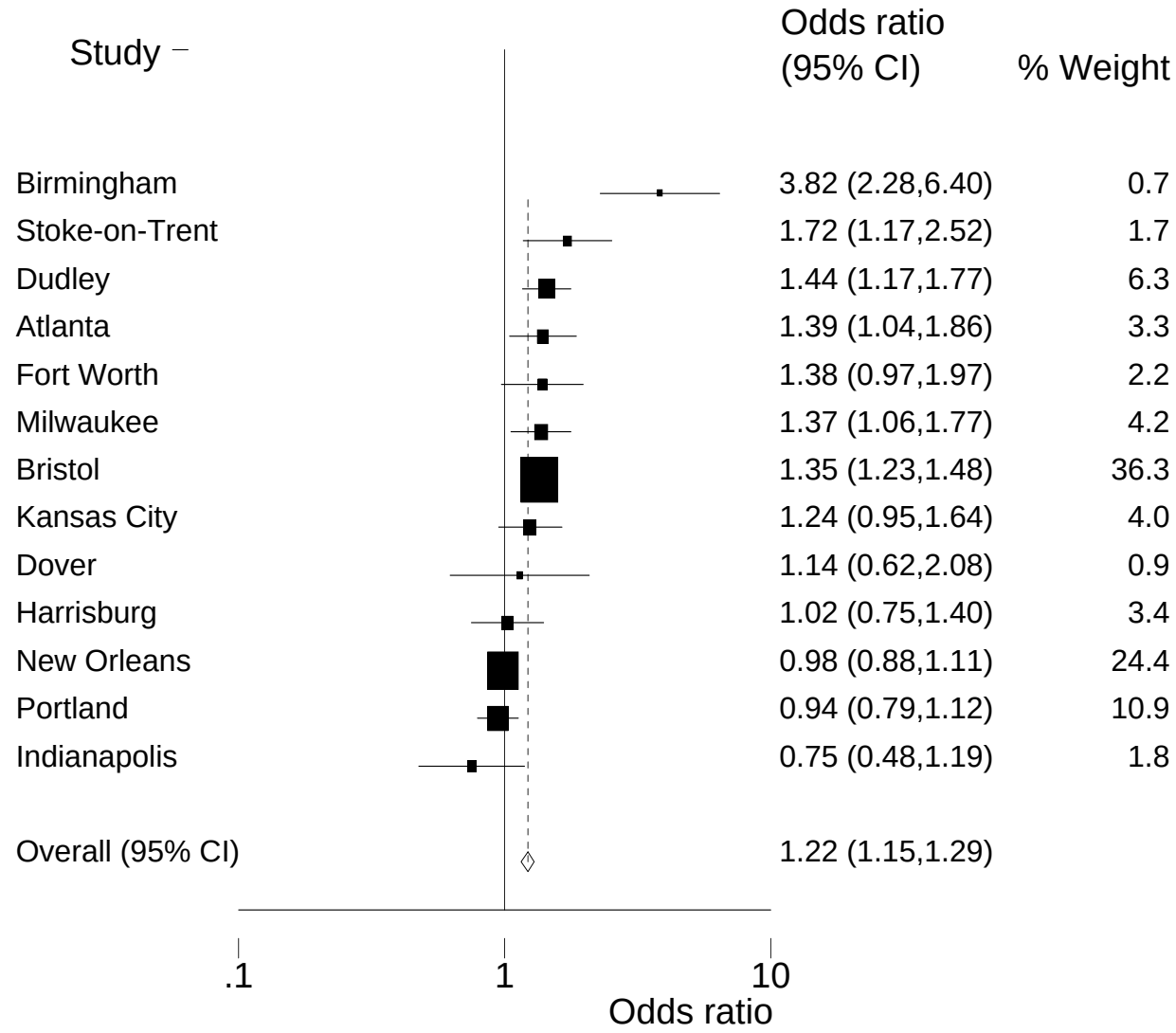
HORS251 compared the ratio of number of crimes before and after in an area that had increased lighting introduced with that of a similar ratio in a ‘control’ area which had no change in lighting. This was done for a number of studies.

	Before	After
Intervention Area	*a*	*b*
Comparison Area	*c*	*d*

The ratio of ratios (a/b)/(c/d) is called by the authors, an ‘odds ratio’ (‘OR’). If the ‘OR’ were observed to be convincingly greater than one then it might be concluded that crime has been reduced in the newly lit area compared with the comparison.

The results of the Farrington and Welsh review are shown in essence in Figure 3.1 of HORS 251. It has been possible to reproduce this ‘forest’ plot using the data for the number of crimes in each of the four settings (treatment/comparison, before/after) for each study, given in the text of HORS251 and the individual data has been combined to generate the combined overall ‘OR’. See below, Figure 1, for the forest plot derived from the data given in HORS251. The overall weighted average ‘OR’ is shown. This kind of statistical synthesis is known as meta-analysis. See Egger et al (2001) for information on systematic reviews and meta-analysis.

Figure 1



Fixed Effect Meta-Analysis as per HORS251

(using STATA metan command)

As the plot shows the 'OR' = 1.22 within a 95% Confidence Interval (1.15, 1.29).

The z –statistic = 7.07 showing a very highly statistical significant effect; p < one in a billion. The heterogeneity statistic, Q = 58.02 with 12 degrees of freedom.

There are small differences between these figures and those given in HORS251; OR=1.25 Confidence Interval (1.18, 1.32) z = 7.78 and Q = 56.91.
The small differences between the two sets of numbers may be due to the handling of loss to follow-up in the 2 household survey studies, Dudley and Stoke-on-Trent, as well as possible rounding errors in calculating absolute numbers from the percentages given in HORS251.

The confidence intervals for the 'ORs' of each of the 13 included studies have been calculated using the simple expression for the variance which is just the sum of the reciprocals of the crime counts in each of the 4 settings. See for example p248 Bland (1995). This expression assumes that the cell counts arise from independent simple random sampling.

The weight of each study's contribution towards the overall combined 'OR' is given by the reciprocal of the variance of the 'OR' for each study and is represented by the area of the square showing the point estimate of the 'OR', in my forest plot, as well as explicitly by the percentage weight.

The expression originally used for variance by Farrington and Welsh in HORS251 may be appropriate in a classic clinical Randomised Controlled Trial (RCT) but is not for crime events, which inevitably involve many correlations, e.g. crimes committed by the same criminal or committed on the same victim (as pointed out in my earlier paper). It is clear that the variability is much greater than the method used in HORS251 requires.

Their method requires that the variance, i.e. standard deviation squared, of the variation should be equal to the mean count. However it can be clearly seen from the Bristol study data, Shaftoe (1994), and from the Birmingham study data, Poyner and Webb (1997), for example, that there is large extra variation much beyond that required for the method used by Farrington and Welsh. These data are shown in Marchant (2004). In fact the variance appears to be an order of magnitude larger than their respective means. The factor by which the variance exceeds the mean is known as the overdispersion factor. This same large overdispersion can be seen in the studies where there is just one measurement point in the before period and one after; see later. The large overdispersion occurs in both comparison and relit areas. Remember that the comparison areas exist to sense the natural variability.

The revised calculation of variance given in the reply

Farrington and Welsh (2004) in their reply show the expression they have used for the revised estimate for the variance of a study's 'OR'. It deals with the case of an experiment of the household survey type. It incorporates the effects of over-dispersion and (just) the correlation of an individual's experience of crime. This expression is credited to Dr. Patricia Altham. However, I maintain that it is not complete for the situation being considered: see below.

Their expression given is:

Corrected variance of the natural logarithm of the 'OR'

Corrected VAR (LOR)= ϕ (VAR -2r[1/$\sqrt{}$ab + 1/$\sqrt{}$cd])

Where ϕ=over-dispersion factor,

VAR = uncorrected variance of (LOR) i.e. that based on independent Poisson sampling=1/a+1/b+1/c+1/d

(where

a= number of crimes committed in the relit area before relighting

b= number of crimes committed in the relit area after relighting

c= number of crimes committed in the 'control' area before relighting in relit area

d= number of crimes committed in the 'control' area after relighting in relit area)

r=correlation coefficient from before to after for households.

The authors use values for the over-dispersion factor and correlation coefficient they estimate for the Dudley household survey study (Painter and Farrington 1997). It is assumed that the values of 'ϕ' and 'r' are the same in both the relit and the comparison areas. It is further assumed that the correction based on them applies across the board to all 13 studies.

I have been able to check the value of ϕ, using the subset of the Dudley data that Prof Farrington kindly sent me. (The data did not include all the associated variables for which data was collected, hence the term subset).

ϕ= 3.6, the value they use, applies to the intervention area before relighting. It is the ratio of the variance of the truncated frequency distribution for that area divided by its mean. The values for the other three time and area combinations were somewhat smaller (Minimum= 2.5).

I have been unable to check the value of 'r' used, because the data I received did not include any of the matching variables that were used to reconstruct the pairings, before and after. Reconstruction was necessary because the investigators failed to link the household addresses before and after at the time of the study, both for Dudley and for Stoke. The value obtained must therefore be additionally uncertain.

However assuming that the value of r was as stated then indeed their chosen values for 'ϕ' and 'r', when inserted in the expression above, seem to justify the assertion made that the revised combined effect for lighting comes out to be statistically significant. However their way of proceeding is incorrect.

Further revision of the calculation of variance

I have derived a similar but extended expression, Marchant and Baxter (to be submitted for publication), which explicitly includes the effect of spatial correlation within one observation period, as well as the correlation of individuals' experience of crime across the time periods. Spatial correlation has to be included as it accounts for the fact that if one household experiences a crime, then the probability that other households in the same area experience crime is also affected. That is crime events in one period of time are not independent, a fact not taken into account in the revised variance expression given by Farrington and Welsh in their reply. This lack of independence relates to the point made in my original paper p445 that "One criminal may be responsible for many crimes and so this one person changing behaviour can cause a large change in the number of crimes committed and recorded".

The modification required to the expression that is used by Farrington and Welsh for the variance of the log 'OR', given above, involves multiplying the right hand side by a factor (Deff) to account for spatial correlation of events. In fact Deff = $1+(n-1)\rho_s$, where n is the number of individuals sampled in one area and ρ_s is the spatial correlation. This is familiar in other contexts and would be called the Design Effect. It takes account of clustering in surveys (Kish 1965) and in 'cluster randomised trials' (Bland 2003). See also Ukoumunne et al. (1999) on the evaluation of health interventions at area and organisation level, which also discusses the design effect and correlation between different individuals within areas.

Including spatial correlation incorporates the obvious effect that any household's changing experience of crime is correlated with that of other households in the area. Thus the over-dispersion is not simply given by ϕ, the ratio of the variance of the distribution of the number of crimes committed per household divided by the mean number of crimes per household, as used by Farrington and Welsh in their revised variance estimates. The revised variance must include the factor Deff, the design effect. When there are correlations and events tend to move in "synchrony", it is as though the sample size is reduced. Thus the design effect makes the effective sample size smaller. (This is exactly the same effect as increasing the variance).

It is possible to extended further the analogy of a treatment for the common cold (Marchant 2004) to exemplify the problem of regression towards the mean, in that sufferers tend to end up in a more average cold-free state. In any properly conducted and analysed clinical trial, it would be wrong to count the number of sneezes for a patient-group rather than the number of sick individuals as this would give the false sense of

having more independent data than actually is the case. Six people out of 10 who are suffering and between them sneeze 66 times do not constitute 66 independent counts. It would be even more wrong to count the viruses in those sneezes. (The same would apply to a disease which caused spots, e.g. measles…don't count the spots, count the sick patients.) This goes to the heart of the misunderstandings of Farrington and Welsh; the effective sample size is much smaller than they think.

The problems with using quantities from the Dudley study

The Dudley study has a desirable feature in that it is a 'before and after household crime survey'. However this makes it different from the vast majority of the studies included in HORS251 which are 'area-level crime count studies'.

The weight for the contribution of a study to the meta-analysis is given by the inverse of its variance. The HORS251 weights are shown in the forest plot Figure 1.

The variances cannot be reliably revised wholesale, using one simple correction factor, as Welsh and Farrington wish. We would need to know what the correction factor is for each study, in order to determine its weight in the meta-analysis.

Another serious problem is to do with the unit of observation. The great majority of the included studies are <u>not</u> of the Dudley type, which is a before-after household crime survey with a defined sample of individual households. In the area crime-count studies, which form the majority of the included studies, we do not know if any of the victims are the same before and after. Some probably are, but what is the unit for correlation here? Perhaps some victims were just passing through. The concept of individual correlation

evaporates. Indeed Farrington and Welsh seem to have problems with recognising the correct unit of analysis; area-level versus individual level. In the crime count studies the unit of analysis is 'area'. See for example 'Unit of….' in the Cochrane Reviewers' Handbook Glossary www.cochrane.org/resources/handbook/glossary.pdf

Overdispersion in crime count studies

In most of the 13 studies in HORS251 there are not many observations over time, in fact there are usually only 2 values; one in the before period and one in the after period. It is possible however to get a variance estimate for crime count for each study using its two data points. The standard expression for variance is used, familiar from basic statistics courses, i.e. the sum of squares of the values minus the sample mean, divided by the number of observations minus one. This sample variance can be divided by the sample mean to give the ratio of variance to mean, i.e. the observed overdispersion, for each study. This seems to have been done by Farrington and Welsh, for the 'control' areas as they state, "the average ratio of variance to the mean was 2.07 in these 13 evaluations, showing that on average the variance was about twice the mean". But let us look more closely. My values of overdispersion calculated from HORS251 for each study's comparison area are given in the table below.

Study Name	Overdispersion in Comparison Areas
Atlanta	58.3594
Birmingham	1.5306
Bristol	44.7116
Dover	4.7647
Dudley	4.4420
Fort Worth	0.2934

Harrisburg	1.5748
Indianapolis	0.0400
Kansas City	9.7449
Milwaukee	13.7547
New Orleans	46.6810
Portland	7.7017
Stoke-on-Trent	0.0083

We see that the values are extremely variable with some very high values up to around 60.The arithmetic mean is 15 for these 'control'/comparison areas, which exist to sense the natural variability of crime. That is, these areas sense the changes that might be expected when no intervention is introduced. The geometric mean is around 2, (I get 2.79), the value Farrington and Welsh claim, and presumably this is the meaning of their sentence. "Geometric means were used to summarize ratio variables". But why use geometric means when there is no justification to do so?

It is clear that using the geometric mean will yield a much smaller value than the arithmetic mean with the highly right-skewed data overdispersion data shown above. This is because the geometric mean is the n^{th}-root of the product of the n-values. Thus here the small values, such as 0.04 for Indianapolis, tend to 'cancel out' a large value such as Bristol's 44.7. It is shown below that it is the arithmetic mean of these data which gives the correct estimate of the underlying overdispersion.

What observed overdispersions are to be expected?

Let us consider an area with a constant underlying mean number of crimes per year and a constant underlying variance being D times the value of the mean, that is the overdispersion is D. Let us briefly examine what theory gives.

If there were no overdispersion (crime counts Poisson) the variance = mean, i.e. D=1, as originally used in HORS251, but as I pointed out in Marchant 2004, this is not so. If the counts are 'large' (count>25) the Poisson is virtually indistinguishable from the Normal distribution. Certainly the numbers of crimes in the 13 studies considered constitute 'large' in this sense. However, we know that there is overdispersion. In this circumstance we can multiply the mean by the appropriate overdispersion factor D (D>1) to get the variance of the underlying distribution. Farrington and Welsh say D≈ 4, I say it is something like 4 or 5 times bigger still, i.e. 16 or so.

What would we expect to find if we take repeated samples of size=2, i.e. one observation before and one observation after, from an unchanging Normal distribution. That is a distribution where the underlying mean remains fixed, as does the variance and so their ratio, the overdispersion, is likewise fixed at the value D. We can calculate the mean and variance from each of these sampled pairs, so we get a value of the observed overdispersion D_{obs} for each pair. The values of D_{obs} will not be all the same because they come from samples: i.e. there is sampling variation. It is this sampling distribution of D_{obs} that we are interested in.

It can be shown (Marchant and Baxter) that D_{obs} will be distributed approximately as a Chi-squared distribution, scaled by the underlying fixed overdispersion D. The degrees of freedom (df) of the Chi-squared distribution are equal to1 in this case, because there are just two points, one before and one after.

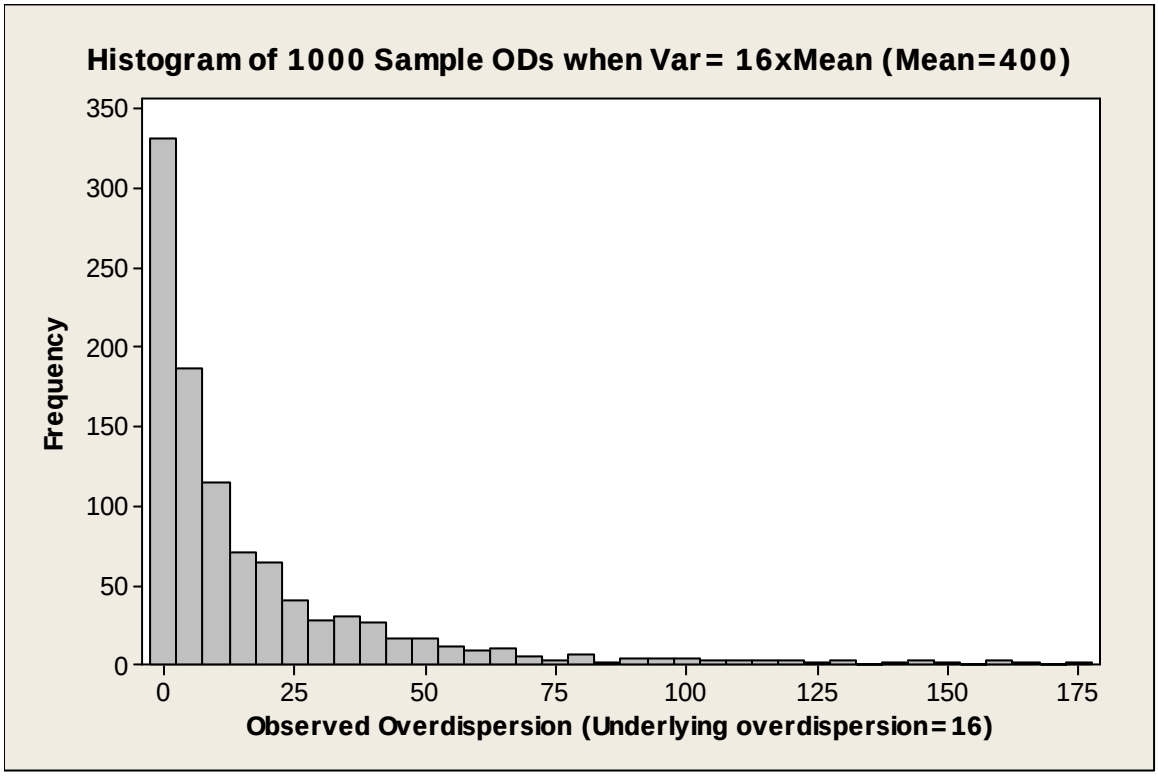
(Note the sampling distribution of the variance, estimated from a Normal distribution is a scaled Chi-squared distribution with df = n-1, where n = sample size. This is the basis of the well known and much used statistical method called Analysis of Variance (ANOVA), so I am not invoking anything particularly new.)

A Chi-squared distribution with one degree of freedom is a very long-tailed distribution so that we must expect to obtain highly dispersed values of D_{obs} when we sample, even though the underlying situation is unchanging. Also we can be sure that, although highly dispersed, the expected value, i.e. arithmetic mean, of D_{obs} gives the underlying overdispersion D. (This is because the mean of a Chi-squared distribution is its degrees of freedom. These are here equal to one, and this must be multiplied by the overdispersion factor, D, the scaling constant.) In statistical parlance D_{obs} is an unbiased estimator of D as on (arithmetic!) average it yields D.

Another entirely equivalent way of writing a Chi-squared distribution which is scaled by a fixed value D is as a Gamma distribution with a scale parameter =2D and a shape parameter = df/2; see Evans, Hastings and Peacock (2000). This formulation was used so that 'theoretical cumulative distribution functions (CDFs)' and 'probability plots' could be constructed, see below, using the statistical package Minitab which includes the Gamma distribution as one of its options for these procedures, but not Chi-squared. Thus we expect the sampling distribution to be a Gamma distribution which has the shape parameter= ½ and the scale parameter = 2D.

To check the above reasoning, a simulation was performed. One thousand pairs of observations were sampled from a Normal distribution with mean = 400 and variance = 16x400 i.e. D = 16. The values of 400 and 16 are typical for the crime studies involved. (These numbers have simple square roots if these are required.) The sample mean and variance and hence overdispersion for each pair was calculated. The histogram of the 1000 sample overdispersions shown below demonstrates the high variability and positive skew, anticipated in the exposition above.

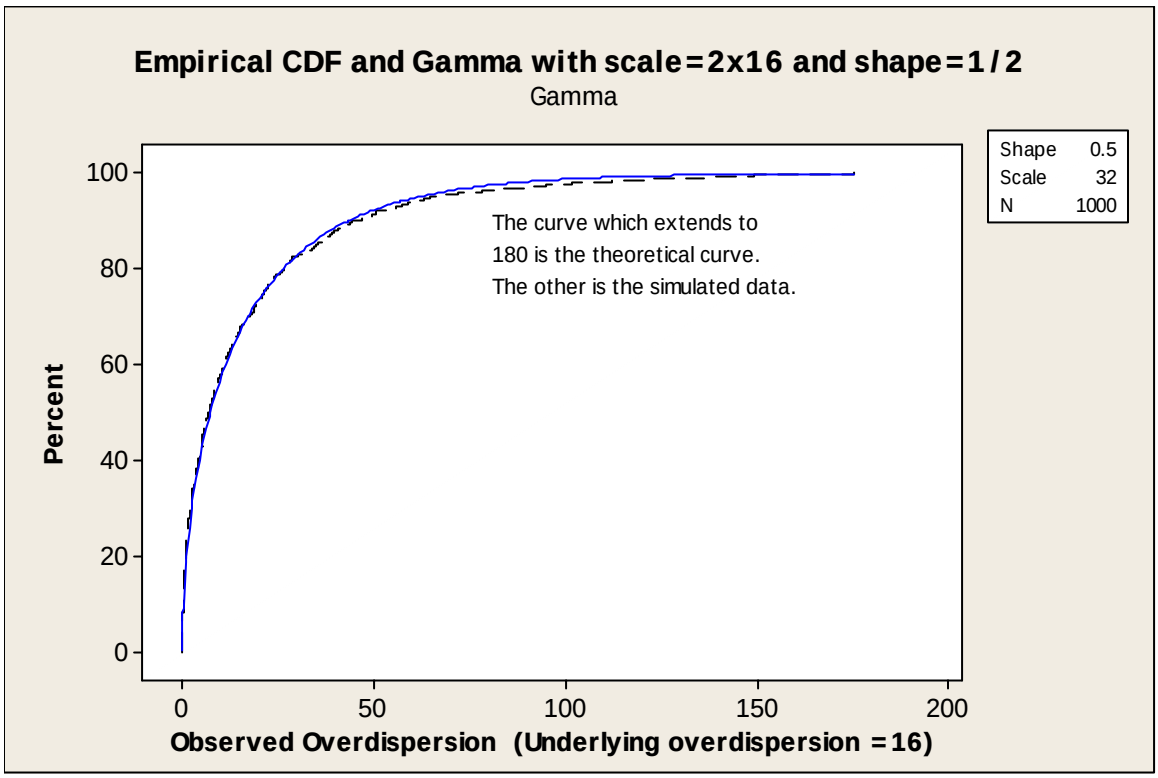
Figure 2



The mean of the 1000 sample overdispersions is 16.877, close to the population value of 16. Note the extreme variability that is shown by the minimum value 0.0000116 and maximum value 175.516.

A way of showing the goodness of fit is to compare the empirical and theoretical cumulative distributions, i.e. the percentage of data values below a given value compared with that expected. This is shown below and confirms that the results are as expected.

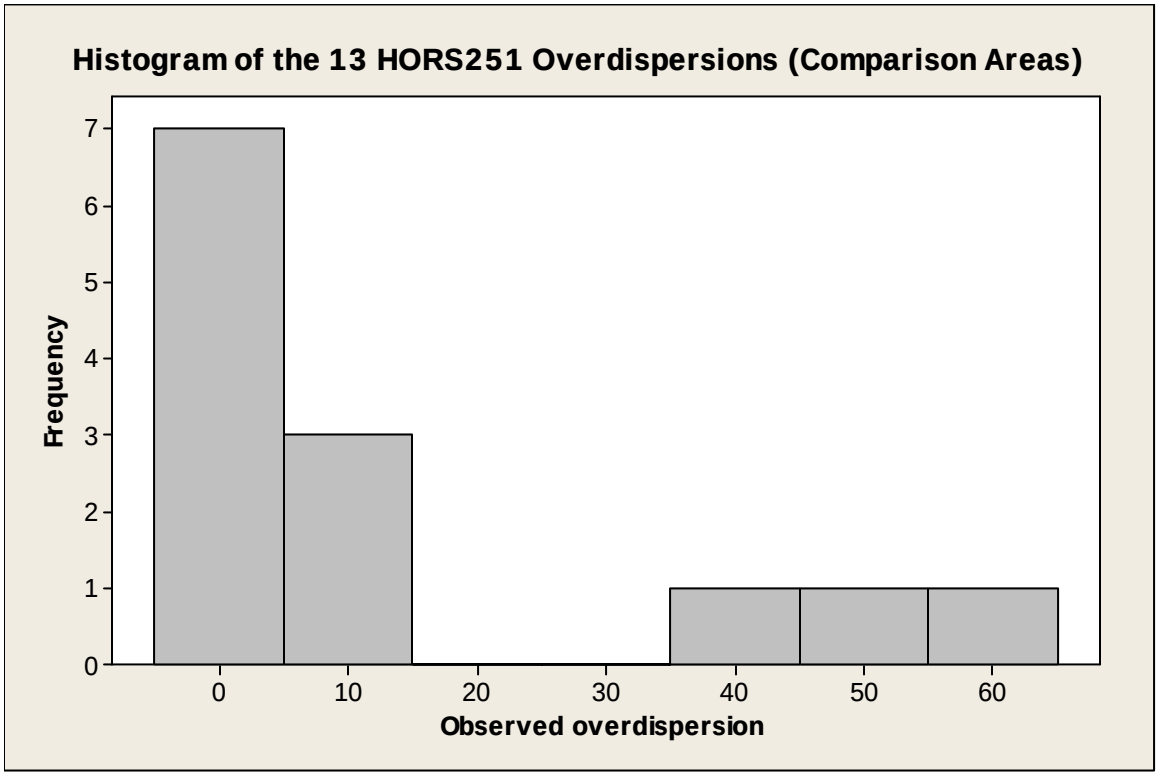
Figure 3



The HORS251 data

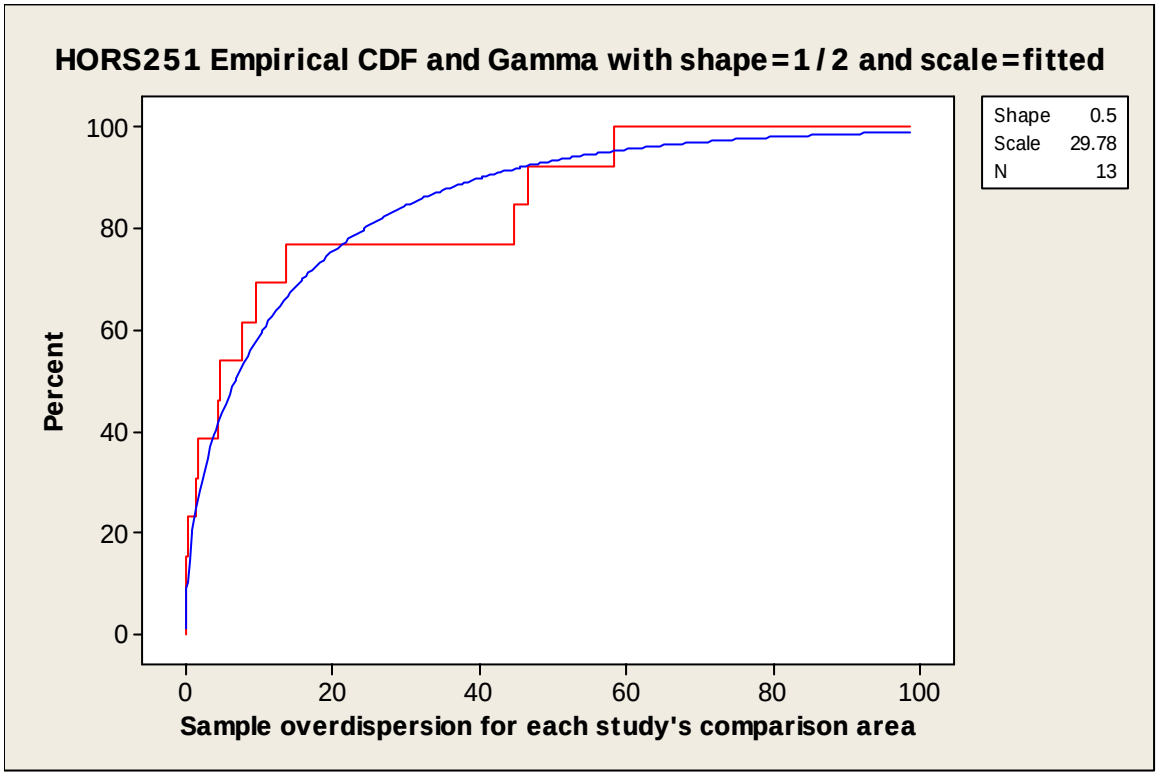
We now do the same with 13 pairs of data for the comparison areas from HORS 251.The histogram shows the large variability noted at the beginning, i.e. that there are overdispersion values up to about 60.

Figure 4



In the present case we do not know the value of the underlying overdispersion, but assuming it is at least approximately the same for all we can plot the cumulative distribution function as before with the Gamma distribution shape = 0.5 and scale = twice the overdispersion, an unknown to be fitted.

Figure 5



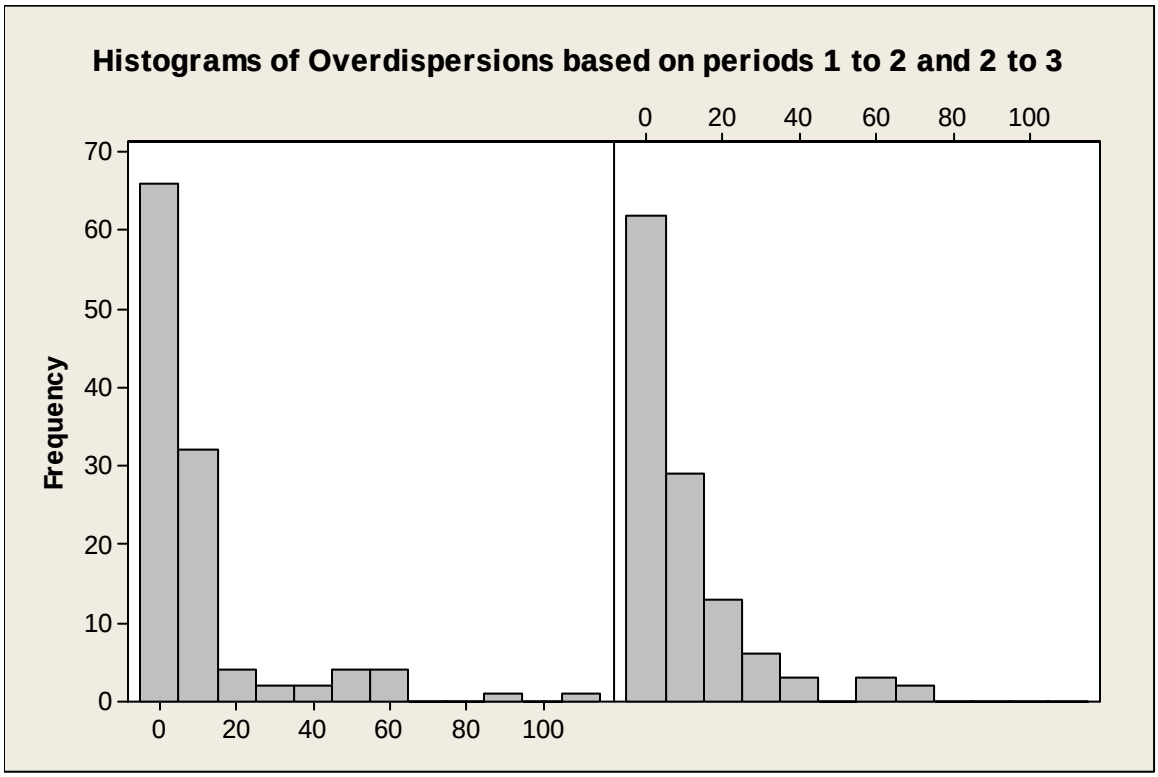
Again this seems a good fit. The fitted scale parameter of 29.78 suggests the overdispersion is 15, just like the arithmetic mean found above. In fact an Anderson-Darling test, provided in Minitab, shows that the fit is formally good.

Overdispersion in other crime data

To assess the generalisability of the above results I asked for area crime count data. Prof Nick Tilley kindly answered my call and provided me with a data set of burglary count data over three successive years from 124 anonymised areas. The data was from a project described in Tilley et al. 1999. I used the two available intervals from the 3 years of data in the data set to calculate overdispersions. A few areas did not have count information for successive periods. One area was anomalous. It had 9 times as many households as the next highest and also a massive D_{obs} of >200, so this case was removed from the data.

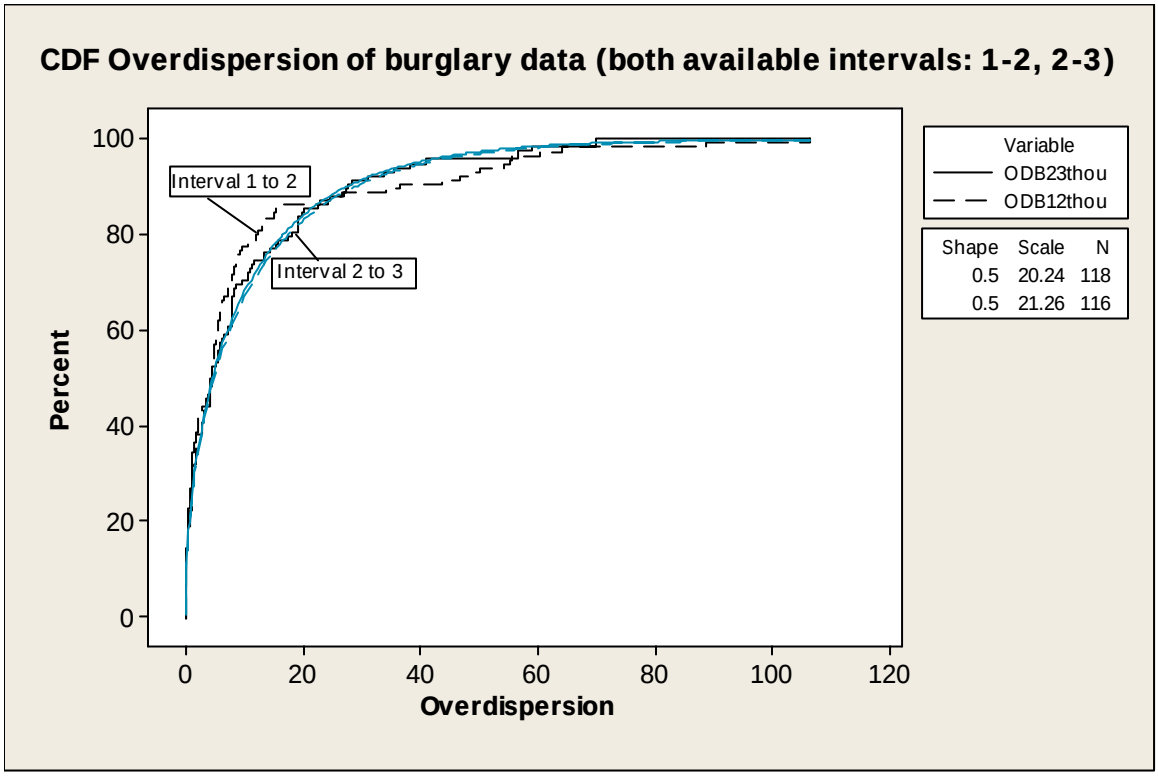
A further small number of areas were unusual in having an overdispersion of zero because the number of burglaries was identical before and after so these were reset to a value of 0.001. The number of useful cases available is 116 and 118 for the first and second intervals respectively. The mean overdispersion is a little more than 10 and the maximum was 10 times this figure. The histograms show the positive skew expected from the long-tailed Gamma sampling distribution.

Figure 6



The CDF plot shows that a Gamma-distribution, with shape parameter =0.5 fits the skewed data reasonably well. The fitted scale parameters = 21 and 20 for the first and second intervals respectively, suggests the overdispersion is 10 or so.

Figure 7



In fact a formal test of fit says that the fit in the first interval is not as good as one would expect if the true underlying distribution was perfectly Gamma(shape =0.5, scale =fitted). In the second interval the fit is very good. It is perhaps hardly surprising that there can be a certain lack of fit, as each area will have its own characteristic overdispersion due to its own specific spatial correlation of crime. That is, what we have is likely to be samples from a mixture of Gamma distributions with different scale parameters, but on average the scale parameter is about 20, i.e. the average overdispersion is about 10.

To summarise overdispersion

The theory shows that it is the arithmetic mean that should characterise overdispersion and not the geometric mean, which is used by Farrington and Welsh. The simulation shows that the theory works.

The work confirms that overdispersion is large, as was pointed out in Marchant (2004). It is imprecisely known. The strong evidence for large overdispersion undermines the possibility expressed by Farrington and Welsh in their 'Other estimates of variance' section that total crimes can be Poisson even if the individual experience is overdispersed. The fact, described on p451 of their reply, that for over 30 years some criminological work has been "dominated by the assumption that commission of crimes can be accurately modelled by a Poisson process" is neither here nor there. I have shown that there is large overdispersion in crime counts; an order of magnitude greater than that for Poisson.

For the lighting studies of HORS251, the best estimate of overdispersion using the comparison areas is 15. (Furthermore, statistical theory shows that because of the limited sample size of 13 this value is indistinguishable from 23, the value obtained from the relit areas, as the standard error appropriate for each is about 7). The small value used of overdispersion by Farrington and Welsh, either of 4 from the Dudley study or of 2 claimed from HORS251-data geometric mean, is clearly a gross underestimate. Furthermore, in any one study it is not possible to say at all precisely what the overdispersion is. This clearly has an impact on the variance of the effect size, which in

turn impacts on the weight that any study contributes in the meta-analysis, along the lines of Figure1.

Thus attempting to measure the effectiveness of a crime reduction method in an area, by this method of counting crimes before and after, is not at all reliable.

Matters become even more dubious when the intervention areas initially have higher crime levels than the comparison areas, as the effect of 'regression towards the mean' will bias the outcome towards showing a reduction in the intervention area. This is examined next.

Regression towards the mean

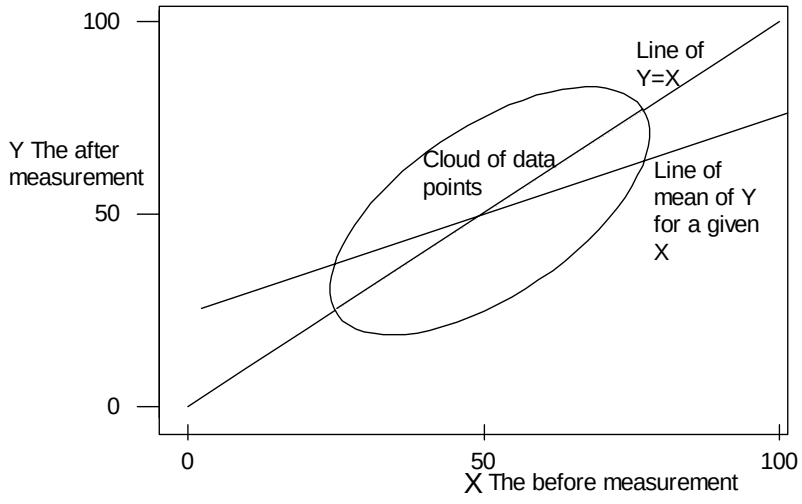


Figure 8: Regression towards the mean

Farrington and Welsh do not counter the argument that regression towards the mean is a problem for their work. It is a problem because comparison is made between 2 groups, which are in different states at the beginning. That is the relighting goes to the one with the higher crime.

Farrington and Welsh are mistaken in thinking that the effect of regression towards the mean is largely to do with test scores. In fact the effect was recognised and named at the end of the 19th century by Francis Galton when he was investigating the relationship of the heights of parents and their children; see Stigler (1986). Such height data are clearly correlated; tall parents produce tall children. Regression towards the mean is a natural and inherent consequence of correlation.

The diagram shows an oval which represents a standard cloud of data points. The X and Y values in the present context are measures of crime on some scale, for a sample of areas at one time X and at a later time Y. The mean and standard deviation of X and Y stay the same. (For Galton these would be parent's height and child's height). The tilt of the oval shows that high crime earlier is associated with high crime later and vice versa; i.e. positive correlation exists. Also on the diagram is the line of equality going diagonally bottom left to top right along the major axis of the oval. (Equality means that its slope = 1). Any point below this line says that the crime fell whereas a point above says the crime increased.

Also on the diagram is a line of shallower slope that gives the mean of Y for a given X. One can easily see that the mean of Y given X is not the line of equality, as taking a vertical slice through the oval will show that the bulk of the distribution lies above the line of equality for an X-value below the mean of X and below the line for an X-value above the mean.

Therefore for a high value before, the expected value, i.e. the mean, after will be below the line of equality, which has slope = 1. For a low value before, the expected value after will be above the line of equality. The regression line through the points is line of mean Y given X. The slope of this line, given by regression formulae, is simply 'r', the Pearson correlation coefficient (as the standard deviations are here the same before and after, for simpler exposition). Therefore there is a tendency to move towards the mean, i.e. to become more average. This is the effect that Galton found, that heights of children are more average than their parents, i.e. they regress towards the mean. (The statistical method of 'regression' owes its name to the discovery by Galton of the effect of regression towards the mean.)

If one applies an intervention to a group with high values of crime (i.e. a bad state before) and a 'control' to a group with low values (i.e. a better state before), one is likely to find that the new treatment is better even if it has no superior effect. This is because the expectation is that the high measurements will become lower and the low measurements become higher. This effect needs to be compensated for even in properly randomised

controlled clinical trials, as there will be a small imbalance between the groups at baseline, just due to chance sampling variation; see Mathews (2000).

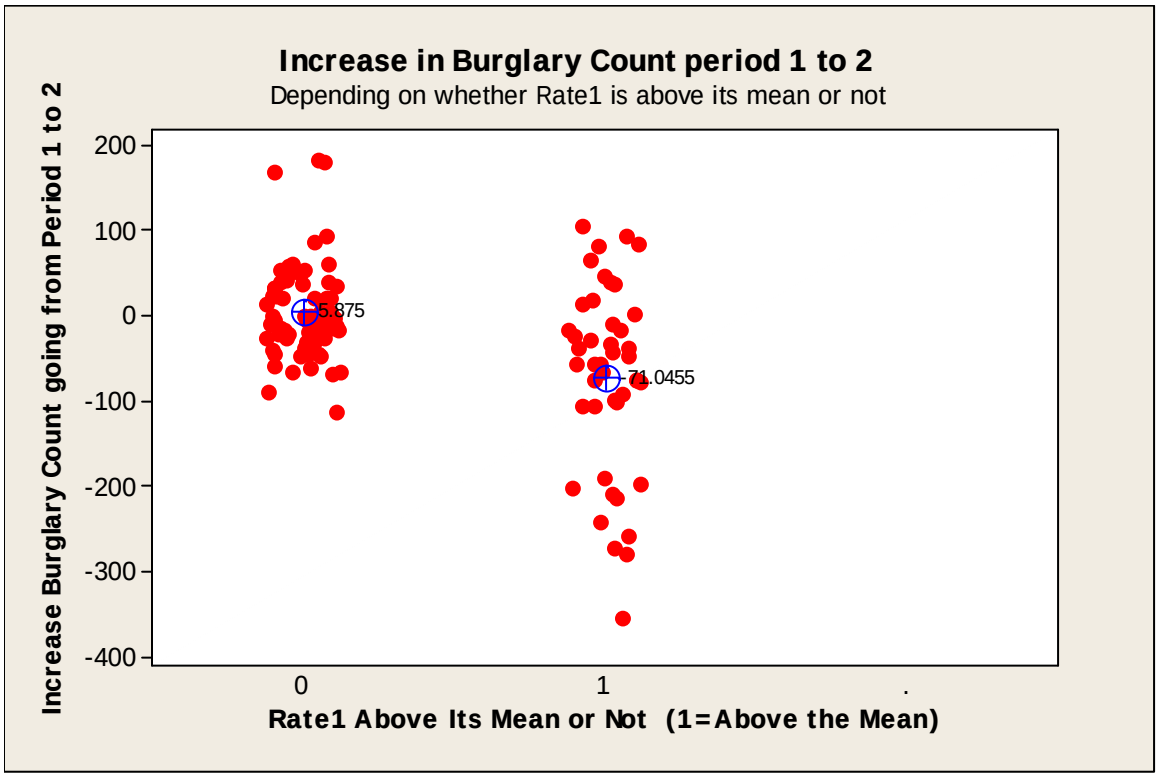
The situation is extreme in any situation when the correlation is zero, as then the expectation for the after measurement is just the average irrespective of the original measurement. The only circumstance in which the regression towards the mean does not occur would be if there were to be perfect correlation.

The approach used by Farrington and Welsh when investigating regression towards the mean in police crime-rate data involves banding. This is not the way to proceed and will mask the effect being looked for especially when correlation is high as is the case and as is expected for these large police areas. However, the burglary data from the 124 areas demonstrates the effect nicely. See next.

Showing regression towards the mean with the burglary data

We divide the cases into 2 groups depending on whether the burglary rate for period one was above or below the mean of the 124 areas, for the first period. Thus we have a high burglary rate group and low burglary rate group at the start. We then calculate the change in the number of burglaries for each area. These changes are plotted for the high and low burglary rate groups

Figure 9



We can clearly see that on average the number of burglaries falls much more in the high burglary rate group compared with the low burglary rate group, exactly as expected from regression towards the mean. The mean drop for the high group is 71 burglaries whereas there is a rise of 6 on average for the low rate group. A similar picture emerges if we plot

change in burglary rate rather than the change in number of burglaries as we have, or use periods 2 and 3 instead of periods 1 and 2.

With the data set being used, burglary rate is falling across the 3 time periods. If instead the average burglary rate were constant we would expect the centres of the two clouds of points to be more equally displaced either side of the position of zero change.

One can readily see here the peril of not recognising regression towards the mean. Had an intervention been given to just the high burglary areas one might wrongly attribute the mean drop of 71, compared with a rise of 6, to the effect of the intervention. However the drop is nothing more than the consequence of correlation that Galton discovered more than a century ago that has come to be known as 'regression towards the mean'.

Other effects of regression towards the mean

Farrington and Welsh claim in HORS251 that lighting reduces crime both day and night in that there is no preferential crime reduction effect at night. Therefore they suggest a theory of crime reduction by increased community pride rather than by increased surveillance. An alternative superior explanation is that the lighting has no effect and high crime both day and night falls simply by regressing towards the mean. That is the overall level of criminality in the high crime rate area moves towards a more average value. (Also, even if the community pride mechanism were true one must surely wonder whether lighting was the most cost effective method of increasing it.)

Painter and Farrington (1999) claim on the basis of the Stoke project that lighting also reduces crime in a surrounding area that does not receive additional lighting. However, the surrounding area, which showed a drop in crime, was a high crime area at the start and so regression towards the mean can explain this naturally.

The individual studies cited

Birmingham

As pointed out previously (Marchant 2004), there is large variation, much larger than Poisson, shown in the before period. Farrington and Welsh counter this observation of mine by stating this variation was due to changes in policing. This may be the case or it may not be, we do not know.

If there were substantial changes in an external factor this would make any experiment like this unreliable. Such vulnerability is a critical weakness of non-randomised studies. It has to be remembered that it was Farrington and Welsh who included this study in their meta-analysis and analysed it as though Poisson variation applied and should have spotted this before their untenable assumption was pointed out. After-the-event justifications are scientifically weak. The fact that the authors of this study (Poyner and Webb 1997) claim a lighting success for their day-time indoor-market study is irrelevant. We have to make a judgement on the facts.

Bristol

Again Farrington and Welsh engage in after-the-event justifications and speculations

It is first claimed, in HORS251, that there is a p= 0.000 000 001 effect for lighting but it now said on the basis of a particular regression model that this has become p=0.011. The model used which yields this is one in which there is a linear trend in crime count over time, and the trend is identical in both areas. It also forces the standard deviation of the residual variability about the fitted line to be identical in the two areas, which surely is not ideal when the areas' crime counts are markedly different. The fitted coefficient given, see their reply page 460, is an increase of 21 crimes per 6-months period. However we also see here that the p-value for this effect is p = 0.068, so the rising crime term in their model is not statistically significant at the standard p = 0.05 level. Removing it and using the simpler model with no time trend gives a non statistically significant p-value for the lighting effect p = 0.063. Going the other way and making a slightly more complex model, in which a linear time trend exists but is now allowed to be different in the two areas, again gives a non statistically significant effect p = 0.118.

When I was first examining the Bristol data, in early 2003, I used a model in which the crime count in the comparison area is used as a covariate along with the lighting increase. This is more reasonable and in the spirit of HORS251 in that the comparison area exists to indicate the underlying fluctuating level of crime, which may not be linear nor indeed any simple function of time. The p-value for lighting increase in this approach is non significant at p = 0.224.

Incidentally on p459 their attempt to use the previous value of crime count to predict the next (i.e. include an autoregressive term) is discussed but the authors say that this turned out to be non-significant. This indicates weak correlation between successive crime counts. This is just the situation in which regression towards the mean has a big impact.

Surely the plot of the data shown in Marchant 2004 (page 443) should convince anyone that there is no strong evidence that lighting reduces crime. Under the null hypothesis that lighting has no effect, is what is seen in the Bristol data so extreme that such an occurrence would only occur 1 in a hundred times by pure chance? I doubt that many will think so. The p = 0.011 that they use to claim an effect is simply an issue of model selection in the face of inherent uncertainty of which potential underlying model is appropriate, if any.

Dudley

In Marchant 2004 I pointed out that the Dudley study had not found a statistically significant effect in terms of its own design criterion as a 2-sided test had not been used. The true 2-sided p-value should be = 0.070 and not p= 0.035. (It needs to be remembered that an important part of the scientific trial method is that one sets the parameters of the experiment including the acceptable error rates at the design stage. Once the trial has been completed it is wrong to try to go back and revise the criterion on which evidence is judged.)

Furthermore in a non-randomised study like Dudley, other factors that impact need to be included. The subset of data from the study, kindly sent to me by Prof. Farrington, included just two extra variables beyond the number of crimes experienced for each household and the relevant area and time period. These were the two variables on which the two areas being compared differed considerably, as discussed in the report of the study, Painter and Farrington (1997). One was whether a police officer had been seen in the area in the last month and the other was the age of the respondent given in 4 age-bands. The single most influential factor in the logistic regression is whether the respondent is aged 60+ or not. (Young people are much more likely to have experienced crime than the old.) When the variable indicating age 60+ or not is included in the model the p-value of the interaction term, which is the indicator of whether lighting had worked, rises to 0.106, moving it further away from statistical significance. (On the other hand, including the 'seeing a police officer' variable had no substantial influence on the results.)

What is critical to the Dudley study's findings is the impact of differential loss to follow up. For the data I have been sent, the following is clear. In the comparison area 77 households were lost out of 448 (i.e.17%), yet 33 of these 77 lost were aged 60+, (i.e. 60+ group fell from 157 to 124). In the relit area on the other hand 14% of the 431 households present at the start were lost, but only 2 of these 59 households lost were aged 60+; (109 dropping to 107). Thus the comparison area becomes relatively younger and hence more crime prone, whilst the opposite is true in the intervention area, during the progress of the experiment.

Sensitivity analysis with this data shows that just one additional person in the data file will alter the p-value given as 0.106 above, to a value in the range 0.096 to 0.116, depending on what area and what time-period and whether that additional person is 60+ or not.

Loss to follow up is always a problem, especially when it is differential, and must introduce more uncertainty in results because of potential statistical bias. One can easily see that under these circumstances a claim that increased lighting reduced crime in Dudley must be unfounded.

So far this discussion for Dudley has treated the crime events as spatially independent. Bringing in the fact that neighbours' experiences are correlated will introduce the 'design effect', discussed earlier and therefore considerable overdispersion, causing the variance to be increased, more than their revised calculations say. This will tend to increase the p-value much further away from statistical significance.

Clearly the authors, who praised the Dudley and Stoke evaluations, quoted in the reply cannot have recognised the fatal shortcomings. These are (1) the massive overdispersion, which is even greater than that now acknowledged by the reply's authors because of previously unrecognised spatial correlation, and (2) the effect of regression towards the mean. It is puzzling how work that failed to link addresses, before and after, on both occasions can be described as a "technical *tour de force*".

Therefore, contrary to the claims of Farrington and Welsh, none of the three studies (Birmingham, Bristol or Dudley) justifies a claim for lighting reducing crime. This is in line with lack of sound evidence from the other lighting studies due to the large overdispersion seen earlier (and consistent with the overdispersion exhibited in the burglary data).

Cost benefit analysis

This involves multiplying two uncertain estimates together, that of the effect and that of the cost/saving. I have shown that the size of the effect of crime reduction is very uncertain and either an underlying reduction crime or an underlying increase in crime is consistent with the findings. Thus performing cost benefit analysis is worse than futile as it appears to lend weight to an unreliable result.

The reply of Farrington and Welsh poses the question of whether a local council should accept paying for more lighting on the grounds that the evidence of the effectiveness of increased street lighting was unconvincing because the confidence interval for the odds ratio included 1.0, i.e. the estimate is also consistent with increased lighting increasing crime. The answer is to do the publicly responsible thing, get good accurate evidence, before spending substantial sums on something which may indeed make matters worse.

Sceptical enquiry

The basis of scientific method is sceptical enquiry. We need good evidence to accept propositions. The trouble with the reply is that it seems to start from the conclusion that lighting beats crime and that anyone of a different more sceptical view must show that lighting does not work. It is unscientific for example to suggest that because the rather complicated and arbitrary models, with many free parameters to be fitted, discussed on p452/3 of the reply, might counter the highlighted problem, that it is alright to conclude that there is not really a problem. What I have shown in the preceding sections, demonstrates that there are indeed very serious problems.

The original HORS251 shows inconsistent appraisal of results of studies. For example, when the US studies give a result that any benefit of lighting is not statistically significant, it is remarked that it is nearly significant. But when the Bristol study shows that robbery increased under the new lighting, this is discounted because it is "affected by small numbers", not withstanding that the confidence intervals, if correctly calculated, take into account sample size.

Also in the case of Bristol, had the data for the year immediately prior to the introduction of the relighting, i.e. periods 2 and 3, been used rather than unnaturally using periods 1 and 2 which leaves a gap of ½ year before the introduction of the new lighting, the effect found would have been only half of that claimed for this study.

A point made in my original article p447 is that of potential funding bias, so that the sources of funding of all the studies needs to be made clear. However this is not addressed at all in their reply, as the required list of funding sources is absent. (Bodenheimer (2000) wrote on the problem existing with research in pharmaceuticals, where industry-funded studies tended to get results more in favour of a company's products than those studies that are independent of such funding when testing the same products). Publication bias, the tendency for 'statistically significant' results to get published whereas non-significant findings tend to disappear without trace also causes research to overestimate positive effects.

Conclusion

Above I have shown that the conclusion that Farrington and Welsh give in their reply, restating their claim that there is a statistically significant reduction in crime when lighting is increased, is clearly wrong. (The conclusion given in their addendum to HORS251 is likewise wrong.)

The conclusion should be:

With the large but unknown and undoubtedly variable overdispersion, the 'weights' to be applied that are necessary for the synthesis cannot be properly calculated. However, the large size of the overdispersion causes the results to be fatally imprecise. Regression towards the mean will work in favour of spurious false claims of the effectiveness of lighting as the high crime areas received the relighting. Thus there is no sound evidence that increased street-lighting does anything to crime. An increase or decrease in crime is consistent with the evidence.

I have shown that other crime (burglary) data supports the existence of large overdispersion and demonstrates the problem of regression towards the mean.

It is mistaken to proceed with 2-area before-after studies thinking of these as the criminological equivalent of a clinical randomised control trial (RCT). Therefore work using the methods of Farrington and Welsh such as their own with the Campbell Collaboration is bound to fail.

What is needed are trials which use the highest standards and where experiments are properly conducted and carried out according to sound statistical design, borrowing the best practice from research in healthcare (e.g. using sound randomisation).

On the other hand sound observational/'epidemiological' methods would also be useful in attempting to find the effect of lighting on crime. Here, the changes in crime-levels of a large number areas could be seen in relation to the changes in lighting levels. This would help establish any underlying relationship between light and crime. Satellite monitoring of changes of the night-time illumination of a large number of areas could be used to tie these changes with the changes in their levels of crime, using multilevel models.

Until better research becomes available it should be clear from the above that the claim that increased street-lighting reduces crime is without any foundation.

Acknowledgement

I should like to express thanks to Norman Powell and Terry Sutton of Leeds Metropolitan University and also to Paul Baxter of the Statistics Department at the University of Leeds for useful discussions. Also thanks go to Prof Nick Tilley from Nottingham Trent University for providing me with the data from 124 areas and to Prof Farrington for sending me the subset of the Dudley data.

References

Bland J.M. (1995) *Introduction to Medical Statistics* 2nd Ed Oxford University Press, Oxford

Bland J.M. (2003) *Cluster Randomised Trials in the Medical Literature*: A talk first presented to the RSS Medical Section and the RSS Liverpool Local Group. www-users.york.ac.uk/~mb55/talks/clusml.htm

Bodenheimer T. (2000) *Uneasy Alliance -- Clinical Investigators and the Pharmaceutical Industry*, The New England Journal of Medicine **342** 1539

Egger M., Davey Smith G. and Altman D., (2001) *Systematic Reviews in Health Care: Meta-analysis in context*, BMJ Publishing, London

Evans M., Hastings N. and Peacock B. (2000), *Statistical Distributions* 3rd Edn. Wiley

Farrington D.P. and Welsh B.C. (2002a) *The Effects of Improved Street Lighting on Crime: A Systematic Review*, Home Office Research Study 251, http://www.homeoffice.gov.uk/rds/pdfs2/hors251.pdf

Farrington D.P. and Welsh B.C., (2002b) Improved Street Lighting and Crime Prevention, *Justice Quarterly* **19** 313

Farrington D.P. and Welsh B.C. (2003) *Addendum added to The Effects of Improved Street Lighting on Crime: A Systematic Review*, Home Office Research Study 251, http://www.homeoffice.gov.uk/rds/pdfs2/hors251.pdf

Farrington D.P. and Welsh B.C. (2004) Measuring the Effects of Improved Street Lighting on Crime: A reply to Dr. Marchant *The British Journal of Criminology* **44** 448-467 http://bjc.oupjournals.org/cgi/content/abstract/44/3/448

Kish L. (1965) *Survey Sampling*, Wiley

Marchant P.R. (2004) A Demonstration that the Claim that Brighter Lighting Reduces Crime is Unfounded *The British Journal of Criminology* **44** 441-447 http://bjc.oupjournals.org/cgi/content/abstract/44/3/441

Marchant P.R. and Baxter P.D. (To be submitted for publication) Some Statistical Issues in the Assessment of Area Crime Reduction Interventions.

Matthews J (2000) *Introduction to Randomised Controlled Clinical Trials*, Arnold

Painter, K. and Farrington, D. P. (1999) Street Lighting and Crime: Diffusion of Benefits in the Stoke-on-Trent Project, *Crime Prevention Studies* **10** 77-122

Painter, K. and Farrington, D. P. (1997) The Crime Reducing Effect of Improved Street Lighting: The Dudley Project, in R.V. Clarke ed., *Situational Crime Prevention: Successful case studies* 209-226 Harrow and Heston, Guilderland NY.

Poyner, B. and Webb B. (1997) Reducing Theft from Shopping Bags in City Center Markets, in '*Situational Crime Prevention: successful case studies'* ed. R.V. Clarke. P83-89, pub. Harrow and Heston.

Shaftoe, H (1994) Easton/Ashley, Bristol: Lighting Improvements, in S. Osborn (ed.) *Housing Safe Communities: An Evaluation of Recent Initiatives* pp72-77, Safe Neighbourhoods Unit, London.

Stigler S M (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900*, Bellknap Harvard

Tilley N., Pease K., Hough M. and Brown R. (1999) *Burglary Prevention: Early Lessons from the Crime Reduction Programme*, Crime Reduction Research series Paper1 London Home Office

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A. C., Burney, P. G. J., and Donner, A. (1999). Evaluation of Health Interventions at Area and Organisation level, *British Medical Journal*, 319: 376-379

http://bmj.bmjjournals.com/cgi/content/full/319/7206/376

Welsh B.C. and Farrington D.P. (2003a) The Effects of Improved Street Lighting on Crime: Protocol for a Systematic Review submitted to the Campbell Crime and Justice Group 3rd revision November

http://www.aic.gov.au/campbellcj/reviews/2003-11-StreetLighting.pdf

Welsh B.C. and Farrington D.P. (2003b), The Effects of Closed Circuit Television Surveillance on Crime: Protocol for a Systematic Review submitted to the Campbell Crime and Justice Group 3rd revision November

http://www.aic.gov.au/campbellcj/reviews/2003-11-CCTV.pdf