Chapter 2

Crime and Statistics

2.1 Attempts of recording crime data in order to cut down crime activity

The increase of criminal activity worldwide calls for the help of statistics in criminal justice. All over the world, scientists of many seemingly unrelated fields are cooperating in order to find methods to prevent or to reduce crime. In a world conference that took place on September 21st -25th 1998 in South Africa, with theme “Modern Criminal Investigation, Organized Crime and Human Rights”, it is noted that crime has become more and more sophisticated, technologically advanced and internationalized. It is also noted that this is a global trend and brings in its wake very complicated issues that have, and will continue to have, very severe implications on the maintenance of law and order. Therefore, there is a definite need for cooperation in order to combat crime.

Governments should provoke and support their police departments and Criminal Justice offices in efforts like the one of the Illinois Criminal Justice Information Authority. The Illinois Criminal Justice Information Authority is a state agency created in 1983 to promote community safety by providing public policymakers, criminal justice professionals and others with information, tools and technology needed to make effective decisions that improve the quality of criminal justice in Illinois. The Authority provides an objective system-wide forum for identifying problems in criminal justice, developing coordinated and cost-effective strategies, and implementing and evaluating solutions to those problems. Two of the Authority’s many responsibilities are serving as a clearinghouse of information and research on criminal justice and undertaking research studies to improve the administration of criminal justice.

Since 1989, the Authority's Research and Analysis Unit has received funds under the federal Anti-Drug Abuse Act of 1998 to document the extent and nature of drug and violent crime in Illinois and the criminal justice system's response to these offenses. As a result of these efforts, the Authority has amassed a large amount of data measuring the extent and nature of drug and violent crime in Illinois and the impact these crimes have had on the criminal justice system. To put this information into the hands of Illinois' criminal justice policymakers in a useful summary format, the Authority's Research and Analysis Unit has developed profiles of the criminal justice system for each county in Illinois. In addition to providing policymakers with an overview of activities across the components of the justice system in their county (law enforcement, courts and corrections), the profiles also provide perspective by including trends experienced in counties with similar population sizes. Moreover, the data is readily available and consistently defined through existing statewide data collection mechanisms. Therefore, it would be of great importance for all countries to develop similar data collection mechanisms, to be able to develop their own county-profiles.

Unfortunately, crime data collection is lacking of such mechanisms in Greece, leading to poor information on crime activity. The lack of sufficient data leads to a small number of analyses to be done, whereas a more complete crime database could be subjected to further analyses, whose inferences would be of tremendous importance for our society.

Moreover, the original information is lost in an effort to tabulate all the offenses that took place during a one-month period. When a crime is reported to a police department, the police officer writes down the date, the time, and the place of the crime occurrence as long as all other available information. Once a month an employee of the Ministry of Public Order visits each police department and tabulates the counts of the reported offenses in a given form. Therefore, the obtained data lack detailed information. Then, the completed forms of each police department are gathered for each police headquarters and the counts for each offense are summed to give a total count of the offense for each unit. These counts are finally summed to give a total count for Greece corresponding to that one-month period.

Thus, the available data is only monthly or annual counts for each offense for Greece and also for each of the police headquarters, which coincide to Greek counties.

The original information can only be found to the incidents books of each police department, where the reported offenses are not categorized. Hence, if one would like to have all information about a reported offense, one should visit all the police departments of the area under consideration (supposing one has been given the permission to do so), and read all the incidents books of each department for the period of interest. This procedure is rather time consuming even for one police department. Note that the yearly book with the counts of the reported offenses for Greece is published at least a year later than the year under consideration.

Unfortunately, Greece lacks of a systematic database, which would be continuously updated online, so that there would be no loss of information and the access to the data would be feasible for the researcher. The first step to the creation of this database could be the use of a computer instead of a writing book in the police departments where all the reported crimes would be recorded using code numbers, so that they could be easily organized.

2.2 Statistical Applications in Criminology

Statistics involvement in criminology is not as recently born as one might think. Although crime has risen the last decades, it seems to have concerned researchers long ago. Grünhut (1951) claims that in criminology a scientific study of crime almost coincided with the beginning of criminal statistics. One is amazed by the way the annual crime statistics were born, so let us give an exact part of Grünhut's text:

> "…Our traditional penal system has as its basis still something of the Benthamite conception of a man being motivated by the balance between the pleasure he expects from the intended offence and the pain he fears

from the impending punishment. Bentham, however, for all his rationalistic generalizations, was a strong believer in empirical observation. When the Penitantiary Bill of 1778 provided for regular statistical returns by the Governors of the intended new penal establishments, Bentham (1778) claimed that he had already begun "sketching out a plan for a collection of Bills of Delinquency with analogy to the Bills of Mortality", though – so he expected – "a little more accurate". Such Bills of Delinquency would be of "excellent use to the legislator", "a political barometer", and should be published in the *Gazette* and printed once a year as a book. The practice of collecting regular returns from criminal courts began in connection with the new forms of public administration which emerged after the French Revolution. It originated in France in 1803. In England, from 1805 to 1856 tables of trials at Assizes and Quarter Sessions were regularly printed as Parliamentary Papers (*Criminal Statistics* (1928), Introduction). In 1827, Peyronnet, Chief of the Paris Police, published the first detailed presentation of French criminal statistics, covering the year 1825. When on February 28th, 1828, Sir Robert Peel moved in the House for a Select Committee to inquire into the increase of criminal commitments and the state of the Police in London, he recommended that "admirable" work to the attention of all those interested in such questions "either in the House or in the Criminal Courts" (*Handsard* (1828), New Series, 788). From 1834, Samuel Redgrave published abstracts of these statistics. In 1857, under the direction of Earl Grey, he edited the first comprehensive publication of *English Criminal Statistics*, covering the year 1856. Police, court and prison returns were combined and issued by the Home Office as part I of the annual *Judicial Statistics*. In the last decade of the century, simultaneously with the Gladstone Committee on Prisons, another Departmental Committee was appointed to revise the Criminal Portion of the *Judicial Statistics.* Their Principal Report of 1893 combines a critical assessment of the methods of criminal statistics with a comprehensive survey of the crime situation in England (*Judicial Statistics,* 1893). It was followed by a Final Report of 1895 and a Supplementary Report of 1899, and determined the form of a separate issue of criminal statistics for England and Wales up to the present time.

A scientific evaluation of the accumulating material began in the thirties of the last century, that first "era of statistical enthusiasm". It was Adolphe Quetelet, Belgian astronomer and mathematician (*Quetelet*, 1831, 1836), and the French lawyer, A. M. Guerry (1833), who resorted to statistics in order to base the study of crime on a solid foundation of empirical observation. For this purpose they deliberately adopted methods from physics and mathematics which had proved so successful in the rise of modern science and technology. Quetelet called his *Study on Man a "physique sociale"*, Guerry claimed to attain *"une certitude complete"*, and the historian of the French Revolution, Jules Michelet, even asserted, "the science of nature and the science of justice are one and the same". Thus, "chance, only a veil for our ignorance" (Quetelet), was to be eliminated, and the multiform phenomena of social life reduced to simple unalterable laws. Those early criminal statisticians were indeed impressed by the regularity and persistence of contemporary crime. Guerry found in France every year the same number of people charged with criminal offences, the same order of frequency among the particular crimes committed, and the same distribution of offences with regard to season, region, sex, age, weapons used and other accompanying circumstances. This might have been a true reflection of the stable years of the French restoration. For Quetelet it was the effect on an "iron budget of crime" in which, as in the monetary budget of the Treasury, the expenses are fixed for mankind, but must be paid in terms of prisons, fetters, and the scaffold. Even when Guerry worked out a comparison between crime in England and France, he used 16 years of English and 22 years of French statistics, not to trace any characteristic trends throughout those periods, but only to compute yearly averages for the design of maps indicating the relative intensity of crime rates in different parts of the two countries…"

It is quite impressive that not only the need to record crime was first stated around 1800, but also that the need to give a graphical presentation of crime activity in a map was also generated around 1830, when everything had to be made by hand, without the assistance of any computer software.

In the sequence we briefly present some more recent works concerning crime statistics. Besides those presented in this paragraph one can see Tarling (1986) for a report of other statistical applications in criminology.

Ogburn (1935) studies the factors that cause some cities to have higher crime rates than others. Since some factors, such as the effectiveness of the police force, could not be measured, Ogburn focused on measurable factors. Such factors were the nativity of population, size of cities, the excess of males, church membership, the extent of manufacturing, the size of family, rates of growth of population, the amount of rent paid, the average earnings per wage earner in manufacturing and the home ownership. The data Ogburn used to indicate crime rates were the total number of offenses known to the police of a city in 1930, for the offenses: criminal homicide, rape, robbery, aggravated assault, burglary and larceny. The data were from the "Uniform crime Reports" of the Department of Justice of Chicago. It is worth mentioned that 1930 was the first year of the service, and Ogburn notices that the data were "…generally considered to be the best indications we have of the number of crimes committed". However, in order to guard somewhat against possible defective reporting during the first year of the service, Ogburn omitted from the study all cities where the total number of offenses known to the police was greatly different in 1931 from the number in 1930.

In order to have some homogeneity in crime rates overcoming the problem of the city size, Ogburn used three samples of cities of different sizes. The first sample was of 16 large cities with a range from 250000 to 578000 inhabitants. The second one was of 24 middle-sized cities from 100000 to 168000 inhabitants. The third one was of 22 still smaller cities from 36000 to 58000 inhabitants. Cities from the Southern states were omitted because of the Negro population, to reduce further the heterogeneity.

For each of the three groups of cities Ogburn correlated the crime rates with about 25 factors, of which 8 or 10 seemed to have an appreciable influence on the variation in crime rates. In seven of these factors Ogburn have accounted for over 50% of the influences causing crime rates to vary. These seven factors were the percentage of foreign born, the percentage of the offspring of foreign born, the sex ratio, the amount of rent paid, the number of church members, the change in the rate of growth and the change in wages. Ogburn noted that the influence seemed to fall in clusters. Those clustering around the immigrant included their offspring, religion, size of family, and manufacturing. These influences operated to lower crime rates. The economic cluster included changes in rates of growth, changes in wages and monthly rentals. The higher economic status was associated with lower crime rates. The third cluster was the sex ratio with such factors as those with which it was correlated. The more of males in a city, the higher was the crime rate.

Ogburn also concluded that of the various economic factors which were favorable to low crime rates, the increases in the rate of growth of cities in population was of more influence than the index used for increase in wages. The factor represented by rent seemed to be less influential. Cities with larger percentages of immigrants and of the offspring of immigrants in their population had lower crime rates. Cities with more church members and larger families also had lower crime rates.

Ahamad (1967) analyzed the number of criminal offenses of eighteen different types for Great Britain for the years 1950-63. The data were analyzed using PCA in order 'to investigate the relationships between different crimes and to determine to what extent the variation in the number of crimes from year to year may be explained by a small number of unrelated factors'. Ahamad (1967) analyzed the (18×18) correlation matrix of the data and found that three components accounted for 92% of the total variation. He interpreted the first component as measuring the general increase in crime due to changes in population structure and social conditions. This interpretation of the first component given by Ahamad (1967) coincides with the interpretation by Chatfield and Collins (1980). However, trying to interpret the second component, Ahamad (1967) suggests that this component may 'reflect changes in recording practice by the police over the period'. Chatfield and Collins (1980) comment this interpretation by Ahamad as rather dubious. Moreover, Ahamad (1967) admits that identification of the third component is 'extremely difficult'.

In Chapter 3 of this thesis we also apply Principal Components Analysis for both the annual and the monthly Greek data.

Kent Borowick (1997) in his Ph.D. dissertation entitled "Analysis of low-probability count data with applications in crime analysis" says that crime was chosen for his research because it has become a measure for quality of life and the public is increasingly concerned with reducing and controlling criminal activity. This dissertation focuses on analytical techniques, which are useful for low probability count data such as crime. Three major topics are discussed. The first topic includes possible distributions of crime data along with a general description of the data and collection procedures. The second topic reviews nonparametric statistics used to test for significant differences in the location of two populations. The third topic introduces the use of process control techniques to crime analysis.

Some of his results are worth mentioned. Three models are suggested to be more appropriate for the crime count data.

The first model is the *Poisson-Binomial* model:

$Y \mid X = x \sim Binomial(x,p)$

$X \sim Poisson(\lambda)$

where

Y is the number of crimes that occur and are reported

X is the number of crimes that actually occur

p is the probability of a crime being reported once it occurs

and λ is the average crime rate per unit time.

This model is based on the assumption that the probability of a crime to occur is independent of the probability that a crime will be reported once it occurs. The obtained distribution of Y under this model is $Poisson(\lambda p)$. However, crime counts do not always follow a Poisson distribution.

The n period hierarchical *Poisson-Binomial* model with a Bayesian prior on the λp parameter is also concerned, that is

$Y_i|\theta \sim Poisson(\theta)$

$Z|\theta \sim Poisson(n\theta)$

$\theta \sim Gamma(z_0+1,n_0)$

where

Y_i is the number of crimes that occur and are reported in the i-th period

Z is the total number of reports for the first n periods (i.e. $Z = \sum_{i=1}^{n} y_i$)

θ is the reporting rate (i.e. $\theta=\lambda p$)

p is the probability of a crime being reported

λ is the average crime rate per designated area

z_0, n_0 are the prior parameters.

Under this model, the resulting distribution of Z with

$$f(z) = \frac{n_0^{z_0+1}}{\Gamma(z_0+1)} \frac{\Gamma(z+z_0+1)}{z!} \frac{n^z}{(n+n_0)^{z+z_0+1}} \tag{2.2.1}$$

$$= \binom{z+z_0}{z} \left(\frac{n_0}{n+n_0}\right)^{z_0+1} \left(\frac{n}{n+n_0}\right)^{z}$$

can be represented as $Z \sim \textit{Negative-Binomial}\left(\frac{n_0}{n+n_0}, z_0+1\right)$, or according to Bernardo and Smith (1994) more convenient Bayesian representation of equation (2.2.1) as $Z \sim \textit{Poisson-Gamma}(z_0+1, n_0, n)$.

The third model is the *n* period hierarchical Poisson-Binomial model with joint priors. Under this model we have:

$Y_i|\lambda,p \sim Poisson(\lambda p)$

$Z|\lambda,p \sim Poisson(n\lambda p)$

$\lambda \sim Gamma(w_0+1,n_0)$

$p \sim Beta(z_0+1,w_0-z_0+1)$

where

Y_i is the number of crimes that occur and are reported in the *i*-th period

Z is the total number of reports for the first n periods (i.e. $Z = \sum_{i=1}^{n} y_i$)

p is the probability of a crime being reported

λ is the average crime rate

w_0, z_0, n_0 are the prior specifications.

The Gamma prior represents the distribution of the total rate for both reported and unreported incidents. The Beta distribution is selected because it gives flexibility over the domain of $0<p<1$. Moreover, it is assumed that λ and p are independently distributed. Under these assumptions, the obtained distribution of Z is given by

$$f(z) = \frac{\Gamma(z+w_0+1)}{\Gamma(w_0+1)} \frac{\Gamma(z+z_0+1)}{\Gamma(z_0+1)} \frac{\Gamma(w_0+2)}{\Gamma(z+w_0+2)} \left(\frac{n}{n_0}\right)^z \frac{1}{z!}$$

$$\times F(z+w_0+1, z+z_0=1; z+w_0+2; -\frac{n}{n_0})$$

All models are interpreted in terms of crime data and fit to crime counts for Waco, Texas. Two goodness-of-fit tests are reviewed and used to test the fit of each marginal distribution to the observed data. Kent Borowick (1997) has shown that the two Bayesian models had the best fit overall.

Furthermore, Kent Borowick (1997) suggests that the use of quality control techniques provides a significant opportunity to improve crime analysis methods using a methodology that is widely utilized in business, industry and government. In particular, he suggests that c-charts are the most suitable quality tools for crime data.

Karlis (2001) introduces an EM algorithm for maximum likelihood estimation for the case of mixed Poisson regression models, aiming to reduce the computational burden for maximum likelihood estimation in those cases. The algorithm makes use of the mixture representation of such models. He especially examines two members of this family, the negative binomial regression models and the Poisson-inverse Gaussian regression model. In order to illustrate the proposed algorithm Karlis (2001) fitted the two above mentioned models in a data set concerning the number of manslaughters in a series of Greek prefectures for 1997.

The model assumes that, for the i-th prefecture the number of manslaughters y_i follows a $Poisson(\lambda_i)$ distribution, while the Poisson rate λ_i depends on a series of covariates through a log-link function. As such variables he used the natural logarithm of the population of the area (in millions), the Gross Domestic Product per capita (in Euros) for each prefecture, the unemployment rate of the prefecture and two dummy variables that show whether the prefecture is at the borders of the country (to account for the increased number of economic refugees crossing the borders) and whether the prefecture has at least one city with population larger than 150000 inhabitants.

A simple Poisson regression model gave bad fit due to the presence of overdispersion. Hence, he fitted mixed Poisson regression models, by assuming a Gamma and an inverse Gaussian mixing distribution for θ. The two models were fitted using the proposed in his paper EM algorithms. In addition, the Poisson–normal model was fitted using the PROC NLMIXED procedure of SAS.

The Poisson-inverse Gaussian model seems to fit the crime data slightly better than does the negative binomial, mainly since it can model better the right tail of the data. However, the fit is similar to that of the Poisson-normal model.

Van der Heijden, Cruyff and Van Houwelingen (2003) try to estimate the number of offenders from police data. Their aim is to estimate the number of offenders never apprehended, using the data about offenders apprehended at least once. They derive these estimates under two assumptions. They assume that the number of apprehensions is a realization of a Poisson distribution, and that the logarithm of the Poisson parameter for an offender is a linear function of covariates. They note, however, that these assumptions need careful considerations as violations may seriously distort the estimate.

Elffers (2003) investigates the neighborhood influence in criminology, suggesting some statistical analyses depending on the neighborhood characteristics under examination. He proposes four questions that may be of interest in criminologists and he divides them according to two criteria. The first one is whether the unit of analysis is the neighborhood as such, or the individual within the neighborhood, and the second one is whether we expect the influence on individuals to be effective only within the neighborhoods themselves, or also in adjacent neighborhoods. Those two criteria lead to the proper statistical analysis for each question under consideration. Elffers (2003) suggests that the first criterion discerns between simple and multilevel analyses, while the second one does so between independent and autocorrelation analyses.

For the case of the between neighborhood influence, Elffers suggests multilevel modeling, emphasizing within the ecological tradition. In the ecological tradition, theory assumes that what is happening to or with an inhabitant is dependent on to the neighborhood in which he or she is living. In this case, neighborhoods are supposed not to influence each other. That implies that the distribution of variables (dependent variable, array of independent variables) differs over neighborhoods. Elffers notes that the simplest statistical model for such a neighborhood influence model is one of *k* populations, allowing a different distribution in each population.

Denoting the dependent variable of citizen *i* in neighborhood *j* with Y_{ij} and its explanatory variables with a *k*-dimensional vector X_{ij} he specifies for each *j* that $Y_{ij} = \alpha_j + X_{ij} \cdot \beta_j + \varepsilon_{ij}$ with regression constants α_j and β_j, which will differ across neighborhoods. Thus, he reduces the analysis to repeating the separate analyses for each neighborhood, and no linking of any neighborhood characteristics forms part of the analysis.

A more refined analysis, incorporating neighborhood characteristics, is possible taking under consideration all individual regression lines simultaneously. In this case we link the mean of the dependent variable in each area to area characteristics. This type of analysis is usually performed by means of multilevel analysis (Snijders and Bosker, 1999), which is sometimes also called hierarchical linear modeling.

The simplest case, known as the *random intercept model*, is a case where we look only at differences in the intercept terms α_j of the separate neighborhood regression lines, assuming that the slopes β_j are equal. In this case, the difference in mean level of the observations in the various areas is partly due to different levels of the regressors. Multi-level analysis determines what parts of the neighborhood differences can be attributed to the distribution of individual characteristics.

A more complex model is where we also allow β_j to be dependent on *j*, the so-called *random slope model.* The difference in levels in different neighborhoods is then partly due to different distribution of *X*-variables, partly due to different α and βs, and for the remainder attributed to other neighborhood influences. Again, multi-level methods separate these influences.

Elffers (2003) also deals with the case of neighborhood influence within the adjacency tradition, according to which modeling neighborhood influence means studying the effect of neighborhood characteristics on each other. Therefore, the dependent as well as the explanatory variables are neighborhood characteristics.

He first examines the case where neighborhood influence is not assumed. Thus, the model's dependent variable is regressed on other area characteristics. Looking into neighborhood influence means now that we try to establish whether the values of the dependent variable in area *i* are not only dependent on area *i*'s explanatory characteristics, but also on characteristics of other areas. This approach to neighborhood influence often calls this influence *spatial dependency*.

The question that now arises is that of which neighborhoods should one suppose to be influencing each other. Elffers states that the more common way of treating that problem is to suppose that neighborhoods influence each other when they are physically adjacent. In order to ensure that such a dependency exists, Elffers computes a so-called spatial autocorrelation statistic, which quantifies whether values observed in adjacent areas are more similar or less similar than may be expected under a null hypothesis of no influence. Coloring the areas with the highest values black and leaving all others white and counting the number of boundaries between two black areas (or two white ones, or a black and a white ones) gives a so-called *join-count-statistic*. The distribution of this statistic (under the null hypothesis of randomization of values over areas, or under the normality assumptions of the values observed without neighboring influence) is asymptotically normal (Moran, 1948; Cliff and Ord, 1973, 1981). Therefore, an approximate t-test can be used for testing the null hypothesis of no spatial correlation.

Elffers (2003) tries to model the manner in which this dependency may operate in case of rejecting the absence. He introduces three autocorrelation approaches; i) the spatial autocorrelation as result of dependency on explanatory variables, ii) the spatial autocorrelation as result of dependency on adjacent explanatory variables, and iii) explaining autocorrelation by dependency on the dependent variable in adjacent areas.

Elffers concludes that the type of analysis that should be used in practice must be guided by criminological theory.

All the above-mentioned show that statisticians can cooperate with criminologists to obtain useful inferences in order to combat crime.