

Prevalence Estimate of Problematic Opiate Consumption in Austria

(second revised edition)

Alfred Uhl, Ludwig Boltzmann Institute for Addiction Research
Dan Seidler, Emergency Department of the Vienna General Hospital

Ludwig Boltzmann Institute for Addiction Research
(LBISucht)
Vienna, December 2001

Comment:

This second version of the article considers some of the readers' comments. A critical review of the comprehensibility of the previous text and corresponding adaptations and additional clarifications were done by Ulrike Kobra. The major change is that some formulas and calculation steps are described in even more detail now.

In the original paper the hidden population was estimated with SPSS and GLIM, which yielded identical results in all models except the saturated models. In the latter cases the SPSS and the GLIM results were reported, in all other cases reference was only made to SPSS. All confidence intervals were calculated with GLIM. In the present version we decided to use SPSS only, taking into account that SPSS is a much more widely spread software than GLIM. The numerical differences due to this change are generally neglectable.

The work described in this publication was commissioned by the Austrian Ministry of Science and Transport supported by UNDCP. The project has also benefited from the European Network to Develop Policy Relevant Models and Socio-Economic Analyses of Drug Use, Consequences and Interventions, funded by TSER/DG12, project ERB 4141 PL 980030 and co-ordinated by the EMCDDA, Lisbon

Contents

| | | |
|-------|--|----|
| 1 | Introduction | 5 |
| 1.1 | Background of the project in a European context | 6 |
| 1.2 | Approaches to estimate hidden populations | 6 |
| 1.3 | Sources of drug related data available in Austria | 7 |
| 1.3.1 | National treatment documentation system | 7 |
| 1.3.2 | Register of drug related deaths | 7 |
| 1.3.3 | Main register of the narcotic drug monitoring agency | 7 |
| 1.3.4 | Register for substitution treatment | 8 |
| 1.3.5 | Emergency ambulance transports related to acute opiate intoxication | 8 |
| 2 | Methodology and statistical procedures | 8 |
| 2.1 | The capture-recapture approach | 8 |
| 2.2 | Log-linear models | 10 |
| 2.3 | The odds-ratio | 10 |
| 2.4 | Confidence intervals | 11 |
| 2.5 | Consecutive capturing processes vs. simultaneous ones | 11 |
| 2.6 | Approaches to estimate the impact of assumption violations | 11 |
| 2.6.1 | Notations used in this report | 11 |
| 2.6.2 | Simple multiplication approach | 12 |
| 2.6.3 | Approximate matrix approach | 12 |
| 2.7 | Statistical packages used | 13 |
| 2.8 | Model Selection | 13 |
| 3 | The empirical project | 14 |
| 3.1 | Criteria of data used in the study | 14 |
| 3.2 | Identifiers | 14 |
| 3.3 | Relevant observation periods | 14 |
| 3.4 | Data selection criteria | 14 |
| 3.5 | Case definition and: <i>are the assumptions required by the capture-recapture approach plausible for the data set?</i> | 16 |
| 3.5.1 | Substitution treatment | 16 |
| 3.5.2 | Drug related deaths data and opiate related emergency ambulance data | 17 |
| 3.5.3 | Opiate related police charges | 17 |
| 3.5.4 | No causal relationship between sampling processes | 17 |
| 3.5.5 | Conclusions in terms of case definition and assumptions | 17 |
| 4 | Problematic opiate consumption in Austria 1995 based on the capture-recapture approach | 18 |
| 4.1 | Previous estimations of problematic opiate consumption in Austria | 18 |
| 4.2 | Results of capture-recapture calculations based on the present study | 18 |
| 4.2.1 | Non-stratified calculation for all Austrians aged 15-54 years | 19 |
| 4.2.2 | Stratification by gender | 22 |
| 4.2.3 | Stratification by age groups | 24 |
| 4.2.4 | Stratification by the nine Austrian provinces | 27 |
| 4.2.5 | The impact of stratification | 31 |
| 4.3 | Additional estimations to judge validity and reliability of the estimation | 31 |
| 4.3.1 | Summary: Estimates for Vienna | 34 |
| 5 | Theoretical considerations | 35 |
| 5.1 | Central aspects related to the capture-recapture approach | 35 |
| 5.1.1 | Advantages of the capture-recapture approach | 35 |
| 5.1.2 | Basic assumptions behind the capture-recapture approach | 35 |
| 5.1.3 | Other important aspects related to the capture-recapture approach | 35 |
| 5.2 | "Best model" chosen by statistical criteria vs. based on substantial considerations | 36 |
| 5.2.1 | Model selection in log-linear models | 36 |
| 5.2.2 | The mechanistic statistical approach and the content oriented approach to model selection | 36 |
| 5.3 | Case definition | 36 |
| 5.4 | What happens if assumption 1 is violated? | 37 |
| 5.4.1 | Scenario 1: Impact of non-cases | 37 |

| | | |
|--------|---|----|
| 5.5 | How likely is it that assumption 2 is violated? | 41 |
| 5.6 | Consecutive capturing vs. simultaneous capturing processes | 41 |
| 5.7 | How likely is it that assumption 3 is violated? | 41 |
| 5.7.1 | Scenario 2: Impact of open population | 42 |
| 5.7.2 | Scenario 3: Impact of drug related deaths – uncompensated | 43 |
| 5.7.3 | Scenario 4: Impact of drug related deaths – compensated | 46 |
| 5.8 | How likely is it that assumption 5 is violated? | 48 |
| 5.8.1 | Scenario 5: Impact of causally related processes | 48 |
| 5.9 | Assumption: No highest-order-interaction | 50 |
| 5.9.1 | Scenario 6: Impact of strongly causally related processes and highest-order interactions | 50 |
| 5.10 | How likely is it that assumption 4 is violated? | 53 |
| 5.10.1 | Scenario 7: Impact of concordant heterogeneity | 54 |
| 5.10.2 | Scenario 8: Impact of discordant heterogeneity | 56 |
| 6 | Plausibility considerations | 58 |
| 6.1 | Comparison to international prevalence estimates | 58 |
| 6.2 | Other prevalence estimation approaches based on Austrian data | 58 |
| 7 | Summary and discussion | 59 |
| 8 | Syntax for calculations and practical examples | 62 |
| 8.1 | Programme to calculate expected contingency tables based on the approximate matrix approach | 62 |
| 8.2 | Example for simple 2-sample-capture-recapture-analyses | 62 |
| 8.3 | SPSS programme code for a 3-sample-capture-recapture-analysis | 62 |
| 8.4 | SPSS programme code for the matrix procedure to construct data according to assumptions | 66 |
| 9 | References | 70 |

1 Introduction

Many interesting models to estimate hidden populations stem from the field of estimating wild animal populations. Without such estimation techniques wild life ecologists could not identify endangered species, they would not know if their attempts to protect populations at risk are successful, nobody could decide on the basis of evidence how much hunting or fishing of a certain species can take, without falling under a critical size. Another area where such models play an increasing role is epidemiology. Questions like “Did information campaigns about AIDS reduce the incidence of HIV-infections significantly? or do we still have to increase our efforts in this field?”, “How many per cent of problematic drug users are using our treatment system?” and many more, depend on such estimates.

Ideal in this context are methods that rely on already existing data sources or on data that can be collected without high costs. The less money decision makers have to spend to get a reliable scientific foundation for their decisions the more money is left to put their decisions into practise.

The task of improving epidemiological techniques to assess the prevalence of drug addiction – particularly the prevalence of opiate addiction – is a task of great national and international importance. This issue is not only of scientific interest, but also of great practical-political importance. International treaties and other obligations require national authorities to supply European and international bodies with reliable statistics on the extent and the development of the national drug problem. Until now many existing statistics in the field of drug abuse have been based on unsystematic and scientifically very questionable estimates derived from official police, court and treatment statistics.

The project underlying this report focused primarily on the “capture-recapture approach”, a very attractive technique in this context. The capture-recapture approach can be applied using existing data registries and does not require the samples to have identical coverage rates. The model behind the approach nevertheless depends on some central assumptions – perfect identification of cases, perfect identification of matching cases, closed population, homogeneity, independence of the sampling processes – and unfortunately these assumptions are hardly ever perfectly correct in real life.

If we want to judge whether capture-recapture estimations are sufficiently exact for a certain purpose we have to have an idea how far the implicit assumption could be violated in a specific situation and furthermore how these violations influence the estimate. In the course of this report we will demonstrate that applying capture-recapture calculations to estimate the number of “problematic opiate users” in a region based on registry data is not very reliable, but sufficiently reliable for certain purposes. We have to be aware that – at least without quantifying the degree of violation to the assumptions and without adjusting the calculations for these violations – we must expect the true population size to be somewhere between 50% less or 100% more than the calculated estimates. Since this refers to a bias and not to random variation this inexactness is naturally not confined by confidence intervals. It exists in addition to this source of variation.

In this project we developed two simple approaches allowing us to construct data based on more realistic assumptions and then to see how the capture-recapture estimates deviated from the true values. In this way it is possible to demonstrate how violations impact on actual estimates. In this project we used some plausible scenarios to demonstrate the magnitude of the possible biases. It is hoped this will lead to intensified empirical research to quantify the relevant violations inherent in available data sources, to more flexible models that allow to enter these more realistic assumptions and eventually to more reliable estimates.

This report is divided into three main sections. The first section (4.2) performs a standard capture-recapture analysis based on the model suggested by the algorithm as best models – as the analysis usually done; the second section is based on sensibility analyses (4.3) and shows that the results produced in the first section have to be looked upon rather critically and interpreted cautiously, and the third section (5.4) analyses data constructed on the basis of plausible scenarios with deviating assumptions and compares the resulting estimates to the values entered into the scenarios; thus allowing us to estimate the resulting biases.

1.1 Background of the project in a European context

In 1996 the Pompidou Group within the European Council and the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) organised a joint conference in Strasbourg with international experts on scientific methods to estimate prevalence on the local and national level. The proceedings published the proceedings (EMCDDA, 1997). Following this conference, EMCDDA started an international expert work group on local prevalence estimates. This project was co-ordinated by the University of Glasgow and finished in early 1998. The project co-ordinators published an EMCDDA report on the theoretical and practical results of the work group (Hay et al., 1997) and added sections on prevalence estimates for about 7 European cities, produced by the participating European experts for cities in their respective countries. These cities were Dublin (Comiskey, 1997), Helsinki (Kaukonen et al., 1997), Rome (D'Ippoliti, 1997), Rotterdam (Smit & Toet, 1997), Setúbal (Freire & Moreira, 1997), Toulouse (Bello, 1997), Vienna (Seidler & Uhl, 1997). Based on this project, methodological guidelines for estimating the prevalence of problem drug use on the local level were published one year later (Hay et al., 1998)

To continue the work on prevalence estimation on the European level EMCDDA initiated a three year project involving 6 subgroups on specific topics ("European Network to Develop Policy Relevant Models and Socio-Economic Analyses"). Uhl is member of the workgroups on local prevalence estimates (co-ordinated by Hay) and on national prevalence estimates (co-ordinated by Kraus) within this EMCDDA project.

On the national level Seidler and Uhl were successful in initiating a project to extend the scope of prevalence estimation to the other eight Austrian provinces and in evaluating the capture-recapture methodology more systematically. This latter project has been supported by the Ministry of Health and by UNDCP. The mentioned EMCDDA projects and the Austrian project worked in close co-operation and benefited from each other. However, this report deals only with the results of the national Austrian project.

1.2 Approaches to estimate hidden populations

There are several methods to estimate prevalence, such as surveys, direct enumeration of cases and indirect methods to estimate hidden populations based on overlap data (Multiplier Methods, Capture-Recapture Approach) or on the data distribution (Truncated Poisson Approach).

Surveys are not really suited for assessing the prevalence of problematic substance use, since surveys commonly miss out on problematic cases (undersampling) and since problematic substance users are commonly reluctant to admit problem behaviour to unknown commercial interviewers.

Direct enumeration of all cases, by so called case-finding studies, is usually a very resource consuming enterprise and commonly not feasible even if unlimited finances were available.

Multiplier methods require only one sample ("drug related death", "treatment data", "number of police charges", etc.) but require empirical information on the ratio between the known population and the hidden population – information that is commonly not readily available. Since these ratios may fluctuate significantly over time and vary from region to region, it is generally not justified to generalise them to other regions and/or time intervals. In other words, if we are interested in reliable estimations for different regions and time intervals, we have to determine the correct ratios for every region and observation period independently. It is common practice to apply multipliers for other regions and/or time intervals. This is justified if we are only interested in very crude estimates – but we have to be aware that they are very crude indeed. If, for example, we base a multiplier approach on "opiate related police charges" and drug related police activities increase dramatically the application of old multipliers naturally leads to a dramatic bias.

More sophisticated estimation approaches, such as the "capture-recapture approach" and the "truncated Poisson approach" allow us to estimate hidden populations without empirical information on ratios and both approaches are not sensitive to changes in the intensity of the registration processes (e.g. more police pressure, more and better treatment facilities, etc.) but they depend on some more or less realistic assumptions.

The present report has two aims. One aim is to estimate the number of problematic opiate users for Austria based on the capture-recapture approach in the usual way. The other aim is to define plausible

scenarios based on implicit assumptions, apply the capture-recapture approach to the data, compare and discuss the validity and reliability of capture-recapture-estimates under various plausible scenarios.

1.3 Sources of drug related data available in Austria

EMCDDA supports the collection of drug related data in the EU Member states through the national Reitox focal points and most national administrations of European countries feel a need to co-operate with these European endeavours. Such data are not only a central necessity to fulfil the mentioned European requirements but also important for evidence based policy decisions and for the evaluation of drug policy related measures and strategies.

1.3.1 National treatment documentation system

A committee of representatives from the Federal administration, the Austrian provinces, the Austrian Reitox focal point and the Ludwig Boltzmann Institute for Addiction Research have been active in conceptualising and implementing a nation-wide Treatment Documentation System for some time by now, but it is hard to foresee whether all parties will eventually agree on details and/or support its practical implementation. If the provinces and the Ministry of Health can agree on a common strategy, they have good grounds to expect good co-operation from the treatment facilities since most of these facilities receive financial support from the federal and/or province administrations.

Currently, without such a national Treatment Documentation System, the only way to produce an overview over the Austrian drug treatment situation is to collect available aggregated data from single treatment facilities or use reports from provinces that aggregate such statistics over the facilities in their province. Naturally there is no chance to exclude double counting between facilities and one has to accept the fact that the aggregated data are very heterogeneous since the reporting schemes between facilities vary considerably.

1.3.2 Register of drug related deaths

The register on drug related deaths is maintained by the Federal Ministry of Health and receives information from police and hospitals based on the toxicological data and autopsy findings collected by the forensic departments. Since the annual number of "drug related deaths" is small compared to other drug related events and since the number of "drug related death" are presented publicly at least annually with some media coverage, the responsible persons try to keep the quality of this register as good as possible. All data on "drug related deaths" are reported to the Narcotic Drug Monitoring Agency, too, after they have been validated and then entered into the main registry of this agency. "**Drug related deaths data**" will be abbreviated with "**d**" in the tables of this report.

1.3.3 Main register of the narcotic drug monitoring agency

The legal basis for the main register of the Narcotic Drug Monitoring Agency is the Narcotic Substances Act 1980. A federal law defines that this register is to include all persons charged by the police with drug related offences, record information on the outcome of these charges including court sentences, record inquiries from governmental bodies concerning the drug careers of individuals and record treatment information of cases where the treatment was initiated by public bodies. The register does not include information on treatment episodes of persons seeking voluntary treatment.

From the setting up of this register in 1980 until August 1998 more than 24,000 events were entered into the register, several of them dating back as late as 1967. Due to the limited number of data entry staff the time lag between reports coming in and data entry increased more and more. In August 1998 the lag had increased to more than 3 years. In the second half of 1999 and beginning of 2000 the register managed to close this data entry gap, but since our project is based on data transmitted to us in the middle of 1999, the latest time period covered is 1995. In the present study we use police data related to opiates. These Data will be referred to as "**opiate related police charges**" in this study and abbreviated with "**p**".

1.3.4 Register for substitution treatment

The Narcotic Drug Monitoring Agency keeps also a separate register on persons under substitution treatment. For any such treatment carried out, the beginning and the end of treatment must be reported to this register. First events covered by this register date back to 1987. This database is also a regular information source for therapists who substitute opiate addicts to prevent patients from collecting their oral substitution drugs from more than one source at the same time. Data from this source will be referred to as "**substitution treatment data**" in this study and abbreviated with "**s**".

1.3.5 Emergency ambulance transports related to acute opiate intoxication

There is no national register on opiate-related emergency ambulance transports, but we could collect a complete record of such transports in Vienna for the period between September 1994 and August 1995. These data will be referred to as "**opiate related emergency ambulance data**" in this study and abbreviated with "**a**".

2 Methodology and statistical procedures

2.1 The capture-recapture approach

The logic behind the capture-recapture approach is very simple. If we want to estimate the size of a fish population in a lake we can capture a certain number of fish at t_1 , mark them, release them and capture a certain number of fish some time later at t_2 (Tab. 1). If we multiply the percentage of marked fish at t_2 (=recaptured fish) with the number of fish captured at t_1 , we have a good estimate for the total number of fish in the lake, given that the following assumption are correct:

- We capture only fish and no other water animals
- all recaptured fish are identified as recaptures,
- the fish population must not change between t_1 and t_2 and
- both samples are true random samples of all fish in the lake

To make it more concrete, if we capture 100 fish in a lake at t_1 , mark them and release them and some time later – at t_2 – capture another 50 fish out of the same lake, and if we find that 20% of the newly captured fish have been previously marked (recaptured), we may conclude that the 100 fish initially captured were a 20%-sample of the total population. This implies that the total fish population in the lake is 500 fish (Tab. 2).

As mentioned above, some basic assumptions must be met to arrive at reliable estimates:

- In order to be able to differentiate fish from other water animals (**only cases according to the case definition are considered**) we need a clear case definition.
- Fish can only be identified reliably as recaptures (**perfect identification of matching cases**) if the markers are specific enough to be recognised at t_2 and if the fish cannot get rid of them.
- The population will only be unchanged at t_2 if no fish are born/hatch, die, migrate into the lake or migrate out of lake in the observation period (**closed population**).
- The first patch of fish captured at t_1 can only be random samples of all fish in the lake if all fish have an equal chance of being captured at t_1 (**homogeneity in terms of equal catchability**) and
- the second patch of fish captured at t_2 can only be random samples of all fish in the lake if the fact that a certain fish has already been caught at t_1 has no influence on his "catchability" at t_2 (**no causal relationship between sampling processes**).

Tab. 1: Capture-recapture method (based on two samples)

| | | sample 1 | | |
|--------------------------------------|---------|--------------------------------|--|---|
| | | present | absent | |
| sample 2 | present | a | b | a + b (present in sample 2) |
| | absent | c | $N_{hid}^1 = ?$ | $c + N_{hid}$ (not in sample 2) |
| | | a + c (present in sample 1) | $b + N_{hid}$ (not in sample 1) | $N_{tot} = a + b + c + N_{hid}$ (total estimate) |
| estimate of hidden population | | | $N_{hid} = b \cdot c / a$ | |
| total population (observed + hidden) | | | $N_{tot} = a + b + c + b \cdot c / a$ | |
| variance of N_{tot} | | | $Var_N = (a+b) \cdot (a+c) \cdot b \cdot c / a^3$ | |
| 95%-confidence interval | | | $ci_{95\%} = N_{tot} \pm Z_{0.975} \cdot Sd_N = N_{tot} \pm 1.96 \sqrt{Var_N}$ | |

Tab. 2: Fish example capture-recapture method (based on two samples)

| | | sample 1 | | |
|--------------------------------------|---------|------------------------------|---|-----------------------------|
| | | present | absent | |
| sample 2 | present | 10 | 40 | 50 (present in sample 2) |
| | absent | 90 | 360 | 450 (not in sample 2) |
| | | 100 (present in sample 1) | 400 (not in sample 1) | 500 (total estimate) |
| estimate of hidden population | | | $N_{hid} = 40 \cdot 90 / 10 = 360$ | |
| total population (observed + hidden) | | | $N_{tot} = 10 + 40 + 90 + 40 \cdot 90 / 10 = 500$ | |
| variance of N_{tot} | | | $Var_N = (10+40) \cdot (10+90) \cdot 40 \cdot 90 / 10^3 = 18,000$ | |
| 95%-confidence interval | | | $ci_{95\%} = 500 \pm 1.96 \sqrt{18,000} = 500 \pm 263 = [237; 763]$ | |

To be realistic we have to accept that these basic assumptions are practically always violated to some degree:

- Since the marking of the fish naturally must not interfere with their life or even kill them, marking has to be done very cautiously and therefore it is quite likely that some recaptured fish will not be identified as recaptures at t_2 (**imperfect identification of matching cases**).
- Even if the distance between t_1 and t_2 is very short, it is very likely that some fish are born/hatch, die, migrate into or out of the lake between during the observation period (**open population**).
- We have no means to catch fish randomly. Some fish are trap-evaders, others trap-seekers, some prefer to stay in areas where catching them is very difficult while others prefer areas where catching them is simple (**heterogeneity in terms of equal catchability**).
- Fish captured at t_1 learn from their experience and avoid to be recaptured at t_2 . The opposite may happen too, fish may get injured when caught or marked at t_1 and recapturing them may even become easier. In both cases there is a **causal relationship between sampling processes**.

As mentioned before, the most striking advantage of the capture-recapture approach is that the estimates are not sensitive to the number of fish captured at t_1 and t_2 . More exact, to capture a small sample does not cause a bias, but naturally estimates based on a large number of catches are more precise (smaller confidence intervals, see chapter 2.4).

If deviations from basic assumptions are small, the resulting estimates will be reliable enough to justify the application of the capture-recapture methodology, but it is obvious that large uncompensated violations render the approach worthless. Some violations that cause a severe bias in a 2-sample-capture-recapture calculations (e.g. a causal relationship between sampling processes) can be partly compensated for by higher order capture-recapture calculations and some other violations (e.g.

1 For notation see chapter 2.6.1.

heterogeneity in terms of equal catchability) can be partly compensated for by a stratified analysis, but not all violations can be compensated.

Generally it is to say that higher order capture-recapture-estimates are preferable to 2-sample-capture-recapture-estimates, since they control part of the violations due to causal relationships between sampling processes and that stratification is superior, since it compensates for some of the heterogeneity.

The nature and impact of the basic assumptions behind the capture-recapture approach will be discussed in more detail in chapter 5.

2.2 Log-linear models

The statistical model to obtain n-sample-capture-recapture estimates are "log-linear models". Somehow log-linear models analyse effects analogous to "analyses of variance". The central difference is that analyses of variance decompose data into additive effects while log-linear models decompose them into multiplicative effects. If additive effects in combined cells do not add up to the observed value in analyses of variance we speak of interaction and the same holds true in log-linear models if multiplicative effects do not multiply to the observed cell frequencies in combined cells.

To make this more concrete:

- If people in "country A" live 10 years longer than in "country B", and if women live 5 years longer than men; women in country A are expected to live 15 (=10+5) years longer than men in "country B". If this is not the case, we speak of interaction between "country of residence" and "gender" within analyses of variance.
- If the chance to be captured is 3 times higher for red fish than for green fish and if the chance to be captured is 4 times higher for male fish than for female fish we expect to catch 12 times (=3*4) more red male fish than green female fish. If this is not the case we speak of an interaction between "gender" and "colour" within log-linear models.

2.3 The odds-ratio

In dealing with rates we are commonly confronted with probabilities and/or odds (=chances).

- If 1 out of 10 persons in "country a" is unemployed and
- if 2 out of 10 persons in "country b" are unemployed
- the probability "p(a)" of a person in "country a" being unemployed is $1 : 10 = 0.10 = 10\%$ and
- the probability "p(b)" of a person in "country b" being unemployed is $2 : 10 = 0.20 = 20\%$.
- The odds "o(a)" of a person in "country a" being unemployed are $1 : 9 = 0.11$ and
- the odds "o(b)" of a person in "country b" being unemployed are $2 : 8 = 0.25$.
- The ratio between probabilities is called relative risk "r". The relative risk of unemployment in "country b" vs. "country a" is $r(b,a) = 0.20 : 0.10 = 2$.
- The ratio between odds is called odds-ratio "or". The odds-ratio for unemployment in "country b" vs. "country a" is $or(b,a) = 0.25 : 0.11 = 2.3$.

According to the above example we could formulate: "The probability of being unemployed in "country b" is 2 times higher than in "country a" and the odds of being unemployed in "country b" are 2.3 times higher than in "country a".

Both, the "relative risk" and the "odds-ratio", range from 0 to infinity. Both coefficients express the degree of correlation between variables. Independence results in a value of "1", coefficients less than 1 result from negative correlations and coefficients greater than 1 result from positive correlations.

The "odds-ratio" has some very nice statistical properties and therefore is essential for many statistical models dealing with rates and proportions, even though the concept of "probability" is much more appealing to non-statisticians.

"Odds" and "probabilities" as well as "odds-ratio" and "relative risk" are numerically almost identical if probabilities are low, but large difference occur if probabilities are high.

2.4 Confidence intervals

If we draw random samples from populations the sample parameters (i.e. observed means, observed rates, etc.) usually deviate more or less from the true population parameters. In large random samples the average deviance is smaller and in small samples it is larger. Confidence intervals help us to get an idea about the plausible impact of random variation due to the sampling process, but they do not help us at all to estimate the impact of systematic errors (biases). This is important to understand in this context since violations of assumptions behind the capture-recapture approach cause biases and not random variation. Confidence intervals therefore do not give us any support at all in this respect. Therefore it is totally inappropriate to expect confidence intervals to help us judge the validity of capture-recapture calculations in the face of violated assumptions.

2.5 Consecutive capturing processes vs. simultaneous ones

Capturing processes sometimes happen in consecutive order and sometimes simultaneously. By "consecutive order" we mean that capturing process 2 starts after capturing process 1 has been completed, that capturing process 3 starts after capturing process 2 has been completed, etc. "simultaneous order" means that the capturing processes happen during the same observation period. Consecutive capturing processes are more common in the field of estimating wild animal populations, simultaneous capturing processes are more common in social science application.

Consecutive capturing processes do not require cases to be marked during the last capturing process, while simultaneous capturing processes require marking during all capturing processes. In simultaneous capturing processes we need to identify cases that have been captured by other capturing processes before as well as those that will be captured by other capturing processes later.

2.6 Approaches to estimate the impact of assumption violations

Possible scenarios relevant for capture-recapture approaches can easily be described based on rates and probabilities. A simple way to assess the impact of violated assumptions in capture-recapture approaches is to define plausible scenarios where basic assumptions are violated, calculate expected cell frequencies for all conditions, ignore the number of hidden cases (those not captured by any of the capturing processes) and estimate hidden cases based on capture-recapture methodology. If the estimate deviates from the true number derived from the scenario we can quantify the amount and the direction of the systematic error (bias). This approach is pursued in chapter 5.

In order to calculate the expected cell frequencies for the scenarios we used two approaches: The **simple multiplication approach** and the more complicated **approximate matrix approach**.

2.6.1 Notations used in this report

Expected cell frequencies are symbolised by the letter n and in brackets whether the case has been captured by any of the involved capturing processes, whereby "+" symbolises a capture and "-" symbolises a non-capture. For example, the number of cases only captured by process P but not by S and D is represented by $n(s-p+d)$. The cell containing $n(s-p+d)$ in the contingency table is referred to as cell "s-p+d".

Tab. 3: Notation for cell frequencies

| | police (p) | | | |
|------------------------|------------------|-------------|------------------|-----------------------|
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| drug related death (d) | present | absent | present | absent |
| present | $n(s+p+d+)$ | $n(s-p+d+)$ | $n(s+p-d+)$ | $n(s-p-d+)$ |
| absent | $n(s+p+d-)$ | $n(s-p+d-)$ | $n(s+p-d-)$ | $n(s-p-d-) = N_{hid}$ |

All cell frequencies add up to the total number of cases symbolised by "N_{tot}" (total population). Those who were not captured by any process "n(s-p-d-)" = "N_{hid}" constitute the hidden population. The term "N_{obs}" symbolises the observed population; i.e. the total population minus the hidden population.

The unconditional probability of being captured by process D ("drug related deaths") is p(d+). The conditional probability of being captured by process D, given that the person has not been captured by processes P ("opiate related police charges") and/or S ("substitution treatment data") previously is symbolised by p(d+/s-,p-).

In case the conditional probability of being captured by process D, given that the case has already been captured by process P previously, differs from the above probability, the factor to describe the change in probability is symbolised by f(p→d). The relationship is as follows:

$$p(d+/s-,p+) = p(d+/s-,p-) * f(p→d)$$

If we do not consider the causal direction of relationship, but simply consider that a certain combination is more likely than expected under independence assumption, the factor to quantify the change in probability is symbolised as f(p,d) which is identical to f(d,p).

2.6.2 Simple multiplication approach

The **simple multiplication approach** calculates the expected cell frequencies based on the scenario using the multiplication theorem for probabilities.

Using the notation defined in 2.6.1 we can calculate the expected value for each cell frequency as shown in the following table.

Tab. 4: Calculation of expected cell frequencies

| | | | | | | | | | | |
|-------------|---|---|---|-------------------|---|-------------------|---|-------------------|---|----------|
| $n(s+p+d+)$ | = | N | * | $p(s+/p-,d-)$ | * | $p(p+/s-,d-)$ | * | $p(d+/s-,p-)$ | * | $f(s,p)$ |
| $n(s+p+d-)$ | = | N | * | $p(s+/p-,d-)$ | * | $p(p+/s-,d-)$ | * | $(1-p(d+/s-,p-))$ | * | $f(s,p)$ |
| $n(s+p-d+)$ | = | N | * | $p(s+/p-,d-)$ | * | $(1-p(p+/s-,d-))$ | * | $p(d+/s-,p-)$ | * | $f(s,d)$ |
| $n(s+p-d-)$ | = | N | * | $p(s+/p-,d-)$ | * | $(1-p(p+/s-,d-))$ | * | $(1-p(d+/s-,p-))$ | | |
| $n(s-p+d+)$ | = | N | * | $(1-p(s+/p-,d-))$ | * | $p(p+/s-,d-)$ | * | $p(d+/s-,p-)$ | * | $f(p,d)$ |
| $n(s-p+d-)$ | = | N | * | $(1-p(s+/p-,d-))$ | * | $p(p+/s-,d-)$ | * | $(1-p(d+/s-,p-))$ | | |
| $n(s-p-d+)$ | = | N | * | $(1-p(s+/p-,d-))$ | * | $(1-p(p+/s-,d-))$ | * | $p(d+/s-,p-)$ | | |
| $n(s-p-d-)$ | = | N | * | $(1-p(s+/p-,d-))$ | * | $(1-p(p+/s-,d-))$ | * | $(1-p(d+/s-,p-))$ | | |

In the case of a scenario based on several homogeneous subsamples that cannot be distinguished from each other, we calculate the expected frequencies for each subsample and aggregate the contingency tables to one total table.

The simple multiplication approach has two disadvantages:

- It does not allow us to consider the impact of open populations, i.e. a certain percentage of cases dropping out of the population and being replaced (violation of the closed population assumption) in case of simultaneous capturing processes, and
- it does not allow us to introduce the direction of relationship in case of causal relationship between sampling processes (violation of the assumption "**no causal relationship between sampling processes**"). If, for example, we look at the relationship between "opiate related police charges" (P) and "drug related deaths" (D) we instantly understand that this indifference of direction constitutes a major problem. Since no case can be charged by the police after "drug related death" the factor f(d→p) has to be zero, while the probability of "drug related death" after being charged by the police may even be elevated, which may be expressed by factor f(p→d) being more than 1.

2.6.3 Approximate matrix approach

In order to be more flexible in designing possible scenarios we also used a more complicated **approximate matrix approach** (see also chapter 8.1), allowing us to calculate a greater variety of possible scenarios.

In this report we restricted ourselves to scenarios based on three capturing processes suitable for a 3-sample-capture-recapture approach. We divided the total observation period into 1,000 time segments of equal length, to approximate the continuous process by 1,000 discrete steps, and calculated the capturing and dropout probabilities per time segments corresponding with the capturing and dropout probabilities defined in the scenarios. An initial vector v_0 had 16 components, representing the expected cell frequencies for all cases still in the population (8 possible combinations given 3 capturing processes) and for those not in the population any more (8 possible combinations given 3 capturing processes as well). The initial vector v_0 consisted of 15 zeros and one cell equal to the total population size N . This represents the fact that initially no case has already been captured by any of the capturing processes or dropped out of the population. Three transition matrixes represent the transition probabilities for the three capturing processes and a fourth matrix represents the transition probabilities out of the population. All three matrixes were applied for each time segment – i.e. 1,000 times – to arrive at the final vector v_{1000} . Finally the expected frequencies per cell representing those cases still in the population at the end of the observation period and those representing the cases dropped out during the observation period were aggregated to arrive at the expected cell frequencies at the end of the observation period.

Based on the approximate matrix algorithm that can be performed using SPSS, it is now easily possible to speculate on the nature and extent of violations to the basic assumptions and to study how the specific set of violations impacts on the resulting estimates.

2.7 Statistical packages used

Simple calculations were done in Microsoft Excel, complicated calculations were done with the Computer Package "SPSS for Windows, Release 6.1.3". The calculations to apply the transition matrixes repeatedly to the initial vector (2.6) were done with the SPSS procedure "MATRIX". Data were aggregated and prepared using the SPSS procedure "AGGREGATE". The main calculations were done using the SPSS procedure "HILOGLINEAR" and "GENLOG".

95% confidence intervals for the estimates in 2-sample-capture-recapture approaches were calculated based on the formula proposed by Bishop et al. (1984, p.233-235) using EXCEL. The confidence intervals for the 3-sample- and 4-sample-capture-recapture estimates were derived from the SPSS procedure "GENLOG".

2.8 Model Selection

There exist many different ways to select the best adjusted model. They can be divided into mechanistic significance oriented approaches and approaches based on theoretical considerations. In the main analysis of this report we used two of the former approaches and one theoretical approach. All three are described in connection with the first main analysis in section 4.2.

The essential decision "best model" chosen by statistical criteria vs. "best model" based on substantial considerations is dealt with in chapter 5.2.

3 The empirical project

3.1 Criteria of data used in the study

As already stated in chapter 1.3, we base this study on "substitution treatment data" ("**S**"), "opiate related police charges" ("**P**"), "drug related deaths data" ("**D**") and "opiate related emergency ambulance data" ("**A**").

3.2 Identifiers

All person-related information in the data was deleted but unique identifiers were introduced to allow matching of identical cases as required to perform capture-recapture calculations.

The set of unique identifiers consisted of

- one letter of the first name
- three letters of the second name
- gender
- date of birth
- province of residence

Additionally we had access to a

- unique personal code within each database to discriminate between different persons with identical identifiers.

3.3 Relevant observation periods

The last complete set of "opiate related police charge data" we received for the project was from August 1995 (see chapter 1.3.3). We had "opiate related emergency ambulance data" only for the one-year-period from September 1994 through August 1995.

Because of these limitations we defined the interval from September 1994 through August 1995 as the observation period. The primary observation period between September 1994 and August 1995, will be referred to as "**t95**" in this study.

"Drug related deaths data" and "substitution treatment data" were available up to more recent dates. For additional calculations we used a sample of "drug related deaths data" for the following time period between September 1995 and August 1996, referred to as "**t96**".

3.4 Data selection criteria

Tab. 5 shows in detail how we arrived from the number of events with complete identifiers at the final sample size used for calculations. Even though we had a very good set of identifiers we found that 11 out of 2,502 relevant unique cases had identical identifiers. This could be identified based on the unique personal code within each database.

The numbers in Tab. 5 differ somewhat from officially published numbers by Haas et al. (1999). Since our observation period was from September till August, while official statistics focus on January through December, we could not compare our numbers directly with the official numbers. To arrive at comparable numbers we had to transform the official statistics into reference numbers.

We calculated the official numbers for one year times $1/3$ plus the official numbers for the following year times $2/3$ in the case of "opiate related police charges" and "drug related deaths data", and we interpolated the value for the last day of August (end of the observation period) concerning "substitution treatment data". "Substitution treatment statistics" represent the number of patients in substitution treatment at the end of each year.

The reference number based on official statistics for the relevant observation periods are given in brackets under the number used for our analyses in Tab. 5. We observe a deviation from the reference numbers representing official statistics.

The main reason for these discrepancies is that we used only

- events involving Austrian residents
- with a full set of identifiers (one letter of the first name, three letters of the second name, gender, date of birth, province of residence)
- between 15 years and 54 years of age.

The number of persons in substitution treatment exceeds the number of published cases. This is probably due to late reporting and late data entry after the official statistics were already processed.

Tab. 5: Description of the data used

| | Substitution "t95" | Police "t95" | Drug Rela- ted Deaths "t95" | Drug Rela- ted Deaths "t96" | Ambulance Vienna "t95" | total sample |
|--|-------------------------------|-------------------------------|-----------------------------------|-----------------------------------|------------------------------|-----------------|
| time span from / to | Sept. 1994 Aug. 1995 | Sept. 1994 Aug. 1995 | Sept. 1994 Aug. 1995 | Sept. 1995 Aug. 1996 | Sept. 1994 Aug. 1995 | |
| relevant events | 2,627 (2,470) ² | 3,810 (4,389) ³ | 202 (244) ⁴ | 186 (234) ⁵ | 803 | 7,628 |
| number of multiple events per case | 0 | 1,094 | 0 | 0 | 236 | 1,330 |
| number of different cases | 2,627 | 2,716 | 202 | 186 | 567 | 6,298 |
| number of identical identifiers referring to different cases | 4 | 3 | 0 | 0 | 0 | 7 |
| out of age range: too young | 0 | 11 | 3 | 2 | 5 | 21 |
| out of age range: too old | 1 | 4 | 0 | 0 | 5 | 10 |
| remaining cases for calculation | 2,622 | 2,698 | 199 | 184 | 557 | 6,260 |

The four different data sources vary quite a bit in the distribution of gender and age. The rate of females is highest among patients in substitution treatment and lowest among drug related deaths. The oldest sample is the one of substitution patients and the youngest cases are those charged by the police and emergency ambulance clients (Tab. 6). This shows very clearly that the assumption that all samples were somehow drawn randomly from the same population is by no means fully correct.

2 $2,264 + (2,573 - 2,264) * 2/3 = 2,470$

3 $4,394 * 1/3 + 4,386 * 2/3 = 4,389$

4 $250 * 1/3 + 241 * 2/3 = 244$

5 $241 * 1/3 + 230 * 2/3 = 234$

Tab. 6: Description of the data set used for calculation – and the reference population

| | Substitution "t95" | Police "t95" | Drug Related Deaths "t95" | Drug Related Deaths "t96" | Ambulance Vienna "t95" | total number of events | 15-54 year olds in the population |
|---------------|--------------------|----------------|---------------------------|---------------------------|------------------------|------------------------|-----------------------------------|
| sample size | 2,622 | 2,698 | 199 | 184 | 557 | 6,260 | 4,608,295 |
| male | 68.9% 1,806 | 79.5% 2,146 | 86.4% 172 | 83.7% 154 | 73.8% 411 | 74.9% 4,689 | 2,346,839 |
| female | 31.1% 816 | 20.5% 552 | 13.6% 27 | 16.3% 30 | 26.2% 146 | 25.1% 1,571 | 2,261,456 |
| Burgenland | 0.4% 10 | 1.1% 30 | 1.0% 2 | 0.5% 1 | | 0.7% 43 | 150,180 |
| Carinthia | 1.8% 46 | 1.6% 43 | 2.5% 5 | 0.0% 0 | | 1.5% 94 | 316,533 |
| Lower Austria | 7.0% 184 | 8.7% 234 | 4.0% 8 | 4.9% 9 | | 6.9% 435 | 841,527 |
| Upper Austria | 7.7% 202 | 6.2% 167 | 10.6% 21 | 6.5% 12 | | 6.4% 402 | 791,522 |
| Salzburg | 2.9% 77 | 2.0% 54 | 1.0% 2 | 2.7% 5 | | 2.2% 138 | 298,893 |
| Styria | 2.0% 53 | 3.2% 86 | 1.5% 3 | 6.0% 11 | | 2.4% 153 | 683,139 |
| Tyrol | 7.1% 187 | 8.0% 215 | 4.5% 9 | 6.5% 12 | | 6.8% 423 | 386,414 |
| Vorarlberg | 8.0% 210 | 10.3% 277 | 11.1% 22 | 10.3% 19 | | 8.4% 528 | 202,260 |
| Vienna | 63.0% 1653 | 59.0% 1592 | 63.8% 127 | 62.5% 115 | 100.0% 557 | 64.6% 4044 | 937,827 |
| 15-24 years | 18.8% 494 | 47.9% 1292 | 37.2% 74 | 31.5% 58 | 60.0% 334 | 36.0% 2252 | 1,018,905 |
| 25-34 years | 54.2% 1420 | 41.1% 1108 | 42.7% 85 | 48.9% 90 | 30.3% 169 | 45.9% 2872 | 1,421,330 |
| 35-54 years | 27.0% 708 | 11.0% 298 | 20.1% 40 | 19.6% 36 | 9.7% 54 | 18.1% 1136 | 2,168,060 |
| average age | 30.42 | 25.79 | 27.87 | 28.15 | 24.1 | 27.71 | – |

3.5 Case definition and: are the assumptions required by the capture-recapture approach plausible for the data set?

In the present study the case definition is "**problematic opiate consumption**". The case definition "problematic opiate consumption" does not include experimental use of opiates, occasional use and/or illicit trafficking without personal abuse. Since the addictive potential of opiates is high, we can assume that the case definition "problematic opiate consumption" in the sense just defined is quite close to "opiate addiction" in extend. The topic "case definition" is discussed in more detail in section 5.3. We will now try to analyse how the case definition corresponds with the implicit definition relevant for the data we rely on.

3.5.1 Substitution treatment

Patients in substitution therapy in Austria are primarily opiate addicts who have been addicted to opiates for longer period of time and who have failed to be successful in abstinence oriented forms of treatment.

We therefore can be quite sure that practically all cases in the register are cases according to the above case definition. At the same time we have to be aware though that the sample does not include cohorts of opiate addicts who reject this form of treatment and that we dramatically under-sample younger addicts with short careers. This violates the central homogeneity assumption in terms of equal catchability (see chapter 2.1 and 5.1).

3.5.2 Drug related deaths data and opiate related emergency ambulance data

Most persons who die of drug related illnesses prematurely, who severely overdose opiates involuntary and/or overdose on purpose (suicidal) very likely have a serious opiate problem. One cannot rule out that some experimenters or occasional users overdose because of a misfortune or because of a depressive episode, but this does not cause a severe quantitative problem in terms of our case definition.

We have to be aware, however, that some addicts have a very high risk to overdose while others have almost no risk to overdose, and that the remaining addicts are located somewhere between these extremes. This again violates the central homogeneity assumption in terms of equal catchability (see chapter 2.1 and 5.1).

3.5.3 Opiate related police charges

The likelihood to be charged by the police in relation to opiates is naturally greatly elevated for addicts who need opiates daily compared to experimenters and occasional users but the police naturally identify and charge experimenters, occasional opiate users and non-opiate-dependent traffickers as well. In other words, the police sample is a sample where we have to expect a relevant proportion of persons who are not cases according to the above case definition.

As in all the other samples, we have to expect quite a large degree of heterogeneity in this sample in terms of unequal catchability as well: Some opiate addicts manage to live a rather inconspicuous life, have very reliable sources for their supply and the chance that the police find out about them is comparatively small. Others look and behave in a way that makes their drug problem very obvious for others, who buy and sell drugs in the open drug scene – directly under the eyes of the police – and therefore are likely to be charged by the police regularly.

3.5.4 No causal relationship between sampling processes

The four data samples we base our calculations on are not really causally independent (see chapter 2.1 and 5.1). Addicts who overdose and are brought to hospital with the ambulance have a much higher risk of dying a drug related death in future events. Persons who die a drug related death cannot enter substitution treatment or overdose afterwards. Conspicuous persons who are much more likely to overdose dramatically and/or to die a drug related death prematurely run a much higher chance of being charged by the police as well. Persons in substitution programmes are less likely to overdose and to die a drug related death than before but on average are in worse physical shape and therefore more likely to die a drug related death than other addicts.

3.5.5 Conclusions in terms of case definition and assumptions

To sum up: Three out of four samples are not bad in terms of exclusively including cases according to the case definition, but they are quite heterogeneous in terms of catchability and they are by no means causally independent from each other. All four samples do not represent closed populations either: In a one year observation period many new events of problematic opiate use may occur in a region, some addicts quit their habit or die and there is fluctuation caused by migration into the region or out of it. Only the assumption of perfect identification of matching cases seem to be fulfilled quite well if the quality of assessing and entering data is good and if the identifiers used are sufficient, as in the case of this study. The implications of these violations as well as methods to compensate for them will be discussed in detail in chapter 5.

4 Problematic opiate consumption in Austria 1995 based on the capture-recapture approach

4.1 Previous estimations of problematic opiate consumption in Austria

In the seventies in a public discussion a popular Viennese drug therapist and a senior police officer were asked to estimate the number of opiate addicts based on their experience. Their first crude estimate was 5,000 opiate addicts in Vienna and 10,000 all over Austria. Since there were no better founded estimates available these numbers were readily accepted by the audience and soon made their way into scientific and semi-scientific publications. From then on scientists could cite respectable scientific publications to justify the formulation of prevalence rates. The estimates received more and more credibility and scientific status. The fact that hardly anybody could trace the foundation and justification of the numbers was no topic since there were no other competing estimates available.

As time went on and more and more new generations of opiates abusers and addicts added to the opiate scene, without all the older ones quitting or dying. It became apparent that prevalence rates for the seventies could not be correct for the nineties as well. Based on the initial estimates and the crude assumption that the number should have around doubled in the last two decades a number amounting to 10,000 opiate addicts in Vienna and to 20,000 in Austria was proposed.

Some experts decided to stick to the initial estimate and neglected the fact that the number of opiate addicts could not have stayed stable over two decades. Others accepted the new estimate. Still others favoured a compromise rate somewhere between the initial and the new estimates. Some rule-of-thumb validation attempts, relating the estimates to the number of drug related deaths, to the number of persons undergoing substitution treatment, to estimates from comparable other nations, etc., supported the idea the true prevalence rate nowadays should be somewhere between the initial estimates and the more recent estimates.

When Seidler and Uhl (1997) attempted a first statistically sophisticated estimation approach for Vienna based on capture-recapture methodology, they estimated 6,747 opiate addicts with a confidence interval ranging from 4,332 to 11,668 persons. This number was exactly between the initial estimate of 5,000 and the recent estimate of 10,000 – and the confidence interval even included the initial and the new estimate as well. In other words, the numbers of Seidler & Uhl were in line with most experts' expectations and therefore did not arouse much criticism.

4.2 Results of capture-recapture calculations based on the present study

The calculations in this section are based on the data samples "substitution treatment data" (s), "opiate related police charges" (p), "drug related deaths data" (d) and "opiate related emergency ambulance data (in Vienna only)". The observation period is September 1994 through August 1995 "t95". The following analyses are based on 3-sample-capture-recapture calculations wherever possible and appropriate. Where the data did not allow 3-sample-capture-recapture calculations due to empty cells in the contingency tables (i.e. the eight Austrian provinces except Vienna) 2-sample-capture-recapture calculations based on "substitution treatment data" and "opiate related police charges" were performed. We performed the analysis as usually done, i.e. we chose the "best model" mechanistically based on certain statistical criteria and not on theoretical considerations. The specific details are explained following Tab. 7 based on a concrete example. A further critical analysis of this approach can be found in chapter 5 ("Theoretical considerations").

4.2.1 Non-stratified calculation for all Austrians aged 15-54 years

If the three samples "drug related deaths data", "opiate related police charges" and "substitution treatment data" were used nation-wide the observed sample size amounted to 5042 known cases (Tab. 7).

Tab. 7: Contingency table for all Austrians aged 15-54 years

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 4,608,295 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 4 | 40 | 28 | 127 |
| absent | 401 | 2,253 | 2,189 | N _{hid} = ? |

The SPSS algorithm "HILOGLINEAR" suggested the "best model" to have one interaction term and that the most appropriate term is "s*p". To consider the best model according to the HILOGLINEAR algorithm is suggested by the Methodological Guidelines to Estimate the Prevalence of Problem Drug Use on the Local Level (Hay et al., 1998).

Using the Pearson's Chi-Square statistics calculated for all possible models – another mechanistic significance oriented approach – we arrived at the conclusion that the best model has one interaction term and that the best adjusted model is the one with "p*d". The basis of this approach is to group the models according to the number of interaction terms, to select the group with the lowest number of interaction terms containing non-significant p-values in order to determine the number of necessary terms and then to select the model with the highest p-value within this group of models.

Based on theoretical calculations a third approach makes even more sense. Due to the fact, that persons who die in the observation period cannot be arrested or treated afterwards, "opiate related police charges" and "substitution" have to be statistically dependent on "drug related deaths". This implies the interaction terms "p*d" and "s*d", even if they turn out statistically independent in the first two mechanistic approaches (compare section 5.7.2). The estimates according to the third approach can be found in Tab. 8 as well.

The syntax and output of the SPSS-calculations for the 3-sample-capture-recapture analyses documented in Tab. 8 can be found in section 8.3, the corresponding simple calculations for the 2-sample-capture-recapture analyses documented in Tab. 8 can be found in section 8.2.

We decided to base the final model selection on the Pearson's Chi-Square basis, but to document the best model according to the HILOGLINEAR algorithm and on theoretical considerations as well. In most cases the first and the second approach yield identical results. All possible models and their statistics are presented (Tab. 8).

Based on this "best model" a 3-sample-capture-recapture analysis estimated the total population of "problematic opiate users" for Austria in the age group between 15 and 54 years of age to be 17,276 cases. This estimate consists of the 5,042 known cases plus 12,234 hidden cases. Since there are 4,608,295 Austrians in the age group between 15 and 54 years of age (Tab. 7), a number of 17,276 "problematic opiate users" is equivalent to a prevalence rate of 0.37%. The likelihood-based asymmetric confidence interval for the preferred model ranges from 15,980 to 18,672 cases (Tab. 8).

6 see Tab. 6

Tab. 8: Analysis for all Austrians aged 15-54 years

| | 3-sample-capture-recapture approach | | | | | | | |
|---------------------------------|---------------------------------------|----------|--------------|--------------|---------------|---------------|------------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence | 8.67 | 3 | 0.034 | 5,042 | 11,834 | 16,876 | 15,657 – 18,137 | 0.37% |
| + s*p ⁷ | 2.51 | 2 | 0.286 | 5,042 | 8,542 | 13,584 | 11,416 – 16,426 | 0.29% |
| + s*d | 8.68 | 2 | 0.013 | 5,042 | 11,868 | 16,910 | 15,657 – 18,269 | 0.37% |
| + p*d ⁸ | 2.28 | 2 | 0.320 | 5,042 | 12,234 | 17,276 | 15,980 – 18,672 | 0.37% |
| + s*p + s*d | 0.22 | 1 | 0.643 | 5,042 | 7,153 | 12,195 | 10,056 – 15,241 | 0.26% |
| + s*p + p*d | 1.23 | 1 | 0.268 | 5,042 | 9,929 | 14,971 | 11,610 – 19,955 | 0.32% |
| + s*d + p*d ⁹ | 2.05 | 1 | 0.153 | 5,042 | 12,299 | 17,341 | 15,980 – 18,809 | 0.38% |
| saturated | 0.00 | 0 | 1.000 | 5,042 | 6,108 | 11,150 | 7,164 – 22,719 | 0.24% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | P | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 4,915 | 12,552 | 17,467 | 16,025 – 18,909 | 0.38% |
| s + d | 0.00 | 0 | 1.000 | 2,789 | 13,517 | 16,306 | 11,162 – 21,450 | 0.35% |
| p + d | 0.00 | 0 | 1.000 | 2,853 | 9,349 | 12,202 | 9,046 – 15,358 | 0.26% |

Based on a number of estimated 17,276 "problematic opiate users", we can quantify the correlation between the three conditions "substitution treatment", "opiate related police charges" and "drug related deaths" based on odds-ratios (see chapter 2.3). An odds-ratio of 1.54 between "opiate related police charges" and "drug related deaths" implies that the chance to die a drug related death in the observation period is 1.54 time higher for "persons charged by the police" than for others (Tab. 9).

Participation in "substitution treatment" is not significantly related to "opiate related police charges" (odds-ratio = 0.98) and there is also no significant relationship between "participation in substitution treatment" and "drug related deaths" (odds-ratio = 1.07) either (Tab. 9).

Tab. 9: Odds-ratios between the samples (all Austrians aged 15-54 years, N=17,276)

| | | |
|----------------------|----------------|----------------------------|
| substitution t95 (s) | | |
| 0.98 | police t95 (p) | |
| 1.07 | 1.54 | drug related death t95 (d) |

The fact that the 2-sample-capture-recapture approach based on "opiate related police charges" and "drug related deaths" yielded a much smaller estimate is perfectly in line with the fact that these two variables are positively correlated (Tab. 8). Positive correlation (odds-ratios greater than 1) between sampling processes leads to an underestimation of the hidden population and negative correlations (odds-ratios less than 1) to overestimation. This bias is commonly claimed to be compensated for by the 3-sample-capture-recapture algorithm but this is only true under certain circumstances. We will deal with this very important issue in chapter 5.

Knowing that "drug related deaths" is causally linked to the other two conditions (see chapter 5.7.2), it makes sense to compare the estimate suggested by the SPSS algorithm to the one based on a model with the interaction terms "s*d + p*d". As can be seen in Tab. 8 the two estimates are almost identical.

7 Best model suggested by the HILOGLINEAR algorithm (compare section 4.2).

8 Best model based on the Pearson Chi-square criterion (compare section 4.2).

9 Since "d" is intrinsically related to "s" and to "p" the model with the interaction terms "s*d" and "p*d" should be preferred in on logical grounds even if the statistical algorithm suggests a different model (compare sections 4.2 and 5.7.2).

This is not unexpected though, since the calculations performed in section 5.7.2 suggest that models with only one of the two relevant interaction terms perform almost equally well.

The rate of "problematic opiate users" annually included in the databases "substitution treatment", "opiate related police charges" and "drug related deaths" are given in Tab. 10.

Tab. 10: The percentage of "cases" included annually in the databases

| | |
|---|-------|
| probability of cases being included in "substitution treatment" annually | 15.2% |
| probability of cases being included in "opiate related police charges" annually | 15.6% |
| probability of cases being included in "drug related deaths" annually | 1.2% |

As discussed in chapter 2.1 homogeneity in terms of equal catchability is a vital assumption for the capture-recapture approach. The only way to eliminate heterogeneity is stratification by relevant variables. Important sources of heterogeneity are, for example, variables such as social integration, patterns of drug use, severity of psychological and physical problems, closeness to the open drug scene, duration of drug careers and many more. Unfortunately those central variables are not included in the databases we had access to for our capture-recapture calculations and since capture-recapture analysis is usually based on available data from registries, this is a common situation in the prevalence estimation based on this methodology. The variables we had access to are gender, age and region of residency. Consequently the only possible strategy for us was to concentrate on these less important variables for stratification and to hope that some relevant sources of heterogeneity are indirectly compensated for using these variables.

4.2.2 Stratification by gender

Tab. 11: Contingency table for male Austrians aged 15-54 years

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 2,346,839 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 3 | 36 | 24 | 109 |
| absent | 286 | 1,821 | 1,493 | N _{hid} = ? |

The analysis sample (observed population) included 3,772 males (Tab. 11) and 1,270 females (Tab. 13). Based on separate 3-sample-capture-recapture calculations the number of male opiate addicts was estimated to be 10,001 (Tab. 12) and the number of female cases to be 3,874 (Tab. 14). This implies that roughly 1/4 of problematic opiate users in Austria are female and 3/4 are male.

If we add the separate estimates for males and females (Tab. 12, Tab. 14), we arrive at 13,875 cases, which is 19.7% less than the non-stratified estimate of 17,276 cases (Tab. 8, Tab. 39). The discrepancy is largely caused by the fact that the "best model" suggested for the male sub-population included interaction between "substitution treatment" and "opiate related police charges", while the chosen "best model" for the overall analysis suggested interaction between "opiate related police charges" and "drug related deaths".

Tab. 12: Analysis for male Austrians aged 15-54 years

| | 3-sample-capture-recapture approach | | | | | | | |
|--------------------------------|---------------------------------------|----------|--------------|--------------|--------------|---------------|------------------------|--------------|
| | chisq | df | p | observed | estimated | Total | 95% ci | prevalence |
| independence | 8.28 | 3 | 0.041 | 3,772 | 9,022 | 12,794 | 11,715 – 13,971 | 0.55% |
| + s*p⁷⁺⁸ | 1.49 | 2 | 0.475 | 3,772 | 6,229 | 10,001 | 8,355 – 12,291 | 0.43% |
| + s*d | 8.00 | 2 | 0.018 | 3,772 | 9,111 | 12,883 | 11,794 – 14,073 | 0.55% |
| + p*d | 3.33 | 2 | 0.189 | 3,772 | 9,378 | 13,150 | 12,039 – 14,493 | 0.56% |
| + s*p + s*d | 0.49 | 1 | 0.485 | 3,772 | 5,514 | 9,286 | 7,562 – 11,794 | 0.40% |
| + s*p + p*d | 1.14 | 1 | 0.286 | 3,772 | 6,780 | 10,552 | 8,131 – 14,387 | 0.45% |
| + s*d + p*d⁹ | 2.44 | 1 | 0.118 | 3,772 | 9,506 | 13,278 | 12,122 – 14,601 | 0.57% |
| saturated | 0.00 | 0 | 1.000 | 3,772 | 4,070 | 7,842 | 5,008 – 17,132 | 0.33% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | Total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 3,663 | 9,748 | 13,411 | 12,093 – 14,729 | 0.57% |
| s + d | 0.00 | 0 | 1.000 | 1,951 | 9,554 | 11,505 | 7,550 – 15,460 | 0.49% |
| p + d | 0.00 | 0 | 1.000 | 2,279 | 7,185 | 9,464 | 6,876 – 12,052 | 0.40% |

Knowing that "drug related deaths" is causally linked to the other two conditions (see chapter 5.7.2), we can conclude on a theoretical basis that there have to be relevant interactions between "drug related deaths" and the other two conditions. If we decide to base the calculations on the model with the two interactions "s*d" and "p*d" and disregard the best models suggested by the SPSS algorithm the non-stratified estimate turns out almost identical: 17,341 instead of 17,276 and the gender stratified analysis is 17,162¹⁰ instead of 13,875¹¹. In other words, the non-stratified analysis and the stratified turn out

¹⁰ 13,278+3,884 = 17,162

practically identical if we decide to choose models different than those suggested based on statistical criteria (Tab. 39).

We may conclude that these two relevant interactions did not reach statistical significance due to random variation and/or an insufficient sample size (see chapter 5.2).

Tab. 13: Contingency table for female Austrians aged 15-54 years

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 2,261,456 | police t95 (p) | | | |
| | Present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | Absent | present | absent |
| present | 1 | 4 | 4 | 18 |
| absent | 115 | 432 | 696 | N _{hid} = ? |

Tab. 14: Analysis for Austrians aged 15-54 years – female

| | 3-sample-capture-recapture approach | | | | | | | |
|------------------------------------|---------------------------------------|----------|--------------|--------------|--------------|--------------|----------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence ⁷⁺⁸ | 0.52 | 3 | 0.915 | 1,270 | 2,604 | 3,874 | 3,371 – 4,499 | 0.17% |
| + s*p | 0.49 | 2 | 0.784 | 1,270 | 2,486 | 3,756 | 2,389 – 6,811 | 0.17% |
| + s*d | 0.41 | 2 | 0.814 | 1,270 | 2,585 | 3,855 | 3,350 – 4,467 | 0.17% |
| + p*d | 0.11 | 2 | 0.949 | 1,270 | 2,633 | 3,903 | 3,392 – 4,532 | 0.17% |
| + s*p + s*d | 0.14 | 1 | 0.710 | 1,270 | 1,944 | 3,214 | 1,922 – 7,038 | 0.14% |
| + s*p + p*d | 0.00 | 1 | 0.955 | 1,270 | 3,132 | 4,402 | 2,324 – 10,591 | 0.19% |
| + s*d + p*d ⁹ | 0.01 | 1 | 0.925 | 1,270 | 2,614 | 3,884 | 3,371 – 4,532 | 0.17% |
| saturated | 0.00 | 0 | 1.000 | 1,270 | 3,574 | 4,844 | 1,694 – 31,301 | 0.21% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 1,252 | 2,631 | 3,883 | 3,301 – 4,465 | 0.17% |
| s + d | 0.00 | 0 | 1.000 | 838 | 3,568 | 4,406 | 930 – 7,882 | 0.19% |
| p + d | 0.00 | 0 | 1.000 | 574 | 2,407 | 2,981 | 633 – 5,329 | 0.13% |

11 10,001+3,874 = 13,875

4.2.3 Stratification by age groups

Another stratification was performed by age groups. Splitting the sample into four ten-year cohorts resulted in too few cases in the oldest age cohort. We therefore had to collapse the last two age groups into a twenty-year cohort from 35 to 54 years of age.

Tab. 15: Contingency table for Austrians aged 15-24 years

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 1,018,905 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 2 | 22 | 2 | 48 |
| absent | 163 | 1,105 | 327 | N _{hid} = ? |

Tab. 16: Analysis for Austrians aged 15-24 years

| | 3-sample-capture-recapture approach | | | | | | | |
|------------------------------------|---------------------------------------|----------|--------------|--------------|--------------|--------------|----------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence ⁷⁺⁸ | 3.85 | 3 | 0.279 | 1,669 | 2,302 | 3,971 | 3,589 – 4,421 | 0.39% |
| + s*p | 3.09 | 2 | 0.213 | 1,669 | 2,945 | 4,614 | 3,495 – 6,439 | 0.45% |
| + s*d | 0.61 | 2 | 0.738 | 1,669 | 2,228 | 3,897 | 3,514 – 4,339 | 0.38% |
| + p*d | 3.84 | 2 | 0.147 | 1,669 | 2,301 | 3,970 | 3,570 – 4,476 | 0.39% |
| + s*p + s*d | 0.50 | 1 | 0.480 | 1,669 | 2,411 | 4,080 | 3,120 – 5,693 | 0.40% |
| + s*p + p*d | 0.43 | 1 | 0.511 | 1,669 | 7,848 | 9,517 | 3,570 – 34,202 | 0.93% |
| + s*d + p*d ⁹ | 0.60 | 1 | 0.440 | 1,669 | 2,217 | 3,886 | 3,495 – 4,366 | 0.38% |
| saturated | 0.00 | 0 | 1.000 | 1,669 | 4,773 | 6,442 | 2,434 – 31,402 | 0.63% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 1,621 | 2,247 | 3,868 | 3,418 – 4,318 | 0.38% |
| s + d | 0.00 | 0 | 1.000 | 564 | 8,575 | 9,139 | 464 – 17,814 | 0.90% |
| p + d | 0.00 | 0 | 1.000 | 1,342 | 2,642 | 3,984 | 2,686 – 5,282 | 0.39% |

Tab. 17: Contingency table for Austrians aged 25-34 years

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 1,421,330 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 2 | 13 | 18 | 52 |
| absent | 180 | 913 | 1,220 | N _{hid} = ? |

Tab. 18: Analysis for Austrians aged 25-34 years

| | 3-sample-capture-recapture approach | | | | | | | |
|------------------------------------|---------------------------------------|----------|--------------|--------------|--------------|--------------|----------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence ⁷⁺⁸ | 5.33 | 3 | 0.149 | 2,398 | 5,844 | 8,242 | 7,412 – 9,234 | 0.58% |
| + s*p | 0.14 | 2 | 0.930 | 2,398 | 3,645 | 6,043 | 4,743 – 8,051 | 0.43% |
| + s*d | 2.95 | 2 | 0.229 | 2,398 | 6,040 | 8,438 | 7,513 – 9,513 | 0.59% |
| + p*d | 4.13 | 2 | 0.127 | 2,398 | 5,966 | 8,364 | 7,462 – 9,442 | 0.59% |
| + s*p + s*d | 0.14 | 1 | 0.704 | 2,398 | 3,652 | 6,050 | 4,376 – 9,099 | 0.43% |
| + s*p + p*d | 0.11 | 1 | 0.745 | 2,398 | 3,525 | 5,923 | 4,457 – 8,461 | 0.42% |
| + s*d + p*d ⁹ | 1.05 | 1 | 0.305 | 2,398 | 6,188 | 8,586 | 7,617 – 9,730 | 0.60% |
| saturated | 0.00 | 0 | 1.000 | 2,398 | 3,246 | 5,644 | 3,155 – 16,303 | 0.40% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 2,346 | 6,299 | 8,645 | 7,573 – 9,717 | 0.61% |
| s + d | 0.00 | 0 | 1.000 | 1,485 | 4,550 | 6,035 | 3,738 – 8,332 | 0.42% |
| p + d | 0.00 | 0 | 1.000 | 1,178 | 5,101 | 6,279 | 3,415 – 9,143 | 0.44% |

Tab. 19: Contingency table for Austrians aged 35-54 years

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 2,168,060 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 5 | 8 | 27 |
| absent | 58 | 235 | 642 | N _{hid} = ? |

Tab. 20: Analysis for Austrians aged 35-54 years

| | 3-sample-capture-recapture approach | | | | | | | |
|------------------------------------|---------------------------------------|----------|--------------|------------|--------------|--------------|----------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence ⁷⁺⁸ | 2.87 | 3 | 0.412 | 975 | 2,530 | 3,505 | 2,895 – 4,303 | 0.16% |
| + s*p | 1.80 | 2 | 0.406 | 975 | 1,942 | 2,917 | 1,977 – 4,765 | 0.13% |
| + s*d | 2.87 | 2 | 0.238 | 975 | 2,525 | 3,500 | 2,876 – 4,336 | 0.16% |
| + p*d | 1.44 | 2 | 0.487 | 975 | 2,597 | 3,572 | 2,934 – 4,404 | 0.16% |
| + s*p + s*d | 0.72 | 1 | 0.396 | 975 | 1,269 | 2,244 | 1,458 – 4,303 | 0.10% |
| + s*p + p*d | 1.23 | 1 | 0.268 | 975 | 2,167 | 3,142 | 1,957 – 5,745 | 0.14% |
| + s*d + p*d ⁹ | 1.43 | 1 | 0.232 | 975 | 2,601 | 3,576 | 2,914 – 4,473 | 0.16% |
| saturated | 0.00 | 0 | 1.000 | 975 | 761 | 1,736 | 1,012 – 16,342 | 0.08% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 948 | 2,690 | 3,638 | 2,833 – 4,443 | 0.17% |
| s + d | 0.00 | 0 | 1.000 | 740 | 2,800 | 3,540 | 1,358 – 5,722 | 0.16% |
| p + d | 0.00 | 0 | 1.000 | 333 | 2,051 | 2,384 | 446 – 4,322 | 0.11% |

The highest prevalence rate of 0.58% was found in the cohort of the 25 to 34 year olds (Tab. 18). The next highest prevalence rate 0.39% was found in the 15 to 24 year olds (Tab. 16). The lowest prevalence rate of 0.16% was found in the oldest cohort of the 35 through 54 year olds (Tab. 20).

If we add up the individual estimates for the three age groups we arrive at 15,718¹² "problematic opiate users" in Austria, which is 9.0% less than the non-stratified estimate of 17,276 cases (Tab. 39). If we base the stratification on the model with the interactions "s*d" and "p*d", we yield a stratified estimate of 16,048¹³ which is somewhat closer to the non-stratified estimate of 17,341 based on this model.

¹² 3,971+8,242+3,505 = 15,718

¹³ 3,886+8,586+3,576 = 16,048

4.2.4 Stratification by the nine Austrian provinces

The estimates for eight of the nine Austrian provinces were based on an insufficient number for a 3-sample-capture-recapture approach and therefore we had to resort to 2-sample approaches based on "substitution treatment" and "opiate related police charges" (Tab. 21 –

Tab. 36). A 3-sample approach was only feasible for Vienna (

Tab. 38).

Tab. 21: Contingency table for Burgenland

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 150,180 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 0 | 1 | 1 |
| absent | 4 | 26 | 5 | N _{hid} = ? |

Tab. 22: Analysis for Burgenland

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|----------|------------|
| s + p | 0.00 | 0 | 1.000 | 36 | 39 | 75 | 22 – 128 | 0.05% |

Tab. 23: Contingency table for Carinthia

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 316,533 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 0 | 0 | 0 |
| absent | 15 | 28 | 31 | N _{hid} = ? |

Tab. 24: Analysis for Carinthia

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|----------|------------|
| s + p | 0.00 | 0 | 1.000 | 74 | 58 | 132 | 88 – 176 | 0.04% |

Tab. 25: Contingency table for Lower Austria

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 841,527 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 1 | 2 | 6 |
| absent | 29 | 204 | 153 | N _{hid} = ? |

Tab. 26: Analysis for Lower Austria

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|---------------|------------|
| s + p | 0.00 | 0 | 1.000 | 389 | 1,096 | 1,485 | 1,021 – 1,949 | 0.18% |

Tab. 27: Contingency table for Upper Austria

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 791,522 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 1 | 0 | 7 | 4 |
| absent | 17 | 149 | 177 | N _{hid} = ? |

Tab. 28: Analysis for Upper Austria

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|---------------|------------|
| s + p | 0.00 | 0 | 1.000 | 351 | 1,523 | 1,874 | 1,093 – 2,655 | 0.24% |

Tab. 29: Contingency table for Salzburg

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 298,893 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 2 | 3 | 0 |
| absent | 9 | 43 | 65 | N _{hid} = ? |

Tab. 30: Analysis for Salzburg

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|-----------|------------|
| s + p | 0.00 | 0 | 1.000 | 122 | 340 | 462 | 203 – 721 | 0.15% |

Tab. 31: Contingency table for Styria

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 683,139 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 1 | 0 | 10 |
| absent | 7 | 78 | 46 | N _{hid} = ? |

Tab. 32: Analysis for Styria

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|-------------|------------|
| s + p | 0.00 | 0 | 1.000 | 132 | 519 | 651 | 220 – 1,082 | 0.10% |

Tab. 33: Contingency table for Tyrol

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 386,414 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 0 | 1 | 1 | 10 |
| absent | 13 | 201 | 173 | N _{hid} = ? |

Tab. 34: Analysis for Tyrol

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|---------------|------------|
| s + p | 0.00 | 0 | 1.000 | 389 | 2,704 | 3,093 | 1,521 – 4,665 | 0.80% |

Tab. 35: Contingency table for Vorarlberg

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 202,260 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 2 | 2 | 3 | 12 |
| absent | 71 | 202 | 134 | N _{hid} = ? |

Tab. 36: Analysis for Vorarlberg

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|-------|-------|----|-------|----------|-----------|-------|-----------|------------|
| s + p | 0.00 | 0 | 1.000 | 414 | 383 | 797 | 670 – 924 | 0.39% |

Tab. 37: Contingency table for Vienna

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 937,827 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 1 | 33 | 12 | 81 |
| absent | 236 | 1,322 | 1,404 | N _{hid} = ? |

Tab. 38: Analysis for Vienna

| | 3-sample-capture-recapture approach | | | | | | | |
|--------------------------------|---------------------------------------|----------|--------------|--------------|--------------|----------------------------|----------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence | 22.19 | 3 | 0.000 | 3,089 | 7,555 | 10,644 | 9,657 – 11,780 | 1.13% |
| + s*p | 13.69 | 2 | 0.001 | 3,089 | 5,216 | 8,305 | 6,730 – 10,569 | 0.89% |
| + s*d | 18.03 | 2 | 0.000 | 3,089 | 7,330 | 10,419 | 9,463 – 11,523 | 1.11% |
| + p*d⁸ | 4.21 | 2 | 0.122 | 3,089 | 8,097 | 11,186 | 10,063 – 12,410 | 1.19% |
| + s*p + s*d⁷ | 0.47 | 1 | 0.493 | 3,089 | 3,245 | 6,334 | 5,254 – 7,955 | 0.68% |
| + s*p + p*d | 3.91 | 1 | 0.048 | 3,089 | 9,477 | 12,566 | 8,256 – 20,590 | 1.34% |
| + s*d + p*d⁹ | 2.69 | 1 | 0.101 | 3,089 | 7,864 | 10,953 | 9,857 – 12,225 | 1.17% |
| saturated | 0.00 | 0 | 1.000 | 3,089 | 2,293 | 5,382 | 3,488 – 16,184 | 0.57% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 3,008 | 8,096 | 11,104¹⁴ | 9,897 – 12,311 | 1.18% |
| s + d | 0.00 | 0 | 1.000 | 1,767 | 14,382 | 16,149 | 7,865 – 24,433 | 1.72% |
| p + d | 0.00 | 0 | 1.000 | 1,685 | 4,262 | 5,947 | 4,255 – 7,639 | 0.63% |

14 This number corresponds to the 2-sample capture-recapture analyses performed for the other 8 Austrian provinces.

If we add up the individual estimates for the nine Austrian provinces based on 2-sample-capture-recapture analyses we arrive at 18,871¹⁵ "problematic opiate users" in Austria, which is 9.2% more than the non-stratified estimate of 17,276 cases (Tab. 39).

If we add up the individual estimates for the eight Austrian provinces based on 2-sample-capture-recapture analyses and the estimate for Vienna, where such an analysis is possible, on the 3-sample capture-recapture model suggested by SPSS we arrive at 14,101¹⁶ "problematic opiate users" in all of Austria, which is 18.4% less than the non-stratified estimate of 17,276 cases (Tab. 39).

Knowing that "drug related deaths" is causally linked to the other two conditions (see chapter 5.7.2), we can conclude on a theoretical basis that there have to be relevant interactions between "drug related deaths" and the other two conditions. In the 2-sample-capture-recapture analyses, where "drug related deaths" are not included, this fact does not matter, but this is very relevant in the 3-sample capture-recapture analysis for Vienna.

If we add up the individual estimates for the eight Austrian provinces based on two sample capture-recapture analyses and base the calculation for Vienna on the 3-sample model with the two interactions "s*d" and "p*d", disregarding the best models suggested by the SPSS algorithm, the stratified estimate results in 18,720 which is 8.0% more than the non-stratified analysis based on the model with the interaction terms "s*d" and "p*d" (Tab. 39).

It is quite obvious from these results that the 3-capture-recapture model suggested by the SPSS algorithm for Vienna has to be questioned seriously. Only the estimates based on 2-sample-capture-recapture approaches and the one, where the two interactions "s*d" and "p*d" are considered in a 3-sample-capture-recapture model for Vienna are credible.

15 $75+132+1,485+1,874+462+651+2,704+384+11,104 = 18,871$

16 $75+132+1,485+1,874+462+651+2,704+384+6,334 = 14,101$

4.2.5 The impact of stratification

Tab. 39: The impact of stratification on the estimation – summed up – based on the best model according to the chi-square criterion resp. on the model with s*d and p*d

| Type of capture-recapture models (crms) | stratification by | non-stratified analysis for Austria | stratified analysis for Austria | absolute discrepancy | relative discrepancy |
|--|-------------------|-------------------------------------|---------------------------------|----------------------|----------------------|
| 3 sample crms suggested by SPSS | gender | 17,276 ¹⁷ | 13,875 | -3,402 | -19.7% |
| | age | 17,276 ¹⁷ | 15,718 | -1,558 | -9.0% |
| 2 samples crms only | province | 17,276 ¹⁷ | 18,871 | 1,595 | 9.2% |
| 2 samples crms for 8 provinces 3 sample crms suggested by chi-square criterion for Vienna | province | 17,276 ¹⁷ | 14,101 | -3,175 | -18.4% |
| 3 sample crms with interactions "s*d" and "p*d" | gender | 17,341 ¹⁸ | 17,162 | -179 | -1.0% |
| | age | 17,341 ¹⁸ | 16,048 | -1,293 | -7.5% |
| 2 samples crms for 8 provinces 3 sample crm with interactions "s*d" and "p*d" for Vienna | province | 17,341 ¹⁸ | 18,720 | 1,379 | 8.0% |

Stratification by gender, by age and by province resulted in a variety of estimates. The estimates ranged from 13,875 through 18,871, which is equivalent to 19.7% less and 9.2% more than the non-stratified estimate. If we base our estimates on the model with the interactions "s*d" and "p*d" all stratified estimates are much closer to the non-stratified estimate (Tab. 39).

4.3 Additional estimations to judge validity and reliability of the estimation

In chapter 4.1 we mentioned that Seidler and Uhl (1997) estimated the number of "problematic opiate users" in Vienna to be 6,747 on the basis of data from the same observation period (September 1994 through December 1994) based on a 3-sample-capture-recapture approach. One would expect rather similar estimates on related data for the same time period (Tab. 41 and Tab. 44).

Tab. 40: Contingency table for Vienna based on Seidler & Uhl (1997)

| | | | | |
|---|---------------------------|--------|---------------------------|----------------------|
| population size ⁶ 937,827 | police 94 (p) | | | |
| | present | | absent | |
| | ambulance/hospital 94 (a) | | ambulance/hospital 94 (a) | |
| drug related death 95 (d) | present | absent | present | absent |
| present | 5 | 13 | 11 | 13 |
| absent | 49 | 658 | 179 | N _{hid} = ? |

Source: Seidler & Uhl (1997)

17 Estimate based on the "best model" according to the chi-square criterion (Tab. 8)

18 Estimate based on the model with the interactions "s*d" and "p*d" (Tab. 8)

Tab. 41: Analysis for Vienna based on Seidler & Uhl (1997)

| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
|----------|-------|----|-------|----------|-----------|-------|--------------|------------|
| +a*d+a*p | 0.76 | 1 | 0.382 | 928 | 5,819 | 6,747 | 4,332-11,668 | 0.91% |

Source: Seidler & Uhl (1997)

An estimated number of 6,747 [4,332; 11,668] "problematic opiate users" in Vienna is almost identical to the 6,334 [5,254; 7,955] based on the 3-sample-capture-recapture model suggested by SPSS in the present study (Tab. 38). If we consider details we realize that the fact that "drug related deaths" is causally linked to the other two conditions (see chapter 5.7.2) is compensated in the first estimate but not in the second one. This seemingly nice correspondence thus has to be put into perspective.

Seidler and Uhl (1997) adjusted for the nature of drug related deaths by using "drug related deaths" from the consecutive year. In the present study one way to compensate for the resulting bias would be to use the model based on the two interactions "s*d" and "p*d". This approach results in an estimate of 10,953 [9,897; 12,311] (Tab. 38). If we adjust like Seidler and Uhl (1997) did, by using "drug related deaths" from the consecutive year (Tab. 43) we arrive at 8,236 [6,614; 10,636] cases.

Tab. 42: Contingency table for all Viennese aged 15-54 years using a sample of drug related deaths from the consecutive year

| population size ⁶ 937,827 | police t95 (p) | | | |
|---|----------------------|--------|----------------------|----------------------|
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t96 (d) | present | absent | present | absent |
| present | 3 | 18 | 21 | 73 |
| absent | 234 | 1337 | 1394 | N _{hid} = ? |

Tab. 43: Analysis for all Viennese aged 15-54 years using a sample of drug related deaths from the consecutive year

| | 3-sample-capture-recapture approach | | | | | | | |
|----------------------------------|---------------------------------------|----------|--------------|--------------|--------------|--------------|-----------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence ⁸ | 4.72 | 3 | 0.193 | 3,081 | 7,587 | 10,668 | 9,715 – 11,772 | 1.14% |
| + s*p ⁷ | 0.15 | 2 | 0.927 | 3,081 | 5,155 | 8,236 | 6,614 – 10,636 | 0.88% |
| + s*d | 2.11 | 2 | 0.348 | 3,081 | 7,807 | 10,888 | 9,849 – 12,126 | 1.16% |
| + p*d | 3.81 | 2 | 0.149 | 3,081 | 7,721 | 10,802 | 9,782 – 11,947 | 1.15% |
| + s*p + s*d | 0.07 | 1 | 0.796 | 3,081 | 5,422 | 8,503 | 6,310 – 12,217 | 0.91% |
| + s*p + p*d | 0.01 | 1 | 0.938 | 3,081 | 4,849 | 7,930 | 6,062 – 10,945 | 0.85% |
| + s*d + p*d | 0.67 | 1 | 0.412 | 3,081 | 7,970 | 11,051 | 9,986 – 12,309 | 1.18% |
| saturated | 0 | 0 | 1.000 | 3,081 | 5,148 | 8,229 | 4,561 – 20,935 | 0.88% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 3,007 | 8,090 | 11,097 | 9,891 – 12,303 | 1.18% |
| s + d | 0.00 | 0 | 1.000 | 1,743 | 6,173 | 7,916 | 5,119 – 10,712 | 0.84% |
| p + d | 0.00 | 0 | 1.000 | 1,686 | 7,032 | 8,718 | 5,369 – 12,067 | 0.93% |

In order to check the validity of our more recent estimate through a further calculation, we tried a 4-sample-capture-recapture approach for Vienna including "opiate related emergency ambulance data" additionally. Since a 4-sample-capture-recapture approach allows too many different models to treat them individually, we chose the one suggested by the HILOGLINEAR algorithm. SPSS suggested the

"best model" to be the one with the interaction terms $s*p*a + s*d + d*a$. Based on this model the total number of problematic opiate users for Vienna was estimated to be 7,393 [5,831; 9,918] (Tab. 45).

Tab. 44: Contingency table for Vienna including "opiate related emergency ambulance data"

| | | | | | | | | |
|---|----------------------|--------|----------------------|--------|----------------------|--------|----------------------|----------------------|
| population size ⁶ 937,827 | ambulance t95 (a) | | | | | | | |
| | present | | | | absent | | | |
| | police t95 (p) | | | | police t95 (p) | | | |
| | present | | absent | | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent | present | absent | present | absent |
| present | 1 | 13 | 1 | 15 | 0 | 20 | 11 | 66 |
| absent | 30 | 141 | 30 | 326 | 206 | 1,181 | 1,374 | N _{hid} = ? |

Tab. 45: Analysis for Vienna including "opiate related emergency ambulance data" additionally

| | | | | | |
|-----------------------|----------|-----------|-------|-------------|------------|
| | observed | estimated | total | 95% ci | prevalence |
| + $s*p*a + s*d + d*a$ | 3,415 | 3,978 | 7,393 | 5,831–9,918 | 0.79% |

Tab. 46: Odds-ratios between the samples (Vienna including "opiate related emergency ambulance data", N=7,393)

| | | | |
|----------------------|----------------|----------------------------|-------------------|
| substitution t95 (s) | | | |
| 0.54 | police t95 (p) | | |
| 0.39 | 1.34 | drug related death t95 (d) | |
| 0.41 | 1.92 | 3.95 | ambulance t95 (a) |

An analysis of the correlation matrix makes evident that "opiate related emergency ambulance transports" is highly correlated with "drug related deaths" (or = 3.95), while "substitution treatment" is hardly related to either of the other three conditions (Tab. 46)

Having four data samples available it makes sense to try still another combination of variables. If we suspect that the low rate of "drug related deaths data" might cause some estimation problems we can use "substitution treatment data", "opiate related police charges" and "opiate related emergency ambulance data" for a 3-sample approach. The results again deviated dramatically from the previous results. The algorithm suggested the "best model" to be the one with both interactions involving "opiate related police charges" (Tab. 48).

Tab. 47: Contingency table for Vienna including "opiate related emergency ambulance data" excluding "drug related deaths data"

| | | | | |
|---|----------------------|--------|----------------------|----------------------|
| population size ⁶ 937,827 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| ambulance t96 (a) | present | absent | present | absent |
| present | 31 | 154 | 31 | 341 |
| absent | 206 | 1,201 | 1,385 | N _{hid} = ? |

Tab. 48: Analysis for Vienna including "opiate related emergency ambulance data" excluding "drug related deaths data"

| | 3-sample-capture-recapture approach | | | | | | | |
|----------------------------------|---------------------------------------|----------|--------------|--------------|---------------|---------------|------------------------|--------------|
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| independence | 143.56 | 3 | 0.000 | 3,349 | 6,108 | 9,457 | 8,781 – 10,185 | 1.01% |
| + s*p | 100.90 | 2 | 0.000 | 3,349 | 4,408 | 7,757 | 7,064 – 8,620 | 0.83% |
| + s*a | 127.00 | 2 | 0.000 | 3,349 | 5,397 | 8,746 | 8,119 – 9,473 | 0.93% |
| + p*a | 11.61 | 2 | 0.003 | 3,349 | 8,765 | 12,114 | 10,980 – 13,446 | 1.29% |
| + s*p + s*a | 66.70 | 1 | 0.000 | 3,349 | 2,659 | 6,008 | 5,535 – 6,578 | 0.64% |
| + s*p + p*a⁷⁺⁸ | 0.58 | 1 | 0.447 | 3,349 | 15,236 | 18,585 | 13,858 – 25,375 | 1.98% |
| + s*a + p*a | 8.86 | 1 | 0.003 | 3,349 | 8,075 | 11,424 | 10,254 – 12,763 | 1.22% |
| saturated | 0 | 0 | 1.000 | 3,349 | 17,819 | 21,168 | 13,650 – 34,295 | 2.26% |
| | 2-sample-capture-recapture approaches | | | | | | | |
| | chisq | df | p | observed | estimated | total | 95% ci | prevalence |
| s + p | 0.00 | 0 | 1.000 | 3,008 | 8,096 | 11,104 | 9,897 – 12,311 | 1.18% |
| s + d | 0.00 | 0 | 1.000 | 2,148 | 12,702 | 14,850 | 11,432 – 18,269 | 1.58% |
| p + d | 0.00 | 0 | 1.000 | 1,964 | 2,829 | 4,793 | 4,263 – 5,324 | 0.51% |

4.3.1 Summary: Estimates for Vienna

- Seidler and Uhl (1997) estimated **6,747** "problematic opiate users" in Vienna based on "opiate related police charges" and "opiate related emergency ambulance data" and "drug related deaths" from the consecutive year (Tab. 41).
- Using "substitution treatment data" and "opiate related police charges" and "drug related deaths" from the same time period the model suggested by SPSS produced an estimate of **6,334** cases (Tab. 38). This estimate is highly questionable though, since the problems arising from using "drug related deaths" together with other variables were not adjusted for.
- If the peculiar nature of "drug related deaths" was accounted for by using the model based on the two interactions "s*d" and "p*d" we arrived at **10,953** cases (Tab. 38).
- A simple 2-sample-capture-recapture approach based on "substitution treatment data" and "opiate related police charges" yielded an estimate of **11,104** cases (Tab. 38).
- If "drug related deaths" were used from the consecutive year – as in Seidler & Uhl's 1997 estimate – in the present study the SPSS suggested model produced an estimate of **8,236** cases (Tab. 43).
- Based on a 4-sample-capture-recapture using the three mentioned variables and additionally including "opiate related emergency ambulance data" – all from the same time period – we arrived at **7,393** cases (Tab. 45).
- If we dropped drug related deaths from this model and used the remaining three samples for a 3-sample-capture-recapture approach we arrived at **18,585** cases (Tab. 48).

The contradicting results for Vienna demand good explanations and give reason to suspect that the capture-recapture approach in the field of illicit drug epidemiology is not as valid and reliable as some experts would like. Possible explanations for biases will be analysed and discussed systematically in chapter 5.

5 Theoretical considerations

5.1 Central aspects related to the capture-recapture approach

5.1.1 Advantages of the capture-recapture approach

Before starting to deal with central assumption behind the capture-recapture approach and analysing the possible impact of violated assumptions, we will once more recapitulate the basic advantages of the approach: The most striking advantage is that the expected estimates are independent of the intensity of the capturing processes. In other words, if, for example, we want to estimate the prevalence of opiate related problems using data sources such as "opiate related police charges", "treatment data", increased or decreased police activity, increased or decreased attractiveness of treatment facilities and the like do not cause systematic errors (biases). Only the precision of estimates is better if the capturing process is more intense. A higher number of known cases results in smaller confidence intervals (chapter 2.4).

The second striking advantage of the capture-recapture approach is that we can use data from existing registers ("opiate related police charges", "substitution treatment data", "drug related deaths data" etc.) and do not need additional empirical research results to assess the proportion between known and hidden cases.

5.1.2 Basic assumptions behind the capture-recapture approach

In chapter 2.1 we have already stated that the 2-sample-capture-recapture approach is based on the following 5 assumptions:

- **assumption 1: only cases in line with the case definition are considered**
- **assumption 2: perfect identification of matching cases**
- **assumption 3: closed population**
- **assumption 4: homogeneity in terms of equal catchability**
- **assumption 5: no causal relationship between sampling processes**

5.1.3 Other important aspects related to the capture-recapture approach

An important aspect in this context is the **case definition**; i.e. the question what phenomenon do we want to assess?

We also should reflect whether the **sequence of the sampling processes** has an impact on the estimates (see chapter 2.5). In the traditional application of the capture-recapture approach the capturing processes happened consecutively¹⁹ while in most social science applications the capturing processes happen simultaneously (e.g. police charges and treatment sequences are recorded in the same observation period).

If there is a **causal relationship between sampling processes**, which violates assumption 5, we can expect that this violation is compensated in higher order capture-recapture approaches through the inclusion of appropriate interaction terms.

We have to be aware though, that we cannot possibly include all interactions possible. Due to the fact that we do not have access to the cell including the hidden cases – this cell is to be estimated in capture-recapture calculations -- it is logically not possible to estimate the highest order interaction. This problem is practically handled by assuming that there is **no highest-order interaction**, but this is not a realistic assumption in many situations – particularly if the capturing processes are highly correlated.

19 Fish are captured at t1, marked and released and some time later – t2 – a new sample is captured.

5.2 "Best model" chosen by statistical criteria vs. based on substantial considerations

5.2.1 Model selection in log-linear models

A correct estimate is only possible if we base our capture-recapture calculations on a model that represents the true situation. The models we can choose from differ in the kind and number of interactions. Based on three samples we have nine possible models.

- The model assuming that all capturing processes are independent, which is formally equivalent to no interactions (**independence model**),
- three models with one second-order-interaction,
- three models with two second-order-interaction
- one model with all three possible second-order-interactions (**saturated model** in capture-recapture analyses)
- one model with the third-order-interaction (**saturated model** in standard log linear models)

Since the last of these models cannot be estimated in capture-recapture analysis due to one missing cell, we have to assume that there is no third-order-interaction, as explained in chapter 5.1.3, and this leaves us with only 8 remaining models. In this case the model with **all three possible second-order-interactions** is the **saturated model**.

Additionally we could investigate all three possible 2-sample-capture-recapture approaches. In these cases the highest order interaction, which is at the same time the only possible one, cannot be estimated and we have to assume that the capturing processes are independent. Total independence of the capturing processes is rather implausible, though, in the social sciences and therefore 2-sample-capture-recapture approaches should only be considered as last resort if higher order capture-recapture approaches are not feasible. The estimates based on different models usually vary quite a bit. Therefore it is essential to choose the correct model.

5.2.2 The mechanistic statistical approach and the content oriented approach to model selection

The common approach to model selection is to choose the "best model" based on statistical criteria. The intrinsic problem with mechanistic selection of "best models" based on statistical algorithms is that "not significant" is by no means equivalent to "non-existing". Statisticians are aware of this problem but commonly do not really care a lot as long as they get away with it and non-statisticians usually are not even aware of the problem. We have to move concepts like "second-order errors", "sample-size considerations", "statistical power" and related concepts much more into the focus of our daily research considerations. Due to a lack of statistical power, commonly the application of mechanical statistical algorithms to choose "best models" leads to arbitrariness. If the basic concepts mentioned above are not considered adequately contradictions cannot be explained rationally and do not lead to improvement but to a rapid change from one scientifically based result to the next one. That way research has more in common with fast changing fashions than with the ideal of a growing body of scientific knowledge.

The alternative method – to choose models based on substantial considerations – is common practice in research as well. The latter approach may result in "scientific opportunism" if we are not careful. We are to choose content oriented models based on existing research findings, formulate arguments and decisions explicitly and apply these models on independent new data sets. If content oriented model selection is performed post hoc – that is after data are already available and particularly after first analyses have already been performed – scientific opportunism is almost inevitable.

In this essential conflict between two problematic alternatives it is sensible to try a compromise between the mechanistic statistical approach and the content oriented approach. The basis should be openness to perceive contradictions rather than to hide them and to be aware that adjusting assumptions systematically until we arrive at expected results is not "the art of scientific modelling", as some persons perceive it, but a very serious form of opportunistic insincerity.

5.3 Case definition

Common sense says that we have to define phenomena before we can measure them. If we look critically into human and social sciences we instantly realise that practical work of scientists usually

deviates quite a bit from this common-sense notion. Psychologists, economists, social scientists, and many other professions have often only quite vague ideas about what they want to measure in the beginning. Then they start to construct measurement techniques and spend much time finding out what constructs or dimensions their new techniques assess. This holds true for the construction of most intelligence tests and personality questionnaires; this holds true for many important economic indexes and this certainly also holds true for case definitions related to prevalence estimates based on the capture-recapture approach.

The merit of the capture-recapture approach is that it makes use of existing data sources. In estimating prevalence by this approach we usually have hardly any chance to decide on a case definition. We have to take what is in the registers. We may speculate before and after our calculations on an implicit case definition inherent to the data and on the implications of varying case definitions in different data sources, but we unfortunately hardly ever are in a position to collect all data using precise selection criteria and collecting detailed background information to base our work on.

5.4 What happens if assumption 1 is violated?

Assumption 1 is that only cases in line with the case definition are included in the data samples. In chapter 0 we argued that the samples "substitution treatment" and "drug related deaths" consist to a very high degree of cases according to the case definition "**problematic opiate consumption**", while we have to accept that a certain percentage of cases with "opiate related police charges" are non-cases according to the case definition.

It is not easy to estimate how far assumption 1 is violated; i.e. how many "opiate related police charges" in Austria are not related to "problematic opiate use". Empirical research on this matter would be of great help, but since no empirical research data are available on this issue, all we can do is speculate on the possible scope of this violation. It is quite plausible to assume that the number of opiate experimenters, occasional opiate users and opiate traffickers outweighs the number of "problematic opiate users" in the population. Since the likelihood of being charged by the police is quite high for opiate addicts, who are in permanent contact with opiate trafficking and the scene, and low for experimenters and occasional users, it is plausible, however, that a large majority of the cases charged by the police are cases according to the case definition. Until we have some empirical evidence on this issue all we can do is crudely speculate on the rate of non-cases.

In order to demonstrate the possible impact of this violation, we calculated expected cell frequencies assuming that that 20% (Scenario 1) of "opiate related police charges" are not cases according to our case definition, dropped the number in the cell representing the hidden cases and estimated the hidden population based on a 3-sample-capture-recapture model. (Tab. 7 and Tab. 8).

5.4.1 Scenario 1: Impact of non-cases

Scenario 1 is based on the following assumptions²⁰:

- In a certain year the number of "problematic opiate users" $N = 100,000$,
- the probability of cases being included in "substitution treatment" $p(s+) = 15\%$,
- the probability of cases being included in "opiate related police charges" $p(p+) = 15\%$,
- the probability of cases being included in "drug related deaths" $p(d+) = 1\%$,
- "substitution treatment data" and "opiate related police charges" are uncorrelated (odds-ratio = 1)
- "substitution treatment data" and "drug related deaths data" are uncorrelated (odds-ratio = 1)
- "opiate related police charges" and "drug related deaths data" are uncorrelated (odds-ratio = 1)
- **Violation of assumption 1 ("Only cases in line with the case definition are considered."):**
20% of the "opiate related police charges" are not cases according to our case definition.

Under the above assumptions we calculated the expected cell frequencies (Tab. 49 – Tab. 51) based on the **simple multiplication approach** (chapter 2.6.2).

20 For notation see chapter 2.6.1

Tab. 49: Scenario 1: Contingency table for cases

| | | | | |
|-------------|------------------------|----------|------------------|----------|
| N = 100,000 | police (p) | | | |
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| | drug related death (d) | present | absent | present |
| present | 22.5 | 127.5 | 127.5 | 722.5 |
| absent | 2,227.5 | 12,622.5 | 12,622.5 | 71,527.5 |

Tab. 50: Scenario 1: Contingency table for captured non-cases

| | | | | |
|-------------------------|------------------------|---------|------------------|---------|
| N = 3,750 ²¹ | police (p) | | | |
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| | drug related death (d) | present | absent | present |
| present | 0.0 | 0.0 | 0.0 | 0.0 |
| absent | 0.0 | 3,750.0 | 0.0 | 0.0 |

Tab. 51: Scenario 1: Contingency table for cases plus captured non-cases

| | | | | |
|-------------|------------------------|----------|------------------|----------|
| N = 103,750 | police (p) | | | |
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| | drug related death (d) | present | absent | present |
| Present | 22.5 | 127.5 | 127.5 | 722.5 |
| Absent | 2,227.5 | 16,372.5 | 12,622.5 | 71,527.5 |

As we can see in Tab. 52 most of the theoretically possible 3-sample and 2-sample capture-recapture estimates lead to a substantial overestimation of the known cases. SPSS suggests the saturated model as "best model". This leads to a 26.4% overestimation of the true sample size – or the other way round – the true sample size is about 20% less than the estimate.

21 All cases amount to 100,000 of which 15,000 are captured by the police. Additionally 3,750 non-cases are captured by the police. All of the latter are concentrated in the cell "non-drug-related deaths" and "non-substitution". A number of 3,750 is equivalent to 20% of all cases plus police captured non-cases (15,000+3,750=18,750).

Tab. 52: Analysis for scenario 1

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|-------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 8.17 | 3 | 0.043 | 32,223 | 91,510 | 123,732 | 23.7% |
| + s*p | 4.65 | 2 | 0.098 | 32,223 | 81,292 | 113,514 | 13.5% |
| + s*d | 0.00 | 2 | 1.000 | 32,223 | 92,781 | 125,004 | 25.0% |
| + p*d | 8.14 | 2 | 0.017 | 32,223 | 91,436 | 123,659 | 23.7% |
| + s*p + s*d | 0.00 | 1 | 1.000 | 32,223 | 92,781 | 125,004 | 25.0% |
| + s*p + p*d | 1.29 | 1 | 0.256 | 32,223 | 71,525 | 103,747 | 3.7% |
| + s*d + p*d | 0.00 | 1 | 1.000 | 32,223 | 92,781 | 125,004 | 25.0% |
| saturated | 0.00 | 0 | 1.000 | 32,223 | 94,146 | 126,368 | 26.4% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 31,500 | 93,500 | 125,000 | 25.0% |
| s + d | 0.00 | 0 | 1.000 | 15,850 | 84,150 | 100,000 | 0.0% |
| p + d | 0.00 | 0 | 1.000 | 19,600 | 105,400 | 125,000 | 25.0% |

The bias naturally does not only cause an overestimation of the true sample size, but also causes deviation in other parameters. As we can see in Tab. 53, the probabilities to be included in "substitution treatment" or to "drug related deaths" are underestimated and we get an artificial relationship between "substitution treatment" and "drug related deaths" (odds-ratio = 1.31).

Tab. 53: Comparison true model parameters vs. estimated parameters

| deviation from assumption: 20% of "opiate related police charges" are non-cases | true | estimate |
|--|---------|----------|
| total number of cases | 100,000 | 126,386 |
| probability of cases being included in "substitution treatment" | 15.0% | 11.9% |
| probability of cases being included in "opiate related police charges" | 15.0% | 14.8% |
| probability of cases being included in "drug related deaths" | 1.0% | 0.8% |
| odds-ratio "substitution treatment data" with "opiate related police charges" | 1.00 | 1.01 |
| odds-ratio "substitution treatment data" with "drug related deaths data" | 1.00 | 1.31 |
| odds-ratio "opiate related police charges" with "drug related deaths data" | 1.00 | 1.01 |

If we assume that 20% of the "opiate related police charges" are not cases according to the case definition, we could compensate the bias by adjusting the table of observed frequencies accordingly: We observed 2,698 cases "charged by the police opiate related" in our database (Tab. 7). If we assume that 20% of these cases are not cases according to the case definition, we arrive at 540 "non-cases" to be subtracted before the capture-recapture calculations. On logical grounds we can derive that these non-cases must be located in cell "s-p+d"²². We therefore have to reduce n(s-p+d-) from 2,253 down to 1,713 cases. The relevant cell "s-p+d-" in Tab. 54 is marked in bold type.

²² For notation see chapter 2.6.1

Tab. 54: Contingency table for all Austrians aged 15-54 (adjusted according to scenario 1)

| | | | | |
|--|----------------------|--------------|----------------------|----------------------|
| population size ²³ 4,608,295 | police t95 (p) | | | |
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t95 (d) | present | absent | present | absent |
| present | 4 | 40 | 28 | 127 |
| absent | 401 | 1,713 | 2,189 | N _{hid} = ? |

Tab. 55: Analysis all Austrians aged 15-54 years (adjusted according to scenario 1)

| | 3-sample-capture-recapture approach | | | | | | | |
|--------------------------------|-------------------------------------|----------|--------------|--------------|--------------|--------------|-------------------------------|---|
| | chisq | df | p | observed | estimated | total | unadjusted estimates (Tab. 8) | deviation from unadjusted "best model" estimation |
| independence | 10.37 | 3 | 0.016 | 4,502 | 9,182 | 13,684 | 16,876 | -20.8% |
| + s*p | 7.47 | 2 | 0.024 | 4,502 | 7,590 | 12,092 | 13,584 | -30.0% |
| + s*d | 8.68 | 2 | 0.013 | 4,502 | 9,023 | 13,525 | 16,910 | -21.7% |
| + p*d⁸ | 2.81 | 2 | 0.245 | 4,502 | 9,504 | 14,006 | 17,276 | -18.9% |
| + s*p + s*d⁷ | 0.22 | 1 | 0.643 | 4,502 | 5,439 | 9,941 | 12,195 | -42.5% |
| + s*p + p*d | 2.76 | 1 | 0.097 | 4,502 | 9,929 | 14,431 | 14,971 | -16.5% |
| + s*d + p*d⁹ | 2.05 | 1 | 0.153 | 4,502 | 9,351 | 13,853 | 17,341 | -19.8% |
| saturated | 0 | 0 | 1.000 | 4,502 | 4,645 | 9,147 | 11,150 | -47.1% |

| | 2-sample-capture-recapture approaches | | | | | | | |
|-------|---------------------------------------|----|-------|----------|-----------|--------|-------------------------------|---|
| | chisq | df | p | observed | estimated | total | unadjusted estimates (Tab. 8) | deviation from unadjusted "best model" estimation |
| s + p | 0 | 0 | 1.000 | 4,375 | 9,596 | 13,971 | 17,467 | -19,1% |
| s + d | 0 | 0 | 1.000 | 2,789 | 13,517 | 16,306 | 16,306 | -5,6% |
| p + d | 0 | 0 | 1.000 | 2,313 | 7,447 | 9,760 | 12,202 | -43,5% |

If we accept the "best model" as suggested by the SPSS algorithm (with interaction s*p and s*d), the resulting estimate of "problematic opiate users" for Austria amounts to 9,941 (Tab. 55), which is 42.5% less than the non-adjusted estimate of 17,276 "problematic opiate users". If we consider that, due to the nature of "drug related deaths" (see chapter 5.7.2) we should expect a relevant interaction between "drug related deaths" and "substitution treatment" as well as between "drug related deaths" and "opiate related police charges" (interactions s*d and p*d) the estimate is closer to the one without the adjustment (13,853), but also deviates 19.2%. The circumstance that "opiate related police charges" include data of non-cases should on the other hand cause a correlation between "substitution treatment" and "opiate related police charges" (interaction s*p) as well. Consequently there is reason to choose the saturated model (interactions s*p, s*d, p*d) and this model produces an estimate of 9,147 cases only.

It is not easy to draw a conclusion from scenario 1. If we speculate that 20% of the "opiate related police charges" are not cases according to the case definition, we would have to expect an overestimate of the true population size between 3.7% and 26.4% (Tab. 52) depending on the model we choose. If we try to compensate for the possible bias by adjusting the contingency table we based the main calculation on (Tab. 7) accordingly (Tab. 54) we get estimates from 5.6% to 47.1% less. Again, what happens depends on the model we choose. In other words, due to the fact that assumption 1 is certainly violated to some degree, we have to expect a substantial bias.

²³ see Tab. 6

5.5 How likely is it that assumption 2 is violated?

Assumption 2 is "**perfect identification of matching cases**". In this study we were in the extremely lucky situation of having almost perfect identifiers: one letter of the first name, three letters of the second name, gender, date of birth and province of residence. Only 7 out of 6,298 cases (Tab. 5) had identical identifiers, even though the unique personal code within the database made it evident that the cases were actually not identical. Since the data we used are entered and used by government bodies and not only collected for research purposes, we may suspect that the error rate caused by data entry is rather low. To sum up, we may be quite sure that violations to assumption 2 are not very dramatic and can be neglected in our considerations.

On the other hand, it is quite obvious that working with a smaller set of identifiers is not sufficient. Hay et al. (1998) have claimed that forename initial, surname initial, sex and date of birth should be sufficient per region. Since we had more than one region we added region of residence as identifier and reduced from 3 letters of the forename to 1 letter of forename. As a result, the number of erroneous matches increased (different persons with identical set of identifiers) from 0.11% (in numbers: 7 from 2698 cases, compare Tab. 5) to 1.42%, which is almost 12 times as much.

5.6 Consecutive capturing vs. simultaneous capturing processes

The normal situation in capture-recapture approaches to estimate wild animal populations is that a first sample of animals is captured, marked and released in time period t_1 and that capturing a second sample of animals happens in period t_2 after all of sample 1 has been captured, marked and released already. We could refer to this situation as **consecutive capturing process**.

The standard situation in social science applications is that sample 1 and sample 2 are captured in the same time interval. We could refer to this situation as **simultaneous capturing process**.

At first sight the two situations seem to be very different:

- In consecutive capturing processes all cases captured by process 1 are already marked when process 2 starts. Therefore the probability of capturing a previously marked case (re-captures) is identical all over time period 2.
- In simultaneous capturing processes naturally no case has been marked at the beginning of the observation period. Therefore at first sight the probability of recapturing a case seems to be zero for all capturing processes in the beginning of the observation period and to increase constantly as time goes on. This interpretation is incorrect, however. All capturing processes mark cases by assigning identifiers and it is irrelevant for the analysis if a case captured by one process has already been marked by the other process or will be marked later in the observation.

The obvious conclusion is that consecutive and simultaneous capturing processes, even though they seem quite different at first glance, produce identical patterns of captures and recaptures, given the basic assumptions are not violated.

5.7 How likely is it that assumption 3 is violated?

Assumption 3 is the "**closed population assumption**". It is quite obvious that the samples used in our analysis do not represent closed populations. In a one year observation period we may expect quite a few new events of problematic opiate use to occur in a region, that some addicts stop their problem behaviour and that there is relevant migration of "problematic opiate users" into the region or out of it. Since old cases dropping out of the population and new cases entering the population goes on while the capturing process goes on, we cannot use the simple multiplication approach to demonstrate how violations of assumption 3 affect the estimate, but we have to use the matrix approach.

A second form of violation of assumption 3 is the occurrence of drug related death, particularly problematic if "drug related deaths data" are used in the capture-recapture approach.

5.7.1 Scenario 2: Impact of open population

Scenario 2 is based on the following assumptions²⁴:

- At any point of time in a year the number of "problematic opiate users" $N = 100,000$,
- the probability of cases being included in "substitution treatment" $p(s+) = 15\%$,
- the probability of cases being included in "opiate related police charges" $p(p+) = 15\%$,
- the probability of cases being included in "drug related deaths" $p(d+) = 1\%$,
- "substitution treatment data" and "opiate related police charges" are uncorrelated (odds-ratio = 1)
- "substitution treatment data" and "drug related deaths data" are uncorrelated (odds-ratio = 1)
- "opiate related police charges" and "drug related deaths data" are uncorrelated (odds-ratio = 1)
- **Violation of assumption 3 (closed population assumption):**
10% of the population are lost in the course of one year observation period and the lost cases are compensated for by new cases. The total population size is constant.

Under the above assumptions we calculate the expected cell frequencies (Tab. 56) based on the **approximate matrix approach** (chapter 2.6.3).

Tab. 56: Scenario 2: Contingency table for cases all cases remaining, dropped and gained

| N = 100,000 (average) N = 110,000 (sum) | police (p) | | | |
|--|------------------|----------|------------------|----------|
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| drug related death (d) | present | absent | present | absent |
| present | 21.4 | 123.9 | 123.9 | 731.0 |
| absent | 2,161.3 | 12,733.0 | 12,733.0 | 81,372.5 |

If we base a 3-sample-capture-recapture model on the expected cell frequencies of all cases remaining, dropped and included in the population (Tab. 56), SPSS suggests the "best model" is the saturated model. The estimated total sample size based on this model is 106,435 cases instead of the true 100,000. This is an overestimation of 6.4% (Tab. 57). To be very precise, the total number of cases involved sometimes in the observation period is 110,000, but since 10,000 cases leave the population and are replaced by new cases all over the observation period, the average number of cases at any point of time is 100,000.

From scenario 2 we can derive that a violation of the closed population assumption leads to an estimate somewhere between the average number of cases and the total number of cases in the observation period. If we are interested in the average number of cases, if the fluctuation is 10%, as assumed in scenario 2, and if the saturated model is chosen, as suggested by the SPSS algorithm then we have to expect an overestimate of 6.4%.

If we choose other models we must expect an overestimate of around 3.7%. The latter case, that SPSS suggests another model, is likely to occur if the sample size is smaller. If, for example, we base an identical calculation on a population size of 18,000 "problematic opiate users", the SPSS algorithm suggests independence of the three capturing processes resulting in a bias of 3.6%.

24 For notation see chapter 2.6.1

Tab. 57: Analysis for scenario 2

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 0.01 | 3 | 1.000 | 28,627 | 74,997 | 103,624 | 3.6% |
| + s*p | 0.01 | 2 | 0.996 | 28,627 | 75,012 | 103,639 | 3.6% |
| + s*d | 0.01 | 2 | 0.996 | 28,627 | 75,004 | 103,632 | 3.6% |
| + p*d | 0.01 | 2 | 0.996 | 28,627 | 75,004 | 103,632 | 3.6% |
| + s*p + s*d | 0.01 | 1 | 0.934 | 28,627 | 75,124 | 103,752 | 3.8% |
| + s*p + p*d | 0.01 | 1 | 0.934 | 28,627 | 75,124 | 103,752 | 3.8% |
| + s*d + p*d | 0.01 | 1 | 0.934 | 28,627 | 75,012 | 103,639 | 3.6% |
| saturated | 0.00 | 0 | 1.000 | 28,627 | 77,808 | 106,435 | 6.4% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 27,897 | 75,723 | 103,620 | 3.6% |
| s + d | 0.00 | 0 | 1.000 | 15,894 | 87,823 | 103,717 | 3.7% |
| p + d | 0.00 | 0 | 1.000 | 15,894 | 87,823 | 103,717 | 3.7% |

Tab. 58: Comparison true model parameters vs. estimated parameters

| Deviation from assumption closed population: consecutive capturing processes 10% of the population are replaced during the observation period | true | estimated |
|---|---------|-----------|
| total number of cases | 100,000 | 106,435 |
| probability of cases being included in "substitution treatment" | 15.0% | 14.1% |
| probability of cases being included in "opiate related police charges" | 15.0% | 14.1% |
| probability of cases being included in "drug related deaths" | 1.0% | 0.9% |
| odds-ratio "substitution treatment data" with "opiate related police charges" | 1.00 | 1.04 |
| odds-ratio "substitution treatment data" with "drug related deaths data" | 1.00 | 1.03 |
| odds-ratio "opiate related police charges" with "drug related deaths data" | 1.00 | 1.03 |

5.7.2 Scenario 3: Impact of drug related deaths – uncompensated

Scenario 3 is based on the following assumptions²⁵:

- The number of "problematic opiate users" $N = 100,000$,
- the conditional probability of cases being included in "substitution treatment" given that they have not been included in any of the other two samples previously $p(s+/p-,d-) = 15\%$,
- the conditional probability of cases being included in "opiate related police charges" given that they have not been included in any of the other two samples previously $p(p+/s-,d-) = 15\%$,
- the conditional probability of cases being included in "drug related deaths" given that they have not been included in any of the other two samples previously $p(d+/s-,p-) = 1\%$

• Violation of assumption 3 (closed population assumption):

Cases who "die a drug related death" cannot enter "substitution treatment data" or run into "opiate related police charges" afterwards: $f(d \rightarrow s) = f(d \rightarrow p) = 0$

The three conditions "substitution treatment", "opiate related offence" and "drug related deaths" are otherwise independent: $f(s \rightarrow p) = f(p \rightarrow s) = f(s \rightarrow d) = f(p \rightarrow d) = 1$

²⁵ For notation see chapter 2.6.1

Under the above assumptions we calculated the expected cell frequencies (Tab. 59) based on the **approximate matrix approach** (chapter 2.6.3).

Tab. 59: Scenario 3: Contingency table for cases

| | | | | |
|------------------------|------------------|----------|------------------|----------|
| N = 100,000 | police (p) | | | |
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| drug related death (d) | present | absent | present | absent |
| present | 7.8 | 69.2 | 69.2 | 853.8 |
| absent | 2,227.5 | 12,622.5 | 12,622.5 | 71,527.5 |

According to scenario 3 there is an interaction between "substitution treatment" and "drug related deaths" (s*d) and an interaction between "opiate related police charges" and "drug related deaths" (p*d) but no interaction between "substitution treatment" and "opiate related police charges". The true odds-ratios are given in Tab. 61.

Tab. 60: Analysis for scenario 3

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|--------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 80.38 | 3 | 0.000 | 28,473 | 76,734 | 105,206 | 5.2% |
| + s*p | 1.50 | 2 | 0.472 | 28,473 | 160,492 | 188,965 | 89.0% |
| + s*d | 40.60 | 2 | 0.000 | 28,473 | 74,191 | 102,664 | 2.7% |
| + p*d | 40.60 | 2 | 0.000 | 28,473 | 74,191 | 102,664 | 2.7% |
| + s*p + s*d | 1.43 | 1 | 0.233 | 28,473 | 155,796 | 184,268 | 84.3% |
| + s*p + p*d | 1.43 | 1 | 0.233 | 28,473 | 155,796 | 184,268 | 84.3% |
| + s*d + p*d | 0.70 | 1 | 0.404 | 28,473 | 71,525 | 99,997 | 0.0% |
| saturated | 0.00 | 0 | 1.000 | 28,473 | 104,443 | 132,916 | 32.9% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 27,619 | 72,061 | 99,680 | - 0.3% |
| s + d | 0.00 | 0 | 1.000 | 15,850 | 178,057 | 193,907 | 93.9% |
| p + d | 0.00 | 0 | 1.000 | 15,850 | 178,057 | 193,907 | 93.9% |

If we drop the cell with the hidden cases in order to estimate the cell based on a 3-sample-capture-recapture approach, we find that SPSS suggests the saturated model again. If we accept this model as "best model" we arrive at an estimate of 132,916 which is a 32.9% overestimate (Tab. 60). The correct model, i.e. the one with the interaction "s*d and p*d", would have yielded an unbiased estimate, just as all other models not including the interaction term "substitution treatment" with "opiate related police charges" (s*p).

Due to the fact that persons who die cannot enter "substitution treatment" afterwards or be "charged by the police" afterwards, these two conditions correlate negatively with drug related death (odds-ratio = 0.47). According to the assumptions in scenario 3, the three conditions are causally independent otherwise. Because of the bias caused by the nature of "drug related deaths" we observe less negative correlation between the conditions actually correlating negatively and an artificial positive correlation between "substitution treatment" and "opiate related police charges" (odds-ratio = 1.46; Tab. 61).

Tab. 61: Comparison true model parameters vs. estimated parameters

| Deviation from assumption closed population: $f(d \rightarrow s) = f(d \rightarrow p) = 0$ $f(s \rightarrow p) = f(p \rightarrow s) = f(s \rightarrow d) = f(p \rightarrow d) = 1$ | true | estimated |
|--|---------|-----------|
| total number of cases | 100,000 | 132,916 |
| probability of cases being included in "substitution treatment" | 14.9% | 11.2% |
| probability of cases being included in "opiate related police charges" | 14.9% | 11.2% |
| probability of cases being included in "drug related deaths" | 1.0% | 0.8% |
| odds-ratio "substitution treatment data" with "opiate related police charges" | 1.00 | 1.46 |
| odds-ratio "substitution treatment data" with "drug related deaths data" | 0.47 | 0.66 |
| odds-ratio "opiate related police charges" with "drug related deaths data" | 0.47 | 0.66 |

What does this imply for our initial estimate concerning the number of "problematic opiate users" in Austria (Tab. 8)? We based our estimate on a model with an interaction term between "opiate related police charges" and "drug related deaths" $p*d$. According to the estimates in Tab. 60 this would result in a realistic estimate, given the true conditions are similar to scenario 3.

In order to check this on existing data, we did a further 3-sample-capture-recapture analysis using "substitution treatment data" and "opiate related police charges" from 1995 and "drug related deaths data" from the consecutive year. This naturally results in an increased loss of cases between the observation periods, but the problematic one-directional relationship between drug related deaths is controlled by the design.

Tab. 62: Contingency table for all Austrians aged 15-54 years using a sample of drug related deaths from the consecutive year

| population size ⁶ 4,608,295 | police t95 (p) | | | |
|---|----------------------|--------|----------------------|---------------|
| | present | | absent | |
| | substitution t95 (s) | | substitution t95 (s) | |
| drug related death t96 (d) | present | absent | present | absent |
| present | 6 | 25 | 37 | 116 |
| absent | 399 | 2,268 | 2,180 | $N_{hid} = ?$ |

The SPSS algorithm suggested only "substitution treatment" and "opiate related police charges" to be statistically dependent in this 3-sample-capture-recapture approach. Including the interaction term $s*d$ we estimated a total number of 17,336 "problematic opiate users" for Austria (Tab. 63), which is almost identical to the 17,276 cases estimated in the main analysis (Tab. 8).

This result adds support to the idea that the negative impact of using "drug related deaths data" from the same observation period can be largely controlled, if the correct interaction terms are used in the analysis.

To sum up, the fact that drug related death is causally related to the probability of meeting the other conditions constitutes a major problem in capture-recapture analyses. In the worst case this may cause a two-fold overestimation of the true population size. If we choose appropriate interaction terms though, we have a chance to compensate for the bias. To rely on the "best model" suggested by the SPSS algorithm is not a very sensible approach in this context. If we used the saturated model as suggested by SPSS in the course of analysing scenario 3 this would have yielded a 32.9% (Tab. 60) overestimate.

Tab. 63: Analysis for all Austrians aged 15-54 years (adjusted according to scenario 3)

| | 3-sample-capture-recapture approach | | | | | | | |
|--------------------|-------------------------------------|----------|--------------|--------------|---------------|---------------|-------------------------------|---|
| | chisq | df | p | observed | estimated | total | unadjusted estimates (Tab. 8) | deviation from unadjusted "best model" estimation |
| independence | 9.76 | 3 | 0.021 | 5,031 | 11,831 | 16,862 | 16,876 | -2.40% |
| + s*p ⁷ | 2.80 | 2 | 0.247 | 5,031 | 8,268 | 13,299 | 13,584 | -23.02% |
| + s*d ⁸ | 0.61 | 2 | 0.739 | 5,031 | 12,305 | 17,336 | 16,910 | 0.35% |
| + p*d | 9.75 | 2 | 0.008 | 5,031 | 11,881 | 16,912 | 17,276 | -2.11% |
| + s*p + s*d | 0.07 | 1 | 0.785 | 5,031 | 10,524 | 15,555 | 12,195 | -9.96% |
| + s*p + p*d | 0.46 | 1 | 0.496 | 5,031 | 6,835 | 11,866 | 14,971 | -31.32% |
| + s*d + p*d | 0.34 | 1 | 0.562 | 5,031 | 12,392 | 17,423 | 17,341 | 0.85% |
| saturated | 0.00 | 0 | 1.000 | 5,031 | 9,805 | 14,836 | 11,150 | -14.13% |

| | 2-sample-capture-recapture approaches | | | | | | | |
|-------|---------------------------------------|----|-------|----------|-----------|--------|-------------------------------|---|
| | chisq | df | p | observed | estimated | total | unadjusted estimates (Tab. 8) | deviation from unadjusted "best model" estimation |
| s + p | 0.00 | 0 | 1.000 | 4,915 | 12,552 | 17,467 | 17,467 | 1.10% |
| s + d | 0.00 | 0 | 1.000 | 2,763 | 8,457 | 11,220 | 16,306 | -35.06% |
| p + d | 0.00 | 0 | 1.000 | 2,851 | 13,163 | 16,014 | 12,202 | -7.31% |

5.7.3 Scenario 4: Impact of drug related deaths – compensated

Scenario 4 is identical to scenario 3 except that the probability of "drug related death" after having been in "substitution treatment" or after "opiate related police charges" is assumed to be twice as high as under the independence assumption²⁶.

- At any point of time in a year the number of "problematic opiate users" N = 100 000,
- the conditional probability of cases being included in "substitution treatment" given that they have not been included in any of the other two samples previously $p(s+/p-,d-) = 15\%$,
- the conditional probability of cases being included in "opiate related police charges" given that they have not been included in any of the other two samples previously $p(p+/s-,d-) = 15\%$,
- the conditional probability of cases being included in "drug related deaths" given that they have not been included in any of the other two samples previously $p(d+/s-,p-) = 1\%$,

• Violation of assumption 3: (closed population assumption)

Cases who "die drug related" cannot enter "substitution treatment" or be "charged by the police opiate related" afterwards: $f(d \rightarrow s) = f(d \rightarrow p) = 0$

Cases who enter "substitution treatment" or are "charged by the police opiate related" have a two-fold risk of a "drug related death" $f(s \rightarrow d) = f(p \rightarrow d) = 2$

"Substitution treatment" and "opiate related police charges" are independent: $f(s \rightarrow p) = f(p \rightarrow s) = 1$

Under the above assumptions we calculated the expected cell frequencies (Tab. 64) based on the **approximate matrix approach** (chapter 2.6.3).

²⁶ For notation see chapter 2.6.1

Tab. 64: Scenario 4: Contingency table for cases

| | | | | |
|------------------------|------------------|----------|------------------|----------|
| N = 100,000 | police (p) | | | |
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| drug related death (d) | present | absent | present | absent |
| present | 30.9 | 137.9 | 137.9 | 853.8 |
| absent | 2,197.0 | 12,557.5 | 12,557.5 | 71,527.5 |

Since cases who enter "substitution treatment" – according to scenario 4 – have a two-fold probability of "drug related death" in the rest of the observation period and since on the other side the probability of entering "substitution treatment" after dying is zero, these two effects compensate each other totally. So, the combined cell "substitution treatment" and "drug related deaths" is as likely as we would expect under independence assumption. The same holds true for "opiate related police charges" and "drug related deaths".

Tab. 65: Analysis for scenario 4

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|-------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 2.36 | 3 | 0.500 | 28,473 | 72,049 | 100,521 | 0.5% |
| + s*p | 1.68 | 2 | 0.432 | 28,473 | 76,039 | 104,511 | 4.5% |
| + s*d | 2.30 | 2 | 0.316 | 28,473 | 71,919 | 100,392 | 0.4% |
| + p*d | 2.30 | 2 | 0.316 | 28,473 | 71,919 | 100,392 | 0.4% |
| + s*p + s*d | 1.53 | 1 | 0.217 | 28,473 | 77,754 | 106,226 | 6.2% |
| + s*p + p*d | 1.53 | 1 | 0.217 | 28,473 | 77,754 | 106,226 | 6.2% |
| + s*d + p*d | 2.24 | 1 | 0.135 | 28,473 | 71,776 | 100,248 | 0.2% |
| saturated | 0.00 | 0 | 1.000 | 28,473 | 100,459 | 128,931 | 28.9% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 27,619 | 72,345 | 99,964 | 0.0% |
| s + d | 0.00 | 0 | 1.000 | 15,915 | 86,695 | 102,610 | 2.6% |
| p + d | 0.00 | 0 | 1.000 | 15,915 | 86,695 | 102,610 | 2.6% |

In other words, even though we have very complicated causal relationships, due to the fact that the effects compensate for each other, we are confronted with all three conditions being independent in the 3-sample-capture-recapture approach (true odds-ratio around 1; Tab. 66). The correct model therefore is the independence model.

If we drop the cell with the hidden cases in order to estimate it on the basis of a 3-sample-capture-recapture approach, we find that SPSS suggests the saturated model as "best model". This model is the one causing the highest bias though, a 28.9% overestimation (Tab. 65). Again we see that relying on the statistical algorithm to find the most appropriate model may be quite problematic.

Tab. 66: Comparison true model parameters vs. estimated parameters

| Deviation from assumption closed population: $f(d \rightarrow s) = f(d \rightarrow p) = 0$ $f(s \rightarrow d) = f(p \rightarrow d) = 2$ $f(s \rightarrow p) = f(p \rightarrow s) = 1$ | | |
|---|---------|-----------|
| | true | estimated |
| total number of cases | 100,000 | 128,931 |
| probability of cases being included in "substitution treatment" | 14.9% | 11.6% |
| probability of cases being included in "opiate related police charges" | 14.9% | 11.6% |
| probability of cases being included in "drug related deaths" | 1.2% | 0.9% |
| odds-ratio "substitution treatment data" with "opiate related police charges" | 1.00 | 1.40 |
| odds-ratio "substitution treatment data" with "drug related deaths data" | 0.97 | 1.30 |
| odds-ratio "opiate related police charges" with "drug related deaths data" | 0.97 | 1.30 |

The overestimation caused by erroneously assuming a saturated model also causes the estimated odds-ratios to reach 1.30 and 1.40 respectively, even though the true odds-ratios are close to 1.00 (Tab. 66).

The number of cases used for the scenarios (100,000) is relatively high. If we use only 18,000 cases in the scenario, a number more close to the number of "problematic opiate users" in Austria, we naturally find an identical pattern of biases, but due to less statistical power SPSS suggests an independence model. In this very situation the lack of statistical power would result in an unbiased estimate.

5.8 How likely is it that assumption 5 is violated?

Assumption 5 is the "**no causal relationship between sampling processes assumption**". In the previous chapter we already dealt with the fact that "drug related deaths" violate assumption 3, causing a very specific causal relationship to the other capturing processes, which violates assumption 5. We analysed this in two versions. In scenario 3 (see 5.7.2) we dealt with the isolated impact of "drug related deaths" and in scenario 4 we dealt with a possible case that this effect is compensated for by a corresponding positive relationships to other processes (see 5.7.3).

Dependence of the selection processes and correlation of the data is not a great problem in higher order capture-recapture approaches, as long as the dependencies are accounted for by choosing the correct models, i.e. by using the appropriate interaction terms.

It is quite obvious that the processes leading to registration in any of the registers used for this study are more or less causally dependent on each other. Persons who enter substitution treatment are supported by therapists and we may assume that their risk of needing "opiate related emergency ambulance transports" in the rest of the observation period should be less than before entering this treatment. Patients undergoing "substitution treatment" should be less involved in illicit drug trafficking, which should result in fewer "opiate related police charges" than before entering. On the other hand, if the police put pressure on "problematic opiate users" through charging them, this may increase their motivation to enter "substitution treatment". If "problematic opiate users" overdose it is quite common that ambulance and police are involved at the same time. To be known to the police may increase the probability of being charged by the police later on, etc.

Causal relationships of this nature, regardless whether they cause positive or negative correlations, are largely compensated for by capture-recapture approaches, given that the correct interaction terms are chosen. This will be illustrated by scenario 5.

5.8.1 Scenario 5: Impact of causally related processes

Scenario 5 assumes a variety of relationships related to the 3-sample-capture-recapture analysis based on "substitution treatment", "opiate related police charges" and "opiate related emergency ambulance transports" close to the relationships found in our data (Tab. 46, Tab. 48).

- At any point of time in a year the number of "problematic opiate users" $N = 100\ 000$,
- the conditional probability of cases being included in "substitution treatment" given that they have not been included in any of the other two samples previously $p(s+|p-,a-) = 15\%$,

- the conditional probability of cases being included in "opiate related police charges" given that they have not been included in any of the other two samples previously $p(p+/s-,a-) = 15\%$,
- the conditional probability of cases being included in "opiate related emergency ambulance transports" given that they have not been included in any of the other two samples previously $p(a+/s-,p-) = 3\%$,
- **Violation of assumption 5: (no causal relationship between sampling processes)**
For cases in substitution treatment the probability of "opiate related police charges" or of being involved in an "opiate related emergency ambulance transport" is reduced by 50% $f(s \rightarrow p) = f(s \rightarrow a) = 0.5$
- Whether cases were involved in "opiate related emergency ambulance transport" has no influence on the probability of entering "substitution treatment" $f(a \rightarrow s) = 1$
- If there were "opiate related police charges" the motivation to enter "substitution treatment" is increased by 50% $f(p \rightarrow s) = 1.5$
- cases with "opiate related police charges" have a three-fold risk of getting involved in "opiate related emergency ambulance transports" and vice versa $f(p \rightarrow a) = f(a \rightarrow p) = 3$

Under the above assumptions we calculated the expected cell frequencies (Tab. 67) based on the **approximate matrix approach** (chapter 2.6.3).

Tab. 67: Scenario 5: Contingency table for cases

| N = 100,000 | | police (p) | | | |
|---|--|------------------|----------|------------------|----------|
| | | present | | absent | |
| | | substitution (s) | | substitution (s) | |
| opiate related emergency ambulance transports (a) | | present | absent | present | absent |
| present | | 149.5 | 1,035.5 | 262.0 | 1,849.2 |
| absent | | 2,113.4 | 11,506.4 | 13,001.4 | 70,082.5 |

The adjustment to the strong deviation from assumption 5 is almost perfectly compensated for by the 3-sample-capture-recapture analysis (interaction terms $s*a$ and $p*a$). Two of three possible 2-sample-capture-recapture analyses would have yielded absolutely misleading results; an overestimation by 24.4% or an underestimation by 58.8% (Tab. 68).

The number of cases used for the scenarios (100,000) is relatively high. If we use 18,000 cases for the scenario, a number more close to the actual number of "problematic opiate users in Austria" for an identical scenario, we naturally would find an identical pattern of biases. A smaller sample size results in less statistical power and this may result in a reduced number of significant interaction terms. In this specific case the model suggested by SPSS includes only the most pronounced interaction $p*a$. An analysis based this model results in an overestimation of 4% (Tab. 68).

To sum up, violations of assumption 5 are compensated for quite well by higher order capture-recapture approaches, given that the correct interactions are identified and given that the dependency is not too strong. A stronger causal relationship will be analysed in scenario 6 (Tab. 71).

Tab. 68: Analysis for scenario 5

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|--------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 1,017.55 | 3 | 0.000 | 29,918 | 56,630 | 86,548 | -13.5% |
| + s*p | 530.69 | 2 | 0.000 | 29,918 | 34,020 | 63,937 | -36.1% |
| + s*a | 888.13 | 2 | 0.000 | 29,918 | 52,723 | 82,640 | -17.4% |
| + p*a | 19.24 | 2 | 0.000 | 29,918 | 74,102 | 104,020 | 4.0% |
| + s*p + s*a | 160.64 | 1 | 0.000 | 29,918 | 20,548 | 50,465 | -49.5% |
| + s*p + p*a | 7.09 | 1 | 0.008 | 29,918 | 91,748 | 121,665 | 21.7% |
| + s*a + p*a | 0.03 | 1 | 0.866 | 29,918 | 70,785 | 100,702 | 0.7% |
| saturated | 0.00 | 0 | 1.000 | 29,918 | 72,193 | 102,111 | 2.1% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 28,068 | 73,512 | 101,580 | 1.6% |
| s + a | 0.00 | 0 | 1.000 | 18,411 | 105,953 | 124,364 | 24.4% |
| p + a | 0.00 | 0 | 1.000 | 16,916 | 24,265 | 41,181 | -58.8% |

Tab. 69: Comparison true model parameters vs. estimated parameters

| Deviation from assumption "no causal relationship between sampling processes": f(s→p) = f(s→a) = 0.5 f(a→s) = 1 f(p→s) = 1.5 f(p→a) = f(a→p) = 3 | true | estimated |
|--|-----------------------|-----------|
| | total number of cases | 100,000 |
| probability of cases being included in "substitution treatment" | 15.5% | 15.4% |
| probability of cases being included in "opiate related police charges" | 14.8% | 14.7% |
| probability of cases being included in "drug related deaths" | 3.3% | 3.3% |
| odds-ratio "substitution treatment data" with "opiate related police charges" | 0.98 | 0.99 |
| odds-ratio "substitution treatment data" with "drug related deaths data" | 0.77 | 0.78 |
| odds-ratio "opiate related police charges" with "drug related deaths data" | 3.42 | 3.45 |

5.9 Assumption: No highest-order-interaction

In log-linear models based on a complete table, all interactions can be estimated from the data (see chapter 2.2). Since the capture-recapture approach is based on an incomplete table – the cell with the unknown cases is empty and the aim is to estimate the frequency of this cell – it is not possible to estimate the highest-order interaction and we have to assume the highest-order interaction to be zero, which may be correct, or may be incorrect (see chapter 5.1.3). The only way to check this empirically is to identify all hidden cases reliably and to see if the interaction is actually zero – a project that usually is beyond economical, technical and/or epistemological feasibility.

5.9.1 Scenario 6: Impact of strongly causally related processes and highest-order interactions

All we can do is judge the plausibility of the assumption that the highest-order interaction is zero. In order to do this we have to understand the nature of interaction terms in log-linear models. This can be explained relatively easily on the basis of the multiplicative nature of the effects.

Scenario 6 violates **assumption 5 (no causal relationship between sampling processes)** and additionally violates the **secondary assumption that the highest order interaction is zero**, the latter being essential to compensate for the effect of the first violation in a higher order capture-recapture analysis. We introduce the capturing processes A, B and C.

- Given the risk of being captured in group "a+b+" is 3 times higher than expected under the independence assumption – for reasons of simplicity we assume $f(a \rightarrow b) = f(b \rightarrow a) = 3$, and
- the risk of being captured in group "a+c+" is 2.5 times higher than expected under the independence assumption – for reasons of simplicity we assume $f(a \rightarrow c) = f(c \rightarrow a) = 2.5$ and
- the risk of being captured in group "b+c+" is 2 times higher than expected under the independence assumption – for reasons of simplicity we assume $f(b \rightarrow c) = f(c \rightarrow b) = 2$,
- then the risk of being captured in group "a+b+c+" must be 15 times ($= 3 \cdot 2.5 \cdot 2$) higher than expected under the independence assumption, given that there is no third-order-interaction (Tab. 70).

If the true frequency in the cell "a+b+c+" deviates from the expected number, we are confronted with a third-order interaction – but we have no means within a capture-recapture approach to find this out. We can demonstrate what would happen if

- the frequency of the combined cell were only 50% of the value expected, causing a substantial third-order interaction. Due to the reduction of this cell "a+b+c+" by 50% the total population size is reduced to $N = 99,549$ (Tab. 72).

In order to calculate specific cell frequencies based on this scenario we furthermore have to assume a population size and the unconditional probabilities of the processes A, B and C capturing a case, given that it has not previously been captured by any of the other processes.

- At any point of time in a year the number of cases $N = 100,000$,
- the conditional probability of cases being included in "condition A" given that they have not been included in any of the other two conditions previously $p(a+/b-,c-) = 5\%$,
- the conditional probability of cases being included in "condition B" given that they have not been included in any of the other two conditions previously $p(b+/a-,c-) = 10\%$,
- the conditional probability of cases being included in "condition C" given that they have not been included in any of the other two conditions previously $p(c+/a-,b-) = 15\%$,

Tab. 70: Scenario 6: Contingency table for cases based on the assumption that there is no highest order interaction

| N = 100,000 | condition B | | | |
|-------------|-------------|-----------|-------------|-----------|
| | present b+ | | absent b- | |
| | condition A | | condition A | |
| condition C | present a+ | absent a- | present a+ | absent a- |
| present c+ | 902.8 | 3,576.2 | 1,275.2 | 11,103.3 |
| absent c- | 538.2 | 6,706.2 | 3,223.2 | 72,675.0 |

Under the above assumptions we calculated the expected cell frequencies in Tab. 72 based on the **approximate matrix approach** (chapter 2.6.3). In Tab. 70 the frequencies are calculated according to scenario 6 under the assumption that there is no third-order interaction²⁷. Tab. 70 and Tab. 72 are identical except in cell "a+b+c+".

The SPSS algorithm based on the frequencies in Tab. 70 suggests that the saturated model is to be preferred, which is perfectly in line with the assumptions. However, the population size is overestimated by 15.6% (Tab. 71). This makes evident that strong dependencies may be overcompensated for if we base the calculations correctly on the saturated model.

27 The frequency in cell "a+b+c+" has not been multiplied by 0.5 yet.

Tab. 71: Analysis for scenario 6: Contingency table for cases based on the assumption that there is no highest order interaction

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|--------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 561.98 | 3 | 0.000 | 27,325 | 20,309 | 47,634 | -52.4% |
| + a*b | 567.05 | 2 | 0.000 | 27,325 | 20,197 | 47,522 | -52.5% |
| + a*c | 552.03 | 2 | 0.000 | 27,325 | 20,854 | 48,179 | -51.8% |
| + b*c | 472.29 | 2 | 0.000 | 27,325 | 25,377 | 52,702 | -47.3% |
| + a*b + a*c | 553.10 | 1 | 0.000 | 27,325 | 20,821 | 48,146 | -51.9% |
| + a*b + b*c | 415.93 | 1 | 0.000 | 27,325 | 28,066 | 55,391 | -44.6% |
| + a*c + b*c | 283.95 | 1 | 0.000 | 27,325 | 40,163 | 67,488 | -32.5% |
| saturated | 0.00 | 0 | 1.000 | 27,325 | 88,239 | 115,564 | 15.6% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| a + b | 0.00 | 0 | 1.000 | 16,222 | 32,098 | 48,320 | -51.7% |
| a + c | 0.00 | 0 | 1.000 | 20,619 | 25,351 | 45,970 | -54.0% |
| b + c | 0.00 | 0 | 1.000 | 24,102 | 20,021 | 44,123 | -55.9% |

Tab. 72: Scenario 6: Contingency table for cases based on the assumption that there is a highest-order interaction causing the combined cell to be only 50% of what we would expect if there was no highest-order interaction

| N = 99,549 | condition B | | | |
|-------------|-------------|-----------|-------------|-----------|
| | present b+ | | absent b- | |
| | condition A | | condition A | |
| condition C | present a+ | absent a- | present a+ | absent a- |
| present c+ | 451.4 | 3,576.2 | 1,275.2 | 11,103.3 |
| absent c- | 538.2 | 6,706.2 | 3,223.2 | 72,675.0 |

If we assume that there is a third-order interaction as defined in scenario 6, we arrive at Tab. 72. An analysis with SPSS suggests the model with the interaction terms "a*c" and "b*c" as "best model". If we base our capture-recapture estimation on this model we estimate a population size of 67,037, which underestimates the true sample size by 32.7%.

Tab. 73: Analysis for scenario 6: Contingency table for cases based on the assumption that there is a highest-order interaction causing the combined cell to be only 50% of what we would expect if there was no highest order interaction

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|--------|--------------------|
| | chisq | df | p | observed | estimated | total | bias ²⁸ |
| independence | 194.25 | 3 | 0.000 | 26,874 | 23,449 | 50,323 | -49.4% |
| + a*b | 126.01 | 2 | 0.000 | 26,874 | 21,917 | 48,790 | -51.0% |
| + a*c | 186.03 | 2 | 0.000 | 26,874 | 22,916 | 49,790 | -50.0% |
| + b*c | 58.47 | 2 | 0.000 | 26,874 | 30,437 | 57,310 | -42.4% |
| + a*b + a*c | 112.15 | 1 | 0.000 | 26,874 | 20,821 | 47,694 | -52.1% |
| + a*b + b*c | 46.15 | 1 | 0.000 | 26,874 | 28,066 | 54,939 | -44.8% |
| + a*c + b*c | 2.65 | 1 | 0.104 | 26,874 | 40,163 | 67,037 | -32.7% |
| saturated | 0.00 | 0 | 1.000 | 26,874 | 44,135 | 71,008 | -28.7% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| a + b | 0.00 | 0 | 1.000 | 15,770 | 46,740 | 62,510 | -37.2% |
| a + c | 0.00 | 0 | 1.000 | 20,168 | 31,979 | 52,147 | -47.6% |
| b + c | 0.00 | 0 | 1.000 | 23,651 | 22,265 | 45,916 | -53.9% |

The conclusions we can draw from scenario 6 are two-fold. One is that strong dependencies between the capturing processes are not fully compensated for by the capture-recapture algorithm respectively may be overcompensated for. The other one is that a violation of the secondary assumption – that the highest interaction is zero – may lead to a dramatic bias.

If the combined cell "a+b+c+" is less than expected under the assumption "no highest-order interaction" we are confronted with an underestimate – if the opposite is true, we are confronted with an overestimate.

5.10 How likely is it that assumption 4 is violated?

Assumption 4 is the "**homogeneity in terms of equal catchability assumption**". There can be no discussion that this assumption is violated seriously in the field of estimating the prevalence of "problematic opiate users". Some opiate addicts manage to live a rather inconspicuous life, have very reliable sources for their supply and the chances that the police find out about them is comparatively small. Others look and behave in a way that makes their drug problem quite obvious for others, buy and sell drugs on the open drug scene – directly under the eyes of the police – and therefore are likely to be charged regularly. This deviation from assumption 4 could be described as "**concordant heterogeneity**", since it varies concordantly in all samples.

But there is also another form of heterogeneity that could be referred to as "**discordant heterogeneity**". Discordant heterogeneity occurs if some cases have a high probability of being captured by one process and a low probability of being captured by another process and vice versa.

- There are some "problematic opiate users" who definitely refuse to accept "substitution treatment" and at the same time have a high risk of being "charged by the police" or to die a "drug related death". More technically speaking this would mean low catchability in "process S" but high catchability in "process P" and "process D".
- Some relatively healthy "problematic opiate users" are likely to enter "substitution treatment" and, as a consequence to quit illicit trafficking with opiates their risk of being charged by police and their risk

²⁸ Since we reduced the cell "a+b+c+" by 50% the total population size is reduced from 100,000 to 99,549. Therefore the bias in the last column is calculated based on 99,549 rather than 100,000.

of a drug related death is very small. More technically speaking this would mean high catchability in "process S" but low catchability in "process P" and "process D".

- Some "problematic opiate users" in very bad physical shape are likely to enter "substitution treatment" and as a consequence to quit illicit trafficking with opiates their risk of being charged by police is rather low, but due to their bad health conditions (AIDS, hepatitis, etc.) their risk of a drug related death is rather large. More technically speaking this would mean high catchability in "process S" and low catchability in "process P" but high catchability in "process D".

We could think of many more reasons for discordant heterogeneity.

We could control heterogeneity easily if we could divide the samples into homogeneous subsamples and perform separate calculations for each subsample, but since we usually do not have access to the necessary information to divide the sample into homogeneous subsamples, we cannot rule out uncontrolled heterogeneity.

Concordant heterogeneity on the aggregated level causes correlations between the processes that differ in nature from the correlation caused by causal dependencies – the case we dealt with in scenario 5 (5.8.1)

Discordant heterogeneity may happen in a way that the different patterns in the homogeneous subsample balance each other out on the aggregated level. In this case we do not even identify interactions that would make us suspicious.

Realistically speaking we have to expect a mixture of concordant and discordant heterogeneity. We will deal with both forms of heterogeneity separately in the following scenarios 7 and 8. For the sake of simplicity we assume causal independence between the capturing processes.

5.10.1 Scenario 7: Impact of concordant heterogeneity

Scenario 7 is based on the following assumptions²⁹:

- In a certain year the number of "problematic opiate users" $N = 100,000$,
- the probability of cases being included in "substitution treatment" $p(s+) = 14.9\%$,
- the probability of cases being included in "opiate related police charges" $p(p+) = 14.9\%$,
- the probability of cases being included in "opiate related emergency ambulance transports" $p(a+) = 3.2\%$,
- the processes of inclusion in "substitution treatment" and inclusion by "opiate related police charges" are causally unrelated (odds-ratio = 1)
- the processes of inclusion in "substitution treatment" and inclusion in "opiate related emergency ambulance transports" are causally unrelated (odds-ratio = 1)
- the processes of inclusion in "opiate related police charges" and inclusion in "opiate related emergency ambulance transports" are causally unrelated (odds-ratio = 1)

- **Violation of assumption 5: (homogeneity in terms of equal catchability assumption – concordant heterogeneity)**

The population can be divided into 8 homogeneous subgroups with different risks of being captured by the capturing processes. In subgroup 1 the risk is lowest for all processes and in subgroup 8 the risk is highest for all processes. The risk increases linearly from subgroup to subgroup in equal steps. The specific probabilities are:

- The probability of cases from subgroup 1 being included in "substitution treatment" $p(s+) = 3.3\%$.
- The probability of cases from subgroup 1 being included in "opiate related police charges" $p(p+) = 3.3\%$.
- The probability of cases from subgroup 1 being included in "opiate related emergency ambulance transports" $p(a+) = 0.7\%$.
- The probability of cases from subgroup 8 being included in "substitution treatment" $p(s+) = 26.4\%$.
- The probability of cases from subgroup 8 being included in "opiate related police charges" $p(p+) = 26.4\%$.
- The probability of cases from subgroup 8 being included in "opiate related emergency ambulance transports" $p(a+) = 5.6\%$.

29 For notation see chapter 2.6.1

Under the above assumptions we calculate the expected cell frequencies based on the **simple multiplication approach** (chapter 2.6.2) for each of the 8 subgroups and then aggregate the tables (Tab. 74).

Tab. 74: Scenario 7: Contingency table for cases assuming concordant heterogeneity

| | | | | |
|---|------------------|----------|------------------|----------|
| N = 100,000 | police (p) | | | |
| | present | | absent | |
| | substitution (s) | | substitution (s) | |
| opiate related emergency ambulance transports (a) | present | absent | present | absent |
| present | 123.5 | 465.6 | 465.6 | 2,095.4 |
| absent | 2,653.5 | 11,607.5 | 11,607.5 | 70,981.6 |

Tab. 75: Analysis for scenario 7

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|--------|--------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 2.48 | 3 | 0.479 | 29,018 | 50,676 | 79,694 | -20.3% |
| + s*p | 2.31 | 2 | 0.315 | 29,018 | 51,395 | 80,414 | -19.6% |
| + s*a | 2.47 | 2 | 0.292 | 29,018 | 50,721 | 79,740 | -20.3% |
| + p*a | 2.47 | 2 | 0.292 | 29,018 | 50,721 | 79,740 | -20.3% |
| + s*p + s*a | 2.07 | 1 | 0.150 | 29,018 | 52,245 | 81,263 | -18.7% |
| + s*p + p*a | 2.07 | 1 | 0.150 | 29,018 | 52,245 | 81,263 | -18.7% |
| + s*a + p*a | 2.44 | 1 | 0.118 | 29,018 | 50,777 | 79,796 | -20.2% |
| saturated | 0.00 | 0 | 1.000 | 29,018 | 60,743 | 89,761 | -10.2% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 26,923 | 52,489 | 79,412 | -20.6% |
| s + a | 0.00 | 0 | 1.000 | 17,411 | 62,001 | 79,412 | -20.6% |
| p + a | 0.00 | 0 | 1.000 | 17,411 | 62,001 | 79,412 | -20.6% |

Even though the "best model" is the saturated model, the SPSS algorithm suggests the independence model as best. If we base our 3-sample-capture-recapture approach on this model we underestimate the population size by 20%. Using the correct model – the saturated model – the underestimate would have been only 10% (Tab. 75).

As we can see in Tab. 75, all of the theoretically possible 3-sample and 2-sample capture-recapture estimates lead to a substantial underestimation of the true number of cases.

The bias naturally not only causes an underestimation of the true sample size, but also causes deviation in other parameters. As we can see in Tab. 76, the calculated probabilities to be included in any of the processes are inflated from 14.9% and 3.2% to 18.6% and 4.0%.

Tab. 76: Comparison true model parameters vs. estimated parameters

| Deviation from assumption: "homogeneity in terms of equal catchability assumption" concordant heterogeneity | true | estimate |
|---|---------|----------|
| total number of cases | 100,000 | 79,694 |
| probability of cases being included in "substitution treatment" | 14.9% | 18.6% |
| probability of cases being included in "opiate related police charges" | 14.9% | 18.6% |
| probability of cases being included in "opiate related emergency ambulance transports" | 3.2% | 4.0% |
| odds-ratio "substitution treatment" with "opiate related police charges" | 1.39 | 1.01 |
| odds-ratio "substitution treatment " with "opiate related emergency ambulance transports" | 1.33 | 1.00 |
| odds-ratio "opiate related police charges" with "opiate related emergency ambulance transports" | 1.33 | 1.00 |

The number of cases used for the scenarios (100,000) is relatively high. If we use only 18,000 cases in the scenario, a number more close to the number of problematic opiate users in Austria SPSS suggests an independence model as well.

5.10.2 Scenario 8: Impact of discordant heterogeneity

Scenario 8 is based on the following assumptions³⁰:

- In a certain year the number of "problematic opiate users" $N = 100,000$,
- the probability of cases being included in "substitution treatment" $p(s+) = 15\%$,
- the probability of cases being included in "opiate related police charges" $p(p+) = 15\%$,
- the probability of cases being included in "opiate related emergency ambulance transports" $p(a+) = 3\%$,
- the processes inclusion in "substitution treatment " and inclusion by "opiate related police charges" are causally unrelated (odds-ratio = 1)
- the processes inclusion in "substitution treatment " and inclusion by "opiate related emergency ambulance transports" are causally unrelated (odds-ratio = 1)
- the processes inclusion in "opiate related police charges" and inclusion by "opiate related emergency ambulance transports" are causally unrelated (odds-ratio = 1)
- **Violation of assumption 5: (homogeneity in terms of equal catchability assumption – discordant heterogeneity)**
The population can be divided into 8 homogeneous subgroups with two different risks of being captured by the capturing processes: a low and a high risk. The 8 subgroups represent all possible combinations of high and low risk.
- The probability of cases being included in "substitution treatment" $p(s+) = 10\%$ (low) respectively 20% (high).
- The probability of cases being included in "opiate related police charges" $p(p+) = 10\%$ (low) respectively 20% (high).
- The probability of cases being included in "opiate related emergency ambulance transports" $p(a+) = 2\%$ (low) respectively 4% (high).

Under the above assumptions we calculate the expected cell frequencies based on the **simple multiplication approach** (chapter 2.6.2) for each of the 8 subgroups and then aggregate the tables (Tab. 77).

³⁰ For notation see chapter 2.6.1

Tab. 77: Scenario 8: Contingency table for cases assuming discordant heterogeneity

| | | | | | |
|---|--|------------------|----------|------------------|----------|
| N = 100,000 | | police (p) | | | |
| | | present | | absent | |
| opiate related emergency ambulance transports (a) | | substitution (s) | | substitution (s) | |
| | | present | absent | present | absent |
| present | | 67.5 | 382.5 | 382.5 | 2,167.5 |
| absent | | 2,182.5 | 12,367.5 | 12,367.5 | 70,082.5 |

If we base our 3-sample-capture-recapture approach on this model we estimate exactly the correct population size, regardless of the model. The overall pattern is balanced in a way that all effects compensate each other perfectly (Tab. 78).

Tab. 78: Analysis for Scenario 8

| | 3-sample-capture-recapture approach | | | | | | |
|--------------|---------------------------------------|----|-------|----------|-----------|---------|------|
| | chisq | df | p | observed | estimated | total | bias |
| independence | 0.00 | 3 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| + s*p | 0.00 | 2 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| + s*a | 0.00 | 2 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| + p*a | 0.00 | 2 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| + s*p + s*a | 0.00 | 1 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| + s*p + p*a | 0.00 | 1 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| + s*a + p*a | 0.00 | 1 | 1.000 | 29,918 | 70,081 | 99,998 | 0% |
| saturated | 0.00 | 0 | 1.000 | 29,918 | 70,425 | 100,342 | 0% |
| | 2-sample-capture-recapture approaches | | | | | | |
| | chisq | df | p | observed | estimated | total | bias |
| s + p | 0.00 | 0 | 1.000 | 27,750 | 72,250 | 100,000 | 0% |
| s + a | 0.00 | 0 | 1.000 | 17,550 | 82,450 | 100,000 | 0% |
| p + a | 0.00 | 0 | 1.000 | 17,550 | 82,450 | 100,000 | 0% |

We get perfect estimations with all models regardless of the probability we use as low and as high probability. Even in the most extreme case, when the low probability is zero, we get unbiased estimates.

To sum up: the largest bias is to be expected from concordant heterogeneity, less is to be expected if heterogeneity happens discordantly. In the latter case some of the biasing effects balance each other out – they may even be balanced out perfectly, resulting in a perfect estimate, as in scenario 8.

6 Plausibility considerations

6.1 Comparison to international prevalence estimates

Prevalence estimates for problematic opiate consumption in 7 large cities in the age group between 15 and 54 years based on the capture-recapture methodology range from 0.14% in Helsinki to 2.11% in Dublin with a median value of 0.86% for Rome (Hay et al., 1997). If we apply these extremely different rates to the population of Vienna this is equivalent to between 1,313 and 19,788 cases with a median of 8,065 cases. If we take this line of reasoning even further and consider that roughly 50% of the Austrian "problematic opiate users" live in Vienna, the above rates translate into between 2,626 and 39,576 cases in Austria, with a median value of 16,130 cases.

Bühringer et al. (1997) estimated the number of frequent consumers of hard drugs in Germany between 100,000 and 150,000 persons. If we assume that these cases are primarily located in the age group between 15 and 54 years and consider that this age group amounts to 45,000,000 Germans, we get a prevalence rate between 0.2% and 0.3%. Applied to the Austrian population size this rate is equivalent to between 9,217 and 13,826 cases of "problematic opiate users".

Kraus et al. (1999) collected national prevalence rates for the Member States of the European Union and found prevalence rates of "problematic opiate users" for the age range between 15 and 54 years between 0.06% in Finland and 1.1% in the United Kingdom, with a medium value of 0.42% in Ireland. Applied to the Austrian population size this rate is equivalent to between 2,765 and 50,691 cases and a median value of 19,354.

It of course does not make much sense directly to project international rates of "problematic opiate use" to Austria, particularly if countries with a very low prevalence such as Finland and countries with a very high prevalence such as the United Kingdom are considered too, but we get a rough idea of the magnitude of variation in prevalence and prevalence estimates all over Europe. If we think that Austria is neither extreme in one direction or the other, an estimate of 15,000 - 20,000 problematic opiate users in Austria seems plausible, on the basis of the above-mentioned estimates.

6.2 Other prevalence estimation approaches based on Austrian data

Since "problematic opiate users" can be and are charged by the police repeatedly in each observation period it is possible to approximate the number of persons never charged by the police based on the "truncated Poisson approach". For this estimate similar assumptions as for the capture-recapture approach are necessary.

What we need to know for a simple truncated Poisson approach per observation period is

- the number of persons charged once (f1): In "opiate related police charges t95" f1 = 2,058
- the number of persons charged twice (f2): In "opiate related police charges t95" f2 = 423
- the number of persons charged altogether (S): In "opiate related police charges t95" S = 2,698

Smit & Toet (1997) recommended this approach for the estimation of "problematic opiate users" and applied a variety of different approaches to their Rotterdam data. We will apply two simple estimators they used, the one from Zeltermann (1988) and the one from Chao (1989), to our data.

The size of the total population according to Zeltermann can be estimated by
$$N = S + S / [1 - \exp(-2 \cdot f2 / f1)] = 2,698 + 2,698 / [1 - \exp(-2 \cdot 423 / 2,058)] = 10,702$$

The size of the total population according to Chao can be estimated by
$$N = 2 \cdot S + f1 \cdot f1 / (2 \cdot f2) = 2 \cdot 2,698 + 2,053 \cdot 2,053 / (2 \cdot 423) = 10,378.$$

Hay et al. (1998) cite Ghodse et al. (1985), who estimated that 1- 2% of all drug injectors die annually. If we apply this multiplier to "drug related deaths" in Austria, even though our case definition is "problematic opiate users" and not "intravenous drug users", we arrive at 12,200 to 24,400 cases. This estimate is based on a reference value of 244 drug related deaths for the observation period from September 1994 to August 1995 (Tab. 5).

The estimates based on the truncated Poisson approach are very low compared to our main analyses (17,276 cases, Tab. 8) and to what seems plausible extrapolating from other European estimates (see chapter 6.1).

7 Summary and discussion

The great advantage of the capture-recapture approach is that existing data sources may be used without any further empirical work – which makes the approach attractive and cheap. Another great advantage is that the estimate does not directly depend on the intensity of the capturing process. In other words, if we use police data, we do not have to worry if the police pressure is high in one year and low in the other year.

The great disadvantage is that the results depend heavily on basic assumptions, whereby we have to expect that all assumptions are violated more or less severely in the data sets we have access to. We could show that if assumption 1 (only cases according to the case definition are considered) is violated by 20% in one of the samples we may overestimate the number of cases by 26% (Tab. 52), but if we choose the truly best model, there may not be any bias at all.

If the quality of data is good and if there are sufficient identifiers, we do not have to worry too much about violating assumption 2 (perfect identification of matching cases). The use of fewer identifiers than we used in our study, particularly reducing from 3 letters of the forename to initials of the forename only, increases the number of erroneous matches 12-fold and should not be considered if at all possible.

Violation of assumption 3 (closed population) in case of a simultaneous capturing process results in an estimate somewhere between the average number of cases and the total number of cases involved at any time in the observation period. Since the average number is the number of primary interest, we have to expect an overestimate of around 60% of the fluctuation rate (Tab. 57).

Another form of violating assumption 3 (closed population assumption) is caused by including "drug related deaths data". Naturally persons who die cannot be captured by any of the other processes afterwards, which causes a negative dependency between the processes. Depending on the specific model chosen we may get anything between a perfect estimate and a two-fold overestimation (Tab. 60).

If the specific effect of including "drug related deaths data" is compensated for by a causal dependency in the sense that persons captured by other processes have a two-fold risk of a drug related death thereafter, the possible range of biases is reduced, but if we use the saturated model we still may overestimate the true number by 30% (Tab. 65).

Violations of assumption 5 (no causal relationship between sampling processes) can be compensated for in 3-sample-capture-recapture calculations to a large degree if the amount of dependency is not too high and if the correct model is chosen. If inadequate models are chosen, though, the resulting bias can be quite dramatic in either direction. In 2-sample-capture-recapture calculations negative dependency always causes an overestimate of the true population size and a positive dependency causes underestimation. In 3-sample-capture-recapture approaches these biases may be compensated for by some models, under-corrected by others and overcorrected by others. Therefore it is not easy to predict the direction of bias (Tab. 68).

Since we have to rely on incomplete contingency tables – the cell with the hidden population is unknown – we cannot possibly estimate the highest-order interaction. We have to assume that this interaction is zero; an assumption not at all likely in many situations. If the combined cell is less likely than expected, caused by a highest-order interaction, we have to expect a substantial underestimation of the true sample size; if the opposite is true we have to expect an overestimation (Tab. 71, Tab. 73).

If we violate assumption 4 (homogeneity in terms of equal catchability) we have to expect quite a severe bias if the heterogeneity is concordant. This type of heterogeneity results in artificial correlations between the data sets only partly compensated for. The degree of compensation depends on the model chosen. In the specific scenario we calculated, we found an underestimation between 10% and 20% (Tab. 75).

If assumption 4 (homogeneity in terms of equal catchability) is violated through discordant heterogeneity, some biases may balance each other out – if we are lucky, even totally, as in the scenario we calculated through (Tab. 78).

To sum up, the capture-recapture approach is highly susceptible to a variety of biases caused by violated assumptions. Some of these biases can be compensated in higher-order capture-recapture approaches if the correct model is chosen – but it is not easy to find the most appropriate model. Neither a mechanistic selection based on statistical algorithms nor a content oriented selection is to be clearly favoured. We could show that the "best model" suggested by the SPSS algorithm may be very far from

the optimal model and we discussed that to choose models content oriented gives the researchers many possibilities to manipulate their analysis in a way to produce the numbers they initially expected – which is unacceptable scientific opportunism.

If we had chosen an uncritical approach, we could have produced a very plausible number of 17,276 "problematic opiate users" for Austria, which would have been perfectly in line with the expectations (Tab. 8). The fact that the estimate based on the very plausible model with the interaction terms "s*d + p*d" yielded an almost identical estimate of 17,341 cases (Tab. 8) adds to the credibility of the estimate. Stratification by gender, by age and by province (Tab. 39) resulted in estimates ranging from 13,875 through 18,871 (19.7% less and 9.2% more) than the non-stratified estimate, which is not dramatic at all. If we only consider models with the interaction terms "s*d + p*d" in the "3-sample-capture-recapture models", the range of stratification related deviation reduces to between 7.5% less and 8% more (Tab. 39), which is even less dramatic. Based on the available analyses we get the impression that we can be quite confident about the estimate.

If we look at existing estimates for Vienna, we are in a much worse situation. There is one previous estimate existing for the same period of time (Seidler & Uhl, 1997) and there are 4 samples existing instead of 3, since we had an additional "opiate related emergency ambulance data" sample for Vienna only. This allowed us to do several independent estimations – and that way the results were much less convincing. Plausible estimates range from 6,334 to 18,585, for the same period of time, using identical case definitions and partly even the same sources of data. This must cast some serious doubts on the reliability of the method (see chapter 4.3.1). Since several estimates are around 11,000 and since this is between the extremes of 6,747 and 18,585, an estimate of around 11,000 is more plausible than the more extreme estimates, but one should be aware that the basis for this conclusion is not very sound.

In our analysis based on some very plausible scenarios, with situations relatively close to the situations we found in our data, we could demonstrate very convincingly that we have to expect strong biases caused by deviations from the basic assumptions, that some of these biases are compensated for and that others are not, that we have no sound criteria to help us evade heavily biased estimates and that the mechanistic way to choose the best model by statistical criteria is often extremely misleading.

For non-statisticians we want to stress once more that all this is not at all related to confidence intervals. Confidence intervals do not help us at all to estimate the range of possible variation caused by systematic errors but they help us to estimate the plausible range of deviation caused by random errors. Random variation is an additional source of unreliability not dealt with in the theoretical section of this paper at all.

We have to face the fact that we can only formulate very crude prevalence estimates based on capture-recapture methodology and that something like 50% less or 100% more than the estimated number is generally quite plausible. The fact that the application of the capture-recapture approach to social sciences is not highly reliable is well known to many experts. Korf (1997), in the popular EMCCDA monograph on prevalence estimation, warned that this approach fosters under- and over-estimation, and that to determine the exact degree of the resulting bias is still a problem. We developed an approach to quantify possible biases caused by violated assumptions, which is a clear step ahead.

Many experts and policy makers look for methods to assess and compare annual changes in incidence within countries and between countries. However, if we want to monitor annual changes in incidence, e.g. to assess the impact of a modified drug policy or to judge important trends, the capture-recapture approach does not lead us anywhere. An annual reduction of incidence by 20% – without doubt a dramatic improvement – would result only in a prevalence reduction of 2%, if we assume the average time of addiction to be around 10 years, as estimated by Simon et al. (1998) for Germany. It is obviously very unrealistic to expect that an estimation method producing estimates in the range of 50% less to 100% more can help to measure annual changes in incidence.

Nevertheless, the precision the capture-recapture approach can give may be sufficient for many scientific and practical questions. If we, in line with our main analyses, estimate that around 0.4% of the adult population younger than 54 are opiate addicts and compare this rate to an estimated rate of 5% alcoholics in this age group, we get a rough idea of the magnitude of both phenomena – and it does not really make a lot of difference if the true rate of opiate dependency is as low as 0.2% or as high as 0.8%.

To give a very rough estimate for the number of "problematic opiate users" in Austria we could say that no number between 9,000 and 36,000 can be ruled out definitely, on the basis of all the calculations we

did. Considering all aspects jointly we find 15,000 to 20,000 cases to be rather plausible, but this is a judgement by the authors of this report and by no means fully evidence based. For Vienna any number between 5,000 and 20,000 cannot be ruled out on the basis of the information we have, but it seems very likely that the true number is between 7,500 and 10,000. Here again we have to state that this as a judgement by the authors of this report and by no means fully evidence based.

In order to give a correct impression to our audience concerning the lack of precision, we feel it makes sense to propose round numbers like 10,000 for Vienna and 20,000 for Austria – and to stress explicitly that these are very crude estimates indeed.

A rate of around 0.4% “current problematic opiate users” cross-sectional in Austrian between 14 and 54 years of age, can be used to derive a crude estimate of “total-life-time-prevalence of problematic illicit drug use³¹” in Austria. In this estimate we have to account for three factors.

- One is that a certain percentage of problematic illicit drug users do not use opiates, and therefore are not included in this estimate,
- the second factor is that some younger persons who are no “problematic users of illicit drugs” yet will start problematic use patterns later in their lives and
- the third factor is that the mortality rate of “illicit drug users” is substantially elevated, so that the percentage reduces disproportionately in older age cohorts.

Considering these three facts we can crudely estimate that around 1% of Austrian young people will encounter serious drug problems in the course of their lives, given that the overall situation does not undergo any dramatic changes. Just for a comparison: “total-life-time-prevalence of alcoholism” in Austria has been estimated to be around 10% (Uhl & Springer, 1996; Uhl et al., 1999).

The present study was designed to give an impression of possible biases within the standard capture-recapture approach and aims at improving the quality of future capture-recapture analyses. The aim is not to discredit this approach for the context of prevalence estimates in the field of illicit drug use. We are convinced that more reliable and precise capture-recapture estimates can be expected in the future, if specific empirical research allows to quantify deviations from the basic assumptions and if these deviations are considered in more elaborate models.

31 The concept of “total-life-time-prevalence of a problem” refers to problem incidence in the course of the individuals’ life-time. Since the term “life-time-prevalence” is commonly used for “incidence before assessment” in a somewhat misleading manner, we use the word “total-life-time” to indicate that life-time is meant in the literal sense.

8 Syntax for calculations and practical examples

8.1 Programme to calculate expected contingency tables based on the approximate matrix approach

The simple multiplication approach (2.6.2), that is multiplying the cell frequencies and aggregating several contingency tables to one final table, can easily be achieved in Excel or in any other spreadsheet program and will not be explained here in detail.

The structure of the approximate matrix approach (2.6.3) based on SPSS is somewhat more complicated and therefore the syntax will be printed here. All relevant programme parameters are entered into the definition section of the programme. The variables are named different from the notation we defined in chapter 2.6.1 in order to make them compatible with the SPSS syntax:

- n: N_{tot} = total number of cases in population
- t: Number of steps to approximate the continuous process by a discrete process
- sprob: $p(s+/p-,d-)$ = conditional probability of being captured by process s, given that the case has not been previously captured by another process
- pprob: $p(p+/s-,d-)$ = conditional probability of being captured by process p, given that the case has not been previously captured by another process
- dprob: $p(d+/s-,p-)$ = conditional probability of being captured by process d, given that the case has not been previously captured by another process
- lprob: rate of cases leaving the population during the observation period and being replaced
- sp: $f(s \rightarrow p)$ = factor by which the probability to capture p increases after s has been captured
- ps: $f(p \rightarrow s)$ = factor by which the probability to capture s increases after p has been captured
- sd: $f(s \rightarrow d)$ = factor by which the probability to capture d increases after s has been captured
- ds: $f(d \rightarrow s)$ = factor by which the probability to capture s increases after d has been captured
- pd: $f(p \rightarrow d)$ = factor by which the probability to capture d increases after p has been captured
- dp: $f(d \rightarrow p)$ = factor by which the probability to capture p increases after d has been captured

The final vector to base the capture-recapture calculations on is named "v1.sav". The structure of this vector is self-explaining.

8.2 Example for simple 2-sample-capture-recapture-analyses

| | | police t95 | | |
|---------------------|---------|-------------|--------------------|-------------|
| | | present | absent | total |
| substitution t95 | present | a = 405 | b = 2,217 | a+b = 2,622 |
| | absent | c = 2,293 | $N_{hid}^{32} = ?$ | |
| | total | a+c = 2,698 | | |

estimate of hidden population: $N_{hid} = b \cdot c/a = 2,217 \cdot 2,293/405 = 12,552$

total population = $N_{tot} = N_{obs} + N_{hid} = 4,915 + 12,552 = 17,467$

variance of N_{tot} $Var_N = (a+b) \cdot (a+c) \cdot b \cdot c/a^3 = 2,622 \cdot 2,698 \cdot 2,217 \cdot 2,298 / 405^3 = 54,1351$

95%-confidence interval $ci_{95\%} = N_{tot} \pm z_{0.975} \cdot Sd_N = 17,467 \pm 1.96 \cdot \sqrt{54,1351} = [16,025; 18,909]$

8.3 SPSS programme code for a 3-sample-capture-recapture-analysis

SPSS 6.1.3 Example relating to Tab. 7 and Tab. 8 (p. 19). Bold type face indicates SPSS-commands, normal type face indicates relevant printout by SPSS. Italic plus bold type face indicates comments and calculations performed by the authors.

```
data list free /P S D freq cw.
```

32 For notation see chapter 2.6.1.

```

var labels P 'police t95'
        /S 'substitution t95'
        /D 'drug related death t95'
        /freq 'number per cell' .
value labels P S D 1 "present"
                2 "absent" .

```

The Code "1" represents "present", code "2" represents "absent".

```

begin data.
1 1 1 4 1
1 1 2 401 1
1 2 1 40 1
1 2 2 2253 1
2 1 1 28 1
2 1 2 2189 1
2 2 1 127 1
2 2 2 0 0
end data.
weight by freq.

```

crosstabs d by s by p.

D drug related death t95 by S substitution t95

Controlling for P police t95 Value = 1.00 present

| | | S | | Row Total |
|--------|---------|---------|--------|--------------|
| Count | | present | absent | |
| D | present | 1.00 | 2.00 | 44 |
| | absent | 401 | 2253 | 2654 |
| Column | | 405 | 2293 | 2698 |
| Total | | 15.0 | 85.0 | 100.0 |

Controlling for P police t95 Value = 2.00 absent

| | | S | | Row Total |
|--------|---------|---------|--------|--------------|
| Count | | present | absent | |
| D | present | 1.00 | 2.00 | 155 |
| | absent | 2189 | 0 | 2189 |
| Column | | 2217 | 127 | 2344 |
| Total | | 94.6 | 5.4 | 100.0 |

Hiloglinear d(1,2) s(1,2) p(1,2) /method backwards /CRITERIA=ITERATE(1000) /cweight cw /print freq resid /maxorder=2.

***** HIERARCHICAL LOG LINEAR *****
FACTOR Information

Factor Level Label
D 2 drug related death t95
S 2 substitution t95
P 2 police t95

***** HIERARCHICAL LOG LINEAR *****

Backward Elimination (p = .050) for DESIGN 1 with generating class

D*S
D*P
S*P

Likelihood ratio chi square = .42053 DF = 1 P = .517

| If Deleted Simple Effect is | DF | L.R. | Chisq Change | Prob | Iter |
|-----------------------------|----|------|--------------|--------|------|
| D*S | 1 | | 1.049 | .3058 | 13 |
| D*P | 1 | | .000 | 1.0000 | 10 |

Step 1

The best model has generating class

D*S
S*P

Likelihood ratio chi square = .28633 DF = 2 P = .867

| If Deleted Simple Effect is | DF | L.R. | Chisq Change | Prob | Iter |
|-----------------------------|----|------|--------------|-------|------|
| D*S | 1 | | 2.288 | .1304 | 9 |
| S*P | 1 | | 7.561 | .0060 | 16 |

Step 2

The best model has generating class

S*P
D

Likelihood ratio chi square = 2.57440 DF = 3 P = .462

| If Deleted Simple Effect is | DF | L.R. | Chisq Change | Prob | Iter |
|-----------------------------|----|------|--------------|-------|------|
| S*P | 1 | | 5.322 | .0211 | 16 |
| D | 1 | | 6062.504 | .0000 | 2 |

Step 3

The best model has generating class

S*P
D

Likelihood ratio chi square = 2.57440 DF = 3 P = .462

***** HIERARCHICAL LOG LINEAR *****

The final model has generating class

S*P
D

The suggested best model includes the interaction term "S*P"

The Iterative Proportional Fit algorithm converged at iteration 0.

The maximum difference between observed and fitted marginal totals is 1.777

and the convergence criterion is 2.253

Goodness-of-fit test statistics

| | | | |
|-------------------------------|---------|---------------------|----------|
| Likelihood ratio chi square = | 2.57440 | DF (UNADJUSTED) = 3 | P = .462 |
| | | DF (ADJUSTED) = 2 | P = .276 |
| Pearson chi square = | 2.49579 | DF (UNADJUSTED) = 3 | P = .476 |
| | | DF (ADJUSTED) = 2 | P = .287 |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design d s p.

GENERALIZED LOGLINEAR ANALYSIS

Model: Poisson
 Design: Constant + D + S + P
 Maximum number of iterations: 1000
 Relative difference tolerance: .001
 Final relative difference: .0002
 Maximum likelihood estimation converged at iteration 3.

Goodness-of-fit Statistics

| | Chi-Square | DF | Sig. |
|------------------|------------|----|-------|
| Likelihood Ratio | 7.8911 | 3 | .0483 |
| Pearson | 8.6731 | 3 | .0340 |

A Pearson Chi-square value of 0.0340 indicates that the model without any interaction terms deviates significantly from the observed model at the 5% level.

| Parameter | Estimate | SE | Z-value | Asymptotic 95% CI | |
|-----------|----------|-------|---------|-------------------|-------|
| | | | | Lower | Upper |
| 1 | 9.3787 | .0542 | 173.07 | 9.27 | 9.48 |

The first parameter representing the constant term is relevant for the purpose of estimating the missing cell. The estimate as well as the lower and upper confidence bounds are logarithms of the numbers we are interested in. The hidden population $N_{hid} = \exp(9,3787) = 11834$, the lower confidence limit is $\exp(9,27) = 10615$, the upper confidence limit is $\exp(9,48) = 13095$. If we are interested in the total number N_{tot} we have to add the observed cases N_{obs} of 5042 cases to all three numbers arriving at $N_{tot} = 11834 + 5042 = 15657$ and a 95%-confidence interval ranging from $10615 + 5042 = 15657$ through $13095 + 5042 = 18137$.

| | | | | | |
|---|---------|-------|--------|-------|-------|
| 2 | -4.4284 | .0813 | -54.49 | -4.59 | -4.27 |
| 3 | .0000 | . | . | . | . |
| 4 | -1.6930 | .0503 | -33.63 | -1.79 | -1.59 |
| 5 | .0000 | . | . | . | . |
| 6 | -1.6591 | .0505 | -32.88 | -1.76 | -1.56 |
| 7 | .0000 | . | . | . | . |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design d s*p.

| | Chi-Square | DF | Sig. |
|------------------|------------|----|-------|
| Likelihood Ratio | 2.5484 | 2 | .2797 |
| Pearson | 2.5067 | 2 | .2855 |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 9.0528 | .1482 | 61.08 | 8.76 | 9.34 |

$N_{tot} = \exp(9.0528)+5042 = 13584$ ci: $\exp(8.76)+5042 = 11416$ through $\exp(9.34)+5042 = 16426$

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design p s*d.

| | Chi-Square | DF | Sig. |
|------------------|------------|----|-------|
| Likelihood Ratio | 7.8437 | 2 | .0198 |
| Pearson | 8.6804 | 2 | .0130 |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 9.3816 | .0558 | 168.02 | 9.27 | 9.49 |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design s p*d.

| | Chi-Square | DF | Sig. |
|------------------|------------|----|-------|
| Likelihood Ratio | 2.4091 | 2 | .2998 |
| Pearson | 2.2793 | 2 | .3199 |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 9.4120 | .0564 | 166.92 | 9.30 | 9.52 |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design s*d s*p.

| | Chi-Square | DF | Sig. |
|------------------|------------|----|-------|
| Likelihood Ratio | .2276 | 1 | .6333 |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 8.8753 | .1825 | 48.62 | 8.52 | 9.23 |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design s*p p*d.

| Chi-Square | DF | Sig. |
|------------------|--------|-------|
| Likelihood Ratio | 1.3950 | .2376 |
| Pearson | 1.2288 | .2676 |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 9.2032 | .2099 | 43.85 | 8.79 | 9.61 |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design s*d p*d.

| Chi-Square | DF | Sig. |
|------------------|--------|-------|
| Likelihood Ratio | 2.2755 | .1314 |
| Pearson | 2.0451 | .1527 |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 9.4173 | .0583 | 161.64 | 9.30 | 9.53 |

Genlog d s p /cstructure=cw /model=poisson /print estim /plot none /design s*d s*p p*d.

| Chi-Square | DF | Sig. |
|------------------|-------|------|
| Likelihood Ratio | .0000 | . |
| Pearson | .0000 | . |

| Parameter | Estimate | SE | Z-value | Lower | Upper |
|-----------|----------|-------|---------|-------|-------|
| 1 | 8.7174 | .5415 | 16.10 | 7.66 | 9.78 |

8.4 SPSS programme code for the matrix procedure to construct data according to assumptions

*The programme models three capturing processes S, P, and D.

```
input program .
+ loop #i=1 to 16 .
- comp xspd=0 .
- comp xsp=0 .
- comp xsd=0 .
- comp xs=0 .
- comp xpd=0 .
- comp xp=0 .
- comp xd=0 .
- comp x=0 .
- comp yspd=0 .
- comp ysp=0 .
- comp ysd=0 .
- comp ys=0 .
- comp ypd=0 .
- comp yp=0 .
- comp yd=0 .
- comp y=0 .
if (#i eq 1) xspd=1 .
if (#i eq 2) xsp=1 .
if (#i eq 3) xsd=1 .
if (#i eq 4) xs=1 .
if (#i eq 5) xpd=1 .
if (#i eq 6) xp=1 .
if (#i eq 7) xd=1 .
if (#i eq 8) x=1 .
if (#i eq 9) yspd=1 .
if (#i eq 10) ysp=1 .
if (#i eq 11) ysd=1 .
if (#i eq 12) ys=1 .
```

```
if (#i eq 13) ypd=1 .
if (#i eq 14) yp=1 .
if (#i eq 15) yd=1 .
if (#i eq 16) y=1 .

***** definition section: enter your parameters here.
- comp n = 100000 .
- comp t = 1000 .
- comp sprob = 0.05 .
- comp pprob = 0.10 .
- comp dprob = 0.15 .
- comp lprob = 0.00 .
- comp sp = 2 .
- comp ps = 2 .
- comp sd = 2.5 .
- comp ds = 2.5 .
- comp pd = 3 .
- comp dp = 3 .
***** end of definition section.

- comp s=1-(1-sprob)**(1/t) .
- comp p=1-(1-pprob)**(1/t) .
- comp d=1-(1-dprob)**(1/t) .
- comp l=lprob/t .
- end case .
+ end loop .
+ end file .
end input program .
form xspd to y (f12.10) n t (f10.0)
  sprob pprob dprob lprob (f6.3)
  sp ps sd ds pd dp (f6.3) logconst(f9.4).
exec .
list variables n t sprob pprob dprob lprob sp ps sd ds pd dp logconst/cases 1 .
save outfile "c:\null.sav".
exec.

get file "c:\null.sav".
exec.
*transition matrix if s is capturing .
*xspd -> s .
*xsp -> s .
*xsd -> s .
*xs -> s .
*xpd -> s .
if ($casenum eq 5) xspd=s*ps*ds .
if ($casenum eq 5) xpd=1-s*ps*ds .
*xp -> s .
if ($casenum eq 6) xsp=s*ps .
if ($casenum eq 6) xp=1-s*ps .
*xd -> s .
if ($casenum eq 7) xsd=s*ds .
if ($casenum eq 7) xd=1-s*ds .
*x -> s .
if ($casenum eq 8) xs=s .
if ($casenum eq 8) x=1-s .
exec.
save outfile "c:\us.sav" /keep xspd to y.

get file "c:\null.sav".
exec.
*transition matrix if p is capturing .
*xspd -> p .
```

```
*xsp -> p .
*xsd -> p .
if ($casenum eq 3) xspd=p*sp*dp .
if ($casenum eq 3) xsd=1-p*sp*dp .
*xs -> p .
if ($casenum eq 4) xsp=p*sp .
if ($casenum eq 4) xs=1-p*sp .
*xpd -> p .
*xp -> p .
*xd -> p .
if ($casenum eq 7) xpd=p*dp .
if ($casenum eq 7) xd=1-p*dp .
*x -> p .
if ($casenum eq 8) xp=p .
if ($casenum eq 8) x=1-p .
exec.
save outfile "c:\up.sav" /keep xspd to y.
```

```
get file "c:\null.sav".
exec.
*transition matrix if d is capturing .
*xspd -> d .
*xsp -> d .
if ($casenum eq 2) xspd=d*sd*pd .
if ($casenum eq 2) xsp=1-d*sd*pd .
*xsd -> d .
*xs -> d .
if ($casenum eq 4) xsd=d*sd .
if ($casenum eq 4) xs=1-d*sd .
*xpd -> d .
*xp -> d .
if ($casenum eq 6) xpd=d*pd .
if ($casenum eq 6) xp=1-d*pd .
*xd -> d .
*x -> d .
if ($casenum eq 8) xd=d .
if ($casenum eq 8) x=1-d .
exec.
save outfile "c:\ud.sav" /keep xspd to y.
```

```
get file "c:\null.sav".
exec.
*transition matrix for cases are lost and replaced.
*xspd -> 1 .
if ($casenum eq 1) xspd=1-1 .
if ($casenum eq 1) x=1 .
if ($casenum eq 1) yspd=1 .
if ($casenum eq 1) x=1 .
*xsp -> 1 .
if ($casenum eq 2) xsp=1-1 .
if ($casenum eq 2) ysp=1 .
if ($casenum eq 2) x=1 .
*xsd -> 1 .
if ($casenum eq 3) xsd=1-1 .
if ($casenum eq 3) ysd=1 .
if ($casenum eq 3) x=1 .
*xs -> 1 .
if ($casenum eq 4) xs=1-1 .
if ($casenum eq 4) ys=1 .
if ($casenum eq 4) x=1 .
*xpd -> 1 .
if ($casenum eq 5) xpd=1-1 .
```

```
if ($casenum eq 5) ypd=1 .
if ($casenum eq 5) x=1 .
*xp -> 1 .
if ($casenum eq 6) xp=1-1 .
if ($casenum eq 6) yp=1 .
if ($casenum eq 6) x=1 .
*xd -> 1 .
if ($casenum eq 7) xd=1-1 .
if ($casenum eq 7) yd=1 .
if ($casenum eq 7) x=1 .
*x -> 1 .
if ($casenum eq 8) x=1 .
if ($casenum eq 8) y=1 .
exec.
save outfile "c:\ul.sav" /keep xspd to y.

get file "c:\null.sav".
exec.
sele if ($casenum eq 1) .
exec .

save outfile "c:\v0.sav"/keep xsp xsd xs xpd xp xd x n yspd ysp ysd ys ypd yp yd y.
exec .
save outfile "c:\tt.sav"/keep t .
exec .
save outfile "c:\logconst.sav"/keep logconst .
exec .

matrix .
get v0 /file "c:\v0.sav" .
get us /file "c:\us.sav" .
get up /file "c:\up.sav" .
get ud /file "c:\ud.sav" .
get ul /file "c:\ul.sav" .
get tt /file "c:\tt.sav" .
get logconst /file "c:\logconst.sav" .

comp uspd1=us*up*ud*ul .
comp v1=v0*uspd1**tt .
comp v2=v1.
comp v1={v2(1)+v2(9),v2(2)+v2(10),v2(3)+v2(11),v2(4)+v2(12),
v2(5)+v2(13),v2(6)+v2(14),v2(7)+v2(15),v2(8)+v2(16)}.
comp tot0=msum(v0) .
comp tot1= v2(1)+v2(2)+v2(3)+v2(4)+v2(5)+v2(6)+v2(7)+v2(8).
comp estimhid=exp(logconst) .
comp tot1= v2(9)+v2(10)+v2(11)+v2(12)+v2(13)+v2(14)+v2(15)+v2(16).
comp estimtot=v1(1)+v1(2)+v1(3)+v1(4)+v1(5)+v1(6)+v1(7)+exp(logconst) .
comp oddssp =(v1(1)+v1(2))/(v1(3)+v1(4))/(v1(5)+v1(6))*(v1(7)+v1(8)) .
comp oddssd =(v1(1)+v1(3))/(v1(2)+v1(4))/(v1(5)+v1(7))*(v1(6)+v1(8)) .
comp oddspdg=(v1(1)+v1(5))/(v1(2)+v1(6))/(v1(3)+v1(7))*(v1(4)+v1(8)) .
comp oddsspg=(v1(1)+v1(2))/(v1(3)+v1(4))/(v1(5)+v1(6))*(v1(7)+estimhid) .
comp oddssdg=(v1(1)+v1(3))/(v1(2)+v1(4))/(v1(5)+v1(7))*(v1(6)+estimhid) .
comp oddspdg=(v1(1)+v1(5))/(v1(2)+v1(6))/(v1(3)+v1(7))*(v1(4)+estimhid) .
comp props=(v1(1)+v1(2)+v1(3)+v1(4))/tot0 .
comp propp=(v1(1)+v1(2)+v1(5)+v1(6))/tot0 .
comp propd=(v1(1)+v1(3)+v1(5)+v1(7))/tot0 .
comp proppg=(v1(1)+v1(2)+v1(5)+v1(6))/estimtot .
comp propdg=(v1(1)+v1(3)+v1(5)+v1(7))/estimtot .
comp discr=estimtot/tot0 .
save v1 /outfile "c:\v1.sav" .
end matrix .
```

9 References

- Bello, P.Y. (1997): Estimates of Opiate Use in Toulouse Conurbation in 1995. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- Bishop, Y.M.M.; Fienberg, S.E.; Holland, P.W. (1984): Discrete Multivariate Analysis Theory and Practice. MIT Press, Cambridge
- Bühringer, G.; Adelsberger, F.; Heinemann, A.; Kirschner, J.; Knauß, I.; Kraus, L.; Püschel, K.; Simon, R. (1997): Schätzverfahren und Schätzungen 1997 zum Umfang der Drogenproblematik in Deutschland. Sucht, 43, Sonderheft 2
- Chao, A. (1989): Estimating population size for sparse data in capture-recapture experiments. Biometrics, 45, 427-438
- Comiskey, C. (1997): Estimating The Prevalence Of Opiate Drug Use In Dublin, Ireland During 1996. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- D'Ippoliti, D. (1997): Prevalence Estimation of Drug Addiction in Rome. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- EMCCDA (1997): Estimating the Prevalence of Problem Drug Use in Europe. EMCCDA Scientific Monograph , Series N 1, Lisbon
- Freire, S.; Moreira, M. (1997): Estimating the Number of Opiate Use in Setúbal. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- Ghodse, A.; Sheehan, M.; Taylor, C.; Edwards, G. (1985): Death of drug addicts in the United Kingdom, 1967-1981. British Medical Journal, 290, 425-428
- Haas, S.; Guzei, K.; Türscherl, E. (1999): Bericht zur Drogensituation 1999. ÖBIG, Wien
- Hay, G.; McKeganey, N.; Birks, E. (1997): Methodological Pilot Study of Local Level Prevalence Estimates – CT.96.EP.07. European Monitoring Centre for Drugs and Drug Addiction (EMCCDA), Lisbon
- Hay, G.; McKeganey, N.; Wiessing, L.; Hartnoll, R.; Bello, P.Y.; S'Ippoliti, D.; Domingo-Salvany, A.; Frischer, M.; Kraus, L.; Smit, F. (1998): Methodological Guidelines to Estimate the Prevalence of Problem Drug Use on the Local Level. EMCCDA, Lisbon
- Kaukonen, O.; Kinnunen, A.; Partanen, P.; Seppälä, T.; Simpura, J.; Virtanen, A. (1997): Estimating the Number of Amphetamine and Opiate Users in the Greater Helsinki Area. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- Korf, D.J. (1997): Comparison of Different Estimation Methods in the Netherlands. in EMCCDA: Estimating the Prevalence of Problem Drug Use in Europe. EMCCDA Scientific Monograph , Series N 1, Lisbon
- Kraus, L.; Augustin, R.; Kümmler, P.; Wiessing, L.; Hartnoll, R.: (1999) Study to Obtain Comparable National Estimates of Problem Drug Use Prevalence for all EU Member States EMCCDA, Lisbon
- Seidler, D.; Uhl, A. (1997): Estimating the Number of Opiate Users in Vienna. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- Simon, R.; Kraus, L.; Bauernfeind, R. (1998): National Prevalence Estimates – Improvement of Comparability of National Estimates of Addiction Prevalence – Final Report. EMCCDA, Lisbon

- Smit, F.; Toet, J. (1997): Estimating the Number of Opiate Users in Rotterdam Using Statistical Models for Incomplete Count Data. in: Hay, G.; McKeganey, N.; Birks, E. (eds.): Methodological Pilot Study of Local Level Prevalence Estimates. EMCCDA, Lisbon
- Uhl, A.; Kopf, N.; Springer, A.; Eisenbach-Stangl, I.; Beiglböck, W.; Preinsberger, W.; Mader, R. (1999): Handbuch: Alkohol – Österreich: Zahlen, Daten, Fakten, Trends 1999. BMAGS, Wien
- Uhl, A.; Springer, A. (1996): Studie über den Konsum von Alkohol und psychoaktiven Stoffen in Österreich unter Berücksichtigung problematischer Gebrauchsmuster – Repräsentativerhebung 1993/94 Textband. Originalarbeiten, Studien, Forschungsberichte des BMGK, Wien
- Zelterman, D. (1988): Robust estimation in truncated discrete distributions with applications to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18, 225-237