Dichomisation of data: why it may not be such a good idea.

Gabriela Czanner PhD CStat
Department of Biostatistics
Department of Eye and Vision Science

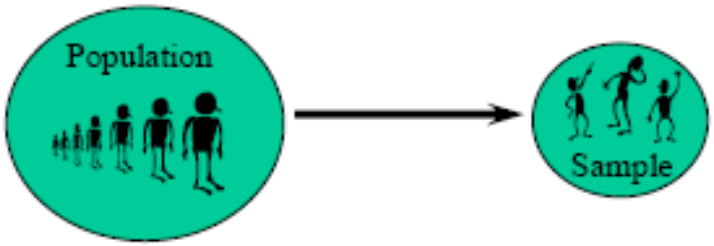
5 March 2014

MERSEY POSTGRADUATE TRAINING PROGRAMME

Workshop Series: Basic Statistics for Eye Researchers and Clinicians

Dichotomisation

IOP [mmHg]	IOP dichotomised
12.2	Low
13.5	Low
16.2	High
18.8	High
19.0	High



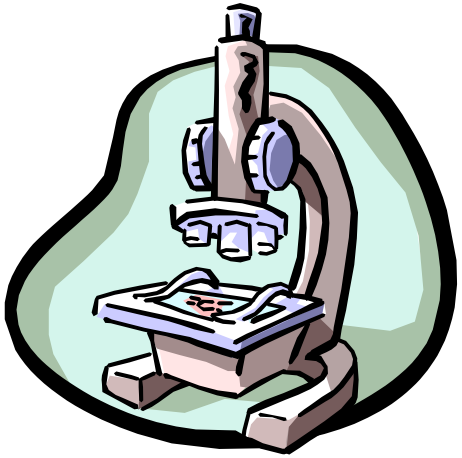
- Dichotomisation = the replacement of the original measured data with two values (e.g. High and Low)
- We often are tempted to dichotomise
  - Examples: Systolic blood pressure: High and Low
- But, is it reasonable to dichotomise?

  The answer is not straightforward.

  It depends on the goal of our research.

Outline

- Part 1: Dichotomisation of outcome measure leads to loss of information and power.
- Part 2: Dichotomisation of the confounding variable leads to biased spurious results.
- Summary: General recommendations.



Main Reference Paper

Title: "Ophthalmic Statistics Note: The perils of dichotomising continuous variables"

- British Journal of Ophthalmology
  - 3rd paper in new series in Education
  - Accepted February 2014
- Authors
  - Cumberland P, Czanner G, Bunce C, Dore CJ, Freemantle N and Garcia-Finana M. On behalf of the Ophthalmic Statistics Group

Part1.

Dichotomisation of outcome measure
leads to loss of information and power

Part1: Consequences of dichotomisation of outcome measure

Consequence 1: loss of descriptive information on the study population

Example: *the nature and extent of differences between individuals is lost when visual acuity is dichotomised as having/not having low vision*

Subjects with similar outcome measures but on either side of the threshold will be described and analysed as different whilst two subjects with values that are on the same side of the threshold, but one near and another a long way from the threshold, will be treated as if they are the same.

Part 1: Consequences of dichotomisation of outcome measure

Consequence 2: impossible to quantify linear relationships

Example: *It is not possible to quantify the change in mmHg of IOP per mmHG of systolic blood pressure (SBP) increase if IOP has been dichotomised.*

Consequence 3: loss of statistical power

Example: see next slides…

Reminder: **Statistical power** = the probability of detecting a true effect of a particular size should it exist.

Example. Loss of statistical power if outcome measure is dichotomised.
Assumptions of our population of patients 60+yrs old: IOP and SBP follow a normal distribution with means 14.5 mmHg and 135 mmHg, and standard deviations 2.4mmHg and 20 mmHg, respectively. Also we assumed a linear change of 0.035 mmHg in IOP per mmHg of SBP.

We conduct a study: The sample size required to detect a significant association (correlation) between IOP and SBP, at the 5% significance level.

	IOP as a Continuous variable	IOP as a Binary variable		
Power to detect association	n_o	cutpoint	n_d	power if $n=n_o$
90%	119	14.5 mmHg	175	73%
		16 mmHg	207	67%
		13 mmHg	212	67%
80%	90	14.5 mmHg	133	61%
		16 mmHg	161	55%
		13 mmHg	162	54%

Example for Problem 4. Loss of statistical power when outcome measure is dichotomised (continued)

Summary:

When IOP is dichotomised, a larger sample size (n_d) is needed to detect a significant association whilst maintaining the same power as an analysis with sample size n_o using IOP as a continuous variable.

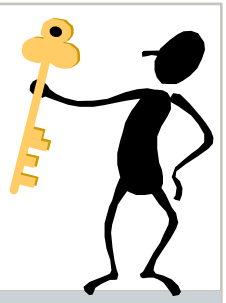
For example, when IOP is analysed as continuous, the sample size required is 119 individuals for a power of 90%.

If IOP is dichotomised using the mean as the cutpoint (14.5 mmHg), then the sample size required to maintain 90% power increases to 175 individuals; 56 additional patients.

If the condition of interest is rare, this increase in the required number of patients might render a study infeasible.

Alternatively, a reduction in power of at least 15% would occur if the sample size remains at n_o=119 and IOP was dichotomised.

Points for consideration



- To dichotomize outcome measure?
  - It is not good practice to power a study, obtain data from a number of patients and then after completing data collection to under-power the analysis by dichotomisation.
  - All decisions regarding cut-points for categorisation should be pre-specified before conducting the analysis, and reasons for such decisions stated when writing a paper.
  - Studies are not comparable if each is using different cutpoint.

Part 2.

Dichotomisation of confounding variables leads to bias in estimated associations.

Part 2: Consequences of dichotomisation of confounding variables

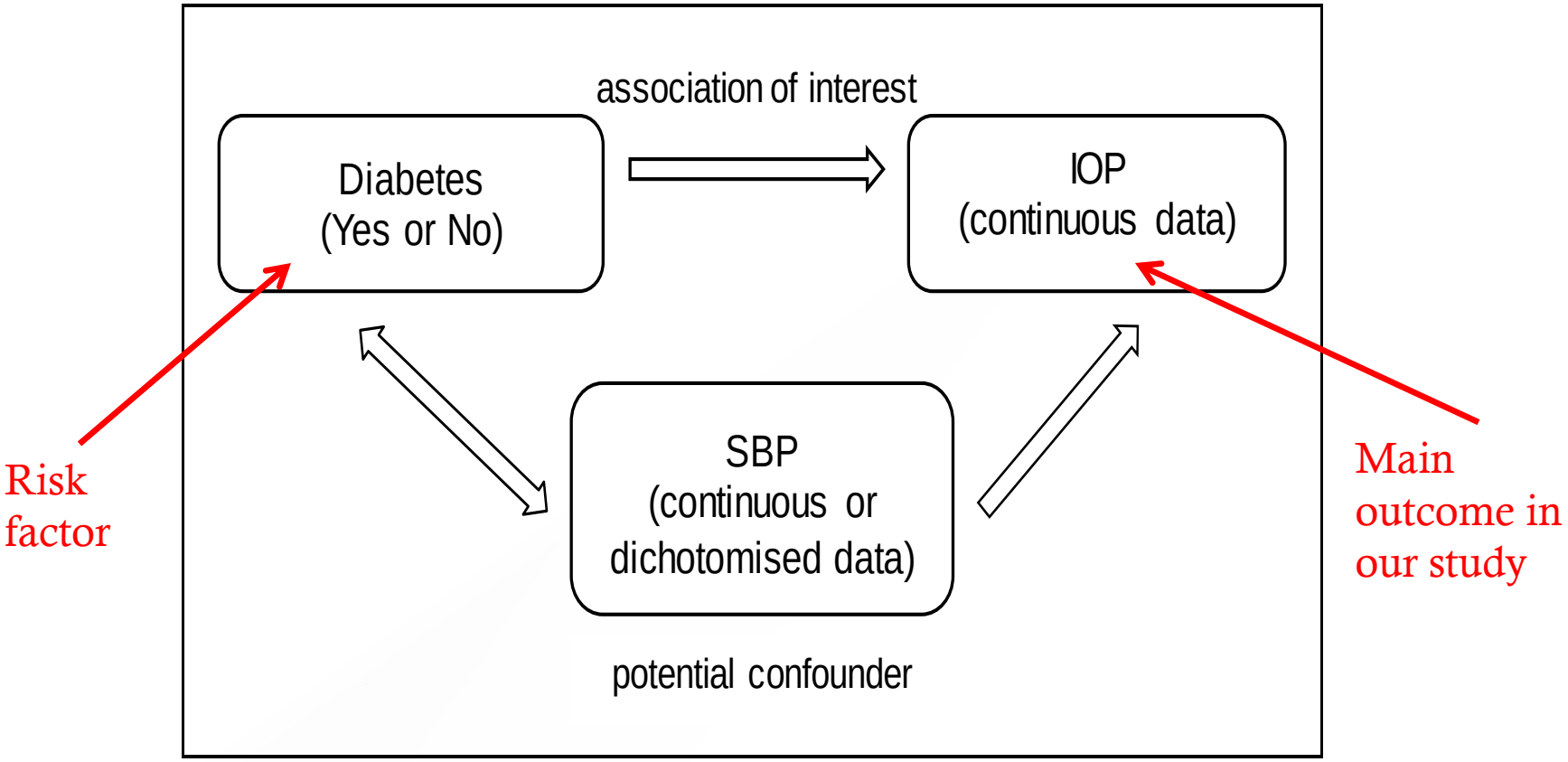
Example*:* We want to study outcome measure IOP, and we want to learn about risk factors for the IOP, e.g. we want to study association between IOP and. having diabetes.

Question: Can we just collect IOP and diabetes status and do the relevant statistical test (i.e. using IOP and diabetes only)?
Answer: No, we need to be aware of confounders (background factors)!

Confounder is a background factor (such as SBP, age or smoking habit) that is associated with the risk factor and with the outcome.

- If confounders are present the estimation of the association of interest between the risk factor and outcome can be **biased**.
- In **clinical trials** we can minimize this bias via **design**: we randomise patients to intervention and control groups to ensure that groups are balanced with regard to the background factors.
- In **epidemiological** and other clinical studies we minimize this bias by accounting for the confounders in the **data analysis**.

Example. Dichotomisation of confounding variable.



SBP is a potential confounder of the association between IOP and Diabetes.
We need to take SBP into account: via appropriate design or data analysis.

Example. Dichotomisation of confounding variable. (continues)

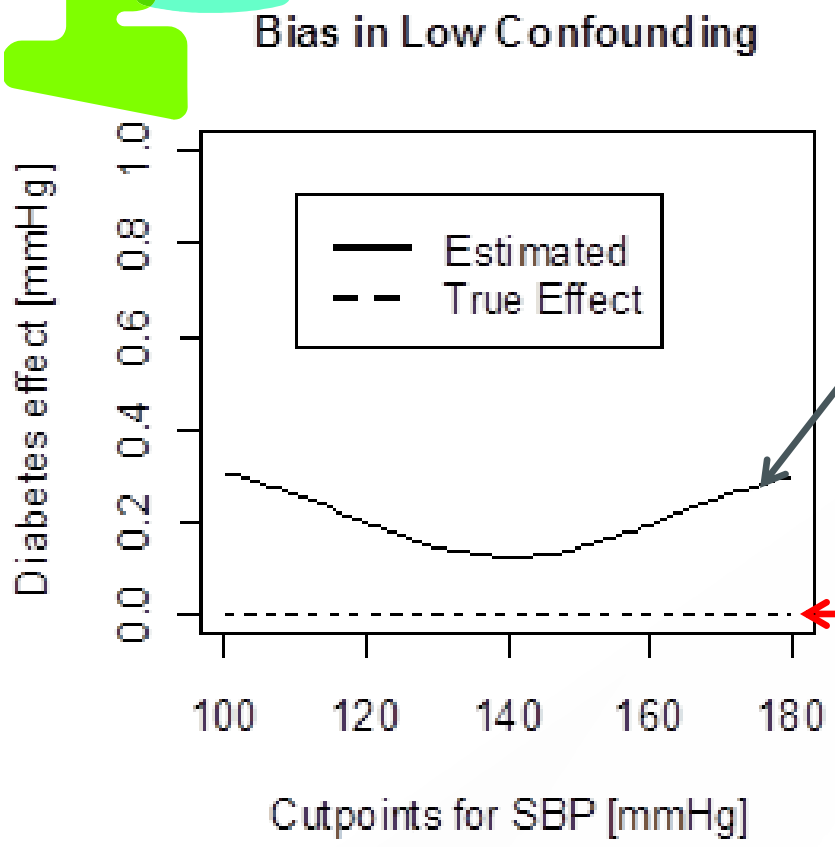
Let the truth about the population is the following:

- IOP [mmHg] data have normal distribution
- IOP increases on average by 0.035mmHg per 1mmHg increase in SBP
- IOP's mean is the same in those with and without diabetes
- SBP [mmHg] follows normal distribution
  - 2 scenarios
    - **Low confounding:** SBP means are 135 and 145mmHg for the non-diabetic and diabetic groups
    - **High confounding:** SBP means are 135 and 155 mmHg for the non-diabetic and diabetic groups

We are conducting a clinical study:
We want to investigate association between IOP and Diabetes. We are collecting data on IOP [mmHg], dichotomised SBP [Low/High] and Diabetes (yes/no).

Question: Will our data analysis truly conclude that there is NO association between IOP and diabetes?

Answer: The estimate of association between IOP and Diabetes will be biased.



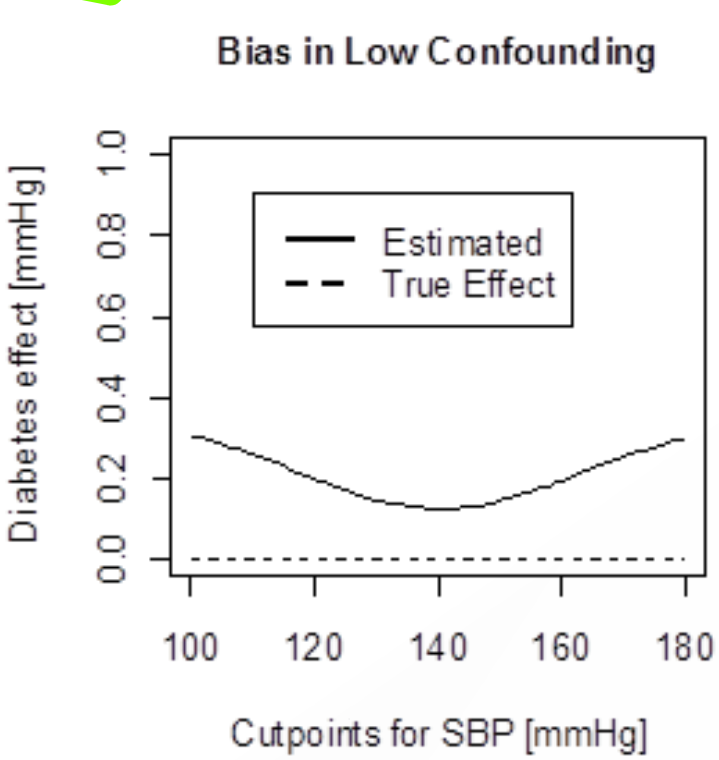
SBP means are 135 and 145 mmHg for the non-diabetic and diabetic groups

Here the confounder SBP is dichotomized, and we use e.g. ANOVA method to study how IOP changes across diabetes groups. This curve shows the means of the estimated changes of IOP between diabetes groups.

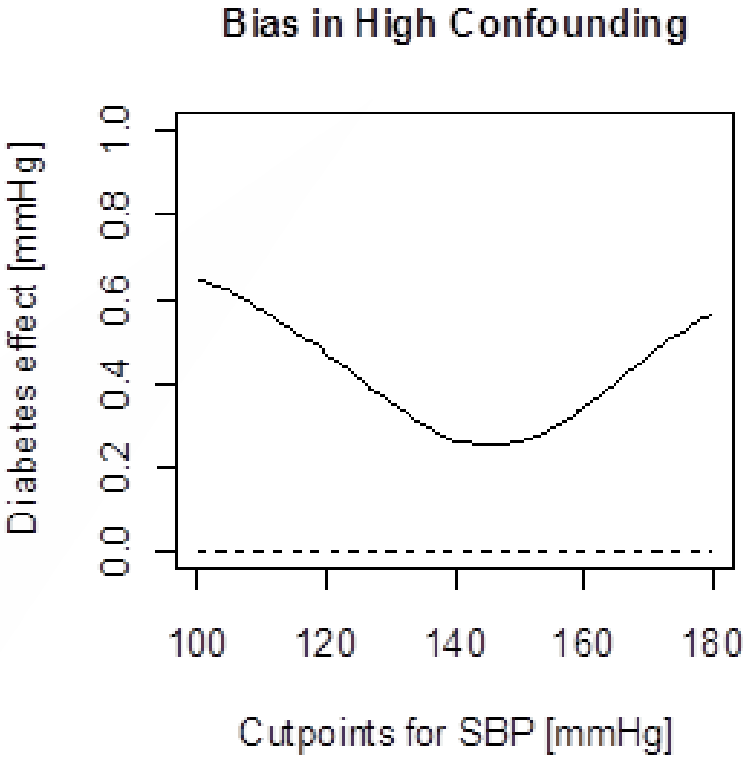
The truth is that there is no effect of diabetes on IOP i.e. IOP difference between diabetic and non-diabetic patients is 0mmHG.

Remember: the bias is a systematic difference between your estimate and the truth. It does not happen due a chance, and will not disappear in large studies.

The bias is larger in high confounding. The bias is smallest at the median SBP.

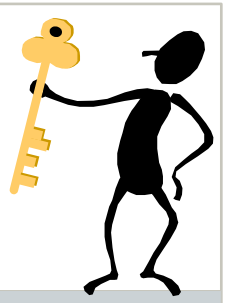


SBP means are 135 and 145 mmHg for the non-diabetic and diabetic groups

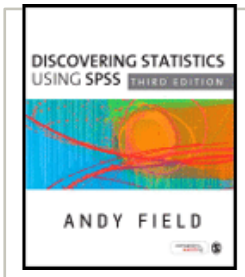


SBP means are 135 and 155 mmHg for the non-diabetic and diabetic groups

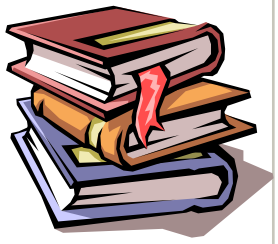
Points for consideration



- To dichotomise the confounding variables?
  - No, do try to avoid the dichotomisation.
  - Be aware of bias that may be introduced by dichotomisation of the continuous confounders.
  - Including the dichotomised confounder in analysis does remove some bias but not all. Magnitude of such bias depends on cutpoint.



General resources



Books

- Practical statistics for medical research by Douglas G. Altman
- Medical Statistics from Scratch by David Bowers

Journals' with series on how to do statistics in clinical research

- American Journal of Ophthalmology has **Series on Statistics**
- British Medical Journal has series **Statistics Notes**

Manual for SPSS statistical software: by Andy Field, Discovering statistics using SPSS

Workshops organized by Biostatistics Department, U of Liverpool

- http://www.liv.ac.uk/translational-medicine/departmentsandgroups/biostatistics/coursesandworkshops/
- Many workshops in month April 2014
- E.g.Validity and reliability of diagnostic tests and other methods of measurement (3 june)
- E.g. Statistical issues in design and analysis of research projects (7 april)

Thank you for your attention

These slides and worksheet can be found on: http://pcwww.liv.ac.uk/~czanner/

Planned future workshops:

- How to make sense of many measured characteristics? Multivariate stats methods
- Ideas are welcome!

Statistical Clinics for ophthalmic clinicians and researchers !
Run by appointment.
Email: czanner@liv.ac.uk
Phone: +44-151-706-4019
Further information: http://pcwww.liv.ac.uk/~czanner/