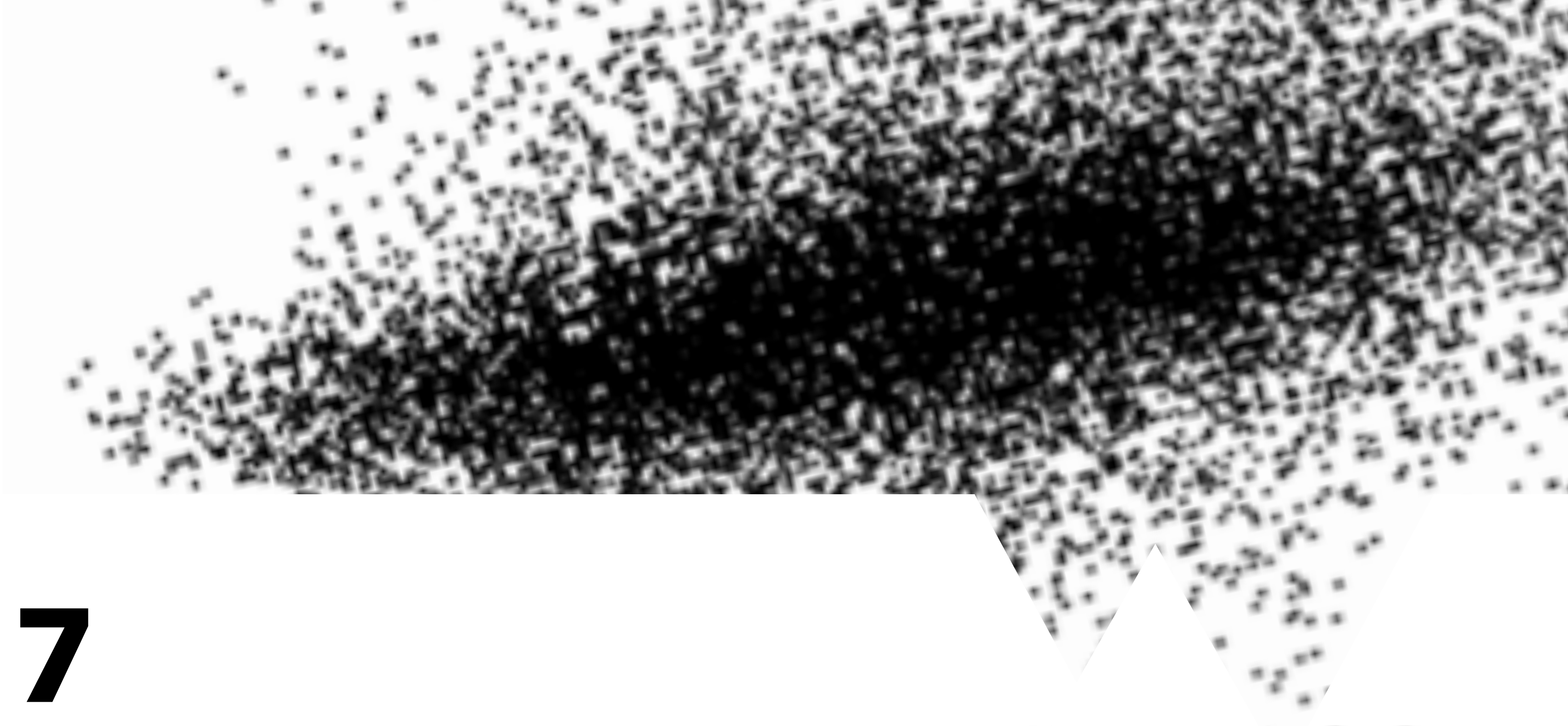


ST 117

4. Linear Regression

WARWICK

Lecture 22
(Week 8)

(Linear) prediction
Regression effect
Test-retest

Regression line (recall)

Regression line:

$$Y - \bar{Y} = r_{XY} \frac{SD_Y}{SD_X}(X - \bar{X})$$

$$Y - \bar{Y} = \frac{Cov(X,Y)}{Var(X)}(X - \bar{X})$$

Calculating the regression line from data:

Y = α+βX, where α and β are estimated by $\hat{y}_i = bx_i + a$

$$a = \bar{y} - b\bar{x} \qquad b = \frac{s_y}{s_x} r_{xy} = \frac{s_{xy}}{s_x^2}$$

Sample Covariance $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Sample Correlation $r_{xy} = \frac{s_{xy}}{s_x s_y}$

See also Theorem and corollary about MLE for the coefficients in handwritten notes

Prediction: y from x

$$y = 0.649x + 23.8$$

Suppose the average height of the parents is 72 inches. What do we predict for the height of the child?

$$y_{pred} = 0.649 \times 72 + 23.8 = 70.5$$

Prediction: both ways

$$y = 0.649x + 23.8$$

Suppose the average height of the parents is 72 inches.
What do we predict for the height of the child?

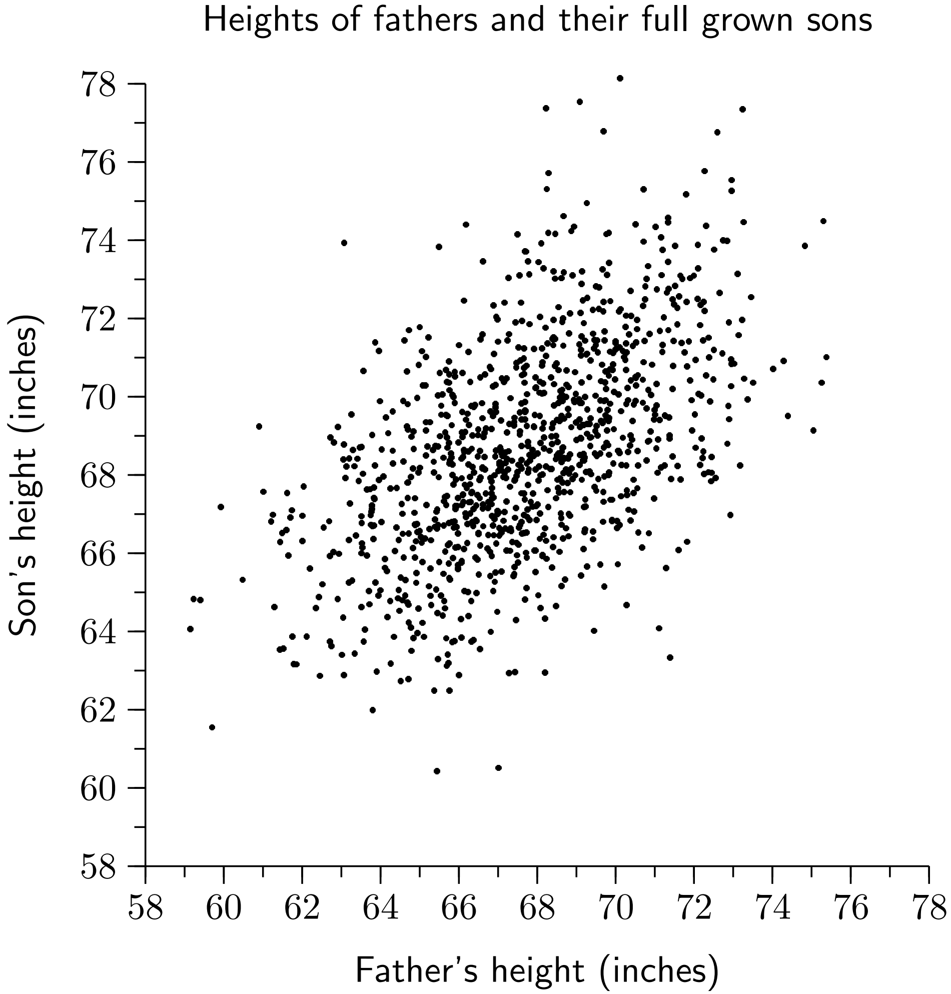
$$y_{pred} = 0.649 \times 72 + 23.8 = 70.5$$

Suppose the child's height is 70.5 inches.
What do we predict for the height of the parents?

$$b = \frac{s_{xy}}{s_y^2} = 0.326 \qquad a = \bar{x} - b\bar{y} = 45.9$$

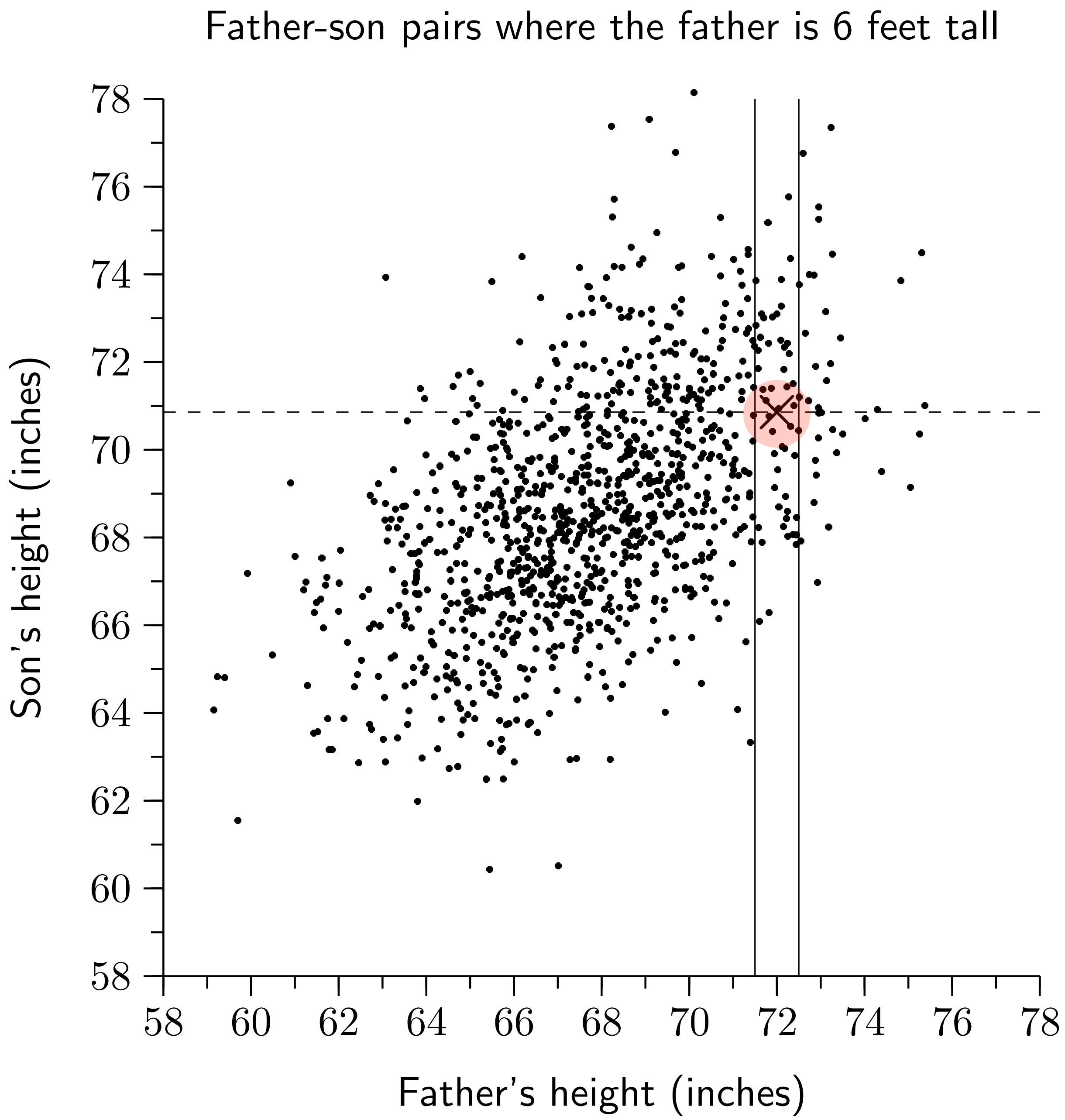
$$x_{pred} = 0.326 \times 70.5 + 45.9 = 68.9$$

Example: **Heights of fathers and sons**

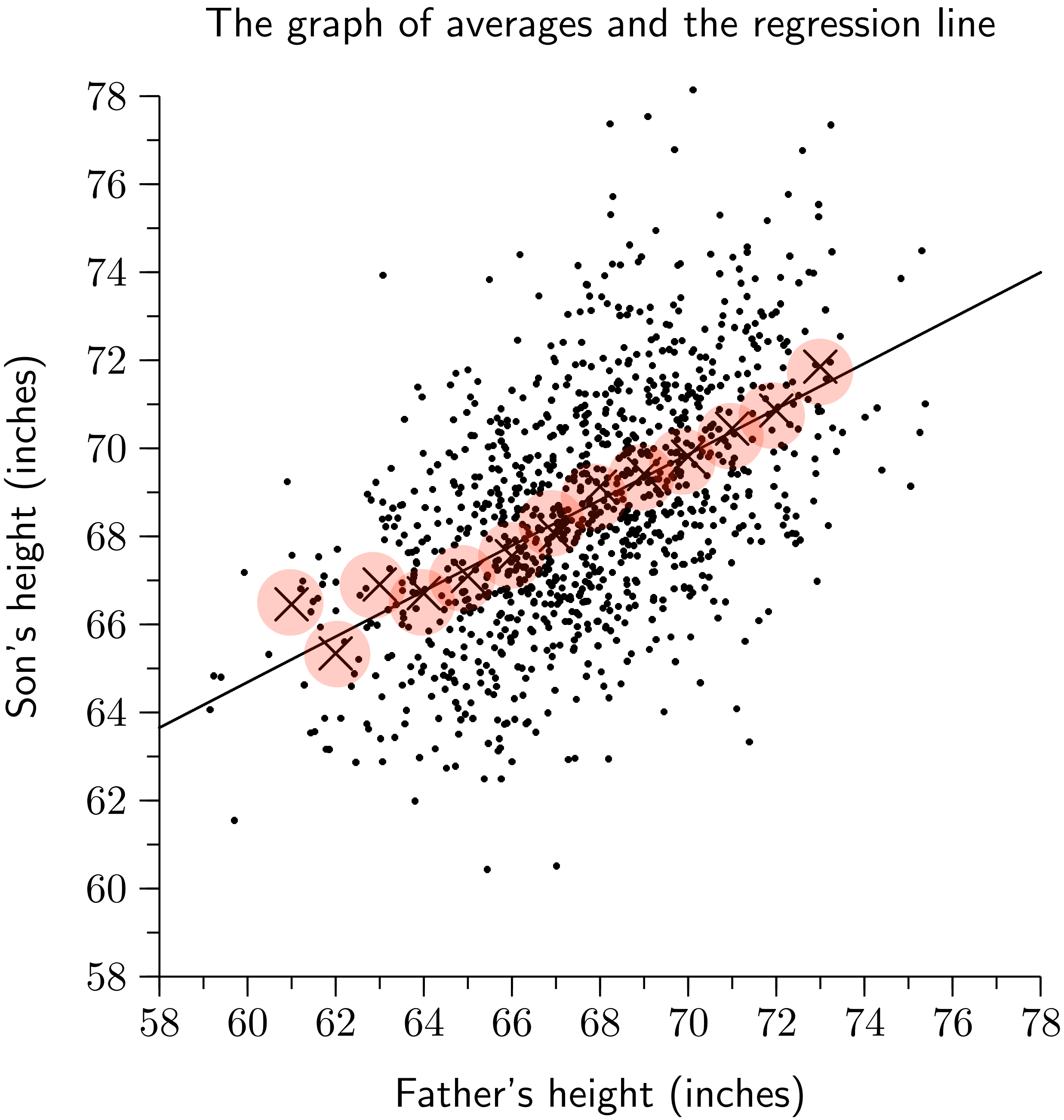


Historical data collection: Heights of 1,078 fathers and their full-grown sons, in England, circa 1900, Pearson and Lee 1903
Available for example at https://www.kaggle.com/datasets/abhilash04/fathersandsonheight
More comprehensive data collection including full families: https://vincentarelbundock.github.io/Rdatasets/doc/HistData/PearsonLee.html

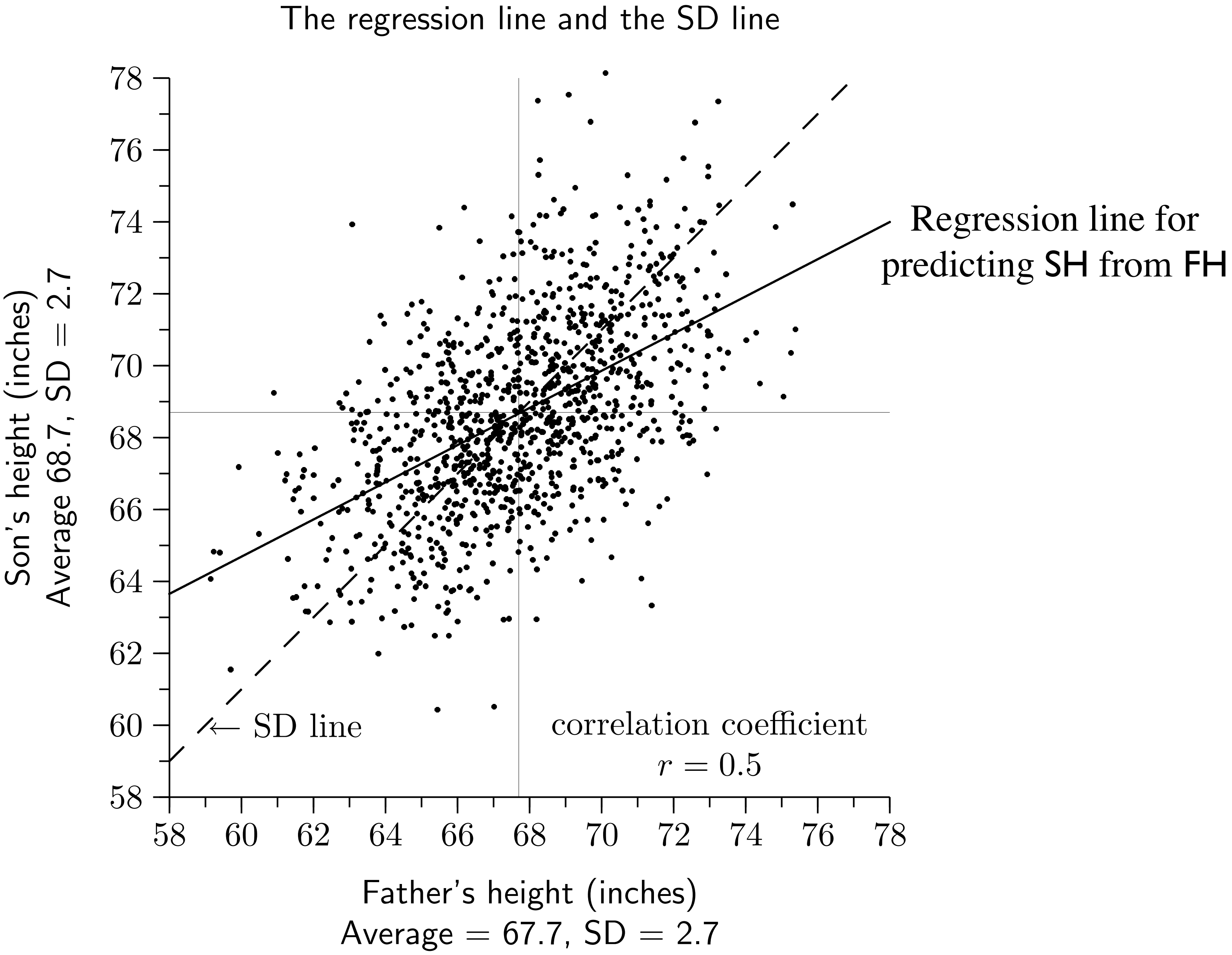
Local means


Father-son pairs where the father is 6 feet tall
Son's height (inches)
Father's height (inches)
58
60
62
64
66
68
70
72
74
76
78

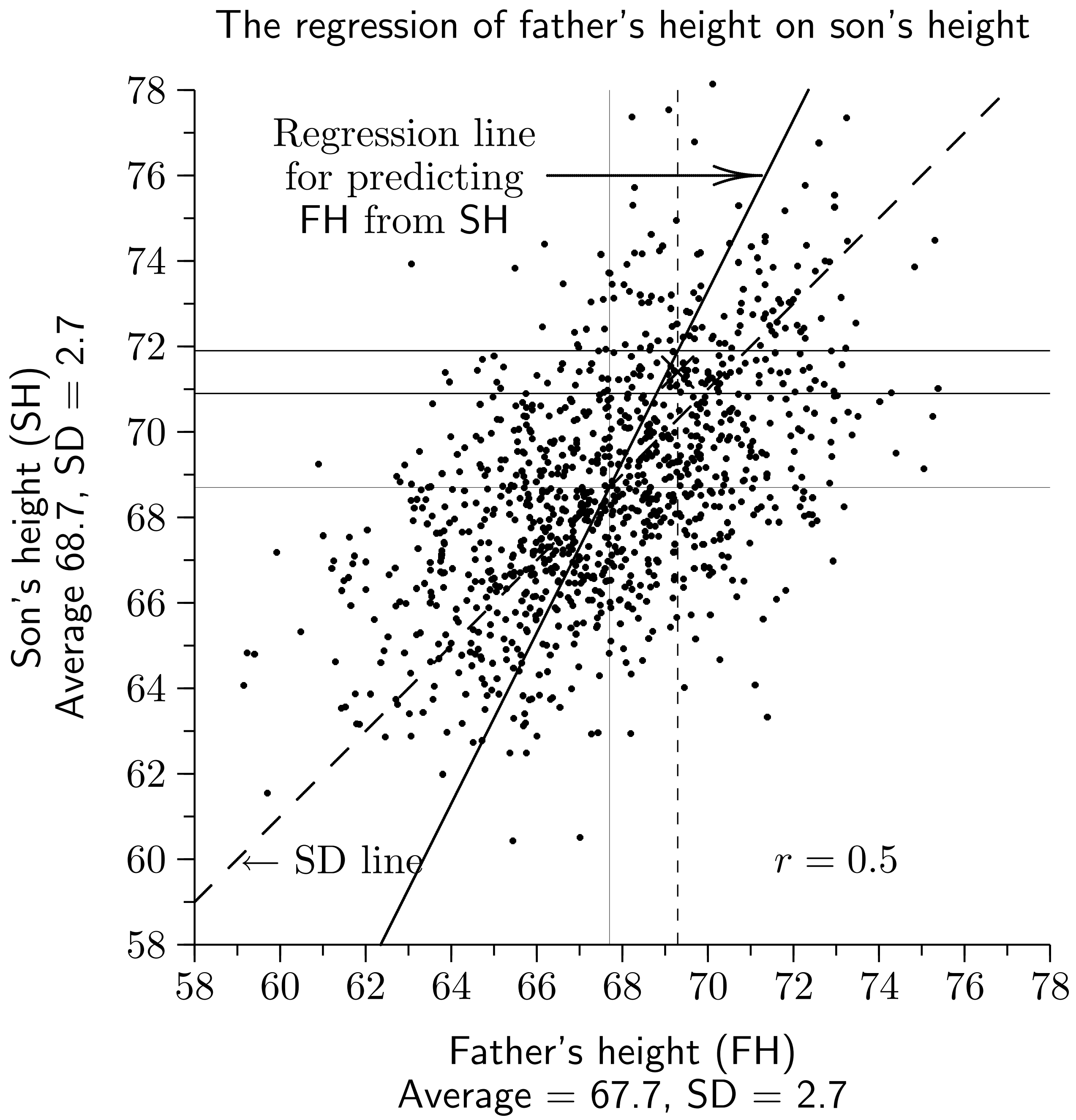
From local means to regression



Regression line and SD line



There are two regression lines



Phenomenon: **Regression to the mean**

Recall prediction:

$$y_{pred} = 0.649 \times \overset{\text{Parent's height}}{72} + 23.8 = \overset{\text{Child's height}}{70.5}$$

Observations in the father-son data:

- Sons of fathers who are taller than average are themselves taller than average, but by *not so much* as the fathers.
- Sons of fathers who are shorter than average are themselves shorter than average, but by *not so much* as the fathers.
- There is a move towards the mean.
- In Galton's terms: "*regression towards mediocrity*".

Why does this happen?

Why? First aspect

A person’s height depends on the heights of both the mother and father.

Very tall men do not generally have children with very tall women, because:

- factors other than height enter into the choice of a mate
- in terms of numbers, there are less very tall women than not very tall women

Illustration: A man who is 2 SDs taller than the average male may marry a women who is 2 SDs taller than the average female. But because attributes other than height matter, the wife is very likely to be a woman who is less than 2 SDs taller than average, just because there are so many more such women.

Thus, most very tall men do not have children with women who are also exceptionally tall. Since there is some aspect of heritability, have sons who are not as exceptionally tall either.

Note: Same arguments for mother, daughters, just different dataset.

Why? Second aspect

Diet, exercise, and other environmental factors influence height, so that observed height is not a perfect reflection of one's genes.

Someone who is very tall is much more likely to be the unusually tall result of what might be less exceptional genes.

Thus the observed height of a very tall person is usually an ***overstatement of his or her genetic height***, which is what determines the expected height of the child.

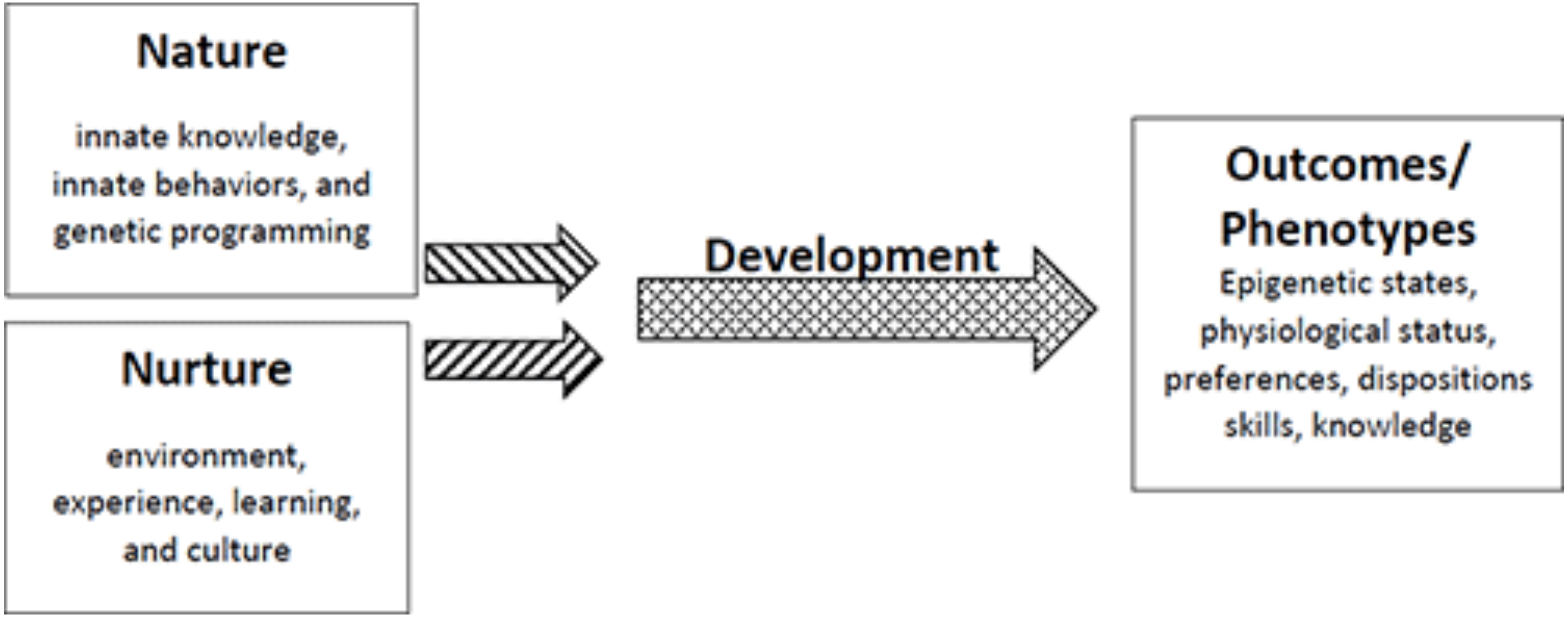
These are aspect of a wider topic (and debate) called: Nature versus Nurture

Sticking with height, an example for the impact of nutrition in the evolution of height of Dutch people in the 19th vs 20th century

Contextual information: **Nature vs Nurture**

Introduction e.g.

https://en.wikipedia.org/wiki/Nature_versus_nurture


Nature
innate knowledge, innate behaviors, and genetic programming
Nurture
environment, experience, learning, and culture
Development
Outcomes/ Phenotypes
Epigenetic states, physiological status, preferences, dispositions skills, knowledge

Source: https://oxfordre.com/psychology/doc/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-518-graphic-001-full.gif

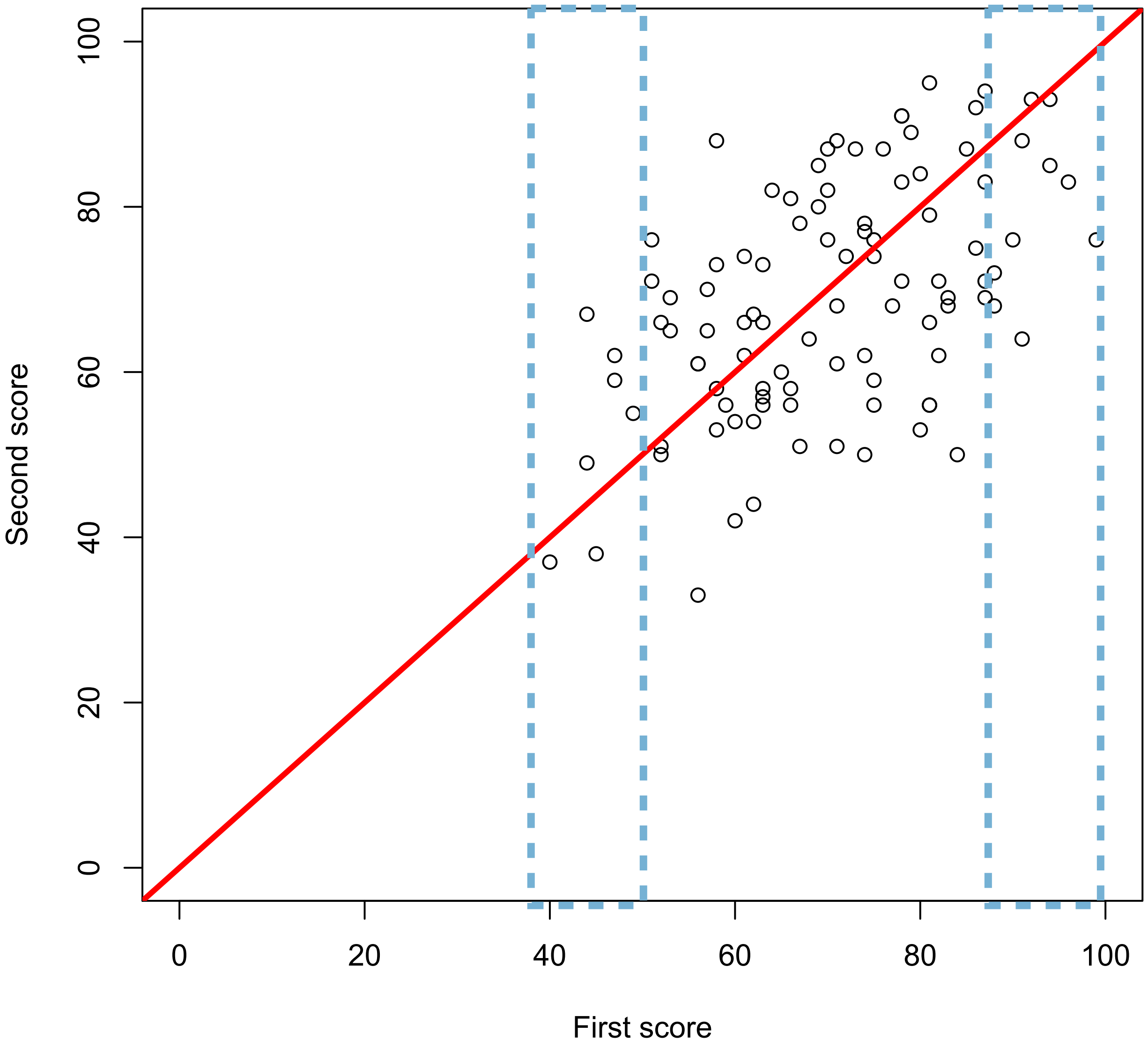
Why? General principle

Large deviations from the mean occur as a ***combination of factors*** some of which can be passed on to the to children via the relationship expressed in correlation, while other are not.

Hence the ***deviations from the mean are expected to shrink in the next generation***.

Test-retest situations: observations and conceptional interpretation

Situation: Students take two tests.

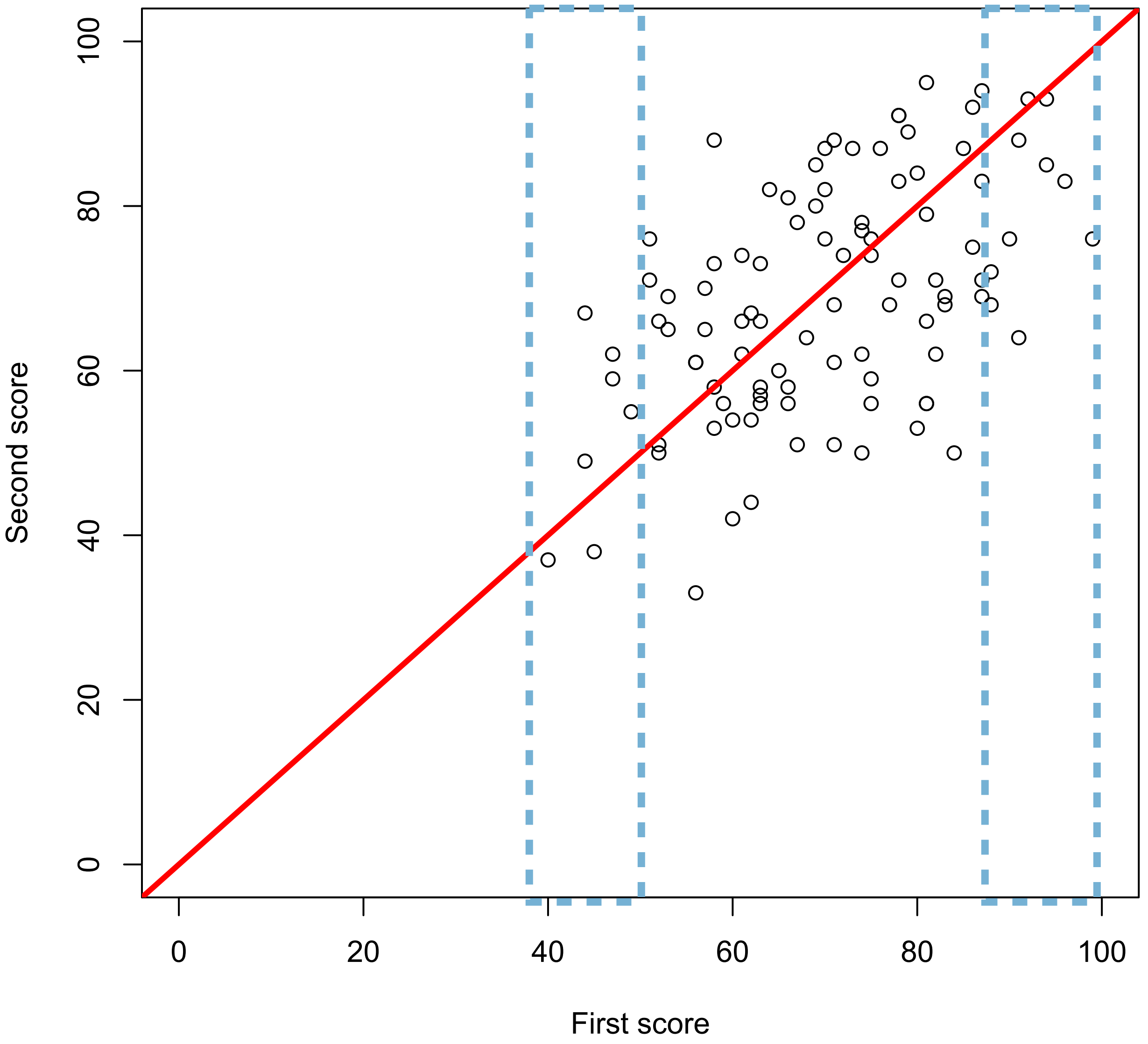


Observations:

- Students who did very well on the first quiz did less well on the second quiz
- Students who did poorly on the first quiz did better on the second quiz.

Test-retest situations: observations and conceptional interpretation

Situation: Students take two tests.



Observations:

- Students who did very well on the first quiz did less well on the second quiz
- Students who did poorly on the first quiz did better on the second quiz.

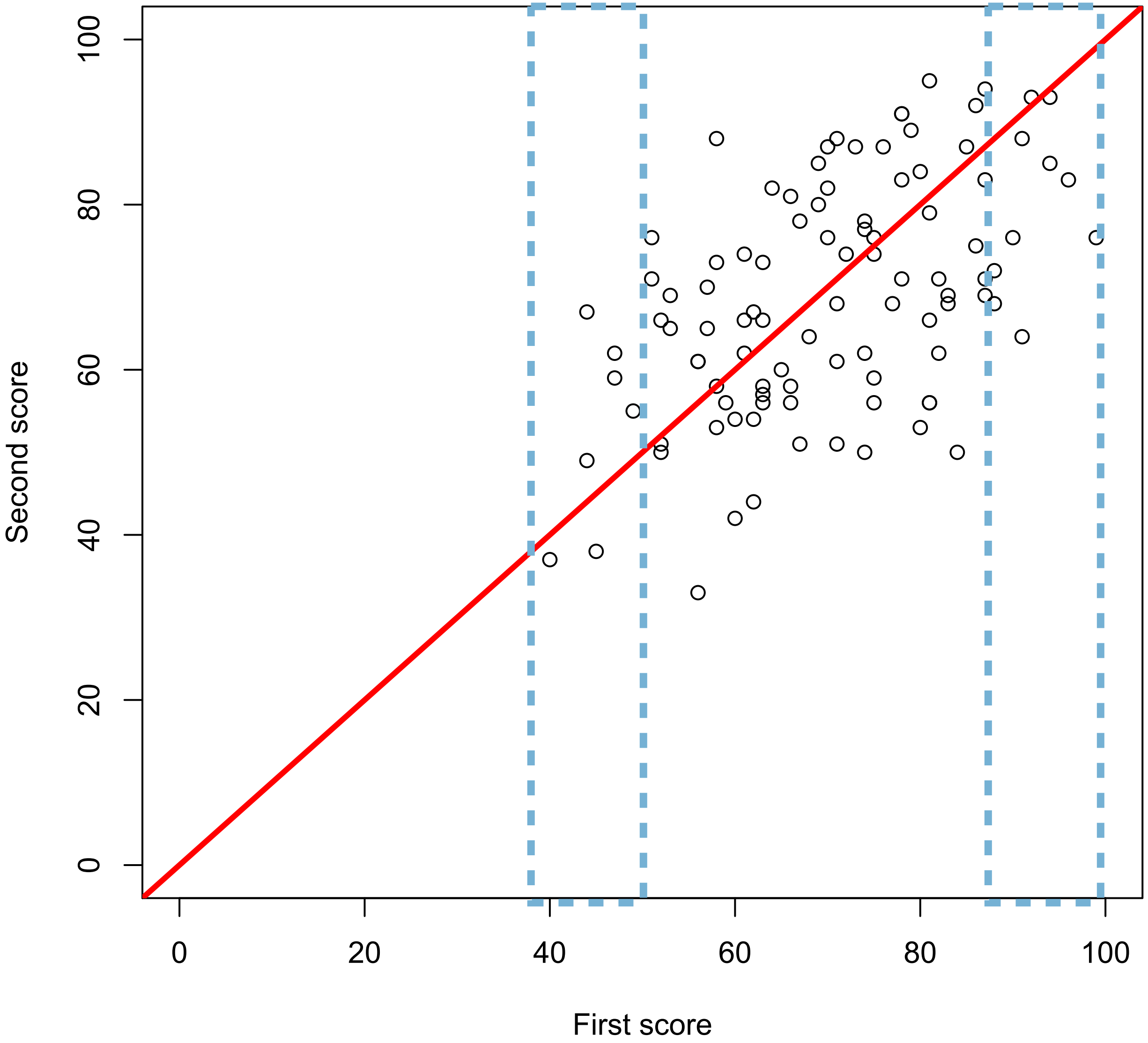
Interpretation:

- Students who did well the first time slacked off.
- Those shocked by a poor score the first time studied hard and improved.

Maybe, but it is a speculation.

Test-retest situations: observations and conceptional interpretation

Situation: Students take two tests.



Observations:

- Students who did very well on the first quiz did less well on the second quiz
- Students who did poorly on the first quiz did better on the second quiz.

Alternative interpretation:

Students who did well the first time had nowhere to go but down...

This is like regression to the mean. In this context also knowns as

"Test-retest effect"

Test-retest: a model based analysis

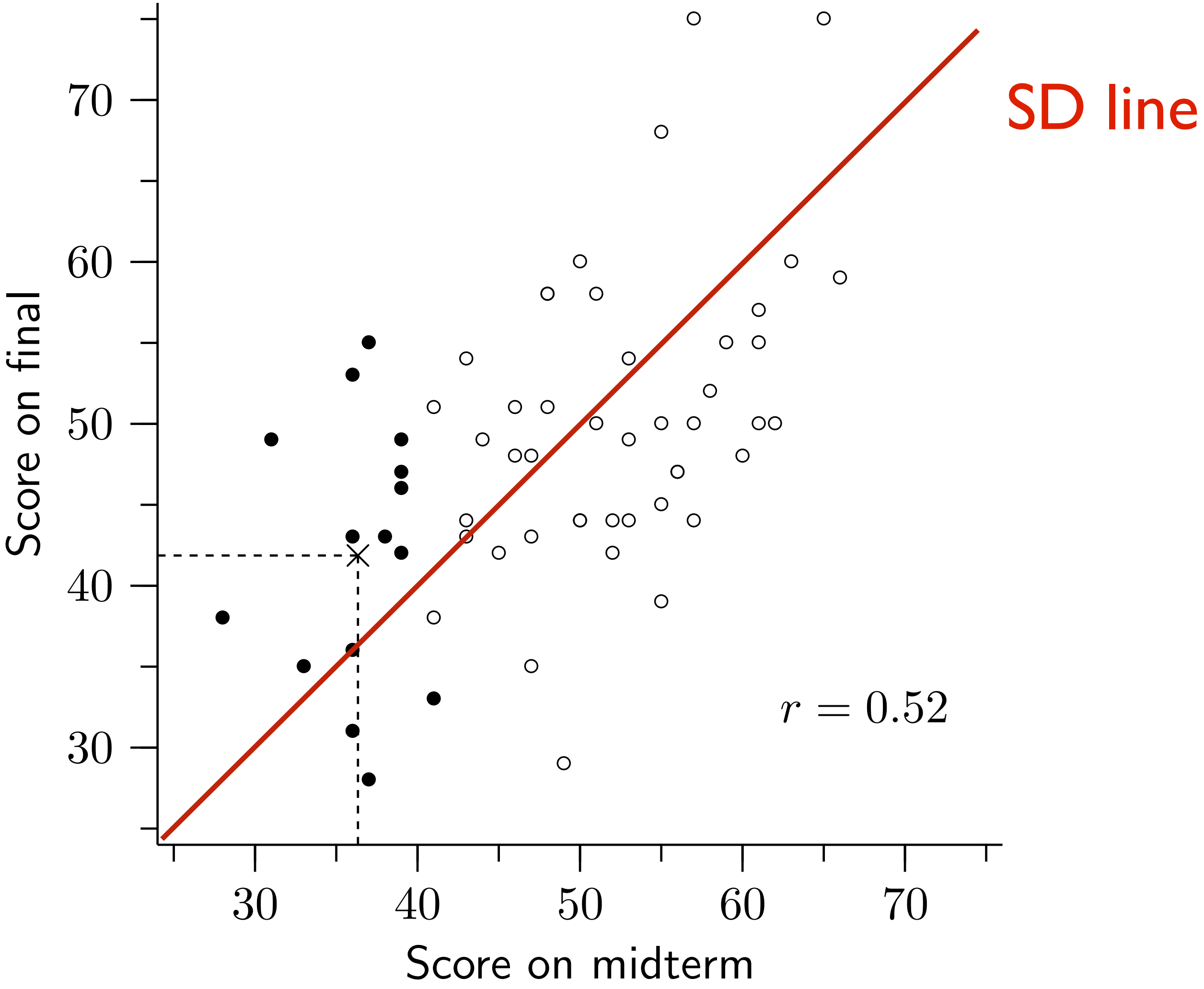
An experiment:

- An instructor standardises her midterm and final so the class average is 50 and the SD is 10 on both tests.
- The correlation between the tests is always around 0.50.
- On one occasion, she took the students who scored in the lower quartile at the midterm and gave them special tutoring.
- On average, they scored about 6 points higher on the final than they did on the midterm.
- Does this show that the special tutoring was effective?

Test-Retest effect

Midterm and final scores

The scores of the students in the lower quartile on the midterm are marked by ●'s; their point of averages is marked by the ×. The scores of the remaining students are marked by ○'s.



Test-retest effect: a model

Model: observed score = true score + chance error

true score: reflects students ability

chance error: accounts for factors like preparedness, anxiety, concentration etc

To check whether this model could explain the effect, create a data set accordingly:

- assign “true scores” to students, ~N(50,10)
- add random errors to the test scores, ~N(0,10)
- plot chance errors against true scores
- label lower quartile student and check *their* error distribution

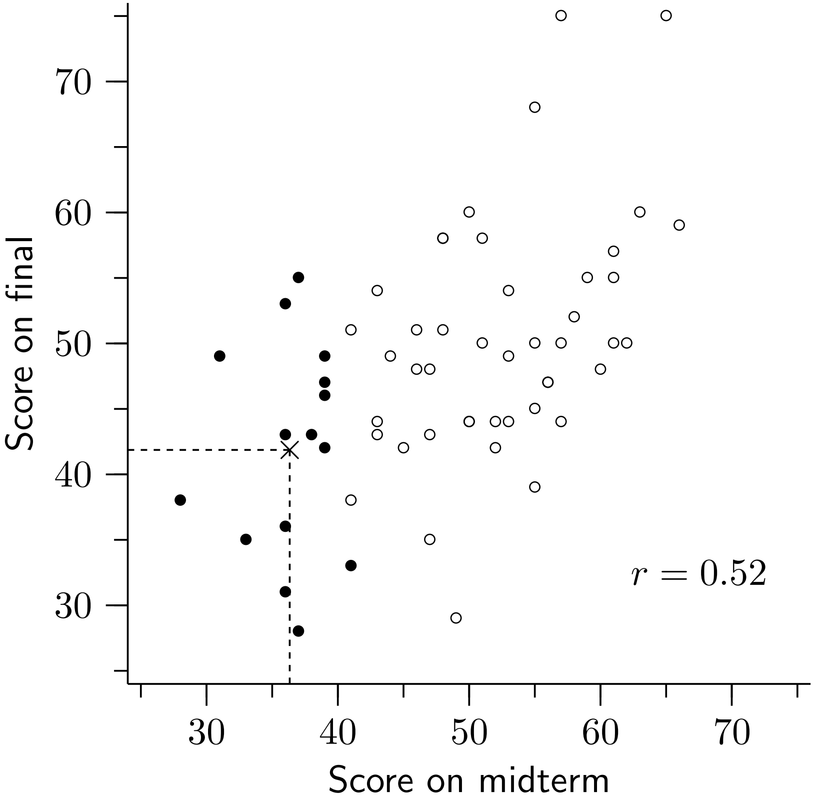
Test-retest effect

Model: observed score = true score + chance error

Lower quartile midterm students had mostly negative chance errors!

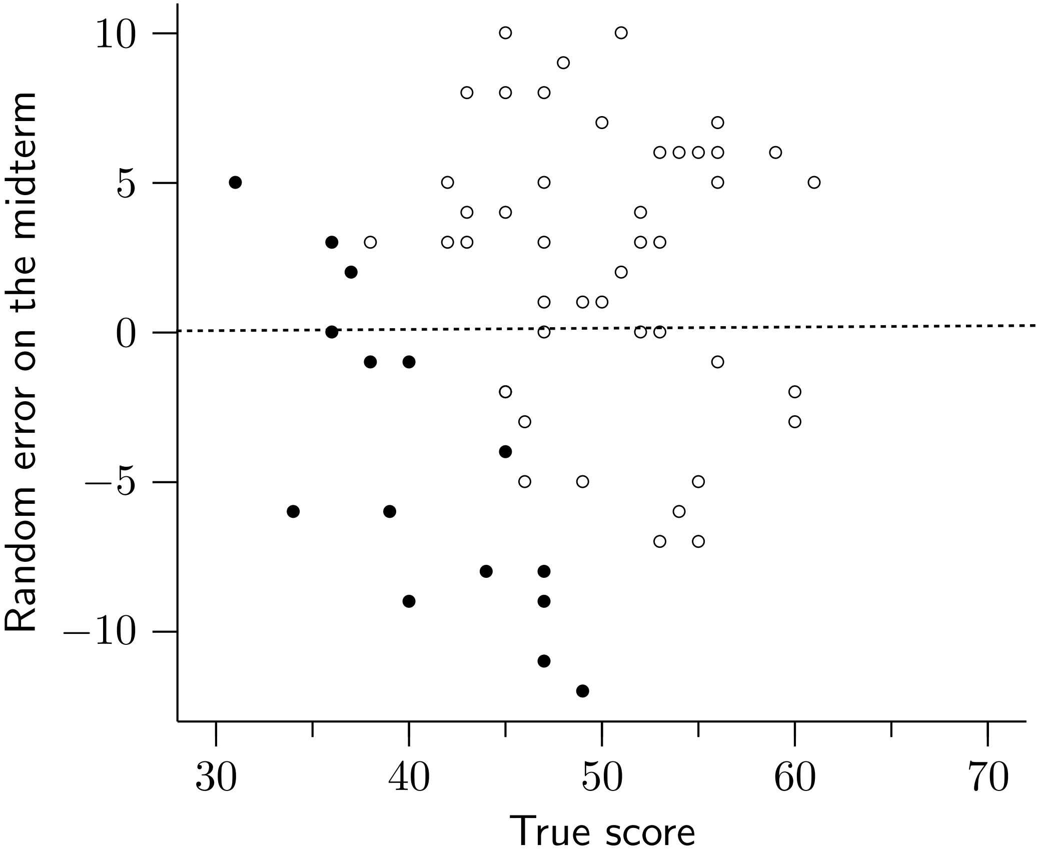
Midterm and final scores

The scores of the students in the lower quartile on the midterm are marked by $\bullet$'s; their point of averages is marked by the $\times$. The scores of the remaining students are marked by $\circ$'s.



Performance on the midterm

Students in the lower quartile on the midterm are represented by $\bullet$'s, the other students by $\circ$'s.



Test-retest effect:

The model offers an explanation for the improvement:

- For students in lower quartile on the midterm, distribution of random errors in midterm is skewed: mostly negative. In other words, their midterm scores understate their true performance.
- Their *average* performance on the final is expected to improve.
- The improvement may be due to special tutoring or it may be due to chance error or it maybe be due to a combination of both.
- This group's average improvement on the final can be predicted by regression applied to this group's midterm.
- In practice, however, we usually do not know all the necessary parameters of the model (e.g. SD of error).

Regression fallacy: Implications

Typical mistake: People attribute a decrease or increase to a systematic cause, based on selected observations of the top or bottom part of data. But the change may be due to chance variation (skewed distribution for selected parts of data

Regression fallacy: Implications

Typical mistake: People attribute a decrease or increase to a systematic cause, based on selected observations of the top or bottom part of data. But the change may be due to chance variation (skewed distribution for selected parts of data

Examples: test scores, pilot performance, safety measures etc

Daniel Kahneman:

"We normally reinforce others when their behaviour is good and punish them when their behaviour is bad. By regression alone, therefore, they are ***most likely to improve after being punished and most likely to deteriorate after being rewarded****.*

Consequently, we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding."

Regression fallacy: Implications

Typical mistake: People attribute a decrease or increase to a systematic cause, based on selected observations of the top or bottom part of data. But the change may be due to chance variation (skewed distribution for selected parts of data

Examples: test scores, pilot performance, safety measures etc

Daniel Kahneman:

"We normally reinforce others when their behaviour is good and punish them when their behaviour is bad. By regression alone, therefore, they are ***most likely to improve after being punished and most likely to deteriorate after being rewarded****.*

Consequently, we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding."

Implications for education, policy evaluation, evaluation of population based health recommendations etc.

Test-retest: What is better study design?

How would you design an experiment to find out whether or not the special tutoring helped?

Test-retest: What is better study design?

How would you design an experiment to find out whether or not the special tutoring helped?

Control group: Some students get special tutoring others do not. Then compare test score differences between groups.

Randomised: Assignment to special tutoring and control groups at random to avoid that other factors interfere (e.g. if special tutoring would be offered to volunteers, motivation would interfere with the results).

Test-retest: What is better study design?

How would you design an experiment to find out whether or not the special tutoring helped?

Control group: Some students get special tutoring others do not. Then compare test score differences between groups.

Test-retest: What is better study design?

How would you design an experiment to find out whether or not the special tutoring helped?

Control group: Some students get special tutoring others do not. Then compare test score differences between groups.

Randomised: Assignment to special tutoring and control groups at random to avoid that other factors interfere (e.g. if special tutoring would be offered to volunteers, motivation would interfere with the results).

Blind (students): All students could receive something called "special tutoring", but for the ones in the control group this would in reality be designed as having no effect (e.g. irrelevant topics presented in incomprehensible ways as "placebo tutoring").

Test-retest: What is better study design?

How would you design an experiment to find out whether or not the special tutoring helped?

Control group: Some students get special tutoring others do not. Then compare test score differences between groups.

Randomised: Assignment to special tutoring and control groups at random to avoid that other factors interfere (e.g. if special tutoring would be offered to volunteers, motivation would interfere with the results).

Blind (students): All students could receive something called "special tutoring", but for the ones in the control group this would in reality be designed as having no effect (e.g. irrelevant topics presented in incomprehensible ways as "placebo tutoring").

Blind (instructor): Instructor marking the exams does not know who received special tutoring.

Test-retest: better study design (ct.)

Is any of this doable in reality? That depends…

Control group: Sounds easy, but is not. Anticipating that students not assigned special tutoring may complain, the exam secretary will not allow this to go ahead. One way to get around this is to allow all students to participate with the understanding that some will receive real special tutoring and others something less effective. However, if the exam results matter, they may try to get each other's tutoring material destroying the study design. And, the exam secretary may still refuse based on legal concerns.

Test-retest: better study design (ct.)

Is any of this doable in reality? That depends…

Control group: Sounds easy, but is not. Anticipating that students not assigned special tutoring may complain, the exam secretary will not allow this to go ahead. One way to get around this is to allow all students to participate with the understanding that some will receive real special tutoring and others something less effective. However, if the exam results matter, they may try to get each other's tutoring material destroying the study design. And, the exam secretary may still refuse based on legal concerns.

Randomised: Easy to implement.

Test-retest: better study design (ct.)

Is any of this doable in reality? That depends…

Control group: Sounds easy, but is not. Anticipating that students not assigned special tutoring may complain, the exam secretary will not allow this to go ahead. One way to get around this is to allow all students to participate with the understanding that some will receive real special tutoring and others something less effective. However, if the exam results matter, they may try to get each other's tutoring material destroying the study design. And, the exam secretary may still refuse based on legal concerns.

Randomised: Easy to implement.

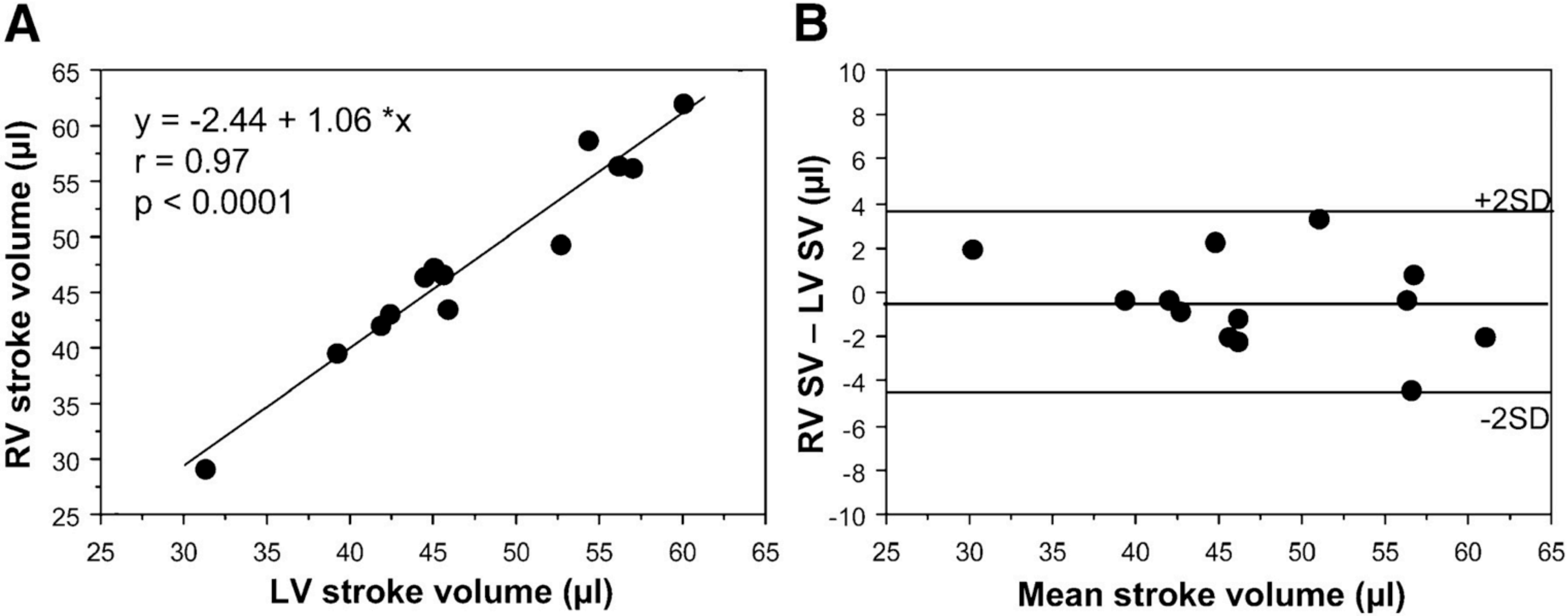
Blind (students): May be hard with students exchanging experiences and the controls realising they don't receive actual special tutoring.

Blind (instructor): Easy to implement as marking is anonymous.

Data visualisation: **Difference vs mean plot**

- For matched pairs data
- Instead of Y vs X plot difference vs mean (or difference vs sum)
- Difference more obvious (w/o tilting your head)
- Shows dependency on intensity
- It’s just a simple transformation
- Traditionally used for comparing different measurement methods, in particularly in medical diagnostic tests etc
- Suitable to access agreement
- Goes back to Tukey (EDA), became very popular in medical application through Bland & Altman’s paper in *Statistics of Medicine*

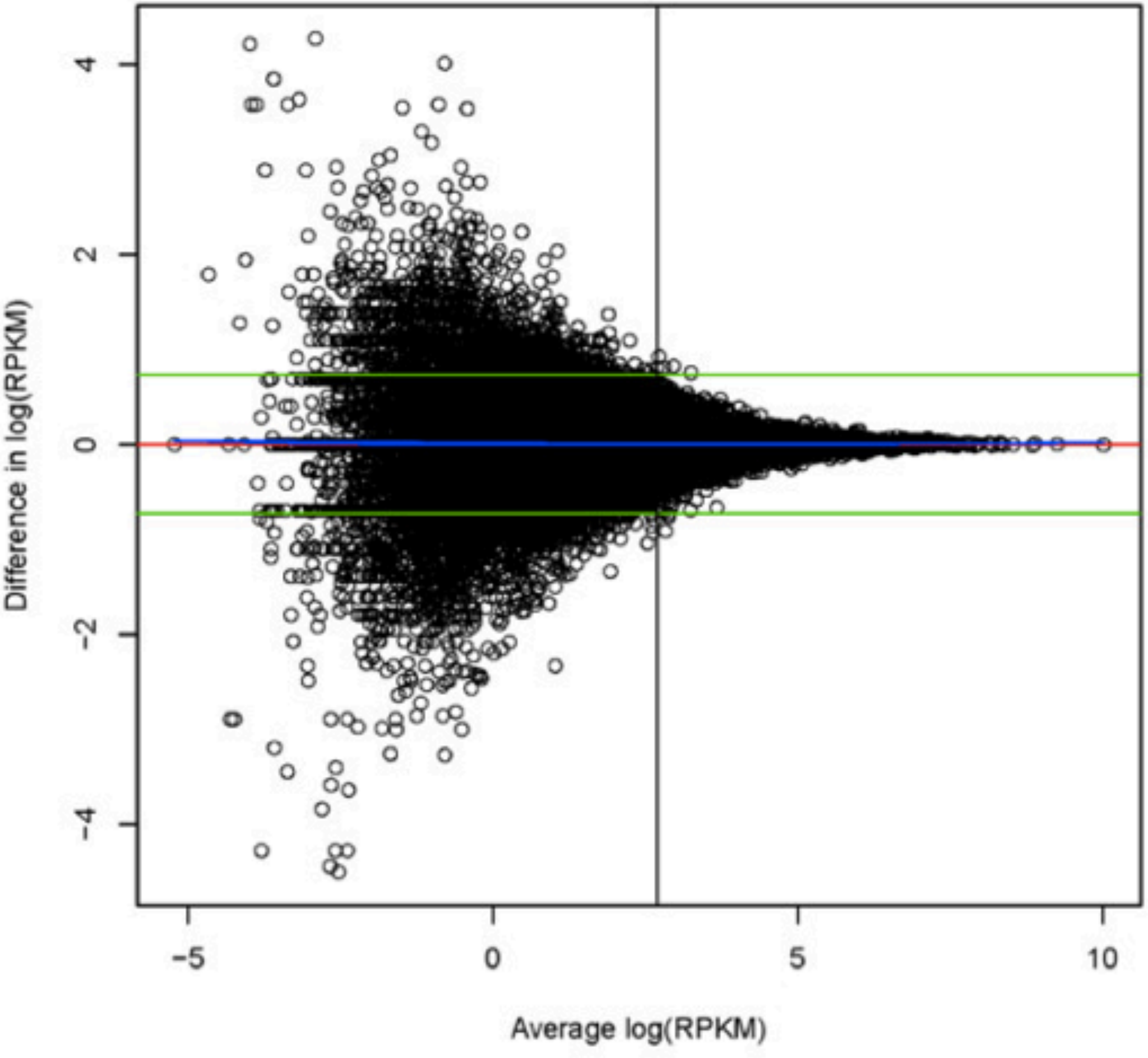
Example with MRI imaging data



A: regression analysis of the comparison between RV and LV stroke volumes (SV) as assessed by MR imaging (MRI). *B*: Bland-Altman analysis for assessment of the agreement between RV and LV SV measurements. Given is the difference between RV and LV SV for each mouse studied (*y*-axis) over the mean of RV and LV SV for each mouse (*x*-axis).

Source: http://ajpheart.physiology.org/content/283/3/H1065

Example with genomic data



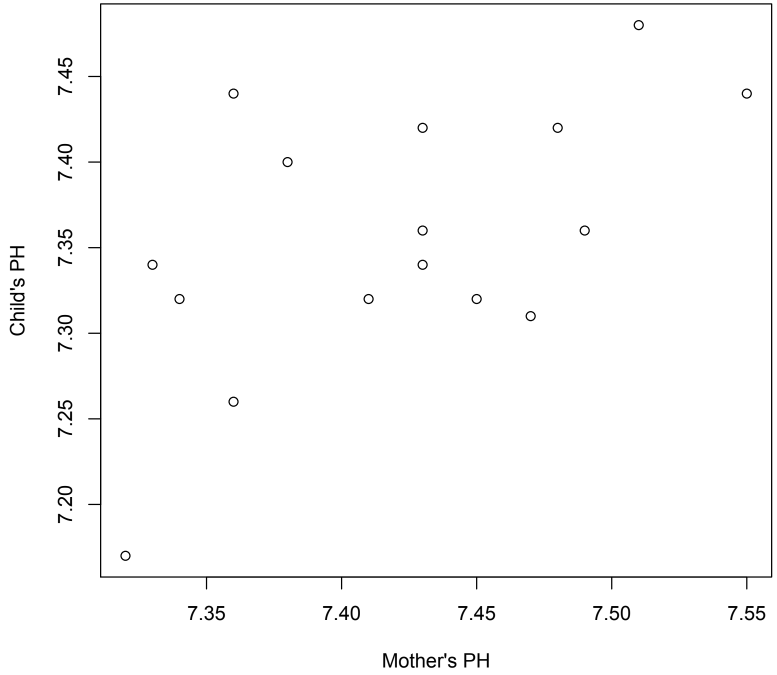
Bland-Altman plot showing level of agreement between technical replicates for natural log transformed RPKM *D. simulans* biological replicate 3. On the Y axis is the difference between technical replicates and on the X axis is the average between technical replicates. Green lines are the average of all differences +/- 1.96 (standard deviation of the differences). The red line is drawn at zero. The blue line is a loess fit. The discrepancy between technical replicates is a function of the estimated expression level. The horizontal line is drawn at an average coverage per nucleotide of 5. Bland-Altman plots for all the remaining comparisons among technical replicates are in Additional file 11.
McIntyre *et al*. *BMC Genomics* 2011 **12**:293

Matched-Pairs Data

Matched pairs arise when the *same* variable is measured on two *matched* experimental units.

- **Example**:

  Mother’s and baby’s blood pH level during labour



Exercise:
Create a difference versus sum plots from this dataset.

To examine the relationship, during labour, of the blood pH-levels of a mother and child. (in pH units: below 7 indicates acidity, above 7 alkalinity)

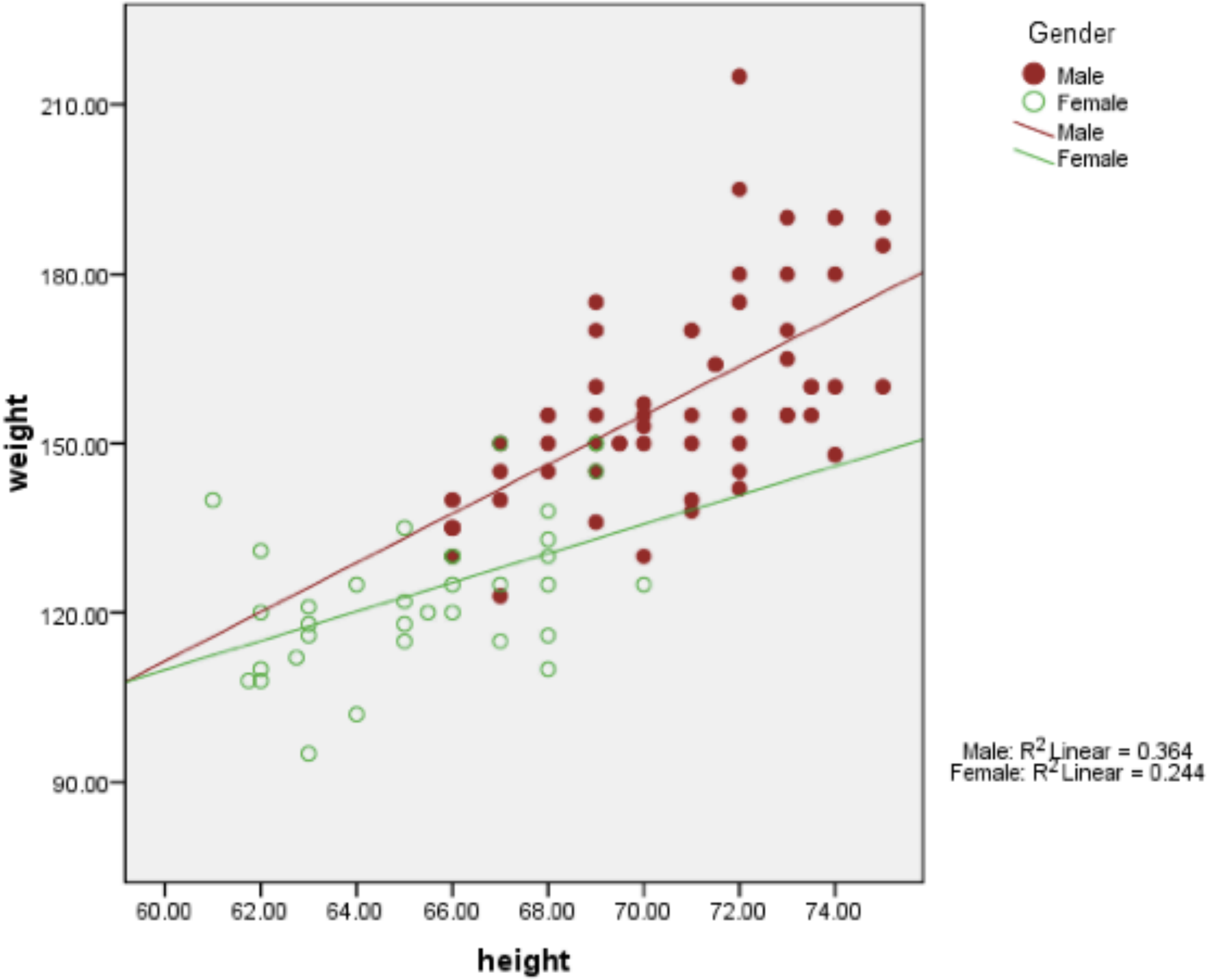
Maternal pH	7.33	7.41	7.49	7.43	7.32	7.43	7.55	7.36
Child pH	7.34	7.32	7.36	7.34	7.17	7.36	7.44	7.26
Maternal pH	7.34	7.45	7.51	7.48	7.38	7.36	7.43	7.47
Child pH	7.32	7.32	7.48	7.42	7.40	7.44	7.42	7.31

Aggregating and stratifying data

Prototype question:

- Observe two variables X and Y
- Population can be divided into a number of groups (known)
- For each group, the correlation between X and Y is 0.6
- What is the correlation, approximately, in the whole population?

Groups could be, for example, by age, occupation, nationality etc


Gender
Male
Female
Male
Female
210.00
180.00
150.00
120.00
90.00
weight
60.00
62.00
64.00
66.00
68.00
70.00
72.00
74.00
height
Male: R^2 Linear = 0.364
Female: R^2 Linear = 0.244

Simpson’s paradox

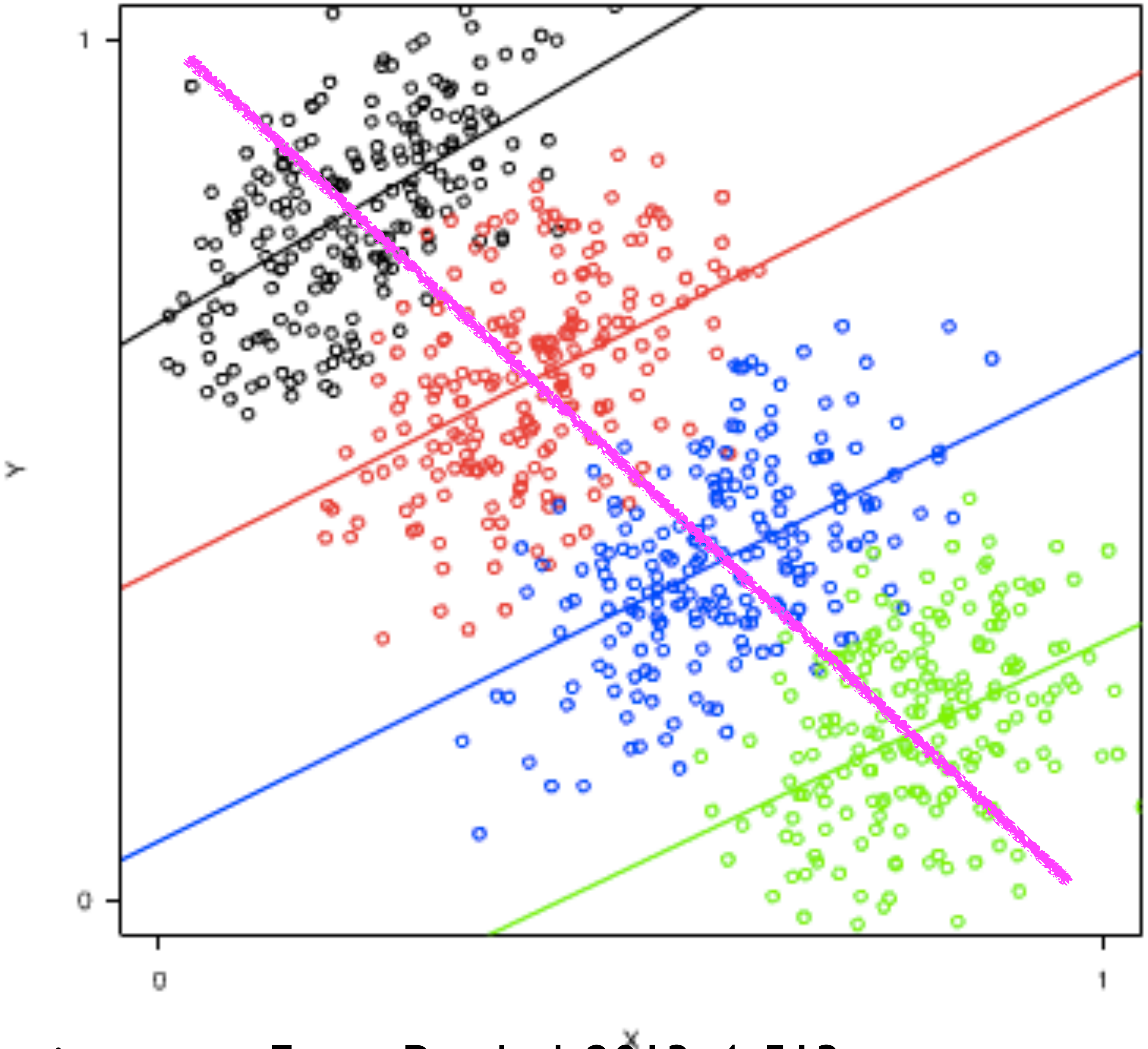


Figure on following pages: Front Psychol. 2013; 4: 513.
Simpson's paradox in psychological science: a practical guide, Rogier A. Kievit et al

http://rogierkievit.com/wp-content/uploads/2013/05/Kievit_Original_Manuscript_7_7.pdf

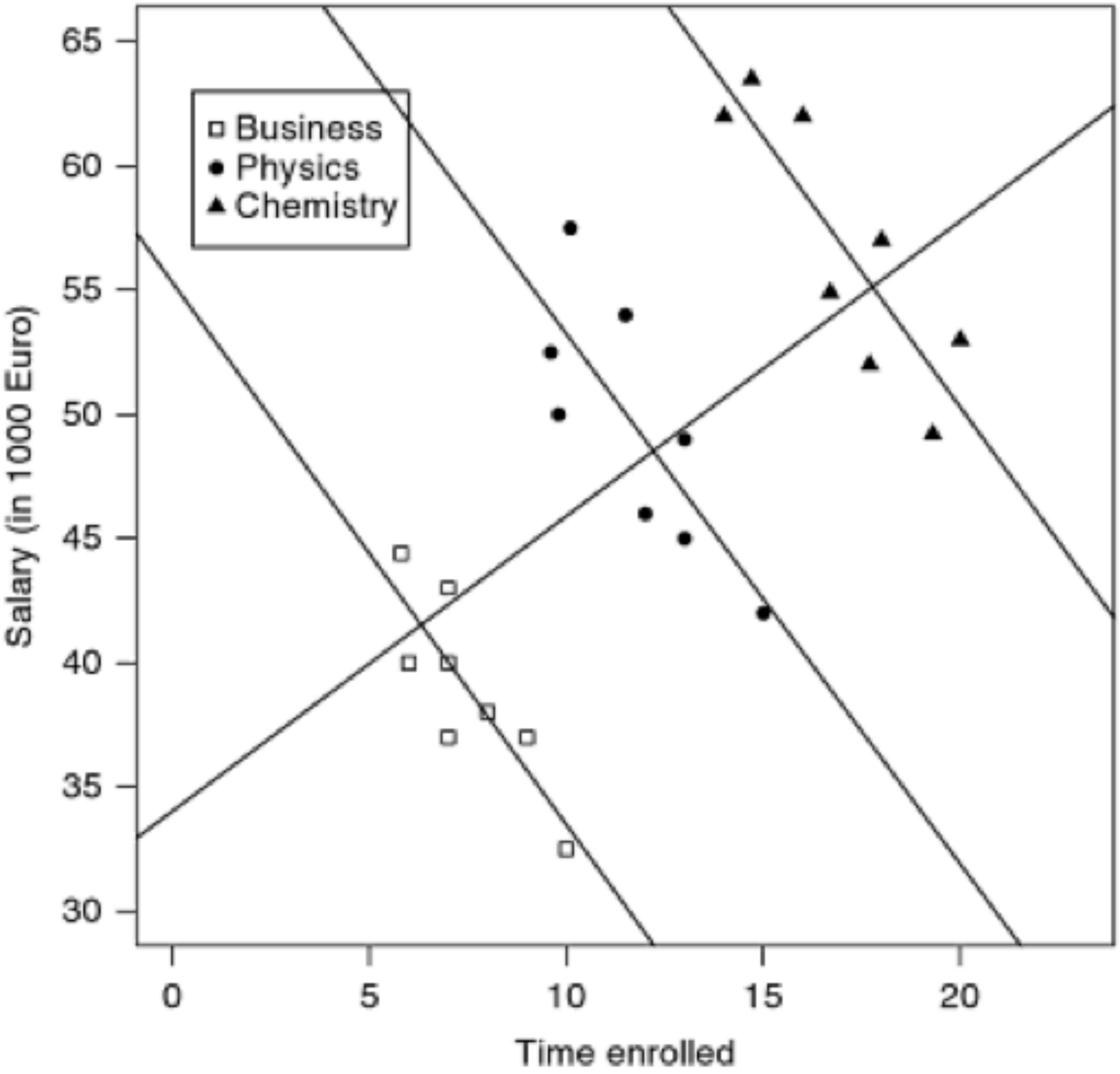
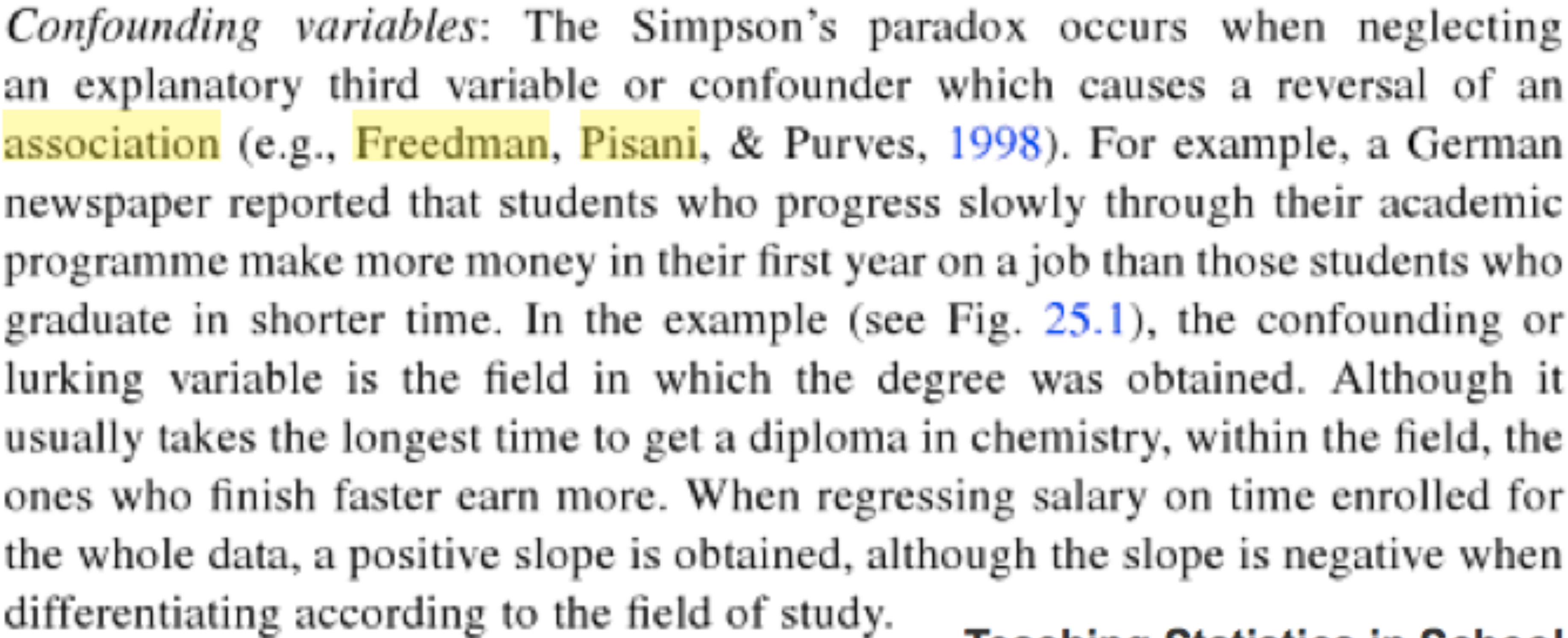


Fig. 25.1 Time enrolled until graduation (in semesters) and salary in first year of employment (in thousand €)



Confounding variables: The Simpson's paradox occurs when neglecting an explanatory third variable or confounder which causes a reversal of an association (e.g., Freedman, Pisani, & Purves, 1998). For example, a German newspaper reported that students who progress slowly through their academic programme make more money in their first year on a job than those students who graduate in shorter time. In the example (see Fig. 25.1), the confounding or lurking variable is the field in which the degree was obtained. Although it usually takes the longest time to get a diploma in chemistry, within the field, the ones who finish faster earn more. When regressing salary on time enrolled for the whole data, a positive slope is obtained, although the slope is negative when differentiating according to the field of study.

Teaching Statistics in School Mathematics-Challenges for Teaching and ...

edited by Carmen Batanero, Gail Burrill,
hris Reading

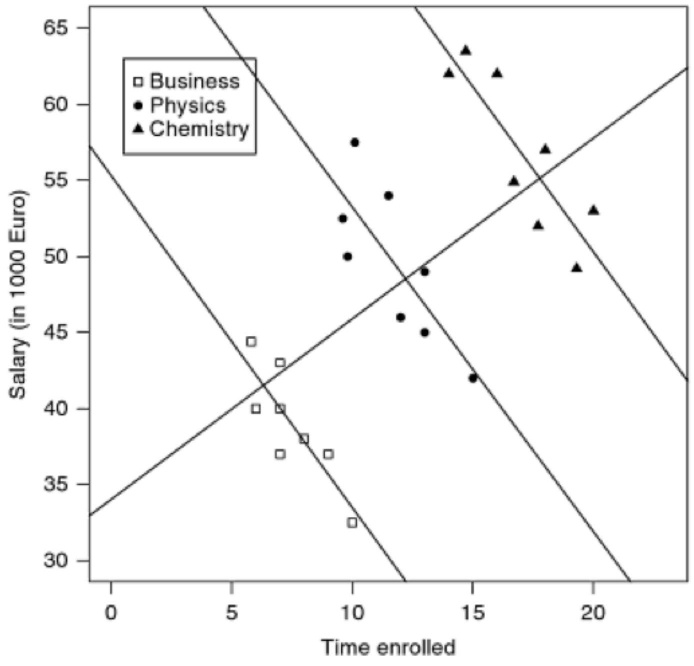


Fig. 25.1 Time enrolled until graduation (in semesters) and salary in first year of employment (in thousand €)

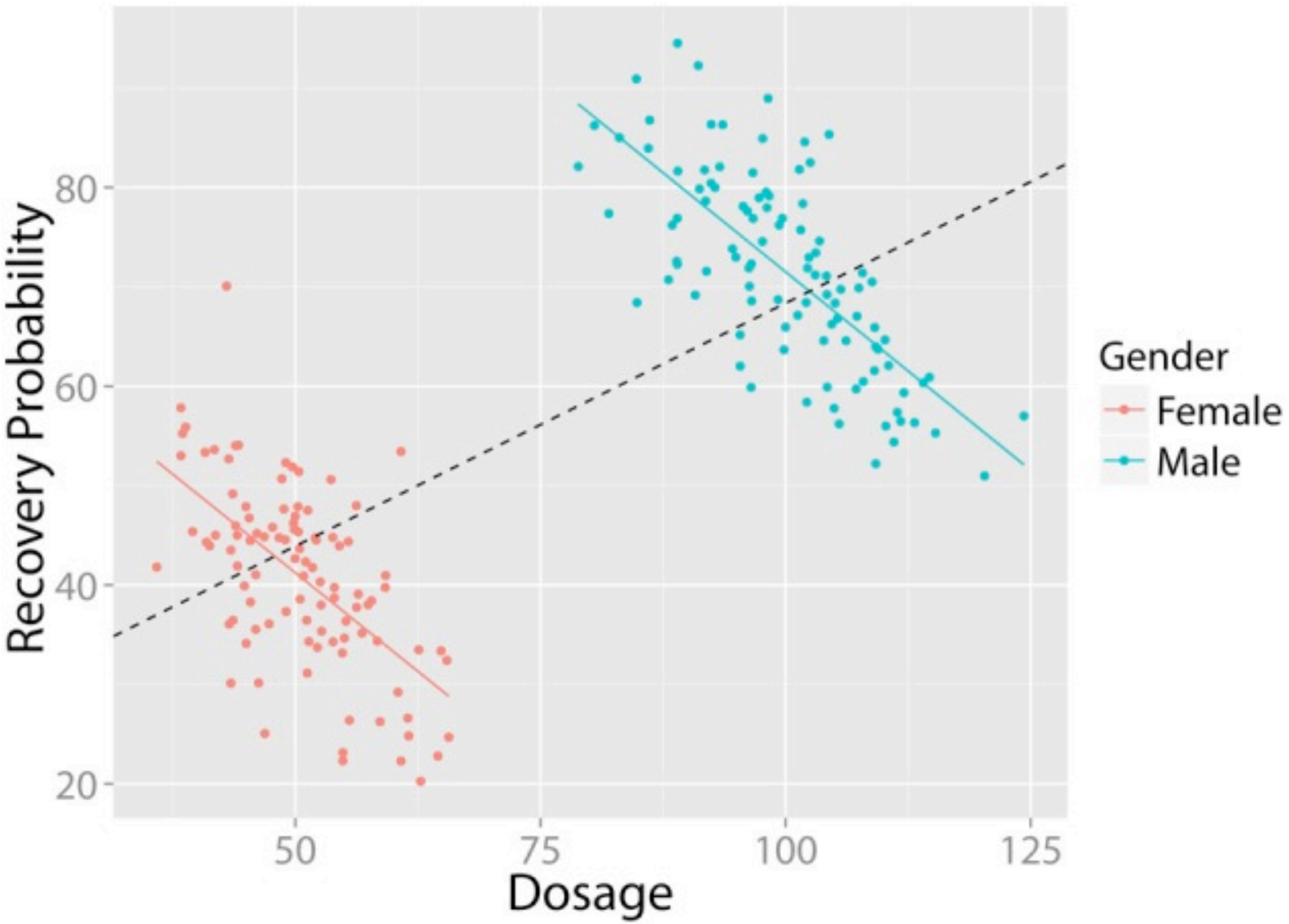
Note: In addition to what is said above, the graduation times for chemistry probably relate to PhD not diploma. For traditional reasons, in Germany, almost all chemistry students stay on for PhD. This is not the case for other degrees.

Typical scenarios for Simpson's paradox

Figure on following figures from a review paper explaining the Simpson's paradox and how to detect and avoid it. This has in mind real-world applications, but has simulated data scenarios making the particular mechanisms very transparent.

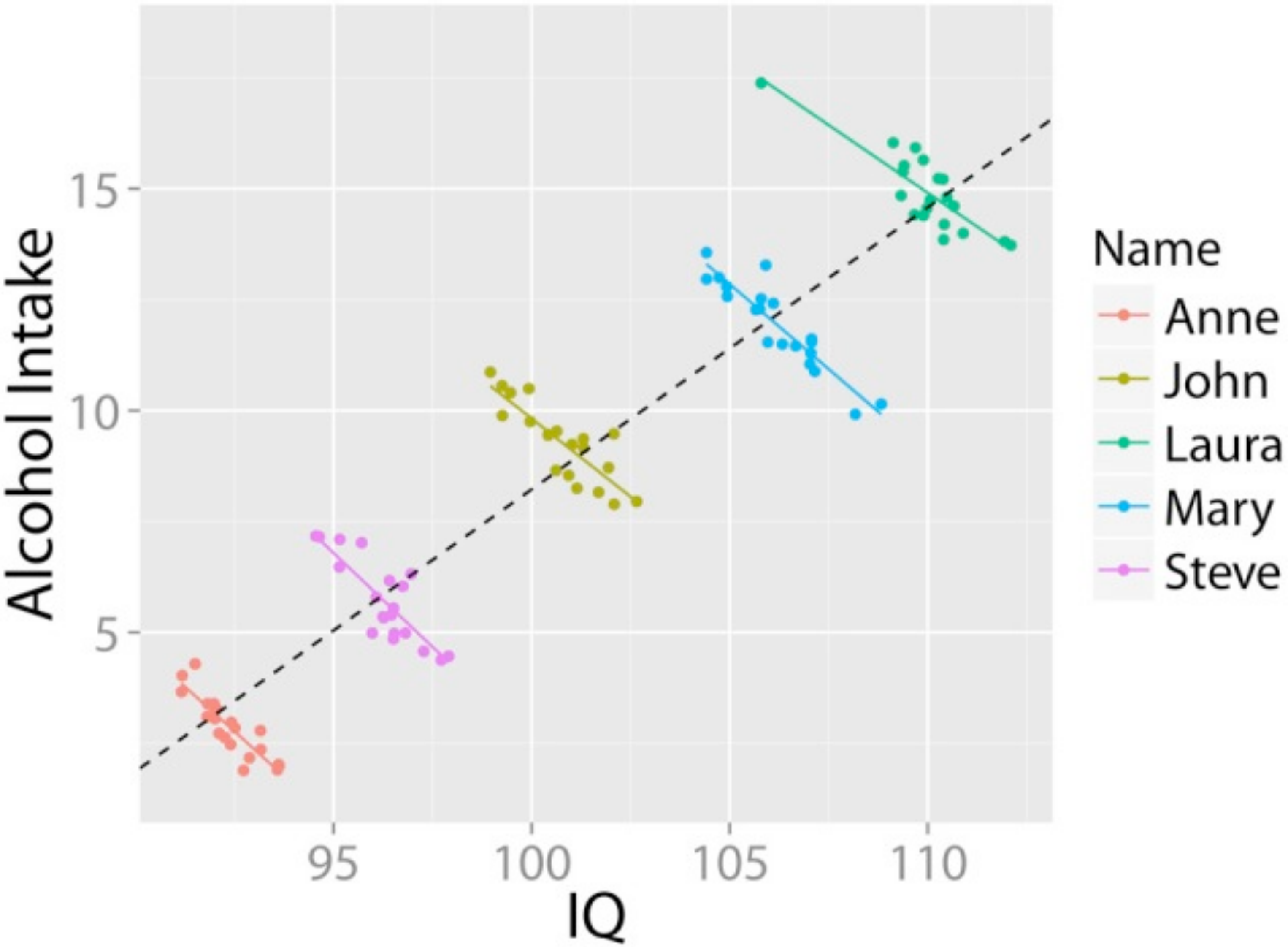
Front Psychol. 2013; 4: 513.
Simpson's paradox in psychological science: a practical guide
Rogier A. Kievit et al

rogierkievit.com/wp-content/uploads/2013/05/Kievit_Original_Manuscript_7_7.pdf

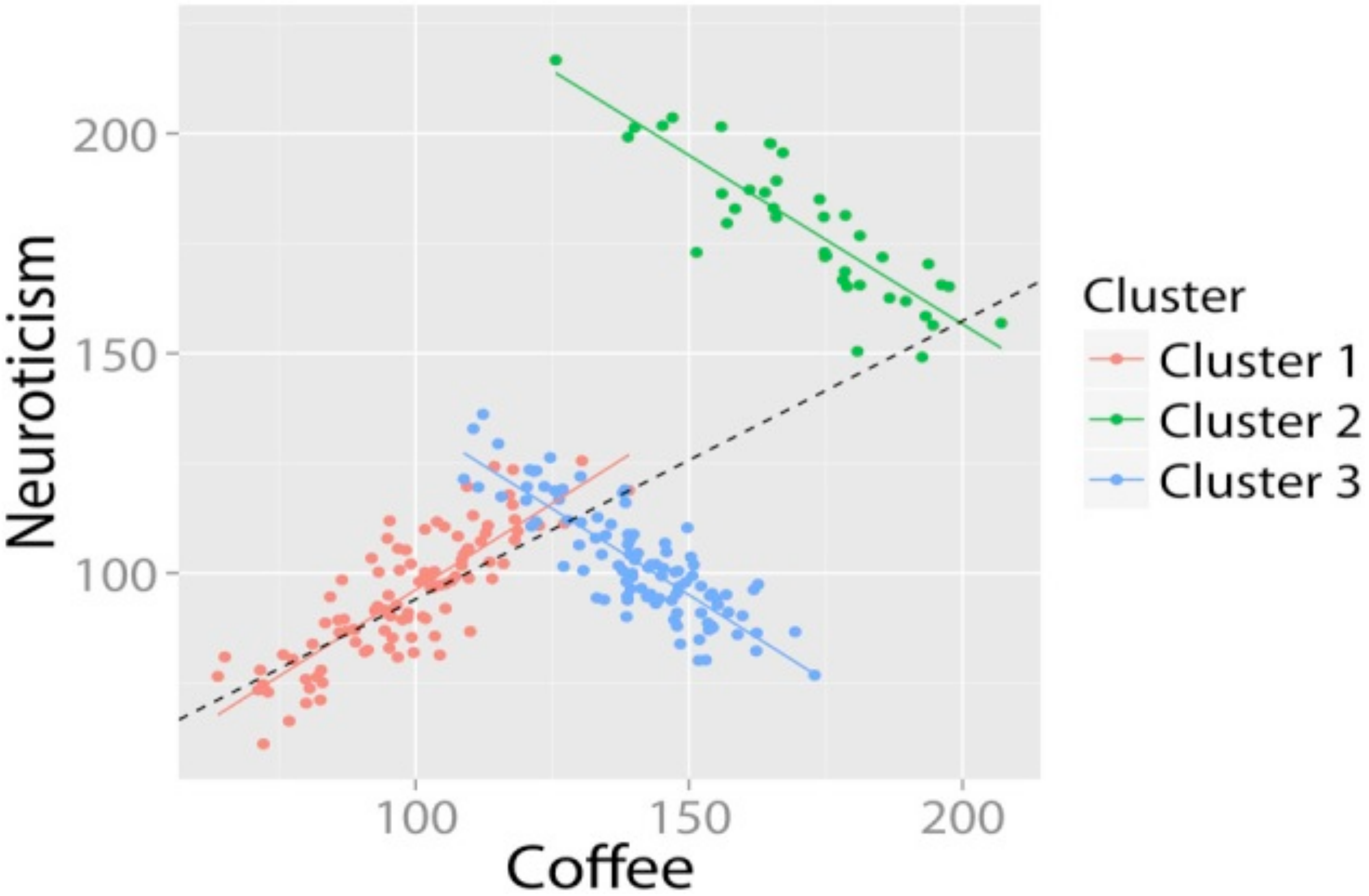


Example of Simpson's Paradox. Despite the fact that there exists a negative relationship between dosage and recovery in both males and females, when grouped together, there exists a positive relationship. All figures created using ggplot2 (Wickham, 2009). Data in arbitrary units.

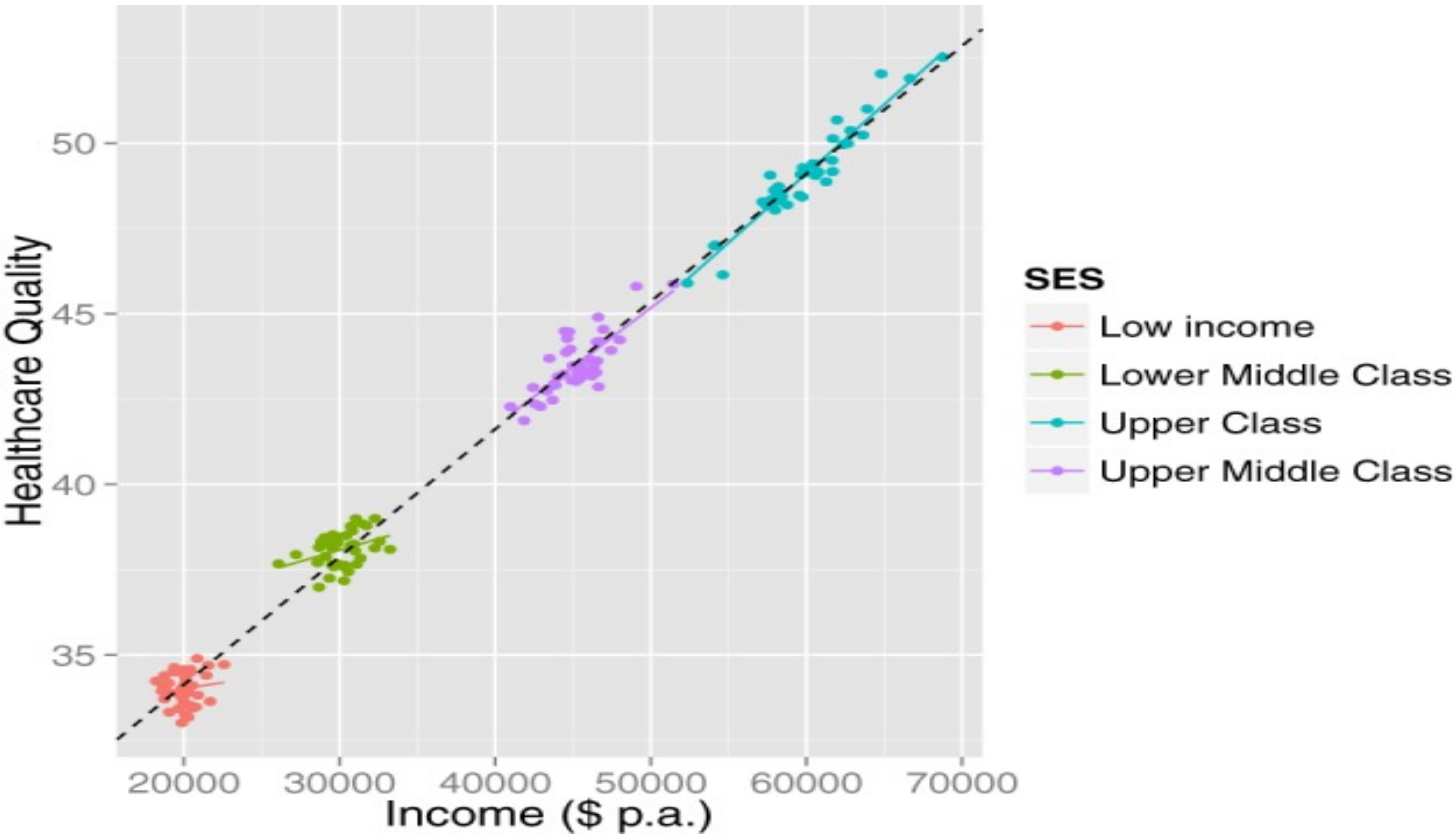
Simpson's paradox in individual differences



Alcohol use and intelligence. Simulated data illustrating that despite a positive correlation at the group level, within each individual there exists a negative relationship between alcohol intake and intelligence. Data in arbitrary units.



Using cluster analysis to uncover Simpson's Paradox. The cluster analysis (correctly) identifies that there are three subclusters, and that the relationship in two of these both deviates significantly from the group mean, and is in the opposite direction. Data in arbitrary units.

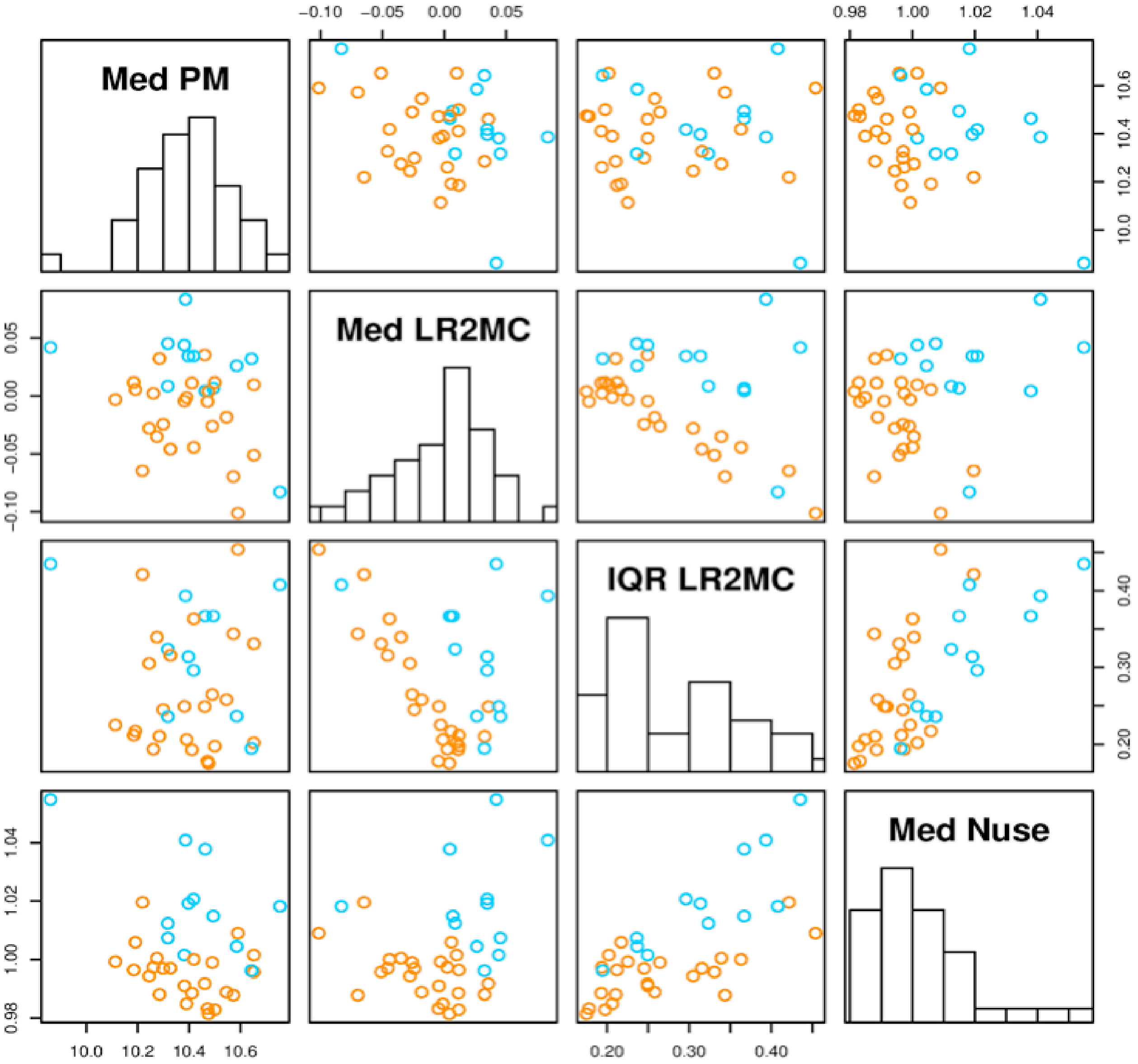


A case when visualizing the data illustrates that although there are separate clusters, the inference is not affected: the relationship between income and healthcare quality is homogeneously positive. The clusters may have arisen due to a sampling artifact or due to naturally occurring patterns in the population (e.g., discontinuous steps in healthcare plans).

Simpson’s paradox

- Counterintuitive feature of data
- Arises when (causal) inferences are drawn across different explanatory levels (e.g. population to subgroups, subgroups to individuals)
- Linear relationships can weaken, disappear or even inverse when aggregating data
- Detect it by labelling data points based on subgroups, exploring alternative categorisations
- Unsupervised detection with clustering methods
- Related/aka: ecological regression, ecological fallacy, Robinson’s paradox (continuous case)

Example from a scientific collaboration



Drosophila time series

Comparison of 4 quality scores for 36 microarrays

Note the effect of the date of the measurement (label in color)

Representation in juries: a question with two answers?

Question: Are Maoris properly represented in the New Zealand's jury pools?

First quick answer: Yes, the overall percentage (Whole new Zealand) is even slightly bigger than what would be expected.

Eligible Population Maori %	Jury Pool Maori %
9.5	10.1

Second attempt: Look at the representation in each of the districts.

Source: Ian Westbrooke (1998)
Simpson's Paradox: An Example in a New Zealand Survey of Jury Composition, CHANCE, 11:2, 40-42

http://www.stats.govt.nz/NR/rdonlyres/5DFEE5C8-A969-4684-B95F-FA2AFEDA649A/0/Simpdox.pdf
http://dx.doi.org/10.1080/09332480.1998.10542093

Simpson's paradox in Berkeley admissions

1973, UC Berkeley was sued for sex discrimination

Graduate School had just accepted, based on departmental decisions:
44% of male applications
35% of female applicants

Investigation by Bickel et al revealed Simpson's paradox:

- Women were (slightly) more likely to be admitted by the individual departments, but more women applied to the departments with higher rejections rates.
- In the aggregate data that amounted to a lower rejection rate for women.

PJ Bickel et al, *Sex Bias in Graduate Admissions: Data from Berkeley,* Science, new series, Vol. 187, no.4175 (Feb. 7, 1975), pp 398-404

Full text at http://www.unc.edu/~nielsen/soci708/cdocs/Berkeley_admissions_bias.pdf

Detecting a lurking variable

% applicants admitted versus % women applicants

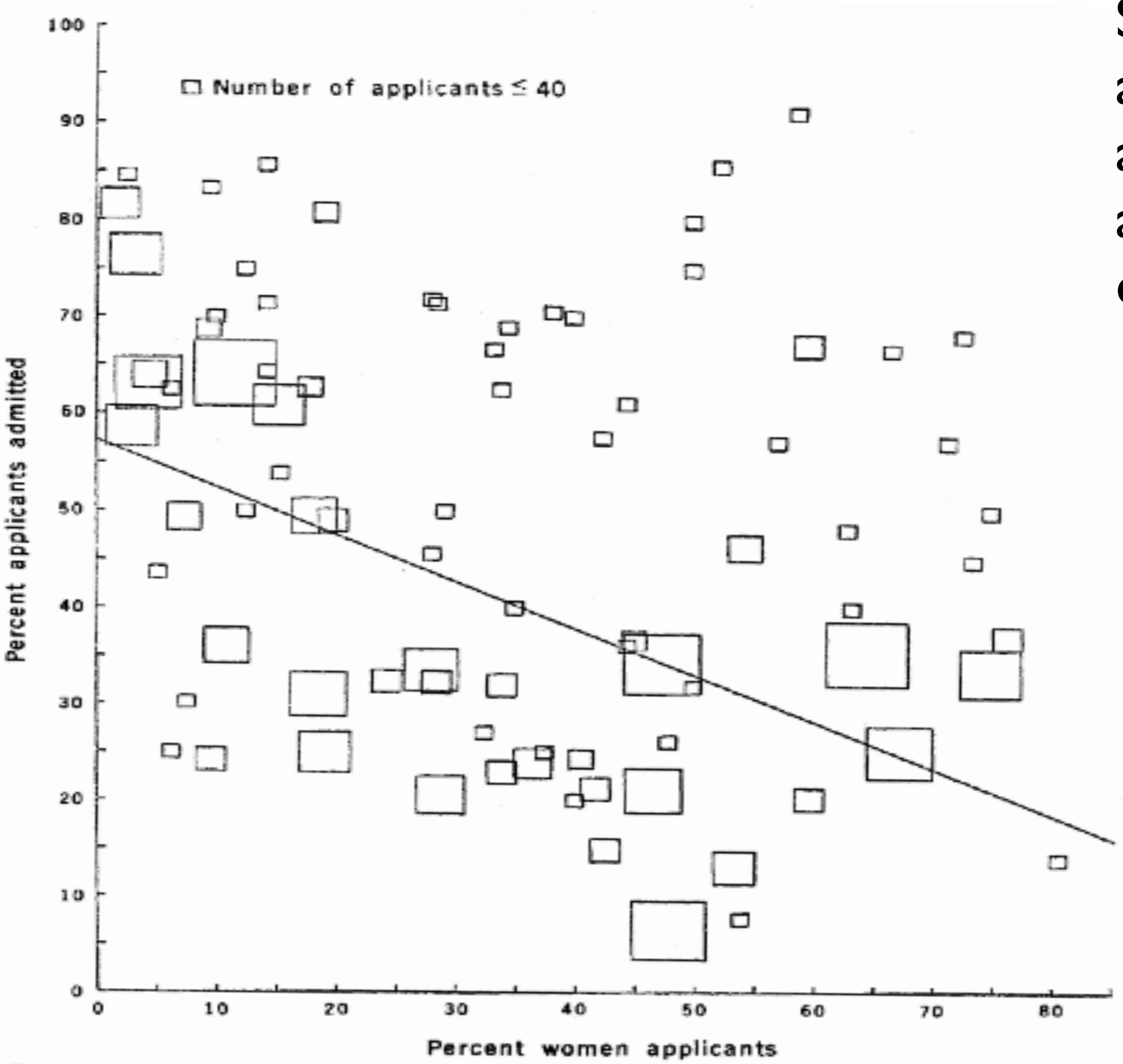


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

Scatter plot of % of women among the applicants (x) and the % of applicants accepted (y) for all 85 departments at Berkeley.

- Negative linear relationship: Percentage admitted decreases with percentage of women
- Relationship is stronger for bigger departments

Source: PJ Bickel et al 1973

Table demonstrating the effect

Further analysis shows that the departments with higher acceptance rates are in engineering/maths/science, while the lower acceptance rates are in the humanities departments. The table below (from PJ Bickel et al, 1973) visualises the Simpson's paradox scenario for the case of two departments.

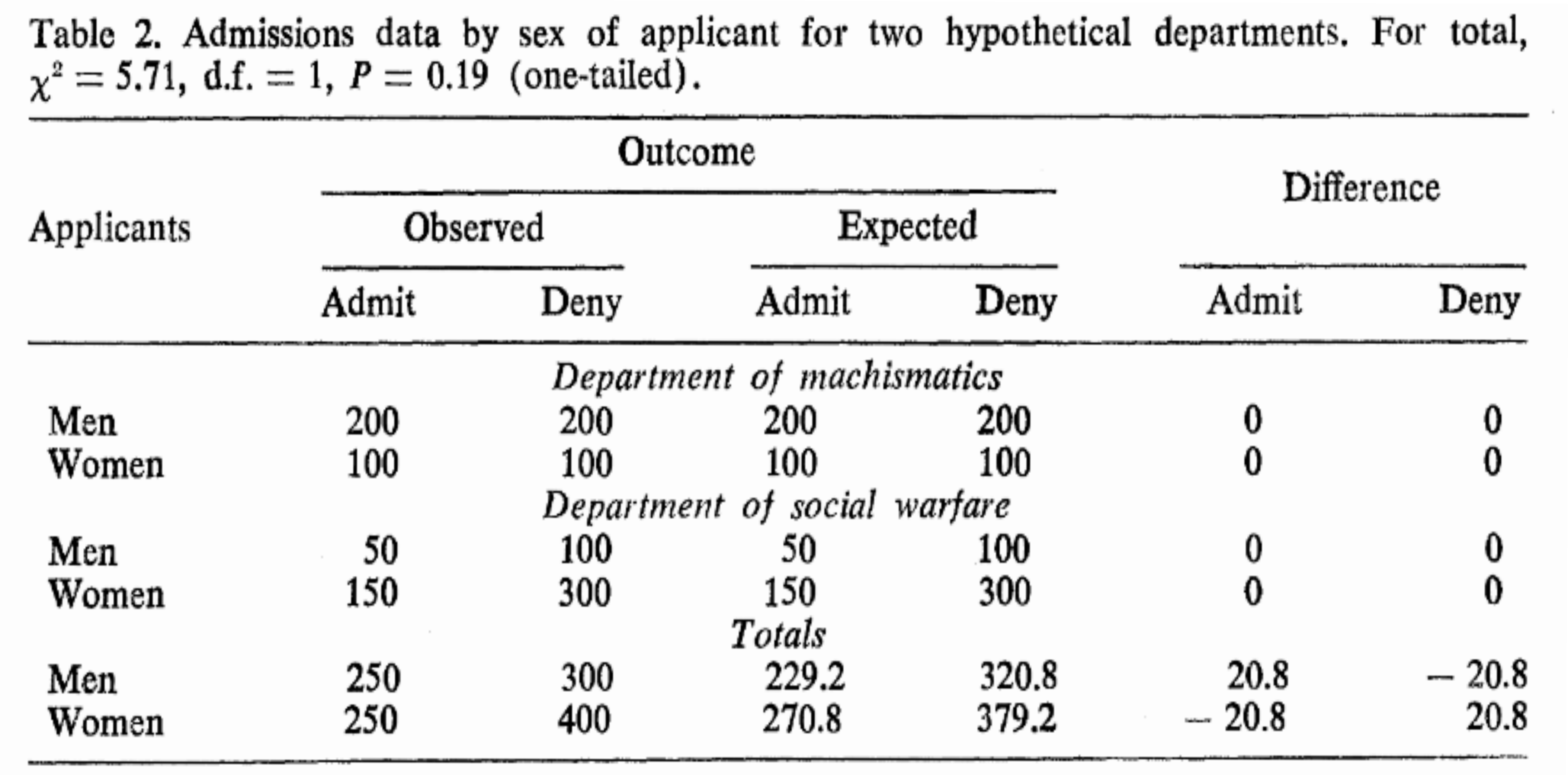
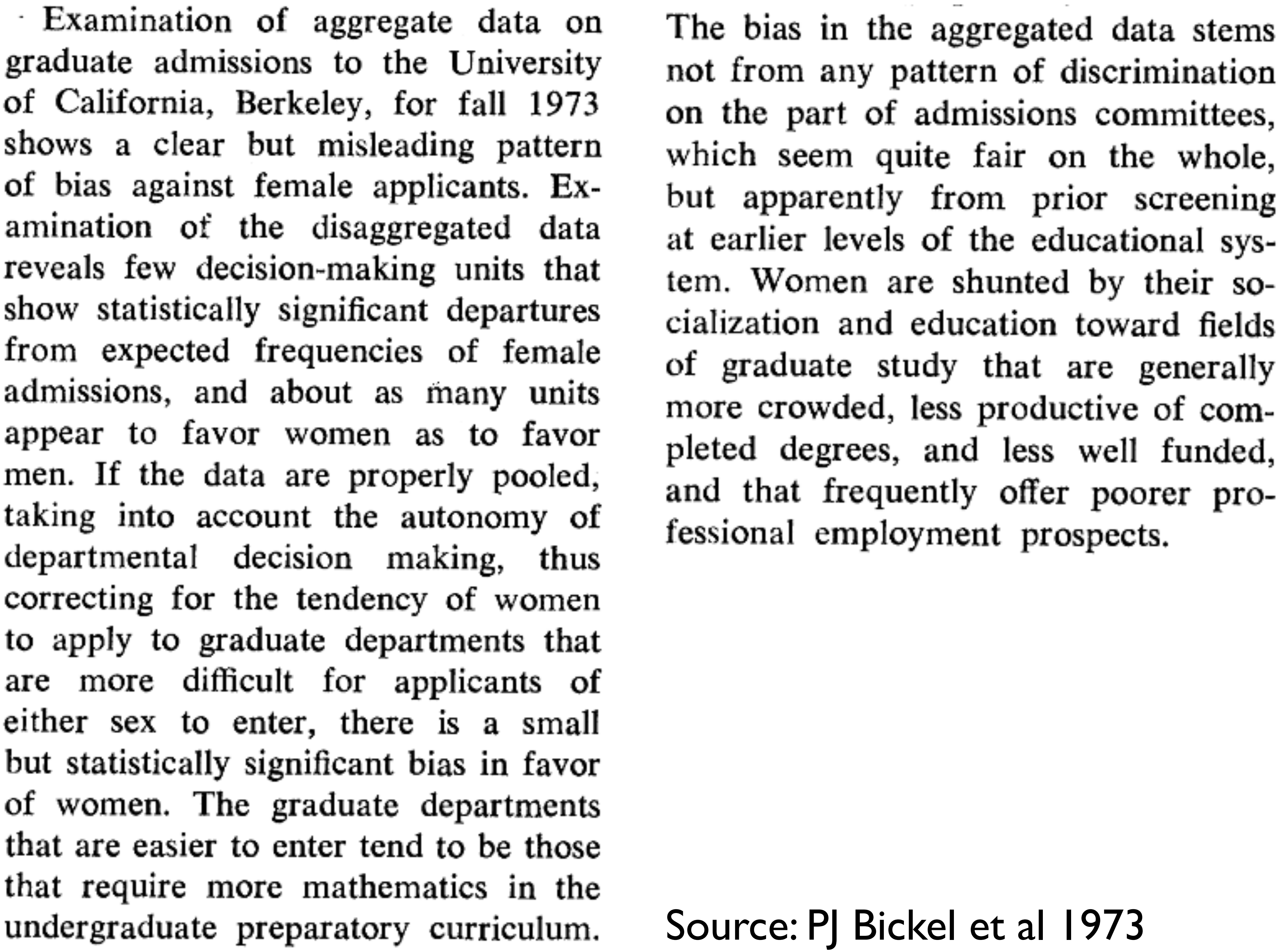


Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. $= 1$, $P = 0.19$ (one-tailed).

	Outcome				Difference	
Applicants	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
			Department of machismatics			
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
			Department of social warfare			
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
			Totals			
Men	250	300	229.2	320.8	20.8	– 20.8
Women	250	400	270.8	379.2	– 20.8	20.8

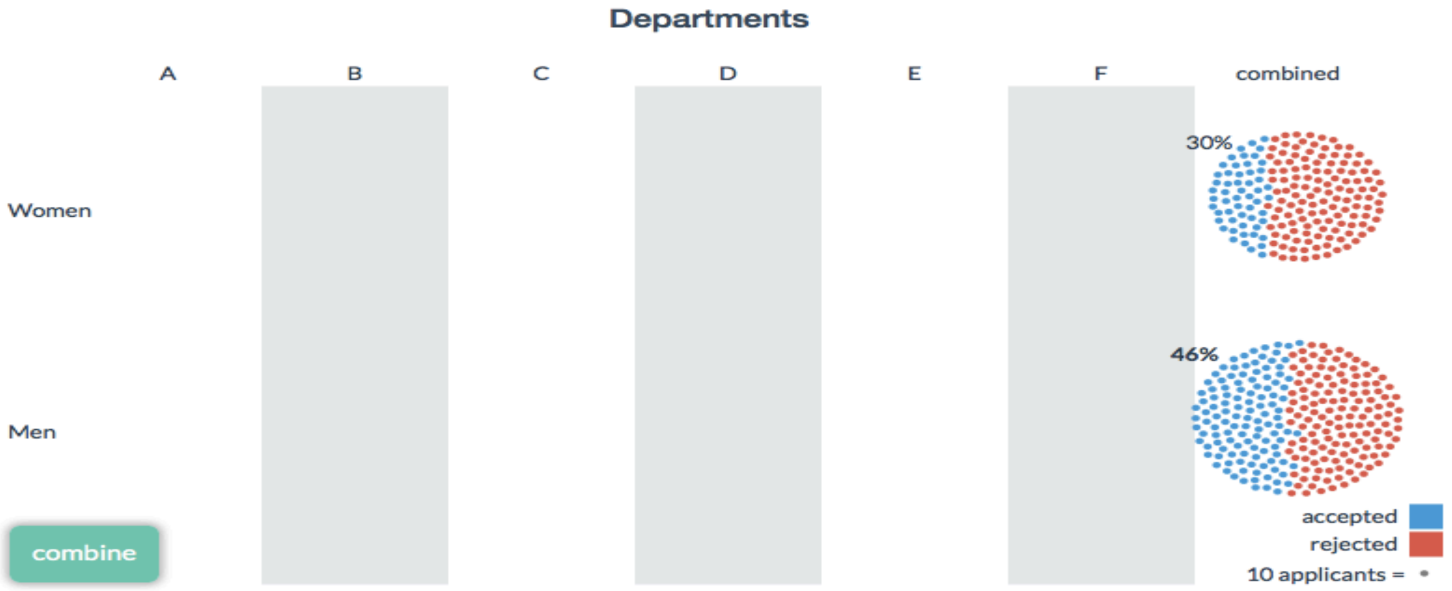
Summary of original paper



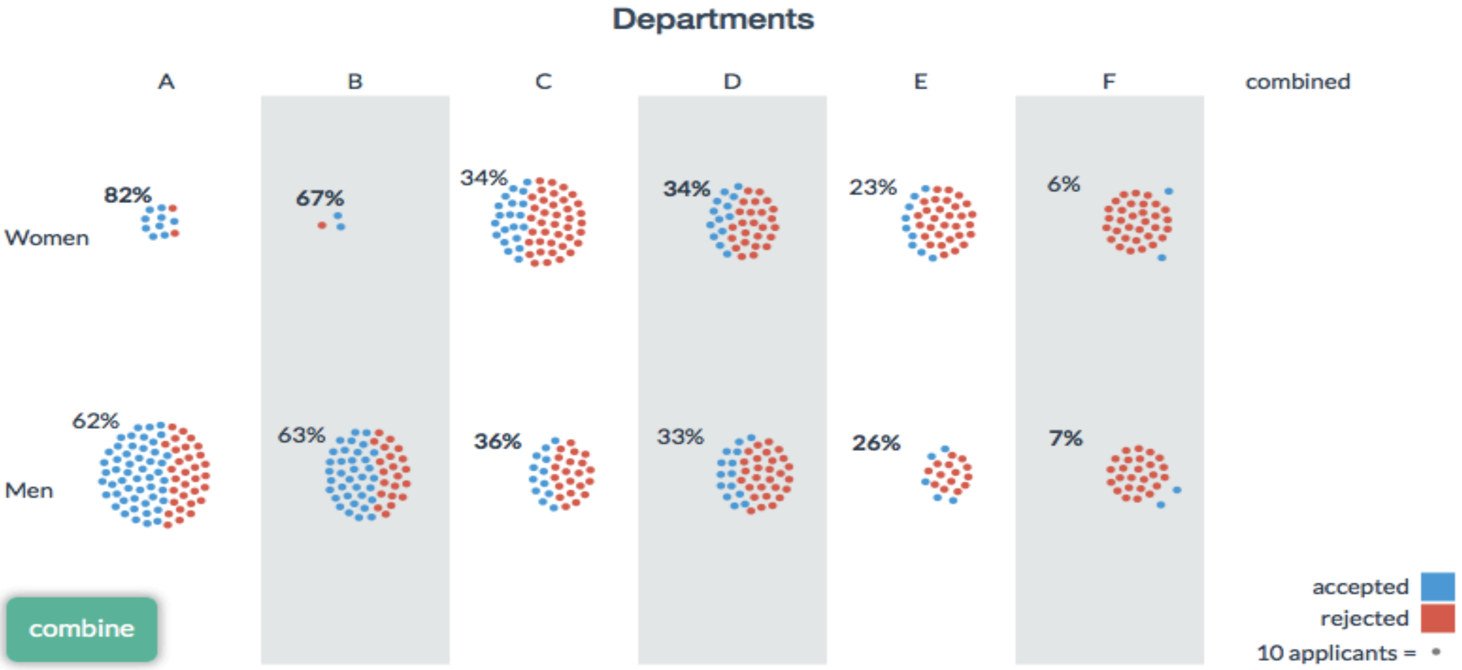
Examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. Examination of the disaggregated data reveals few decision-making units that show statistically significant departures from expected frequencies of female admissions, and about as many units appear to favor women as to favor men. If the data are properly pooled, taking into account the autonomy of departmental decision making, thus correcting for the tendency of women to apply to graduate departments that are more difficult for applicants of either sex to enter, there is a small but statistically significant bias in favor of women. The graduate departments that are easier to enter tend to be those that require more mathematics in the undergraduate preparatory curriculum. The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

Source: PJ Bickel et al 1973

Visualisation of Simpson's paradox in Berkeley grad admissions



Aggregated data



Departmental data

Source: vudlab.com/simpsons/

Simpson's paradox

- explained variable Y (*response*)
- observed explanatory variable X *(predictor)*
- lurking explanatory variable Z (may be known or suspected)

Effect of the observed explanatory variable on the explained variable changes substantially (even qualitatively) when lurking variable is taken into account.

Some continuous examples: weight vs height (lurking: gender), 1st salary vs graduation time (lurking: course), alcohol consumption vs IQ (lurking: individual), health care quality vs salary (social class)

Some discrete examples: Maori representation in New Zealand jury pools, graduate admission and gender