Keele University

Statistics & Experimental Design with R

Barbara Kitchenham
Keele University

Quasi-Experiments

Quasi-Experiments

- Experiments where it is impossible or unethical to apply randomization
  - When factor of interest cannot be changed
    - E.g. gender
    - University education
- Within-subject experiments in SE
  - Difficult to find large number of qualified participants
  - So use individuals as their own control
- Importance
  - Are used to assess impact of program change
    - I.e. major business/social changes
  - In context of SE
    - Adoption of CMM
    - Change from 3GL to OO programming

Causal Inferences

- Quasi-experiments must show
  - Cause Precedes Effect
    - Quasi-experiments manipulate the treatment to ensure that it occurs before the effect
    - Same for randomised experiments
  - Cause co-varies with Effect
    - Covariation is usually established statistically
    - Same for randomised experiments
  - Alternative explanations for the effect are implausible
    - Basic problem for quasi-experiments
    - Cannot argue based on randomisation

Basic Principles for QE Design

- Identification and study of plausible threats to internal validity
  - What threats could plausibly have caused the observed relationship treatment-outcome
- Primacy of control by design
  - Adding design elements aims to prevent threats or provide evidence about them
- Coherent pattern matching
  - A complex prediction made about the outcomes that few alternative explanations can match

Basic Forms of Quasi-Experiment

- Type 1: Experiment-like studies
  - Subjects use different methods under controlled conditions
- Type 2: Large scale surveys of trends
  - Interrupted time series
  - Regression Discontinuity
  - Differences in Differences

Design elements

- Time
  - Most quasi-experiments take place over a time period
- Treatment
  - A policy or method intended to cause some measurable affect to change
- Controls
  - Units not receiving the treatment that are matched in some way to the units receiving treatment
- Pre-test
  - Measurements taken before the treatment condition is applied
- Post-test
  - Measurements taken before the treatment condition is applied

Design Variants

- Post-Test only
  - Introduce change then take one measurement

$$X \qquad O_1$$

  - Weakest possible design
    - No way of knowing whether anything changed
    - No way of knowing what would have happened without the treatment
- All other designs add elements to address these weaknesses

Adding Pre-Test Observations

- Pre-Test-Post-Test

$$O_1 \qquad X \qquad O_2$$

  - Initial observations as a "control"
  - With only one before and after measurement the design is still fairly weak
    - Effect could be associated with some other event
- SE Quasi-Experiment
  - Participants
    - Volunteers from set of available people
      - Read a program and identify defects
      - Receive training in defect detecting method
      - Read another program and identify defects

Pre-test & Post-test Patterns

- Adding more observations and treatment changes strengthens design
  - Pretest-Posttest removing treatment

$$O_1 \quad X \quad O_2 \qquad O_3 \quad \cancel{X} \quad O_4$$

  - Pretest-Post-test Repeated treatment

$$O_1 \quad X \quad O_2 \quad \cancel{X} \quad O_3 \quad X \quad O_4$$

- If the observations follow pattern of interventions
  - Difficult to argue that they are not related
  - But may be vulnerable to a single chance event

Independent Control Groups

- Post-test designs with control group but no pre-test

$$
\begin{array}{cc}
X & O_1 \\
\hline
 & O_1
\end{array}
$$

  - Weak because the groups may differ on more than just treatment
- SE Experiment Example
  - Students volunteer for extra courses on Formal methods
  - Volunteers and non-volunteers compared on examination results
  - Results attributed benefits of Formal methods
- Adding more pre- and post-test measures again strengthens the design

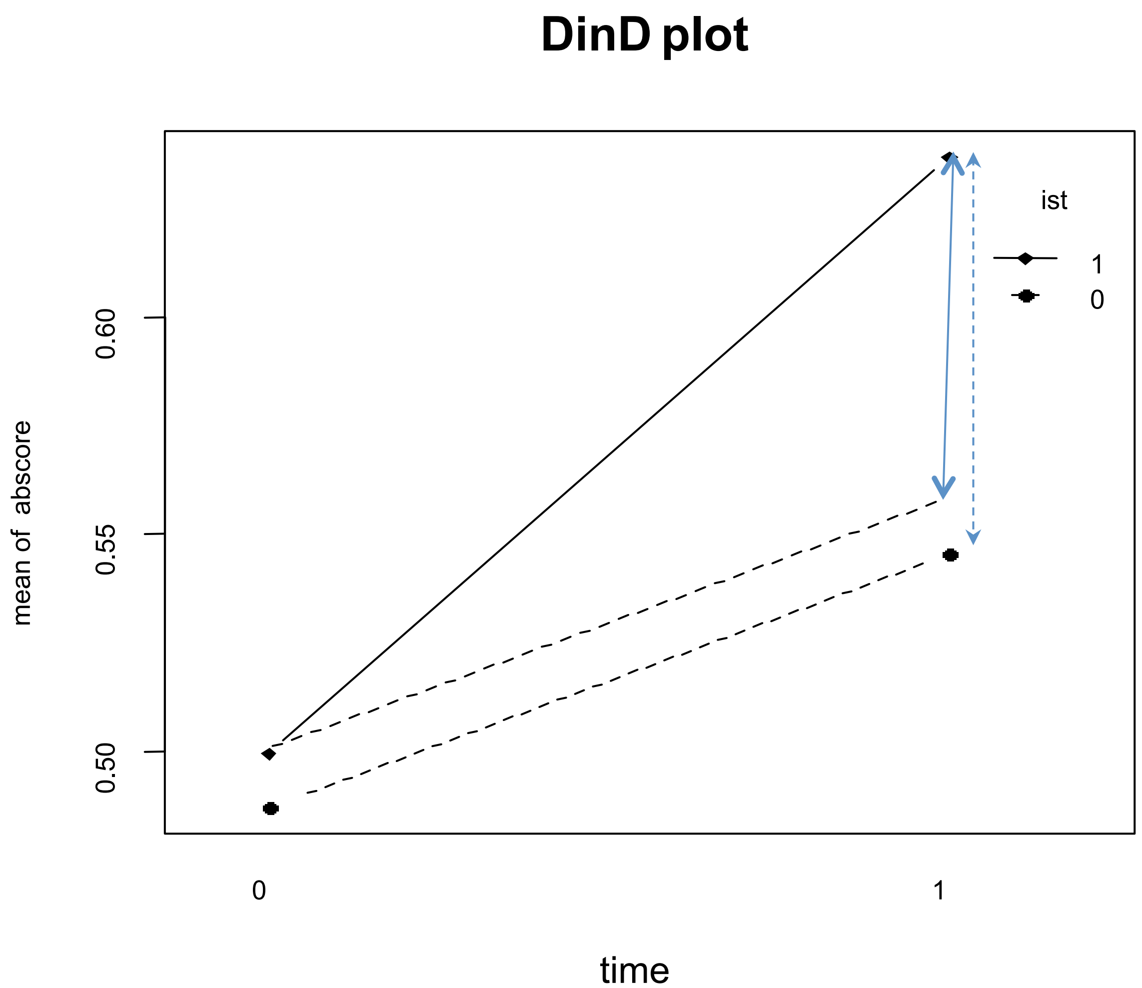
Difference in Differences Designs

- Pre- and Post-tests with controls

$$
\begin{array}{ccc}
O_1 & X & O_2 \\
\hdashline
O_1 & & O_2
\end{array}
$$

- Matched groups with
  - One group receives intervention (T)
  - Other group doesn't (C)
  - Two time periods
    - Before Treatment Time 0
    - After treatment Time 1
- Not a simple two-way analysis
  - Treatment effect based on four group means
    - Effect = T1-C1+ (T0-C0)
    - Period 2 difference adjusted for Period 1 difference

Example



Analysing D-in-D designs

- Can be analysed as a linear combination of mean values
    - Effect = T1-C1+ (T0-C0)
  - Assumes common within-group variance (s^2)
    - For independent groups $s_E^2 = 4\frac{s^2}{n}$
- Alternatively use regression and dummy variables
  - Time (T) is 1 if time period=1 else 0
  - Treatment (Tr) is 1 for treatment group, 0 for control
  - Treated group (TG) is 1 for treatment group in Time Period 1 else 0

Cross-Over Designs

- When comparing two treatments
- Each participant exposed to both treatments
  - Assignment to order randomized

X_1	O_1	X_2	O_2
X_2	O_1	X_1	O_2

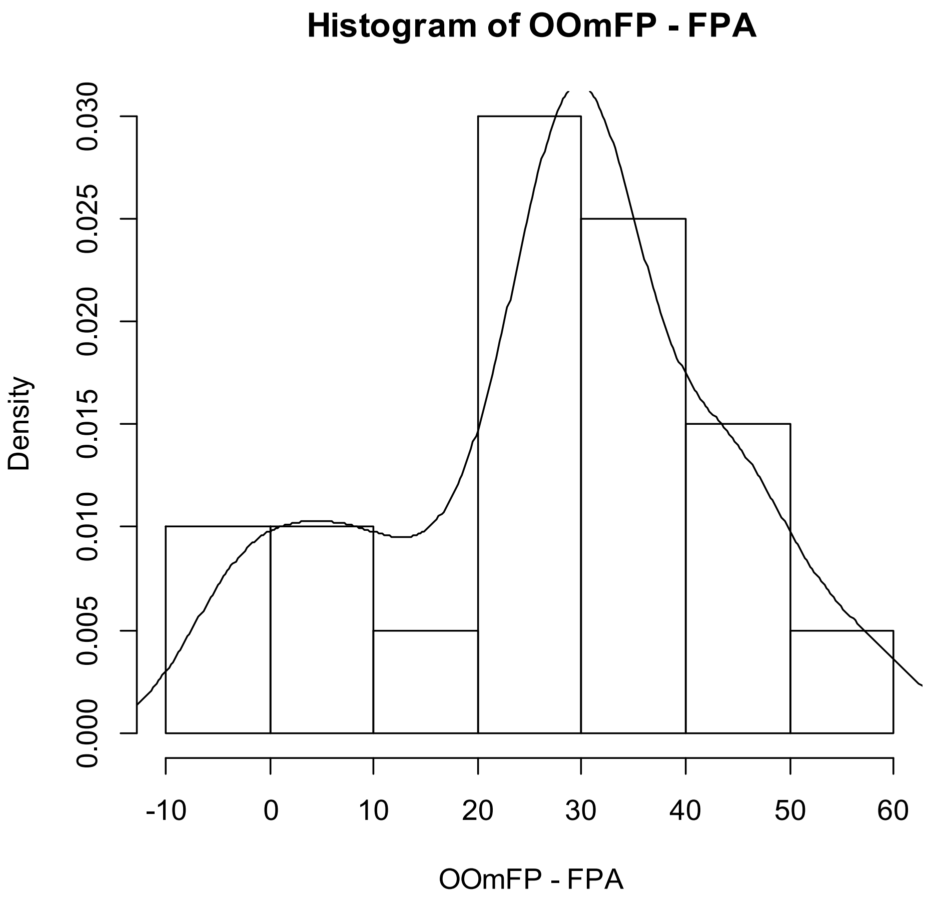
  - Proper analysis removes period effect
    - E.g. general task performance improvement that is independent of treatment
  - Still vulnerable to period×treatment interaction
  - Can be improved by additional pre- and post-tests
- Design is very popular in SE experiments

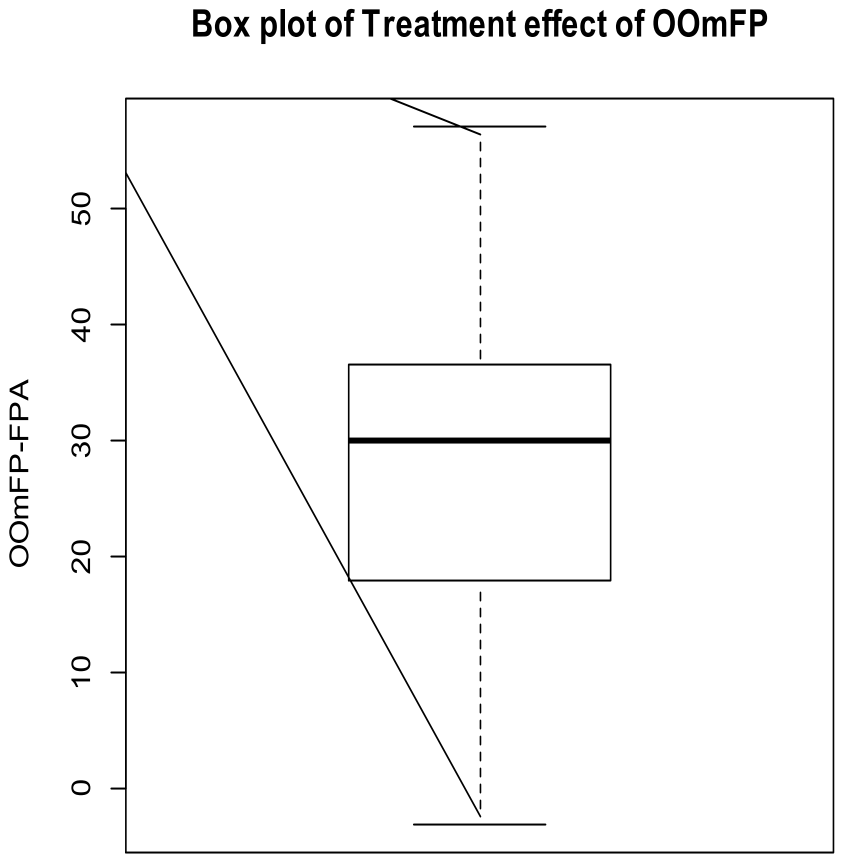
Cross-Over Model

- Model based on
  - π the period effect due to general difference between period 1 and 2
  - τ the treatment effect i.e. difference between T_A and T_B
  - λ_A and λ_B the interaction due to doing A before B and vice-versa – for analysis, assumed approximately 0
  - μ_j the "effect" due to participant j

Participant	Expected Response		Cross-Over Difference	Period Difference
	Period 1	Period 2	T_A-T_B	P2-P1
j	$\mu_j + \tau$ (Treatment A)	$\mu_j + \pi + \lambda_A$ (Treatment B)	$\tau - \pi - \lambda_A$	$\tau - \pi - \lambda_A$
k	μ_k (Treatment B)	$\mu_k + \tau + \pi + \lambda_B$ (Treatment A)	$\tau + \pi + \lambda_B$	$-\tau - \pi - \lambda_B$
Sum			$2\tau + \lambda_B - \lambda_A$	$-2\pi + \lambda_B - \lambda_A$

SE Cross-Over Example

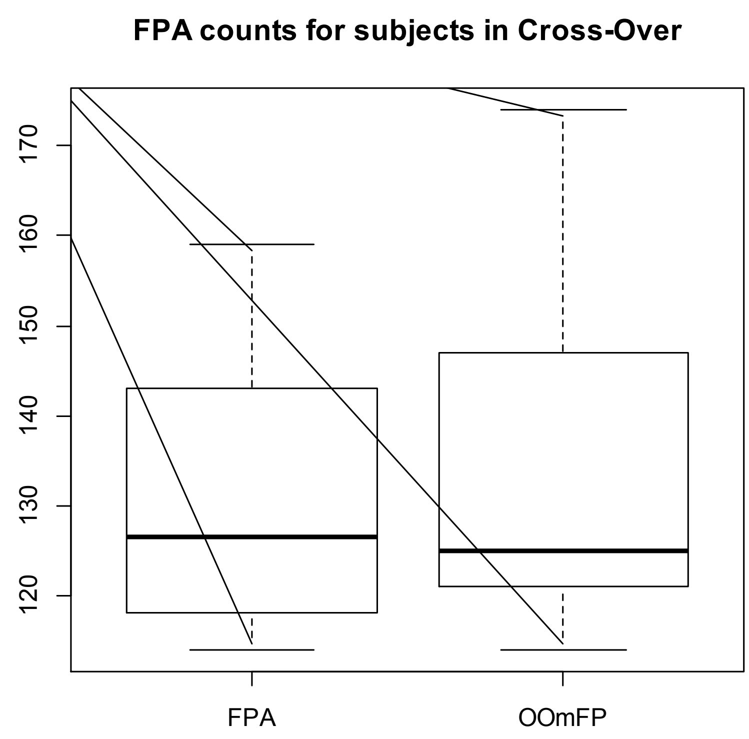




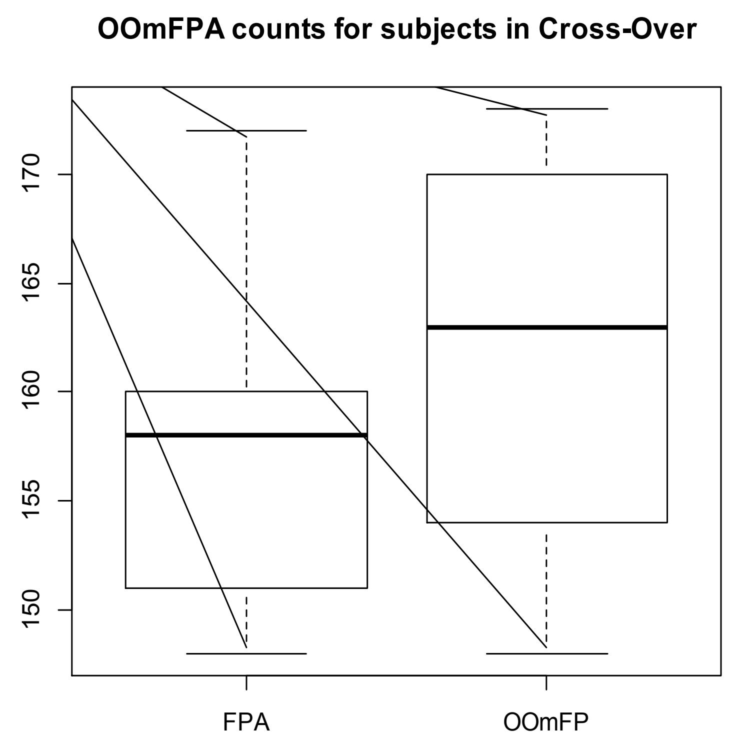
Analysis

- Comparing two FPA versions
- 20 participants count same document
  - 10 used FPA first
  - 10 used OOmFPA first
- Period effect= -0.45
- Treatment effect =27.25
  - Use standard "t" test on Cross-over values (i.e. differences)
    - Variance of Cross-over values=259.04
    - SE treatment effect= 3.6
    - T=7.57 with 19 d.f. Critical Value=2.093 (two-sided, p=0.05)
  - Alternatively use trimmed mean
    - If concerned about non-normal distribution
- Not so simple if groups not same size and period effects significant

Cross-Over Example



Label indicates which treatment was first

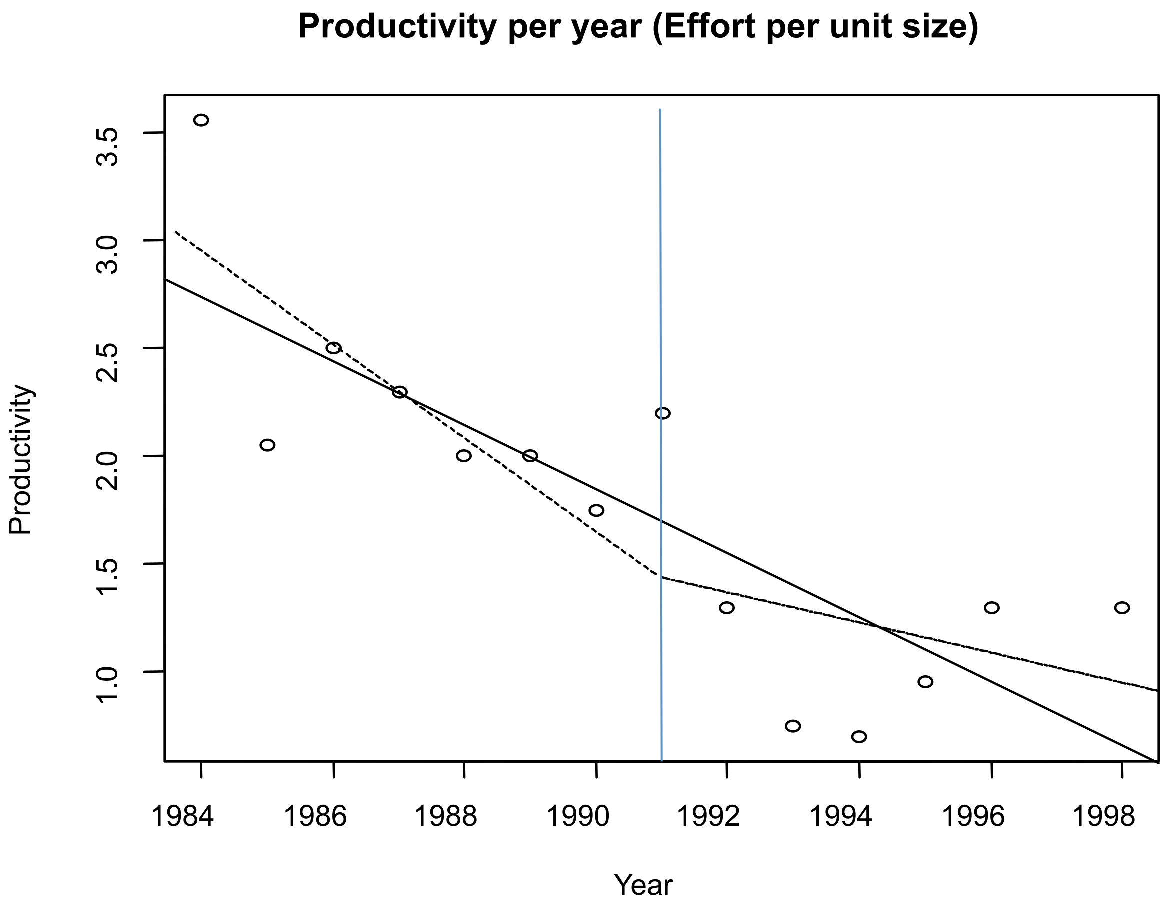


Label indicates which treatment was first

Large Scale Interventions

- Interrupted Time-Series
  - Based on taking observations at many points before and after intervention
    $$O_1 \quad O_2 \quad O_3 \quad O_4 \quad O_5 \quad X \quad O_6 \quad O_7 \quad O_8 \quad O_9 \quad O_{10}$$
  - Estimate Regression lines before and after intervention
  - Look for difference in slope or intercept
- Still may be a confounding effects
  - Need to be listed and accounted for
  - Changes in measurement process could affect results
- As always adding extra elements to design can help
  - E.g plotting another variable that the treatment should NOT effect

SE Example CMM Introduction



Interrupted Time Series Model

- Analyses is based on a specific model

$$Sc_{ijk} = \beta_0 + \beta_1 Year_i + \beta_2 Group_j + \beta_3 TP2Year_i + \epsilon_{ijk}$$

  - $Group_j$ is dummy variable identifying observations record before ($Group_1$=0) or after ($Group_2$=1) the intervention
    - β1 >0 implies a change in intercept
  - $Year_i$ (or any appropriate time period) identifies when the observations were recorded
    - β2 >0 implies a common regression line in the two time periods
  - $TP2Year_i$ refers to each year in the second time period (i.e. when the dummy variable Group=1)
    - B3>0 implies the slope of the regression line is different for the second time period

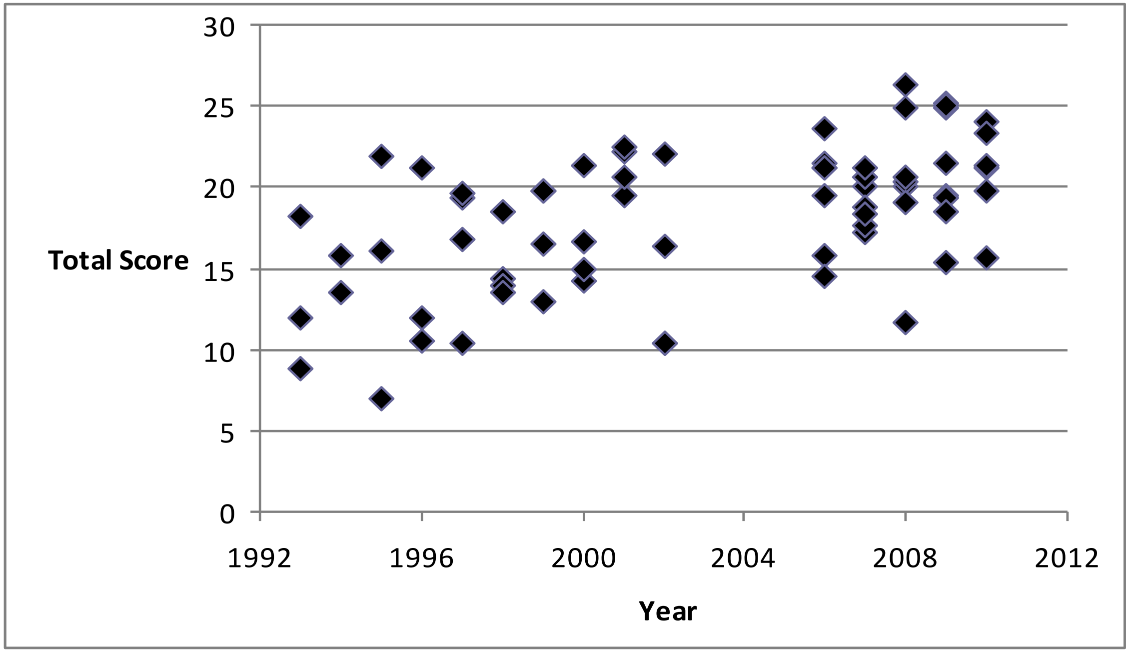
Common Problems with Interrupted Time series

- Gradual rather than abrupt changes
  - So change is not clear cut
- Delayed effects
  - Effects take place some time after change introduced
- Short time series
- Insufficient data points for statistical analysis

SE Example

- Assessing the quality of SE experiments and quasi-experiments
- Investigated whether there was an improvement
  - Due to text book & articles in early 2000's
- Used two measures
  - Subjective assessment
  - Quality scale based on 9 questions
- Evaluated articles from TSE,IST,JSS and ESJ
  - 70 articles in all,
  - Assessed separately by three different people
- Selected papers from years 1993 – 2010
  - Omitted years 2003-2005
  - Because those would be a period of transition

Outcome of Experiment



- Analysis based on average score for each paper
- Only b_1 significantly different from 0
- So common trend before and after 2004

Regression Discontinuity

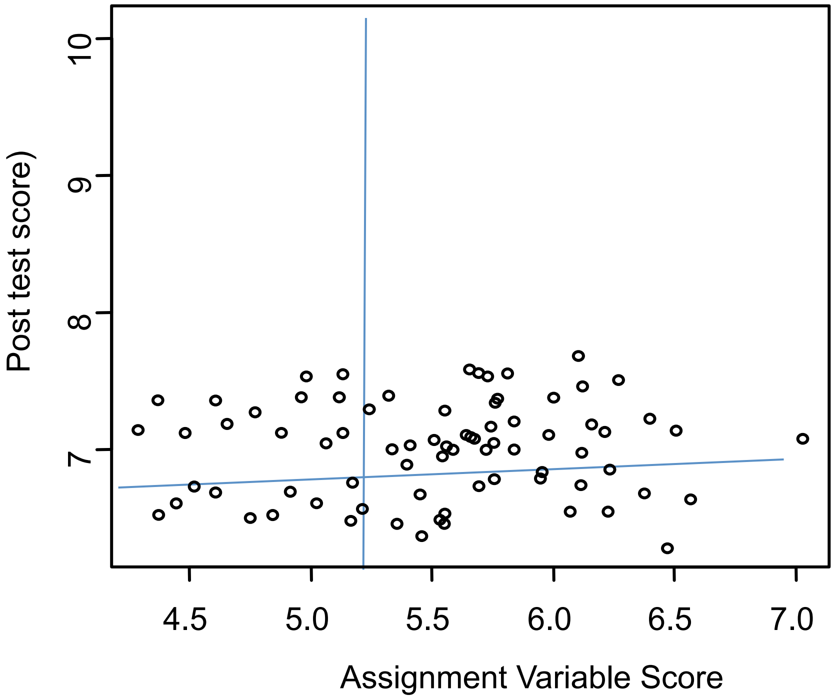
- Experimenter assigns participants to two or more treatment conditions with a post-test
  - The assignment procedure is based on some measurement taken prior to treatment

O_A	C	X	O_2
O_A	C		O_2

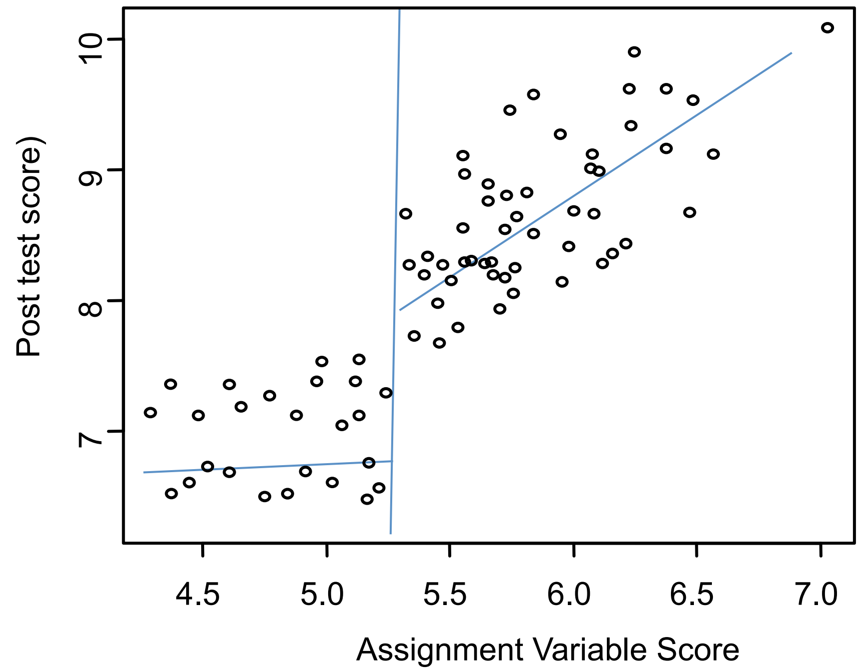
- Control and Treatment group outcomes plotted against post-test measure

Regression Discontinuity

Before



After



Summary

- Quasi-experiments
  - Not second class citizens
  - Often impossible to do randomized experiments
    - Particularly in field
- With appropriate designs
  - Quasi-experiments can be extremely reliable
- Often need specialised analysis to match the specialised design
- Also need to consider how to argue that results can be generalised.