Lecture 7: Quasi-experimental Design

References

- Donald T. Campbell, Julian Stanley, *Experimental and Quasi Experimental Designs for Research*, Houghton Mifflin Company, **1963**
- John W. Creswell, *Educational research: planning, conducting, and evaluating quantitative and qualitative research, 4th ed.*, Pearson Education, **2012**
- John W. Creswell, J. David Creswell, *Research design: qualitative, quantitative, and mixed methods approaches*, SAGE Publications, **2018**
- Christopher J. Millera, Shawna N. Smith and Marianne Pugatch, *Experimental and quasi-experimental designs in implementation research*, Psychiatry Research, vol. 283, **2020**

Content

- Introduction to quasi-experimental design
- Quasi-experimental designs:
  - Nonequivalent Control Group
  - Counterbalanced Design
  - Time-Series Experiment
  - Equivalent Time-Samples
  - Equivalent Materials
  - Multiple Time-Series
  - Separate-Sample Pretest-Posttest
  - Separate-Sample Pretest-Posttest Control Group
  - Recurrent Institutional Cycle
  - Regression-Discontinuity Analysis
- Focus on epidemiological studies
  - Cohort studies
  - Case-control studies

Quasi-experimental designs

The characteristics of quasi-experimental designs:

- Have manipulation or control
- Generally lack randomization
- Are generally prospective in nature
- Are moderate in scientific validity

Quasi-experimental design

The following research designs are believed to be sufficiently probing, however, to be well worth employing where more efficient probes are unavailable.

The notion that experiments never "confirm" theory, while correct, so goes against our attitudes and experiences as scientists as to be almost intolerable.

Varying degrees of "confirmation" are conferred upon a theory through the number of plausible rival hypotheses available to account for the data.

Nonequivalent Control Group

In this design, a popular approach to quasi-experiments, the experimental and the control groups are selected without random assignment. Both groups take a pretest and posttest. Only the experimental group receives the treatment.

Group A	O	_____	X	_____O
Group B	O	_____		O

Do you remember the example with the impact of a code inspection rate on a software quality?
For this experiment we are likely not able to form the equal randomized teams, but have to experiment with existing ones according to the Nonequivalent Control Group Design.

Nonequivalent Control Group (2/3)

- The groups constitute naturally assembled collectives, as similar as availability permits but yet not so similar that one can dispense with the pretest.
- The more similar the experimental and the control groups are in their recruitment, and the more this similarity is confirmed by the scores on the pretest, the more effective this control becomes.
- Thus, we can regard the design as controlling the main effects of *history*, *maturation*, *testing*, and *instrumentation*: the difference for the experimental group between pretest and posttest cannot be explained by main effects of these variables such as would be found affecting both the experimental and the control group.

Nonequivalent Control Group (3/3)

- *Selection-maturation interaction* (or a selection-history interaction, or a selection-testing interaction) could be mistaken for the effect of X, and thus represents a threat to the internal validity of the experiment.

- If either of the comparison groups has been selected for its extreme scores on observed or correlated measures, then a difference in degree of shift from pretest to posttest between the two groups may well be a product of *regression* rather than the effect of X.

Counterbalanced Design

In this design, experimental control is achieved by entering all respondents (or settings) into all treatments.
The Latin-square arrangement is typically employed, in which four experimental treatments are applied in a restrictively randomized manner in turn to four naturally assembled groups.

Group A	X_1 ___ O	___ X_2 ___ O	___ X_3 ___ O	___ X_4 ___ O
Group B	X_2 ___ O	___ X_4 ___ O	___ X_1 ___ O	___ X_3 ___ O
Group C	X_3 ___ O	___ X_1 ___ O	___ X_4 ___ O	___ X_2 ___ O
Group D	X_4 ___ O	___ X_3 ___ O	___ X_2 ___ O	___ X_1 ___ O

It can be a case that we cannot change inspection rate, so we will manipulate with number of LOC per reviewer, and change these limits after each iteration between our subgroups.

Counterbalanced Design (2/3)

- This design is especially preferred where pretests were inappropriate, and designs like **Nonequivalent Control Group Design** were unavailable.
- The counterbalancing was introduced to provide a kind of equation because random assignment is not possible.
- Occasions are likely to produce a main effect due to repeated testing, maturation, practice, and cumulative carry-overs, or transfer.
- History is likewise apt to produce effects for occasions.
- The Latin-square arrangement keeps these main effects from contaminating the main effects of X_s.

Counterbalanced Design (3/3)

- However, where main effects symptomatize significant heterogeneity, one is probably more justified in suspecting significant interactions than when main effects are absent.

- Thus the apparent differences among the effects of the X_s could be a specific complex interaction effect between the group differences and the occasions. Inferences as to effects of X will be dependent upon the plausibility of this rival hypothesis.

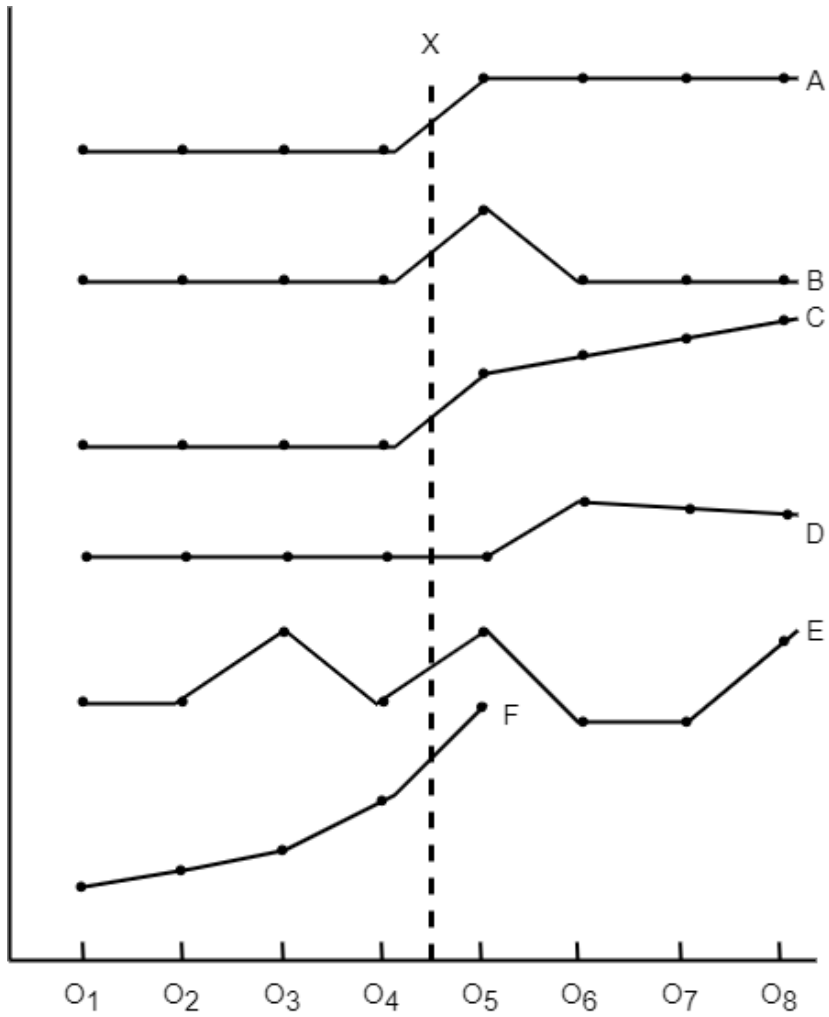
Time-Series Experiment

The essence of the time-series design is the presence of a periodic measurement process on some group or individual and the introduction of an experimental change into this time series of measurements.

$$\textit{Group A} \quad O_1 \text{--} O_2 \text{--} O_3 \text{--} O_4 \text{--} X \text{--} O_5 \text{--} O_6 \text{--} O_7 \text{--} O_8$$

The problem of internal validity boils down to the question of plausible competing hypotheses that offer likely alternate explanations of the shift in the time series other than the effect of X.
In our experiment it can be startup company with one team only. In this case we can look at the defects rate before first release, then change the inspection process and collect data during next release cycle

Time-Series Experiment (2/4)



from Campbell, Stanley pg 38

Time-Series Experiment (3/4)

- Failure to control *history* is the most definite weakness of this design. It might be that not X but some more or less simultaneous events produced the difference in the outcome.

  Experimental isolation is an approach to reduce the effect of history on the experiment.

- The possible concerns about *instrumentation* are related to situations in which a change in the calibration of the measurement device could be misinterpreted as the effect of X.

  To preserve the interpretability of a time series, it would be better to continue to use a somewhat antiquated device rather than to shift to a new instrument.

Time-Series Experiment (4/4)

- *Regression* effects are usually a negatively accelerated function of elapsed time and are therefore implausible as explanations of an effect at O_5 greater than at O_2 or O_4.

- *Selection* as a source of main effects is ruled out if the same specific entities are involved at all observations.

Equivalent Time-Samples

This design can be seen as a form of the **Time-Series Design** with the repeated introduction of the experimental variable.

$$\textit{Group A} \quad X_1 \_\_ O \_\_ X_0 \_\_ O \_\_ X_1 \_\_ O \_\_ X_0 \_\_ O$$

The experiment is applied where the effect of the experimental variable is anticipated to be of transient or reversible character.

For example, we have only one team and less time for experimentation, thus through the number of iterations we change the frequency of inspections from weekly to daily ones and back

Equivalent Time-Samples (contd.)

While the logic of the experiment may be seen as an extension of the time-series experiment, the mode of statistical analysis is more typically similar to that of the two-group experiment:
The significance of the difference between the means of two sets of measures is employed.

- *History* is controlled by presenting X on numerous separate occasions, making extremely unlikely a coincidence of extraneous events.

- The other sources of invalidity are controlled by the same logic as in basic time-series design.

Equivalent Materials

Closely allied to the **Equivalent Time-Samples Design**, but the equivalence of samples of materials to which the experimental variables being compared are applied.

$$\textit{Group A} \quad M_a X_1 \_\_ O \_\_ M_b X_0 \_\_ O \_\_ M_c X_1 \_\_ O \_\_ M_d X_0 \_\_ O,$$

where M_s indicate specific materials, the sample M_a, M_c, etc., being, in sampling terms, equal to the sample M_b, M_b, etc.

For example, we can use different types of the checklists for daily inspections, performed by our experimental team - in this case the checklists are controlled variable

Equivalent Materials (contd.)

- Equivalent materials are required whenever the nature of the experimental variables is such that the effects are enduring and the different treatments and repeats of treatments must be applied to nonidentical content.

- Like in the equivalent time-samples design, this design has internal validity on all points, and in general for the same reasons.

- Reactive arrangements seem to be less certainly involved because of the heterogeneity of the materials and the greater possibility that the subjects will not be aware that they are getting different treatments at different times for different items.

Multiple Time-Series

This design contains within it the **Nonequivalent Control Group Design**, but gains in certainty of interpretation from the multiple measures plotted, as the experimental effect is in a sense twice demonstrated, once against the control and once against the pre-X values in its own series.

Group A O ___ O ___ O ___ X ___ O ___ O ___ O

Group B O ___ O ___ O _____ O ___ O ___ O

To avoid threats of pure Time Series and Nonequivalent Control Group designs, we observe two teams simultaneously during two release cycles with introduction of daily code reviews for the experimental team between cycles

Multiple Time-Series (contd.)

- The *selection-maturation interaction* is controlled to the extent that, if the experimental group showed in general a greater rate of gain, it would show up in the pre-X O_s.

- Because *maturation* is controlled for both experimental and control series, the difference in the selection of the groups operating in conjunction with maturation, instrumentation, or regression, can hardly account for an apparent effect.

- However, an interaction of the selection with *history* is possibile.

Separate-Sample Pretest-Posttest

For large populations it may often happen that although one cannot randomly segregate subgroups for differential experimental treatments, one can exercise something like full experimental control over the when and to whom conduct observations, employing random assignment procedures.

Group A R O _____(X)
Group B R X _____O

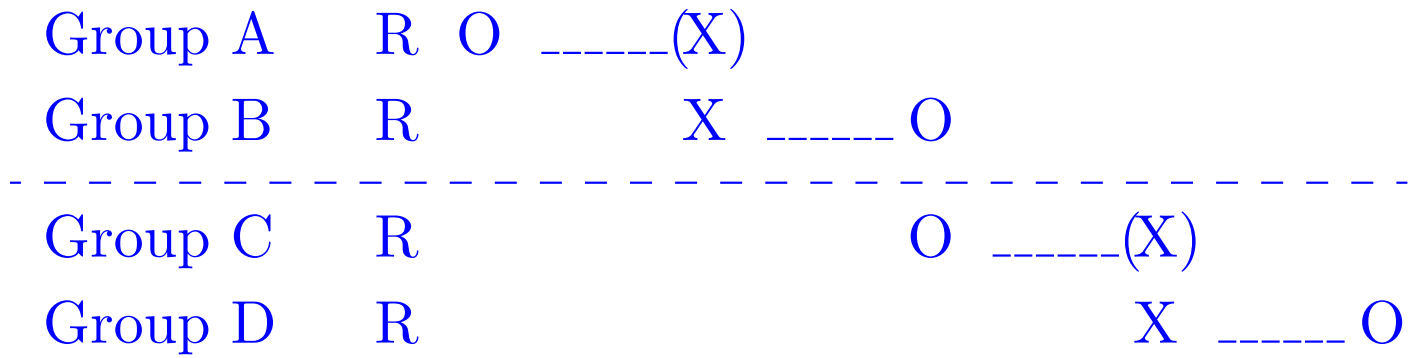
(X) standing for a presentation of X irrelevant to the argument.
One sample is measured prior to the X, an equivalent one subsequent to X.

Example: when the whole organization introduces new standards to software quality, we can collect data from one unit in advance, and then compare them against another unit after policies changing

Separate-Sample Pretest-Posttest (2/4)

The design is not inherently a strong one. Nevertheless, it may frequently be all that is feasible, and is often well worth doing.

- Repeating of this design in different settings at different times controls for *history*:



- But consistent secular historical trends or seasonal cycles remain uncontrolled rival explanations.

Separate-Sample Pretest-Posttest (3/4)

- *Maturation* is unlikely to be invoked as a rival explanation. The samples are large enough and heterogeneous enough so that subsamples of the pretest group differing in maturation can be compared.

- For pretests and posttests separated in time by several months, *mortality* can be a problem.
  To provide an additional confirmation of effect which mortality could not contaminate, the pretest group can be retested:

Group A	R	O_1	(X)	O_2
Group B	R		X	O_3

Separate-Sample Pretest-Posttest (4/4)

It is characteristic of this design that it moves the laboratory into the field situation to which the researcher wishes to generalize, testing the effects of X in its natural setting. In general, it is superior in external validity or generalizability to the ”true” experiments.

Sep.-Sample Pre.-Post. Control Group

It is expected that **Separate-Sample Pretest-Posttest Design** will be used in those settings in which the X, if presented at all, must be presented to the group as a whole. If there are comparable groups from which X can be withheld (*other divisions of the company*), then a control group can be added.

In **The Separate-Sample Pretest-Posttest Control Group Design** the same specific persons are not retested and thus the possible interaction of testing and X is avoided.

Group A	R	O	_____(X)		
Group B	R			X	_____O
Group C	R	O			
Group D	R				O

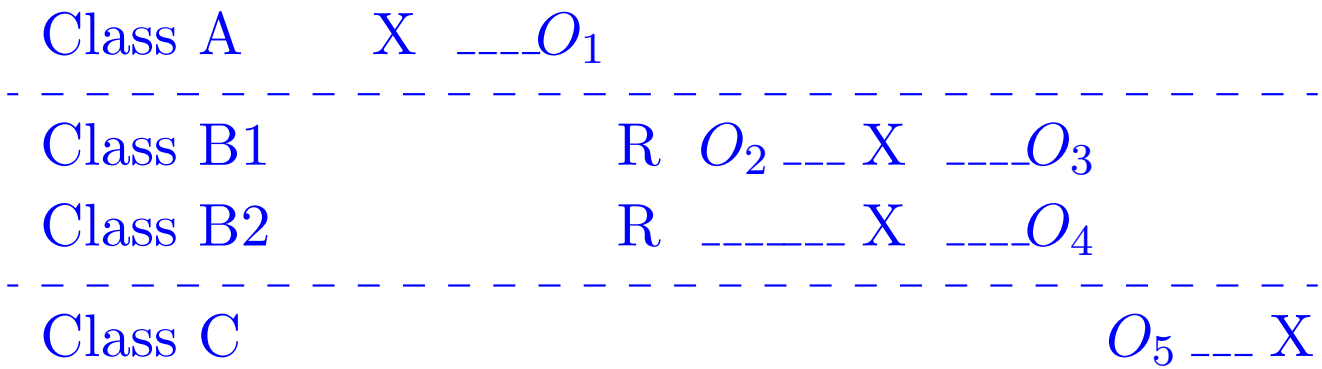
Sep.-Sampl. Pre.-Post. Ctrl. Gr. (contd.)

- The weakness of this design for internal validity comes from the possibility of mistaking for an effect of X a specific local trend in the experimental group which is, in fact, unrelated.

- By increasing the number of the entities involved (organizations, cities, factories, etc.) and by assigning them in some number and with randomization to the experimental and control treatments, the one source of invalidity can be removed, and a true experiment, avoiding the retesting of specific individuals, might be achieved.

Recurrent Institutional Cycle

This design illustrates a strategy for field research in which one starts out with an inadequate design and then adds specific features to control for one or another of the recurrent sources of invalidity.



Recurrent Institutional Cycle (2/3)

- This design combines the "longitudinal" and "cross-sectional" approaches commonly employed in developmental research.
- Comparison between O_1 and O_2 corresponds to the **Static-Group Comparison**, remeasuring the personnel of Class B one cycle later provides the **One Group Pretest-Posttest** segment.
- The effect of X is thus documented in three separate comparisons, $O_1 > O_2$, $O_2 > O_3$ and $O_2 > O_4$.
- The introduction of O_5, that is Class C, tested on the second testing occasion prior to being exposed to X, provides another pre-X measure to be compared with O_4 and O_1, etc., providing a needed redundancy.

Recurrent Institutional Cycle (3/3)

- Such a design lacks the clear-cut control on *history* in the $O_1 > O_2$ and the $O_4 > O_5$ comparisons because of the absence of simultaneity.

- If the cross-sectional and longitudinal comparisons indicate comparable effects of X, this could not be explained away as an interaction between maturation and the selection differences between the classes.

Regression-Discontinuity Analysis

This design is used for illustration of the desirability of exploring in each specific situation all of the implications of a causal hypothesis, seeking novel outcroppings where the hypothesis might be exposed to test.

It is applied in a situation in which ex post facto designs were previously being used.

Let us look by example

Regression-Discontinuity Analysis (2/6)

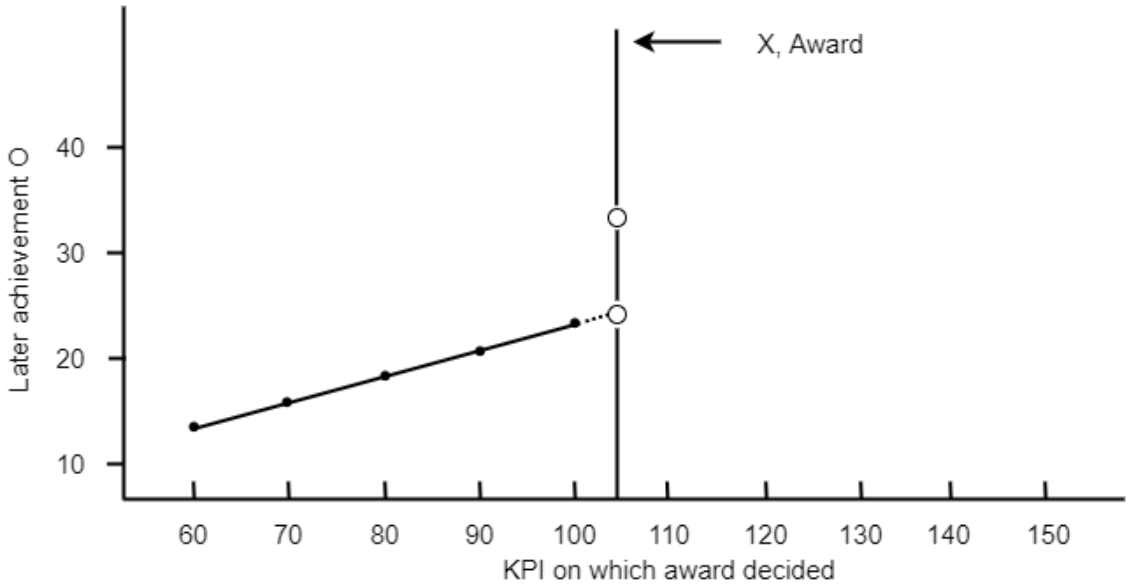
- Awards are made to the most successful employees on the basis of a cutting score on a quantified composite of KPIs. Then applicants receiving and not receiving the award are measured on various O_s representing later achievements, attitudes, etc.

The question is then asked, Did the award make a difference?

- The problem of inference is sticky just because almost all of the qualities leading to eligibility for the award are qualities which would have led to higher performance on these subsequent O_s. We are certain in advance that the recipients would have scored higher on the O_s than the non-recipients, even if the award had not been made.

Regression-Discontinuity Analysis (3/6)

Let us consider a true experiment with **Posttest-Only Control Group Design**



from Campbell, Stanley pg 62

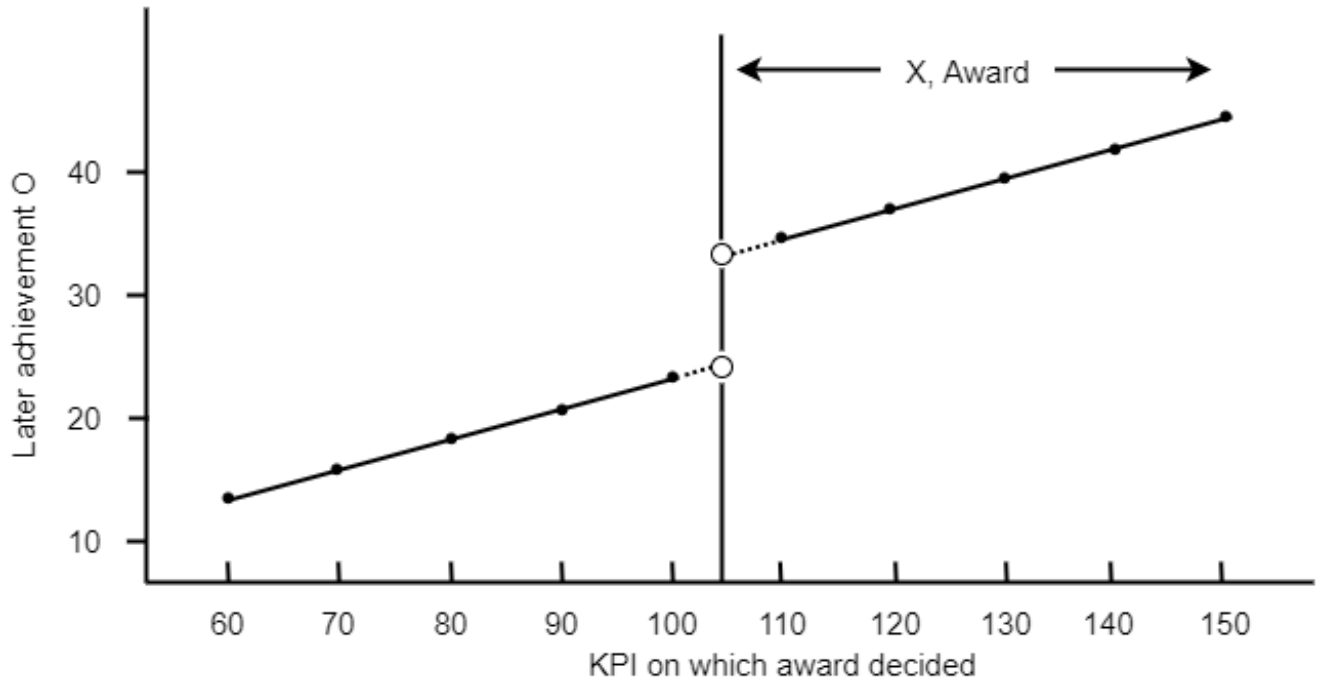
Random assignment would create an award-winning experimental group and a non-winning control. These would presumably perform as the two circle-points at the cutting line.

Regression-Discontinuity Analysis (4/6)

- For this narrow range of abilities, a true experiment would have been achieved.

- **The Regression-Discontinuity Analysis** attempts to substitute for this true experiment by examining the regression line for a discontinuity at the cutting point which the causal hypothesis clearly implies.

Regression-Discontinuity Analysis (5/6)



from Campbell, Stanley pg 62

If the outcome were as diagrammed, and if the circle points represented extrapolations from the two halves of the regression line rather than a randomly split tie-breaking experiment, the evidence of effect would be almost as compelling as in the true experiment.

Regression-Discontinuity Analysis (6/6)

- The hypothesis is clearly one of intercept difference rather than slope, and that the location of the step in the regression line must be right at the X point, no ”lags” or ”spreads”.
- Assumptions of linearity are usually more plausible for such regression data than for time series.
- The most efficient test would be a covariance analysis, in which the award decision score would be the covariate of later achievement, and award and no-award would be the treatment.

Quasi-experimental designs summary (1/3)

	Sources of invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction with selection	Interaction of testing and X	Interaction of selection and X	Reactive arrangements	Multiple interference
Time Series	–	+	+	?	+	+	+	+	–	?	?	
Equivalent Time Samples Design	+	+	+	+	+	+	+	+	–	?	–	–
Equivalent Materials Samples Design	+	+	+	+	+	+	+	+	–	?	?	–
Nonequivalent Control Group Design	+	+	+	+	?	+	+	–	–	?	?	

"–" indicates a definite weakness,

"+" indicates that the factor is controlled,

"?" indicates a possible source of concern,

blank space indicates that the factor is not relevant.

Quasi-experimental designs summary (2/3)

	Sources of invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction with selection	Interaction of testing and X	Interaction of selection and X	Reactive arrangements	Multiple interference
Counterbalanced Designs	+	+	+	+	+	+	+	?	?	?	?	–
Separate-Sample Pretest-Posttest Design	–	–	+	?	+	+	–	–	+	+	+	
Separate-Sample Pretest-Posttest Control Group Design	+	+	+	+	+	+	+	–	+	+	+	
Multiple Time-Series	+	+	+	+	+	+	+	+	–	–	?	

"–" indicates a definite weakness,

"+" indicates that the factor is controlled,

"?" indicates a possible source of concern,

blank space indicates that the factor is not relevant.

Quasi-experimental designs summary (3/3)

	Sources of invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction with selection	Interaction of testing and X	Interaction of selection and X	Reactive arrangements	Multiple interference
Institutional Cycle Design:												
$O_2 > O_1$ & $O_5 > O_4$	+	−	+	+	?	−	?		+	?	+	
$O_2 < O_3$	−	−	−	?	?	+	+		−	?	+	
$O_2 < O_4$	−	−	+	?	?	+	?		+	?	?	
Regression Discontinuity	+	+	+	?	+	+	?	+	+	−	+	+

"–" indicates a definite weakness,

"+" indicates that the factor is controlled,

"?" indicates a possible source of concern,

blank space indicates that the factor is not relevant.

Focus on epidemiological studies

Epidemiological studies play a major role in software engineering:

- they tend to belong to the family of quasi-experimental design
- they are often retrospective
- they often only rely on data collected without intervention of the subjects involved, apart from privacy and confidentiality agreements
- they often do not require any intervention, which could be perceived as detrimental by the management of the organization under observation

Observational vs. Interventional

Epidemiological studies can be classified as:

- observational, when no intervention takes place
- interventional, when some of the subjects make some changes

An observational study occurs when I analyse after the fact the introduction of the daily standup meeting in a team to report to the manager the success of the initiative

Descriptive vs. Analytic

Observational studies can further be subdivided into:

- descriptive, when the distribution of an outcome in a population is studied and a hypothesis is drawn, but there is no attempt to pinpoint the treatment that brought to a specific outcome
- analytical, when an hypothesis is tested to determine the causes or the treatment that lead to a given situation

A descriptive observational study occurs when I collect the perceived needs in software companies, as reported in the first lecture.

Cohort studies (1/2)

A cohort study is a type of analytical study where two aspects are considered:

- an outcome, such as the emergence of a bug in the code,
- an influencing factor, which could be a *risk factor*, such as an event that is considered as a potential cause of the bug.

In summary:

- At the beginning the cohort of subjects under consideration does not have the considered outcome.
- Then, some of the subjects get exposed to the risk factor, for instance working over time
- Thus, the cohort is partitioned in two groups, those exposed to such factor and those not exposed.
- Finally the emergence of the outcome is considered in the two groups.

Cohort studies (2/2)

Note that it is important:

- To follow the entire cohort over time:
  - to determine with precision the subjects exposed to the influencing factor
  - to compare the incidence overtime of the outcome of interest in both the exposed and unexposed subjects.

An example presented in the papers supplied for the research project is the analysis of the factors that may cause the injection of bugs:

- developers were tracked non invasively for more than an year,
- emergent bugs were detected for 2 years,
- then for each bug via the version control system the time of bug injection was determined,
- via the data measured non invasively the cause of the injection of bugs and the context were determined.

Case-control studies (1/2)

A case-control study is a type of analytical study where:

- first, subjects with the outcome of interest are selected,
- a representative of group subjects coming from the same population but without such outcome are selected,
- potential influencing factors are then identified,
- the two groups are then compared to determine the influencing factors that are likely to trigger the outcome of interest.

Case-control studies (2/2)

An ongoing case-control study that we are currently performing relates to remote work:

- the *outcome of interest* is the perceived reduction of performances due to the covid-induced remote work,
- the *subjects* with the outcome of interest were subjects claimed to perceive a reduction in work performances,
- the *control group* from the same population were other developers who did not perceive such reduction,
- the *factors* were the process in use, the number of online meetings, the previous habits related to remote work.

Details in the paper sent in Telegram.

Cohort vs. Case-control (1/3)

The differences between cohort and case-control studies refers to the following aspects:

- the selection of the subjects,
- the speed at which the study can be conducted and the founding needed to complete the study,
- type of outcomes that can be studied
- the quality of the evidence produced by the study

Cohort vs. Case-control (2/3)

In details:

- selection:
  - in cohort studies:
    - the subjects are selected before they have developed the outcome of interest, and
    - the subjects are then classified according to whether or not they were exposed to the possible influencing factor, lastly
    - the subjects are then followed over time to see if they develop the outcome of interest.
  - in case-control studies:
    - subjects are chosen after according to whether they have the outcome or not,
    - then one must assess if the subjects were exposed to the possible influencing factor.

Cohort vs. Case-control (3/3)

- speed:
  - case-control studies can be done more quickly and cost less they measure the exposure to the risk factor and the outcome at the same time;
  - whereas cohort study take long periods of time, and thus more money, to be completed;
- long-term effects:
  - with case-control studies, one can study outcomes with a long induction period, since the starting point of the study is the identification of people with the outcome of interest,
  - this is not possible in cohort study;
- quality of evidence:
  - in a cohort study, we have a superior quality of evidence than in case-control studies, since they usually take place over a longer time,
  - therefore, there is more time to gather all the necessary evidence.

End of lecture 7