Part IV

Other Fundamental Evaluation Methodologies

Chapter 16

Design of Experiments

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the Design of Experiments (in short, DoE).

16.1 Basic Concepts

The basic concepts come from early work of DoE [91, 200, 100, 60, 59].

Experimental Units are objects under study, ranging from an individual, e.g., a person, to aggregated entities (e.g., a class).

Responses are quantitative metrics obtained on the experimental units through measurements, distinct from categorical variables.

Factors represent parameters whose changes in value will result in variations in the response. A factor has several levels. The *factor level* could be both the quantitative variable and the categorical variable.

Interactions are any systematic dependencies between factors.

Experimental Condition Groups are the combinations of all factor levels. If there are n input factors, and each factor i has m_i factor levels, then there are a total of $\prod_{i=1}^{n} m_i$ experimental condition groups.

An *Experiment* is to apply different experimental condition groups to the experimental units with determined and controlled factor levels.

Factor Effects are the response difference between the average of experimental condition group runs at the different levels for corresponding factors.

A *Model Equation* is an equation that establishes the model to characterize the relationships between responses and factor effects.

$l^k r$ *Factorial Designs* select k factors, each factor selects l levels, and each experimental condition group replicates r times. Its model equation calculates all effects of selected factors or their interactions.

$2^{k-p} r$ *Factorial Designs* select k factors, each factor selects 2 levels, with 2^p factors or interactions are confounded to reduce costs, and each experimental condition group

[1] Mr. Chenxi Wang is the primary contributor of this author.

replicates r times. Its model equation calculates the aggregated effects from 2^p factors or interactions.

Full Factorial Designs separately estimate all effects of selected factors or their interactions through the model equation. Full factorial designs include $2^k r$, $l^k r$, and general factorial designs. Here, 2 or l indicates the factor level, k means the number of factors, and r is the number of replications. General factorial design also involves studying k factors with r replications, but the selection of levels for each factor is not restricted to the same fixed value (such as 2 in a $2^k r$ factorial design or l in $l^k r$ factorial design).

Fractional Factorial Designs compute aggregated effects of selected factors or their interactions through the model equation. Classical fractional factorial designs include $2^{k-p} r$ factorial designs.

Residual error represents the portion of the response variation not explained by the model that characterizes the relationships between responses and factor effects.

Homoscedasticity means constant variance, while *Heteroscedasticity* indicates non-constant variance.

16.2 Problem Statement

DoE is a structured, statistical methodology developed to systematically infer a linear model between multiple input factors and output responses within a process or system [200, 91, 60]. This approach has become foundational in fields ranging from engineering and manufacturing to biology and behavioral sciences, providing researchers with a powerful tool set for optimization, quality improvement, and hypothesis testing [100, 59, 60].

Given an experiment unit or a population of experiment units with one or more measurable responses of interest and a set of controllable input factors, each factor can be set to different levels. The factor effects, which are the additive components that constitute the expected value of the response, are unknown.

The DoE problem could be formally stated as "how to strategically select a limited but representative set of factor-level combinations to run, allowing for the efficient and simultaneous estimation of the factor effects, to derive a reliable model that describes the relationship between the response and the factor effects, while minimizing the confounding of uncontrolled or unknown factors."

The inputs of DoE include factors, the corresponding levels for each factor, experimental units, and the responses measured from them. The outputs of DoE include the relationship between each response and factor effects. The constraint is that experimental resources are limited; it is infeasible to run all possible combinations of factor levels due to prohibitive costs or time constraints. Furthermore, the observed response is often contaminated by the uncontrolled or unknown factors that may influence the response, which obscures the factor effects.

16.3 Basic Assumptions

Assumption 16.1 (Linearity)**.** *The relationship between the response variable and the factor effects, including their interactions and random error, can be adequately represented by a linear model. Formally, the response Y is a linear function of the factor effects:*

$$Y = \mu + \beta_1 + \beta_2 + \cdots + \beta_{i,j} + \cdots + \epsilon, \tag{16.1}$$

where Y is the response, μ is the overall mean of all responses, β's are the factor effects, and ϵ is the random error term. A single subscripted β denotes the main effect of each factor, with the subscript indicating the corresponding factor; a multi-subscripted β represents the interaction effect among the factors indicated by its subscripts.

Assumption 16.2 (Additivity of Effects)**.** *The effects of different factors are separable and additive in nature. That is, the change in the response caused by varying one factor, to a first approximation, is independent of the levels of other factors. Any interaction can be explicitly identified and modeled as a separate term in the model equation, preserving the principle of effect separation.*

Assumption 16.3 (Randomization)**.** *The order in which experimental runs are performed should be randomized. Randomization serves to distribute the potential effects of uncontrolled or unknown factors that may influence the response evenly across all experimental groups. This process mitigates the risk of confounding these nuisance factors with the systematic effects of the factors under study, thereby ensuring unbiased estimates of the effects.*

Assumption 16.4. *The residual error terms are independently and identically distributed (i.i.d.). Specifically, they follow a normal distribution with a mean of zero and constant variance σ^2:*

$$\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2). \tag{16.2}$$

This implies independence of errors, homoscedasticity, and normality.

Assumption 16.5. *The variance of the residual errors is constant across all factor levels. Denoted as $Var(\epsilon) = \sigma^2$, this property of homoscedasticity is crucial for the validity of standard hypothesis tests (e.g., F-tests) and the correctness of confidence intervals derived from the model. Heteroscedasticity can invalidate these inferences.*

16.4 Basic Principles

DoE uses a linear model (called the model equation) that sums additive effects from different factors. For each measured response in the experimental units, its effects comprise the overall mean, the main effects of individual factors, interaction effects between factors, and residual random errors—more details in Section 16.5.2.

The Analysis of Variance (ANOVA) methodology decomposes total response variance into components attributable to between-group variance, within-group variance, and the

square of errors, as elaborated in Section 16.5.2. Through ANOVA [167], it quantifies the contribution of each factor to response variance and tests the statistical significance of factors and their interactions.

By systematically combining factor levels into experimental condition groups and applying the experimental condition groups to the experimental units, DoE enables measurement of responses under structured and controlled conditions. It encompasses three effect components: 1) main effects of individual factors, 2) interaction effects between factors, and 3) random errors from chance fluctuations.

Rather than altering one factor at a time, DoE introduces a framework for simultaneously varying multiple factors, thereby enabling the identification of interaction effects and the construction of predictive models. Typically grounded in factorial or fractional factorial designs, DoE incorporates randomization, replication, and blocking to reduce experimental error and enhance the robustness of inferences.

16.5 Methodology

16.5.1 Classical Factorial Design

The factorial design encompasses both full factorial and fractional factorial designs. Full factorial designs separately estimate all main effects and interactions of selected factors through the model equation, while fractional factorial designs compute aggregated effects of multiple factors or their interactions. Although a full factorial design requires higher experimental and computational costs, it provides a precise estimation of individual factor effects and interactions. In contrast, a fractional factorial design reduces experimental costs by estimating aggregated effects, but these are subject to confounding due to the cumulative nature of multiple factor contributions.

- *Full Factorial Designs:* Full factorial designs include $2^k r$, $l^k r$, and general factorial designs. In a $2^k r$ factorial design, each factor selects 2 levels; a $l^k r$ factorial design selects l levels per factor; and a general factorial design allows variable levels across factors. The model equation for full factorial designs accounts for $(2^k - 1)$ total effects, including $\binom{k}{1}$ main effects, and $\sum_{i=2}^{k} \binom{k}{i}$ interaction terms. With r replicates per run, an additional random error term is incorporated. The details are shown in Section 16.5.2. $\binom{k}{i}$ is read as "k choose i."
- *Fractional Factorial Designs:* A fractional factorial design, such as $2^{k-p} r$ factorial design, estimates $(2^{k-p} - 1)$ effects, each representing aggregated effects from 2^p factors or interactions. Similar to a full factorial design, it includes r replicates and a random error term in the model equation.

16.5.2 Model Equations and Analysis of Variance Formulations

We illustrate model equations and analysis of variance (ANOVA) formulations using a general factorial design as a case study. Notably, all other factorial designs are special cases of this approach, sharing the same structural formulation.

Consider a general factorial design investigating k factors influencing the responses. The ith factor has n_i levels, and each experimental run is replicated r times. The model equation decomposes each measurement into a linear additive model comprising: 1) *Overall mean:* The overall average response across all experiment groups; 2) *Main effects:* $\binom{k}{1} = k$ terms representing the individual effect of each factor; 3) *Interaction effects:* $\sum_{i=2}^{k} \binom{k}{i} = 2^k - 1 - k$ terms, including $\binom{k}{2}$ two-factor interactions, $\binom{k}{3}$ three-factor interactions, up to the k-factor interaction; 4) *Random errors:* Accounting for chance fluctuations introduced by repeated measurements.

ANOVA proceeds by squaring and summing all effects, which partition the total variance into distinct components: main effects, interactions, and residual error.

For concreteness, we demonstrate the model equations and the ANOVA formulation with $k = 3$ factors, highlighting how effects are structured and analyzed. Other factorial designs follow an analogous formulation.

In a general factorial design with $k = 3$ factors, Factor A has n_1 levels, Factor B has n_2 levels, and Factor C has n_3 levels. Each experimental combination is replicated r times. The model equation is expressed as:

$$Y_{i,j,k,l} = \mu + a_i + b_j + c_k + d_{i,j} + e_{i,k} + f_{j,k} + g_{i,j,k} + \epsilon_{i,j,k,l}. \tag{16.3}$$

In this model equation, $Y_{i,j,k,l}$ represents the measured response of replicate l at level i of Factor A, level j of Factor B, and level k of Factor C. μ represents the overall mean of all responses. a_i denotes the main effect of Factor A, b_j indicates the main effect of Factor B, and c_k signifies the main effect of Factor C. The two-way interaction effects are captured by $d_{i,j}$ (Factor A × Factor B), $e_{i,k}$ (Factor A × Factor C), and $f_{j,k}$ (Factor B × Factor C), while $g_{i,j,k}$ represents the three-way interaction effect among all three factors. Finally, $\epsilon_{i,j,k,l}$ accounts for the random errors introduced by experimental replication of chance fluctuations.

To calculate these components in the model equation, we first calculate the group means.

$$\overline{Y}_{.,.,.,.} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_2 \times n_3 \times r}, \tag{16.4}$$

$$\overline{Y}_{i,.,.,.} = \frac{\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_2 \times n_3 \times r}, \tag{16.5}$$

$$\overline{Y}_{.,j,.,.} = \frac{\sum_{i=1}^{n_1}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_3 \times r}, \tag{16.6}$$

$$\overline{Y}_{.,.,k,.} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_2 \times r}, \tag{16.7}$$

$$\overline{Y}_{i,j,.,.} = \frac{\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_3 \times r}, \tag{16.8}$$

$$\overline{Y}_{i,.,k,.} = \frac{\sum_{j=1}^{n_2}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_2 \times r}, \tag{16.9}$$

$$\overline{Y}_{.,j,k,.} = \frac{\sum_{i=1}^{n_1} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times r}, \tag{16.10}$$

$$\overline{Y}_{i,j,k,.} = \frac{\sum_{l=1}^{r} Y_{i,j,k,l}}{r}. \tag{16.11}$$

Using these group means, the components of the model equation in Equation 16.3 are calculated as:

$$\mu = \overline{Y}_{.,.,.,.}, \tag{16.12}$$

$$a_i = \overline{Y}_{i,.,.,.} - \overline{Y}_{.,.,.,.}, \tag{16.13}$$

$$b_j = \overline{Y}_{.,j,.,.} - \overline{Y}_{.,.,.,.}, \tag{16.14}$$

$$c_k = \overline{Y}_{.,.,k,.} - \overline{Y}_{.,.,.,.}, \tag{16.15}$$

$$d_{i,j} = \overline{Y}_{i,j,.,.} - \overline{Y}_{i,.,.,.} - \overline{Y}_{.,j,.,.} + \overline{Y}_{.,.,.,.}, \tag{16.16}$$

$$e_{i,k} = \overline{Y}_{i,.,k,.} - \overline{Y}_{i,.,.,.} - \overline{Y}_{.,.,k,.} + \overline{Y}_{.,.,.,.}, \tag{16.17}$$

$$f_{j,k} = \overline{Y}_{.,j,k,.} - \overline{Y}_{.,j,.,.} - \overline{Y}_{.,.,k,.} + \overline{Y}_{.,.,.,.}, \tag{16.18}$$

$$g_{i,j,k} = \overline{Y}_{i,j,k,.} - \overline{Y}_{i,j,.,.} - \overline{Y}_{i,.,k,.} - \overline{Y}_{.,j,k,.} + \overline{Y}_{i,.,.,.} + \overline{Y}_{.,j,.,.} + \overline{Y}_{.,.,k,.} - \overline{Y}_{.,.,.,.}, \tag{16.19}$$

$$\epsilon_{i,j,k,l} = Y_{i,j,k,l} - \overline{Y}_{i,j,k,.}. \tag{16.20}$$

By relocating the overall mean term μ from the right side of Equation 16.3 to the left, we derive the following equation:

$$Y_{i,j,k,l} - \mu = a_i + b_j + c_k + d_{i,j} + e_{i,k} + f_{j,k} + g_{i,j,k} + \epsilon_{i,j,k,l}. \tag{16.21}$$

Squaring and summing across all responses for both sides of the equation, the left side represents the total variance of the responses, while the right side decomposes into the sum of squares (SS) for each effect. Notably, cross-product terms among distinct effects sum to zero due to orthogonal design properties. Thus, the total variance is additively partitioned into contributions from main effects, interactions, and random errors. Denoting SS as the sum of squares, we have the following equations:

$$SSY = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{r} Y_{i,j,k,l}^2, \tag{16.22}$$

$$SS0 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}\mu^2, \tag{16.23}$$

$$SST = SSY - SS0, \tag{16.24}$$

$$SSA = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}a_i^2, \tag{16.25}$$

$$SSB = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}b_j^2, \tag{16.26}$$

$$SSC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}c_k^2, \tag{16.27}$$

$$SSAB = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}d_{i,j}^2, \tag{16.28}$$

$$SSAC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}e_{i,k}^2, \tag{16.29}$$

$$SSBC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}f_{j,k}^2, \tag{16.30}$$

$$SSABC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}g_{i,j,k}^2, \tag{16.31}$$

$$SSE = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}\epsilon_{i,j,k,l}^2. \tag{16.32}$$

So, we have:

$$SSY - SS0 = SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC + SSE. \tag{16.33}$$

To quantify the contribution of each effect to the response variance, we calculate the ratio of the specific effect's sum of squares (SS) to the total variance (SST). By dividing the sum of squares (SS) for each effect by its corresponding degrees of freedom (df), we obtain the mean squares (MS). The degrees of freedom (df) for each sum of squares are shown in Table 16.1.

We calculate the mean squares (MS) of effects as follows: So, we have:

$$MSA = \frac{SSA}{n_1 - 1}, \tag{16.34}$$

SS	*df*
SSY	$n_1 \times n_2 \times n_3 \times r$
SS0	1
SST	$n_1 \times n_2 \times n_3 \times r - 1$
SSA	$n_1 - 1$
SSB	$n_2 - 1$
SSC	$n_3 - 1$
SSAB	$(n_1 - 1) \times (n_2 - 1)$
SSAC	$(n_1 - 1) \times (n_3 - 1)$
SSBC	$(n_2 - 1) \times (n_3 - 1)$
SSABC	$(n_1 - 1) \times (n_2 - 1) \times (n_3 - 1)$
SSE	$n_1 \times n_2 \times n_3 \times (r - 1)$

Table 16.1: The degrees of freedom (df) for each sum of squares (SS).

$$MSB = \frac{SSB}{n_2 - 1}, \tag{16.35}$$

$$MSC = \frac{SSC}{n_3 - 1}, \tag{16.36}$$

$$MSAB = \frac{SSAB}{(n_1 - 1) \times (n_2 - 1)}, \tag{16.37}$$

$$MSAC = \frac{SSAC}{(n_1 - 1) \times (n_3 - 1)}, \tag{16.38}$$

$$MSBC = \frac{SSBC}{(n_2 - 1) \times (n_3 - 1)}, \tag{16.39}$$

$$MSABC = \frac{SSABC}{(n_1 - 1) \times (n_2 - 1) \times (n_3 - 1)}, \tag{16.40}$$

$$MSE = \frac{SSE}{n_1 \times n_2 \times n_3 \times (r - 1)}. \tag{16.41}$$

The F-statistic is calculated as the ratio of mean squares (MS) for any two effects among Equation 16.34 and Equation 16.41. By comparing this ratio to the critical F-value from the F-distribution with corresponding degrees of freedom (df), an F-test can assess the statistical significance of the two effects. For example, we calculate $F_c = \frac{MSA}{MSE}$. At a 0.05 significance level, we refer to the F-distribution with $df_1 = n_1 - 1$ numerator and $df_2 = n_1 \times n_2 \times n_3 \times (r - 1)$ denominator degrees of freedom to obtain the critical value F_v corresponding to 95% cumulative probability. If $F_c \geq F_v$, we conclude that the main effect of Factor A is statistically significant relative to random errors.

16.6 Examples: Infer the Effects of Apple Origins on Purchasing Prices

We then conduct a case study on the apple purchasing price, performing ANOVA and F-tests within the DoE framework.

Design of Experiment

Case study

The price of purchasing a box (5kg) of apples is related to the combination of its origin and size. There are two regions of origin: Qixian County in Henan Province (Origin 1) and Yantai City in Shandong Province (Origin 2), and three types of size: small (Size 1), medium (Size 2), and large (Size 3). The corresponding apple prices are shown in the table below.
We will analyze the effects of the origins and sizes of the apples on the purchasing price, and determine whether the origin or the size has a more statistically significant influence on the price. These factors and their corresponding levels in the example will be reused throughout all chapters of Part IV.

	Size 1	Size 2	Size 3
Origin 1	¥25	¥45	¥50
Origin 2	¥55	¥65	¥90

Table 16.2: Price of purchasing one box (5kg) of apples.

Calculation

The model equation of this example is:

$$Y_{i,j} = \mu + a_i + b_j + c_{i,j} + \epsilon_{i,j,k}.$$

Where μ is the overall mean effect, a_i is the effect of origin, b_j is the effect of size, $c_{i,j}$ is the effect of interaction between origin and size, $\epsilon_{i,j,k}$ is the effect of random errors. In this case, the price has no chance of fluctuations, so $\epsilon_{i,j,k} = 0$.
We calculate each mean as follows:

$$\overline{Y}_{.,.} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{3} Y_{i,j}}{2 \times 3},$$
$$\overline{Y}_{i,.} = \frac{\sum_{j=1}^{3} Y_{i,j}}{3},$$
$$\overline{Y}_{.,j} = \frac{\sum_{i=1}^{2} Y_{i,j}}{2}.$$

Then each effect in the model equation is:

$$\begin{aligned}\mu &= \overline{Y}_{.,.},\\ a_i &= \overline{Y}_{i,.} - \overline{Y}_{.,.},\\ b_j &= \overline{Y}_{.,j} - \overline{Y}_{.,.},\\ c_{i,j} &= Y_{i,j} - \overline{Y}_{i,.} - \overline{Y}_{.,j} + \overline{Y}_{.,.}.\end{aligned}$$

We can then calculate the sum of squares of each effect as follows:

$$\begin{aligned}SSY &= \sum_{i=1}^{2}\sum_{j=1}^{3} Y_{i,j}^2 = 20,500,\\ SS0 &= \sum_{i=1}^{2}\sum_{j=1}^{3} \mu^2 = 18,150,\\ SST &= SSY - SS0 = 2,350,\\ SSA &= \sum_{i=1}^{2}\sum_{j=1}^{3} a_i^2 = 1,350,\\ SSB &= \sum_{i=1}^{2}\sum_{j=1}^{3} b_j^2 = 900,\\ SSAB &= \sum_{i=1}^{2}\sum_{j=1}^{3} c_{i,j}^2 = 100.\end{aligned}$$

Factor	Contribution
Origin	57.45%
Size	38.30%
Origin-Size	4.25%

Table 16.3: ANOVA of the price of a box (5kg) of apples.

Conclusion

For this example, the degrees of freedom for each factor are shown in the following Table.

Factor	df
Origin	1
Size	2
Origin-Size	2

Table 16.4: Degrees of freedom of each factor .

Then we can calculate the mean of the square of each effect as follows:

$$MSA = \frac{SSA}{df_a} = \frac{1,350}{1} = 1,350,$$
$$MSB = \frac{SSB}{df_b} = \frac{900}{2} = 450,$$
$$MSAB = \frac{SSAB}{df_{ab}} = \frac{100}{2} = 50.$$

We compare the origin to the size for the price of purchasing a box (5kg) of apples with a significance of 95%. The computed F-value f-compute is:

$$f\text{-compute} = \frac{MSA}{MSB} = \frac{1,350}{450} = 3.$$

The critical value in the F-distribution with 1 numerator and 2 denominator degrees of freedom at 95% significance is 18.51.

$$f\text{-table} = 18.51,$$
$$f\text{-compute} < f\text{-table}.$$

Therefore, at the 95% significance, we cannot conclude that origin is more statistically significant than size for the price of a box (5kg) of apples.

16.7 Limitations

DoE is a valuable tool for analyzing the effect relationships between responses and factor effects, enabling the calculation of the variance contribution of different factors to the response as well as their statistical significance. However, DoE also has limitations that researchers should be mindful of when applying it.

Firstly, DoE requires exhaustive experimentation across all experimental condition groups composed of specific level values for each investigated factor. Without responses from any single experimental condition group, ANOVA and F-tests cannot be conducted.

After selecting factors and their level values, a grid can be plotted where factors represent different dimensions, and experimental condition groups correspond to points on the grid. During experimentation, responses measured for each experimental condition group are then filled into their respective positions on the grid. Any missing response data at any point on the grid precludes ANOVA and F-tests.

Secondly, as the number of factors selected in DoE increases, the costs associated with conducting experiments and analyzing data grow exponentially. Moreover, ANOVA and F-tests, which require calculating the sum of squares (SSE) of responses across different experimental condition groups, incur higher computational costs compared to methods that simply calculate means. Consequently, the experimental costs of DoE remain a critical factor to be balanced.

Finally, conducting ANOVA and F-tests in DoE requires testing that the corresponding statistical assumptions are met, which necessitates validating whether these assumptions hold for the response variables.

16.8 Summary

The DoE employs a model equation to characterize the effects between factors and responses. Its model equation quantifies: 1) the overall mean of all responses, 2) main effects for individual factors, 3) factors' interaction effects 4) random errors due to chance fluctuations. By grouping experimental responses according to factor levels and calculating corresponding group means, component values for each model equation term can be estimated. The SS for all modeled effects (including errors) partitions total response variance, enabling ANOVA and an F-test to be conducted for statistical analysis.

Chapter 17

Randomized Controlled Trials

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of Randomized Controlled Trials (in short, RCTs).

17.1 Basic Concepts

The definitions of the basic concepts are cited from the early work of RCTs [191, 141, 59, 60].

An Intervention or Treatment is the precisely defined procedure whose effects or safety are under investigation. This can be a pharmacological agent (drug), a medical device, a surgical technique, a behavioral therapy, an educational program, or any other maneuver administered to participants to elicit a measurable response.

A Participant or Unit is an individual who meets the predefined eligibility criteria for a study and is formally enrolled to undergo the investigative procedures, data collection, and follow-up as outlined in the trial protocol.

Outcomes are quantitative metrics obtained from the participants through measurements, distinct from categorical variables.

A Treatment Group is the cohort of participants who are randomly allocated to receive the treatment, which is the primary focus of the investigation.

A Control Group is the cohort of participants who are randomly allocated to receive a comparator against which the treatment is evaluated. This comparator can be a placebo (an inert substance), a standard-of-care treatment, a different treatment, or no treatment, depending on the research question and ethical considerations.

Randomization means every participant has an equal chance of being assigned to either a Treatment Group or Control Group.

An Investigator is a qualified individual responsible for the conduct of the RCTs at a trial site. The Principal Investigator (PI) bears the overall responsibility for the ethical and protocol-directed execution of the study, including management of the investigative team, and protection of participant rights and data integrity.

[1] Mr. Chenxi Wang is the primary contributor.

The *Controlled Comparison* is a fundamental methodological principle in RCTs whereby the outcomes of the Treatment Group are systematically evaluated against those of the Control Group. The primary objective is to isolate the causal effect of the treatment by holding constant, through the research design, the influence of all other extraneous variables.

A Blinding is a methodological procedure in RCTs whereby one or more parties involved in the research are kept unaware of the treatment assignment (i.e., whether a participant is in the Treatment Group or the Control Group). The primary purpose is to prevent the introduction of conscious or unconscious bias that could influence the perceived outcomes, behaviors, or interpretations of the results.

The *Average Treatment Effect, in short ATE,* is the expected effect difference in outcomes between the Treatment Group and the Control Group. For any participant i, there are two potential outcomes: the observed outcome $Y_i(1)$ if assigned to the Treatment Group and the observed outcome $Y_i(0)$ if assigned to the Control Group. However, both potential outcomes for the same participant i cannot be observed simultaneously. RCTs utilize randomization to calculate an unbiased estimator of the average treatment effect.

The *External Validity* refers to the extent to which the results of RCTs can be generalized and applied beyond the research experimental setting (to real-world setting), and this extent can be evaluated through the correlation between RCT results and those obtained in real-world setting.

The *Internal Validity* refers to the degree to which RCTs can reliably demonstrate the results, meaning that the calculated ATE is indeed attributable to the difference between the treatment and control procedure, rather than to other confounding factors. By employing randomization, controlled comparison between Treatment Group and Control Group, and blinding, RCTs ensure the internal validity.

17.2 Problem Statement

RCTs are a systematic methodology for inferring the relationship between interventions and outcomes [191, 141].

Given a population of participants and a proposed treatment as inputs, each participant i possesses two potential outcomes, $Y_i(1)$ when receiving the treatment and $Y_i(0)$ when receiving the control. The challenges lie in that only one of these outcomes can ever be observed for the same participant i. Please note that in several cases, we could observe both outcomes, as we have discussed in Section 14.

The RCTs problem could be formulated as "how to use a random mechanism (randomization) to assign participants to either the Treatment Group or Control Group, providing an unbiased estimate of the Average Treatment Effect (ATE)." The inputs for RCTs are participants, the Treatment Group, the Control Group, and the observed outcome in each participant. Under the constraint of randomly assigning participants into the Treatment Group and Control Group, RCTs output the ATE.

17.3 Basic Assumptions

Assumption 17.1 (Stable Unit Treatment Value Assumption (SUTVA) [43, 158])**.** *The potential outcomes for any unit or participant, e.g., an individual patient, are unaffected by the particular treatment assignment of other units or participants. This assumption comprises two key components:*

1. *No Interference: The treatment assignment of one unit does not influence the outcome of any other unit. This ensures the independence of units, a prerequisite for standard statistical inference.*
2. *No Hidden Variations of Treatment: Each treatment regime is identical for every unit assigned to it. There are no multiple, unspecified versions of the treatment or control that could differentially affect the outcomes.*

SUTVA is the most fundamental assumption underpinning RCTs, as it allows us to define a unique potential outcome for each unit under each treatment state.

Assumption 17.2 (Ignorability)**.** *This assumption is also known as unconfoundedness or exchangeability. It states that the assignment of treatment is conditionally independent of the potential outcomes, given the observed covariates. Formally,*

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X, \tag{17.1}$$

where $Y(1)$ and $Y(0)$ are the potential outcomes under the Treatment Group and the Control Group, respectively, T is the treatment assignment indicator, and X represents a vector of pre-treatment covariates. The symbol "$\mid$" is a mathematical notation denoting "under the condition of," and "$A \mid B$" signifies "A under the condition of B." The symbol "$\perp\!\!\!\perp$" represents "mutual independence," with "$A \perp\!\!\!\perp B$" indicating that "A and B are mutually independent."

Randomization is the physical design feature that ensures this assumption is met. It renders the Treatment Group and the Control Group statistically equivalent, or exchangeable, in expectation, not only with respect to measured covariates X but also with respect to all unmeasured factors. Consequently, any systematic difference in observed outcomes can be causally attributed to the treatment.

Assumption 17.3 (Consistency)**.** *The observed outcome for a participant who received a specific treatment level is precisely the potential outcome for that participant under that same treatment level. Formally, if $T_i = t$, then $Y_i^{obs} = Y_i(t)$.*

This assumption implies that the treatment is well-defined and that there is no ambiguity in its application or measurement across all participants. It links the conceptual potential outcomes to the actually observed data.

17.4 Basic Principles

The primary motivation for employing RCTs lies in the need to eliminate confounding bias by ensuring that the Treatment Group and Control Group are statistically equivalent at baseline. This is achieved through the random assignment of participants to experimental conditions, thereby ensuring that both observed and unobserved covariates are, on average, balanced across groups.

RCTs typically involve the comparison of one or more treatment interventions against a control condition, often incorporating blinding and placebo controls to minimize expectancy and measurement biases. The methodology is characterized by a clearly defined protocol, prespecified outcomes, and rigorous adherence to statistical principles such as intention-to-treat analysis. Widely applied in clinical medicine, social sciences, and public policy evaluation, RCTs enable researchers to make high-confidence claims about the effect of interventions under controlled and replicable conditions[191].

Randomization is the heart of the RCTs. It ensures that every participant has an equal chance of being assigned to either the Treatment Group or the Control Group. This is crucial because it ensures the groups are, on average, similar in every way at the start of the study, not just in age or gender, but also in countless other factors we can't even measure (like genetics, diet, or natural resilience). RCTs control for other factors (known and unknown) beyond the Treatment/Control Group through randomization, reducing effects to a matter of chance fluctuations.

Because the groups are comparable, any significant difference in the outcome at the end of the study can be more confidently attributed to the treatment itself, rather than to pre-existing differences between the groups.

17.5 Methodology

17.5.1 The Average Treatment Effect Formulation

We can formally define the controlled comparison made in RCTs. The goal is to estimate the Average Treatment Effect (ATE), which is the average causal effect of the treatment across the entire study population.

We begin by defining the key steps in RCTs:

- Let $i = 1, 2, ..., N$ represent each of the N total participants in the trials.
- The treatment assignment for Participant i is denoted by T_i:

$$T_i = \begin{cases} 1, & \text{if Participant } i \text{ is assigned to the Treatment Group,} \\ 0, & \text{if Participant } i \text{ is assigned to the Control Group.} \end{cases} \tag{17.2}$$

- If $T_i = t$, the observed outcome for Participant i (e.g., the decrease in apple purchasing price) is denoted by $Y_i(t)$.

Based on the above discussion, the observed outcome for Participant i depends on the treatment assignment:

$$Y_i^{obs} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0). \tag{17.3}$$

The fundamental advantage of randomization is that it allows for a simple and powerful comparison. The Average Treatment Effect (ATE) is estimated by calculating the difference in the average outcome between the two groups:

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)], \tag{17.4}$$

where:

- $E[Y_i(1)]$ is the expected value (arithmetic average) of the outcomes for all participants in the Treatment Group.
- $E[Y_i(0)]$ is the expected value of the outcomes for all participants in the Control Group.

In practice, we calculate the sample averages to estimate the ATE:

$$\widehat{ATE} = \frac{1}{N_1} \sum_{i:T_i=1} Y_i(1) - \frac{1}{N_0} \sum_{i:T_i=0} Y_i(0). \tag{17.5}$$

Where:

- N_1 is the number of participants in the Treatment Group.
- N_0 is the number of participants in the Control Group.
- $N = N_0 + N_1$

17.6 Example: Infer the Effects of Apple Origins on Purchasing Prices

We then conduct a case study of performing RCTs to investigate the apple origin trial on the purchasing price.

Randomized Controlled Trails

Case study

- Presume that 10,000 ($N = 10,000$) apple seeds in your study.
- You use a computer to randomly assign 5,000 ($N_1 = 5,000$) apple seeds for planting in Qixian County, Henan Province (*Treatment Group*).
- The other 5,000 ($N_0 = 5,000$) seeds were planted in Yantai City, Shandong Province (*Control Group*).

After 10 years, you measure the average purchasing price per kilogram for the apples grown in both groups.
Suppose the obtained average purchasing price per kilogram is shown in the following Table.

Group	T_i	$E[Y_i(T_i)]$
Treatment Group	1	¥8/kg
Control Group	0	¥14/kg

Table 17.1: The average purchasing price per-kilogram in two trials.

Calculation

Applying the Formula to the Example:

- The average purchasing price per-kilogram in the Treatment Group ($E[Y_i(1)$) was ¥8/kg.
- The average purchasing price per-kilogram in the Control Group ($E[Y_i(0)]$) was ¥14/kg.
- Using Equation 17.5: $\widehat{ATE} = 8 - 14 = -6$.

Conclusion

This suggests apples grown in Qixian County, Henan Province had an average purchasing price of ¥6/kg *lower* than that of apples grown in Yantai City, Shandong Province. Because of randomization, we can be confident that this difference is likely due to the origin itself and not other confounding factors.

17.7 Limitations

RCTs are widely regarded as the gold standard for establishing causal inference, yet they are subject to several important limitations that researchers must acknowledge.

Firstly, the high costs and complex logistics associated with properly conducting an

RCT often pose significant barriers. Implementing randomization, ensuring adherence to protocols, and maintaining blinding over a sufficient follow-up period require substantial financial resources, time, and operational effort, which can be prohibitive for many research questions.

Secondly, ethical and practical constraints frequently limit the applicability of RCTs. It is ethically impermissible to randomize participants to interventions known or strongly believed to be harmful. Furthermore, for research on long-term outcomes or rare diseases, it may be practically infeasible to recruit and retain an adequate sample size over the necessary timeframe.

Thirdly, the issue of generalizability (external validity) is a critical concern. The highly controlled conditions and specific participant population of an RCT may not be representative of real-world settings or broader patient groups. Consequently, findings from an RCT, while internally valid, may not translate directly to routine clinical practice.

Finally, RCTs are vulnerable to specific methodological biases if not meticulously designed and executed. These include attrition bias, if dropout rates (the proportion of participants who leave the study) differ systematically between groups, and a failure of blinding, which can introduce performance and detection bias. The analysis must also adhere to the intention-to-treat principle (analyzing all participants in their originally assigned groups regardless of what treatment they actually received) to preserve the integrity of the randomization, which can complicate the interpretation of the actual treatment effect received.

17.8 Summary

The RCTs are considered the gold standard for a simple reason: the design provides the clearest possible answer to the question “What would have happened otherwise?” The Control Group shows what happens in the absence of the intervention, and randomization ensures this comparison is fair. While other study designs can find associations, a well-conducted RCTs provides the strongest evidence for a causal relationship. This makes it an indispensable tool for advancing reliable knowledge in science and medicine. However, it is important to note that RCTs require a sufficient sample size due to randomization to reliably estimate causal effects.

Chapter 18

Quasi-experiments

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the quasi-experiments.

18.1 Basic Concepts

We will reuse most concepts defined in Chapter 17, and only introduce those concepts in quasi-experiments different from those of RCT.

A Treatment/ Controlled Group denotes the intervention or manipulation that is applied to the participants or groups. The explicit manipulation of this variable is the characteristic that links quasi-experimental research to real-world experiments. There are slight differences between the concepts of Treatment/Controlled in quasi-experiments from those of RCTs. In quasi-experiments, *a control Group is frequently nonequivalent because participants were not assigned through randomization.* According to Campbell et al. [29], "the concept of the application of an experimental mode of analysis and interpretation to bodies of data not meeting the full requirements of experimental control because experimental units are not assigned at random to at least two 'treatment' conditions." [2]

Cause refers to the differences between the Treatment Group and the Control Group in quasi-experiments on pre-existing groups, while *effect* refers to the differences in outcomes between the Treatment Group and the Control Group in the context of quasi-experiments.

Internal Validity refers to the influence in a certain specific experimental instance.

External Validity refers to how the influence can be generalized to populations, settings, treatments, and measurements [29].

Natural Group refers to research units that are pre-existing organizational structures in the real world [119]. Researchers are constrained to assign the treatment to the

[1] Mr Hongxiao Li is the primary contributor.

[2] According to the original author's intention, the 'two treatment conditions' should be the treatment group and the control group.

entire unit because they cannot subdivide these natural groups into randomly equivalent subgroups for either treatment or control. The use of natural grouping is a physical constraint that necessitates a quasi-experimental approach due to selection biases from the real world.

Confounding Variable is not strictly defined in the existing quasi-experiment studies. According to the description of Thyer et al. [203], confounding variables are referred to as other variables than the control and treatment, which are external and might make changes to the outcome if left uncontrolled on the dependent variable, thereby undermining the ability to claim a causal relationship between the treatment and the outcome. As a limitation, a quasi-experiment cannot completely infer the causal effect.

18.2 Problem Statement

Quasi-experimental designs [150] are a research design used to estimate causal relationships when random assignment of participants to experimental groups is not feasible or ethical. Unlike RCTs, it lacks full control over variables but still compares groups exposed to different conditions.

A quasi-experiment problem can be stated as "when a random mechanism (randomization) is not available, how to estimate the Average Treatment Effect (ATE)." Especially, quasi experiments focus on the outcome in a time period.

Same as RCTs, the inputs for quasi-experiments are participants, the treatment group, the control group, and the observed outcome in each participant. The outputs are the approximated effects of the Treatment Group rather than its causal effects.

18.3 Basic Assumptions

The basic assumptions of quasi-experiments are mostly the same as those of RCT, with the two major exceptions as follows:

Assumption 18.1 (Non-independence [29])**.** *Most quasi-experiments are conducted in real-world scenarios, where the correlation between individuals is stronger. This makes the assumption of no interference harder to guarantee.*

Meanwhile, the implementation of treatments may vary across scenarios (e.g., the same policy implemented in different hospitals), leading to a higher risk of violating the requirement of no hidden treatment variation.

Assumption 18.2 (Approximate unconfoundedness [29])**.** *There is no random assignment process. The formation of Treatment and Control Groups is usually related to individual characteristics, leading to the existence of confounding variables. So the unconfoundedness assumption is not fully satisfied.*

18.4 Basic Principles

Quasi-experiments provide a structured framework for approximating causal effects where randomization is infeasible or unethical. The principles of quasi-experiments explicitly distinguish between causal and associational relationships by carefully identifying sources of variation in a way that mimics randomization. These frameworks incorporate prior domain knowledge to design and justify the selection of comparison groups, ensuring that they approximate counterfactual scenarios. Quasi-experiments estimate the effect of intervention by leveraging natural or real-world variations to estimate the effects of interventions.

18.5 Methodology

Quasi-experimental designs [150] evolve from the simple before-and-after design through the following five key strategies, each targeting specific threats to internal validity.

1. *Add a non-randomized control group* [150]. This method is like RCTs. However, the partition is sometimes impossible. Therefore, a quasi-experiment setting is used.

In this and the following parts, we use Notation O to represent a measurement, and Notation X to represent an intervention. If there are two lines, they represent two groups. Corresponding to Notation X in a group, a blank in another group represents a non-intervention. This method can be represented as follows:

O X O
O O

Example [150]: To lower reindeer herders' high injury rates, three geographically divided groups were set: Group A got preventive measure letters, Group B got info from trained health staff, and the control group got nothing.

For a pre-period (1985) and a post-period (1987), the accident rates dropped for all: A:18.7→15.1; B:21→14.9; control:19.2→14.6. According to the standard deviation threshold requirement, no significant group difference emerged, so the interventions were deemed ineffective.

2. *Take more measurements* [150]. Instead of a single before-and-after measurement, multiple baseline measurements establish a pre-intervention, and multiple post-intervention measurements establish a post-intervention. Evaluation on several more measurements is more reliable.

This method can be represented as follows:

O O O X O O O

shows an example for a single before-and-after measurement and

O O O X O O O
O O O O O O

shows an example for multiple baseline measurements.

Example [150]: A food factory did safety training for two departments. It first took 3-4 weeks for safety behavior baseline checks (via checklist) for both departments. Then, after the post-intervention, including education, safety lists, and feedback, checks contin-

ued. The results showed safety behaviors changed after intervention in each department, strengthening the links between interventions and behaviors.

3. *Stagger the introduction of the intervention* [150]. The term "stagger" means all partitioned groups eventually receive the intervention, but at different times, with each group serving as a comparison for others. This minimizes chronological effects: coincidental events aligning with one intervention are plausible, but multiple such coincidences are unlikely.

For instance, the interventions of salary improvement for different groups are applied at different times. Therefore, the effects' discrepancy of salary improvement is less interfered with by occasional incidents, e.g., natural disasters or economic fluctuations. This method can be represented as follows:

O O O X O O O O O O
O O O O O O X O O O

Example [150]: The same food factory launched safety training first in the wrapping department, then later in the make-up department. Each had baseline checks before training. A staggered launch lets departments act as comparisons, reducing the chance of extraneous events causing results.

4. *Reverse the intervention* [150]. Compare the effects of the intervention and that after reversing the intervention. If outcomes revert to baseline levels after reversal, the intervention's causality is confirmed. This method eliminates chronological and occasional threats and errors involved in the testing procedure, but it is only suitable when the intervention is reversible. With the symbol $-X$ standing for the reversed intervention, this method can be represented as follows:

O O O X O O O -X O O O

Example [150]: A company plans a participatory ergonomics program with task modifications, new equipment, and worker education. To measure its impact, the coordinator will compare symptoms and injuries before and after, while tracking equipment purchases and self-reports on tasks and stressors to rule out other factors.

5. *Measure multiple outcomes* [150]. There are two approaches. The first one is to add intervening outcome measures. Track implementation and short or intermediate-term effects to distinguish between ineffective interventions and flawed implementation. For example, a "train-the-trainer" program's lack of injury reduction was attributed to poor implementation rather than ineffective techniques. Second, add related but untargeted outcomes. Measure outcomes similar to the main target but unaffected by the intervention. This method can be represented as follows, where O_1/O_2 stands for two measured quantities:

O_1/O_2 X O_1/O_2

Example 1 [150]: Add intervening measures: A company's ergonomics program tracked ultimate outcomes (symptoms/injuries) and added checks like equipment purchase records and work task reports. This helped tell if failure was from a bad program or poor implementation.

Example 2 [150]: Add untargeted measures: Oil platforms tracked tong-related injuries (target of new equipment) and non-tong injuries (untargeted). No change in

non-tong injuries meant that the tong injury changes were likely from the equipment. A grocery ergonomic intervention found targeted (neck/back) discomfort dropped, but untargeted (arm/wrist) didn't, reducing validity threats.

Besides the above, analytical methods such as matching, propensity score, and difference-in-differences are required to control for observable confounders as much as possible, but unobservable confounders cannot be addressed.

18.6 Example: Infer the Effects of Apple Origins on Purchasing Prices

Quasi experiments

Case study

This example uses the first strategy. To investigate how the apple origin influences market prices, experiments ought to be conducted across two origins (designated as the control group and treatment group). Nevertheless, given the insufficient sample size of the selected origins (due to policy or practical constraints), supplementary sample data were collected both from the origins themselves and their adjacent areas. The design leverages a natural characteristic: for the origins with a small sample size, geographically adjacent areas —often sharing identical climate, soil, and farming practices—can serve as their reference. The quasi-experiment design approximately ensures that the only systematic difference between the two groups is "origin label (X)", while confounding factors (A, B, C) remain balanced, where A stands for sweetness (12, 14, 16 Brix), B stands for apple size (small, medium, large), and C stands for shelf-life (3 days, a week, two weeks).

1. *Treatment Group*
   - Scope: Apple growers located within 5 km of the Origin 1.
   - Label: Apples are marketed with the target origin certification ($X = 1$).

2. *Control Group*
   - Scope: Apple growers located within 5 km of the Origin 2.
   - Label: Apples are marketed with the target origin certification ($X = 2$).

Table 18.6 is the price per kilogram of apples from two origins.

Group	T_i	$E[Y_i(T_i)]$
Treatment Group	1	¥8/kg
Control Group	0	¥14/kg

Table 18.1: The average apple per-kilogram purchasing price in two trials.

Calculation

Before analyzing the price effect, it is necessary to verify that A, B, C are balanced between the two groups, to rule out the impact of non-origin factors on price.

1. *Data Collection*
   - For each sampled apple orchard, collect 50 representative apples.
   - Measure indicators: Sweetness (A, via Brix refractometer), size (B, via single fruit weight in grams), shelf life (C, via days of normal storage without decay).

2. *Balance Test Methods*
   - Descriptive statistics: Calculate the mean, standard deviation, and median of A, B, C for both groups.
   - Statistical significance test: Use an independent samples t-test to verify that the mean difference of A, B, C between the two groups is not statistically significant (standard threshold: p-value > 0.05).
   - Distribution visualization: Draw kernel density plots for A, B, C; overlapping curves indicate balanced distributions.

After confirming the balance of confounding factors, the effect of origin (X) on purchase price (Y) is calculated through direct comparison:

1. *Price Data Collection*
   - Track the same batch of merchants to collect on-site purchase prices of apples from both groups.
   - Record Y as the unit price of apples meeting the same basic quality standards.

2. *Effect Calculation*
   1. Calculate the average purchase price of the treatment group ($\bar{Y}_1$) and the control group ($\bar{Y}_0$).
   2. Compute the effect: Effect $= \bar{Y}_1 - \bar{Y}_0 = 8 - 14 = -6$.
   Explanation: The value represents the additional price brought by the "target origin label", as confounding factors (A, B, C) have been balanced by geographic matching.

Conclusion

The result reports apples grown in Origin 1 had an average purchasing price of ¥6/kg *lower* than that of apples grown in Origin 2. Therefore, it is confident that this difference is likely due to the origin itself and not other confounding factors.

18.7 Limitations

The primary limitation of quasi-experiments is the lack of internal validity, let alone external validity. The lack of internal validity comes from the following multiple factors.

1. *Lack of randomization*: In quasi-experiments, objects are not randomly assigned to Experimental and Control Groups. Grouping is based on pre-existing conditions as a weak substitute. This lack of randomization leads to selection bias, where systematic differences between the groups may exist independently of the Treatment and Control.

2. *Effect of confoundings*: Without full control over the experimental environment, other confoundings may simultaneously interfere with both the independent and dependent variables, masking or exaggerating the true effect of the independent variable.

3. *History effects*: As quasi-experiments are often deployed in real-world scenarios, some external events that occur during the study can influence the dependent variable, regardless of the experimental treatment. These events can confound the results. For example, a policy intervention during a specific period might be interfered with by simultaneous economic changes.

18.8 Summary

Quasi-experiment is a research design that approximates the methodological rigor of a true experiment. Quasi-experiments enable researchers to approximate the effects of the Treatment group. This distinguishes them from purely observational studies, as quasi-experiments often involve deliberate manipulation of an independent variable or exploitation of natural interventions. Although a quasi-experiment is useful in social science and some other fields, its limitations of internal validity and generalization failure are inherent. The approximated effects are not a rigorous causal effect.

Chapter 19

Structural Causal Model

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the structural causal model (in short, SCM).

19.1 Basic Concepts

Endogenous Variables are the observed variables of interest.

Exogenous Variables originate outside the model and are not influenced by any other variables within the system..

A Directed Acyclic Graph (DAG) is a representation used in graph theory where nodes represent variables and directed edges represent causal directionality [202, 12]. The graph must be acyclic, meaning no path of arrows can lead back to a node already included in that path. The absence of an arrow between two variables constitutes an empirical assumption.

An Intervention is mathematically formalized using the $\mathbf{do}(X = x)$ operator, which refers to a physical action setting the variable X to a fixed value x in its value range. Pearl et al. first formalized this concept in the definition of Causal Bayesian Network [139]. Deleting the structural equation for X and replacing it with the fixed value equation $X = x$ in the original SCM, noted as M, results in a modified sub-model M_x.

Causal Effect is defined as the probability distribution $\mathbb{P}(Y = y|\mathbf{do}(X = x))$, which is the distribution of Y in sub-model M_x. This concept is implicitly defined by Pearl et al. [139] in the original text as: "A causal structure or relationship or mechanism serves as a blueprint for forming a 'causal model' – a precise specification of how each variable is influenced by its parents in the DAG."

A Direct Cause is implied by a direct causal relationship $X \to Y$ that exists if X is a parent of Y in the causal graph and appears as an argument in the structural equation for Y. This relationship quantifies the causal effect of X on Y, asserting that a change in X results in a corresponding change in Y regardless of the values taken by other variables in the model. The variable Y is called the effect of X in this context [139, 141].

[1] The primary contributor is Mr. Hongxiao Li.

A Potential Cause X of Y is defined graphically, where X is an ancestor node of Y, meaning a directed path exists between them. Operationally, X is considered a cause of Y if the intervention $\mathbf{do}(X = x)$ alters the probability distribution of Y, $\mathbb{P}(Y|\mathbf{do}(X = x))$, thus allowing for the inference of relationships that remain invariant under external change.

The following is the original formalization of Pearl et al. [139]:
"A variable X has a potential causal influence on another variable Y (that is inferable from P) if the following conditions hold:

1. X and Y are dependent in every context.
2. There exists a variable Z and a context S such that:
   (i) X and Z are independent given S (i.e., $X \perp\!\!\!\perp Z \mid S$). The notation $\perp\!\!\!\perp$ means independent.
   (ii) Z and Y are dependent given S (i.e., $Z \not\perp\!\!\!\perp Y \mid S$)." The notation $\not\perp\!\!\!\perp$ means dependent.

Counterfactual refers to the value Y "would have taken" under a hypothetical condition $X = x$, are defined using the intervention sub-model M_x [139] by Pearl et al. The counterfactual $Y_x(u)$ with background factors $U = u$ is defined as the value of Y computed in the sub-model M_x: $Y_x(u) \triangleq Y_{M_x}(u)$. This formulation means that the probability of a counterfactual $\mathbb{P}(Y_x = y)$ is the probability assigned to the set of variables U that satisfy $Y_{M_x}(u) = y$. The symbol $\triangleq$ is used for definition.

19.2 Problem Statement

SCM [89, 141] is a foundational statistical and conceptual framework that uses causal graphs and structural equations to represent and infer causal relationships between variables. This approach is widely used in fields such as epidemiology, economics, computer science, and social sciences for tasks like causal discovery, mediation analysis, and policy evaluation.

The SCM problem could be stated as "how to model the causal mechanisms underlying a system in what formal language to enable the estimation of causal effects that go beyond mere correlational patterns observed in data."

The inputs to this problem consist of a DAG encoding the causal structure, and a distribution over exogenous noise variables. The output is the estimation of causal effects among those inputs. The constraints require that the graph must be acyclic, each variable has a unique structural equation, and exogenous noise terms satisfy specified independence assumptions.

19.3 Basic Assumptions

The structural causal model method relies on the following three assumptions.

Assumption 19.1 (Causal DAG [139]). *It is assumed that the causal relationships between variables can be represented by a DAG, and the graph is acyclic (i.e., no variable can influence itself through a causal path).*

SCMs are represented using a directed acyclic graph (DAG), where:

- *Nodes represent variables.*
- *Directed edges represent causal relationships.*

Assumption 19.2 (Structural Equations [139]). *It is assumed that each variable X_i is determined as a function of its direct causes (parents in the DAG) and an independent noise term U_i:*

$$X_i = f_i(\mathbf{Pa}(X_i), U_i), \tag{19.1}$$

where:

- $\mathbf{Pa}(X_i)$ *represents the direct causes (parents) of X_i in the DAG.*
- f_i *is a deterministic function describing how X_i depends on its parents.*
- U_i *is an independent noise term capturing exogenous influences.*

Assumption 19.3 (Causal Markov Condition [78, 139]). *It is assumed that each variable X_i is conditionally independent of its non-descendants (any other variables) given its parents in a causal DAG:*

$$X_i \perp\!\!\!\perp \textit{Non-Descendants of } X_i \mid \mathbf{Pa}(X_i). \tag{19.2}$$

This assumption allows the global probability distribution to be factorized based on the local causal relationships.

19.4 Basic Principles

SCM provides a formal language for encoding causal assumptions, deriving causal inferences, and quantifying the effects of interventions. SCM moves beyond purely statistical associations by explicitly incorporating causal knowledge through a structured system of equations and a corresponding graphical representation.

The formal language explicitly distinguishes between causal and associational relationships, encodes prior domain knowledge into the structure of causal graphs where edges represent direct causal effects, defines and computes counterfactuals through interventions that simulate external manipulations of variables, and answers "what-if" questions. These frameworks typically involve a set of structural equations that quantify how each variable is determined by its direct causes and unobserved noise terms, which capture unmeasured factors and random variation, while also clarifying the distinction between confounding variables that affect both cause and effect and mediators/moderators that transmit or modify causal effects within a coherent causal hierarchy.

19.5 Methodoology

In this section, big letters stand for spaces and small letters stand for the element in the spaces.

19.5.1 Abilities of SCM

SCM is formalized using a triple (V, W, F), where:

- V is a set of endogenous variables.
- W is a set of exogenous variables, which are not caused by any variable in V.
- F is a set of functions $f_1, f_2, \cdots, f_n$ that assign each variable $v_i \in V$ a value based on the values of its causal parents and an exogenous variable W_i, i.e., $v_i = f_i(pa_i, w_i)$.

It provides the following abilities.

- *Explicit Causal Assumptions:* The graphical model forces researchers to make their causal assumptions transparent and testable. The implications of these assumptions (e.g., conditional independencies) can then be checked against the data.
- *Handling of Confounding:* SCM provides a clear framework for defining, identifying, and adjusting for confounding variables, which are common sources of bias in observational studies.
- *Unification of Concepts:* SCM unifies concepts from statistics, graph theory, and potential outcomes into a single coherent framework, facilitating a deeper understanding of causality.
- *Counterfactual Reasoning:* It provides a formal semantics for answering counterfactual questions, which are central to explanation, attribution, and legal reasoning.

19.5.2 Fundamental Methodology

Basic Equations: The core of an SCM is its structural equations. For each endogenous variable V_i, we have:

$$V_i := f_i(PA_i, W_i), \tag{19.3}$$

where PA_i are the direct parents of v_i in the graph and W_i is an exogenous variable. The assignment operator := signifies a causal assignment, not just a mathematical equality.

The do-operator and Intervention: The effect of an intervention that sets a variable X to a value x is represented by the *do*-operator, $\mathbf{do}(X = x)$. This operation modifies the SCM by replacing the structural equation for X with the equation $X = x$. The resulting probability distribution, denoted as $\mathbb{P}(Y|\mathbf{do}(X = x))$ or $\mathbb{P}_x(y)$, is the *interventional distribution*.

Identification Condition: A central question in SCM is whether a causal effect, $\mathbb{P}(Y|\mathbf{do}(X = x))$, can be uniquely determined from the observed data distribution and the causal graph G. This is the problem of *identification*. A key identifiability result is the *back-door criterion*: if a set of variables Z satisfies the back-door criterion relative to (X, Y) (i.e., Z blocks all back-door paths from X to Y and contains no descendants of X), then the causal effect is identified by the adjustment formula:

$$\mathbb{P}(Y|\mathbf{do}(X = x)) = \sum_z \mathbb{P}(Y|X = x, Z = z)\mathbb{P}(Z = z). \tag{19.4}$$

The formula represents the causal effect of an intervention $\mathbf{do}(X = x)$ on Y. It can be interpreted as a weighted sum over all possible values of the mediator variable Z. The formula is composed of two parts:
1. $\mathbb{P}(Y|X = x, Z = z)$: The conditional probability of Y given $X = x$ and $Z = z$.
2. $\mathbb{P}(Z = z)$: The natural distribution of Z without any intervention.

This formula shows that the causal effect $\mathbb{P}(Y|\mathbf{do}(X = x))$ can be computed by adjusting for the mediator Z. It is commonly used in causal inference to estimate intervention effects.

Another powerful identifiability tool is the *front-door criterion*, which can be used even in the presence of unmeasured confounders.

If there is no confounding variables that have paths to both X and Y, we call paths from X to Y textitfront-door paths. In this case, adjustment on variable Z is not feasible because it changes the distribution of X and Y. In this case, the causal effect is computed with the following formula according to the total probability theorem:

$$\mathbb{P}(Y|\mathbf{do}(X = x)) = \mathbb{P}(Y|X = x) = \sum_z \mathbb{P}(Y|X = x, Z = z). \tag{19.5}$$

19.5.3 Procedures

Evaluations using the SCM method follow the following six steps. The fifth step is essential.

1. **Define the Problem**: Define the research problem, factors, and causal assumptions.
2. **Construct a Causal Graph**: Use a DAG to represent causal relationships among variables, marking confounders.
3. **Determine Identifiability**: Check if causal effects can be estimated.
4. **Data Processing**: Collect data and handle missing values, outliers, etc.
5. **Causal Inference**: Compute causal effect using the structure formula, for example 19.4 or 19.5.
6. **Validation and Interpretation**: Assess the reliability of results and get conclusions.

19.6 Example: Infer the Effects of Apple Origins on Purchasing Prices

Structural Causal Model

Case study

In agricultural economic research, identifying the true effect of product origin on market prices is crucial for guiding production layout and optimizing resource allocation. However, this causal inference is often complicated by multiple non-independent influencing factors, such as the intrinsic quality attributes of agricultural products. This example focuses on the causal effect of *apple origin* (X) on *purchase price* (Y). Other influencing factors include sweetness (A), apple size (B), and shelf life (C), where A, B, and C are not mutually independent. Table 19.6 is the price of apples for different variable values.

Variables	(12,4,3)	(12,4,7)	(12,6,3)	(12,6,7)	(14,4,3)	(14,4,7)	(14,6,3)	(14,6,7)
Price 1	¥5	¥6	¥9	¥10	¥7	¥9	¥10	¥16
Price 2	¥11	¥13	¥16	¥16	¥13	¥16	¥18	¥19
Rarity 1	0.25	0.25	0.2	0.1	0.05	0.05	0.05	0.05
Rarity 2	0.2	0.2	0.2	0.15	0.1	0.05	0.03	0.07

Table 19.1: The price (in Yuan(¥)) and rarity (denoted with probability weight) of apples from two origins. "Variables" stands for (Brix, kg, days).

- *Endogenous Variables*: Variables influenced by other variables within the model.
  - X (Apple Origin): 1 = target origin, 0 = other origins;
  - Y (Apple Purchase Price): Continuous variable (unit: e.g., yuan/kg);
  - A (Apple Sweetness): Continuous variable (unit: e.g., Brix value);
  - B (Apple Size): Continuous variable (unit: e.g., diameter in mm or weight in g per fruit);
  - C (Apple Shelf Life): Continuous variable (unit: e.g., days).

- *Exogenous Variables*: Variables not influenced by model-internal variables, serving only as inputs.
  - W_X (Unobserved factors for X): e.g., trace elements in soil of the origin, climate stability;
  - W_Y (Unobserved factors for Y): e.g., short-term capital pressure of buyers, sudden market demand;
  - W_A (Unobserved factors for A): e.g., genetic differences of apple tree varieties, sunlight duration during growth;
  - W_B (Unobserved factors for B): e.g., rainfall during growth, uniformity of fertilization;
  - W_C (Unobserved factors for C): e.g., preliminary processing after picking, storage conditions before transportation.

- *Structural Functions (F)*: Define causal relationships between endogenous variables, reflecting non-independence among A, B, and C.

$$
\begin{aligned}
X &:= f_X(W_X) \\
A &:= f_A(X, W_A) \\
B &:= f_B(X, A, W_B) \\
C &:= f_C(A, B, W_C) \\
Y &:= f_Y(X, A, B, C, W_Y)
\end{aligned}
$$

The causal relationships between multiple variables in this example is presented in Figure 19.1.

Calculation

- *Front-Door Paths*: Including all paths from X to Y. The causal flows between them include: $X \to A \to B \to C \to Y$; $X \to B$; $A \to C$; $X \to Y$; $A \to Y$; $B \to Y$. All exogenous variables W point to their corresponding endogenous variables (e.g., $W_A \to A$).

- *Condition Satisfaction*: Compute with all concerning variables $\{A, B, C\}$ (sweetness, size, shelf life).

1. There are no back-door paths from X to Y, therefore non-causal associations are not involved;
2. There are no ancestors of X in endogenous variables (ancestors refer to variables that can influence X. A, B, and C are influenced by X, not influencing X);
3. Thus, the interventional distribution $\mathbb{P}(Y = y \mid \mathbf{do}(X = x))$ is identifiable.

- *Interventional Distribution*: Calculate the distribution of purchase price under a specific origin ($\mathbf{do}(X = x)$) with formula 19.5 and without cutting-off the paths between X and Y:

$$\begin{aligned}&\mathbb{P}(Y = y \mid \mathbf{do}(X = x))\\ &= \sum_{a,b,c} \mathbb{P}(Y = y \mid X = x, A = a, B = b, C = c)\end{aligned}$$

  Since A, B, and C are not independent, the joint probability $\mathbb{P}(A, B, C)$ is used instead of the product of their marginal probabilities.

- *Average Treatment Effect (ATE)*: Measures the average causal effect difference of "target origin" vs. "other origins" on purchase price:

$$\text{ATE} = \mathbb{E}[Y \mid \mathbf{do}(X = 1)] - \mathbb{E}[Y \mid \mathbf{do}(X = 0)]$$

- *Calculation*: Substitute the concrete values into ATE to get a computable expression:

$$\begin{aligned}&\text{ATE} =\\ &\sum_{a,b,c} \mathbb{E}[Y \mid X = 1, A = a, B = b, C = c]\\ &- \sum_{a,b,c} \mathbb{E}[Y \mid X = 0, A = a, B = b, C = c]\\ &= (5 \times 0.25 + 6 \times 0.25 + 9 \times 0.2 + 10 \times 0.1 + 7 \times 0.05 + 9 \times 0.05\\ &+ 10 \times 0.05 + 16 \times 0.05) - (11 \times 0.2 + 13 \times 0.2 + 16 \times 0.2 + 16 \times 0.15\\ &+ 13 \times 0.1 + 16 \times 0.05 + 18 \times 0.03 + 19 \times 0.07)\\ &= -6.72\end{aligned}$$

Conclusion

The ATE quantifies the average causal effect of *apple origin* on *purchase price*. By computing with the mediator variables (sweetness A, size B, and shelf life C), this effect excludes the interference of confounding factors and only reflects the true impact of "different origins" on the purchase price. The result reports apples grown in Origin 1 had an average purchasing price of ¥6.72/kg *lower* than that of apples grown in Origin 2. Therefore, it is confident that this difference is likely due to the origin itself and not other confounding factors.

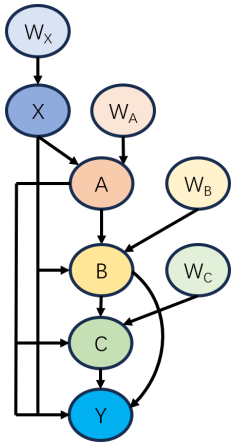

Figure 19.1: The causal relationship model between multiple variables in this example.

19.7 Limitations

SCM is a robust framework for understanding and analyzing causal relationships. However, SCM has the following limitations.

Firstly, it completely relies on the correctness of the causal assumptions as a DAG. If the assumptions are incorrect or not confident, the resulting conclusions may be invalid or misleading.

Secondly, SCM also requires sufficient background knowledge to justify the causal structure, as the evaluation results of SCM cannot specify the true causal-effect relationship.

Thirdly, SCM often necessitates large datasets, particularly in complex models with many variables. Measurement error or unobserved confounding can further compromise the validity of SCM results, as the framework assumes that all relevant variables are correctly measured and included.

Lastly, the computation will be cost-consuming or even infeasible when dealing with high-dimensional data or when attempting to estimate effects in a complicated DAG.

19.8 Summary

Structural causal model is a comprehensive framework for representing and analyzing causal effects among variables. SCM extends traditional statistical association analysis by introducing a framework that represents causal relationships through a system of structural equations and a corresponding graphical model. This framework is applied in fields such as epidemiology, economics, computer science, and social sciences for tasks including causal discovery, mediation analysis, and evaluation of intervention effects. The limitations of SCM is the high cost of data collecting and the unreliable assumptions and causal diagrams.

Chapter 20

Structural Equation Model

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the Structural Equation Model (in short, SEM).

20.1 Basic Concepts

Observed Variables are the variables whose values can be directly obtained in a study [96], such as survey responses, test scores, or any other quantifiable data points. In SEM, they serve as the observed indicators, that is, the observed variables at the statistical level, for latent variables, and their sample covariance matrix provides the empirical basis for a model estimation.

Latent Variables are unobserved variables that can only be inferred indirectly from a theoretical model [96]. They usually represent unobservable constructs, such as psychological traits (e.g., motivation, intelligence) or sociological concepts (e.g., social influence, group norms).

Factor. In SEM, the factor represents the latent variable at the statistical level and is identified through its systematic relationships with observed indicators [96].

Exogenous Variables are "given from the outside [96]," and serve as the theoretical model's inputs or causes. They can be *observed or latent variables* and may co-vary with each other. They influence endogenous variables.

Endogenous Variables are "accounted for by the model [96]." Their variance is explained by other variables within the model. They can be *observed or latent variables*.

Factor Analysis aims to describe the covariance relationships among many variables in terms of a few underlying factors. [95]

A Measurement Model is a model linking latent variables to their observed indicators [21]. It specifies how each latent variable is operationally defined and measured. The measurement model in SEM is essentially a confirmatory factor analysis that specifies how observed variables reflect their underlying latent variables.

[1] The primary contributor is Dr. Lei Wang.

Confirmatory Factor Analysis (CFA) "deals specifically with measurement models—that is, the relationships between observed measures or indicators, e.g., test items, test scores, behavioral observation ratings, and latent variables or factors [24]."

Path Analysis is a statistical method that estimates effects among a set of observed variables based on a hypothesized causal structure, which is visually represented in a path diagram where unidirectional arrows denote the theorized directional influences [218].

A Structural Model specifies the hypothesized causal relationships among variables, particularly between latent variables [96]. In this context, causal relationships refer to the directional influences where changes in an independent variable are theorized to produce changes in a dependent variable. A structural model is a broader conceptual framework that generalizes and extends the principles of path analysis.

A Theoretical Model is a formal specification of hypothesized relationships, which is defined by a measurement model and a structural model.

20.2 Problem Statement

SEM [96, 207, 87, 5, 101] is a powerful statistical technique that allows researchers to analyze complex relationships between variables. SEM is widely used in fields requiring multivariate analysis, especially in the behavioral sciences, where it is used to model psychological, social, and economic constructs.

SEM is a general method for estimating the unknown coefficients in a set of linear structural equations [96]. The SEM problem could be stated as "how to create and validate a theoretical model that includes multiple abstract concepts (latent variables) and their complex causal relationships based on observable data." SEM operationalizes the theoretical model by estimating parameters to quantify the direction and strength of influences, illustrating how exogenous variables affect endogenous variables. Specifically, it simultaneously addresses two questions: "How to accurately measure abstract concepts?" and "What are the causal network relationships between these concepts?"

This problem can be further formulated as a detailed analysis problem. It begins with two primary inputs: the empirical data, which consist of measurements of the observed variables (e.g., survey responses or test scores), and the theoretical model specification. The theoretical model specification articulates the hypothesized relationships using the fundamental concepts of SEM: it defines the latent variables (unobservable constructs) and links them to their observed variables (indicators), while also positing the causal network between exogenous variables (the external inputs or causes, which can be observed or latent) and endogenous variables (the outcomes whose variance is explained by the model). The model is subject to five key constraints, as detailed in Section 20.3. Then, the core task of SEM is to estimate the unknown coefficients in this system of linear structural equations, testing how well the specified model explains the observed data.

20.3 Basic Assumptions

Assumption 20.1 (Independence of Measurement Errors)**.** *The error terms associated with the observed indicators have an expected value of zero, are uncorrelated with each other, and are uncorrelated with the latent constructs they are intended to measure.*

Assumption 20.2 (Multivariate Normal Distribution)**.** *The observed variables exhibit a multivariate normal distribution.*

Assumption 20.3 (Linearity)**.** *SEM typically assumes linear relationships across its components, including linear associations between latent variables and their observed variables, as well as between latent variables themselves.*

Assumption 20.4 (Independence of Structural Residuals and Exogenous Variables)**.** *The structural disturbance terms (errors in equations) have an expected value of zero and are uncorrelated with the exogenous latent variables in the model.*

Assumption 20.5 (Model Identification)**.** *The model must be identified, meaning that there is a unique set of parameter values consistent with the population covariance matrix of the observed variables.*

20.4 Basic Principles

Structural Equation Modeling (SEM) is a statistical framework that combines *Confirmatory Factor Analysis (CFA)* and *Path Analysis* to test theoretical models by evaluating the fit between the covariance structure predicted by the model is represented by its covariance matrix. and the covariance structure observed in empirical data. SEM allows for the simultaneous estimation of both measurement and structural components, providing a comprehensive view of the relationships between latent and observed variables.

A Measurement Model defines how latent variables are measured by observed indicators. CFA is used to validate the hypothesis that a set of observed variables represents an underlying concept. The *Structural Model* specifies the causal relationships between latent variables or between latent and observed variables. The structural models direct and indirect effects in a way that is similar to a system of simultaneous regression equations, allowing for complex interdependencies.

A key strength of SEM is its ability to account for measurement error, unlike traditional regression models that assume independent variables are measured without error. SEM incorporates error terms for both latent and observed variables, resulting in more accurate and unbiased estimates.

The parameters of the SEM include factor loadings (coefficients that measure the strength and direction of a latent variable's influence on its observed indicators), path coefficients (coefficients that represent the strength and direction of the direct influence of one variable on another), variances (primarily including the variances of exogenous

variables, residuals, and measurement errors), and covariances (mainly referring to the associations between exogenous variables, which are often standardized into correlation coefficients for interpretation).

To estimate these parameters, SEM relies primarily on statistical estimation methods that compare the hypothesized model with the observed covariance matrix of the data. The most commonly used method is *Maximum Likelihood (ML)* estimation, which seeks parameter values that maximize the likelihood of observing the data given the assumed model structure. The fit between the hypothesized model and observed data is tested using statistical indices such as the chi-square test (χ^2), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR).

20.5 Methodology

The SEM methodology involves a process, beginning with the formal specification of the model, followed by parameter estimation, and concluding with model evaluation.

SEM is formally represented by a set of linear equations, which are typically divided into two components:

Measurement Model. The measurement component specifies how the latent variables are reflected in the observable indicators. Formally, the relationships for exogenous and endogenous measurements are given by:

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \delta, \tag{20.1}$$

$$\mathbf{y} = \Lambda_y \boldsymbol{\eta} + \epsilon, \tag{20.2}$$

where $\mathbf{x}$ denotes the observed indicators associated with the exogenous latent variables $\boldsymbol{\xi}$, and $\mathbf{y}$ denotes the observed indicators associated with the endogenous latent variables $\boldsymbol{\eta}$. The matrices Λ_x and Λ_y represent the corresponding factor loadings, while δ and ϵ capture measurement errors for $\mathbf{x}$ and $\mathbf{y}$, respectively.

Structural Model. The structural component characterizes the causal relations among the latent variables. Specifically, the endogenous latent variables are determined by both other endogenous factors and the exogenous inputs:

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \zeta, \tag{20.3}$$

where $\boldsymbol{\eta}$ represents the vector of endogenous latent variables and $\boldsymbol{\xi}$ denotes the exogenous latent variables. The matrix B captures the regression relations among endogenous variables, whereas Γ represents the effects of exogenous variables on endogenous ones. The term ζ accounts for structural residuals(disturbances) not explained by the model.

These equations form the foundation of SEM, enabling the simultaneous modeling of the measurement structure of latent constructs and the structural (causal or associative) relationships among them.

The primary method for estimating parameters in SEM is *Maximum Likelihood (ML) estimation*, which assumes that the observed variables follow a *multivariate normal distribution*. ML estimates model parameters—such as factor loadings, path coefficients, variances, and covariances—by minimizing the difference between the observed sample covariance matrix and the covariance matrix implied by the hypothesized model. Under the assumption of multivariate normality, ML provides efficient and unbiased parameter estimates.

The ML estimation method also generates key *goodness-of-fit indices* that help test how well the model reproduces the observed data structure. After estimating the model, researchers evaluate its overall fit using a combination of *goodness-of-fit indices*. While the *Chi-square (χ^2) statistic* tests the exact fit between the observed and model-implied covariance matrices, it is highly sensitive to sample size and is therefore interpreted cautiously. More practically, researchers rely on *absolute fit indices* (e.g., *RMSEA*, *SRMR*) and *incremental fit indices* (e.g., *CFI*) to test model adequacy. Commonly used fit criteria include:

- *CFI (Comparative Fit Index)*: Values greater than 0.90 are generally regarded as indicative of an acceptable model fit, whereas values exceeding 0.95 denote a good fit.
- *RMSEA (Root Mean Square Error of Approximation)*: Values below 0.08 typically indicate an acceptable fit, while those below 0.05 reflect a close or good fit.
- *SRMR (Standardized Root Mean Square Residual)*: Values below 0.08 are commonly interpreted as indicative of an acceptable fit.

These indices should be interpreted in combination, taking into account model complexity, theoretical expectations, and sample characteristics.

20.6 Example: Infer the Effects of Apple Origins on Purchasing Prices

This example uses a hybrid SEM approach to explore how *the apple origins* influence the purchasing price through key quality characteristics. We employ a mixed-variable design where an origin is modeled as a latent construct, while apple characteristics and prices are directly observed. This approach balances methodological rigor with practical interpretability.

Structural Equation Model

Variable Specification

The model uses a combination of latent and manifest variables based on agricultural measurement standards:

- *Origin (OR)*: *Latent variable* representing regional cultivation advantage, measured by 3 observed indicators:
  - OR1: Climate suitability (1-5 scale, 5 = most suitable for apple growth)
  - OR2: Soil organic matter content (g/kg, 15-25 g/kg = optimal range)
  - OR3: Average annual temperature (°C, 8-12 °C = ideal for sugar accumulation)
- *Sweetness (SW)*: *Observed variable* measured in Brix (continuous):
  - Typical range: 10-16 Brix for commercial apples
- *Size (SI)*: *Observed variable* (continuous):
  - Typical range: 120-220 g for commercial apples
- *Shelf Life (SL)*: *Observed variable* measured in days (continuous):
  - Typical range: 3-14 days under normal temperature storage
- *Purchasing Price (PP)*: *Observed variable* measured in ¥/kg:
  - Directly observed market transaction price

Note: In the path model, SW, SI, SL, and PP function as endogenous variables (being influenced by other variables), while OR serves as the exogenous latent variable.

Theoretical Framework and Hypotheses

Based on agricultural economics research, we hypothesize that superior origin conditions directly enhance apple quality attributes, which in turn increase market value. The proposed causal pathways are:

$$OR \to SW, \quad OR \to SI, \quad OR \to SL$$
$$SW \to PP, \quad SI \to PP, \quad SL \to PP$$

Theoretical Rationale:

- Better origin conditions (optimal climate, soil, temperature) promote sugar accumulation (↑ sweetness), fruit development (↑ size), and structural integrity (↑ shelf life)
- Higher quality attributes command price premiums in the market due to consumer preferences for sweeter, larger, and longer-lasting apples
- Origin may also exert a direct price effect through regional branding and reputation effects

Model Specification

1. the Measurement Model (for the only exogenous latent variable *Origin*):

$$OR1 = \lambda_{OR1} OR + \delta_{OR1}, \quad (\lambda_{OR1} = 1, \text{ marker variable for identification})$$
$$OR2 = \lambda_{OR2} OR + \delta_{OR2}$$
$$OR3 = \lambda_{OR3} OR + \delta_{OR3}$$

where λ = factor loading, δ = measurement error. The outcome of the measurement model—the factor scores of the latent variable Origin (OR)—is computed and subsequently treated as observable data within the structural model for the estimation of path coefficients and future predictions.

2. the Structural Model (path analysis among all variables):

$$SW = \beta_1 OR + \zeta_1 \qquad \text{(Origin} \rightarrow \text{Sweetness)} \tag{20.4}$$

$$SI = \beta_2 OR + \zeta_2 \qquad \text{(Origin} \rightarrow \text{Size)} \tag{20.5}$$

$$SL = \beta_3 OR + \zeta_3 \qquad \text{(Origin} \rightarrow \text{Shelf Life)} \tag{20.6}$$

$$PP = \gamma_1 OR + \gamma_2 SI + \gamma_3 SW + \gamma_4 SL + \zeta_4 \qquad \text{(Total effects} \rightarrow \text{Price)} \tag{20.7}$$

where β, γ = path coefficients, ζ = structural disturbances.

Parameter Estimation and Model Identification

Estimation Results (hypothetical but realistic parameters based on agricultural research):

- **Measurement model parameters:**

$$\begin{aligned} \lambda_{OR1} &= 1.00 \quad \text{(fixed for identification)} \\ \lambda_{OR2} = 0.75, &\quad \lambda_{OR3} = 0.68 \quad \text{(freely estimated)} \\ \text{Var}(\delta_{OR1}) = 0.20, &\quad \text{Var}(\delta_{OR2}) = 0.25, \quad \text{Var}(\delta_{OR3}) = 0.30 \end{aligned}$$

- **Structural model parameters:**

$$\begin{aligned} \beta_1 = 0.55, &\quad \beta_2 = 0.48, \quad \beta_3 = 0.62 \\ \gamma_1 = 0.25, &\quad \gamma_2 = 0.30, \quad \gamma_3 = 0.35, \quad \gamma_4 = 0.20 \end{aligned}$$

All parameters are statistically significant at $p < 0.05$, supporting the hypothesized relationships.

Numerical Prediction Example

This example demonstrates how the model can be used for prediction. With estimated parameters, the structural equation for price is:

$$\hat{PP} = 0.25 \times OR + 0.30 \times SI + 0.35 \times SW + 0.20 \times SL + \zeta_4$$

For apples with specific quality attributes:

$$SW = 14 \text{ Brix}, \quad SI = 165g, \quad SL = 7 \text{ days}$$

For a region with above-average origin conditions (where the origin condition score, OR = 3.8, is calculated as a weighted combination of the region's OR1, OR2, and OR3 values, with the weights determined by the measurement model parameters), the predicted price component derived from observed factors is:

$$\begin{aligned} \hat{PP} &= 0.25 \times 3.8 + 0.30 \times 1 + 0.35 \times 14 + 0.20 \times 7 \\ &= 0.95 + 0.30 + 4.90 + 1.40 = 7.55 \text{ ¥/kg} \end{aligned}$$

Accounting for unmodeled factors ($\zeta_4 = -1.05$), the final predicted price is:

$$\hat{PP} = 7.55 - 1.05 = 6.50 \text{ ¥/kg}$$

Effect Decomposition Analysis

We decompose the total effect of the apple origins on the purchasing price to quantify mediation pathways.

1. Indirect effects:

$$\begin{aligned} \text{Through sweetness:} &\quad \beta_1 \times \gamma_3 = 0.55 \times 0.35 = 0.1925 \\ \text{Through size:} &\quad \beta_2 \times \gamma_2 = 0.48 \times 0.30 = 0.1440 \\ \text{Through shelf life:} &\quad \beta_3 \times \gamma_4 = 0.62 \times 0.20 = 0.1240 \\ \text{Total indirect effect:} &\quad 0.1925 + 0.1440 + 0.1240 = 0.4605 \end{aligned}$$

2. Direct effects :

$$\text{Direct effect} = \gamma_1 = 0.25$$

3. Total effect:

$$\text{Total effect} = \text{Direct} + \text{Indirect} = 0.25 + 0.4605 = 0.7105$$

Based on the standardized results, the *total effect* of origin advantage (OR) on price (PP) is 0.7105. This means that when OR increases by one standard deviation, PP increases by 0.71 standard deviations on average. The *direct effect* is 0.25, and the *indirect effect* is 0.4605—mediated through sweetness (0.1925), size (0.1440), and shelf life (0.1240)—corresponding to a *direct share* of 35.2% and an *indirect share* of 64.8%.Figure 20.1 is the graphical representation.

Model Fit and Validation

Goodness-of-Fit Indices (hypothetical results):

- $\chi^2/df = 2.15$ (acceptable: < 3.0)
- CFI = 0.94 (good fit: > 0.90)
- RMSEA = 0.06 (good fit: < 0.08)
- SRMR = 0.05 (good fit: < 0.08)

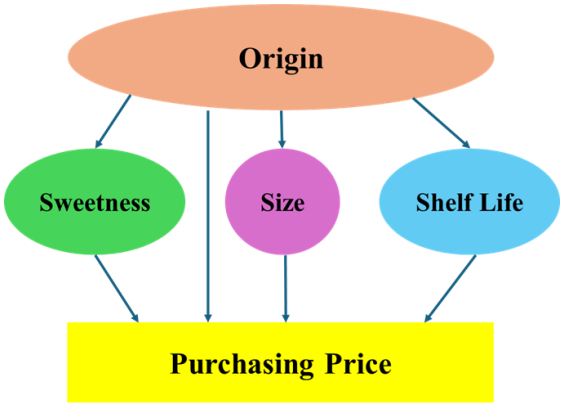

Figure 20.1: SEM path diagram for apple pricing model

In summary, this example demonstrates SEM's applicability in agricultural price analysis—by standardizing indicators (10-16 Brix sweetness, 120-220g size, 3-14 day shelf life), the model's results are more actionable for producers and marketers. The impact decomposition clarifies that the apple origin influences the purchasing price mainly through improving sweetness. This insight can help regions focus on sugar accumulation techniques to enhance apple market value.

20.7 Limitations

SEM is a powerful tool. However, SEM has limitations that researchers must consider. It generally requires many assumptions to be satisfied, including linearity, normality, and independence of errors. It also generally requires a large sample size, especially for complex models, as small samples can lead to unreliable results. Model misspecification can also produce misleading conclusions, emphasizing the importance of grounding model design in theoretical and empirical evidence. Furthermore, SEM can be computationally intensive, particularly when incorporating nonlinear relationships or categorical variables, necessitating robust software and computational resources. Lastly, while SEM can identify associations between variables, it does not inherently establish causality, so causal inferences should be made cautiously.

20.8 Summary

Structural Equation Model (SEM) is a powerful tool used to analyze complex relationships in multivariate data. It allows researchers to model latent variables, handle measurement errors, and examine the structure of relationships across various fields, such as psychology, sociology, education, and marketing. SEM is particularly useful for understanding how multiple variables interact simultaneously, making it invaluable in the behavioral sciences. However, it is essential to pay close attention to model spec-

ification, data quality, and sample size to avoid common pitfalls such as overfitting or misspecification.

Chapter 21

The Potential Outcome Theory

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and a case study of the potential outcome (PO) theory.

21.1 Basic Concepts

A Unit refers to "physical objects at particular points in time, e.g., plots of land, individual people, one person at repeated points in time" [158].

A Treatment represents the intervention or condition assigned to a unit, often denoted by a binary indicator T, where $T = 1$ denotes the active treatment and $T = 0$ the control treatment [158]. Generally, PO uses treatment to denote "active treatment" and control to denote "control treatment," respectively [158].

An Assignment Mechanism refers to "a probabilistic model for the treatment each unit receives as a function of covariates and potential outcomes" [158].

A Covariate represents "a variable that takes its values before the treatment assignment or cannot be affected by the treatment, such as the sex of the unit [158]." Covariates are also called pretreatment variables [155].

A Posttreatment Variable refers to "the variables describing experimental units that might be recorded after the assignment of treatment [155]." These variables may be influenced by the treatment itself or by other post-assignment factors. For example, in an agricultural experiment where the treatment is the use of a specific fertilizer, the *chlorophyll content* or *fruit size* measured at harvest would be posttreatment variables, since they are recorded after the fertilizer has been applied [155].

A Mediator is a specific type of posttreatment variable that not only occurs after the treatment but also transmits part of the treatment's causal effect to the outcome [157]. Formally, for each unit i, let $M_i(t)$ denote the potential outcome of the mediator under treatment level t, and let $Y_i(t, m)$ denote the potential outcome as a function of both the treatment t and the mediator m. Under this framework, the *natural direct effect (NDE)*

[1] The primary contributor is Dr. Wanling Gao.

and *natural indirect effect (NIE)* are defined as [140, 157]:

$$NDE = \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))], \tag{21.1}$$

$$NIE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]. \tag{21.2}$$

The mediator thus represents a causal pathway through which the treatment influences the outcome, beyond the direct effect of treatment itself.

PO refers to the outcome that would be observed for a unit at a particular point in time after a specific action (i.e., treatment or control). Specifically, $Y(1)$ represents the PO under the active treatment and $Y(0)$ under the control treatment. Only one of these outcomes can be observed in reality, depending on the treatment actually received [158].

A Counterfactual refers to the unobserved potential outcome for a unit under a treatment condition different from the one it actually received. Since each unit can experience only one treatment state, the counterfactual outcome remains unobserved and represents the core challenge of causal inference, estimating what would have happened under the alternative treatment scenario.

Causal Effects are defined as "comparisons of POs under different treatments on a common set of units [158]."

21.2 Problem Statement

This section first illustrates the fundamental problem of causal inference under the PO framework and then describes its problem formulation.

21.2.1 Fundamental Problem of Causal Inference: Missing Data Problem

A crucial implication of this framework is that causal inference is inherently a missing data problem [89]: for each unit, at most one of the potential outcomes is observed, while the other remains counterfactual. This creates the central challenge of causal inference —how to estimate treatment effects when one of the outcomes is always unobserved.

More fundamentally, at any given time point, a unit cannot simultaneously both receive and not receive a treatment, nor can it be exposed to multiple alternative treatments at once. Only one realized path is observed, while all other potential outcomes remain unobserved counterfactuals. Although a unit can indeed receive different treatments across different time periods, the states of the unit are not interchangeable across time. Outcomes at later time points may be influenced not only by the treatment assigned at that time, but also by earlier interventions or by evolving characteristics of the unit itself. For example, if a patient takes a drug at time t, the observed effect at time $t+1$ may reflect not only the drug's direct causal impact, but also prior exposures, accumulated biological responses, or progression of the underlying disease. These time-varying factors act as confounders, complicating the attribution of observed differences solely to the treatment of interest.

21.2.2 Problem Statement

Given a population of observational or experimental units, each of which can receive one of several treatment conditions ($T \in \{0, 1\}$), let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes that unit i would exhibit under the treatment and control conditions, respectively. For each unit, only one potential outcome can be observed, depending on the treatment actually assigned, while the other remains unobserved.

The unknown quantity is the unobserved potential outcome for each unit, which prevents direct observation of the individual-level treatment effect ($Y_i(1) - Y_i(0)$). This limitation defines the core challenge of the potential outcome framework: the impossibility of simultaneously observing both potential outcomes for the same unit.

The PO problem is stated as "how to infer or estimate the causal effect of the treatment on the outcome based on observed data comprising the treatment assignment, covariates, and realized outcomes." It begins with one primary input: the observed data, consisting of the realized outcome for each unit together with its treatment assignment and pre-treatment covariates. The model is subject to a set of well-established constraints as detailed in Section 21.3 that govern the relationship between the potential outcomes, the assignment mechanism, and the observed data. Given these ingredients, the core task of causal inference under the potential outcomes framework is to identify and estimate causal estimands—such as the average treatment effect (ATE) or related population-level contrasts—using only the observed data together with the assumptions encoded in the PO. The framework thereby provides a disciplined way to formalize what aspects of the causal effect are identifiable and to what extent the observed data support inferences about the unobserved potential outcomes.

Inputs. The observed data $\mathcal{D}$ consist of

$$\mathcal{D} = \{(Y_i^{\text{obs}}, T_i, X_i)\}_{i=1}^{n}, \tag{21.3}$$

where Y_i^{obs} is the realized outcome for each unit i, $T_i \in \{0, 1\}$ indicates its corresponding treatment assignment, and X_i denotes pre-treatment covariates.

Outputs (Causal Estimands). The target quantities are causal effects defined in terms of the joint distribution of the potential outcomes, including:

$$\text{ATE: } \mathbb{E}[Y_i(1) - Y_i(0)], \tag{21.4}$$

$$\text{ATT: } \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1], \tag{21.5}$$

and covariate-conditional effects such as the CATE,

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]. \tag{21.6}$$

Constraints (Assumptions). The PO is subject to three key constraints, as detailed in Section 21.3.

21.3 Basic Assumptions

To make causal effects identifiable and estimable, the PO framework relies on three foundational assumptions:

Assumption 21.1 (Stable Unit Treatment Value Assumption (SUTVA) [43, 158].)**.** *A fundamental assumption in causal inference is SUTVA. Each unit's PO depends only on its own treatment status; there is no interference between units [43, 158], and treatments are well-defined. This principle ensures that PO is well-defined and comparable across units. It consists of two essential components: (1) no interference between units, and (2) no hidden versions of treatments [158].*

First, the no-interference condition requires that the treatment assigned to one unit does not alter the PO of another. Consider, for example, the evaluation of a personalized recommendation system in an online learning platform. If we assume that the courses recommended to Student A do not influence Student B's learning outcomes, then this part of SUTVA is satisfied. In practice, however, this assumption may fail. Suppose the platform has limited seats in certain courses; if Student A is recommended and enrolls in a popular class, Student B may be excluded from it, thereby indirectly affecting B's outcome. In such settings, interference needs to be explicitly addressed to avoid biased causal estimates.

Second, the assumption that there is no hidden version of treatments requires that each treatment level be uniquely and consistently defined. If we define the treatment as "participating in a training program," it is crucial that this program is homogeneous across participants. If some trainees attend a three-day workshop while others undergo a three-month intensive curriculum, then the label "training" actually corresponds to multiple distinct interventions. In this case, potential outcomes cannot be uniquely mapped to a single treatment level, violating SUTVA. A common strategy to resolve this issue is to refine the definition of treatment—distinguishing, for instance, between "short-term training" and "long-term training," so that each POS corresponds clearly to a well-defined intervention.

In summary, SUTVA provides the conceptual clarity needed for defining and estimating causal effects. Without it, the PO may be confounded by interference across units or ambiguities in treatment definitions. Although in many applied settings SUTVA may only hold approximately, recognizing its role and limitations is an indispensable step in credible causal inference.

Assumption 21.2 (Ignorability/Unconfoundedness [151, 154])**.** *Conditional on observed covariates X, treatment assignment is independent of potential outcomes $(Y(0), Y(1))$:*

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X. \tag{21.7}$$

This assumption is the cornerstone for establishing causality from observational data. The symbol "$\mid$" is a mathematical notation denoting "under the condition of," and "$A \mid B$"

signifies "A under the condition of B." The symbol "⫫" represents "mutual independence," with "$A \perp\!\!\!\perp B$" indicating that "A and B are mutually independent." The formula asserts that, once we account for (or condition on) all relevant pre-treatment covariates X, the mechanism that determines whether a unit receives the treatment ($T = 1$) or control ($T = 0$) becomes independent of what the outcomes would have been under either scenario. In essence, after controlling for X, the treatment groups are comparable, and any remaining difference in outcomes can be attributed to the treatment itself. This is equivalent to saying there are no unmeasured confounders: all common causes of T and Y are captured in X.

Consider an online learning platform assessing the effect of a new premium interactive module ($T = 1$) versus the standard module ($T = 0$) on final exam scores (Y), where students self-select into the premium module. The ignorability assumption requires that we can measure all covariates X (e.g., prior GPA, time spent on the platform, past quiz scores) that influence both a student's decision to enroll and their eventual exam score. If this assumption holds, then among students who are identical on these measured covariates X, choosing the premium module is as good as a random assignment. This allows for a fair comparison, as the groups are balanced with respect to both observed and, critically, any unobserved factors that are correlated with the observed X.

Assumption 21.3 (Positivity/Overlap [151])**.** *For all possible values of covariates X, the probability of treatment assignment is strictly between 0 and 1:*

$$0 < P(T = 1 \mid X) < 1. \tag{21.8}$$

This ensures sufficient overlap in covariate distributions between treatment and control groups. In the online learning platform example, this assumption requires that for every student profile (e.g., all ranges of prior achievement levels), there exists a non-zero probability of both enrolling and not enrolling in the new module. If the platform restricts module access based on certain criteria (e.g., prohibiting low-performing students from enrolling), the assumption is violated for the affected subpopulations, rendering causal inference infeasible in those strata due to a lack of counterfactual data.

These assumptions provide the theoretical scaffolding that allows researchers to link observed or experimental data to causal effects.

21.4 Origin of Ideas

PO, also known as the Neyman–Rubin Causal Model (NRCM), is one of the most influential paradigms in modern causal inference. Its intellectual roots trace back to Jerzy Neyman (1923) [183, 156], who first introduced the concept of PO in the context of randomized experiments in agricultural research.

In his 1923 doctoral thesis, Jerzy Neyman introduced the notion of "potential yields" to analyze agricultural field experiments. He imagined that for each plot of land, we could think of the potential yield that would result if that plot were planted with each of

the different varieties. In reality, however, only one crop variety can actually be planted on a given plot, so we only get to observe one of these potential yields. The other yields remain unobserved. This way of thinking —that each unit (in this case, each plot of land) has multiple possible outcomes, only one of which is realized depending on the treatment it receives (which crop is planted) —is the basic idea behind the modern PO framework.

Neyman then asked a practical question: if we want to compare the average yields of two varieties, how do we calculate the uncertainty (variance) of that comparison, given that we cannot observe all the potential yields?

Neyman derived a formula for the variance of the difference between the observed averages. This variance depends not only on the variability of yields across plots but also on the correlation r between the potential yields of the two varieties on the same plot. Here, r represents how similarly the two varieties would have performed on the same piece of land if each had been planted. The challenge is that r cannot be directly observed—since a plot can host only one variety—so it introduces uncertainty into the variance estimate.

To address this uncertainty, Neyman adopted a conservative strategy, setting $r = 1$. While this assumption may appear counterintuitive, it functions as a *principled lower-bound estimate* for the variance. Specifically, the variance of the difference between the sample means of two varieties can be expressed as

$$\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma_1^2}{p_1} + \frac{\sigma_2^2}{p_2} - 2r\frac{\sigma_1\sigma_2}{p}, \tag{21.9}$$

where σ_1^2 and σ_2^2 are the variances of the potential yields for the two varieties, r is the correlation between their potential yields on the same plot, and p is the number of plots. The corresponding standard error (SE) is given by

$$\mathrm{SE} = \sqrt{\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2)}, \tag{21.10}$$

and the t-statistic for testing the difference in means is

$$t = \frac{\text{Estimated Difference}}{\mathrm{SE}}. \tag{21.11}$$

Since r is unobservable, setting $r = 1$ maximizes the covariance term $2r\sigma_1\sigma_2/n$, which in turn *reduces the subtracted portion in the variance formula* and thus yields a *larger estimated variance*. A larger variance increases the standard error, producing a more conservative t-statistic. This ensures that the comparison does not overstate significance: if the observed difference is non-significant, one can be reasonably confident that the lack of effect is genuine, minimizing the risk of false-positive conclusions. Conversely, any significant finding should still be interpreted with caution, acknowledging that the true correlation may differ from the assumed upper bound.

This idea of "potential yield" was Neyman's original way of formalizing the fact that outcomes depend on the treatment assigned. Decades later, Donald Rubin (1972 [154],

1974, 1978 [155], 2004 [157], 2005 [158], 2006 [159]) reformulated and generalized this concept under the name PO, which has since become the foundation of modern causal inference. Rubin's contribution transformed Neyman's idea into a general statistical methodology for causal reasoning, applicable well beyond randomized trials.

21.5 Basic Principles

Rubin elegantly reframes this causal inference issue as a systematic attempt to recover or approximate these missing counterfactuals through careful design, appropriate modeling, or both.

In Rubin's formalization, each unit i in a study is associated with a pair of potential outcomes: $(Y_i(1), Y_i(0))$, where $Y_i(1)$ represents the outcome if the unit receives the treatment, and $Y_i(0)$ represents the outcome if the unit does not. The observed outcome is given by:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \tag{21.12}$$

where $T_i \in \{0, 1\}$ is the treatment assignment indicator.

This representation makes explicit the fact that only one of the two outcomes can ever be observed for each unit. The causal effect at the individual level is thus defined as the *Individual Treatment Effect*:

$$\tau_i = Y_i(1) - Y_i(0), \tag{21.13}$$

while population-level effects, such as the *Average Treatment Effect (ATE)*, are defined as:

$$\tau = \mathbb{E}[Y(1) - Y(0)]. \tag{21.14}$$

Since individual effects are generally unidentifiable, most analyses focus on population-level averages such as the ATE, or subgroup-level effects like the Conditional Average Treatment Effect (CATE).

21.6 Methodology

Within the PO framework, the estimation of causal effects proceeds from the fundamental problem of missing counterfactuals—only one of $Y_i(1)$ or $Y_i(0)$ is observable for each unit i. Therefore, the methodological focus lies in constructing valid comparisons that approximate the distribution of the missing PO.

The methodology [155, 160, 154] for estimating causal effects can be broadly divided into two settings: *randomized experiments* and *observational studies*.

21.6.1 Randomized Experiments

In a randomized experiment, treatment assignment T_i is independent of potential outcomes $(Y_i(1), Y_i(0))$. This design guarantees the condition of ignorability by construction:

$$(Y(1), Y(0)) \perp\!\!\!\perp T. \tag{21.15}$$

Under this setting, unbiased estimation of the average treatment effect (ATE) can be achieved directly through sample means:

$$\hat{\tau}_{ATE} = \bar{Y}_{T=1} - \bar{Y}_{T=0}, \tag{21.16}$$

where $\bar{Y}_{T=1}$ and $\bar{Y}_{T=0}$ are the sample averages of observed outcomes in the treated and control groups, respectively. The randomization ensures that differences in sample means consistently estimate $E[Y(1) - Y(0)]$ without further adjustment. The sampling variance can be estimated using standard methods for independent samples, allowing for inference through confidence intervals and hypothesis testing.

21.6.2 Observational Studies

When randomization is not possible, treatment assignment may depend on pre-treatment covariates X. In this setting, unbiased estimation requires conditioning on X to recover the independence between treatment and potential outcomes:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X. \tag{21.17}$$

The estimation then proceeds by comparing treated and control units with identical or similar covariate values.

Several classical estimation strategies fall under this approach:

(a) Covariate Adjustment (Regression Estimation)

A parametric model is specified for the conditional expectation of Y given T and X:

$$\mathbb{E}[Y \mid T, X] = \alpha + \tau T + f(X), \tag{21.18}$$

where $f(X)$ captures the effect of covariates. The coefficient τ provides an estimate of the causal effect, assuming the model is correctly specified and ignorability holds.

(b) Matching Estimation

Units in the treatment and control groups are paired based on similarity in X, thereby approximating the counterfactual outcomes. The causal effect is estimated by averaging outcome differences across matched pairs:

$$\hat{\tau}_{match} = \frac{1}{N_T} \sum_{i \in T=1} (Y_i - Y_{j(i)}), \tag{21.19}$$

where $j(i)$ denotes the matched control unit for treated unit i. Matching directly implements the principle of comparing "statistical twins" under equivalent covariate conditions.

(c) Propensity Score Methods

Rubin and Rosenbaum [151] introduced the *propensity score*—the probability of receiving treatment given covariates:

$$e(X) = P(T = 1 \mid X). \tag{21.20}$$

Conditioning on $e(X)$ rather than X itself preserves the ignorability property:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid e(X). \tag{21.21}$$

This allows estimation through several equivalent procedures:

- *Stratification or subclassification:* dividing the sample into strata based on estimated propensity scores and computing within-stratum mean differences;
- *Weighting:* using inverse probability weighting (IPW) to balance covariate distributions between groups;
- *Matching:* pairing treated and control units based on the estimated propensity score.

Each approach produces an unbiased estimate of $\mathbb{E}[Y(1) - Y(0)]$ under correct specification of $e(X)$.

21.6.3 Inference and Uncertainty Estimation

Uncertainty in estimated causal effects arises from both sampling variability and the estimation of counterfactuals. Standard errors are computed under the assumption of independent units, and Rubin proposed using repeated sampling logic [156] or the Bayesian posterior variance [155, 158] to quantify uncertainty. In Bayesian implementations, the PO and parameters are treated as random variables, and the posterior distribution of the causal effect provides direct probabilistic statements about uncertainty.

21.6.4 Censoring and Principal Stratification

In many empirical studies, especially in biomedical, social, and economic research, causal inference is complicated by the fact that some post-treatment outcomes may be *undefined* for certain units. This situation, often referred to as *censoring* or *truncation*, occurs whenever the measurement of the outcome of interest is not possible for all units due to intermediate events, which may include death, dropout, non-response, or other absorbing events [159]. Such censoring violates the standard potential outcome assumption that both potential outcomes, $Y_i(1)$ and $Y_i(0)$, are well-defined for every unit i.

To address this problem, Rubin [159] introduced the framework of *principal stratification*, which defines causal effects within subpopulations (principal strata) characterized by the joint potential outcomes of the intermediate post-treatment variable. Let $S_i(1)$ and $S_i(0)$ denote the potential outcome of an intermediate variable under treatment and control, respectively. The pair $(S_i(1), S_i(0))$ defines principal strata that classify

units according to how the intermediate event would manifest under both treatment conditions.

For example, if the intermediate variable is survival, the strata are:

$$\begin{aligned} &\text{Always-survivors:} && S_i(1) = 1,\ S_i(0) = 1, \\ &\text{Protected:} && S_i(1) = 1,\ S_i(0) = 0, \\ &\text{Harmed:} && S_i(1) = 0,\ S_i(0) = 1, \\ &\text{Never-survivors:} && S_i(1) = 0,\ S_i(0) = 0. \end{aligned} \tag{21.22}$$

In general, causal effects on the post-treatment outcome Y are only well-defined for units in strata where $Y_i(1)$ and $Y_i(0)$ are both defined. For the survival example, this corresponds to the *always-survivors* stratum, and the stratum-specific causal effect is:

$$\tau_{AS} = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i(1) = S_i(0) = 1]. \tag{21.23}$$

Since stratum membership $(S_i(1), S_i(0))$ is never fully observed (each unit experiences only one treatment level), estimation requires assumptions or auxiliary information. Covariates predictive of stratum membership can be used in modeling approaches to infer causal effects probabilistically, for example via Bayesian or likelihood-based methods.

Key considerations in this framework include:

- *Stratum-specific estimands:* When post-treatment outcomes are undefined for part of the population, the overall average treatment effect (ATE) may not be well-defined. Researchers should specify the principal stratum of interest.
- *Separating effects:* Treatments may affect both the intermediate variable (e.g., survival or censoring event) and the post-treatment outcome. Principal stratification conceptually disentangles these effects, though identification often requires strong assumptions.

In practice, inference often involves joint modeling of the intermediate variable and the outcome, or sensitivity analyses to assess the robustness of conclusions under alternative specifications of unobserved strata.

In summary, censoring due to any intermediate post-treatment event poses challenges for causal inference under the PO framework. Principal stratification provides a coherent approach to define and estimate causal effects within well-defined subpopulations, preserving the logic of PO while acknowledging the limits of identifiability in the presence of post-treatment truncation.

21.6.5 Conclusion

In summary, the PO methodology defines causal effects as contrasts between well-defined potential outcomes and provides a coherent framework for estimating these effects under both randomized and non-randomized designs. Through conditioning, matching, or weighting, it reconstructs the missing counterfactuals under explicit assumptions, yielding interpretable and statistically principled estimates of causal effects. Its strength

lies in its conceptual clarity: it separates the *definition of causality* from the *estimation procedure*, ensuring that all empirical analyses remain grounded in a transparent counterfactual logic.

By framing causal inference as a missing data problem governed by clear assumptions, the PO framework provides a unifying statistical paradigm. RCTs naturally satisfy the ignorability assumption by design, while observational studies require careful adjustment strategies—such as matching, regression, propensity scores, or instrumental variables, which are discussed in Chapter 22—to approximate randomization. This versatility explains why the PO framework has become the dominant language of modern causal inference, offering both clarity in conceptual reasoning and rigor in statistical estimation.

21.7 Example: Infer the Effects of Apple Origins on Purchasing Prices

This example reformulates the apple purchasing price problem under the PO framework originally proposed by Neyman (1923) and Rubin (1974). The goal is to estimate the causal effect of *apple origin (PL)* on the *purchasing price (APP)* of apples, both directly and indirectly through key mediating characteristics: *sweetness (SW)*, *size (AS)*, and *shelf life (SL)*.

Units, Treatment, and Potential Outcomes

Each apple batch or shipment represents a distinct *unit* $i \in \{1, 2, \ldots, N\}$. The *treatment variable* PL_i denotes the *origin status* of apples, representing whether they come from a favorable cultivation region (e.g., advantageous climate or soil). For simplicity, we assume two treatment levels:

$$PL_i \in \{0, 1\}, \quad 0 = \text{less favorable origin}, \quad 1 = \text{favorable origin}.$$

For each unit i, the PO for purchasing price is defined as:

$$Y_i(PL_i) = \textit{Purchasing price that would be observed if treatment } = PL_i.$$

However, only one of these potential outcomes is observed:

$$Y_i^{obs} = Y_i(PL_i^{obs}),$$

while the counterfactual outcome $Y_i(1 - PL_i^{obs})$ remains unobserved.

Mediators and Outcome

To capture the mechanisms through which *the apple origin* influences the purchasing price outcome, we define a vector of mediating variables for each unit:

$$M_i = (SW_i, AS_i, SL_i),$$

corresponding to sweetness, size, and shelf life.
Each mediator is itself affected by the treatment, and thus has its own potential outcomes:

$$SW_i(PL_i), \quad AS_i(PL_i), \quad SL_i(PL_i).$$

The PO for purchasing price can therefore be represented as:

$$Y_i(PL_i, M_i(PL_i)),$$

indicating that the apple origins may influence purchasing price both directly (via PL_i) and indirectly (via M_i).

Causal Estimands

Within the PO framework, the *Average Treatment Effect (ATE)* of the apple origin on the purchasing price is defined as:

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)],$$

which represents the expected causal effect on the purchasing price when switching from a less favorable to a favorable origin.
To decompose this total effect into direct and indirect components, we define the *Natural Direct Effect (NDE)* and *Natural Indirect Effect (NIE)* as:

$$NDE = \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))],$$

$$NIE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))].$$

- The *NDE* measures how the price would change if the treatment changed, but mediators were fixed at their counterfactual values under the less favorable origin. - The *NIE* measures how the price changes solely due to treatment-induced changes in mediators, holding the treatment level fixed.
By definition:

$$ATE = NDE + NIE.$$

Covariates and Assignment Mechanism

We denote X_i as a vector of pre-treatment covariates (e.g., production scale, harvest period, market conditions) that may confound the treatment–outcome relationship. The *assignment mechanism* $P(PL_i = 1 \mid X_i)$ governs how treatment is assigned given these covariates. Proper adjustment for X_i ensures that the causal estimands are identifiable under standard assumptions.

Assumptions for Identification

Identification of ATE, NDE, and NIE requires the following PO assumptions:

- *Stable Unit Treatment Value Assumption (SUTVA):* Each unit's PO depends only on its own treatment and mediators, implying no interference between units and no multiple treatment versions:

$$Y_i(PL_i, M_i(PL_i)) \text{ and } M_i(PL_i) \text{ are well-defined for all } i.$$

- *Ignorability (Unconfoundedness):* Conditional on covariates X_i, treatment assignment is independent of PO and mediators:

$$\{Y_i(1), Y_i(0), M_i(1), M_i(0)\} \perp\!\!\!\perp PL_i \mid X_i.$$

- *Positivity:* Each treatment level occurs with positive probability given the covariates:

$$0 < P(PL_i = 1 \mid X_i) < 1.$$

Numerical Illustration

Assume a binary treatment scenario with hypothetical means calibrated from agricultural data:

$$\mathbb{E}[Y_i(1)] = 6.8 \text{ ¥/kg}, \quad \mathbb{E}[Y_i(0)] = 6.1 \text{ ¥/kg}.$$

Then:

$$ATE = 6.8 - 6.1 = 0.7 \text{ ¥/kg}.$$

Further decomposition yields:

$$NDE = 0.25 \text{ ¥/kg}, \quad NIE = 0.45 \text{ ¥/kg}.$$

Thus, about 64% of the total effect is mediated through improvements in *sweetness, size, and shelf life*, while 36% reflects the direct effect attributable to origin-specific market advantages.

Interpretation

The PO framework explicitly distinguishes *potential outcomes* from *observed outcomes*, enabling formal reasoning about *counterfactuals*—for example, the purchasing price an apple batch would have had if grown in a different origin.
This causal decomposition clarifies that:

- The *direct effect* captures regional or brand-related advantages associated with favorable origins;
- The *indirect effect* captures quality improvements—greater sweetness, larger size, and longer shelf life—induced by better cultivation conditions.

Overall, even under identical market conditions, transitioning to a favorable origin raises average apple prices by approximately 0.7 ¥/kg, primarily through mediator-driven quality effects.

21.8 Limitations

The PO framework provides a rigorous foundation for defining and estimating causal effects. However, several limitations should be acknowledged.

First, the framework relies on key identification assumptions, such as *ignorability* (no unmeasured confounding), *positivity*, and the *Stable Unit Treatment Value Assumption (SUTVA)*, which are often untestable in practice. Violations of these assumptions, for instance, due to hidden confounders or interference between units, can lead to biased causal estimates.

Second, PO-based analyses are typically data-intensive: reliable estimation of counterfactual quantities requires large and well-balanced samples, particularly when the treatment or mediator is continuous or high-dimensional.

Third, while the framework conceptually separates treatment, outcome, and covariates, it cannot on its own determine the correct causal structure; misspecifying covariates or mediators may distort the estimated causal effects. Furthermore, estimation of indirect (mediated) effects introduces additional assumptions, such as sequential ignorability, that are even harder to verify empirically.

Finally, although the PO framework offers a precise language for causal reasoning, it does not itself provide causal identification without strong assumptions or experimental control. Therefore, causal interpretations under the PO framework should always be supported by substantive theory, domain knowledge, and sensitivity analyses.

21.9 Summary

The Potential Outcome framework is a fundamental approach for defining and estimating causal effects. It formalizes causal questions by comparing outcomes that would be observed under different treatment conditions, thereby reframing causal inference as

a missing-data problem. The Potential Outcome approach is especially powerful because it connects causal reasoning to explicit assumptions, such as consistency, SUTVA, and ignorability, that are required to recover unobserved counterfactuals from available data. When applied carefully, it enables researchers to quantify causal effects, evaluate treatment strategies, and understand heterogeneity across populations. Nevertheless, rigorous causal inference within this framework depends critically on study design, assumption validity, and the quality of observed covariates. Violations such as unmeasured confounding, interference, or selection bias can compromise the interpretation of estimated causal effects. Thus, while the Potential Outcome framework provides a robust theoretical foundation for causal analysis, its practical success requires disciplined implementation and careful attention to methodological detail.

Chapter 22

Instrumental Variables

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the instrumental variables (in short, IV).

22.1 Basic Concepts

An Outcome Variable (Y) is the dependent variable whose causal relationship with the endogenous variable X under investigation [217].

An Endogenous Variable (X) is the independent variable presumed to have a causal impact on an outcome variable Y [217].

Ordinary Regression is a statistical method that fits a linear relationship between the explanatory variable and the outcome variable by minimizing the sum of squared differences between observed and predicted values of the outcome variable [53]. Ordinary regression is typically estimated using ordinary least squares. In IV, the endogenous variable X is the explanatory variable, and Y is the outcome variable.

Endogeneity refers to the situation where an explanatory variable in a statistical model is correlated with the error term [217]. For instance, endogenous explanatory variable X may be correlated with unobserved confounders, causing endogeneity [141].

An IV (Z) serves to isolate the component of variation in the endogenous variable X that is exogenous to the outcome variable Y, and the IV must also be independent of any unobserved confounders [141].

22.2 Problem Statement

IV is a statistical method commonly used in econometrics and data analysis to estimate causal relationships, particularly when conducting controlled experiments is not feasible [217, 149, 166, 141].

The IV problem could be stated as "how to address the *endogeneity* in estimating the causal influence of variable X on variable Y, formally expressed as $Y = \alpha + \beta X + \epsilon$."

[1] The primary contributor is Dr. Lei Wang.

Endogeneity arises when X is correlated with the error term ϵ, leading to biased and inconsistent estimates of β.

The input to the IV problem consists of observed data on the endogenous variable X, the outcome Y, and a proposed IV Z. A valid solution critically depends on whether Z satisfies three key constraints, detailed in Section 22.3. When these constraints are met, the output of the IV procedure is a consistent estimate of the causal parameter β.

22.3 Basic Assumptions

Assumption 22.1 (Relevance)**.** *An IV Z must be correlated with the endogenous variable X. That is, the IV must have a non-zero impact on X (the IV must shift X significantly).*

Assumption 22.2 (Exogeneity)**.** *An IVZ must be uncorrelated with the error term. This ensures that the variation in X induced by Z is exogenous and does not suffer from the same bias as the endogenous variable X.*

Assumption 22.3 (Exclusion Restriction)**.** *An IV Z must only influence the outcome Y through its influence on X and should have no direct influence on Y. If Z directly affects Y, then the estimate of the causal influence will be biased.*

22.4 Basic Principles

IV provides a powerful solution to the problem of *endogeneity* in causal inference. Endogeneity arises when an explanatory variable X is correlated with the error term in a regression model, leading to biased estimates of its causal effect on an outcome Y. A primary source of endogeneity is *confounding*—when unobserved variables U influence both X and Y, creating a spurious association. The IV method addresses this by introducing an instrument Z that serves as an exogenous source of variation. As shown in Figure 22.1, the core idea is to use a IV Z that is correlated with the endogenous variable X but independent of the confounders U, and that affects the outcome Y only through X.

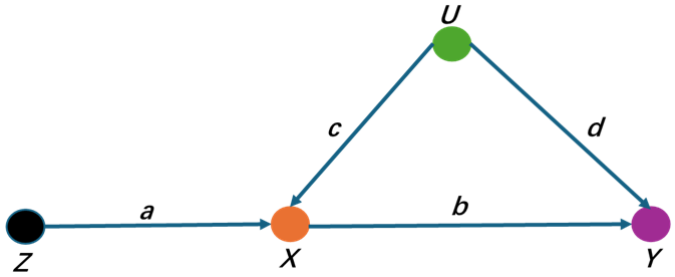

Figure 22.1: The IV framework: An IV Z addresses endogeneity by providing exogenous variation in X. Z is independent of confounders U [141].

The most common implementation of IV is *two-stage least squares*. This procedure involves two stages:

- *First stage*: The endogenous variable X is regressed on the IV Z (and any other exogenous controls) to obtain its predicted values, $\hat{X}$.
- *Second stage*: The outcome Y is regressed on the predicted values $\hat{X}$. This step isolates the component of X that is driven by the exogenous IV, thereby purging the bias caused by confounders U.

While confounding is a major application, IV methods are also essential for addressing other forms of endogeneity, including:

- *Measurement error*: When X is measured with error, leading to attenuated estimates.
- *Simultaneity (reverse causality)*: When X and Y mutually influence each other.
- *Omitted variables*: A general case encompassing confounding, when any unobserved factor affects both X and Y.

Historical applications—from John Snow's investigation of cholera transmission to Philip Wright's analysis of supply curves—demonstrate how well-chosen IV can uncover causal relationships [141]. In summary, the IV framework provides a rigorous approach for causal estimation in non-experimental settings. However, its validity critically depends on satisfying the core assumptions of IV relevance, exogeneity, and the exclusion restriction. Violations of these assumptions can lead to misleading conclusions, necessitating careful design and testing.

22.5 Methodology

The basic model of IV is a regression model. Suppose we have a regression model where Y is the dependent variable, X is the endogenous explanatory variable, and Z is IV. The basic model we wish to estimate is:

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{22.1}$$

- Y: The outcome variable.
- X: The endogenous explanatory variable.
- β_0: Intercept term, representing the expected value of Y when $X = 0$.
- β_1: The coefficient of X, representing the impact of a one-unit change in X on Y.
- ϵ: Error term, representing all factors or random errors not explained by X.

However, due to endogeneity (i.e., X is correlated with the error term ϵ), we cannot directly estimate β_1. To address this issue, we introduce an IV Z, which is correlated with X but uncorrelated with ϵ. The standard estimation procedure of IV is *two-stage*.

Stage 1: Regressing the Endogenous Explanatory Variable In the first stage, we predict X using Z, i.e., the regression:

$$X = \pi_0 + \pi_1 Z + u. \tag{22.2}$$

- X: The endogenous explanatory variable.
- Z: IV.
- π_0: Intercept term, representing the expected value of X when $Z = 0$.
- π_1: The coefficient of IV Z, representing the impact of a one-unit change in Z on X.
- u: Error term, representing the part of X not explained by Z.

The goal of this regression is to explain the variation in X through IV Z.

Stage 2: Using the Predicted Values for Regression In the second stage, we replace X with the predicted value $\hat{X}$ obtained from the first stage regression (which is a linear combination of Z), and then estimate the relationship between Y and $\hat{X}$:

$$Y = \beta_0 + \beta_1 \hat{X} + \nu. \tag{22.3}$$

- Y: The outcome variable.
- $\hat{X}$: The predicted value of X obtained from the first-stage regression, which is a linear combination of Z.
- β_0: Intercept term, representing the expected value of Y when $\hat{X} = 0$.
- β_1: New regression coefficient, representing the impact of the predicted $\hat{X}$ on Y in the second stage regression.
- ν: New error term, representing the random error component in the second-stage regression, which is uncorrelated with ϵ.

Since $\hat{X}$ is predicted using Z and Z is not correlated with ϵ, this second stage regression provides a consistent estimate of β_1.

22.6 Example: Infer the Effects of Apple Origins on Purchasing Prices

This example illustrates the application of the IV method to address endogeneity arising from an *unobserved confounder*. We aim to estimate the causal effect of *the apple origin* on the purchase price. The primary source of endogeneity is *unobserved consumer perception* (e.g., the belief that local apples are fresher or higher quality). This perception influences both the likelihood of an apple being marketed as local and the price consumers are willing to pay. Since we cannot measure this perception directly, we use an IV to isolate exogenous variation in origin.

Variables

- Y: purchase price (per kilogram, in ¥).
- X: Apple origin (binary variable, 0 = other regions, 1 = local origin). This is the *endogenous* variable.
- Z: Agricultural policy for subsidized farming (binary variable, 0 = no policy, 1 = policy in place). This is an IV.
- U: *Unobserved* consumer perception of local produce quality.

Hypothetical Data

We collected data from 1,000 apple shipments.

- Y: Apple purchase price ranges from 5.00 to 15.00 ¥/kg.
- X: Apple origin is 0 (Other regions) or 1 (Local).
- Z: Agricultural policy is 0 (No policy) or 1 (Policy in place).

Assume the following sample data: - 500 samples from the apples from the other regions ($X = 0$) - 500 samples from the locally grown apples ($X = 1$) - 60% of the samples (600 shipments) are from the regions subject to the agricultural policy ($Z = 1$)

Problem Setup: Unobserved Confounding

We wish to estimate the causal model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where the error term ϵ contains the unobserved consumer perception U. This creates endogeneity because U influences both X (Apple origin) and Y (Purchasing price), leading to a correlation between X and ϵ. For example, regions with strong positive perception (U) will have more local apples (X) and higher prices (Y), creating a

spurious correlation. An ordinary least squares regression of Y on X would yield a biased estimate of β_1.

An IV Solution

We introduce an IV Z, an agricultural subsidy policy (the government program that reduces the cost of local agricultural production to encourage local farming), to resolve the endogeneity. The validity of Z rests on two assumptions:

- *Relevance:* The policy Z directly influences the apple origin X. Subsidies lower the cost of local production, making it more likely for apples to be grown and sold locally ($X = 1$). Hence, Z is correlated with X.
- *Exclusion Restriction:* The policy Z affects the purchasing price Y *only* through its effect on the supply of local apples X. The policy itself is unrelated to unobserved consumer perception U, and does not directly affect market prices other than by shifting the origin composition.

IV provides a source of exogenous variation in X that is independent of the confounder U.

Two-Stage Least Squares Implementation

Stage 1: Isolating Exogenous Variation in the Apple Origins
In the first stage, we regress the endogenous variable X on the IV Z. This isolates the variation in the apple origin that is driven solely by the exogenous policy:

$$X_i = \pi_0 + \pi_1 Z_i + u_i,$$

where X_i indicates the local origin (1) or not (0), and Z_i indicates the presence of the policy (1) or not (0). The regression is estimated using ordinary least squares, producing the fitted values $\hat{X}_i$, which represent the component of the apple origin that is driven by the policy and is therefore exogenous to the unobserved confounders in ϵ.
Stage 2: Estimating the Causal Effect on Purchasing Price
In the second stage of the two-stage least squares procedure, we regress the purchasing price Y_i on the predicted (exogenous) values of the apple origin $\hat{X}_i$ from the first stage:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \nu_i.$$

This stage yields a consistent estimate of the causal parameter β_1, as it uses only the exogenous variation in X induced by the IV Z.
Illustrative Numerical Results
Assume the first-stage ordinary least squares regression yields:

$$X_i = 0.20 + 0.65 Z_i + u_i$$

This suggests the policy increases the probability of the local apple origin by 65 percentage points.
Substituting $\hat{X}_i$ into the second-stage regression gives the two-stage least squares estimate:

$$Y_i = 8.50 + 2.10\hat{X}_i + \nu_i.$$

This implies that the causal effect of an apple being locally grown is to increase its price by an average of *2.10 ¥/kg*.

Conclusion and Interpretation

The two-stage least squares procedure provides a consistent estimate of the causal impact of the apple origins on the purchase prices. The first stage isolates exogenous variation in the apple origin due to the agricultural policy, while the second stage corrects for endogeneity and yields consistent parameter estimates. Under the assumptions of instrument relevance and exclusion, the coefficient $\hat{\beta}_1 = 2.10$ indicates that *locally grown apples are priced approximately 2.10 ¥ higher per kilogram* than non-local apples.
However, it is important to note that this estimate represents the Local Average Treatment Effect (LATE), which applies specifically to the subset of apples whose origin is influenced by the agricultural policy. This means that the estimated effect reflects the price difference between locally grown and non-local apples only in regions where the policy has altered the origin of apples. In other words, this result is not a general estimate for all local apples but specifically for those whose production decisions are affected by the policy. Thus, the estimate $\hat{\beta}_1 = 2.10$ reflects the causal effect of being locally grown on price for apples in areas where the policy has induced changes in production decisions.

The above example demonstrates how IV can be used to address endogeneity issues and estimate the causal impact of the apple origin on the purchase price, while also verifying the indirect impact of an agricultural policy on purchasing prices. Through this method, we can draw more reliable conclusions about the impact of an agricultural policy on market outcomes, such as apple pricing.

22.7 Limitations

Despite their power, IV methods have important limitations. First, finding a valid IV is difficult because few variables in practice simultaneously satisfy relevance, exclusion, and exogeneity. Second, a weak IV, which is only weakly correlated with the variable of interest (X), can lead to biased estimates and large standard errors. Third, IV estimates typically identify a local average treatment effect (LATE) for the subpopulation affected by IV, limiting the generalizability of the results. Finally, IV estimates are sensitive to model specification, and incorrect models or inappropriate covariates can compromise the validity of the estimates.

22.8 Summary

Instrumental Variables (IV) methods are widely used in econometrics to address endogeneity and estimate causal relationships when controlled experiments are not feasible. By using an instrument—a variable correlated with the endogenous explanatory variable but not with the error term—IV methods help isolate the causal effect of the explanatory variable on the outcome. However, IV methods have limitations, including challenges in finding valid instruments, sensitivity to model specification, and the potential for weak instruments to lead to biased estimates.