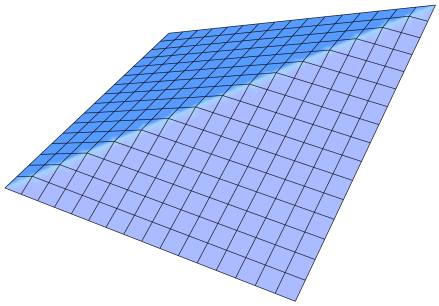


STAT 332: Sampling and Experimental Design

Prof. Ryan P. Browne • Winter 2016 • University of Waterloo
TR 10:00–11:20AM in AL 113

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

1. Survey Sampling Foundations

This topic develops the populations, parameters, sampling frames, and probability mechanisms that underlie survey inference. A recurring distinction is the difference between a number computed from the observed sample and the random quantity that would result from repeating the

sampling procedure. The former is an estimate; the latter is an estimator.

Methods of Data Collection

In an observational study, data are obtained through reading, counting, or measurement under normal working conditions. In an experimental study, normal working conditions are deliberately changed so that the resulting response can be studied.

Survey sampling studies finite populations through observational methods. Experimental design instead studies how deliberate changes to factors affect a response, with the collection of possible experimental outcomes viewed as conceptually infinite. In both settings, the mechanism that produces the data is part of the statistical argument: a probability sampling design justifies survey inference, while random treatment assignment justifies causal comparisons.

Survey Sampling Definitions

An **observational unit**, or **element**, is an object or individual on which we take a measurement. The **target population** is the collection of observational units that we want to study. We usually write this finite population as

$$U = \{1, 2, 3, \ldots, N\}.$$

Example 0.1 Examples of target populations include the students in STAT 332, the population of Canada, and a collection of tax files.

The **sampled population** is the collection of observational units that could be in the sample. Equivalently, it is the population from which the sample is actually taken. The **sampling frame** is the list of sampling units available to the sampling procedure.

Example 0.2 For a telephone survey, the sampling frame may be a list of households. For an educational survey, the sampling frame may be a list of schools.

The **sampling unit** is the unit that is actually selected, such as a household or a school. A sampling unit need not be the same as an observational unit: a household may be sampled even if measurements are taken on the people within that household.

Population Parameters

Let the complete list of individuals be

$$U = \{1, 2, 3, \ldots, N\},$$

where N is the population size. The study variable, or response of interest, is denoted by y. For example, y may represent height or income. The value y_i is the yearly income for individual i, giving the finite set $\{y_1, \ldots, y_N\}$.

Definition 0.3 The **population total** is

$$\tau = \sum_{i=1}^{N} y_i.$$

The **population average** is

$$\mu = \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{\tau}{N}.$$

The **population variance** is

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2.$$

Since $\bar{Y} = \mu$, the population variance has the working form

$$\begin{aligned}
\sigma^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2 \\
&= \frac{1}{N-1} \sum_{i=1}^{N} (y_i^2 - 2y_i\mu + \mu^2) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - 2\mu \sum_{i=1}^{N} y_i + N\mu^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - 2N\mu^2 + N\mu^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 - N\mu^2 \right).
\end{aligned}$$

Population Proportion

Population proportions can be treated as averages of indicator variables. For example, suppose that "low income" means $y \leq 17000$.

Definition 0.4 Let m be the number of indices i such that $y_i \leq 17000$. The **population proportion** of low-income individuals is

$$p = \frac{m}{N}.$$

For each $i = 1, \ldots, N$, define the indicator

$$\mathbb{1}_i = \begin{cases} 1, & \text{if } y_i \leq 17000, \\ 0, & \text{otherwise.} \end{cases}$$

Because $\mathbb{1}_i^2 = \mathbb{1}_i$ for an indicator, the corresponding population mean and variance are

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{1}_i &= m, \\
\mu_{\mathbb{1}} &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_i = \frac{m}{N} = p, \\
\sigma_{\mathbb{1}}^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (\mathbb{1}_i - \mu_{\mathbb{1}})^2 \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} \mathbb{1}_i^2 - N\mu_{\mathbb{1}}^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} \mathbb{1}_i - N\mu_{\mathbb{1}}^2 \right) \\
&= \frac{1}{N-1} (N\mu_{\mathbb{1}} - N\mu_{\mathbb{1}}^2) \\
&= \frac{N}{N-1} (p - p^2) \\
&= \frac{N}{N-1} p(1-p) \approx p(1-p) \quad \text{if } N \text{ is large.}
\end{aligned}$$

There are several practical reasons to sample instead of taking a census, which studies every observational unit. Sampling is usually less costly and can be completed more quickly. Moreover,

concentrating data collection effort on fewer units can make a carefully planned survey more accurate and reliable than a poorly executed census.

The collections involved in a survey need not coincide. Figure 1 distinguishes the target population and sampling frame from the sampled population. It also identifies nonrespondents and ineligible frame entries. Ideally, the frame covers the target population exactly; in practice, undercoverage, ineligible entries, and nonresponse prevent this.

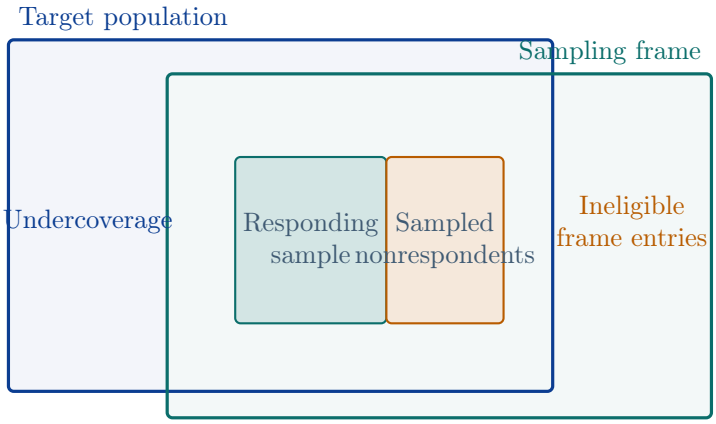


FIGURE 1

Population Variance and Auxiliary Information

The finite population variance in Definition 0.3 measures the spread of the response values in the population:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2.$$

Here μ is the population average,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

The quantity $\hat{\mu}$ denotes a sample estimate of the population average, while $\tilde{\mu}$ denotes the corresponding estimator.

For simple random sampling without replacement, the sampling design itself supplies the variance formula

$$\operatorname{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sigma^2.$$

This is a design based statement: the population values are regarded as fixed, and randomness enters only through the choice of sample. Ratio and regression estimators are model assisted

because a relationship with an auxiliary variable suggests how to use known population information more efficiently, while repeated sampling properties are still evaluated under the sampling design.

The ratio estimator uses auxiliary information about each element. Suppose the response is a block's weight and the auxiliary variable is its perimeter. Then μ_y is the target population average weight, μ_x is the known target population average perimeter, $\bar{y}$ is the sample average of the weights, and $\bar{x}$ is the sample average of the perimeters. Assuming $\mu_x \neq 0$ and that the realized sample has $\bar{x} \neq 0$, the basic ratio approximation is

$$\frac{\mu_y}{\mu_x} \approx \frac{\bar{y}}{\bar{x}}.$$

Thus

$$\hat{\mu}_{\text{ratio}} = \frac{\bar{y}}{\bar{x}} \cdot \mu_x = \left(\frac{\mu_x}{\bar{x}}\right) \cdot \bar{y}.$$

The regression estimator instead models the response through an auxiliary variable. A sample-centered parametrization is

$$y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i.$$

Figure 2 shows why auxiliary information can improve estimation. A fitted relationship between perimeter and weight permits the known population average perimeter to inform the estimate of the population average weight.

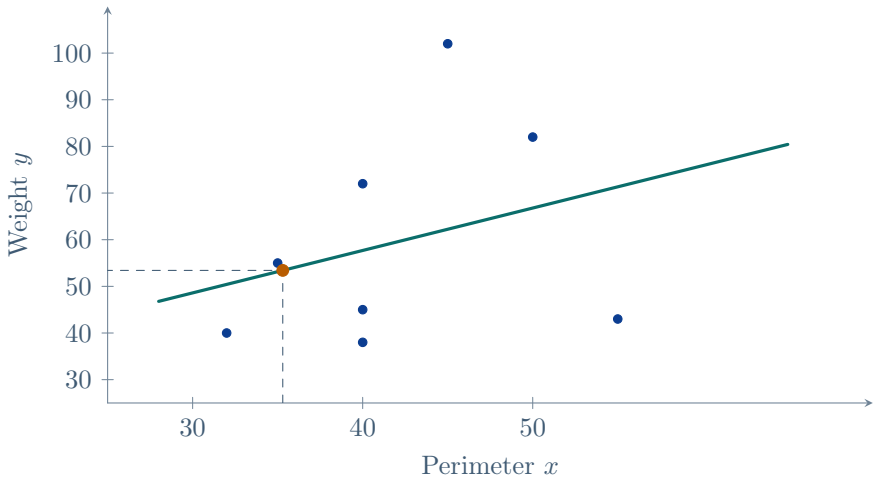


FIGURE 2

For the plotted sample, $\bar{x} = 42.125$, $\bar{y} = 59.625$, and $\hat{\beta} \approx 0.9094$. The known auxiliary mean and the corresponding regression estimate are therefore

$$\mu_x = 35.3 \qquad \text{and} \qquad \hat{\mu}_{\text{reg}} = \bar{y} + \hat{\beta}(\mu_x - \bar{x}) \approx 53.42.$$

The least squares coefficients used in this adjustment are

$$\hat{\alpha} = \bar{y} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in s}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

Sample Definitions

Let $U = \{1, \ldots, N\}$ be the finite study population. A sample is a subset of the study population, denoted by $s \subset U$, and its size is denoted by $n = |s|$. The sample data consist of the observed response values y_i for the elements $i \in s$. If auxiliary information is also observed, then the sample data are

$$\{(x_i, y_i) : i \in s\}.$$

Definition 0.5 The **sample average** is

$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i.$$

For $n \geq 2$, the **sample variance** is

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2.$$

The sample variance has the working form

$$s^2 = \frac{1}{n-1} \left[\sum_{i \in s} y_i^2 - n\bar{y}^2 \right].$$

The sampling design is the probability distribution on the collection of possible samples. Two probabilities extracted from this distribution will be used repeatedly.

Definition 0.6 The quantity $\mathbb{P}(s)$ is the probability that the sample s is selected. The **inclusion probability** of element i is

$$\pi_i = \mathbb{P}(i \in s),$$

and the **joint inclusion probability** of distinct elements i and j is

$$\pi_{ij} = \mathbb{P}(i \in s \text{ and } j \in s).$$

Example 0.7 Let $N = 3$ and $U = \{1, 2, 3\}$. For the sampling design below, compute the inclusion probability of element 1, the joint inclusion probability of elements 1 and 2, the sampling distribution of the sample average, and the expected value of the corresponding estimator.

The possible nonempty candidate samples are

$$s_1 = \{1\}, \quad s_2 = \{2\}, \quad s_3 = \{3\}, \quad s_4 = \{1, 2\},$$
$$s_5 = \{1, 3\}, \quad s_6 = \{2, 3\}, \quad \text{and} \quad s_7 = \{1, 2, 3\}, \text{ the census.}$$

One possible sampling design, viewed as a probability distribution on this sample space, is

s	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$\mathbb{P}(s)$	0	0	0	1/4	1/4	1/4	1/4

The inclusion probability for element 1 is

$$\pi_1 = \mathbb{P}(1 \in s) = \frac{1}{4} + \frac{1}{4} + 0 + \frac{1}{4} = \frac{3}{4}.$$

The joint inclusion probability for elements 1 and 2 is

$$\pi_{12} = \mathbb{P}(1 \in s \text{ and } 2 \in s) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

The sampling distribution of the sample average is as follows.

s	s_4	s_5	s_6	s_7
$\bar{y}$	$(y_1 + y_2)/2$	$(y_1 + y_3)/2$	$(y_2 + y_3)/2$	$(y_1 + y_2 + y_3)/3$
$\mathbb{P}(\bar{y})$	1/4	1/4	1/4	1/4

The sample average is therefore a discrete random variable. If the design instead assigns

equal probability to all samples of a fixed size $n = 2$, it becomes simple random sampling without replacement.

To compute the expected value of the sample average, recall that a discrete random variable X satisfies

$$\mathbb{E}[X] = \sum_{x \in S_X} x \cdot \mathbb{P}(X = x).$$

The estimate is $\hat{\mu} = \bar{y}$, and the corresponding estimator is $\tilde{\mu}$. Its expectation is

$$\begin{aligned}
\mathbb{E}[\tilde{\mu}] &= \sum_{i=1}^{7} \bar{y}(s_i)\mathbb{P}(s = s_i) \\
&= \sum_{i=4}^{7} \bar{y}(s_i)\frac{1}{4} \\
&= \frac{1}{4}\left[\frac{y_1 + y_2}{2} + \frac{y_1 + y_3}{2} + \frac{y_2 + y_3}{2} + \frac{y_1 + y_2 + y_3}{3}\right] \\
&= \frac{1}{3}[y_1 + y_2 + y_3] \\
&= \mu.
\end{aligned}$$

Thus the sample average is unbiased under this design.

For comparison, another valid sampling design is given by the following probabilities.

s	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$\mathbb{P}(s)$	2/5	0	0	0	2/5	1/5	0

Under this second design, $\pi_1 = 4/5$, $\pi_2 = 1/5$, and $\pi_3 = 3/5$. The sample average is no longer unbiased in general, since

$$\mathbb{E}[\tilde{\mu}] = \frac{3}{5}y_1 + \frac{1}{10}y_2 + \frac{3}{10}y_3,$$

which need not equal $(y_1+y_2+y_3)/3$. Unequal inclusion probabilities therefore require an estimator that accounts for the sampling design.

Proposition 0.8 Suppose that every population unit has a positive inclusion probability.

The **Horvitz–Thompson estimator** of the population total,

$$\tilde{\tau}_{\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i},$$

is unbiased under the design: $\mathbb{E}[\tilde{\tau}_{\mathrm{HT}}] = \tau$.

Proof. Let $\mathbb{1}_i$ indicate whether unit i is included in the sample. Since $\mathbb{E}[\mathbb{1}_i] = \pi_i$,

$$\mathbb{E}[\tilde{\tau}_{\mathrm{HT}}] = \mathbb{E}\left[\sum_{i=1}^{N} \frac{\mathbb{1}_i y_i}{\pi_i}\right] = \sum_{i=1}^{N} \frac{y_i}{\pi_i}\mathbb{E}[\mathbb{1}_i] = \sum_{i=1}^{N} y_i = \tau.$$

□

This result explains the central role of inclusion probabilities. When all $\pi_i = n/N$, dividing the Horvitz–Thompson total by N recovers the ordinary sample average.

2. Simple, Systematic, and Cluster Sampling

Probability samples can be drawn in several ways, and each design produces its own estimator variance. This topic begins with simple random sampling, then compares sampling with replacement, systematic sampling, and one stage cluster sampling.

Simple Random Sampling Without Replacement

In simple random sampling without replacement, the sampling frame is the simple list $U = \{1, 2, \ldots, N\}$, and the sample size is fixed at n. The scheme draws individuals from the population without replacement, with every subset of size n equally likely. As a probability measure on possible samples,

$$\mathbb{P}(s) = \begin{cases} 1/\binom{N}{n}, & \text{if } s \text{ has size } n, \\ 0, & \text{otherwise.} \end{cases}$$

The inclusion probabilities can be computed by counting the number of samples of size n that contain the required elements. Since

$$\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1},$$

we have

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\frac{N}{n}\binom{N-1}{n-1}} = \frac{n}{N}.$$

Similarly, for $i \neq j$,

$$\pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}.$$

Our estimate for the population average is the sample average, $\hat{\mu} = \bar{y}$.

Its corresponding estimator is $\tilde{\mu}$. The estimator $\tilde{\mu}$ is the random variable obtained by applying the sample average to a randomly selected sample.

Proposition 0.9 Under simple random sampling without replacement,

$$\mathbb{E}[\tilde{\mu}] = \mu \qquad \text{and} \qquad \text{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}.$$

Thus $\tilde{\mu}$ is unbiased for the population average. The quantity σ^2 is the population variance:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2.$$

Indicator method. Let

$$\mathbb{1}_i = \begin{cases} 1, & \text{if individual } i \text{ is in the sample,} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\mathbb{E}\left[\mathbb{1}_i\right] = 0 \cdot \mathbb{P}(\mathbb{1}_i = 0) + 1 \cdot \mathbb{P}(\mathbb{1}_i = 1) = \mathbb{P}(\mathbb{1}_i = 1) = \pi_i$$

and

$$\text{Var}(\mathbb{1}_i) = \mathbb{E}\left[\mathbb{1}_i^2\right] - \mathbb{E}\left[\mathbb{1}_i\right]^2 = \underbrace{\mathbb{E}\left[\mathbb{1}_i\right]}_{\pi_i} - \pi_i^2 = \pi_i(1 - \pi_i).$$

Because the sample contains exactly n units, $\mathbb{1}_1 + \mathbb{1}_2 + \cdots + \mathbb{1}_N = n$. Moreover,

$$\tilde{\mu} = \frac{1}{n}\sum_{i\in s} y_i = \frac{1}{n}\sum_{i=1}^{N} \underbrace{\mathbb{1}_i}_{\text{random}} \underbrace{y_i}_{\text{nonrandom}}.$$

Therefore,

$$\mathbb{E}[\tilde{\mu}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N}\mathbb{1}_i y_i\right] = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{E}[\mathbb{1}_i] = \frac{1}{\cancel{n}}\sum_{i=1}^{N} y_i \frac{\cancel{n}}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu.$$

For the variance calculation, recall that a linear combination of dependent random variables $X_1, \ldots, X_N$ satisfies

$$\operatorname{Var}\left(\sum_{i=1}^{N} a_i X_i\right) = \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} a_i a_j \operatorname{Cov}(X_i, X_j).$$

Applying this identity gives

$$\operatorname{Var}(\tilde{\mu}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{N}\mathbb{1}_i y_i\right) \tag{0.1}$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \operatorname{Var}(\mathbb{1}_i) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_i y_j \operatorname{Cov}(\mathbb{1}_i, \mathbb{1}_j)\right]. \tag{0.2}$$

Using $\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ and $\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, we obtain

$$\operatorname{Var}(\mathbb{1}_i) = \pi_i(1-\pi_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

$$\operatorname{Cov}(\mathbb{1}_i, \mathbb{1}_j) = \mathbb{E}\left[\mathbb{1}_i \mathbb{1}_j\right] - \mathbb{E}\left[\mathbb{1}_i\right]\mathbb{E}\left[\mathbb{1}_j\right] = \pi_{ij} - \pi_i\pi_j = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n}{N}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}.$$

Substituting these expressions into (0.2) gives

$$\operatorname{Var}(\tilde{\mu}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \frac{n}{N}\left(1 - \frac{n}{N}\right) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_i y_j \left(-\frac{n}{N}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\right)\right]$$

$$= \frac{1}{n^2}\frac{n}{N}\left(1 - \frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N} y_i(N\mu - y_i)\right]$$

$$= \frac{1}{n}\frac{1}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}N\mu \cdot N\mu + \frac{1}{N-1}\sum_{i=1}^{N} y_i^2\right].$$

Here the double sum simplifies because

$$\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} y_i y_j = \sum_{i=1}^{N} y_i \sum_{\substack{j=1\\ j\neq i}}^{N} y_j = \sum_{i=1}^{N} y_i \left(\sum_{k=1}^{N} y_k - y_i\right) = \sum_{i=1}^{N} y_i(N\mu - y_i).$$

Indeed,

$$y_1 + \cdots + y_{i-1} + y_{i+1} + \cdots + y_N = \sum_{j=1}^{N} y_j - y_i.$$

Finally, $1 + 1/(N-1) = N/(N-1)$, so

$$\begin{aligned}
\mathrm{Var}(\tilde{\mu}) &= \frac{1}{n}\frac{1}{N}\left(1-\frac{n}{N}\right)\left[\frac{N}{N-1}\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}N^2\mu^2\right] \\
&= \frac{1}{n}\frac{\cancel{N}}{\cancel{N}}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N\mu^2\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\underbrace{\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2}_{\sigma^2} \\
&= \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}.
\end{aligned}$$

□

We have

$$\mathrm{Var}(\tilde{\mu}) = \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} = (1-f)\frac{\sigma^2}{n},$$

where $f = n/N$ is called the **sampling fraction** and $(1-f)$ is called the **finite population correction factor**.

Proposition 0.10 Under simple random sampling without replacement,

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2$$

is unbiased for the finite population variance σ^2.

Proof. The usual sum-of-squares identity gives

$$\sum_{i\in s}(y_i-\bar{y})^2=\sum_{i\in s}(y_i-\mu)^2-n(\bar{y}-\mu)^2.$$

Every population unit has inclusion probability n/N, so

$$\mathbb{E}\left[\sum_{i\in s}(y_i-\mu)^2\right]=\frac{n}{N}\sum_{i=1}^{N}(y_i-\mu)^2=\frac{n(N-1)}{N}\sigma^2.$$

By Proposition 0.9,

$$n\mathbb{E}[(\bar{y}-\mu)^2]=n\,\mathrm{Var}(\tilde{\mu})=\frac{N-n}{N}\sigma^2.$$

Subtracting the two expectations yields

$$\mathbb{E}\left[\sum_{i\in s}(y_i-\bar{y})^2\right]=(n-1)\sigma^2,$$

and division by $n-1$ completes the proof. □

Consequently, the observed sample variance $\hat{\sigma}^2$ yields the unbiased plug-in estimate

$$\widehat{\mathrm{Var}}(\tilde{\mu})=\left(1-\frac{n}{N}\right)\frac{\hat{\sigma}^2}{n}$$

of $\mathrm{Var}(\tilde{\mu})$.

Definition 0.11 The **standard error** of $\hat{\mu}$ is

$$\text{standard error}(\hat{\mu})=\sqrt{\widehat{\mathrm{Var}}(\tilde{\mu})}=\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{\sigma}^2}{n}}.$$

If N, n, and $N-n$ are large, and no small collection of population units dominates the total variation, then a finite population central limit theorem gives

$$\frac{\tilde{\mu}-\mu}{\sqrt{(1-f)\dfrac{\tilde{\sigma}^2}{n}}}\overset{\text{approximately}}{\sim}\mathcal{N}(0,1).$$

A large sample $100(1-\alpha)\%$ confidence interval for the population average μ is

$$\hat{\mu} \pm c \cdot \text{standard error}(\hat{\mu}).$$

For a 95% confidence interval, we use $c = 1.96$. This value is used because, if $Z \sim \mathcal{N}(0,1)$, then

$$\mathbb{P}(|Z| \leq 1.96) = 0.95 \qquad \text{and} \qquad \mathbb{P}(|Z| > 1.96) = 0.05.$$

For a $100(1-\alpha)\%$ confidence interval, we need a c such that

$$\mathbb{P}(|Z| \leq c) = 1 - \alpha \qquad \text{and} \qquad \mathbb{P}(|Z| > c) = \alpha.$$

A 90% confidence procedure has the following repeated-sampling interpretation. If we were to repeat a survey 100 times and construct a 90% confidence interval for μ from each sample, then we would expect about 90 of the intervals to cover μ and about 10 not to cover it. Figure 3 illustrates this distinction: the blue intervals cover the fixed population mean, whereas the orange intervals miss it.

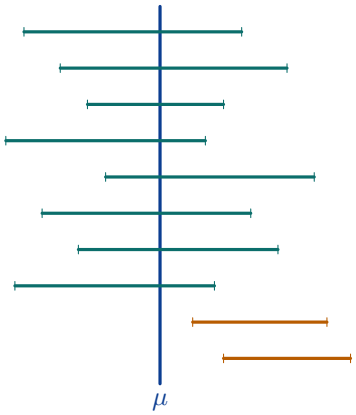


FIGURE 3

For the population total, the true value, point estimate, and estimator are respectively

$$\tau = N\mu, \qquad \hat{\tau} = N\hat{\mu} \qquad \text{and} \qquad \tilde{\tau} = N\tilde{\mu}.$$

It follows from Proposition 0.9 that

$$\mathbb{E}[\tilde{\tau}] = N\mathbb{E}[\tilde{\mu}] = \tau \qquad \text{and} \qquad \text{Var}(\tilde{\tau}) = N^2 \,\text{Var}(\tilde{\mu}).$$

The standard error of the estimator for the population total is

$$\text{standard error}(\hat{\tau}) = N \cdot \text{standard error}(\hat{\mu}).$$

For simple random sampling without replacement,

$$\text{standard error}(\hat{\tau}) = N\sqrt{(1-f)\frac{\hat{\sigma}^2}{n}}.$$

A $100(1-\alpha)\%$ large sample confidence interval for τ may therefore be written as

$$\hat{\tau} \pm c \text{ standard error}(\hat{\tau}) = N\hat{\mu} \pm cN \text{ standard error}(\hat{\mu}).$$

Sample Size

We now ask how large the sample should be.

We can state the precision in terms of the length of a confidence interval. If the desired length is 2ℓ, then

$$\text{the confidence interval for } \mu \text{ will be } \pm\ell.$$

For simple random sampling without replacement, we have

$$\ell = c \cdot \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}.$$

Solving for n gives

$$n = \frac{1}{\frac{1}{N} + \frac{\ell^2}{c^2\sigma^2}} \overset{N \text{ is large}}{\approx} \frac{c^2\sigma^2}{\ell^2}.$$

The calculated value must be rounded up so that the requested precision is attained.

The sample size calculation therefore requires the following quantities.

1. Specify the desired margin of error ℓ, so the interval has length 2ℓ.
2. Specify the confidence level $1-\alpha$, which determines the critical value c.
3. Supply an estimate of the population variance, possibly from a pilot survey.

Simple Random Sampling with Replacement

For simple random sampling with replacement, the sampling frame is $U = \{1, \ldots, N\}$, the sample size is fixed at n, and the population units are drawn independently with replacement.

When N is large, simple random sampling without replacement and simple random sampling with replacement are approximately equivalent.

To estimate the population average μ, we will use the sample average $\hat{\mu} = \bar{y}$.

Example 0.12 Let $N = 3$, $n = 2$, and $U = \{1, 2, 3\}$, with response values $\{y_1, y_2, y_3\}$. The six distinct unordered samples that can arise with replacement are shown in Figure 4. The three samples with unequal elements are also precisely the samples available without replacement.

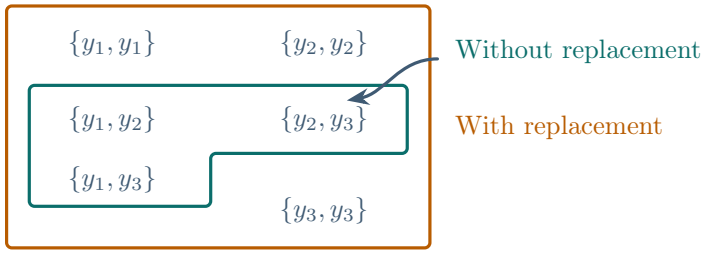


FIGURE 4

Although Figure 4 lists unordered samples, the sampling mechanism consists of two ordered independent draws. Thus a sample with a repeated value has probability 1/9, whereas a sample with two distinct values has probability 2/9, because the distinct values can be drawn in either order.

Proposition 0.13 Under simple random sampling with replacement,

$$\mathbb{E}[\tilde{\mu}] = \mu \qquad \text{and} \qquad \operatorname{Var}(\tilde{\mu}) = \left(1 - \frac{1}{N}\right) \frac{\sigma^2}{n}.$$

Here

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2.$$

Independent draw method. Let I be uniform on $\{1, \ldots, N\}$, and define $Z = y_I$. Its probability mass function is

$$f_Z(z) = \mathbb{P}(Z = z) = \frac{1}{N}|\{i : y_i = z\}|.$$

Thus repeated population response values automatically contribute their combined probability. Direct calculation gives

$$\mathbb{E}[Z] = \sum_{i=1}^{N} y_i \frac{1}{N} = \mu,$$

$$\mathrm{Var}(Z) = \mathbb{E}[(Z-\mu)^2] = \sum_{i=1}^{N} (y_i - \mu)^2 \frac{1}{N} = \frac{N-1}{N}\sigma^2.$$

Let $Z_1, \ldots, Z_n$ be independent random variables with this common distribution. The estimator for sampling with replacement is

$$\tilde{\mu}_{\mathrm{wr}} = \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

Its expectation is

$$\mathbb{E}[\tilde{\mu}_{\mathrm{wr}}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Z_i] = \mu.$$

Since the draws are independent,

$$\mathrm{Var}(\tilde{\mu}_{\mathrm{wr}}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Z_i) = \frac{1}{n}\mathrm{Var}(Z) = \left(1 - \frac{1}{N}\right)\frac{\sigma^2}{n}.$$

□

For $n \geq 2$, the ordinary sample variance estimator for the draws,

$$\tilde{\sigma}^2_{\mathrm{draw}} = \frac{1}{n-1}\sum_{j=1}^{n} (Z_j - \bar{Z})^2,$$

is unbiased for $\mathrm{Var}(Z) = (1 - 1/N)\sigma^2$, not for the finite population variance σ^2. Consequently, its realized value gives

$$\widehat{\mathrm{Var}}(\tilde{\mu}_{\mathrm{wr}}) = \frac{\hat{\sigma}^2_{\mathrm{draw}}}{n}.$$

For $n > 1$, simple random sampling without replacement is more efficient than sampling with replacement. Indeed, Propositions 0.9 and 0.13 give the respective variance factors $1 - n/N$ and

$1 - 1/N$, and $1 - n/N < 1 - 1/N$.

Inference for Means and Proportions

Under simple random sampling without replacement, μ is the population average, $\hat{\mu}$ is a point estimate for μ based on one sample, and $\tilde{\mu}$ is the random variable that represents $\hat{\mu}$ under repeated sampling. In this setting, $\hat{\mu} = \bar{y}$.

As a summary, Proposition 0.9 and the plug-in variance estimate give

$$\begin{aligned} \mathbb{E}[\tilde{\mu}] &= \mu, \\ \mathrm{Var}(\tilde{\mu}) &= \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}, \\ \widehat{\mathrm{Var}}(\tilde{\mu}) &= \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}^2}{n}, \end{aligned}$$

where $\hat{\sigma}^2$ is the sample variance. The population and sample variances are

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y})^2.$$

Using the central limit theorem for a finite population, a large sample 95% confidence interval is

$$\hat{\mu} \pm c \cdot \text{standard error}(\hat{\mu}).$$

Since the standard error is the square root of the estimated repeated-sampling variance, this interval can be written as

$$\hat{\mu} \pm c\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}^2}{n}}.$$

For a population proportion, encode the event of interest with the indicator

$$y_i = \mathbb{1}(\text{the event of interest holds for unit } i).$$

The population and sample proportions are then

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = p \qquad \text{and} \qquad \hat{\mu} = \hat{p} = \frac{1}{n}\sum_{i \in s} y_i.$$

The corresponding population and sample variances are

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2 = \frac{N}{N-1}p(1-p) \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\bar{y})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

Consequently,

$$\operatorname{Var}(\tilde{p}) = \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} \qquad \text{and} \qquad \widehat{\operatorname{Var}}(\tilde{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}.$$

A confidence interval for p based on the sample is therefore

$$\hat{p} \pm c\,\text{standard error}(\hat{p}) = \hat{p} \pm c\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}}.$$

Theorem 0.14 (Student's t-distribution) Let $n \geq 2$, and let $X_1, \ldots, X_n$ be independent and identically distributed with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, where μ is unknown and $\sigma^2 > 0$. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Then

$$\frac{\bar{X}-\mu}{\sqrt{\tilde{\sigma}^2/n}} \sim t_{n-1}.$$

Derivation. The standardized sample mean and scaled sample variance satisfy

$$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad W = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1},$$

and Z and W are independent. By the definition of the t-distribution,

$$\frac{Z}{\sqrt{W/(n-1)}} \sim t_{n-1}.$$

Substitution gives

$$\frac{Z}{\sqrt{W/(n-1)}} = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\Bigg/\sqrt{\frac{(n-1)\tilde{\sigma}^2}{\sigma^2}\cdot\frac{1}{n-1}} = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\sqrt{\frac{\sigma^2}{\tilde{\sigma}^2}} = \frac{\bar{X}-\mu}{\sqrt{\tilde{\sigma}^2/n}}.$$

□

As n grows, the t_{n-1} distribution approaches the standard normal distribution.

Remark 0.15 The exact Student t result assumes an independent random sample from a normal population. It is not an exact finite population result for simple random sampling without replacement. In survey sampling, the normal interval above is instead justified by a finite population central limit theorem and the estimated design variance.

Systematic Sampling

In **systematic sampling**, the sampling frame is the ordered list $U = \{1, \dots, N\}$. Assume that $N = nk$, and fix the sample size at n. The sampling interval is $k = N/n$. The scheme selects a random start r from $\{1, \dots, k\}$ with probability $1/k$, and then samples the units

$$r, r+k, r+2k, \dots, r+(n-1)k.$$

Arranging the ordered frame in n rows and k columns makes the design transparent:

$$\begin{bmatrix} 1 & 2 & 3 & \dots & k-1 & k \\ k+1 & k+2 & k+3 & & 2k-1 & 2k \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ nk-(k-1) & & & & nk-1 & nk \end{bmatrix}$$

Each column forms one possible sample, and the random start chooses one of the k columns. There are therefore k candidate samples $s_1, \dots, s_k$. The point estimate $\hat{\mu}$ for the population average is the sample average, whose sampling distribution is summarized by

$$\left[\begin{array}{c|ccccc} r & 1 & 2 & \dots & k \\ \mathbb{P}(s = s_r) & 1/k & & & 1/k \\ \bar{y} & \dfrac{\sum_{i=1}^{n} y_{ik-(k-1)}}{n} & & & \dfrac{\sum_{i=1}^{n} y_{ik}}{n} \end{array}\right]$$

The estimator is unbiased because

$$\begin{aligned} \mathbb{E}[\tilde{\mu}] &= \sum_{r=1}^{k} \bar{y}(s_r)\mathbb{P}(s = s_r) = \frac{1}{k}\sum_{r=1}^{k} \bar{y}(s_r) \\ &= \frac{1}{nk}\sum_{r=1}^{k}\sum_{i \in s_r} y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu. \end{aligned}$$

Its design variance is

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}) &= \mathbb{E}\left[(\bar{y}(s_r) - \mathbb{E}[\bar{y}(s_r)])^2\right] \\ &= \sum_{r=1}^{k} (\bar{y}(s_r) - \mu)^2 \mathbb{P}(s = s_r) \\ &= \frac{1}{k} \sum_{r=1}^{k} (\bar{y}(s_r) - \mu)^2.\end{aligned}$$

Systematic samples are easy to collect because only one random start is required. When the ordered responses follow a gradual trend, as in Figure 5, the regularly spaced observations spread the sample across the population and may be more efficient than simple random sampling without replacement. The orange points represent one systematic sample.

FIGURE 5

The opposite occurs when the response has a period that aligns with the sampling interval. In Figure 6, the highlighted sample repeatedly selects the same phase of the oscillation and therefore badly misrepresents the population.

Finally, variance estimation is difficult because a single random start yields only one systematic sample. In general, there is no variance estimator based on that one sample alone that is unbiased under the design.

One Stage Cluster Sampling

In **one stage cluster sampling**, the sampling frame is a list of clusters denoted by $\{1, 2, \ldots, N\}$. The sample consists of n clusters, drawn using simple random sampling without replacement, and every element in each sampled cluster is observed.

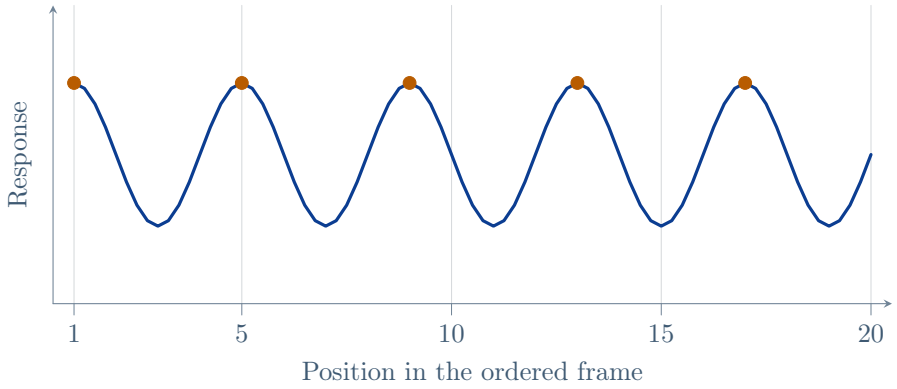


FIGURE 6

Figure 7 emphasizes the two levels of the design. Randomization selects entire clusters, after which all observational units inside each selected cluster are measured.

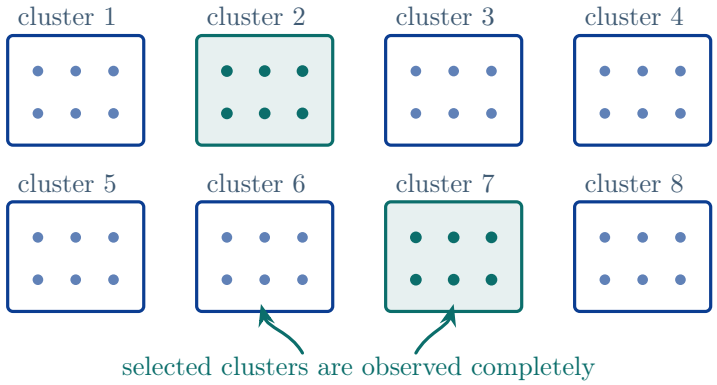


FIGURE 7

Notation 0.16 Here N is the number of clusters, M is the number of individuals in each cluster, and the population size is $N \cdot M = T$. The cluster average is

$$\mu_i = \bar{y}_i = \frac{1}{M} \sum_{j=1}^{M} y_{ij}.$$

Here y_{ij} is the response of interest from individual j in cluster i, where $i = 1, \ldots, N$ and $j = 1, \ldots, M$.

The population average is

$$\mu = \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{M}\sum_{j=1}^{M} y_{ij}\right] = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i.$$

The sample data are y_{ij}, where $i \in s$ and $j = 1, \ldots, M$, or equivalently the cluster means $\bar{y}_i$, where $i \in s$.

The point estimate for the population average is

$$\hat{\mu} = \frac{1}{nM}\sum_{i\in s}\sum_{j=1}^{M} y_{ij} = \frac{1}{n}\sum_{i\in s}\bar{y}_i.$$

Proposition 0.17 For one stage cluster sampling with equal cluster size,

$$\mathbb{E}[\tilde{\mu}] = \mu \qquad \text{and} \qquad \mathrm{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_c^2}{n}.$$

Proof. Set $z_i = \bar{y}_i$. Then

$$\tilde{\mu} = \frac{1}{n}\sum_{i\in s}\bar{y}_i = \frac{1}{n}\sum_{i\in s} z_i.$$

This is simple random sampling without replacement from the finite population $z_1, \ldots, z_N$. Therefore, Proposition 0.9 gives

$$\begin{aligned}\mathbb{E}[\tilde{\mu}] = \mu_z &= \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i \\ &= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \mu.\end{aligned}$$

Moreover,

$$\mathrm{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_z^2}{n},$$

where

$$\sigma_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \mu_z)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_i - \mu)^2.$$

This is precisely the asserted variance with $\sigma_c^2 = \sigma_z^2$. □

Thus

$$\sigma_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_i - \mu)^2 \qquad \text{and} \qquad \hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i\in s}(\bar{y}_i - \hat{\mu})^2.$$

Therefore, the estimated variance of the cluster sampling estimator is

$$\widehat{\mathrm{Var}}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_c^2}{n}.$$

Cluster sampling is inexpensive when units within the same cluster can be observed together, but statistical efficiency depends on how similar those units are. Under a common exchangeable correlation model with cluster size M and correlation ρ within clusters, the variance inflation relative to independent observations is approximately

$$\text{design effect} = 1 + (M-1)\rho.$$

Positive correlation within clusters therefore reduces the effective amount of information in a sample of nM observations. This is why good clusters for one stage sampling resemble small versions of the whole population rather than internally homogeneous groups.

The results in Propositions 0.9, 0.13 and 0.17 show that an estimator cannot be evaluated apart from its sampling design. Simple random sampling without replacement gains the finite population correction, systematic sampling can exploit a helpful ordering but is vulnerable to periodicity, and cluster sampling trades collection cost against within-cluster redundancy.

3. Ratio, Regression, and Stratified Sampling

Auxiliary information and population structure can substantially improve survey precision. Ratio and regression estimators exploit a related explanatory variable, while stratification divides the frame into internally homogeneous groups.

Ratio Estimation

There are two main reasons to use the ratio estimator. First, the population quantity of interest may itself be a ratio. Second, we may be able to improve estimation by using explanatory variables.

We first consider estimating a ratio. Suppose we are interested in the average yield in bushels of

grain per acre. The population consists of fields of different sizes. Let y_i be the number of bushels of grain harvested in field i, and let x_i be the acreage of field i. Then μ_y is the average yield in bushels per field, μ_x is the average acreage per field, and, provided $\mu_x \neq 0$, the average yield in bushels per acre is

$$\theta = \frac{\mu_y}{\mu_x}.$$

For a sample s with $\hat{\mu}_x = \bar{x} \neq 0$, the observed data are $\{(x_i, y_i) : i \in s\}$, and the ratio point estimate is

$$\hat{\theta} = \frac{\hat{\mu}_y}{\hat{\mu}_x} = \frac{\bar{y}}{\bar{x}}.$$

The corresponding estimator is $\tilde{\theta} = \tilde{\mu}_y/\tilde{\mu}_x$. Unlike a sample mean, a ratio of unbiased estimators is not generally unbiased. We therefore derive first order approximations to its expectation and variance, then use a Gaussian approximation for inference.

Recall that the first order Taylor approximation of a differentiable function $f(x, y)$ at (x_0, y_0) is

$$f(x,y) \approx f(x_0,y_0) + \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} (x - x_0) + \left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} (y - y_0).$$

We apply this approximation to $f(x, y) = y/x$. In this case,

$$\frac{\partial f}{\partial x} = -\frac{y}{x^2} \qquad \text{and} \qquad \frac{\partial f}{\partial y} = \frac{1}{x},$$

so

$$\frac{y}{x} \approx \frac{y_0}{x_0} + \frac{1}{x_0}(y - y_0) - \frac{y_0}{x_0^2}(x - x_0).$$

Setting $(x, y) = (\tilde{\mu}_x, \tilde{\mu}_y)$ and $(x_0, y_0) = (\mu_x, \mu_y)$ yields

$$\tilde{\theta} = \frac{\tilde{\mu}_y}{\tilde{\mu}_x} \approx \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\tilde{\mu}_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\tilde{\mu}_x - \mu_x).$$

Because both sample means are unbiased under simple random sampling without replacement,

$$\mathbb{E}\left[\frac{\tilde{\mu}_y}{\tilde{\mu}_x}\right] \approx \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}\mathbb{E}\left[\underbrace{\tilde{\mu}_y - \mu_y}_{=0}\right] - \frac{\mu_y}{\mu_x^2}\mathbb{E}\left[\underbrace{\tilde{\mu}_x - \mu_x}_{=0}\right].$$

Therefore,

$$\mathbb{E}\left[\tilde{\theta}\right] \approx \frac{\mu_y}{\mu_x},$$

so $\tilde{\theta}$ is approximately unbiased. Its first order variance is

$$\begin{aligned}
\operatorname{Var}(\tilde{\theta}) &= \operatorname{Var}\left(\frac{\tilde{\mu}_y}{\tilde{\mu}_x}\right) \\
&\approx \operatorname{Var}\left(\frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\tilde{\mu}_y - \mu_y) - \frac{\mu_y}{\mu_x^2}(\tilde{\mu}_x - \mu_x)\right) \\
&= \operatorname{Var}\left(\frac{1}{\mu_x}\left[\tilde{\mu}_y - \frac{\mu_y}{\mu_x}\tilde{\mu}_x\right]\right) \\
&= \frac{1}{\mu_x^2}\operatorname{Var}(\tilde{\mu}_y - \theta\tilde{\mu}_x).
\end{aligned}$$

Let $r_i = y_i - \theta x_i$. The corresponding population, sample, and estimator means are

$$\mu_r = \mu_y - \theta\mu_x, \qquad \hat{\mu}_r = \hat{\mu}_y - \theta\hat{\mu}_x \qquad \text{and} \qquad \tilde{\mu}_r = \tilde{\mu}_y - \theta\tilde{\mu}_x.$$

Because $\theta = \mu_y/\mu_x$, the population residual mean is $\mu_r = 0$. Moreover, $\mathbb{E}[\tilde{\mu}_r] = \mu_r$, and Proposition 0.9 gives

$$\operatorname{Var}(\tilde{\mu}_r) = \left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}.$$

The finite population variance of the residuals is

$$\sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(r_i - \mu_r)^2.$$

To estimate this variance from the sample, begin with the sample residual variance and replace the unknown ratio θ by $\hat{\theta}$:

$$\begin{aligned}
\hat{\sigma}_r^2 &= \frac{1}{n-1}\sum_{i\in s}(r_i - \hat{\mu}_r)^2 \\
&= \frac{1}{n-1}\sum_{i\in s}(r_i - \bar{r})^2 \\
&= \frac{1}{n-1}\sum_{i\in s}(y_i - \theta x_i - (\bar{y} - \theta\bar{x}))^2 \\
&\approx \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i - \bar{y} + \hat{\theta}\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i - \bar{y} + \underbrace{\frac{\bar{y}}{\bar{x}}\cdot\bar{x}}_{=\bar{y}})^2.
\end{aligned}$$

Thus the residual variance estimate is

$$\hat{\sigma}_r^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i)^2.$$

The estimated sampling variance of the residual mean used in this linearization is

$$\widehat{\mathrm{Var}}(\tilde{\mu}_r) = \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_r^2}{n}.$$

Consequently,

$$\widehat{\mathrm{Var}}(\tilde{\theta}) = \frac{1}{\hat{\mu}_x^2}\left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_r^2}{n}. \tag{0.3}$$

To construct a large sample confidence interval, use the fact that $\tilde{\theta}$ is approximately Gaussian. The interval has the form

$$\hat{\theta} \pm c \,\text{standard error}(\hat{\theta}),$$

where the standard error is calculated as

$$\text{standard error}(\hat{\theta}) = \sqrt{\widehat{\mathrm{Var}}(\tilde{\theta})} = \sqrt{\frac{1}{\hat{\mu}_x^2}\left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_r^2}{n}}.$$

Here

$$\hat{\sigma}_r^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i)^2.$$

We now consider ratio estimation of the population average μ_y. Suppose the survey also records auxiliary values (x_i) whose population average μ_x is known. For example, demographic surveys can use gender ratios or age distributions obtained from a census.

Use this information to adjust $\bar{y}$, the usual simple-random-sampling estimate, based on the difference between the sample and population averages of the explanatory variable.

The resulting point estimate of μ_y is

$$\hat{\mu}_{\text{ratio}} = \frac{\hat{\mu}_y}{\hat{\mu}_x}\mu_x = \frac{\mu_x}{\hat{\mu}_x}\hat{\mu}_y = \hat{\theta}\mu_x.$$

Indeed, since

$$\frac{\mu_y}{\mu_x} \approx \frac{\hat{\mu}_y}{\hat{\mu}_x},$$

we obtain

$$\hat{\mu}_{\text{ratio}} = \frac{\bar{y}}{\bar{x}}\mu_x.$$

For $\tilde{\mu}_{\text{ratio}} = \tilde{\theta}\mu_x$, the estimator is approximately unbiased because

$$\mathbb{E}[\tilde{\mu}_{\text{ratio}}] = \mu_x \mathbb{E}[\tilde{\theta}] \approx \mu_x \frac{\mu_y}{\mu_x} = \mu_y.$$

Its approximate variance is

$$\text{Var}(\tilde{\mu}_{\text{ratio}}) = \text{Var}(\mu_x \tilde{\theta}) = \mu_x^2 \,\text{Var}(\tilde{\theta}) \approx \mu_x^2 \frac{1}{\mu_x^2} \left(1 - \frac{n}{N}\right) \frac{\sigma_r^2}{n} = \left(1 - \frac{n}{N}\right) \frac{\sigma_r^2}{n}.$$

The variance estimator for $\tilde{\mu}_{\text{ratio}}$ is

$$\widehat{\text{Var}}(\tilde{\mu}_{\text{ratio}}) = \left(1 - \frac{n}{N}\right) \frac{\hat{\sigma}_r^2}{n}, \tag{0.4}$$

where

$$\hat{\sigma}_r^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{\theta} x_i)^2.$$

For comparison, the ratio estimator and the ordinary estimator have approximate variances

$$\text{Var}(\tilde{\mu}_{\text{ratio}}) \approx \left(1 - \frac{n}{N}\right) \frac{\sigma_r^2}{n},$$
$$\text{Var}(\tilde{\mu}_y) = \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n}.$$

Under this first order approximation, the ratio estimator is more efficient precisely when

$$\text{Var}(\tilde{\mu}_{\text{ratio}}) < \text{Var}(\tilde{\mu}_y),$$

which is equivalent to

$$\sigma_r^2 < \sigma_y^2.$$

Using the residual variance identity derived below, this condition is equivalent to

$$\theta^2 \sigma_x^2 - 2\theta S_{xy} + \sigma_y^2 < \sigma_y^2 \qquad \text{and} \qquad \theta^2 \sigma_x^2 < 2\theta S_{xy},$$

where S_{xy} is the finite population covariance

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y).$$

For the positive valued yield and acreage variables in this example, $\theta > 0$, $\sigma_x > 0$, and $\sigma_y > 0$.

Dividing by $\theta\sigma_x\sigma_y$ and using $\theta = \mu_y/\mu_x$ gives

$$\theta\frac{\sigma_x}{\sigma_y} < 2\frac{S_{xy}}{\sigma_x\sigma_y} \qquad \text{and} \qquad \frac{\mu_y}{\mu_x}\frac{\sigma_x}{\sigma_y} < 2\frac{S_{xy}}{\sigma_x\sigma_y}.$$

Define the finite population correlation by $\rho_{xy} = S_{xy}/(\sigma_x\sigma_y)$. The efficiency condition can then be written as

$$\frac{\sigma_x/\mu_x}{\sigma_y/\mu_y} < 2\rho_{xy}.$$

The left side is the ratio of the coefficients of variation of x and y. If this ratio is approximately 1, then the ratio estimator is more efficient when $1/2 < \rho_{xy}$.

To verify the efficiency condition, expand the residual variance:

$$\begin{aligned}
\sigma_r^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(r_i - \mu_r)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \theta x_i)^2 \quad (\text{since } \mu_r = 0) \\
&= \theta^2\sigma_x^2 - 2\theta S_{xy} + \sigma_y^2.
\end{aligned}$$

Figure 8 gives the geometric intuition: the residuals $y_i - \theta x_i$ are small when a line through the origin fits the population points well.

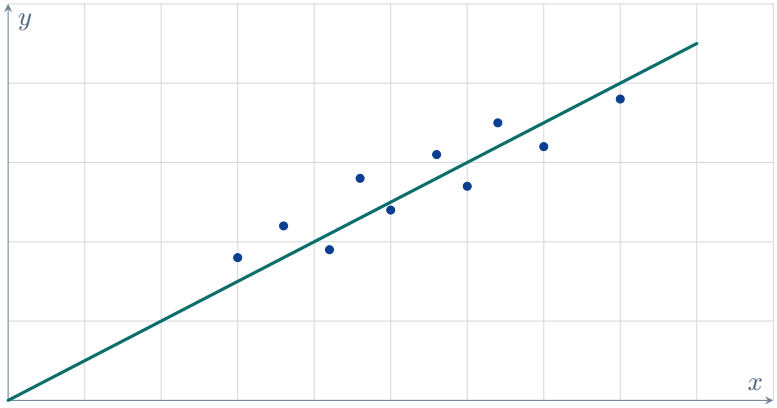


FIGURE 8

Thus, if a line through the origin fits the population data well, the ratio estimator tends to be more efficient than the unadjusted sample average.

Definition 0.18 For a random variable with nonzero mean, the **coefficient of variation** is

$$\frac{\text{standard deviation}(X)}{\mathbb{E}[X]} = \frac{\sqrt{\text{Var}(X)}}{\mathbb{E}[X]}.$$

Regression Estimation

When the fitted line need not pass through the origin, we use the centered working model

$$Y_i = \alpha + \beta(x_i - \bar{x}) + R_i \qquad \text{and} \qquad R_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

The model motivates the adjustment, but the repeated-sampling calculations below remain design based: the population pairs (x_i, y_i) are fixed, and randomness comes from the selected sample.

Least squares gives

$$\hat{\alpha} = \bar{y} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum_{i\in s}(x_i - \bar{x})y_i}{\sum_{i\in s}(x_i - \bar{x})^2} = \frac{\sum_{i\in s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in s}(x_i - \bar{x})^2}.$$

Regression estimation also requires the population average μ_x of the auxiliary variable. The resulting point estimate is the adjusted sample mean

$$\hat{\mu}_{\text{reg}} = \hat{\alpha} + \hat{\beta}(\mu_x - \hat{\mu}_x) = \hat{\mu}_y + \hat{\beta}(\mu_x - \hat{\mu}_x).$$

The corresponding estimator is

$$\tilde{\mu}_{\text{reg}} = \tilde{\mu}_y + \tilde{\beta}(\mu_x - \tilde{\mu}_x).$$

Subtracting μ_y and separating the contribution of β, we obtain

$$\begin{aligned}
\tilde{\mu}_{\text{reg}} - \mu_y &= \tilde{\mu}_y - \mu_y + \tilde{\beta}(\mu_x - \tilde{\mu}_x) \\
&= \tilde{\mu}_y - \mu_y + (\tilde{\beta} - \beta + \beta)(\mu_x - \tilde{\mu}_x) \\
&= \tilde{\mu}_y - \mu_y + \beta(\mu_x - \tilde{\mu}_x) + (\tilde{\beta} - \beta)(\mu_x - \tilde{\mu}_x).
\end{aligned}$$

The final term is a product of two estimation errors and is therefore of smaller order than the other terms in a large sample. Neglecting it gives

$$\tilde{\mu}_{\text{reg}} - \mu_y \approx \tilde{\mu}_y - \mu_y + \beta(\mu_x - \tilde{\mu}_x).$$

The regression estimator is therefore approximately unbiased:

$$\mathbb{E}[\tilde{\mu}_{\text{reg}} - \mu_y] \approx \mathbb{E}[\tilde{\mu}_y - \mu_y] + \beta\mathbb{E}[\mu_x - \tilde{\mu}_x] = 0.$$

Similarly,

$$\mathrm{Var}(\tilde{\mu}_{\mathrm{reg}}) = \mathrm{Var}(\tilde{\mu}_{\mathrm{reg}} - \mu_y) \approx \mathrm{Var}(\tilde{\mu}_y - \mu_y + \beta(\mu_x - \tilde{\mu}_x)).$$

Define $e_i = y_i - \mu_y + \beta(\mu_x - x_i)$. Its population, sample, and estimator averages satisfy

$$\begin{aligned}
\mu_e &= \frac{1}{N}\sum_{i=1}^{N} e_i = \mu_y - \mu_y + \beta(\mu_x - \mu_x) = 0, \\
\hat{\mu}_e &= \frac{1}{n}\sum_{i\in s} e_i = \hat{\mu}_y - \mu_y + \beta(\mu_x - \hat{\mu}_x), \\
\tilde{\mu}_e &= \tilde{\mu}_y - \mu_y + \beta(\mu_x - \tilde{\mu}_x).
\end{aligned}$$

Consequently,

$$\mathbb{E}\left[\tilde{\mu}_e\right] = \mu_e = 0 \qquad \text{and} \qquad \mathrm{Var}(\tilde{\mu}_e) = \left(1 - \frac{n}{N}\right)\frac{\sigma_e^2}{n}.$$

The corresponding finite population residual variance is

$$\sigma_e^2 = \frac{1}{N-1}\sum_{i=1}^{N}(e_i - \mu_e)^2 = \frac{1}{N-1}\sum_{i=1}^{N} e_i^2 = \frac{1}{N-1}\sum_{i=1}^{N}[y_i - \mu_y + \beta(\mu_x - x_i)]^2.$$

Replacing the unknown quantities by their sample estimates gives the design based residual variance estimate

$$\begin{aligned}
\hat{\sigma}_e^2 &= \frac{1}{n-1}\sum_{i\in s}[y_i - \hat{\mu}_y + \hat{\beta}(\hat{\mu}_x - x_i)]^2 \\
&= \frac{1}{n-1}\sum_{i\in s}[y_i - \hat{\mu}_y - \hat{\beta}(x_i - \hat{\mu}_x)]^2 \\
&= \frac{1}{n-1}\sum_{i\in s}[y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2 \\
&= \frac{1}{n-1}\sum_{i\in s}[y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})]^2.
\end{aligned}$$

If the regression model itself is used for exact model based inference, two regression coefficients have been fitted, so the usual unbiased estimator of the model error variance instead divides the residual sum of squares by $n-2$. The denominator $n-1$ above belongs to the first order survey variance calculation.

Figure 9 illustrates the adjustment. The ordinary sample mean is located at $\bar{x}$, whereas the regression estimate evaluates the fitted relationship at the known auxiliary variable mean μ_x. The widening band also illustrates a setting with an approximately constant coefficient of variation.

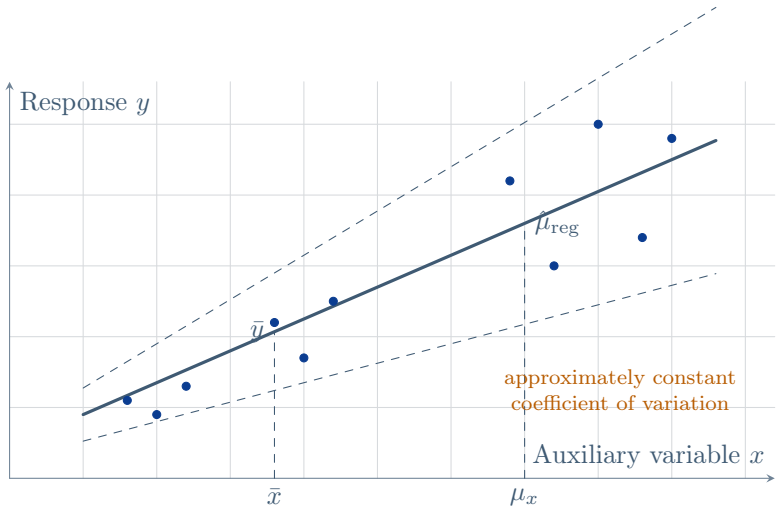


FIGURE 9

For regression estimation, comparison with the unadjusted sample average reduces to

$$\text{Var}(\tilde{\mu}_{\text{reg}}) \leq \text{Var}(\tilde{\mu}_y) \qquad \text{and} \qquad \sigma_e^2 \leq \sigma_y^2.$$

Since $e_i = (y_i - \mu_y) - \beta(x_i - \mu_x)$, the regression residual variance is

$$\sigma_e^2 = \frac{1}{N-1}\sum_{i=1}^{N}[(y_i - \mu_y) - \beta(x_i - \mu_x)]^2 = \sigma_y^2 - 2\beta S_{xy} + \beta^2\sigma_x^2.$$

If $\sigma_x^2 > 0$, the finite population least squares slope is $\beta = S_{xy}/\sigma_x^2$. Substitution gives

$$\sigma_e^2 = \sigma_y^2 - \frac{S_{xy}^2}{\sigma_x^2} = \sigma_y^2(1 - \rho_{xy}^2) \leq \sigma_y^2.$$

Thus the ideal regression adjustment is at least as efficient as the unadjusted estimator, with a larger improvement when the auxiliary variable is more strongly correlated with the response.

In summary, the regression point estimate is

$$\hat{\mu}_{\text{reg}} = \hat{\mu}_y + \hat{\beta}(\mu_x - \hat{\mu}_x) = \hat{\alpha} + \hat{\beta}(\mu_x - \hat{\mu}_x).$$

The approximate variance estimator is

$$\widehat{\text{Var}}(\tilde{\mu}_{\text{reg}}) \approx \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_e^2}{n},$$

where the residual variance estimate is

$$\hat{\sigma}_e^2 = \frac{1}{n-1}\sum_{i\in s}[y_i - \hat{\alpha} - \hat{\beta}(x_i - \hat{\mu}_x)]^2 = \frac{1}{n-1}\sum_{i\in s}[y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2.$$

Regression estimation in this form uses continuous response and explanatory variables and requires the population average of the explanatory variable. The method extends to multiple explanatory variables and nonlinear relationships. For example, when $x > 0$ and $y > 0$, the power relationship $y = x^\alpha$ becomes linear after taking logarithms:

$$\log y = \alpha \log x.$$

Stratified Random Sampling

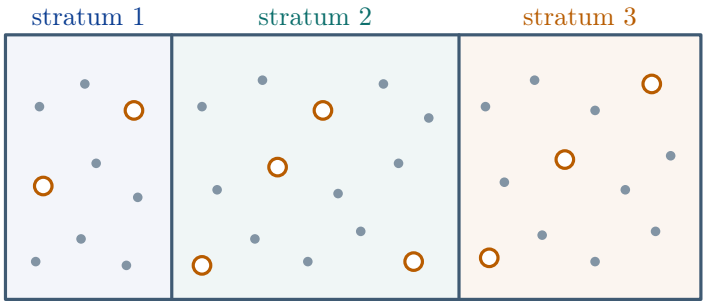
The sampling frame is partitioned into H disjoint lists $U_1, U_2, \ldots, U_H$, called strata. The stratum U_h contains N_h individuals, and

$$U_i \cap U_j = \emptyset \quad \text{whenever } i \neq j.$$

Thus each element belongs to exactly one stratum, and the complete frame is $U = U_1 \cup \cdots \cup U_H$.

Example 0.19 Examples of strata include provinces for polling, home faculties for a student survey at the University of Waterloo, and lists of schools in the Waterloo region.

The purpose of stratification is to guarantee representation from every group and, when units are relatively similar within each group, to reduce estimator variance. Figure 10 shows independent samples drawn from three differently sized strata.



sampling is randomized separately within each stratum

FIGURE 10

The total number of individuals is known to be $N = N_1 + \cdots + N_H$. Let y_{hj} denote the response of the jth element in stratum h. For stratum h, the population quantities of interest are the average

$$\mu_h = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$$

and the variance

$$\sigma_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2.$$

The population average is therefore

$$\mu = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{N_h} \sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h = \sum_{h=1}^{H} W_h \mu_h,$$

where $W_h = N_h/N$ is the **stratum weight**.

The population variance is

$$\sigma^2 = \frac{1}{N - 1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \mu)^2.$$

The sampling scheme draws n_h individuals from stratum h by simple random sampling without replacement. The complete sample is $s = s_1 \cup \cdots \cup s_H$, its size is $n = n_1 + \cdots + n_H$, and $w_h = n_h/n$ is the sample proportion drawn from stratum h.

For stratum h, the sample average is defined when $n_h \geq 1$, and the sample variance is defined when $n_h \geq 2$:

$$\hat{\mu}_h = \bar{y}_h = \frac{1}{n_h} \sum_{j \in s_h} y_{hj} \qquad \text{and} \qquad \hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} (y_{hj} - \bar{y}_h)^2.$$

The corresponding estimator is $\tilde{\mu}_h$. By Proposition 0.9,

$$\mathbb{E}\left[\tilde{\mu}_h\right] = \mu_h \qquad \text{and} \qquad \operatorname{Var}(\tilde{\mu}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}.$$

Proposition 0.20 The point estimate and corresponding estimator for the population

average are

$$\hat{\mu}_s = \sum_{h=1}^{H} W_h \hat{\mu}_h \qquad \text{and} \qquad \tilde{\mu}_s = \sum_{h=1}^{H} W_h \tilde{\mu}_h.$$

Then

$$\mathbb{E}[\tilde{\mu}_s] = \mu \qquad \text{and} \qquad \operatorname{Var}(\tilde{\mu}_s) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}.$$

Proof. Linearity of expectation gives

$$\begin{aligned}\mathbb{E}\left[\tilde{\mu}_s\right] &= \sum_{h=1}^{H} W_h \mathbb{E}\left[\tilde{\mu}_h\right] = \sum_{h=1}^{H} W_h \mu_h \\ &= \sum_{h=1}^{H} \frac{N_h}{N} \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj} = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \mu.\end{aligned}$$

Because sampling is independent across strata,

$$\operatorname{Var}(\tilde{\mu}_s) = \operatorname{Var}\left(\sum_{h=1}^{H} W_h \tilde{\mu}_h\right) = \sum_{h=1}^{H} W_h^2 \operatorname{Var}(\tilde{\mu}_h) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}.$$

□

Provided $n_h \geq 2$ in every stratum, the point estimate of this variance is

$$\widehat{\operatorname{Var}}(\tilde{\mu}_s) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{\sigma}_h^2}{n_h}.$$

Three natural questions guide the remainder of the discussion:

1. Can ratio or regression estimation be combined with stratified sampling?
2. When is stratified sampling more efficient than simple random sampling without replacement?
3. How should observations be allocated to strata?

For (1), the answer is yes. Since the strata are sampled independently, each stratum average may be estimated using the best available method, including ratio or regression estimation, before the

estimates are combined using the stratum weights.

To answer (2), consider proportional allocation, in which the sampling weights equal the stratum weights. Thus

$$W_h = w_h \iff \frac{N_h}{N} = \frac{n_h}{n},$$

so

$$\frac{n}{N} = \frac{n_h}{N_h} \qquad \text{and} \qquad n_h \propto N_h.$$

Under this allocation,

$$\begin{aligned}
\text{Var}(\tilde{\mu}_{\text{prop}}) &= \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} \\
&= \sum_{h=1}^{H} W_h \left(1 - \frac{n_h}{N_h}\right) \frac{N_h}{N} \frac{1}{n_h} \cdot \sigma_h^2 \\
&= \sum_{h=1}^{H} W_h \left(1 - \frac{n}{N}\right) \frac{N_h}{N} \frac{1}{n_h} \cdot \sigma_h^2 \\
&= \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} W_h \sigma_h^2.
\end{aligned}$$

For simple random sampling without replacement,

$$\text{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}.$$

To compare the two variances, partition the total sum of squares into within-stratum and between-stratum components:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N-1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 \\
&= \frac{1}{N-1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{hj} - \mu_h + \mu_h - \mu)^2 \\
&= \frac{1}{N-1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} \left[(y_{hj} - \mu_h)^2 + 2(y_{hj} - \mu_h)(\mu_h - \mu) + (\mu_h - \mu)^2\right] \\
&= \frac{1}{N-1} \sum_{h=1}^{H} (N_h - 1)\sigma_h^2 + \frac{2}{N-1} \sum_{h=1}^{H} (\mu_h - \mu) \sum_{j=1}^{N_h} (y_{hj} - \mu_h)
\end{aligned}$$

$$+ \frac{1}{N-1}\sum_{h=1}^{H} N_h(\mu_h - \mu)^2.$$

The cross term vanishes because

$$\sum_{j=1}^{N_h}(y_{hj} - \mu_h) = \sum_{j=1}^{N_h} y_{hj} - N_h\mu_h = 0.$$

Continuing the derivation gives

$$\begin{aligned}\sigma^2 &= \frac{1}{N-1}\sum_{h=1}^{H}(N_h - 1)\sigma_h^2 + \frac{1}{N-1}\sum_{h=1}^{H} N_h(\mu_h - \mu)^2 \\ &= \sum_{h=1}^{H}\frac{N_h - 1}{N-1}\sigma_h^2 + \sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_h - \mu)^2.\end{aligned}$$

If all N_h are large, then

$$\sigma^2 \approx \sum_{h=1}^{H} W_h\sigma_h^2 + \sum_{h=1}^{H} W_h(\mu_h - \mu)^2.$$

The second term is nonnegative. Thus, for large strata, proportional allocation $n_h = nW_h$ is at least as efficient as simple random sampling without replacement, and it is strictly more efficient when the stratum means are not all equal. Stratification is therefore most useful when the strata separate populations with different means while remaining internally homogeneous. Question (3) leads to the optimal allocation calculation below.

Optimal Allocation

We want to choose $n_1, \ldots, n_H$ to minimize the stratified variance $\mathrm{Var}(\tilde{\mu}_s)$, subject to a fixed total sample size:

$$\min_{n_1,\ldots,n_H} \mathrm{Var}(\tilde{\mu}_s) \quad \text{such that} \quad n = n_1 + \cdots + n_H$$

The corresponding Lagrangian is

$$L(n_1, \ldots, n_H, \lambda) = \sum_{h=1}^{H} W_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} + \lambda(n_1 + \cdots + n_H - n).$$

The finite population correction can be rewritten as

$$\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} = \frac{N_h - n_h}{N_h n_h} = \frac{1}{n_h} - \frac{1}{N_h}.$$

The Lagrangian is therefore

$$L(n_1, \ldots, n_H, \lambda) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2 + \lambda(n_1 + \cdots + n_H - n).$$

Taking the partial derivative with respect to n_h gives

$$\frac{\partial L}{\partial n_h} = -\frac{W_h^2 \sigma_h^2}{n_h^2} + \lambda.$$

Setting $\frac{\partial L}{\partial n_h} = 0$ gives

$$\begin{aligned} \frac{W_h^2}{n_h^2} \cdot \sigma_h^2 &= \lambda, \\ n_h^2 &= \frac{W_h^2 \sigma_h^2}{\lambda}, \\ n_h &= \frac{W_h \cdot \sigma_h}{\sqrt{\lambda}}, \end{aligned}$$

so $n_h \propto W_h \cdot \sigma_h$.

We solve for λ by setting $\frac{\partial L}{\partial \lambda} = 0$. This gives

$$\frac{\partial L}{\partial \lambda} = n_1 + \cdots + n_H - n = 0.$$

Using the formulas for $n_1, \ldots, n_H$, we obtain

$$\frac{W_1 \sigma_1}{\sqrt{\lambda}} + \cdots + \frac{W_H \sigma_H}{\sqrt{\lambda}} = n,$$

and hence

$$\sqrt{\lambda} = \frac{1}{n} \left(W_1 \sigma_1 + \cdots + W_H \sigma_H \right).$$

Therefore, the optimal allocation is

$$n_h = n \frac{W_h \sigma_h}{\sum_{\ell=1}^{H} W_\ell \sigma_\ell}. \tag{0.5}$$

The denominator is the weighted sum of the stratum population standard deviations.

The allocation in (0.5) is often called **Neyman allocation**. It assigns more observations to strata that are larger or more variable. If the cost of observing one unit in stratum h is c_h, the same

Lagrange multiplier argument under a fixed total cost gives

$$n_h \propto \frac{W_h \sigma_h}{\sqrt{c_h}}.$$

Thus a stratum receives fewer observations when its units are comparatively expensive to measure. If all stratum standard deviations are equal, say

$$\sigma_1 = \cdots = \sigma_H = c,$$

then

$$n_h = n \frac{W_h}{\sum_{\ell=1}^{H} W_\ell} = n W_h.$$

If all stratum weights are equal, say

$$W_1 = \cdots = W_H = c,$$

then

$$n_h = n \frac{\sigma_h}{\sum_{\ell=1}^{H} \sigma_\ell}, \qquad \text{and} \qquad n_h \propto \sigma_h.$$

The derivation treats the n_h's as positive continuous quantities and gives the unconstrained optimum. An implemented allocation must use integers and satisfy $1 \leq n_h \leq N_h$ (or $n_h \geq 2$ where a within-stratum variance estimate is required); strata that hit a bound are fixed there before the remaining sample is reallocated.

Post-Stratification

Suppose a discrete explanatory variable divides the population into H groups whose population proportions $W_1, \ldots, W_H$ are known. Group membership is absent from the sampling frame, so a simple random sample is selected without replacement and the group label x_i is recorded together with the response y_i for each sampled unit. The observed data are therefore

$$\{(y_i, x_i) : i \in s\}.$$

For group h, let $s_h = \{i \in s : x_i = h\}$ and $n_h = |s_h|$. The post-stratified point estimate and its corresponding estimator combine the group means using the known population weights:

$$\hat{\mu}_{\text{post}} = \sum_{h=1}^{H} W_h \hat{\mu}_h \qquad \text{and} \qquad \tilde{\mu}_{\text{post}} = \sum_{h=1}^{H} W_h \tilde{\mu}_h,$$

where the realized group mean is

$$\hat{\mu}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i.$$

Because group membership becomes known only after selection, the post-stratum sample sizes $n_1, \ldots, n_H$ are random. The estimator is defined only when every required group is represented, so the calculations below are conditional on $n_h > 0$ for each h.

To study the repeated-sampling behaviour of the estimator, first recall the law of total expectation. For discrete random variables X and Y,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in S_x} \sum_{y \in S_y} x \cdot \mathbb{P}(X = x, Y = y) \\
&= \sum_{x \in S_x} \sum_{y \in S_y} x \cdot \mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y) \\
&= \sum_{y \in S_y} \mathbb{E}[X \mid Y = y] \mathbb{P}(Y = y).
\end{aligned}$$

If $f(y) = \mathbb{E}[X \mid Y = y]$, then

$$\sum_{y \in S_y} f(y) \mathbb{P}(Y = y) = \mathbb{E}[f(Y)] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

Applying this identity to the post-stratified estimator gives

$$\begin{aligned}
\mathbb{E}[\tilde{\mu}_{\text{post}}] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{h=1}^{H} W_h \tilde{\mu}_h \,\middle|\, n_1, \ldots, n_H \right] \right] \\
&= \mathbb{E}[\mu] = \mu.
\end{aligned}$$

The inner expectation uses Proposition 0.20, conditional on the realized post-stratum sample sizes.

For every possible value y,

$$\operatorname{Var}(X \mid Y = y) = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2.$$

Applying the law of total expectation to X^2, and then adding and subtracting $\mathbb{E}[\mathbb{E}[X \mid Y]^2]$, gives

$$\begin{aligned}
\operatorname{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}[\mathbb{E}[X^2 \mid Y]] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\
&= \mathbb{E}\left[\mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2\right] + \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\
&= \mathbb{E}[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(\mathbb{E}[X \mid Y]).
\end{aligned}$$

This identity is the law of total variance.

For the post-stratified estimator, the same identity yields

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}_{\text{post}}) &= \mathbb{E}[\mathrm{Var}(\tilde{\mu}_{\text{post}} \mid n_1, \ldots, n_H)] + \mathrm{Var}(\mathbb{E}[\tilde{\mu}_{\text{post}} \mid n_1, \ldots, n_H]) \\ &= \mathbb{E}\left[\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}\right] + \mathrm{Var}(\mu) \\ &= \sum_{h=1}^{H} W_h^2 \left(\mathbb{E}\left[\frac{1}{n_h}\right] - \frac{1}{N_h}\right) \sigma_h^2.\end{aligned}$$

For large n, the random counts are concentrated near their expected values. Replacing the expectation of each reciprocal count by the reciprocal of its realized value motivates the approximation

$$\mathrm{Var}(\tilde{\mu}_{\text{post}}) \approx \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \sigma_h^2.$$

Provided every realized $n_h \geq 2$, replacing the unknown stratum variances by their sample estimates gives

$$\widehat{\mathrm{Var}}(\tilde{\mu}_{\text{post}}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \hat{\sigma}_h^2.$$

Nonresponse

Nonresponse creates bias when respondents differ systematically from nonrespondents. Figure 11 separates the target population into these two groups.

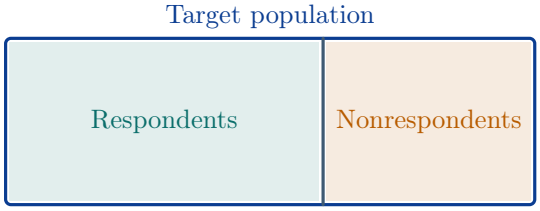


FIGURE 11

For nonresponse bias, suppose the target population of size N consists of N_R respondents and N_M nonrespondents.

$$\mu = W_R \mu_R + W_M \mu_M = \frac{N_R}{N} \mu_R + \frac{N_M}{N} \mu_M.$$

If only respondents are observed, the estimator targets μ_R rather than μ. Its bias is

$$\text{Bias}(\tilde{\mu}) = \mathbb{E}[\tilde{\mu}] - \mu = \mu_R - \mu = W_M(\mu_R - \mu_M) = \frac{N_M}{N}(\mu_R - \mu_M).$$

To reduce nonresponse bias, we may follow up with nonrespondents using a different mode of contact, such as an in-person interview instead of a phone interview or a phone interview instead of a web survey. Another option is to increase the incentive.

One correction strategy uses two phases. In the first phase, draw a simple random sample without replacement of size n, obtaining n_R respondents and n_M nonrespondents. In the second phase, draw a simple random subsample of size m from the n_M nonrespondents and pursue a more intensive follow-up.

This is post-stratification with N_R and N_M as the two unknown strata. We estimate the strata weights by

$$\hat{W}_R = \frac{n_R}{n} \qquad \text{and} \qquad \hat{W}_M = \frac{n_M}{n}.$$

The estimate of the population average is

$$\hat{\mu}_{\text{follow-up}} = \hat{W}_R\hat{\mu}_R + \hat{W}_M\hat{\mu}_m.$$

This adjustment removes the bias only if the intensive follow-up produces a representative sample of the original nonrespondents. A second nonresponse mechanism within the follow-up group can leave residual bias.

Nonresponse is distinct from frame error. As Figure 12 shows, the sampling frame may omit some members of the target population and may also contain units outside it.

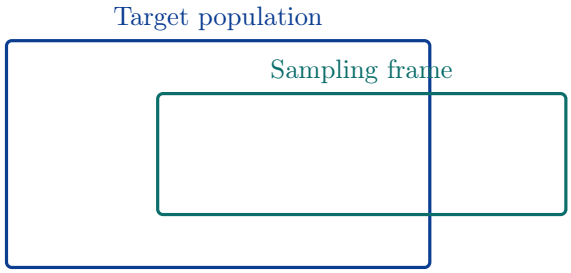


FIGURE 12

Survey Estimator Summary

The principal survey sampling estimators may now be compared in a common notation. Under simple random sampling without replacement, Proposition 0.9 gives

$$\hat{\mu} = \bar{y}, \qquad \mathbb{E}[\tilde{\mu}] = \mu, \qquad \text{and} \qquad \operatorname{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n},$$

where

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2.$$

The variance estimate and its associated standard error are

$$\widehat{\operatorname{Var}}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}^2}{n}, \qquad \text{and} \qquad \text{standard error}(\hat{\mu}) = \sqrt{\widehat{\operatorname{Var}}(\tilde{\mu})},$$

where

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\mu})^2.$$

For one stage cluster sampling with N clusters of common size M, the population average can be expressed as an average of cluster means:

$$\mu = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i\cdot}.$$

The sample consists of n complete clusters, so the point estimate is

$$\hat{\mu} = \frac{1}{n}\sum_{i\in s}\bar{y}_{i\cdot}.$$

By Proposition 0.17,

$$\operatorname{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_c^2}{n}, \qquad \text{and} \qquad \sigma_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i\cdot} - \mu)^2.$$

Ratio estimation requires measurements of an auxiliary variable x_i. When the target parameter is

$$\theta = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x},$$

the point estimate is $\hat{\theta} = \hat{\mu}_y/\hat{\mu}_x$, and its corresponding estimator is approximately unbiased. The linearization formulas are

$$\mathrm{Var}(\tilde{\theta}) \approx \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}, \qquad \text{and} \qquad \sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \theta x_i)^2.$$

As recorded in (0.3), the sample residuals $y_i - \hat{\theta}x_i$ provide the variance estimate

$$\widehat{\mathrm{Var}}(\tilde{\theta}) = \frac{1}{\hat{\mu}_x^2}\left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_r^2}{n}, \qquad \text{and} \qquad \hat{\sigma}_r^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i)^2.$$

If the target is instead μ_y and the population auxiliary mean μ_x is known, then the ratio point estimate is $\hat{\mu}_{\text{ratio}} = \hat{\theta}\mu_x$. Consequently,

$$\mathrm{Var}(\tilde{\mu}_{\text{ratio}}) \approx \left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}, \qquad \text{and} \qquad \widehat{\mathrm{Var}}(\tilde{\mu}_{\text{ratio}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_r^2}{n},$$

in agreement with (0.4).

Regression estimation also targets μ_y when μ_x is known. Its point estimate is

$$\hat{\mu}_{\text{reg}} = \bar{y} + \hat{\beta}(\mu_x - \bar{x}), \qquad \text{and} \qquad \hat{\beta} = \frac{\sum_{i\in s}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i\in s}(x_i - \bar{x})^2}.$$

The regression estimator is approximately unbiased, and its estimated variance is

$$\widehat{\mathrm{Var}}(\tilde{\mu}_{\text{reg}}) \approx \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_{\text{reg}}^2}{n},$$

where

$$\hat{\sigma}_{\text{reg}}^2 = \frac{1}{n-1}\sum_{i\in s}\left[y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})\right]^2.$$

Finally, stratified sampling applies simple random sampling without replacement independently within each stratum. With known stratum weights $W_h = N_h/N$,

$$\mu = \sum_{h=1}^{H} W_h\mu_h, \qquad \text{and} \qquad \hat{\mu}_s = \sum_{h=1}^{H} W_h\hat{\mu}_h.$$

The expectation and variance are given by Proposition 0.20. Under proportional allocation, $n_h/n = W_h$, so

$$\mathrm{Var}(\tilde{\mu}_{\text{prop}}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H} W_h\sigma_h^2.$$

Optimal allocation instead assigns observations according to (0.5); equivalently, n_h is proportional to $N_h\sigma_h$. Thus larger or more variable strata receive more observations.

These methods all use auxiliary structure in different ways. The Horvitz–Thompson estimator in Proposition 0.8 corrects unequal selection probabilities directly; ratio and regression estimators use measured auxiliary variables; and the stratified estimator in Proposition 0.20 uses known group membership and group sizes. The relevant standard error must always be derived from the design that actually generated the sample.

4. Experimental Design and Treatment Comparisons

Experimental design concerns the deliberate assignment of treatments to experimental units. Randomization supports causal comparison, replication measures experimental variation, and blocking removes variation attributable to known nuisance factors.

Experimental Design

The conceptual population is the collection of all repetitions of the experiment that could be performed under the same conditions. Unlike a finite survey population, this population is usually infinite.

Example 0.21 A gardener wants to determine how fertilizer and water affect the yield of tomato plants. Twelve plants are available. The fertilizer factor has two levels, standard fertilizer A and new fertilizer B. The water factor has three levels: half a cup, one cup, and one and a half cups, denoted by w_1, w_2, and w_3. Design an experiment whose response is the yield from each plant.

Solution. There are six treatment combinations. Since twelve plants are available, each treatment can be replicated twice:

$$(A,w_1),(A,w_1), \qquad (A,w_2),(A,w_2), \qquad (A,w_3),(A,w_3),$$
$$(B,w_1),(B,w_1), \qquad (B,w_2),(B,w_2), \qquad \text{and} \qquad (B,w_3),(B,w_3).$$

The twelve treatment labels are randomly assigned to the twelve plants, as represented in Figure 13.

When position, sunlight, or soil conditions vary systematically, the plants can instead be arranged into relatively homogeneous groups. Figure 14 illustrates two possible ways to form two blocks of

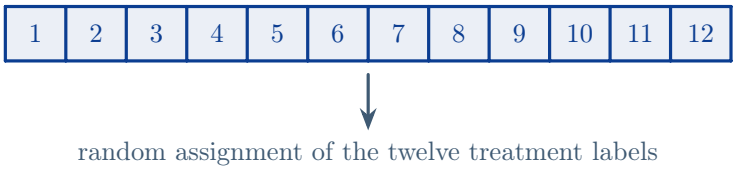


Figure 13

six plants before randomizing the six treatments within each block.

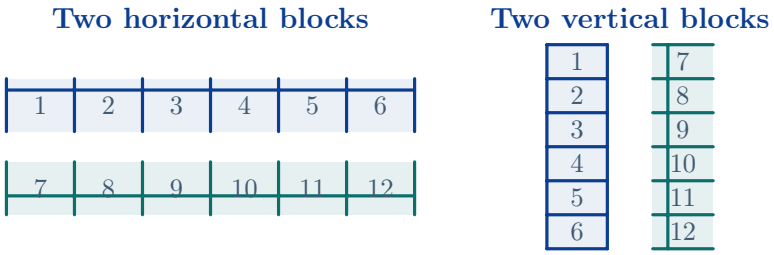


Figure 14

Placing one complete set of six treatments in each group is an instance of **blocking**. □

The example illustrates the decisions that must be made before data are collected. A useful design sequence is as follows:

1. State the scientific objective.
2. Define the response and how it will be measured.
3. Choose the factors and their levels. For a continuous factor, the chosen levels should be sufficiently separated to reveal a scientifically meaningful response without extending beyond the practical range. Illustrative triples include $(1/2, 1, 3/2)$, $(0.99, 1, 1.01)$, and $(0, 1, 100)$; their very different spacing emphasizes that the scale must be chosen deliberately.
4. Specify the experimental units, replication, blocking, and randomization scheme.
5. Conduct the experiment according to the plan and record departures from it.

Basic Experimental Terminology

A **factor** is a single explanatory variable, or input, that will be changed in an experiment. A factor can be quantitative or qualitative; examples include the amount of water and the type of fertilizer.

The **factor levels** are the values assigned to a factor. For example, fertilizer may have 2 levels and water may have 3 levels.

A **treatment** is a combination of factor levels that can be applied to an experimental unit. For example, an experiment with 2 fertilizer levels and 3 water levels has $2 \cdot 3 = 6$ treatments.

An **experimental unit** is an object or individual to which a treatment is applied.

The **response** is the outcome or observation. Although the methods developed below focus on continuous responses, a response may also be ordinal, such as a rank; a count, such as the number of tomatoes; binary, such as survival or death; or a structured object, such as an image, map, or graph. Figure 15 depicts a small response in the form of a graph.

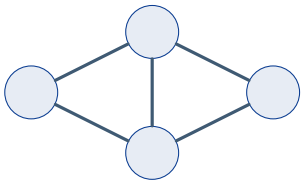


Figure 15

Fundamental Principles of Experimental Design

The fundamental principles of experimental design are blocking, replication, and randomization.

Replication occurs when a treatment is applied independently to several experimental units. Replication makes experimental variation estimable and improves the precision of treatment comparisons. Repeated measurements on the same experimental unit do not provide independent replication of the treatment.

A **treatment effect** measures the change in the expected response relative to a specified baseline, often a control treatment or the grand mean.

Randomization means assigning treatments to experimental units according to a specified probability mechanism. Because experimental units are not identical, a systematic assignment could align a treatment with an uncontrolled feature of the units. Randomization prevents subjective allocation and balances both observed and unobserved features on average, providing the basis for causal comparison.

Blocking means grouping experimental units that are known to be similar before treatments are assigned. Treatments are compared within blocks, so variation between blocks is removed from the

comparison. Randomization must still be performed independently within each block. Blocking increases efficiency when the blocking variable is genuinely related to the response.

A useful design maxim is:

Block what you can and randomize what you cannot.

The single treatment model provides the foundation for estimation and inference in more elaborate designs.

Example 0.22 Last year the gardener used fertilizer A and the same amount of water on 12 plants, then weighed all the tomatoes from each plant. The response Y_i is the yield from plant i.

For plant i, consider the model

$$Y_i = \mu + R_i, \quad i = 1, \dots, 12.$$

Here μ is the long-run average yield, while R_i represents experimental error arising from position, measurement, sunlight, and other uncontrolled sources. We assume that the errors are mutually independent and satisfy

$$\mathbb{E}[R_i] = 0 \qquad \text{and} \qquad \operatorname{Var}(R_i) = \sigma^2.$$

Exact inference under the normal model additionally requires $R_i \sim \mathcal{N}(0, \sigma^2)$.

More generally, write $y_1, \dots, y_n$ for the observed yields, with $n = 12$ in this experiment. The least squares estimate of μ is the value that minimizes the residual sum of squares,

$$\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Solving this minimization problem gives the mean estimate, and the residuals give the unbiased estimate of the error variance:

$$\hat{\mu} = \bar{y} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Before the data are observed, the corresponding estimator and its first two moments are

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \qquad \mathbb{E}[\tilde{\mu}] = \mu \qquad \text{and} \qquad \operatorname{Var}(\tilde{\mu}) = \frac{\sigma^2}{n}.$$

Under normal errors, $\tilde{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$ exactly; under weaker moment conditions, this distribution is approximately normal for large n.

Under normal errors, the scaled sample variance has the chi-square distribution

$$\frac{\tilde{\sigma}^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}.$$

For the twelve plants, suppose the sample average is 1483 grams and the sample standard deviation is 115.1 grams. The product label claims that each plant should yield 1500 grams. The inferential question is whether the observed data provide evidence that the true mean yield differs from this advertised value.

Remark 0.23 Suppose $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_k$ are independent. Then the statistic

$$T = \frac{Z}{\sqrt{U/k}}.$$

has a t_k distribution, where k is the degrees of freedom.

The resulting test statistic has a t-distribution:

$$\frac{\tilde{\mu} - \mu}{\tilde{\sigma}/\sqrt{n}} \sim t_{n-1}.$$

This pivot produces confidence intervals and hypothesis tests.

A $(1-\alpha) \times 100\%$ confidence interval is

$$\hat{\mu} \pm c \cdot \text{standard error}(\hat{\mu}),$$

where c is such that $\mathbb{P}(|T_{n-1}| \leq c) = 1 - \alpha$.

Example 0.24 Construct a 95% confidence interval for the mean yield using $n = 12$, $\hat{\mu} = 1483$ grams, and $\hat{\sigma} = 115.1$ grams.

Solution. The critical value is determined by

$$\mathbb{P}(T_{11} \leq 2.20) = 0.975, \qquad \text{and} \qquad \mathbb{P}(|T_{11}| \leq 2.20) = 0.95.$$

Therefore, the exact normal theory interval is

$$1483 \pm 2.20 \cdot \frac{115.1}{\sqrt{12}} = 1483 \pm 73.1,$$

or approximately $(1409.9, 1556.1)$ grams. □

A p-value is the probability, calculated under the null hypothesis H_0, of obtaining a test statistic at least as extreme as the observed value in the direction specified by the alternative hypothesis.

$$t_{\text{obs}} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{1483 - 1500}{115.1/\sqrt{12}} = -0.511.$$

For $\mu_0 = 1500$, the possible alternatives lead to the following tail probabilities.

1. For $H_0 : \mu = 1500$ against $H_A : \mu \neq 1500$, the two-sided p-value is

$$p\text{-value} = \mathbb{P}\left(|T_{11}| \geq |t_{\text{obs}}|\right).$$

2. For $H_0 : \mu = 1500$ against $H_A : \mu < 1500$, the lower-tail p-value is

$$p\text{-value} = \mathbb{P}(T_{11} \leq t_{\text{obs}}).$$

3. For $H_0 : \mu = 1500$ against $H_A : \mu > 1500$, the upper-tail p-value is

$$p\text{-value} = \mathbb{P}(T_{11} \geq t_{\text{obs}}).$$

The observations must be independent with a common mean and variance. Exact t-inference also requires normal errors; a normal quantile–quantile plot of the residuals helps assess this condition.

Comparisons of Two Treatments Without Blocking

The gardener wants to compare two fertilizers, A and B. The gardener applies fertilizer A to 6 plants and fertilizer B to 6 plants, using the same amount of water.

Writing Y_{ij} for the response of plant j under treatment i, consider the model

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad i = 1, 2, \quad j = 1, \ldots, n_i.$$

μ is the long-run average yield of tomatoes from each plant. The parameter τ_1 is the treatment effect from fertilizer A, and τ_2 is the treatment effect from fertilizer B. We impose the constraint

$$\tau_1 + \tau_2 = 0.$$

For two treatments, we could parameterize the model more simply, but this constrained form extends directly to experiments with many treatments. The variable R_{ij} represents experimental error. The errors are independent, with mean zero and common variance σ^2. For exact normal theory inference, they are also assumed to be normally distributed.

The long-run average for plants using fertilizer A is $\mu + \tau_1$, while the long-run average for plants using fertilizer B is $\mu + \tau_2$. The difference between A and B is

$$(\mu + \tau_1) - (\mu + \tau_2) = \underbrace{\tau_1 - \tau_2}_{\text{quantity of interest}},$$

which is the parameter of interest.

An equivalent parameterization assigns each treatment its own mean:

$$Y_{1j} = \mu_1 + R_{1j} \qquad \text{and} \qquad Y_{2j} = \mu_2 + R_{2j},$$

where all errors are independent, centered, and have variance σ^2. In this parameterization, the quantity of interest is $\mu_1 - \mu_2$.

The observed data are $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$.

The least squares point estimates are obtained by minimizing

$$\min_{\mu,\tau_1,\tau_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 \quad \text{subject to} \quad \tau_1 + \tau_2 = 0.$$

Equivalently, we minimize the Lagrangian

$$L(\mu, \tau_1, \tau_2, \lambda) = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 + \lambda(\tau_1 + \tau_2).$$

Solving the normal equations gives the fitted treatment means $\hat{\mu} + \hat{\tau}_i = \bar{y}_{i\bullet}$. Under the constraint $\tau_1 + \tau_2 = 0$, the individual parameter estimates are

$$\hat{\mu} = \frac{\bar{y}_{1\bullet} + \bar{y}_{2\bullet}}{2},$$

$$\hat{\tau}_1 = \bar{y}_{1\bullet} - \hat{\mu},$$
$$\hat{\tau}_2 = \bar{y}_{2\bullet} - \hat{\mu},$$

where the treatment sample mean is

$$\bar{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

The bullet indicates averaging over that index. Thus $\bar{y}_{i\bullet}$ averages over replicates within treatment i. In the balanced fertilizer experiment, $n_1 = n_2 = 6$, so $\hat{\mu}$ also equals the ordinary grand mean of all twelve observations. The treatment difference is $\hat{\tau}_1 - \hat{\tau}_2 = \bar{y}_{1\bullet} - \bar{y}_{2\bullet}$, whether or not the replicate counts are equal.

An unbiased estimate of the common error variance σ^2 is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 \\ &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2.\end{aligned}$$

Under normal errors, the treatment means have the distributions

$$\bar{Y}_{1\bullet} \sim \mathcal{N}(\mu + \tau_1, \sigma^2/n_1),$$
$$\bar{Y}_{2\bullet} \sim \mathcal{N}(\mu + \tau_2, \sigma^2/n_2).$$

The corresponding difference estimator satisfies

$$\begin{aligned}\tilde{\tau}_1 - \tilde{\tau}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} &\sim \mathcal{N}\left((\mu + \tau_1) - (\mu + \tau_2), \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \\ &\sim \mathcal{N}\left(\tau_1 - \tau_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).\end{aligned}$$

The variance estimates within treatments are

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2.$$

Under normal errors, their random counterparts are independent and satisfy

$$\frac{(n_i - 1)\tilde{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}, \quad i = 1, 2,$$

and hence

$$\frac{(n_1-1)\tilde{\sigma}_1^2}{\sigma^2}+\frac{(n_2-1)\tilde{\sigma}_2^2}{\sigma^2}=\frac{(n_1-1)\tilde{\sigma}_1^2+(n_2-1)\tilde{\sigma}_2^2}{\sigma^2}\sim\chi^2_{n_1+n_2-2}.$$

With

$$\hat{\sigma}^2=\frac{1}{n_1+n_2-2}\left[(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2\right],$$

we have

$$\frac{(n_1+n_2-2)\tilde{\sigma}^2}{\sigma^2}\sim\chi^2_{n_1+n_2-2}.$$

Under normal errors, the pivotal statistic is

$$\frac{(\tilde{\tau}_1-\tilde{\tau}_2)-(\tau_1-\tau_2)}{\tilde{\sigma}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t_{n_1+n_2-2}. \tag{0.6}$$

The pooled two-sample procedure requires independent normal errors with equal variances. Normality may be assessed with a quantile–quantile plot of the residuals

$$r_{ij}=y_{ij}-\bar{y}_{i\bullet},\qquad\text{and}\qquad\hat{\sigma}^2=\frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}r_{ij}^2.$$

Side-by-side data or residual plots help assess the equal variance assumption; a formal variance ratio test is developed next.

If $X\sim\chi^2_n$ and $Y\sim\chi^2_m$ are independent, then

$$F=\frac{X/n}{Y/m}\sim F_{n,m}.$$

Remark 0.25 If $F\sim F_{n,m}$, then $1/F\sim F_{m,n}$.

Under the equal variance null hypothesis, the two independent scaled sample variances have the chi-square distributions displayed above. Therefore,

$$\frac{\left(\dfrac{(n_1-1)\tilde{\sigma}_1^2}{\sigma^2}\right)/(n_1-1)}{\left(\dfrac{(n_2-1)\tilde{\sigma}_2^2}{\sigma^2}\right)/(n_2-1)}=\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_2^2}\sim F_{n_1-1,n_2-1}.$$

The observed variance ratio is $F_{\text{obs}} = \hat{\sigma}_1^2 / \hat{\sigma}_2^2$.

The F-distribution is supported on the positive real line and is generally right-skewed. Figure 16 shows the two rejection regions for a two-sided variance ratio test.

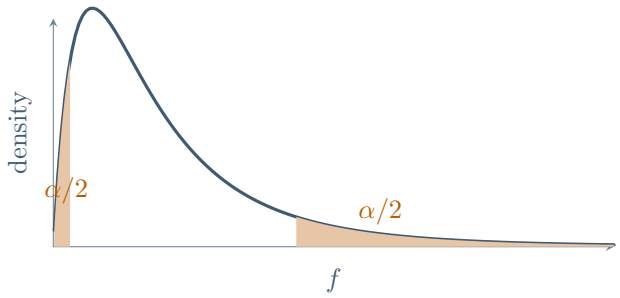


FIGURE 16

For a significance level α, we reject $H_0 : \sigma_1^2/\sigma_2^2 = 1$ against $H_A : \sigma_1^2/\sigma_2^2 \neq 1$ when

$$F_{\text{obs}} > F_{1-\alpha/2,(n_1-1,n_2-1)} \qquad \text{or} \qquad F_{\text{obs}} < F_{\alpha/2,(n_1-1,n_2-1)}.$$

When the variances differ, a large sample normal approximation gives

$$\frac{(\tilde{\tau}_1 - \tilde{\tau}_2) - (\tau_1 - \tau_2)}{\sqrt{\dfrac{\tilde{\sigma}_1^2}{n_1} + \dfrac{\tilde{\sigma}_2^2}{n_2}}} \overset{\text{approximately}}{\sim} \mathcal{N}(0,1).$$

For smaller samples, the Welch procedure replaces the normal reference distribution by

$$\frac{(\tilde{\tau}_1 - \tilde{\tau}_2) - (\tau_1 - \tau_2)}{\sqrt{\dfrac{\tilde{\sigma}_1^2}{n_1} + \dfrac{\tilde{\sigma}_2^2}{n_2}}} \overset{\text{approximately}}{\sim} t_K,$$

where the **Welch–Satterthwaite degrees of freedom** are estimated by

$$K = \frac{\left(\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{\hat{\sigma}_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{\hat{\sigma}_2^2}{n_2}\right)^2}.$$

The value of K will not generally be an integer. It may be used directly in software; when only integer critical value tables are available, rounding down gives a conservative procedure.

Remark 0.26 The preliminary variance ratio test should not be used mechanically to choose between the pooled and Welch procedures. Such a two-stage rule does not preserve the advertised significance level. When equal variances are not justified by the design or subject matter knowledge, the Welch procedure is generally the safer comparison.

The **Poggendorff illusion** provides a concrete setting for a designed experiment. A participant adjusts a response marker to indicate where an interrupted diagonal appears to continue; the signed displacement d is the response. A representative configuration appears in Figure 17.

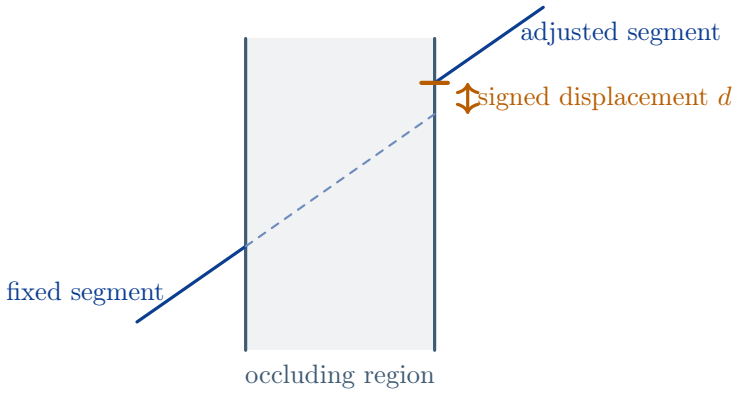


FIGURE 17

The geometry in Figure 18 isolates the visual elements that can be varied experimentally.

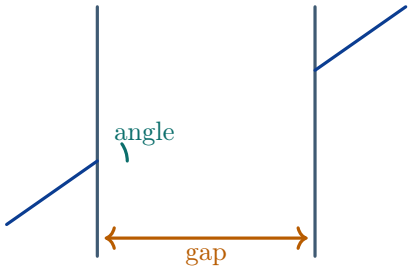


FIGURE 18

1. The distance between the parallel lines, illustrated together with the diagonal angle in Figure 19.

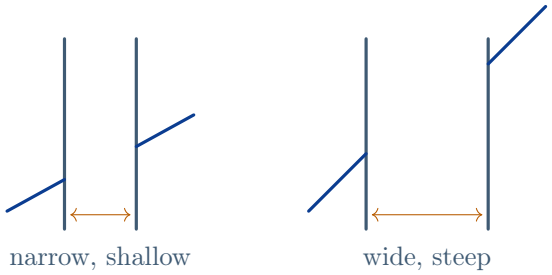


Figure 19

2. The angle at which the diagonal meets the parallel lines.
3. The orientation of the entire configuration.

Alternative orientations and response markings are displayed in Figure 20.

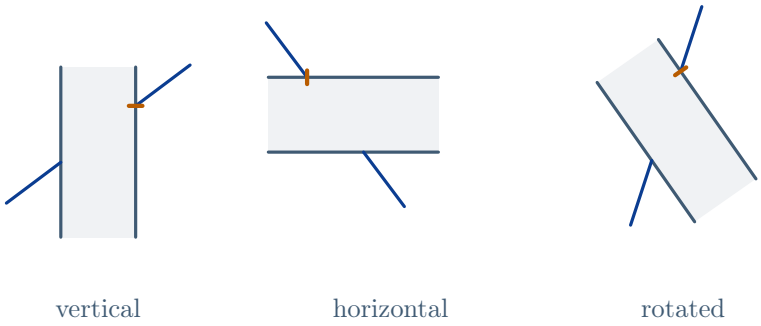


Figure 20

The levels are chosen as follows.

1. Use two levels for each geometric factor.
2. Randomize the order in which a participant sees the configurations. In software, a reproducible randomization may be generated with `sample` after setting a seed with `set.seed`.
3. Hold orientation fixed when it is not a factor of interest.
4. Record whether the participant wears glasses as a two-level subject characteristic.

Comparisons of Two Treatments with Blocking

The gardener next compares fertilizers A and B using six pots that hold two plants each. Each pot is a block, and one plant in every pot receives each fertilizer. The fertilizer positions are randomized independently within pots, as in Figure 21.

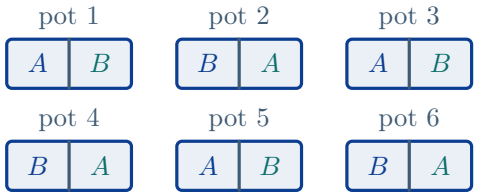


FIGURE 21

For treatment i in block j, consider the additive model

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}, \quad i = 1, 2, \quad j = 1, \ldots, n,$$

where $n = 6$ in the fertilizer experiment. The errors are mutually independent, with $\mathbb{E}[R_{ij}] = 0$ and $\text{Var}(R_{ij}) = \sigma^2$. Normality is imposed when exact small sample inference is required.

The two constraints are

$$\tau_1 + \tau_2 = 0 \qquad \text{and} \qquad \sum_{j=1}^{n} \beta_j = 0.$$

Here τ_i is the treatment effect and β_j is the block effect. The model is additive: the difference between the expected responses to the two treatments is the same in every block. A treatment-by-block interaction would violate this assumption.

Define $D_j = Y_{1j} - Y_{2j}$, the difference from each block j.

$$D_j = (\mu + \tau_1 + \beta_j + R_{1j}) - (\mu + \tau_2 + \beta_j + R_{2j}) = \tau_1 - \tau_2 + E_j,$$

where $E_j = R_{1j} - R_{2j}$, so $\mathbb{E}[E_j] = 0$ and $\text{Var}(E_j) = 2\sigma^2$.

The observed data are paired by block:

$$\{(y_{11}, y_{21}), (y_{12}, y_{22}), \ldots, (y_{1n}, y_{2n})\}.$$

The least squares point estimates are obtained by minimizing

$$\min_{\mu,\tau_1,\tau_2,\beta_1,\ldots,\beta_n} \sum_{i=1}^{2}\sum_{j=1}^{n}(y_{ij}-\mu-\tau_i-\beta_j)^2 \quad \text{subject to} \quad \begin{cases}\tau_1+\tau_2=0,\\ \sum_{j=1}^{n}\beta_j=0.\end{cases}$$

Equivalently, minimize the Lagrangian

$$L(\mu,\tau_1,\tau_2,\beta_1,\ldots,\beta_n,\lambda_1,\lambda_2)=\sum_{i=1}^{2}\sum_{j=1}^{n}(y_{ij}-\mu-\tau_i-\beta_j)^2 + \lambda_1(\tau_1+\tau_2)+\lambda_2\left(\sum_{j=1}^{n}\beta_j\right).$$

Solving the constrained least squares problem gives

$$\begin{aligned}\hat{\mu}&=\bar{y}_{\bullet\bullet},\\ \hat{\tau}_1&=\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet},\\ \hat{\tau}_2&=\bar{y}_{2\bullet}-\bar{y}_{\bullet\bullet},\\ \hat{\beta}_j&=\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet}.\end{aligned}$$

The estimated treatment difference is

$$\begin{aligned}\hat{\tau}_1-\hat{\tau}_2&=\bar{y}_{1\bullet}-\bar{y}_{2\bullet}\\ &=\frac{1}{n}\sum_{j=1}^{n}y_{1j}-\frac{1}{n}\sum_{j=1}^{n}y_{2j}\\ &=\frac{1}{n}\sum_{j=1}^{n}(y_{1j}-y_{2j})\\ &=\frac{1}{n}\sum_{j=1}^{n}d_j\\ &=\bar{d}.\end{aligned}$$

The number of free parameters is

$$\begin{aligned}&\text{overall mean} + \text{treatment effects} - \text{treatment constraint}\\ &\qquad + \text{block effects} - \text{block constraint}\\ &=1+2-1+n-1\\ &=n+1.\end{aligned}$$

Thus the residual degrees of freedom are $2n - (n+1) = n - 1$, and the error variance estimate is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{2n-(n+1)} \sum_{i=1}^{2} \sum_{j=1}^{n} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{2} \sum_{j=1}^{n} \left[y_{ij} - \bar{y}_{\bullet\bullet} - (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) - (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) \right]^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{2} \sum_{j=1}^{n} \left[y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet} \right]^2 .\end{aligned}$$

The residuals in block j simplify to

$$\begin{aligned} r_{1j} &= y_{1j} - \bar{y}_{1\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet} = \frac{d_j - \bar{d}}{2}, \\ r_{2j} &= y_{2j} - \bar{y}_{2\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet} = -\frac{d_j - \bar{d}}{2}. \end{aligned}$$

Consequently,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-1} \sum_{j=1}^{n} \left(r_{1j}^2 + r_{2j}^2 \right) \\ &= \frac{1}{2(n-1)} \sum_{j=1}^{n} (d_j - \bar{d})^2 = \frac{\hat{\sigma}_d^2}{2},\end{aligned}$$

where $\hat{\sigma}_d^2$ is the sample variance of the paired differences.

The treatment averages satisfy

$$\mathbb{E}[\bar{Y}_{i\bullet}] = \mu + \tau_i \qquad \text{and} \qquad \operatorname{Var}(\bar{Y}_{i\bullet}) = \frac{\sigma^2}{n},$$

because the block effects average to zero under the constraint. Under normal errors, the estimator of the treatment difference therefore satisfies

$$\tilde{\tau}_1 - \tilde{\tau}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} = \bar{D} \sim \mathcal{N}\left(\tau_1 - \tau_2, \frac{2\sigma^2}{n} \right).$$

Independently,

$$\frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \text{and} \qquad \frac{(n-1)\tilde{\sigma}_d^2}{2\sigma^2} \sim \chi^2_{n-1}.$$

The test statistic is

$$\frac{(\tilde{\tau}_1 - \tilde{\tau}_2) - (\tau_1 - \tau_2)}{\tilde{\sigma}\sqrt{\frac{2}{n}}} \sim t_{n-1}. \tag{0.7}$$

Thus the blocked comparison of two treatments is exactly a one sample problem for the paired differences. Writing $\delta = \tau_1 - \tau_2$ and $\sigma_d^2 = 2\sigma^2$, the required model for exact small sample inference is

$$D_j = \delta + E_j, \quad j = 1, \ldots, n \qquad \text{and} \qquad E_j \sim \mathcal{N}(0, \sigma_d^2),$$

with independent errors $E_1, \ldots, E_n$.

The corresponding confidence interval is

$$\hat{\tau}_1 - \hat{\tau}_2 \pm t_{1-\alpha/2,n-1}\hat{\sigma}\sqrt{\frac{2}{n}} = \bar{d} \pm t_{1-\alpha/2,n-1}\frac{\hat{\sigma}_d}{\sqrt{n}}.$$

The unblocked pivot in (0.6) compares variation between two independent groups, whereas the paired pivot in (0.7) uses only differences within blocks. Effective blocking reduces the latter standard error by removing stable differences among blocks.

5. Analysis of Variance, Randomized Blocks, and Factorial Designs

Analysis of variance extends comparisons of two treatments to several treatments and partitions observed variation into interpretable sources. Randomized blocks account for nuisance variation, while factorial designs reveal both main effects and interactions.

Comparing Many Treatments

There are two main types of designs for comparing many treatments. A completely randomized design has no blocking. A randomized block design uses blocking, and each treatment is repeated once per block.

Completely Randomized Design

The gardener wants to compare the yield from using fertilizers A, B, and C on 12 plants.

The questions of interest are whether there are any differences among the treatments and which fertilizer is best.

A design is **balanced** if every treatment is applied to the same number of experimental units.

For treatment i and replicate j, consider the model

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, r.$$

In the fertilizer example, $t = 3$ and $r = 4$.

Here μ is the overall long-run average yield of tomatoes from each plant across the three fertilizers, and τ_i is the treatment effect. The treatment effects satisfy the constraint

$$\sum_{i=1}^{t} \tau_i = 0.$$

The errors are independent, have mean zero, and have common variance σ^2. Normality is additionally assumed for exact analysis-of-variance tests and confidence intervals.

The observed responses are indexed by treatment and replicate:

$$\{y_{ij} : i = 1, \ldots, t,\ j = 1, \ldots, r\}.$$

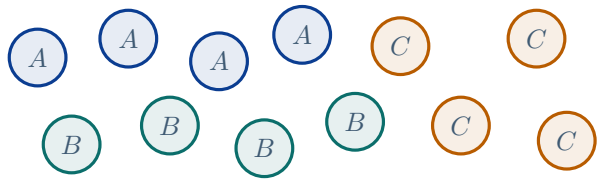
The point estimates are obtained by minimizing

$$\min_{\mu, \tau_1, \ldots, \tau_t} \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \mu - \tau_i)^2 \quad \text{subject to} \quad \sum_{i=1}^{t} \tau_i = 0.$$

Solving the constrained least squares problem gives

$$\hat{\mu} = \bar{y}_{\bullet\bullet} \qquad \text{and} \qquad \hat{\tau}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}.$$

Figure 22 shows twelve experimental units in an irregular spatial layout. A completely randomized design assigns four units to each fertilizer without using location to constrain the assignment.



one randomized assignment: four units per treatment

FIGURE 22

The difference between any two treatments is

$$\hat{\tau}_i - \hat{\tau}_k = \bar{y}_{i\bullet} - \bar{y}_{k\bullet}.$$

The number of free parameters is

$$\text{overall mean} + \text{treatment effects} - \text{treatment constraint} = 1 + t - 1 = t.$$

The residual degrees of freedom are therefore $rt - t = t(r-1)$, and an estimate of the error variance is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{rt-t}\sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 \\ &= \frac{1}{t(r-1)}\sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{i\bullet})^2.\end{aligned}$$

The treatment averages satisfy

$$\mathbb{E}[\bar{Y}_{i\bullet}] = \mu + \tau_i \qquad \text{and} \qquad \operatorname{Var}(\bar{Y}_{i\bullet}) = \frac{\sigma^2}{r}.$$

Under normal errors, $\bar{Y}_{i\bullet} \sim \mathcal{N}(\mu + \tau_i, \sigma^2/r)$.

For the variance estimator,

$$\frac{t(r-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{t(r-1)}.$$

This variance estimator is independent of every contrast estimator.

Definition 0.27 A **contrast** θ is any linear combination of the treatment effects for which the coefficients sum to zero:

$$\theta = \sum_{i=1}^{t} a_i \tau_i \quad \text{such that} \quad \sum_{i=1}^{t} a_i = 0.$$

Example 0.28 The contrast $\tau_1 - \tau_2$ is the difference between treatments 1 and 2. Its coefficients sum to zero:

$$a_1 + a_2 + a_3 + \cdots + a_t = 1 - 1 + 0 + \cdots + 0 = 0.$$

Example 0.29 The contrast $\tau_1 - (\tau_2 + \tau_3)/2$ compares the effect of treatment 1 with the average effect of treatments 2 and 3.

Example 0.30 The coefficients for $\tau_1 - (\tau_2 + \tau_3)/2$ satisfy the contrast condition:

$$a_1 + a_2 + a_3 + a_4 + \cdots + a_t = 1 - 1/2 - 1/2 + 0 + \cdots + 0 = 0.$$

The point estimate of a contrast is

$$\hat{\theta} = \sum_{i=1}^{t} a_i \hat{\tau}_i = \sum_{i=1}^{t} a_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) = \sum_{i=1}^{t} a_i \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} \underbrace{\sum_{i=1}^{t} a_i}_{=0} = \sum_{i=1}^{t} a_i \bar{y}_{i\bullet}.$$

The corresponding estimator is

$$\tilde{\theta} = \sum_{i=1}^{t} a_i \tilde{\tau}_i = \sum_{i=1}^{t} a_i \bar{Y}_{i\bullet}.$$

It is unbiased for the contrast:

$$\mathbb{E}\left[\tilde{\theta}\right] = \mathbb{E}\left[\sum_{i=1}^{t} a_i \bar{Y}_{i\bullet}\right] = \sum_{i=1}^{t} a_i \mathbb{E}\left[\bar{Y}_{i\bullet}\right] = \sum_{i=1}^{t} a_i (\mu + \tau_i) = \sum_{i=1}^{t} a_i \tau_i.$$

Its variance is

$$\operatorname{Var}(\tilde{\theta}) = \operatorname{Var}\left(\sum_{i=1}^{t} a_i \bar{Y}_{i\bullet}\right) = \sum_{i=1}^{t} a_i^2 \operatorname{Var}(\bar{Y}_{i\bullet}) = \sum_{i=1}^{t} a_i^2 \frac{\sigma^2}{r} = \frac{\sigma^2}{r} \sum_{i=1}^{t} a_i^2.$$

Under normal errors,

$$\tilde{\theta} \sim \mathcal{N}\left(\sum_{i=1}^{t} a_i \tau_i, \frac{\sigma^2}{r} \sum_{i=1}^{t} a_i^2\right).$$

The test statistic is

$$\frac{\sum_{i=1}^{t} a_i \tilde{\tau}_i - \sum_{i=1}^{t} a_i \tau_i}{\frac{\tilde{\sigma}}{\sqrt{r}} \cdot \sqrt{\sum_{i=1}^{t} a_i^2}} \sim t_{t(r-1)}.$$

This pivot can be used to construct a confidence interval or perform a hypothesis test for any prespecified contrast.

Analysis of Variance

The analysis of variance question asks whether there are any differences among the fertilizers:

$$\tau_1 = \tau_2 = \tau_3 = \cdots = \tau_t.$$

A single t-statistic tests only one linear restriction, so when $t > 2$ it cannot test equality of all treatment means at once.

A t-test has one equality constraint:

$$H_0 : \tau_1 - \tau_2 = 0 \qquad \text{and} \qquad H_A : \tau_1 - \tau_2 \neq 0.$$

By contrast, this analysis of variance test has $t - 1$ equality constraints:

$$\begin{aligned} H_0 : \tau_1 - \tau_2 &= 0, \\ \tau_2 - \tau_3 &= 0, \\ &\vdots \\ \tau_{t-1} - \tau_t &= 0. \end{aligned} \qquad H_A : \text{at least one constraint fails},$$

Equivalently,

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t \qquad \text{and} \qquad H_A : \text{at least one } \tau_i \text{ is not equal to the others.}$$

The new test statistic is based on estimators of the variance.

The first variance estimate is $\tilde{\sigma}^2$. This estimator is unbiased for σ^2 for all values of the treatment effects $\tau_1, \ldots, \tau_t$, even if all treatment effects are equal to zero. Thus it is unbiased for σ^2 whether the null hypothesis is true or false.

The second variance estimate uses the treatment means. If the null hypothesis is true, then

$$\tau_1 = \tau_2 = \cdots = \tau_t = 0.$$

The equality of all treatment effects, together with the constraint that they sum to zero, implies that every τ_i is zero. Thus $\mu + \tau_i = \mu$, and under normal errors,

$$\bar{Y}_{i\bullet} \sim \mathcal{N}(\mu, \sigma^2/r), \quad i = 1, \ldots, t.$$

The variation among the observed treatment means is measured by

$$\frac{\sum_{i=1}^{t}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}{t-1} = \frac{\sum_{i=1}^{t}\hat{\tau}_i^2}{t-1}.$$

Consequently,

$$r\frac{\sum_{i=1}^{t}\tilde{\tau}_i^2}{t-1}$$

is an unbiased estimator of σ^2 when the null hypothesis is true.

Proposition 0.31 For a balanced completely randomized design,

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{treat}} + \mathrm{SS}_{\mathrm{res}},$$

where

$$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{\bullet\bullet})^2,$$

$$\mathrm{SS}_{\mathrm{treat}} = r\sum_{i=1}^{t}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2,$$

$$\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{i\bullet})^2.$$

Proof. For each observation, add and subtract its treatment average and then expand:

$$\begin{aligned}\sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{\bullet\bullet})^2 &= \sum_{i=1}^{t}\sum_{j=1}^{r}(\underbrace{y_{ij}-\bar{y}_{i\bullet}}_{a}+\underbrace{\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet}}_{b})^2 \\ &= \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i\bullet})^2 + 2\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})\underbrace{\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i\bullet})}_{=0} + \sum_{i=1}^{t}\sum_{j=1}^{r}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 \\ &= \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i\bullet})^2 + r\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2.\end{aligned}$$

The cross term vanishes because the residuals sum to zero within each treatment. The last line is precisely the claimed decomposition. □

Under the null hypothesis of no treatment effect, the treatment sum of squares gives the chi-square statistic

$$\frac{r\sum_{i=1}^{t}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}{\sigma^2} = \frac{r\sum_{i=1}^{t}\tilde{\tau}_i^2}{\sigma^2} \sim \chi^2_{t-1}.$$

The residual sum of squares gives the independent chi-square statistic

$$\frac{t(r-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{t(r-1)}.$$

Remark 0.32 The two estimators above are independent.

The test statistic is

$$\frac{\left(\dfrac{r\sum_{i=1}^{t}\tilde{\tau}_i^2}{\sigma^2}\right)/(t-1)}{\left(\dfrac{t(r-1)\tilde{\sigma}^2}{\sigma^2}\right)/\big[t(r-1)\big]} \sim F_{t-1,\,t(r-1)}.$$

Writing a mean square as its sum of squares divided by its degrees of freedom gives the following analysis of variance table.

Source	Sum of squares	Degrees of freedom	Mean square	F_{obs}
Treatments	SS_{treat}	$t-1$	MS_{treat}	$\text{MS}_{\text{treat}} / \text{MS}_{\text{res}}$
Residuals	SS_{res}	$t(r-1)$	MS_{res}	
Total	SS_{total}	$tr-1$		

We can show that

$$\begin{aligned}\mathbb{E}\left[\text{MS}_{\text{treat}}\right] &= \mathbb{E}\left[\frac{r\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2}{t-1}\right] \\ &= \sigma^2 + r\frac{\sum_{i=1}^{t}\tau_i^2}{t-1}.\end{aligned}$$

Thus, if at least one treatment effect is nonzero, the treatment mean square tends to exceed the residual mean square and the resulting F-ratio tends to be large.

For a chosen significance level α, the upper-tail critical value c satisfies

$$\mathbb{P}(F_{t-1,\,t(r-1)} \geq c) = \alpha.$$

The p-value is

$$p\text{-value} = \mathbb{P}(F_{t-1,\,t(r-1)} \geq F_{\text{obs}}).$$

The analysis requires independent, normally distributed errors with a common variance. Normality is assessed from the residuals $r_{ij} = y_{ij} - \bar{y}_{i\bullet}$, while side-by-side residual plots help assess whether the variance is constant across treatments.

Exercise 0.33 In a small investigation, three treatments A, B, and C were compared by assigning six units at random to each treatment. The data are shown below.

	1	2	3	4	5	6	Average	Standard deviation
A	11.82	13.03	10.78	14.31	14.21	8.56	12.12	2.22
B	15.46	12.90	14.88	13.75	18.59	13.80	14.90	2.02
C	15.43	14.28	14.76	12.07	13.80	13.46	13.97	1.16

The overall sample average is 13.66, and the sample standard deviation of all eighteen observations is 2.11.

1. State the model and define any parameters introduced.
2. Complete the analysis of variance table and determine, at level 0.05, whether there is evidence of a difference among the treatments.
3. Treatment A is the control. Test whether the average effect of treatments B and C exceeds the control effect, and explain why a one-sided test is appropriate.
4. Find a 95% confidence interval for the difference between the effects of treatments B and C.

Solution. For (1), the completely randomized model is

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad i = 1, 2, 3, \quad j = 1, \ldots, 6,$$

where μ is the grand mean, τ_i is the effect of treatment i, and $\tau_1 + \tau_2 + \tau_3 = 0$. The errors are independent with mean zero and common variance σ^2; normality is required for the exact tests below.

For (2), the residual sum of squares can be recovered from the three within-treatment standard deviations:

$$\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{3} \sum_{j=1}^{6} (y_{ij} - \bar{y}_{i\bullet})^2 = 51.764750.$$

Direct calculation from the eighteen observations gives $\mathrm{SS}_{\mathrm{total}} = 75.765494$, so $\mathrm{SS}_{\mathrm{treat}} = 24.000744$. Hence

Source	Sum of squares	Degrees of freedom	Mean square	F_{obs}
Treatments	24.000744	2	12.000372	3.477
Residuals	51.764750	15	3.450983	
Total	75.765494	17		

Since the 0.95 quantile of $F_{2,15}$ is approximately 3.68, the observed value does not exceed the critical value. Equivalently, the p-value is approximately 0.0574. At level 0.05, there is insufficient evidence that the treatment means differ.

For (3), the planned contrast is

$$\theta = -\tau_1 + \frac{\tau_2}{2} + \frac{\tau_3}{2}, \qquad \text{and} \qquad (a_1, a_2, a_3) = (-1, 1/2, 1/2).$$

The one-sided alternative $H_A : \theta > 0$ expresses the scientific question that the average of the two new treatments exceeds the control. The estimate and its standard error are

$$\hat{\theta} = -12.1183 + \frac{14.8967}{2} + \frac{13.9667}{2} = 2.3133,$$
$$\text{standard error}(\hat{\theta}) = \sqrt{3.450983\left(\frac{1 + 1/4 + 1/4}{6}\right)} \approx 0.9288.$$

Thus $t_{\text{obs}} \approx 2.491$ with 15 degrees of freedom, giving a one-sided p-value of approximately 0.0125. There is evidence that the average effect of treatments B and C exceeds the control effect.

For (4), take $\theta = \tau_2 - \tau_3$. The estimate is $14.90 - 13.97 = 0.93$, and

$$\text{standard error}(\hat{\theta}) = \sqrt{3.450983\left(\frac{1^2 + (-1)^2}{6}\right)} \approx 1.073.$$

Since $t_{0.975,15} \approx 2.131$, the confidence interval is

$$0.93 \pm 2.131(1.073) \approx (-1.36, 3.22).$$

The interval contains zero, which is consistent with no difference between treatments B and C. □

In an unbalanced completely randomized design, treatment i has r_i replicates. The model is

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, r_i.$$

The identifying constraint is

$$\sum_{i=1}^{t} r_i \tau_i = 0.$$

The quantity $\mu + \tau_i$ is the mean response under treatment i, and τ_i is its treatment effect. Under the weighted constraint,

$$\mu = \frac{\sum_{i=1}^{t} r_i \mathbb{E}[Y_{ij}]}{\sum_{i=1}^{t} r_i},$$

so μ is the average of the treatment means weighted by their replicate counts.

The parameter estimates are

$$\hat{\mu} = \bar{y}_{\bullet\bullet} \qquad \text{and} \qquad \hat{\tau}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}.$$

The degrees of freedom for the residuals are

$$\sum_{i=1}^{t} r_i - (t + 1 - 1) = \sum_{i=1}^{t} r_i - t = \sum_{i=1}^{t} (r_i - 1).$$

Define

$$\begin{aligned} \text{SS}_{\text{treat}} &= \sum_{i=1}^{t} r_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2, \\ \text{SS}_{\text{res}} &= \sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i\bullet})^2, \\ \text{SS}_{\text{total}} &= \sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2. \end{aligned}$$

The analysis of variance table is then

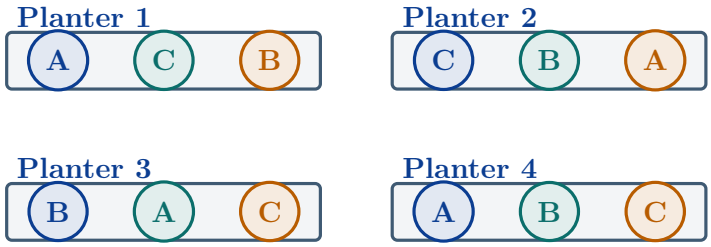
Source	Sum of squares	Degrees of freedom	Mean square	F_{obs}
Treatments	SS_{treat}	$t - 1$	$\text{SS}_{\text{treat}} / (t-1)$	$\text{MS}_{\text{treat}} / \text{MS}_{\text{res}}$
Residuals	SS_{res}	$\sum_{i=1}^{t} (r_i - 1)$	$\text{SS}_{\text{res}} / \sum_{i=1}^{t} (r_i - 1)$	
Total	SS_{total}	$\sum_{i=1}^{t} r_i - 1$		

The errors must be independent and normally distributed with a common variance. Residual plots and normal quantile–quantile plots are used to assess these conditions.

Randomized Block Designs

In a randomized block design, each treatment is repeated once in each block.

Example 0.34 The gardener wants to compare yield from three different fertilizers, A, B, and C. The gardener has four planters, and each planter holds three plants. The four planters are blocks in Figure 23; the three fertilizers are randomly assigned to positions within each planter.



Each block receives every fertilizer once; assignments are randomized independently within blocks.

FIGURE 23

Example 0.35 A baker wants to compare three cookie recipes. The ingredients arrive in batches, and three recipes can be made per batch.

In the notation below, there are t treatments and b blocks. Allocation is randomized independently within each block, and the plan is balanced. Blocking is useful when units within a block are expected to be more alike than units from different blocks: the analysis removes systematic block-to-block variation before comparing treatments.

The questions of interest are whether there are differences among the treatment effects and which treatment is best. We are not interested in differences among the blocks.

For treatment i in block j, the randomized block model is

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, b.$$

Here μ is the overall long-run average response across treatments and blocks, τ_i is the treatment effect, and β_j is the block effect. The constraints are

$$\sum_{i=1}^{t} \tau_i = 0 \qquad \text{and} \qquad \sum_{j=1}^{b} \beta_j = 0,$$

so the parameters are identifiable.

The errors R_{ij} are mutually independent with mean zero and common variance σ^2. Normality is additionally assumed for the exact small sample tests and confidence intervals below.

The model is additive: there is no interaction between treatments and blocks. Because there is only one observation in each treatment–block cell, such an interaction cannot be estimated separately and would be absorbed into the residual term.

The observed data are $\{y_{ij} : i = 1, \ldots, t \text{ and } j = 1, \ldots, b\}$.

The point estimates are

$$\begin{aligned}\hat{\mu} &= \bar{y}_{\bullet\bullet},\\ \hat{\tau}_i &= \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet},\\ \hat{\beta}_j &= \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}.\end{aligned}$$

Here the averages are

$$\begin{aligned}\bar{y}_{\bullet\bullet} &= \frac{1}{tb}\sum_{i=1}^{t}\sum_{j=1}^{b} y_{ij},\\ \bar{y}_{i\bullet} &= \frac{1}{b}\sum_{j=1}^{b} y_{ij},\\ \bar{y}_{\bullet j} &= \frac{1}{t}\sum_{i=1}^{t} y_{ij}.\end{aligned}$$

The residual degrees of freedom are

$$tb - (1 + t - 1 + b - 1) = tb - t - b + 1 = (t - 1)(b - 1).$$

Proposition 0.36 Define

$$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{\bullet\bullet})^2,$$

$$\mathrm{SS}_{\text{treat}} = b\sum_{i=1}^{t}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2,$$

$$\mathrm{SS}_{\text{block}} = t\sum_{j=1}^{b}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2,$$

$$\mathrm{SS}_{\text{res}} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})^2.$$

Then

$$\mathrm{SS}_{\text{total}} = \mathrm{SS}_{\text{treat}} + \mathrm{SS}_{\text{block}} + \mathrm{SS}_{\text{res}}\,. \tag{0.8}$$

Proof. The residual variance estimator is

$$\hat{\sigma}^2 = \frac{1}{(t-1)(b-1)}\sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j)^2.$$

$$(t-1)(b-1)\hat{\sigma}^2 = \sum_{i=1}^{t}\sum_{j=1}^{b}\left[(y_{ij} - \bar{y}_{\bullet\bullet}) - (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) - (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})\right]^2.$$

To expand the square, use

$$(a+b+c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc.$$

Thus,

$$\begin{aligned}(t-1)(b-1)\hat{\sigma}^2 = \sum_{i=1}^{t}\sum_{j=1}^{b}\Big[&\underbrace{(y_{ij} - \bar{y}_{\bullet\bullet})^2}_{(1)} + \underbrace{(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}_{(2)} + \underbrace{(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2}_{(3)} \\ &\underbrace{-2(y_{ij} - \bar{y}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})}_{(4)}\underbrace{-2(y_{ij} - \bar{y}_{\bullet\bullet})(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})}_{(5)} \\ &\underbrace{+2(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})}_{(6)}\Big].\end{aligned}$$

Each of the six terms in the expansion can be identified explicitly.

1. The first term is the total sum of squares:

$$\sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{\bullet\bullet})^2 = \mathrm{SS}_{\text{total}}\,.$$

2. The second term is the treatment sum of squares:

$$\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 = b\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 = \mathrm{SS}_{\text{treat}}\,.$$

3. The third term is the block sum of squares:

$$\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})^2 = t\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})^2 = \mathrm{SS}_{\text{block}}\,.$$

4. The fourth term simplifies to $-2\,\mathrm{SS}_{\text{treat}}$:

$$\begin{aligned}
&-2\sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{\bullet\bullet})(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})\\
&\quad=-2\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})\sum_{j=1}^{b}(y_{ij}-\bar{y}_{\bullet\bullet})\\
&\quad=-2b\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 = -2\,\mathrm{SS}_{\text{treat}}\,.
\end{aligned}$$

5. Similarly, the fifth term simplifies to $-2\,\mathrm{SS}_{\text{block}}$:

$$\begin{aligned}
&-2\sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{\bullet\bullet})(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})\\
&\quad=-2\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})\sum_{i=1}^{t}(y_{ij}-\bar{y}_{\bullet\bullet})\\
&\quad=-2t\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})^2 = -2\,\mathrm{SS}_{\text{block}}\,.
\end{aligned}$$

6. The sixth term is zero because both centered marginal means sum to zero:

$$\begin{aligned} &2\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet}) \\ &\quad = 2\left[\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})\right]\left[\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})\right] = 0. \end{aligned}$$

Substituting (1), (2), (3), (4), (5), and (6) into the expansion gives

$$\begin{aligned} (t-1)(b-1)\hat{\sigma}^2 &= \mathrm{SS}_{\text{total}} + \mathrm{SS}_{\text{treat}} + \mathrm{SS}_{\text{block}} - 2\,\mathrm{SS}_{\text{treat}} - 2\,\mathrm{SS}_{\text{block}} \\ &= \mathrm{SS}_{\text{total}} - \mathrm{SS}_{\text{treat}} - \mathrm{SS}_{\text{block}} \\ &= \mathrm{SS}_{\text{res}}\,. \end{aligned}$$

Rearranging proves (0.8). □

Under the normal model, each treatment mean has the exact distribution

$$\bar{Y}_{i\bullet} \sim \mathcal{N}\left(\mu+\tau_i, \frac{\sigma^2}{b}\right).$$

For any contrast $\theta = \sum_{i=1}^{t} a_i\tau_i$ whose coefficients satisfy $\sum_{i=1}^{t} a_i = 0$, the corresponding estimator satisfies

$$\tilde{\theta} = \sum_{i=1}^{t} a_i\bar{Y}_{i\bullet} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{b}\sum_{i=1}^{t} a_i^2\right).$$

For the variance estimator,

$$\frac{(t-1)(b-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{(t-1)(b-1)}.$$

The variance estimator and any contrast estimator are independent.

If the hypothesis

$$\tau_1 = \cdots = \tau_t = 0$$

is true, then

$$\bar{Y}_{i\bullet} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{b}\right),$$

and therefore

$$\frac{b\sum_{i=1}^{t}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}{\sigma^2}\sim\chi^2_{t-1}.$$

This is independent of the other variance estimator.

For contrast test statistics,

$$\frac{\tilde{\theta}-\theta}{\tilde{\sigma}\sqrt{\frac{\sum_{i=1}^{t}a_i^2}{b}}}\sim t_{(t-1)(b-1)}.$$

For analysis of variance, ask whether there are differences among the treatments:

$$H_0:\tau_1=\cdots=\tau_t=0 \qquad \text{and} \qquad H_A:\tau_i\neq 0 \text{ for at least one } i.$$

The observed F-statistic is

$$F_{\text{obs}}=\frac{b\sum_{i=1}^{t}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2/(t-1)}{\hat{\sigma}^2}=\frac{\text{MS}_{\text{treat}}}{\text{MS}_{\text{res}}}.$$

Under the null hypothesis, the corresponding random statistic has the $F_{t-1,(t-1)(b-1)}$ distribution. Using the sums of squares in Proposition 0.36, define

$$\text{MS}_{\text{treat}}=\frac{\text{SS}_{\text{treat}}}{t-1}, \qquad \text{MS}_{\text{block}}=\frac{\text{SS}_{\text{block}}}{b-1} \qquad \text{and} \qquad \text{MS}_{\text{res}}=\frac{\text{SS}_{\text{res}}}{(t-1)(b-1)}.$$

The analysis of variance table is

Source	Sum of squares	Degrees of freedom	Mean square	F_{obs}
Treatments	SS_{treat}	$t-1$	MS_{treat}	$\text{MS}_{\text{treat}}\,/\,\text{MS}_{\text{res}}$
Blocks	SS_{block}	$b-1$	MS_{block}	
Residuals	SS_{res}	$(t-1)(b-1)$	MS_{res}	
Total	SS_{total}	$tb-1$		

The errors must be independent and normally distributed with a common variance, and the treatment and block effects must be additive. Residual plots by treatment and by block help assess constant variance and reveal patterns that may indicate nonadditivity. Blocking is most effec-

tive when block differences explain substantial response variation; otherwise, spending degrees of freedom on blocks may provide little precision gain.

An **incomplete block design** is useful when there are too many treatments to include every treatment in each block. For example, four wines can be compared in tasting sessions of size three by using all $\binom{4}{3} = 4$ possible triples:

Block	Treatments
1	A, B, C
2	A, B, D
3	A, C, D
4	B, C, D

Every treatment occurs in three blocks, and every pair occurs together in two blocks, so this is a balanced incomplete block design.

A **Latin square** blocks in two directions. For example, four torch designs can be compared while controlling for both the batch of material and the individual who makes each torch:

	Individual			
Batch	1	2	3	4
1	A	B	C	D
2	D	A	B	C
3	C	D	A	B
4	B	C	D	A

Each treatment appears once in every row and once in every column. Thus treatment comparisons are separated from both blocking variables under the additive model.

Factorial Treatment Structure and Interaction

Treatments have a **factorial structure** when each treatment is defined by one level of each of two or more factors. An **interaction** occurs when the effect of one factor on the response depends on the level of another factor. Parallel profiles indicate no interaction; nonparallel profiles indicate interaction, as illustrated in Figure 24.

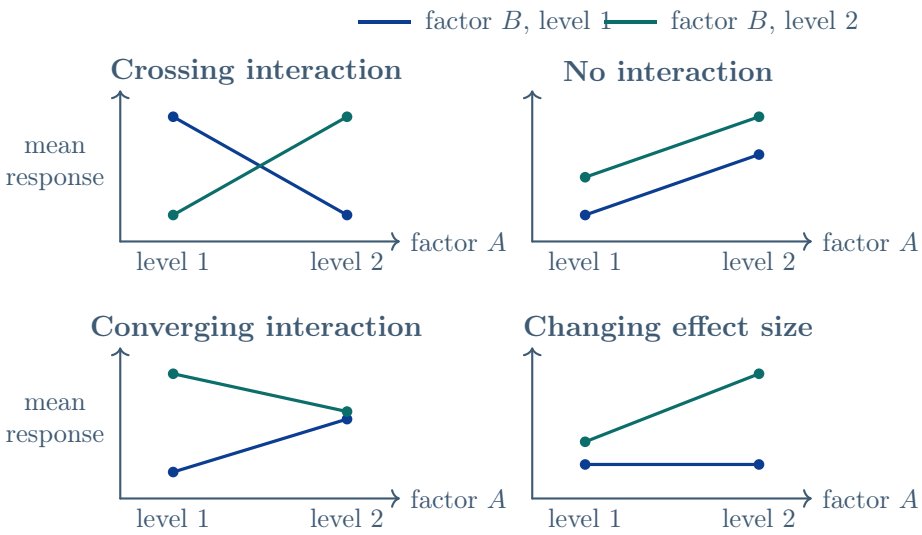


FIGURE 24

The same graphical criterion applies when a factor has more than two levels. Figure 25 compares parallel and nonparallel profiles in a 3×2 design.

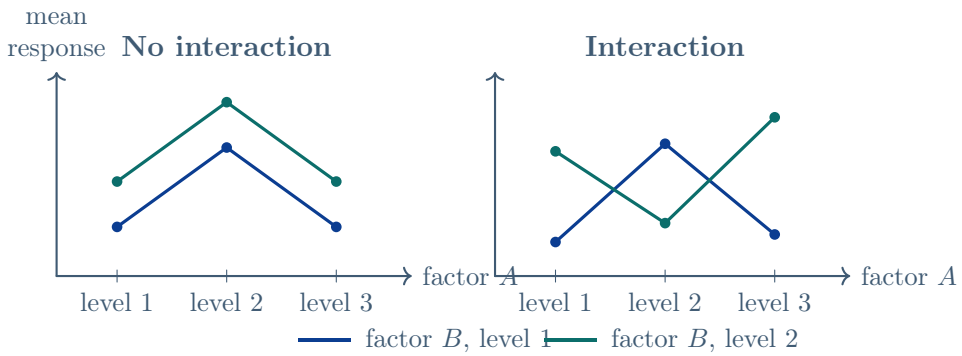


FIGURE 25

Consider a rose propagation experiment with two nutrient solutions and two cutting methods. There are eight independent replicates at each of the four treatment combinations. The observed cell means are

Treatment	Solution	Cutting method	Cell mean	Estimated treatment mean
1	1	1	11.06	$\hat{\mu} + \hat{\tau}_1$
2	1	2	12.65	$\hat{\mu} + \hat{\tau}_2$
3	2	1	13.19	$\hat{\mu} + \hat{\tau}_3$
4	2	2	11.09	$\hat{\mu} + \hat{\tau}_4$

The difference between cutting methods is $\tau_2 - \tau_1$ under solution 1 and $\tau_4 - \tau_3$ under solution 2. Their difference,

$$\theta = (\tau_4 - \tau_3) - (\tau_2 - \tau_1),$$

is the interaction contrast. Its coefficients are $(a_1, a_2, a_3, a_4) = (1, -1, -1, 1)$, which sum to zero, so θ is indeed a treatment contrast. The estimate is

$$\hat{\theta} = (\hat{\tau}_4 - \hat{\tau}_3) - (\hat{\tau}_2 - \hat{\tau}_1) = (11.09 - 13.19) - (12.65 - 11.06) = -3.69.$$

Its variance is

$$\mathrm{Var}(\tilde{\theta}) = \sigma^2 \frac{\sum_{i=1}^4 a_i^2}{r} = \sigma^2 \frac{4}{8} = \frac{\sigma^2}{2}.$$

The observed test statistic is

$$t_{\mathrm{obs}} = \frac{(\hat{\theta} - \theta_0)}{\hat{\sigma}\sqrt{1/2}} = \frac{-3.69}{\sqrt{2.674/2}} = -3.19.$$

Here 2.674 is $\hat{\sigma}^2$ from the output. We compare this value to T_{28}.

The null hypothesis $H_0 : \theta = 0$ states that there is no interaction.

The two-sided p-value is

$$\mathbb{P}(|T_{28}| \geq 3.19) \approx 0.0035.$$

Therefore, there is strong evidence of an interaction.

Arranging the same means by factor level makes the interaction visible: the preferred cutting method reverses between the two solutions.

	Cutting method 1	Cutting method 2	Marginal mean
Solution 1	11.06	12.65	11.86
Solution 2	13.19	11.09	12.14
Marginal mean	12.13	11.87	12.00

We now consider a factorial design using two factors and two levels, also called a 2^2 design. There are four treatments. The two factors A and B each have two levels, and there are n replicates of each treatment. Thus the number of observations is $4n$, and the design is balanced.

For a 2^2 factorial design, the model is

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + R_{ijk}, \quad i = 1, 2, \quad j = 1, 2, \quad k = 1, \ldots, n.$$

Here τ_i is the main effect of level i of factor A, β_j is the main effect of level j of factor B, and γ_{ij} is their interaction effect. The errors are mutually independent with mean zero and common variance σ^2; normality is assumed for exact analysis of variance inference. The mean at cell (i, j) is

$$\mu_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij}.$$

The constraints are

1. The main effects satisfy
$$\sum_{i=1}^{2} \tau_i = 0 \qquad \text{and} \qquad \sum_{j=1}^{2} \beta_j = 0.$$
2. The interaction effects sum to zero over either factor:
$$\sum_{i=1}^{2} \gamma_{ij} = 0 \quad \text{for all } j = 1, 2,$$
and
$$\sum_{j=1}^{2} \gamma_{ij} = 0 \quad \text{for all } i = 1, 2.$$

In a 2^2 design, these constraints become

1. The main effect constraints are $\tau_1 + \tau_2 = 0$ and $\beta_1 + \beta_2 = 0$.
2. The interaction table has zero row and column totals:

	Factor B	
Factor A	γ_{11}	γ_{12}
	γ_{21}	γ_{22}

Explicitly,

$$\gamma_{11} + \gamma_{21} = 0, \qquad \gamma_{12} + \gamma_{22} = 0,$$
$$\gamma_{11} + \gamma_{12} = 0, \qquad \gamma_{21} + \gamma_{22} = 0.$$

Equivalently, let

$$\boldsymbol{\gamma} = \begin{bmatrix}\gamma_{11} & \gamma_{21} & \gamma_{12} & \gamma_{22}\end{bmatrix}^{\mathsf{T}} \qquad \text{and} \qquad \boldsymbol{C} = \begin{bmatrix}1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\\ 1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\end{bmatrix}.$$

Then the interaction constraints are $\boldsymbol{C}\boldsymbol{\gamma} = \mathbf{0}$. The matrix $\boldsymbol{C}$ has rank 3, because the sum of its first two rows equals the sum of its last two rows.

Among the interaction constraints, the number that are independent is the number of levels of factor A plus the number of levels of factor B, minus one. One of the row- or column-sum constraints follows from the others.

The point estimates are

$$\begin{aligned}\hat{\mu} &= \bar{y}_{\cdots},\\ \hat{\tau}_i &= \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots},\\ \hat{\beta}_j &= \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots},\\ \hat{\gamma}_{ij} &= \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}.\end{aligned}$$

Consequently, each observation has the fitted decomposition

$$y_{ijk} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \hat{\gamma}_{ij} + r_{ijk},$$

where $r_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}$.

The conventional sign table below displays the four treatments and the contrasts that estimate the two main effects and the interaction. The identity column I estimates the overall mean.

Treatment combination	Factor levels (A, B)	Label	I	A	B	AB
A low, B low	$(-,-)$	(1)	$+$	$-$	$-$	$+$
A high, B low	$(+,-)$	a	$+$	$+$	$-$	$-$
A low, B high	$(-,+)$	b	$+$	$-$	$+$	$-$
A high, B high	$(+,+)$	ab	$+$	$+$	$+$	$+$

The A, B, and AB columns give the signs of the corresponding orthogonal contrasts. Orthogonality is valuable because, in a balanced design, these three comparisons partition the treatment sum of squares into distinct contributions from the two main effects and the interaction.

Factorial Design with Two Factors and Blocks

Suppose factor A has t_a levels, factor B has t_b levels, and every one of the $t_a t_b$ treatment combinations is applied once within each of b blocks. Thus there are $t_a t_b b$ observations. For example, six treatment combinations in five blocks produce 30 observations.

The additive block model with factorial treatment structure is

$$Y_{ijk} = \mu + \alpha_i + \lambda_j + \gamma_{ij} + \beta_k + R_{ijk}, \quad \begin{matrix} i=1,\ldots,t_a,\ j=1,\ldots,t_b, \\ k=1,\ldots,b. \end{matrix} \tag{0.9}$$

Here α_i and λ_j are main effects, γ_{ij} is the AB interaction, and β_k is a block effect. The errors are mutually independent with mean zero and common variance σ^2, and normality is required for exact small sample inference.

The model assumes that block effects are additive. With one observation for each treatment combination in each block, interactions between blocks and treatment factors cannot be estimated separately and are absorbed into the residual term.

The identifying constraints are

$$\sum_{i=1}^{t_a} \alpha_i = 0, \qquad \sum_{j=1}^{t_b} \lambda_j = 0, \qquad \sum_{k=1}^{b} \beta_k = 0,$$

$$\sum_{i=1}^{t_a} \gamma_{ij} = 0 \text{ for each } j, \qquad \sum_{j=1}^{t_b} \gamma_{ij} = 0 \text{ for each } i.$$

The least squares estimates are

$$\hat{\mu} = \bar{y}_{\cdots}, \qquad \hat{\alpha}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}, \qquad \hat{\lambda}_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots},$$

$$\hat{\beta}_k = \bar{y}_{\cdot\cdot k} - \bar{y}_{\cdots}, \qquad \hat{\gamma}_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}.$$

Define the treatment, main effect, and interaction sums of squares by

$$\mathrm{SS}_{\text{treat}} = b \sum_{i=1}^{t_a} \sum_{j=1}^{t_b} (\bar{y}_{ij\cdot} - \bar{y}_{\cdots})^2,$$

$$\mathrm{SS}_A = b t_b \sum_{i=1}^{t_a} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2, \qquad \mathrm{SS}_B = b t_a \sum_{j=1}^{t_b} (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2,$$

$$\mathrm{SS}_{AB} = b\sum_{i=1}^{t_a}\sum_{j=1}^{t_b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2.$$

Proposition 0.37 In the balanced factorial design with two factors,

$$\mathrm{SS}_{\text{treat}} = \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{AB}\,. \tag{0.10}$$

The corresponding degrees of freedom satisfy

$$t_a t_b - 1 = (t_a - 1) + (t_b - 1) + (t_a - 1)(t_b - 1).$$

Proof. For each cell, write

$$\begin{aligned}\bar{y}_{ij\cdot} - \bar{y}_{\cdots} = (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})\\ + (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}).\end{aligned}$$

After squaring and summing over i and j, all three cross product sums vanish. This follows from the zero sums of the centered marginal means and of the estimated interaction effects over either index. The three squared terms are precisely SS_A, SS_B, and SS_{AB}, which proves (0.10). The degrees-of-freedom identity follows by direct expansion. □

The remaining sums of squares are

$$\begin{aligned}\mathrm{SS}_{\text{block}} &= t_a t_b \sum_{k=1}^{b}(\bar{y}_{\cdot\cdot k} - \bar{y}_{\cdots})^2,\\ \mathrm{SS}_{\text{res}} &= \sum_{i=1}^{t_a}\sum_{j=1}^{t_b}\sum_{k=1}^{b}(y_{ijk} - \bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot k} + \bar{y}_{\cdots})^2,\\ \mathrm{SS}_{\text{total}} &= \sum_{i=1}^{t_a}\sum_{j=1}^{t_b}\sum_{k=1}^{b}(y_{ijk} - \bar{y}_{\cdots})^2.\end{aligned}$$

The residual degrees of freedom are

$$t_a t_b b - 1 - (t_a t_b - 1) - (b - 1) = (t_a t_b - 1)(b - 1).$$

With each mean square equal to its sum of squares divided by its degrees of freedom, the analysis of variance table is

Source	Sum of squares	Degrees of freedom	Mean square	F_{obs}
Treatments	SS_{treat}	$t_a t_b - 1$	MS_{treat}	$\text{MS}_{\text{treat}} \,/\, \text{MS}_{\text{res}}$
Factor A	SS_A	$t_a - 1$	MS_A	$\text{MS}_A \,/\, \text{MS}_{\text{res}}$
Factor B	SS_B	$t_b - 1$	MS_B	$\text{MS}_B \,/\, \text{MS}_{\text{res}}$
Interaction AB	SS_{AB}	$(t_a - 1)(t_b - 1)$	MS_{AB}	$\text{MS}_{AB} \,/\, \text{MS}_{\text{res}}$
Blocks	SS_{block}	$b - 1$	MS_{block}	
Residuals	SS_{res}	$(t_a t_b - 1)(b - 1)$	MS_{res}	
Total	SS_{total}	$t_a t_b b - 1$		

For the interaction test,

$$H_0 : \gamma_{ij} = 0 \text{ for all } i, j \qquad \text{and} \qquad H_A : \gamma_{ij} \neq 0 \text{ for at least one pair } (i, j),$$

and

$$F_{\text{obs}} = \frac{\text{MS}_{AB}}{\text{MS}_{\text{res}}} \sim F_{(t_a-1)(t_b-1),\,(t_a t_b-1)(b-1)} \quad \text{under } H_0.$$

The interaction should be examined before the main effects. When the interaction is substantial, a single average effect of factor A or factor B can conceal important differences across the levels of the other factor. If the interaction is negligible, the main effect tests use the same residual mean square in the denominator and the corresponding factor mean square in the numerator.

Factorial structure makes it possible to estimate several main effects and their interactions from one experiment. Figure 26 compares a factorial strategy with one-factor-at-a-time experiments.

Example 0.38 For the rose experiment, the control, or typical setting, is solution 1 with cutting method 1.

One strategy performs two separate two-arm experiments, changing one factor at a time from the control setting. Assigning two observations to each arm uses four control observations and two observations at each off-diagonal combination, for a total of eight. A factorial strategy instead assigns two observations to each of all four combinations, also using eight observations. Only the factorial design directly observes solution 2 with cutting method 2 and can estimate the interaction. The allocations are compared in Figure 26.

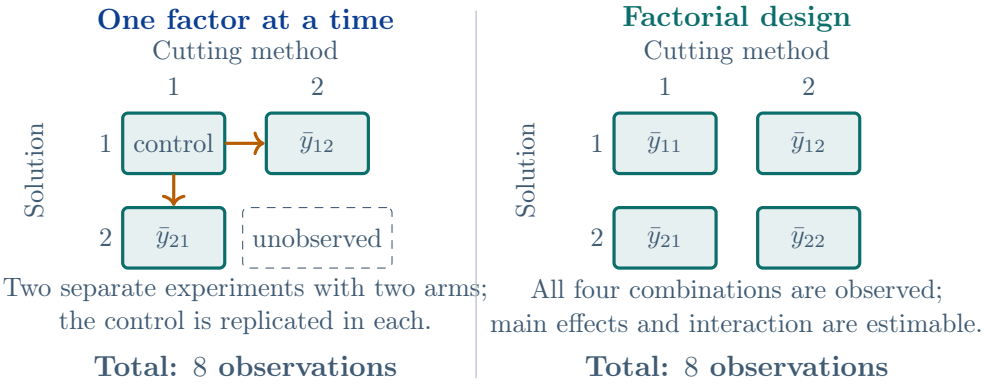


FIGURE 26

Confounding

Two effects are **confounded** when the allocation makes their contributions indistinguishable. For example, suppose that only the two diagonal combinations below are observed:

	Cutting method	
Solution	1	2
1	$\bar{y}_{11}$	—
2	—	$\bar{y}_{22}$

The difference $\bar{y}_{22} - \bar{y}_{11}$ changes both factors at once, so it cannot identify either factor's contribution or estimate their interaction. Randomization protects against uncontrolled variables, but it cannot recover comparisons absent from the design.

The number of combinations grows quickly: three factors with two levels require $2^3 = 8$ combinations, whereas five require $2^5 = 32$. A **fractional factorial design** deliberately aliases selected effects and runs only a chosen fraction. It is useful only when scientific assumptions make the aliased effects plausibly separable.

The **effect hierarchy principle** treats lower order effects as more plausible than high order interactions. Declaring a three-way or higher order interaction negligible is nevertheless a modelling assumption, not a consequence of the design. The alias structure must be reported so that every estimated effect has a clear interpretation.

Slight nonnormality is usually acceptable, but residual plots should still be checked for outliers and other evidence that the error structure in the experimental design is inappropriate.

Experimental Design Summary

Randomization supports treatment comparison, replication measures experimental variation, and blocking accounts for known nuisance factors. The identities in Propositions 0.31, 0.36 and 0.37 give the corresponding decompositions for completely randomized, randomized block, and factorial designs. Their scientific value still depends on a credible assignment mechanism, the correct experimental unit, prespecified comparisons, and appropriate residual diagnostics.