University of Waterloo

STAT 332
Sampling and Experimental Design

Course Notes

Spring 2018

by

JOCK MACKAY

STEFAN STEINER

MATTHIAS SCHONLAU

The material in these notes is derived and combined from the STAT 371 and STAT 372 notes of Jock MacKay, Stefan Steiner and Matthias Schonloau. Should there be any issues with the combination they are due to Ryan Browne.

Please email Ryan Browne (ryan.browne@uwaterloo.ca) with any errors or points of clarification. These notes are a work in progress.

Data Sets You can download all data sets in the notes and exercises from the zip file in the Content Folder on Learn!

Contents

Chapter 1

Fundamentals of Experimental Plans

In this and subsequent chapters, we describe the use, design and analysis of **experimental plans**. Recall that unlike an **observational plan**, in an experimental plan we deliberately change one or more of the process inputs. We use experimental plans to assess the effect of changing an existing process or the design of a product or service. Experiments are especially useful when we want to compare a new condition to an existing one, for example, to compare a new drug to the drug currently in use via a clinical trial.

We use experimental plans to better understand the relationship between the manipulated inputs and the output. We can conduct an experimental plan where we deliberately manipulate either or both fixed and varying inputs. Note that there will always be varying inputs (not controlled in the experiment) that will vary as the experiment is conducted. We must use experimental plans (rather than observational plans) to investigate the effect of changing fixed inputs since by definition they do not change unless we change them. We can also investigate the effect of varying inputs using an experimental plan but may also gain useful process knowledge regarding the effect of changes in the varying inputs with observational plans. Typically observational plans are cheaper and easier to conduct than experimental plans.

Generally experimental plans have a variety of possible goals including investigating the effect of changing inputs on the process output average and/or variability. In Chapters 1–3 we explore the use of experimental plans to investigate the effects on the process average of changing two or more fixed inputs. We explore other uses of experimental plans in the context of process improvement in the course Stat 435.

To start we use two examples for context to introduce the language of experimen-

tal plans. Remember that in an experimental plan, the people conducting the investigation deliberately change one or more input variates on the sampled units before the response variate is measured.

Example 1 (oral insulin trial): Compare effects of taking oral insulin daily versus no insulin with the goal of preventing the onset of type 1 diabetes.

Example 2 (pricing investigation): Assess the effect on profit of changing the price of one or both of cereals B and C while leaving the price of a third cereal A alone.

Some Language

Definition 1.1. Factor: *a single explanatory variate (input) that will be changed or set on each unit (e.g. people in the insulin trial and stores in the pricing investigation) in the sample. e.g. amount of daily insulin, price of cereal B, price of cereal C*

Definition 1.2. Factor Levels: *the set of values assigned to any factor in the plan. e.g. amount of insulin: low dose or none (2 levels) e.g. price of B: current or +5%, price of C: current or +5% (2 levels for each factor)*

Definition 1.3. Treatment: *a combination of the levels of the factors that can be applied to a unit e.g. current price of B, +5% price of C applied in a particular store. Example 1 has two treatments; Example 2 has four.*

Factors can be:

- **quantitative** e.g. dose of insulin in mg, price change %;
- **qualitative** e.g. colour, design 1 or design 2, ...

Response variates can be

- **continuous** e.g. profit at a given store over a specified period.
- **ordinal** e.g. ranking on scale of 1 to 5
- **count** e.g. number of flaws in a roll of paper
- **binary** e.g. type I diabetes occurs within 5 years or not
- **other** (categorical, image, map, scatter plot ...)

For this course, we restrict attention to experimental plans where we treat the response variate (output) as continuous.

1.1 Fundamental Principles of Experimental Plans (Stat 231 Review)

In designing any experimental plan, we should consider the following three ideas.

1. **Blocking**
   Blocking involves forming groups **(blocks)** of units in which one or more explanatory variate is held fixed while different treatments are applied to the units within the group.

   For example, in the pricing investigation, the sample consisted of 300 stores. Stores were formed into blocks of size four based on sales volume over a three week period prior to the application of the treatment. That is all stores in one block had very similar prior sales volumes. The four treatments were applied to one store in each block.

   Blocking

   - prevents **confounding** due to those explanatory variates held fixed within the block, and
   - improves the precision of the conclusions (e.g. shortens the length of confidence intervals) assuming the blocking variate is well chosen and has an effect on the response.

   Recall that **confounding** between two inputs occurs when the two inputs vary together in the investigation. If two inputs are confounded we can not tell which is the cause of any observed change in the (average) output. Note also that blocking is sometimes not possible due to cost or for logistical reasons such as carryover effects.

2. **Replication**
   Replication involves applying each treatment to more than one unit in the sample.

   For example, in the pricing investigation, each treatment was applied to 75 stores (once in each of the 75 blocks).

   Replication

   - helps to avoid sample error, and
   - lets us estimate the precision of our conclusions.

   Substantial replication is sometimes not possible due to high cost especially if there are many treatments.

3. **Random assignment (random allocation)**
   Random assignment is the process of assigning the treatments to the sample units using a probabilistic mechanism.

   For example, in the diabetes trial, patients were assigned to one of the two treatments (insulin or placebo) at random as they entered the investigation.

   In the pricing investigation, within each block of 4 stores, the 4 treatments were assigned to the stores at random (each store had the same chance of getting each treatment)

   Random assignment

   - together with replication reduces the risk of confounding by unknown explanatory variates, and
   - generates an analysis method - see exercise 12 at the end of the Chapter.

   In the diabetes trial, suppose 500 people receive each treatment. Because of the random assignment, the two treatment groups are likely to be balanced on average for unmeasured explanatory variates such as family history of diabetes several generations earlier. This balancing means that we avoid confounding by this unmeasured explanatory variate.

   Random assignment may not be possible due to cost or complexity of changing the treatment from unit to unit in the sample.

1.2 Comparative Experimental Plans Without Blocking

In Stat 231, we considered the models and analyses for comparative experimental plans with and without blocking. We include the following for review.

Example 3 (Adapted from `http://www.dack.com/web/flashVhtml/`.) In a usability test, an experimental plan was devised to compare two versions (V1 versus V2) of a company web site with respect to the time it takes to get information from the site. The response variate was the time to complete a set of questions such as

Question 1: How much does the men's fish tie cost?

Issue 1: What is an appropriate question or set of questions?
Twenty subjects were randomly formed into two groups (one for each web site version) of 10. Each subject completed the questions and the time was measured in seconds. Note that there is no blocking.

Issue 2: How should the subjects be selected?
Issue 3: When should the question(s) be answered?

The data are given below and in the file *ed_example1.txt*.

subject	1	2	3	4	5	6	7	8	9	10	average	stdev
V1	209	207	213	183	206	202	208	182	191	178	197.9	13.1
V2	215	239	237	228	229	238	256	210	232	229	231.3	12.9

To analyze the data, a model to describe the repeated execution of the plan is

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad R_{ij} \sim G(0, \sigma), \quad i = 1, 2\ ,\ j = 1, \ldots, 10$$

where the R_{ij} are independent and $\tau_1 + \tau_2 = 0$. We call τ_1, τ_2 the **treatment effects** to represent the increase (or decrease) from the overall average response if we use version 1 or 2.

Note that we use $(\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$ to compare the two treatments.

We have written the model for two treatments in an unnecessarily complicated way with a constraint because we will need to do so when we consider models for experimental plans with many treatments. We can write the model in many equivalent ways. For example, we could write

$$Y_{1j} = \mu + \delta + R_{1j}, \quad Y_{2j} = \mu + R_{2j}, \quad R_{ij} \sim G(0, \sigma), \quad i = 1, 2\ ,\ j = 1, \ldots, 10$$

where the parameter $\delta = \tau_1 - \tau_2$. This model is equivalent to the one above because we will get the same estimate for the parameter of interest $(\mu + \delta) - \mu = \delta = \tau_1 - \tau_2$.

We represent the observed values by y_{ij} and apply least squares to estimate the unknown parameters $\mu,\ \tau_1,\ \tau_2$ and σ by minimizing

$$\begin{aligned} W(\mu, \tau_1, \tau_2) &= \sum_{i,j} \left(y_{ij} - (\mu + \tau_i)\right)^2 \\ &= \sum_{i,j} \left(y_{ij} - \mu - \tau_i\right)^2 \end{aligned}$$

with respect to $\mu,\ \tau_1,\ \tau_2$, subject to the constraint. Using the method of **Lagrange Multipliers** from multi-variable calculus, we minimize

$$V(\mu, \tau_1, \tau_2, \lambda) = \sum_{i,j} \left(y_{ij} - \mu - \tau_i\right)^2 + \lambda(\tau_1 + \tau_2)$$

with respect to $\mu,\ \tau_1,\ \tau_2,\ \lambda$, ignoring the constraint.

After some calculus (see the exercises you need to be able to do this), we get $\hat{\mu} = \overline{y}_{++}$ (the overall average) and $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$ (the treatment i average minus the overall average). More importantly, $\hat{\tau}_1 - \hat{\tau}_2 = \overline{y}_{1+} - \overline{y}_{2+}$ estimates the difference in the treatment effects.

Note that in the example $\overline{y}_{i+} = \sum_{i=1}^{10} y_{ij}/10$, the treatment average for treatment i. This notation for averages carries throughout the course. The + sign in the subscript position indicates that we have added the variate values over that subscript. The bar over the variate symbol indicates that we have averaged the variate values over all subscripts with a + sign.
From the data, we have

$$\begin{aligned} \hat{\mu} &= \overline{y}_{++} = 214.6 \\ \hat{\tau}_1 &= \overline{y}_{1+} - \overline{y}_{++} = -16.7 \\ \hat{\tau}_2 &= \overline{y}_{2+} - \overline{y}_{++} = 16.7 \\ \hat{\delta} &= \hat{\tau}_1 - \hat{\tau}_2 = -33.4 \end{aligned}$$

and the estimate of σ is

$$\widehat{\sigma} = \sqrt{\frac{W(\hat{\mu}, \hat{\tau}_1, \hat{\tau}_2)}{20-2}} = \sqrt{\frac{\sum_{i,j} \left[y_{ij} - (\hat{\mu} + \hat{\tau}_i)\right]^2}{20-2}} = 13.0.$$

The estimated residuals $\hat{r}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i$ (used to check the model fit) are plotted in Figure 1.1. See the section on using R for the commands that produced this plot. From the plot, the variability within each treatment is about the same and there are no exceptional values (outliers).

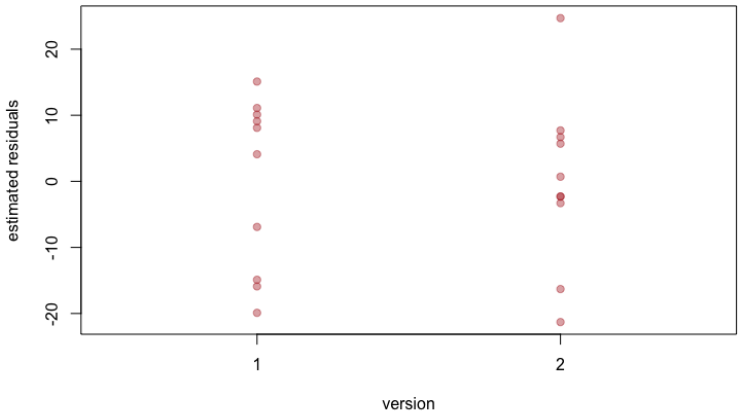


Figure 1.1: Estimated Residuals from Usability Trial

Formal Analysis Procedures (Confidence Intervals, Tests of Significance)

The estimators $\tilde{\mu}, \tilde{\tau}_1, \tilde{\tau}_2, \tilde{\sigma}$ are random variables that describe the behaviour of the corresponding estimates (according to the model) if the plan were to be repeated over and over (i.e. you repeatedly select another 20 people at random, randomly assign 10 each version and measure the time to complete the tasks). By adopting the model we have that

$$\begin{aligned} Y_{ij} &\sim G(\mu + \tau_i, \sigma), \\ \text{so that} \quad \overline{Y}_{i+} &\sim G\left(\mu + \tau_i, \frac{\sigma}{\sqrt{10}}\right), \end{aligned}$$

by the rules for linear combinations of independent identically distributed Gaussian random variables. To compare treatments we have

$$\begin{aligned} \tilde{\tau}_1 - \tilde{\tau}_2 &= \overline{Y}_{1+} - \overline{Y}_{2+} \\ &\sim G\left((\mu + \tau_1) - (\mu + \tau_2), \sqrt{\frac{\sigma^2}{10} + \frac{\sigma^2}{10}}\right), \text{ by the same linear combination rules} \\ &\sim G\left(\tau_1 - \tau_2, \sigma\sqrt{\frac{1}{10} + \frac{1}{10}}\right) \end{aligned}$$

and

$$\frac{\tilde{\sigma}^2(20-2)}{\sigma^2} \sim \chi^2_{20-2},$$

so then we have

$$\frac{(\tilde{\tau}_1 - \tilde{\tau}_2) - (\tau_1 - \tau_2)}{\tilde{\sigma}\sqrt{\frac{1}{10} + \frac{1}{10}}} \sim t_{18}.$$

Question 1: What is the estimated size of the difference between the two treatments? How precisely have we estimated this difference?

A 95% confidence interval for the difference in treatment effects is

$$\hat{\tau}_1 - \hat{\tau}_2 \pm 2.10\hat{\sigma}\sqrt{\frac{1}{10} + \frac{1}{10}}, \quad \text{or} \quad -33.4 \pm 12.2,$$

where

$$\begin{aligned} \hat{\tau}_1 - \hat{\tau}_2 &= (-16.7) - 16.7 = -33.4, \\ \hat{\sigma} &= 13.0 \text{ from above, and} \\ c &= 2.101 \end{aligned}$$

for right-tail probability 0.025 with a t-distribution with 18 degrees of freedom.

Recall that the general form of a confidence interval (for all parameters other than σ) is

$$estimate \pm \ c \times estimated\ standard\ deviation\ (estimator)$$

where the constant c is found in the t-tables based on the confidence level and the degrees of freedom.

We conclude with high confidence that users can complete the tasks with version 1 faster (on average) than version 2 by somewhere between 21 and 45 seconds. This is about 10-20% faster and we conclude that this percentage improvement will carry over to other tasks. There are severe limitations to this conclusion because of the set of tasks and the subjects selected.

For practice and review, we ask

Question 2: Is there any evidence of a difference between the two treatments? To carry out a **test of significance** (also called a **hypothesis test**)

1. Suppose that there is no difference, i.e. suppose $\tau_1 - \tau_2 = 0$ (This is our **null hypothesis**, H_0).
2. Find the estimate of $\tau_1 - \tau_2$ and the standard deviation of the corresponding estimator. The estimate of $\tau_1 - \tau_2$ is $\hat{\tau}_1 - \hat{\tau}_2 = -33.4$, and the standard deviation of the corresponding estimator is

$$stdev(\tilde{\tau}_1 - \tilde{\tau}_2) = \sigma\sqrt{\frac{1}{10} + \frac{1}{10}}.$$

3. Calculate the discrepancy measure

$$d = \frac{|(\hat{\tau}_1 - \hat{\tau}_2) - 0|}{\hat{\sigma}\sqrt{\frac{1}{10} + \frac{1}{10}}} = 5.76.$$

4. Find the p-value

$$p = P(D \geq d) = P\left(|t_{18}| \geq 5.76\right) \leq 0.001.$$

   Using the t_{18} table, we get

$$\begin{aligned} P\left(|t_{18}| \geq 3.9216\right) &= 2(0.0005) = 0.001, \text{ which means that} \\ P\left(|t_{18}| \geq 5.76\right) &\leq 0.001. \end{aligned}$$

5. Interpret the result. Since the p-value is so small, there is strong evidence that $\tau_1 - \tau_2 \neq 0$.

We reach the same conclusion with the test of significance as with the confidence interval. There is very strong evidence that, on average, users can complete tasks faster with version 1 than with version 2. The limitations discussed above apply equally to this conclusion.

USING R
For this simple plan with only two treatments, we can get the plot, confidence interval and results of the t-test using the following R code. Using the R commands

```
a<-read.table('Data/ed_example1.txt',header=T)
attach(a)

estres<-resid(lm(time~version))
stripchart( estres ~ version, vertical=T,xlab='version',
     ylab='estimated residual',
     main='Estimated Residuals vs Version',
     col=adjustcolor("firebrick", 0.4), xlim=c(0.5,2.5), pch=19)

t.test(time~version,var.equal=T)
```

we obtain figure 1.1 and the following

```
Two Sample t-test
data:  time by version
t = -5.7611, df = 18, p-value = 1.845e-05
alternative hypothesis: true difference in means
     is not equal to 0
95 percent confidence interval:
 -45.58011 -21.21989
sample estimates:
mean in group v1 mean in group v2
           197.9            231.3
```

1.3 Comparative Experimental Plans With Blocking

We now look at the second comparative plan that uses blocking to prevent **confounding** and increase the precision of the comparison.

Example 2 (two treatments with blocking) In the painting of two-colour plastic fascias, the prime coat (black in this case) was applied to the whole part. Some of the black area was masked with paper taped to the fascia. Then, the color coat was added. When the masking was removed, there was occasionally a residual pattern, called ghosting, on the matte black surface under the tape. A simplified process map is given in Figure 1.2.

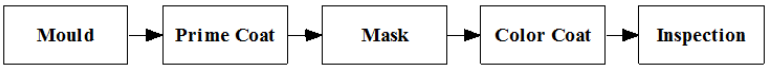


Figure 1.2: Process Map of the Fascia Manufacturing

Customers can detect ghosting if present since the black surface of the fascia was prominent. The plant could have reworked the surface to remove the ghosting; however, this would have added cost and slowed production. A process engineer was assigned the task to eliminate the ghosting problem.

The process operators visually judged ghosting on a scale of 0 to 10 (0 = no ghosting, 10 = heavy ghosting). The plant reworked fascias with a ghosting score greater than 2 and reluctantly shipped those with a score of ≤ 2. The engineer was convinced that the cause of the problem was some environmental factor such as ambient temperature or humidity that changed from day to day and could not be easily controlled. She also knew that the ghosting appeared under the tape during the baking of the color coat and that the problem seemed more frequent with certain colors. Another tape supplier claimed that his product would not generate ghosting. She decided to investigate whether the new source of tape was more robust to the effects of the environmental factors and colour than was the current brand of tape.

Plan:
Because of the high cost of scrap, she decided to use 15 fascias that had been scrapped upstream of the masking operation for her study population. She also decided to use only those fascias produced on a hot humid day and to paint these fascias with the color having the most frequent ghosting problem. Furthermore, she planned to mask two small areas, labeled I and II on the primed surface on both sides of each fascia with the two different tapes. See Figure 1.3.

On each side of the fascia she assigned two tapes, the current (treatment 1) and new (treatment 2) at random to I or II. Then she painted the fascia and measured the degree of ghosting. The data are shown below in table 1.1 and are available in the file *ed_example2.txt*.

Figure 1.3: Arrangement of Treatments in Two Blocks per Fascia

Table 1.1: The fascia and measured the degree of ghosting.

fascia	side	1	2	fascia	side	1	2
1	L	4	0	9	L	1	2
1	R	4	3	9	R	4	1
2	L	5	5	10	L	5	3
2	R	9	3	10	R	5	2
3	L	6	2	11	L	7	4
3	R	6	2	11	R	5	4
4	L	5	3	12	L	1	0
4	R	5	3	12	R	2	1
5	L	0	0	13	L	1	2
5	R	1	0	13	R	3	2
6	L	1	0	14	L	1	0
6	R	2	0	14	R	1	1
7	L	6	5	15	L	3	2
7	R	4	1	15	R	3	1
8	L	1	1	average		3.47	1.8
8	R	3	1				

Note that we have 15 pieces, with two sides each, for a total of 30 blocks.

Analysis: The model is

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij},\ R_{ij} \sim G(0, \sigma),\ i = 1, 2\ ,\ j = 1, \ldots, 30$$

where the R_{ij} are independent and $\sum_i \tau_i = 0,\ \sum_j \beta_j = 0$. As before, we call τ_1, τ_2 the **treatment effects** and $\beta_1, \ldots, \beta_{30}$ the **block effects**.

Note that with the proposed model the treatment effects are additive, i.e. for any block, each treatment has the same effect. As well, we are applying a continuous model to a response variate (namely ghosting effect) that is ordinal.

We can express the questions of interest in terms of the difference of the treatment effects, $\tau_1 - \tau_2$.

We represent the observed values of ghosting effect by y_{ij}. Using least squares, we estimate the unknown model parameters by minimizing

$$\begin{aligned} W(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_{30}) &= \sum_{i,j} (y_{ij} - (\mu + \tau_i + \beta_j))^2 \\ &= \sum_{i,j} (y_{ij} - \mu - \tau_i - \beta_j)^2 \end{aligned}$$

subject to the constraints, $\sum_i \tau_i = 0$ and $\sum_j \beta_j = 0$. Again using the method of **Lagrange multipliers**, we minimize

$$V(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_{30}; \lambda_1, \lambda_2) = W(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_{30}) + \lambda_1 \sum_i \tau_i + \lambda_2 \sum_j \beta_j$$

with respect to the 35 parameters. After some calculus (that I know you would love to do), we get

$$\hat{\mu} = \overline{y}_{++}, \quad \hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++} \quad \text{and} \quad \hat{\beta}_j = \overline{y}_{+j} - \overline{y}_{++}.$$

Doing the calculations, the difference in treatment effects is estimated by

$$\hat{\tau}_1 - \hat{\tau}_2 = \overline{y}_{1+} - \overline{y}_{2+} = 1.67.$$

The estimate of the standard deviation σ is

$$\begin{aligned} \hat{\sigma} &= \sqrt{\frac{\sum_{i,j} \left[y_{ij} - \overline{y}_{++} - (\overline{y}_{i+} - \overline{y}_{++}) - (\overline{y}_{+j} - \overline{y}_{++}) \right]^2}{60 - \underbrace{1}_{\mu} - \underbrace{(2-1)}_{\tau_i s} - \underbrace{(30-1)}_{\beta_j s}}} \\ &= \sqrt{\frac{\sum_{i,j} \left[y_{ij} - \overline{y}_{i+} - \overline{y}_{+j} + \overline{y}_{++}) \right]^2}{29}} = 1.12. \end{aligned}$$

After a little algebra (see the exercises), we can show that

$$\hat{\sigma} = \frac{1}{\sqrt{2}} \sqrt{\frac{\sum_j (d_j - \overline{d}_+)^2}{29}}$$

where $d_j = y_{1j} - y_{2j}$. Recall in Stat 231, we analyzed the differences for this model. We can show this results in an identical confidence interval for $\theta = \tau_1 - \tau_2$.

Formal Analysis Procedures

The estimator of $\theta = \tau_1 - \tau_2$ is $\tilde{\theta} = (\overline{Y}_{1+} - \overline{Y}_{++}) - (\overline{Y}_{2+} - \overline{Y}_{++}) = \overline{Y}_{1+} - \overline{Y}_{2+} = \overline{D}$, the difference of the averages. We have

$$\begin{aligned} D_j &= Y_{1j} - Y_{2j} \\ &= (\mu + \tau_1 + \beta_j + R_{1j}) - (\mu + \tau_2 + \beta_j + R_{2j}) \\ &= (\tau_1 - \tau_2) + (R_{1j} - R_{2j}) \\ &= \theta + R_{1j} - R_{2j}. \end{aligned}$$

Recall that $R_{ij} \sim G(0, \sigma)$, so that $D_j = \theta + R_{1j} - R_{2j} \sim G\left(\theta, \sqrt{2}\sigma\right)$, and thus

$$\tilde{\theta} \sim G\left(\theta, \frac{\sqrt{2}\sigma}{\sqrt{30}}\right) \quad \text{and} \quad \frac{(29) \times \tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{29}$$

Confidence Interval A 95% confidence interval for θ is $\hat{\theta} \pm c\hat{\sigma}\sqrt{\frac{2}{30}}$ where the constant $c = 2.045$ comes from the t-table with 29 degrees of freedom. The interval is 1.67 ± 0.59. We are confident that the difference in average ghosting level of the two tapes falls in the range 1.67 ± 0.59.

The engineer decided to switch tape supplier since the new tape appeared to give less ghosting on average and the cost was the same. This turned out to be a good decision since the overall rework cost due to ghosting was reduced by more than 70%.

Test of Significance (again for practice only) Is there any evidence of a difference in the two tapes?

1. Suppose that $\theta = \tau_1 - \tau_2 = 0$ (This is our **null hypothesis**, H_0).

2. The estimate of $\theta = \tau_1 - \tau_2$ is $\hat{\theta} = \hat{\tau}_1 - \hat{\tau}_2 = 1.67$.

3. The discrepancy measure is

$$d = \frac{|\hat{\theta} - 0|}{\hat{\sigma}\sqrt{\frac{2}{30}}} = 5.77.$$

   By a Theorem from Stat 231, the corresponding estimator has a t_{29} distribution.

4. The p-value is

$$p = P(D \geq d) = P\left(|t_{29}| \geq 5.77\right) \leq 0.001.$$

   Using the t_{29} table, we get

$$\begin{aligned} P\left(|t_{29}| \geq 3.6594\right) &= 2(0.0005) = 0.001, \text{ so that} \\ P\left(|t_{29}| \geq 5.77\right) &\leq 0.001. \end{aligned}$$

5. There is very strong evidence that $\theta \neq 0$.

That is, there is very strong evidence of a difference in the average ghosting associated with the two tapes. To determine which tape is better (on average) we

need to examine the individual values of $\hat{\tau}_1$ and $\hat{\tau}_2$ (or carefully interpret $\hat{\theta}$ realizing we arbitrarily defined it so that positive values mean that $\hat{\tau}_1$ is larger than $\hat{\tau}_2$).

USING R

The data are stored in the file *ed_example2.txt*. Reading and displaying (the first few lines only) of the data file, we have the R code (full code in the file ed_example2.R):

```
a<-read.table('ed_example2.txt',header=T)
attach(a)
a
```

with partial output

```
   fascia side block treatment ghosting
1       1    L     1         1        4
2       1    L     1         2        0
3       1    R     2         1        4
4       1    R     2         2        3
5       2    L     3         1        5
```

Here we have included a variate to indicate the block that runs from 1 to 30. We need to ensure that R treats the block as a categorical variate so we use the code

```
block <-factor(block)
```

We can then produce the estimate of the standard deviation and the t-test for the hypothesis that there is no treatment effect with the code

```
b <-lm(ghosting~block+treatment)
summary(b)
```

The output includes many unneeded results. We focus on the following few lines

```
Coefficients:
             Estimate Std. Error    t value    P(>|t|)

block30    5.903e-16   1.119e+00    5.27e-16   1.00000
treatment  -1.667e+00  2.890e-01   -5.767      3.02e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.119 on 29 degrees of freedom
Multiple R-Squared: 0.8546,     Adjusted R-squared: 0.7042
F-statistic: 5.683 on 30 and 29 DF,  p-value: 5.246e-06
```

The estimate $\hat{\sigma} = 1.12$ is given by the Residual standard error along with the corresponding degrees of freedom. The estimate of the difference of treatment effects is $\hat{\tau}_2 - \hat{\tau}_1 = -1.67$ (note the order of $\hat{\tau}_1$ and $\hat{\tau}_2$ is reversed from what we did earlier), the corresponding standard error and the results from a t-test of the hypothesis of no treatment difference are given on the treatment line of the output.

The estimated residuals are available using

```
resid(b)
```

and we can examine the suitability of the model by plotting the estimated residuals in various ways.

One-sided versus Two-sided Tests of Significance

In the tape example above, we asked the question “Is there any evidence of a difference in the two tapes?” Suppose instead we had initially asked the question “Is there any evidence of that the new tape is better than the current tape?” In terms of the treatment effects, the first question asks if there is evidence that $\theta = \tau_1 - \tau_2 \neq 0$ and the second if there is evidence that $\theta > 0$.

The change in the form of the question (from $\theta \neq 0$ to $\theta > 0$) changes the discrepancy in the test of significance and leads to a **one-sided test**. Remember that we want the discrepancy to be large **if the null hypothesis is not true**. We re-visit the example above with the new question.

Is there any evidence that $\theta > 0$?

1. Suppose that $\theta = \tau_1 - \tau_2 = 0$. This is our null hypothesis, with the alternative hypothesis that $\theta > 0$.
2. The estimate of $\theta = \tau_1 - \tau_2$ is $\hat{\theta} = \hat{\tau}_1 - \hat{\tau}_2 = 1.67$ as before.

3. The discrepancy measure is

$$d = \frac{\hat{\theta} - 0}{\hat{\sigma}\sqrt{\frac{2}{30}}} = 5.77 \text{ as before.}$$

   [Note d will be large only if $\hat{\theta}$ is large and positive.] The corresponding estimator has a t_{29} distribution.

4. The p-value is

$$p = P(D \geq d) = P\left(t_{29} \geq 5.77\right) \leq 0.0001.$$

   [We calculate the p-value in **one tail** of the distribution only.]

5. There is very strong evidence that $\theta > 0$.

The only difference between one- and two-sided tests is the discrepancy measure. The steps and interpretation are identical. We decide to use a one- or two-sided test based on the form of the question asked about the parameter. Look for words such "greater than", "less than" to indicate a one-sided test and "different from" to indicate a two-sided test.

Summary

- We have looked at two designs to compare two treatments, one with blocking and one without.
- Both designs use replication and randomization to help prevent confounding.
- We model the treatment effects, the change from the population average induced by the use of a particular treatment. For two treatments $\tau_1 + \tau_2 = 0$.
- We estimate the model parameters using least squares (including the constraints).
- The parameter estimates are linear combinations of the response variates. Accordingly, the estimators are Gaussian with mean equal to the corresponding parameter and standard deviation equal to a multiple of σ.
- The estimate of σ is the square root of the minimum value of the least squares function divided by the degrees of freedom (number of observations number of parameters + number of constraints).
- The general form of a confidence interval (for all parameters other than σ) is

$$\textit{estimate} \pm c \times \textit{estimated standard deviation } (\textit{estimator}).$$

- The general form of the discrepancy measure for testing $\theta = \theta_0$ with a two-sided alternative is

$$d = \frac{|\hat{\theta} - \theta_0|}{\text{estimated standard deviation } (\tilde{\theta})}$$

and with a one-sided alternative $\theta > \theta_0$ is

$$d = \frac{\hat{\theta} - \theta_0}{\text{estimated standard deviation } (\tilde{\theta})}$$

1.4 Exercises

1. Look at the description of the investigation to compare two versions of the same Web page (`http://www.dack.com/web/flashVhtml/`). Use the notion of study error to criticize the conclusion that the Flash version is inferior to the HTML version.

2. In a marketing investigation, a large supermarket chain wants to look at the effect of shelf placement for three brands of coffee. There are three shelf positions top, middle and bottom. For a given treatment, the response variate is the total sales ($) over a one week period. There are 24 stores available.

   (a) Explain why there are 6 different treatments.

   (b) Describe an experimental plan that does not involve blocking be sure to discuss randomization and replication.

   (c) Repeat part 2b but include blocking in your plan.

   (d) Compare the two plans.

3. To compare a distance education option versus a traditional lecture version of a course, a common examination is administered and the grades of the students are compared. The data are summarized below.

style	**number of students**	**average**	**st dev**
lecture	47	71.3	10.2
distance ed	36	68.7	11.3

   (a) Is there any evidence of that performance is worse with distance education version? Be sure write down the model you use.

   (b) This is not an experimental plan. Why? Does this observation have any impact on your conclusion in part 3a?

4. To compare the bias in two measurement systems I and II, a sample of 10 parts is selected from production over one day. Each part is measured on both measurement systems in a random order. The data are shown below and stored in the file *ed_exercise4.txt* in row/column format.

part	**system I**	**system II**
1	3.2	3.1
2	5.6	5.8
3	1.2	1.2
4	3.8	3.6
5	7.2	7.7
6	4.2	4.2
7	3.4	3.6
8	5.2	5.6
9	2.8	3.1
10	2.7	2.8

   (a) Write a model for the repeated application of the Plan that treats parts as blocks. What parameter in the model can you use to estimate the relative bias of the two systems?

   (b) Estimate the relative bias and give a 95% confidence interval.

   (c) Plot the estimated residuals against the average measured value for each part. Does it appear that the bias changes with part size?

   (d) Consider an alternate plan in which 20 parts are selected and randomly divided into two groups of 10. Each group is measured by one system only. Why would it be more difficult to detect a bias with this plan? Note that if the measurement is destructive we must use a plan such as this.

5. Suppose in an experimental plan to compare two treatments without the use of blocking, the investigator has funds to measure $2r$ units. The plan is **balanced** if r units are allocated to each treatment.

   (a) For what allocation, will the standard deviation of the estimator of the treatment difference be smallest?

   (b) What is the impact of this choice on the confidence interval for the treatment difference?

6. Let's do some calculus.

   (a) For the unblocked model, find the least squares estimates of μ, τ_1, τ_2 that minimize

$$W(\mu, \tau_1, \tau_2) = \sum_{i,j} (y_{ij} - \mu - \tau_i)^2$$

subject to the constraint $\tau_1 + \tau_2 = 0$.

(b) For the model with b blocks, find the least squares estimates of $\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b$ that minimize

$$W(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b) = \sum_{i,j} (y_{ij} - \mu - \tau_i - \beta_j)^2$$

subject to the constraints $\sum_i \tau_i = 0,\ \sum_j \beta_j = 0$.

7. Alternate forms for $\hat{\sigma}$ - these are useful for calculation purposes.

(a) In the unblocked plan, show that the estimate of the residual standard deviation can be written as

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where n_i, s_i are the sample size and sample standard deviation for treatment i respectively. Interpret this formula in words.

(b) In the plan for comparing two treatments with n blocks, show that

$$\hat{\sigma} = \frac{1}{\sqrt{2}} \sqrt{\frac{\sum_j (d_j - \overline{d}_+)^2}{n - 1}}$$

where $d_j = y_{1j} - y_{2j}$ is the difference of the response variates within block j. Hint: For two numbers a_1, a_2, show that $stdev(a_1, a_2) = \frac{|a_1 - a_2|}{\sqrt{2}}$ and then, with $a_i = y_{ij} - \overline{y}_{i+}$, use this result in each block.

8. A technical assistance center (TAC) provides advice to mechanics who are trying to repair a car. The mechanic who cannot solve a problem telephones the TAC and is connected with a highly trained technician who tries to diagnose the problem and provide advice for a solution. To reduce cost, the management of the TAC decides to investigate a change. In the current system, the mechanic calls the center and the technician collects and enters information into a database about the vehicle identification, mileage and the dealership. Then the technician asks the mechanic about the problem. In the proposed system, the mechanic will enter the information about the vehicle, mileage and dealership using the keys on the telephone. Then the technician will be contacted and can start immediately to ask about the problem. The TAC management estimates that the technicians time costs about twice that of the mechanic. The computer system can automatically measure the time from the start of the call and connection to the technician. Carefully explain how you would design an experimental plan to investigate the cost saving available from the proposed system.

9. To investigate a new packaging plan, the producer of a consumer product placed the current packaging and the new packaging side by side in 15 stores for one week and recorded the sales (in $) of each type of package. Different stores charged different amounts for the product but kept the price of each packaging style the same. The total sales to the nearest dollar are shown below. The average weekly sales from the past year were also recorded. The data are stored in the file *ed_exercise9.txt* and given below

store	**old package**	**new package**	**average weekly sales**
1	1406	1351	2888
2	1134	1135	2416
3	1124	1607	2684
4	497	484	816
5	621	1084	1688
6	309	726	1256
7	1172	1732	3008
8	1599	1045	2560
9	693	913	1464
10	800	533	1220
11	965	868	1724
12	445	816	1240
13	1357	1418	3076
14	868	1033	1904
15	899	1389	2112

   (a) Write a brief report for your manager that includes the results of a formal analysis of the data and a conclusion. The key issue is to estimate the change in sales if the new packaging is adopted.

   (b) Technical issue: Does the treatment effect depend on the past average sales?

   (c) The original report contained a 95% confidence interval to describe the change (on average) due to the new packaging. A manager reading the report asked the analyst for a non-technical explanation of the interval. What would you have written?

   (d) The company adopted the new packaging for all stores. After a brief increase in sales, they were disappointed in the results. In a review of the trial and results, the analyst pointed out that it might have been better to put only a single packaging in each store. Explain.

10. A manufacturing organization is planning to train a large number of employees in problem solving, continuous improvement and the use of simple statistical methods. After considerable research, cost comparisons etc,

the training manager has narrowed the choice of training method to two. Method 1 involves external consultants who will deliver classroom training combined with role playing and hands-on exercises. Method 2 uses individual on-line training with machine generated feedback. Before committing to either approach, the manager decides to compare the two methods with an experiment. She is especially interested in knowing if the classroom training is more effective since it is more costly. She randomly assigns 24 employees into two groups of 12. She then contracts to have each group use one of the two methods. Each employee is given a test immediately after the training is over (score is y1) and then another test two weeks later (score is y2). Carry out an analysis of the data given in the file *ed_exercise10.txt* and write a brief report on your findings.

11. For some parameter θ, suppose we have an estimator $\tilde{\theta} \sim G(\theta, k\sigma)$ and an estimate of σ with q degrees of freedom. Prove that a particular value of θ, say $\theta = \theta_0$, falls within the $100(1-p)\%$ confidence interval if and only if the p-value for the hypothesis $\theta = \theta_0$ is greater than or equal to p.

12. Suppose that we have eight subjects (this is a deliberately chosen small number to avoid a lot of computation) and want to compare two treatments. We randomly split the subjects into two groups and assign each group a treatment. The data are shown below. Here we look at an alternate model and analysis based on the random assignment of the treatments.

Treatment A	**Treatment B**
6.3	5.9
5.4	6.6
5.7	6.7
5.8	6.1

The Model: For every subject we suppose that there is a fixed value for the response variate for each treatment. That is, for each subject we have a pair (y_a, y_b) corresponding to the two values of the response variate. In the investigation, we get to see exactly one of the two elements of the pair depending on which treatment is applied according to the random allocation.

(a) How many different ways can the groups be formed?

The Analysis: If there is no treatment difference, then we hypothesize that $y_a = y_b$ for every subject. We have observed a difference in treatment average $|5.80 - 6.325| = 0.525$.

(b) Using this as a discrepancy measure, what is the probability that we would see such a large difference if our hypothesis is true?

(c) What conclusion do you draw based on the significance level found in part 12b?

We can extend the model to allow for a treatment difference. Suppose that $y_a = y_b + \theta$ for every subject so that suppose that θ represents the treatment effect.

(d) How would you test the hypothesis that $\theta = \theta_0$?

(e) How would you find a confidence interval for θ? [use a generalization of the result in question 3 - you do not need to prove this generalization which in fact is true.]

(f) Suppose that we had 50 subjects in each group. How can you carry out the analyses described above?

Note that all of the above is based on a model derived solely from the random assignment of the treatments. We will exploit the same idea when we look at the analysis of surveys.

Chapter 2

Experimental Plans for More Than Two Treatments

Now we look at the design and analysis for experimental plans with more than two treatments. There are two basic designs corresponding to the plans we used for two treatments (no blocking versus blocking):

- Completely randomized design - no blocking is used
- Randomized block design - each treatment is repeated once in each block

The major new feature of this section is the introduction to the analysis of variance (ANOVA), a methodology used to test hypotheses that involve two or more restrictions on the parameters.

2.1 Completely Randomized Designs

Suppose that we have $t \geq 2$ treatments to compare. In a completely randomized design we randomly assign a number of units to each treatment. We looked at several examples with $t = 2$ in Chapter 1.

Example 1
A battery manufacturer decided to conduct an investigation to compare five different designs of a new type of battery. To make the comparison, a standard test was developed. The response variate was the watt-hours (a measure of energy) delivered by the battery under the test conditions. The larger the watt-hours, the longer the battery should last under conditions similar to those in the test. Ten batteries of each design were produced by companys R&D lab. We label the five designs A, B C, D and E. The batteries were then tested in a random order. The data are stored in the file *crd_example1.txt* and are given below:

A	B	C	D	E
1.55	1.11	1.24	1.38	1.05
1.23	1.26	1.11	1.50	1.17
1.28	1.24	1.53	1.20	1.15
1.44	1.24	1.21	1.19	1.21
1.29	1.05	1.30	1.32	1.07
1.56	1.28	1.30	1.04	1.28
1.43	1.15	1.09	1.11	1.14
1.22	1.23	1.06	1.05	1.42
1.36	1.24	1.00	1.19	1.29
1.32	1.26	1.20	1.10	1.25

Using the R commands below (and data stored in the file *crd_example1.txt*)

```
require(ggplot2)
qplot(battery,power, data=batterydata)
```

we get Figure 2.1

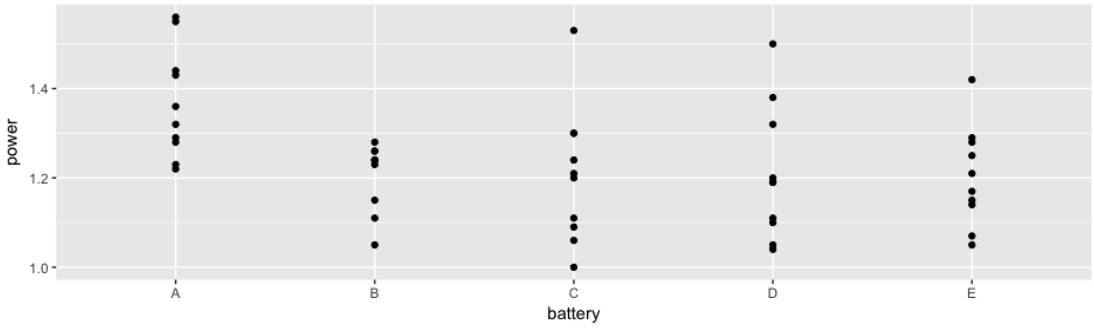


Figure 2.1: Power versus Battery Type

The questions of interest:

1. Are there any differences among the performances of the battery designs?
2. Does energy decrease on average with design cost?
3. Which battery design is the best?

Example 2 (based on Seawater tolerance and gene expression in two strains of Atlantic salmon smolts, T.D. Singer et al., Canadian Journal of Fisheries, 59, 2002)

Atlantic salmon are raised in fish farms throughout the world. In the wild, the salmon breed in fresh water and the young fish (smolts) migrate to the sea after about one year. Here we describe an experimental plan to compare the difference

in response of the two strains of smolt to the movement from fresh to salt water. As in most expensive investigations, there were several response variates; we deal only with the plasma concentration of the ion Ca^{+2}. There were two factors of interest:

Strain of smolt:	native versus farm-raised
Time after exposure to seawater:	0, 6, 24, 96, 336 hours

The questions of interest (not necessarily in order of importance) are:

1. Is there any evidence of a difference amongst the treatments?
2. Is there any difference, on average, between the native and farmed fish?
3. Is there any difference in the strains after 336 hours?
4. Is the pattern of variation over time different for the two groups?

Forty native and forty farm-raised smolts were placed in the same large tank of fresh water to acclimatize. Then the freshwater was replaced with salt water over a 1 hour period. At each of the five time points, eight fish from each group were selected at random and sacrificed. Then the Ca^{+2} concentration in each fish was measured. There were 10 treatments denoted by a combination of the two factors, the strain (N or F) and time, so, for example, N96 corresponds to the native group of smolts after 96 hours. The data are stored in the file *crd_example2.txt*.

n0	**n6**	**n24**	**n96**	**n336**	**f0**	**f6**	**f24**	**f96**	**f336**
1.31	1.70	1.23	1.24	1.20	1.85	1.68	1.32	1.50	1.01
1.57	1.96	1.39	1.50	1.14	1.21	1.05	1.37	1.07	1.30
1.09	1.67	0.67	0.96	1.69	0.82	1.64	0.95	1.52	0.99
1.16	1.04	1.34	0.69	0.89	0.92	1.32	1.71	1.77	0.81
1.62	1.32	1.52	0.92	0.95	1.50	1.92	1.08	1.64	1.08
1.75	1.10	1.16	0.84	1.17	1.66	1.35	1.02	0.98	1.37
1.02	0.84	1.27	1.06	0.87	1.49	1.95	0.95	1.65	NA
1.63	1.57	1.20	1.41	1.52	1.32	0.74	NA	2.01	NA

Note that three fish died (NA is the R code for missing) before the measurements were taken. We return to these two examples below.

Plan Issues

- If the number of units given to each treatment is the same, we say the design is **balanced**. With a balanced design, we estimate each treatment effect with the same precision. In general, we prefer balanced designs unless we want to estimate certain effects more precisely than others. The salmon example is not balanced because of the loss of fish.

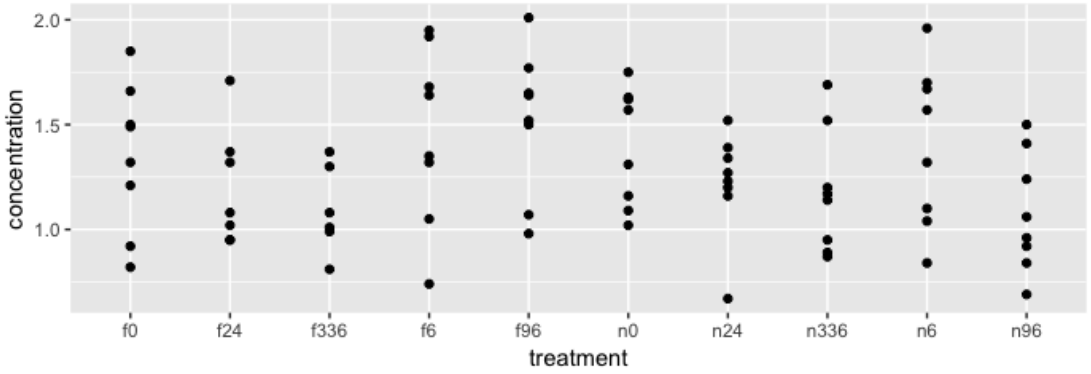


Figure 2.2: Ca^{+2} concentration in each fish versus treatment

- A key issue is sample size. How many units should we use for each treatment? In a relative sense, to improve the precision by a factor of 2 we need to increase the sample size by a factor of 4. To predict the precision of any comparison before the plan is executed, we need an estimate of the underlying standard deviation that is difficult to come by. We return to this point after the analysis section. In most cases, the cost of the units and measurement dominate the choice of sample size.
- In many problems, one of the treatments is a "control" which represents the current situation (e.g. current treatment for a disease, the current process settings etc.)
- You may have noticed that example 2 is not technically an experimental plan. The investigators did not assign the strain to each smolt. The strain is determined elsewhere and is not under the control of the investigator. In such cases, we lose the protection that random assignment provides against confounding. We proceed with the analysis as if the plan is experimental and in the conclusion, remember that we did not allocate the smolts at random to the treatments.

Model and Analysis (Balanced Plan)
Suppose we have a balanced experimental plan with r observations for each of t treatments. We model the repeated execution of the plan by:

$$Y_{ij} = \mu + \tau_i + R_{ij}, \qquad R_{ij} \sim G(0, \sigma), \qquad i = 1, \ldots, t, \ j = 1, \ldots, r.$$

Note that this model is identical to that used for the case of 2 treatments. Since the problem is to compare the treatments, in terms of the parameters, we are interested in assessing $\tau_i - \tau_k$ for various i and k.

In the model, we have $E(Y_{ij}) = \mu + \tau_i$. Note that if we subtract a constant (say 3) from μ and add the same constant to each τ_i, the individual parameters are changed but $E(Y_{ij}) = (\mu - 3) + (\tau_i + 3)$ and $(\tau_i + 3) - (\tau_k + 3) = \tau_i - \tau_k$ are not. This means that we have too many parameters in the model. To deal with this issue, we add a constraint that will make the parameters uniquely identifiable.

The form of the constraint changes the interpretation of the parameters. For example,

1. Suppose we let $\sum_i \tau_i = 0$. Then we have for j fixed,

$$\sum_i E(Y_{ij}) = (\mu + \tau_1) + \cdots + (\mu + \tau_t) = t\mu$$

   or equivalently

$$\mu = \frac{\sum_i E\left(Y_{ij}\right)}{t}.$$

   so that μ represents the average response across the t treatments. This is usually not a parameter of interest since we can apply only one treatment at a time. We interpret τ_i as the increase over this average when we apply treatment i, and hence τ_i is called the **treatment effect**.

2. We can consider another form for the constraint such as $\tau_1 = 0$. Now we have $\mu = E(Y_{1j})$ and $\tau_i = E(Y_{ij}) - E(Y_{1j})$ represents the increased average response for treatment i over treatment 1.

Any such constraint will do but, in the balanced case, using $\sum_i \tau_i = 0$ makes the arithmetic easier so we will use this constraint.

To estimate the model parameters, we use least squares and a Lagrange multiplier to deal with the constraint. That is, ignoring the constraint, we find the critical values of the function,

$$V(\mu, \tau_1, \ldots, \tau_t, \lambda) = \sum_{i,j} (y_{ij} - \mu - \tau_i)^2 + \lambda \sum_i \tau_i$$

by setting to zero the partial derivatives with respect to $\mu, \tau_1, \ldots, \tau_t, \lambda$ and then solving the system of linear equations to find the parameter estimates. After some algebra that you should be able to reproduce, we have

$$\begin{aligned} \hat{\mu} &= \quad \overline{y}_{++}, \qquad \text{the overall average and} \\ \hat{\tau}_i &= \overline{y}_{i+} - \overline{y}_{++}, \quad \text{the treatment } i \text{ average – the overall average.} \end{aligned} \tag{2.1}$$

More importantly, $\hat{\tau}_i - \hat{\tau}_k = \overline{y}_{i+} - \overline{y}_{k+}$ is the estimate of the difference of the treatment effects. Note that this estimate is the difference of treatment averages.

Re-arranging equation (2.1) gives

$$\begin{aligned} \hat{\tau}_i &= \overline{y}_{i+} - \overline{y}_{++} \\ \hat{\tau}_i &= \overline{y}_{i+} - \hat{\mu} \\ \hat{\mu} + \hat{\tau}_i &= \overline{y}_{i+} \\ y_{ij} - (\hat{\mu} + \hat{\tau}_i) &= y_{ij} - \overline{y}_{i+}, \end{aligned} \tag{2.2}$$

and hence the estimate of the standard deviation is

$$\hat{\sigma} = \sqrt{\frac{\sum_{i,j}(y_{ij} - \hat{\mu} - \hat{\tau}_i)^2}{\underbrace{rt}_{\#\text{ observations}} - \underbrace{(1+t)}_{\#\text{ parameters}} + \underbrace{1}_{\#\text{ constraints}}}} = \sqrt{\frac{\sum_{i,j}(y_{ij} - \overline{y}_{i+})^2}{t(r-1)}} \tag{2.3}$$

with the equation on the numerators of line (2.3) holding because of line (2.2), with $t(r-1)$ degrees of freedom. We can assess the adequacy of the model informally by various plots involving the estimated residuals: $\hat{r}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i$.

For the battery example (denoting designs A, B, C, D, E as 1, 2, 3, 4, 5), we have (with $t = 5,\ r = 10$)

$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$	$\hat{\mu}$	$\hat{\sigma}$
0.130	-0.032	-0.034	-0.030	-0.035	1.238	0.126

with 45 degrees of freedom. We can compare any two designs i and k by assessing $\tau_i - \tau_k$. Note that the corresponding estimator is

$$\tilde{\tau}_i - \tilde{\tau}_k = \overline{Y}_{i+} - \overline{Y}_{k+} \sim G\left(\tau_i - \tau_k, \sigma\sqrt{\frac{1}{10} + \frac{1}{10}}\right) \quad \text{and} \quad \frac{45 \times \tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{45}.$$

Note for this example we have $t = 5$, $r = 10$ and $t(r-1) = 45$. We can use this estimator for formal analysis of questions involving the comparison of any two designs. The difference $\tau_i - \tau_k$ is an example of a **contrast** – see below.

2.1.1 Contrasts

A **contrast** is any linear combination of the treatment effects where the sum of the coefficients is zero. Two examples are:

1. $\tau_1 - \tau_2$: the difference between treatment 1 and 2
2. $\tau_1 - (\tau_2 + \tau_3)/2$: the effect of treatment 1 versus the average effect of treatments 2 and 3

For any contrast, we can construct tests of hypothesis or confidence intervals using the t distribution as in Stat 231. Note that for any contrast $\theta = \sum_i a_i \tau_i$ the corresponding estimate is $\hat{\theta} = \sum_i a_i \hat{\tau}_i = \sum_i a_i \overline{y}_{i+}$ since $\sum_i a_i = 0$. As shown by

$$\begin{aligned}\sum_i a_i \widehat{\tau}_i &= \sum_i a_i (\overline{y}_{i+} - \overline{y}_{++}) \\ &= \sum_i a_i \overline{y}_{i+} - \overline{y}_{++} \underbrace{\sum_i a_i}_{=0} \\ &= \sum_i a_i \overline{y}_{i+}.\end{aligned}$$

In the battery example, designs A and B are relatively expensive to produce. Is there any evidence that these batteries designs have more average energy than designs C, D and E?

The table of average battery energies is

A	**B**	**C**	**D**	**E**
1.368	1.206	1.204	1.208	1.203

Consider the contrast $\theta = (\tau_1 + \tau_2)/2 - (\tau_3 + \tau_4 + \tau_5)/3$. We want to see if there is evidence that $\theta > 0$. Note the form of this question – we want a **one-sided** test of significance. The corresponding estimator is

$$\begin{aligned}\tilde{\theta} &= \frac{\tilde{\tau}_1 + \tilde{\tau}_2}{2} - \frac{\tilde{\tau}_3 + \tilde{\tau}_4 + \tilde{\tau}_5}{3} \\ &= \frac{\overline{Y}_{1+} + \overline{Y}_{2+}}{2} - \frac{\overline{Y}_{3+} + \overline{Y}_{4+} + \overline{Y}_{5+}}{3}.\end{aligned}$$

The terms involving $\overline{Y}_{++}$ cancel because the sum of the coefficients is 0.
Note that, for every treatment average we have

$$\overline{Y}_{i+} \sim G\left(\mu + \tau_i, \frac{\sigma}{\sqrt{r}}\right)$$

so that the distribution of the estimator is

$$\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\frac{1}{40} + \frac{1}{40} + \frac{1}{90} + \frac{1}{90} + \frac{1}{90}}\right),$$

and we can carry out the test of hypothesis using a t-test.

1. The null hypothesis is $H_0 : \theta = 0$ with the alternative hypothesis $H_a : \theta > 0$.

2. The estimates are

$$\begin{aligned}\hat{\theta} &= 0.082, \text{ and} \\ \hat{\sigma} &= 0.126.\end{aligned}$$

3. The discrepancy measure is $d = \frac{\hat{\theta}-0}{\hat{\sigma}\sqrt{\frac{1}{12}}} = 2.25$.

   Note the lack of absolute value. We want the discrepancy to be large only if $\theta > 0$.

4. The p-value is $p = P(t_{45} \geq 2.25) = 0.015$.

5. There is strong evidence that $\theta > 0$.

We conclude that there is strong evidence that the higher priced batteries produce more energy on average. From the data, battery A is what makes the difference.

In general, a contrast has the form $\theta = \sum_i a_i \tau_i$ where $\sum_i a_i = 0$. The constraint on the coefficients ensures that the estimate $\widehat{\theta}$ is the same linear combination $\sum_i a_i \widehat{\tau}_i$ of the treatment averages. In terms of the estimator, we have

$$\tilde{\theta} = \sum_i a_i \overline{Y}_{i+} \sim G\left(\theta, \sigma\sqrt{\frac{\sum_i a_i^2}{r}}\right)$$

For any contrast, we can construct tests of significance or confidence intervals using the t distribution, as in Stat 231.

2.1.2 Analysis of Variance (ANOVA)

In the battery example, the first question is "Are there any differences among performances of the batteries?" In terms of the model parameters, we are asking if there are any differences among $\tau_1,\ \tau_2,\ \tau_3,\ \tau_4,\ \tau_5$. In formal terms, we want to test the hypothesis that there are no differences or mathematically that $\tau_1 = \ \tau_2 = \tau_3 = \ \tau_4 = \ \tau_5 = 0$ (remembering the constraint $\sum_i a_i = 0$). This hypothesis is different than any we have seen before because it involves restrictions on all five parameters simultaneously. We cannot use a t-test for this problem.

Here we develop a new hypothesis test called the **analysis of variance**. The idea is to estimate the residual standard deviation σ (actually σ^2) in two ways.

Estimate 1: From least squares, we have the estimate $\hat{\sigma}^2$ that is valid whether or not the hypothesis is true. This estimate depends only on the correctness of the overall model.

Estimate 2: Suppose that the hypothesis $\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0$ is true. We have $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$ where $\overline{y}_{++} = (\overline{y}_{1+} + \cdots + \overline{y}_{t+})/t$ (recall there are t treatments). Under the hypothesis, we know that $\overline{Y}_{i+} \sim G\left(\mu, \frac{\sigma}{\sqrt{r}}\right)$ (recall, there are r observations for each treatment), and hence we know that

$$\frac{\sum_i \hat{\tau}_i^2}{t-1} = \frac{\sum_i (\overline{y}_{i+} - \overline{y}_{++})^2}{t-1}$$

is an estimate of σ^2/r because the expression on the right is the square of the sample standard deviation of $\overline{y}_{1+}, \ldots, \overline{y}_{t+}$.

Finally we have that

$$r\frac{\sum_i \hat{\tau}_i^2}{t-1} = \frac{\sum_i r(\overline{y}_{i+} - \overline{y}_{++})^2}{t-1}$$

is an estimate of σ^2 when the hypothesis is true.

We use the ratio of these two estimates of σ^2 as the discrepancy measure in the hypothesis test. The ratio will be close to 1 if the hypothesis is true. It is convenient to lay out all of the algebra (or the corresponding numerical quantities) in the ANOVA table. See Table 1. Note that the mean square (ms) column gives the estimates of σ^2.

You can show easily that the total sum of squares is the sum of the treatment and residual sum of squares – see the exercises. This fact was useful when the quantities in the table were calculated by hand.

Table 2.1: ANOVA Table for Completely Randomized Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	$\sum_i r(\overline{y}_{i+} - \overline{y}_{++})^2$	$t-1$	$\frac{\sum_i r(\overline{y}_{i+} - \overline{y}_{++})^2}{t-1}$ (1)	$\frac{(1)}{(2)}$
Residual	$\sum_{i,j}(y_{ij} - \overline{y}_{i+})^2$	$t(r-1)$	$\frac{\sum_{i,j}(y_{ij} - \overline{y}_{i+})^2}{t(r-1)}$ (2)	
Total	$\sum_{i,j}(y_{ij} - \overline{y}_{++})^2$	$tr-1$		

The numerical ANOVA table for the battery example is given in Table 2.2.

The ratio of the treatment mean square to the residual mean square is larger than 1. To assess the size of the discrepancy, we calculate a p-value. That is, we look

at the corresponding estimators and see how often we get a discrepancy this large if the hypothesis is true.

In terms of the estimators, assuming the hypothesis is true, we have

Table 2.2: ANOVA Table for the Battery Example

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	0.212	4	0.053	3.35
Residual	0.713	45	0.016	
Total	0.925	49		

$$\frac{\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2/(t-1)}{\sum_{i,j}(Y_{ij} - \overline{Y}_{i+})^2/t(r-1)} \sim F_{t-1,t(r-1)}$$

Note that we have assumed without proof here that the numerator and denominator are independent random variables.

The **F-distribution** is a continuous probability distribution defined for non-negative values that has two associated degrees of freedom (for the numerator and denominator). Generally the F-distribution is skewed as illustrated in Figure 2.3.

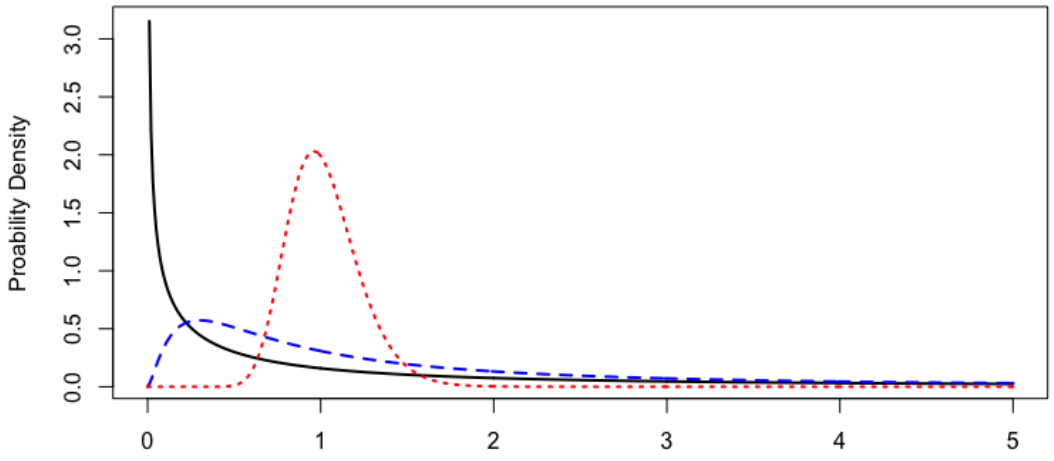


Figure 2.3: The Probability Density Functions for the F-distribution; Black solid line $F_{1,1}$, blue dashed line $F_{5,2}$, red dotted line $F_{100,100}$.

We can show, (see the exercises) whether or not the hypothesis is true, that

$$E\left[\frac{\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2}{t-1}\right] = \sigma^2 + r\frac{\sum_i \tau_i^2}{t-1}.$$

This means that if there are differences amongst the treatments, we expect to see large values of the discrepancy (as then $\sum_i \tau_i^2$ will be large). The p-value is

$$p = P(F_{t-1,t(r-1)} \geq \text{ observed discrepancy}).$$

Tables for the F distribution are provided with the course notes. For a small selection of probabilities p (0.10, 0.05, 0.01), the tables give the value of the constant c so that

$$P(F_{j,k} \geq c) = p.$$

The columns of the table give the numerator degrees of freedom and the rows the denominator degrees of freedom. For example

$$P(F_{4,50} \geq 2.56) = 0.05.$$

If the numerator or denominator degrees of freedom or the probability do not match those given in the table, you can interpolate to get the p-value.

In the battery example, the discrepancy measure has observed value 3.35. If the hypothesis is true, then the ratio of the corresponding estimators has an F distribution with 4 and 45 degrees of freedom so we have

$$p = P(F_{4,45} \geq 3.35) \approx 0.018.$$

There is some evidence of differences in average energy amongst the battery designs.

This so-called F-test is useful in more complicated experimental plans to look at hypotheses that involve two or more simultaneous restrictions on the parameters.

Notes

1. **Sample size:** We briefly discussed the issue of the number of replications required for each treatment in the discussion of the Plan. The key drivers behind choosing the sample size are

    - the question(s) being asked
    - the required precision of the conclusions
    - cost and ethical considerations.

Suppose the primary question deals with comparing two treatment effects. A confidence interval for the corresponding parameter $\theta = \tau_i - \tau_j$ is

$$\hat{\theta} \pm c \times \hat{\sigma}\sqrt{\frac{2}{r}}$$

The precision of the conclusion about θ can be measured by the width of the confidence interval, that is $2c\hat{\sigma}\sqrt{\frac{2}{r}}$. This width depends on the confidence level (helps to determine c), $\hat{\sigma}$ (and hence σ, the underlying residual standard deviation) and the number of replications per treatment r (determines the degrees of freedom and the multiplier in the standard deviation of the estimator).

If, before carrying out the Plan, we guess at $\hat{\sigma}$, perhaps based on similar investigations and have specified level of confidence, then we can determine r so that the confidence interval for θ will have specified length.

2. Multiple comparisons: We can ask any number of questions about linear combinations of the treatment effects, each of which corresponds to a contrast. We need to be careful if we then carry out a significance test for each contrast. It is easy to distort the conclusions because of the problem of **multiple comparisons**, a topic beyond the scope of this course. If the ANOVA shows that there are no differences among the treatments, then it is improper to search for contrasts that are significantly different.

3. Picking the best treatment: In the battery example, we were interested in picking the best and worst designs. To do so, we pick the designs that have the highest and lowest average energy. However, to carry out a formal procedure to test that these choices are correct is beyond what we can do in this course. Ranking of treatments based on the data is an interesting statistical problem.

Using R

The data are stored in the file *crd_example1.txt*. We can get the estimate of $\hat{\sigma} = 0.1259$ and the ANOVA with the following R commands

```
b <- lm(energy~battery)
summary(b)
anova(b)
```

and we obtain the following output

```
Call:
lm(formula = energy ~ battery)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2040 -0.0955 -0.0060  0.0695  0.3260

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.36800    0.03980  34.372  < 2e-16 ***
batteryB    -0.16200    0.05629  -2.878  0.00610 **
batteryC    -0.16400    0.05629  -2.914  0.00554 **
batteryD    -0.16000    0.05629  -2.843  0.00670 **
batteryE    -0.16500    0.05629  -2.931  0.00529 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1259 on 45 degrees of freedom
Multiple R-squared: 0.2293,     Adjusted R-squared: 0.1608
F-statistic: 3.347 on 4 and 45 DF,  p-value: 0.01757

Analysis of Variance Table

Response: energy
          Df  Sum Sq  Mean Sq F value  Pr(>F)
battery    4 0.21205 0.053012  3.3467 0.01757 *
Residuals 45 0.71281 0.015840
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In the output, the coefficients batteryB, batteryC, batteryD, batteryE, give the estimates of $\hat{\tau}_B - \hat{\tau}_A$,$\hat{\tau}_C - \hat{\tau}_A$, $\hat{\tau}_D - \hat{\tau}_A$, $\hat{\tau}_E - \hat{\tau}_A$, respectively. The easiest way to get the treatment averages is with R command

```
tapply(energy,battery,mean)
```

which outputs

```
  A      B      C       D      E
1.368  1.206  1.204   1.208  1.203
```

2.2 Model and Analysis (Unbalanced Plans)

A completely randomized design is **unbalanced** if the number of units per treatment is not the same for all treatments. The salmon experiment is unbalanced. In practice, we like to keep the plan balanced because all comparisons of two treatments have the same precision. We plan many investigations to be balanced but for some unexpected reason, units are lost and we are left with an unbalanced plan. We can have difficult interpretation problems if the treatment is the cause of the lost observations.

Suppose that we have r_i units for treatment i. We consider the model

$$Y_{ij} = \mu + \tau_i + R_{ij}, \qquad R_{ij} \sim G(0, \sigma), \qquad i = 1, \ldots, t, \ j = 1, \ldots, r_i$$

where we now change the constraint to $\sum_i r_i \, \tau_i = 0$. Some things to note.

- $E(Y_{ij}) = \mu + \tau_i$ represents the average response for treatment i.
- $E(Y_{ij}) - E(Y_{kj}) = \tau_i - \tau_k$ represents the difference in effects of treatments i and k.
- The constraint changes the interpretation of the parameter μ because $\mu = \sum_i r_i E(Y_{ij}) / \sum_i r_i$ is the **weighted average** of the treatment means. We are not interested in this parameter.
- The important point is that we interpret the treatment effects τ_i exactly as we did in the balanced case.

In a comparative investigation, all of the questions of interest are expressed in terms of contrasts formulated in terms of the τ_is.

To estimate the parameters, we apply least squares, subject to the constraint. That is, subject to $\sum_i r_i \tau_i = 0$, we minimize

$$\sum_i \sum_{j=1}^{r_i} (y_{ij} - \mu - \tau_i)^2$$

The parameter estimates are (see the exercises):

$$\hat{\mu} = \overline{y}_{++}, \qquad \hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}, \qquad \hat{\sigma} = \sqrt{\frac{\sum_i \sum_{j=1}^{r_i} (y_{ij} - \overline{y}_{i+})^2}{\sum_i (r_i - 1)}}.$$

Remarks:

1. The choice of the constraint ensures that the estimates of the treatment effects are the same as in the balanced design.

2. The degrees of freedom $\sum_i (r_i - 1)$ are accumulated within each treatment.
3. As well, note that a contrast defined on the treatment effects is a linear combination of the treatment averages as in the balanced case.

For the salmon investigation, we have $\hat{\mu} = 1.295$, $\hat{\sigma} = 0.315$ (with 67 degrees of freedom) and the estimated treatment effects in the following table:

treatment	treatment average	treatment effect	sample size
n0	1.394	0.099	8
n6	1.400	0.105	8
n24	1.223	-0.072	8
n96	1.078	-0.217	8
n336	1.179	-0.116	8
f0	1.346	0.051	8
f6	1.456	0.161	8
f24	1.200	-0.095	7
f96	1.518	0.223	8
f336	1.093	-0.201	6

We can better display the treatment averages on a scatterplot. From this plot, we see that the major difference between the strains occurs at the fourth time period. This picture tells the whole story except for dealing with the variation. Are the averages significantly different or not?

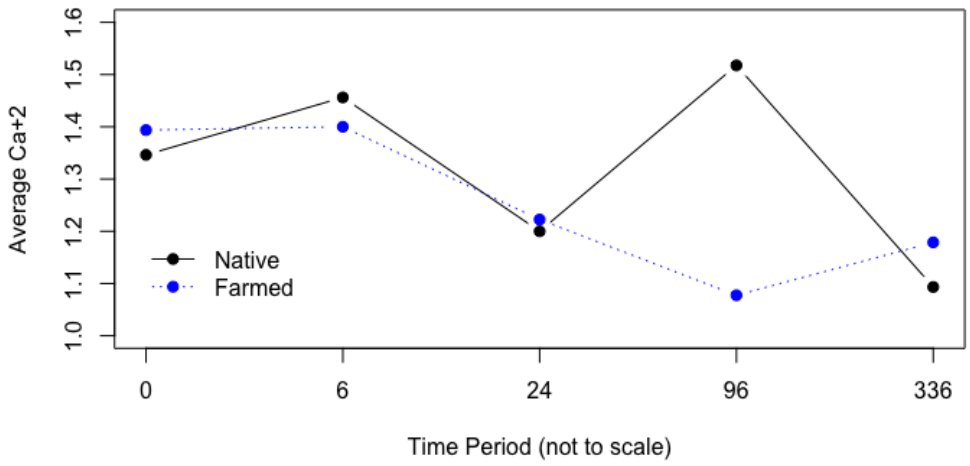


Figure 2.4: Average plasma concentration of the ion Ca^{+2} from farmed and native Atlantic salmon smolts versus time period; Black solid line native and blue dashed line farmed.

We can compare the average responses at 96 hours (i.e. time 4) by assessing the difference $\theta = \tau_4 - \tau_9$.

The estimate is $\hat{\theta} = \overline{y}_{4+} - \overline{y}_{9+} = -0.44$ and $\hat{\sigma} = 0.315$. The corresponding estimator is

$$\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\frac{1}{8} + \frac{1}{8}}\right)$$

and hence a 95% confidence interval for $\theta = \tau_4 - \tau_9$ is $-0.44 \pm 2.00(0.315)/2$ or -0.44 ± 0.32. Note that 0 does **not** lie in this confidence interval.

There is a relatively large difference in average levels of Ca^{+2} between the two strains at 96 hours. We can compare any other time periods in the same way.

For completeness, we present the ANOVA (table 2.3) for testing the hypothesis of no treatment differences. The basis for the test remains the same. We estimate the residual variance in two ways. First, using the full model, we get the estimate from least squares that does not depend on any hypothesis about the treatment effects. Second, if there are no treatment effects ($\tau_i = 0$ for every i), then the treatment mean square is an estimate of the residual variance (see the exercises). The ratio has an F distribution as before.

Table 2.3: ANOVA Table for Unbalanced Completely Randomized Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	$\sum_i r_i(\overline{y}_{i+} - \overline{y}_{++})^2$	$t-1$	$\frac{\sum_i r_i(\overline{y}_{i+}-\overline{y}_{++})^2}{t-1}$ (1)	$\frac{(1)}{(2)}$
Residual	$\sum_{i,j}(y_{ij} - \overline{y}_{i+})^2$	$\sum_i(r_i - 1)$	$\frac{\sum_{i,j}(y_{ij}-\overline{y}_{i+})^2}{\sum_i(r_i-1)}$ (2)	
Total	$\sum_{i,j}(y_{ij} - \overline{y}_{++})^2$	$\sum_i r_i - 1$		

The ANOVA table for the salmon investigation is shown in table 2.4 .

Table 2.4: ANOVA Table for the Salmon Example

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	1.6271	9	0.1808	1.76
Residual	6.8636	67	0.1024	
Total	8.4907	76		

The p-value is $P(F_{9,67} \geq 1.76) = 0.09$ so there is weak evidence of differences among the treatments.

Using R
Read the data as usual. The variate names are *concentration* and *treatment*. Missing values in R are assigned the symbol `NA`. Many functions will not work unless told to ignore the missing values with the parameter `na.rm=T`. For example to get the treatment means, use

```
fish = read.csv("Data/crd_example2.txt", sep="\t" )
attach(fish)
tapply(concentration, treatment, mean, na.rm=T)
```

To produce the ANOVA and the estimate of $\hat{\sigma}$ using

```
model <- lm(concentration ~ treatment)
anova(model)
summary(model)
```

2.3 Randomized Block Designs

In this section of the notes, we look at the comparison of t treatments when we form blocks. We consider the simplest situation with b blocks where we apply all t treatments in a random order within each block. Note that this plan is balanced. The blocks are groups of t units selected to have similar values of one or more explanatory variates. If feasible, we prefer a plan with blocking because we can better control confounding by holding an explanatory variate fixed within each block. We also expect that the precision of the comparison of any two treatments will be greater than that for the completely randomized design with b replications per treatment.

Example 3
A company planned to replace specialized software that is used throughout the organization. To decide which product to purchase, they first narrowed their choices to four available products (here called designs A-D) based on cost, reliability, service history and outside recommendations. To further help with the choice, the selection team then planned a trial to compare the four products in-house. The team selected 6 different tasks typical of those for which the software was used. They then chose 24 individuals from the company and assigned them to the tasks in groups of four at random. Within each group, one person used each product to complete the task. All individuals received a brief training program on the use of the software. The team measured the time to finish the task and also asked each person to complete a questionnaire about their personal reaction to the software. Here we concentrate on the times to complete the tasks, shown below in minutes. The data are stored in the file *rbd_example3.txt*.

task	brand	time	task	brand	time
1	A	6.5	4	A	13.4
1	B	10	4	B	12.9
1	C	5	4	C	12.1
1	D	6.8	4	D	15.6
2	A	14.1	5	A	6.6
2	B	14.5	5	B	10
2	C	14.4	5	C	6.8
2	D	15	5	D	8.8
3	A	9.9	6	A	8.2
3	B	13.2	6	B	7.4
3	C	10.5	6	C	6.9
3	D	12.2	6	D	11.4

The tasks are the blocks. The questions of interest are:

1. Are there differences among the brands?
2. Which brand has the best performance?

Note that we are not interested in the differences among the blocks. The questions are about treatment differences.

Model and Analysis
We write the model in general but the example has $t = 4,\ b = 6$.

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}, \qquad R_{ij} \sim G(0, \sigma), \qquad i = 1, \ldots, t, \quad j = 1, \ldots, b$$

where $\sum_i \tau_i = 0$ and $\sum_j \beta_j = 0$ are the constraints we choose so that all of the parameters are identifiable. This is a simple extension of the model we used with two treatments in Chapter 1. Applying least squares and using Lagrange multipliers, we get the same estimates as before.

$$\hat{\mu} = \overline{y}_{++}, \qquad \hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++} \qquad \hat{\beta}_j = \overline{y}_{+j} - \overline{y}_{++}$$

and

$$\begin{aligned}
\hat{\sigma} &= \sqrt{\frac{\sum_{i,j}[y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j]^2}{\underbrace{bt}_{\#\text{ observations}} - \underbrace{1}_{\text{parameter } \mu} - \underbrace{(t-1)}_{\text{parameters } \tau_i\text{, 1 constant}} - \underbrace{(b-1)}_{\text{parameters } \beta_j\text{, 1 constant}}}} \\
&= \sqrt{\frac{\sum_{i,j}[y_{ij} - \overline{y}_{++} - (\overline{y}_{i+} - \overline{y}_{++}) - (\overline{y}_{+j} - \overline{y}_{++})]^2}{bt - 1 - (t-1) - (b-1)}} \\
&= \sqrt{\frac{\sum_{i,j}(y_{ij} - \overline{y}_{i+} - \overline{y}_{+j} + \overline{y}_{++})^2}{(t-1)(b-1)}}
\end{aligned} \tag{2.4}$$

with $(t-1)(b-1)$ degrees of freedom. We obtain the simplified the denominator from line (2.4) as follows:

$$\begin{aligned} bt-1-(t-1)-(b-1) &= bt-1-t+1-b+1 \\ &= (bt-t)-(b-1) \\ &= t(b-1)-(b-1) \\ &= (t-1)(b-1). \end{aligned}$$

The sum of squares residuals is

$$\sum_{i,j}(y_{ij}-\overline{y}_{i+}-\overline{y}_{+j}+\overline{y}_{++})^2$$

and from (2.4), we have the relation

$$\sum_{i,j}(y_{ij}-\overline{y}_{i+}-\overline{y}_{+j}+\overline{y}_{++})^2=\sum_{i,j}[(y_{ij}-\overline{y}_{++})-(\overline{y}_{i+}-\overline{y}_{++})-(\overline{y}_{+j}-\overline{y}_{++})]^2,$$

we can expand the term on the right hand into the sum of 6 terms; three squares and three cross products.

The sum of squares of the treatment effects is

$$\sum_{i,j}(\overline{y}_{i+}-\overline{y}_{++})^2=b\sum_{i}(\overline{y}_{i+}-\overline{y}_{++})^2$$

and the cross product term

$$\begin{aligned} -2\sum_{i,j}(y_{ij}-\overline{y}_{++})(\overline{y}_{i+}-\overline{y}_{++}) &= -2\sum_{i}(\overline{y}_{i+}-\overline{y}_{++})\sum_{j}(y_{ij}-\overline{y}_{++}) \\ &= -2\sum_{i}(\overline{y}_{i+}-\overline{y}_{++})b(y_{i+}-\overline{y}_{++}) \\ &= -2b\sum_{i}(\overline{y}_{i+}-\overline{y}_{++})^2. \end{aligned}$$

We get the same pattern for the sum of squares of the block effects

$$\sum_{i,j}(-(\overline{y}_{+j}-\overline{y}_{++}))^2=t\sum_{j}(\overline{y}_{+j}-\overline{y}_{++})^2$$

and the cross product term

$$\begin{aligned} -2\sum_{i,j}(y_{ij}-\overline{y}_{++})(\overline{y}_{+j}-\overline{y}_{++}) &= -2\sum_{j}(\overline{y}_{+j}-\overline{y}_{++})\sum_{i}(y_{ij}-\overline{y}_{++}) \\ &= -2\sum_{j}(\overline{y}_{+j}-\overline{y}_{++})t(\overline{y}_{+j}-\overline{y}_{++}) \\ &= -2t\sum_{j}(\overline{y}_{+j}-\overline{y}_{++})^2. \end{aligned}$$

Finally the last cross product term is

$$2\sum_{i,j}(\overline{y}_{i+}-\overline{y}_{++})(\overline{y}_{+j}-\overline{y}_{++}) = 2\sum_{i}(\overline{y}_{i+}-\overline{y}_{++})\underbrace{\sum_{j}\left(\overline{y}_{+j}-\overline{y}_{++}\right)}_{=0} \quad (2.5)$$
$$= 0.$$

Show that $\sum_j \left(\overline{y}_{+j}-\overline{y}_{++}\right)=0$.

Now combining the results we have that the residual sums of squares equals

$$\sum_{i,j}[(y_{ij}-\overline{y}_{++})-(\overline{y}_{i+}-\overline{y}_{++})-(\overline{y}_{+j}-\overline{y}_{++})]^2$$
$$= \sum_{i,j}(y_{ij}-\overline{y}_{++})^2 - b\sum_{i}(\overline{y}_{i+}-\overline{y}_{++})^2 - t\sum_{j}(\overline{y}_{+j}-\overline{y}_{++})^2$$

We arrange these sums of squares in the ANOVA table as before.

Table 2.5: ANOVA Table for Completely Randomized Block Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	$b\sum_i(\overline{y}_{i+}-\overline{y}_{++})^2$	$t-1$	$\frac{b\sum_i(\overline{y}_{i+}-\overline{y}_{++})^2}{t-1}$ (1)	$\frac{(1)}{(2)}$
Blocks	$t\sum_j(\overline{y}_{+j}-\overline{y}_{++})^2$	$b-1$		
Residual	$\sum_{i,j}(y_{ij}-\overline{y}_{i+}-\overline{y}_{+j}+\overline{y}_{++})^2$	$(t-1)(b-1)$	$\frac{\text{Residual}}{(t-1)(b-1)}$ (2)	
Total	$\sum_{i,j}(y_{ij}-\overline{y}_{++})^2$	$tb-1$		

Note that the estimate $\hat{\sigma}$ is the square root of the residual mean square given in table 2.5.

As suggested by this table, we can test the hypothesis that there are no differences among the treatments i.e. $\tau_1 = \ldots = \tau_t = 0$, using the ratio of the treatment

mean square to the residual mean square as the discrepancy measure and the appropriate F distribution. We justify the test as before. From the least squares calculation, we know that the residual mean square is an estimate of σ^2 regardless of any hypothesis about the treatment effects. If the hypothesis $\tau_1 = \cdots = \tau_t = 0$ is true, then you can easily show that the estimators

$$\overline{Y}_{i+} \sim G\left(\mu, \frac{\sigma}{\sqrt{b}}\right)$$

(see the exercises) and that (as before)

$$\frac{b\sum_i(\overline{Y}_{i+} - \overline{Y}_{++})^2}{t-1} \sim \sigma^2 \frac{\chi^2_{t-1}}{t-1}.$$

We also have, for any contrast $\theta = \sum_i a_i\tau_i$ where $\sum_i a_i = 0$, the result

$$\tilde{\theta} = \sum_i a_i \overline{Y}_{i+} \sim G\left(\theta, \sigma\sqrt{\frac{\sum_i a_i^2}{b}}\right)$$

which we use to test hypotheses or build confidence intervals for θ.

Returning to the example, we have the treatment and block averages as shown below. We can easily calculate the appropriate sums of squares in the ANOVA table by first finding the sample standard deviations of the raw data, the treatment and block averages. Then we adjust these results appropriately.

	brand				
task	**A**	**B**	**C**	**D**	**average**
1	6.5	10	5	6.8	7.08
2	14.1	14.5	14.4	15	14.5
3	9.9	13.2	10.5	12.2	11.45
4	13.4	12.9	12.1	15.6	13.5
5	6.6	10	6.8	8.8	8.05
6	8.2	7.4	6.9	11.4	8.48
average	9.78	11.33	9.28	11.63	10.51

The treatment effects can be calculated directly from the last row of the data table. The completed ANOVA table is

From the ANOVA, the estimate of σ is $\hat{\sigma} = \sqrt{1.59} = 1.26$. Note that we are **not interested in the block effects**. Note that the standard output from R will calculate a mean square for tasks in the numerical ANOVA given above. We have left it out to emphasize that in the example we are not interested in testing the significance of the block effect. If we found that the block effect was small it would suggest that blocking was not that helpful in increasing the precision of the conclusions and perhaps was not worth the effect. In the software example we see

Table 2.6: ANOVA Table for the Software Example

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Brands	23.84	3	7.95	5.00
Tasks	190.94	5		
Residual	23.82	15	1.59	
Total	238.60	23		

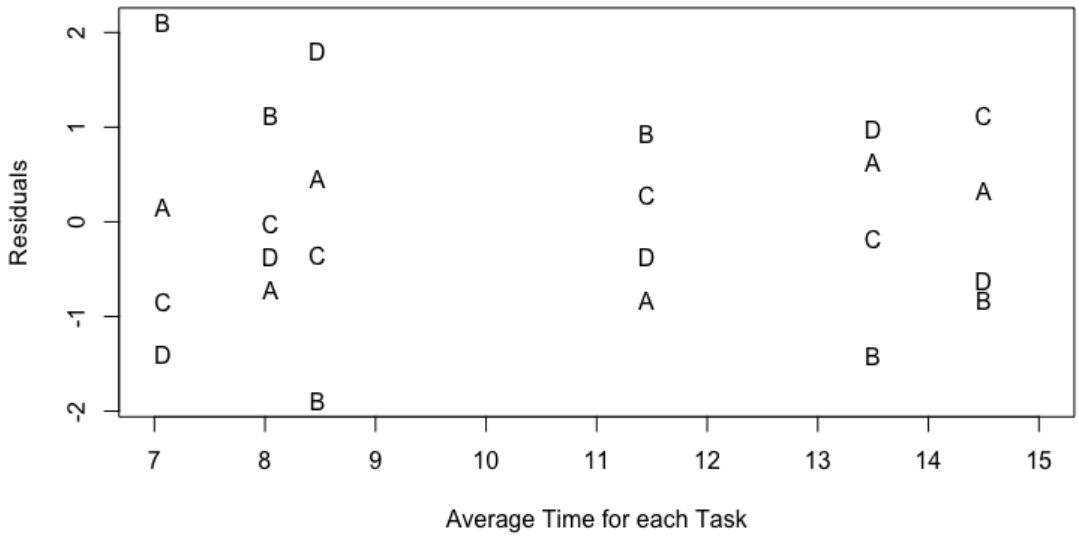
however that blocking was very effective.

The p-value for the hypothesis of no differences among the treatments is $P(F_{3,15} \geq 5.00) = 0.013$ so there is strong evidence of differences among the brands. Based on the table of averages, we conclude that brand C has the lowest average time to complete the tasks.

Notes

- Once we have more than two treatments, we cannot simplify the modeling and analysis of the block design by taking the treatment differences within each block.

- In the model, we assume that the treatment and block effects are additive. So in the example, we are assuming that the difference on average between any two brands is the same for each task. We might expect that the differences would be larger for more complex tasks. We can examine this issue informally by plotting the estimated residuals against the task average times. We use a separate symbol for each brand. Looking at the plot there is no pattern showing greater differences among the brands for tasks that take longer.



- Here we have looked at the simplest form of blocking. There are many variations on this theme.
  - In a field experiment, where we have 4 varieties to compare, we arrange the field in a square of 16 units so that we can simultaneously create blocks in two directions. We then assign the varieties to the units so that each variety appears in exactly one row and one column. This is called a **Latin Square design**. Here is one example.

	column 1	**column 2**	**column 3**	**column 4**
row 1	A	C	B	D
row 2	D	B	A	C
row 3	C	A	D	B
row 4	B	D	C	A

  - We may decide to replicate the entire plan on another day or at another site. In the software example, we might select another 24 people and run the entire experiment over on another day. We can consider the day as another blocking factor.
  - We may have many treatments and cannot build blocks large enough. In this case, we use an incomplete block design where not all treatments appear in every block.

- We can understand how blocking helps to avoid confounding since we hold one explanatory variate fixed within each block. In the example, if we had not blocked, we might have assigned more difficult tasks to one brand than another, by chance, and it may have then appeared that the brand that happened to get the easier tasks was better.

Blocking also helps to improve the precision of the comparison of the treatments. In the ANOVA table, we can see that the block mean square is quite large due to variation in the difficulty of the tasks. If we had not blocked, this variation would have been part of the estimate of the underlying standard deviation (which would thus have been much larger). Hence in comparing treatments, the estimators for contrasts of the treatment effects would have had larger variability.

Using R

We can use R to get the treatment averages, ANOVA and the estimate of $\hat{\sigma}$. For the example, the variate names are *task, brand, time*. Once you have read the data, the first step is to set *task* to be a factor (i.e. a categorical variate with 6 levels). The R code is

```
task <- factor(task)
model <- lm(time ~ brand+task)
summary(model)
anova(model)
```

The estimated residuals can be found using the command `resid(model)`.

2.4 Exercises

1. Practice with the F distribution.

   (a) Suppose $U \sim F_{6,24}$. Find c so that $P(U \geq c) = 0.05$.

   (b) Estimate $P(U \geq 3)$.

   (c) What is the distribution of $\frac{1}{U}$?

   (d) Find d so that $P(U \leq d) = 0.05$.

   (e) Show that if $W \sim t_k$ then W^2 has an F-distribution. What are the degrees of freedom?

2. In a small investigation, three treatments A,B,C were compared by assigning 6 units at random to each treatment. The data are in the file *crd_rbd_exercise2.txt.*

	A	B	C
	11.82	15.46	15.43
	13.03	12.90	14.28
	10.78	14.88	14.76
	14.31	13.75	12.07
	14.21	18.59	13.80
	8.56	13.80	13.46
average	12.12	14.90	13.97
st. dev.	2.22	2.02	1.16

(a) Calculate the ANOVA table.

(b) Is there any evidence of a difference among the treatments?

(c) Treatment A was a control. Is there any evidence that the average effect of treatments B and C exceeds the effect of the control? [Use a one-sided test here - why?]

(d) Find a 95% confidence interval for the difference between the effects of B and C.

3. Suppose we have a balanced design with t treatments and r units per treatment. If we represent the data by $y_{ij}, i = 1, \ldots t, j = 1, \ldots, r$, show algebraically that the sum of the treatments sum of squares and the residual sum of squares is the total sum of squares. That is, show that

$$\sum_i r(\overline{y}_{i+} - \overline{y}_{++})^2 + \sum_{i,j}(y_{ij} - \overline{y}_{i+})^2 = \sum_{i,j}(y_{ij} - \overline{y}_{++})^2.$$

Explain this expression in terms of the variation in the data.

4. Consider the model for the balanced completely randomized design

$$Y_{ij} = \mu + \tau_i + R_{ij}, \qquad R_{ij} \sim G(0, \sigma), \qquad i = 1, \ldots, t, \quad j = 1, \ldots, r$$

where $\sum_i \tau_i = 0$. Show that $E\left[\frac{\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2}{t-1}\right] = \sigma^2 + r\frac{\sum_i \tau_i^2}{t-1}$.

Hint: First show that the numerator can be written as $\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2 = \sum_i r\overline{Y}_{i+}^2 - rt\overline{Y}_{++}^2$ and then exploit the fact that for any random variable U, $E[U^2] = Var(U) + E(U)^2$.

5. For the imbalanced design with model

$$Y_{ij} = \mu + \tau_i + R_{ij}, \qquad R_{ij} \sim G(0, \sigma), \qquad i = 1, \ldots, t, \quad j = 1, \ldots, r_i$$

and constraint $\sum_i r_i \tau_i = 0$, show that

$$\hat{\mu} = \overline{y}_{++}, \qquad \hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}, \qquad \hat{\sigma} = \sqrt{\frac{\sum_i \sum_{j=1}^{r_i}(y_{ij} - \overline{y}_{i+})^2}{\sum_i (r_i - 1)}}.$$

6. In an experiment to compare five treatments A,B,C,D,E, 8 units were randomly assigned to each treatment. A partial ANOVA table for the data is shown below, along with the treatment averages.

Table 2.7: A partially completed ANOVA Table for Completely Randomized Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	3.505			
Residual				
Total	57.170			

treatment	**A**	**B**	**C**	**D**	**E**
average	10.06	10.41	9.95	9.86	9.49

(a) Complete the ANOVA table.

(b) Is there any evidence of a difference among the treatments?

(c) Having completed part 6b, a novice statistician looked at the table of averages and noticed that treatment B had the highest average and treatment E the lowest. She decided to see if treatment B was significantly greater, on average, than treatment E since those averages were farthest apart. Carry out the test of hypothesis.

(d) Explain why the test in part 6c is misleading.

7. In an industrial study, there were two factors, feed rate and temperature with two levels each (called low and high). For each of the four treatments, 12 parts were produced. The response variate of interest was surface finish, a measure of how smooth the part is. The data are given below. Note that the lower case letter in the definition of the treatment indicates the factor was at its low level. A partial ANOVA table is also given below. The data are given in the file *crd_rdb_exercise7.txt*.

	ft	fT	Ft	FT
	0.53	0.76	0.46	0.58
	0.35	0.75	0.62	0.4
	0.37	0.31	0.82	0.66
	0.47	0.66	0.67	0.51
	0.99	0.74	0.94	0.45
	0.41	0.73	0.4	0.73
	0.04	0.61	0.53	0.93
	0.32	0.56	0.79	0.89
	0.36	0.84	0.33	0.46
	0.62	0.62	0.7	0.36
	0.31	0.67	0.51	0.65
	0.34	0.68	0.74	0.79
average	0.426	0.661	0.626	0.618

Table 2.8: A partially completed ANOVA Table for Completely Randomized Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	0.4054			
Residual	1.5311			
Total	1.9360			

(a) Is there any evidence of a difference among the treatments?

(b) Consider the following questions:

- Does increasing the level of the feed rate effect increase the average surface finish?
- Does changing the level of temperature effect the average surface finish?
- Is the effect of changing the temperature the same for both levels of feed rate?

Construct appropriate contrasts and carry out the necessary hypothesis test to examine each of the questions.

(c) It was decided to run the process at the settings FT. Find a 95% confidence interval for the average surface finish at this setting.

(d) The treatments were not assigned to the units at random. Instead, the factor levels were set and 12 parts were run off. Briefly discuss the consequences of this decision as it affects your answers to parts 7a, 7b and 7c.

8. In an investigation to understand the impact of packaging on sales of a consumer product, a large number of new designs was considered and tested with small focus groups. As a result, four new designs (here called B,C,D and E) were considered as possibilities for further testing. Forty stores, all having close to the same historical sales, were available and it was decided to test each of the 5 designs (four new plus the current, denoted A) in eight stores each. The designs were assigned to the stores at random and the total sales for one week was selected as the response variate. The major questions of interest are:

   - Are there significant differences among the designs?
   - Are the new designs better on average than the current design?
   - Design E is predicted to be the best. What average sales can be expected from this design?

   Write an executive summary to address these questions. Be sure to include one table or graph that can be used to support your conclusions. Also include an appendix that provides the technical back-up for your conclusions. The data are stored in the file *crd_rdb_exercise8.txt.*

A	**B**	**C**	**D**	**E**
525	499	500	512	535
518	498	480	490	525
523	525	515	527	570
470	502	473	506	529
492	537	493	505	508
540	527	484	496	519
506	516	527	530	529
517	543	488	500	523

9. One of the assumptions that we have made since Stat 231 is that $\hat{\sigma}$, the square root of the sum of squares of the estimated residuals divided by the degrees of freedom, is an estimate of the standard deviation σ.

   (a) For any set of numbers $u_1, \ldots, u_n$ with average $\overline{u}$, show that $\sum_i (u_i - \overline{u})^2 = \sum_i u_i^2 - n\overline{u}^2$.

   (b) For the balanced completely randomized design, show that $E(\tilde{\sigma}^2) = \sigma^2$. [Hint: use the result from part 9a and the fact that for any random variable U, we have $Var(U) = E[U^2] - (E(U))^2$]

   (c) Suppose W is any positive random variable such that $E(W^2) = \lambda^2$. Prove that $E(W) \leq \lambda$ with equality if and only if W is constant.

   (d) Is $E(\tilde{\sigma}) = \sigma$?

10. Suppose you were asked to help plan an investigation to compare five types of road paint with respect to their durability. The response variate is the time from application of a test stripe on a highway until the paint has lost 90% of its "brightness". The rest of the planning team has no formal training in statistical planning.

    (a) Explain the notion of blocking in this context.

    (b) Why would you recommend blocking?

    (c) How might blocking be implemented?

11. A seed company has produced three new varieties of corn using genetic modification. They plan to field test these varieties in south-western Ontario using their current best seller as a control. The trial is for one growing season only. The response variate is the yield, measured in kilograms per hectare. The company has available 60 two hectare test plots scattered throughout the region.

    (a) How would you recommend that blocking could be used in this trial?

    (b) Discuss two reasons for blocking using this context.

12. In another packaging trial, an advertising firm wants to compare two features, colour and image to see if changing either factor impacts sales. There are 6 treatments (3 colours and two images). 30 stores are arranged in blocks of size 6 based on geographic location and within each block, the six treatments are assigned at random. The total sale over a two week period is the response variate. The data are shown below and given in the file *crd_rdb_exercise12.txt*.

	treatment						
block	**1**	**2**	**3**	**4**	**5**	**6**	**average**
1	1014	980	894	958	822	938	934.3
2	872	762	795	828	832	792	813.5
3	991	1133	941	1048	868	1054	1005.8
4	827	807	891	768	700	833	804.3
5	621	516	513	586	492	670	566.3
average	865.0	839.6	806.8	837.6	742.8	857.4	824.9

    (a) Complete the ANOVA table for this investigation.

    (b) Is there any evidence of a difference among the treatments?

    The factor levels for the treatments are shown below.

treatment	**colour**	**image**
1	red	1
2	red	2
3	blue	1
4	blue	2
5	green	1
6	green	2

(c) Find a 95% confidence interval for the contrast that compares the two images. What do you conclude?

(d) Is there any evidence that there is a difference in average sales for the two images if the colour is red? Is blue? Is green? What do you conclude?

13. I once read a textbook on experimental design that strongly recommended that you should check the significance of the block effects using the ANOVA and an appropriate F test.

(a) Using the data from Example 1 in this chapter, show how this can be done.

(b) The text then recommended that, if the block effects were significantly different, your investigation was poorly planned and you should start over. What do you think of this advice?

14. Starting with the randomized block model,

(a) Show that $\overline{Y}_{i+} \sim G\left(\mu + \tau_i, \frac{\sigma}{\sqrt{b}}\right)$

(b) Find the standard deviation of the estimator corresponding to the difference of two treatment effects $\tau_1 - \tau_2$.

Chapter 3

Factorial Treatment Structure and Interaction

In many investigations, the treatments have a **factorial** structure. That is, each treatment is defined by the levels of two or more **factors**. Here we restrict our attention to treatments defined by two factors and to balanced designs.

A key idea for such treatment structures is that of an **interaction**. We say that there is an **interaction** between two factors if the effect on the response of changing one factor depends on the level of the second. Note that interaction is defined in terms of effects on the response.

3.1 Two Factors At Two Levels

Example 1
A commercial flower grower sells cut roses to florists' shops. The grower is interested in finding a way to treat the roses so that they will last as long as possible. She plans a small investigation to look at two different ways of cutting the stems and two different solutions in which the cut stems are stored for shipping. The response variate is the time from cutting until the rose starts to lose petals, measured in days. The two factors define four treatments. The grower selects 32 stems from a single variety and, at random, allocates 8 stems to each treatment.

The questions of interest are:

1. Is there any evidence of a difference among the treatments?
2. Is the effect of the cutting method the same for each solution type?

The data (stored in the file *fac_example1.txt*), the treatment averages and the ANOVA table are shown below.

Treatment	Solution	Cut method	Time to petal loss	Treatment	Solution	Cut method	Time to petal loss
1	1	1	13.8	3	2	1	13.3
1	1	1	10.4	3	2	1	11.4
1	1	1	11.5	3	2	1	11.3
1	1	1	8.3	3	2	1	13.4
1	1	1	12.3	3	2	1	15.5
1	1	1	9.6	3	2	1	14.1
1	1	1	11.3	3	2	1	15
1	1	1	11.3	3	2	1	11.5
2	1	2	11.1	4	2	2	9.7
2	1	2	14.6	4	2	2	13.3
2	1	2	14.8	4	2	2	10.6
2	1	2	14	4	2	2	12.2
2	1	2	13.3	4	2	2	9
2	1	2	10.9	4	2	2	11.7
2	1	2	10.1	4	2	2	11.6
2	1	2	12.4	4	2	2	10.6

Treatment	**Solution**	**Cut method**	**Average**
1	1	1	11.06
2	1	2	12.65
3	2	1	13.19
4	2	2	11.09

Table 3.1: ANOVA Table for a Completely Randomized Design using the Commercial Flower Data.

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**	**p-value**
Treatments	28.353	3	9.451	3.53	0.027
Residual	74.896	28	2.675		
Total	103.250	31			

From the ANOVA table, we have $\hat{\sigma} = \sqrt{2.675} = 1.636$ with 28 degrees of freedom.

We use the ANOVA table to address the question about the differences among the treatments. Looking at the F ratio and the significance level, we see that there is some evidence of a difference among the treatments.

We can see from the table of averages, that for solution 1, changing the cutting method from 1 to 2 increases the average time whereas for solution 2, changing the cutting method from 1 to 2 decreases the average time. In other words, there appears to be interaction.

We can test this idea formally using the contrast $\theta = (\tau_4 - \tau_3) - (\tau_2 - \tau_1)$. The estimate is

$$\hat{\theta} = (11.09 - 13.19) - (12.65 - 11.06) = -3.69.$$

and the standard deviation of the corresponding estimator $\tilde{\theta}$ is

$$stdev\left(\tilde{\theta}\right) = \sigma\sqrt{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}.$$

To test the hypothesis that $\theta = 0$, the discrepancy (two-sided) is $d = \frac{|\hat{\theta}-0|}{\hat{\sigma}\sqrt{1/2}} = 3.19$ and the p-value is $P(|t_{28}| \geq 3.19) = 0.003$, so there is very strong evidence of an interaction.

We can demonstrate the interaction clearly if we re-organize the table of averages as follows.

	cutting method 1	**cutting method 2**	**average**
solution 1	11.06	12.65	11.86
solution 2	13.19	11.09	12.14
average	12.13	11.87	12.00

By looking at the rows, we can see explicitly the differing effects of changing the cutting method for each level of solution. Note that the idea of interaction is symmetric. We also see that there is a difference in the effects of changing solution for each cutting method. You can re-order the contrast

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

to make the same point.

Another way to display the interaction is the **interaction plot** that is a picture of the above table. Figure 3.1 displays the interaction plot for the commercial flower data. The horizontal axis is the level of one factor and the vertical axis is the average response. We plot the four averages and join with a straight line the two averages corresponding to the same level of the second factor. Non-parallel lines indicate the presence of interaction. Note that traditionally the interaction plot shows averages, however, it can be useful to also plot the individual values that make up the averages. Including the individual values allows us more easily visually assess how important the interaction effect is.

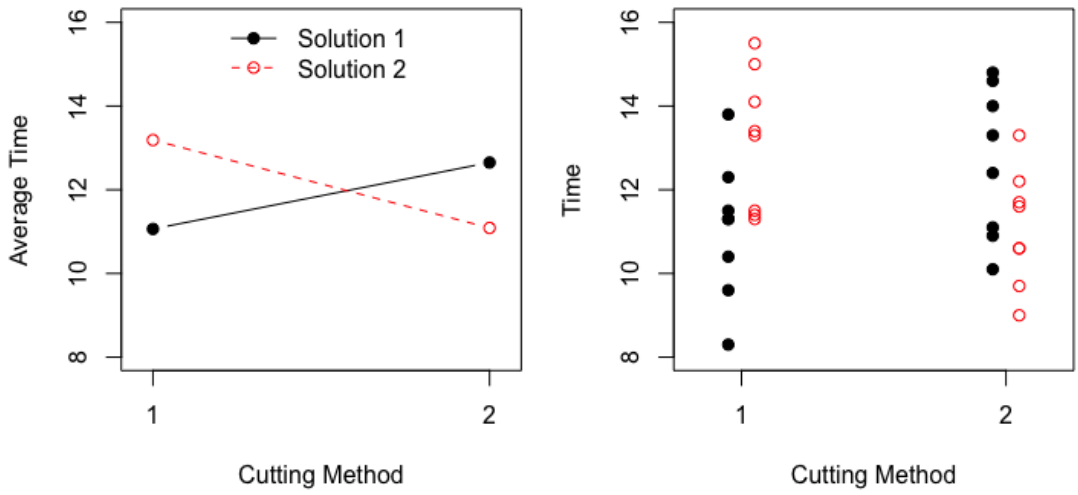


Figure 3.1: The interaction plot for average time to first petal drop for the factors solution and cutting method. The left panel has the average and the right panel displays all the points.

We can carry out the same analysis for any experimental plan with treatments specified by two factors with two levels each. The analysis is more complicated if we have more factors or two factors with one having more than two levels.

Re-create the interaction plot use the following R code.

```
flower <- read.table('fac_example1.txt', header=T)
attach(flower)
interaction.plot(method, sol, time,
    type='b', pch=c(19,1), col=1:2, lty=1:2)
```

3.2 Two Factors, One With Two Levels, One With Three Levels

Example 2
In Chapter 2, Exercise 12, we described a packaging trial with six treatments formed from two factors (three colours and two images). Thirty stores were arranged in blocks of size six based on geographic location and within each block, the six treatments were assigned at random. The total sales over a two week period was the response variate. The factor levels for the treatments are:

treatment	colour	image
1	red	1
2	red	2
3	green	1
4	green	2
5	blue	1
6	blue	2

Here the questions of interest (among others) were:

1. Is there evidence of a difference among the treatments?
2. Is the effect of changing images the same for each colour?

The data (stored in the file *fac_example2.txt*), the treatment and block averages and ANOVA table are shown below.

	treatment						
block	**1**	**2**	**3**	**4**	**5**	**6**	**average**
1	1014	980	894	958	822	938	934.3
2	872	762	795	828	832	792	813.5
3	991	1133	941	1048	868	1054	1005.8
4	827	807	891	768	700	833	804.3
5	621	516	513	586	492	670	566.3
average	865.0	839.6	806.8	837.6	742.8	857.4	824.9

Table 3.2: An ANOVA Table for the packing trail data using Completely Randomized Block Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**	**p-value**
Treatments	50548	5	10110	2.89	0.040
Blocks	672733	4	0.040		
Residual	69910	20	3496		
Total	793191	29			

Here $\hat{\sigma} = \sqrt{3496} = 59.1$ with 20 degrees of freedom. From the ANOVA, we see that there is some evidence of differences among the treatments.

In this example, we cannot use a single contrast to see if there is evidence of interaction since we are asking about how the effect of changing images changes over three different colours. No contrast corresponds to $\tau_2 - \tau_1 = \tau_4 - \tau_3 = \tau_6 - \tau_5$,

a two-dimensional comparison.

To deal with this issue formally, we partition the treatment sum of squares in the ANOVA table. We change the notation to keep track of the factor levels for each treatment. We write the model as

$$Y_{abj} = \mu + \tau_{ab} + \beta_j + R_{abj}, \qquad R_{abj} \sim G(0, \sigma)$$

where the subscript $a = 1, 2, 3$ indexes the colour, $b = 1, 2$ indexes the image factor levels and $j = 1, 2, 3, 4, 5$ indexes the blocks. The constraints are $\sum_{a,b} \tau_{ab} = 0$ and $\sum_j \beta_j = 0$.

In the original notation, we have the treatment effects $\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6$. With the new notation, we re-label the six treatment effects $\tau_{11}, \tau_{12}, \tau_{21}, \tau_{22}, \tau_{31}, \tau_{32}$.

Recall that in this example we have $r = 5$ blocks. Hence in the new notation, the treatment sum of squares is $5 \sum_{ab} (\overline{y}_{ab+} - \overline{y}_{+++})^2$.

To generate the partition of this sum of squares, consider the table of treatment averages.

Colour	**Image 1**	**Image 2**	**average**
Red	$\overline{y}_{11+}$	$\overline{y}_{12+}$	$\overline{y}_{1++}$
Green	$\overline{y}_{21+}$	$\overline{y}_{22+}$	$\overline{y}_{2++}$
Blue	$\overline{y}_{31+}$	$\overline{y}_{32+}$	$\overline{y}_{3++}$
average	$\overline{y}_{+1+}$	$\overline{y}_{+2+}$	$\overline{y}_{+++}$

We can pretend that this array corresponds to a randomized block design with two "treatments" (image 1 versus image 2) and 3 "blocks" corresponding to the colours. Recall we can decompose the residual sum of square (SS)

$$\begin{aligned}
& \text{SS Residual} \\
= & \sum_{ab} \left[(\overline{y}_{ab+} - \overline{y}_{+++}) - \underbrace{(\overline{y}_{+b+} - \overline{y}_{+++})}_{\text{treatments}} - \underbrace{(\overline{y}_{a++} - \overline{y}_{+++})}_{\text{blocks}} \right]^2 \\
= & \sum_{ab} \left[\overline{y}_{ab+} - \overline{y}_{+b+} - \overline{y}_{a++} + \overline{y}_{+++}) \right]^2 \\
= & \sum_{ab} (\overline{y}_{ab+} - \overline{y}_{+++})^2 - 3 \sum_{b} (\overline{y}_{+b+} - \overline{y}_{+++})^2 - 2 \sum_{a} (\overline{y}_{a++} - \overline{y}_{+++})^2,
\end{aligned}$$

so that

$$\underbrace{\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})^2}_{\text{SS Total}}$$
$$= \underbrace{\sum_{ab}\left[\overline{y}_{ab+}-\overline{y}_{+b+}-\overline{y}_{a++}+\overline{y}_{+++})\right]^2}_{\text{SS interaction}} + 3\underbrace{\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2}_{\text{SS image}} + 2\underbrace{\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2}_{\text{SS colour}}$$

The "total sum of squares" for the table of averages (ignoring the individual values in each block) is given by the last equation, and is proportional to the original treatment sum of squares. We can assign sources to each of the sum of squares on the right side. The first measures the interaction, the second the variation among the image averages and the third the variation between the colour averages. We list the terms in the ANOVA table in this order to remind you to look at the interaction first – if there is evidence of interaction, you must look at the factors together and the first and second terms in the breakup of the treatment sum of squares are not relevant.

Table 3.3: ANOVA Table for Constructed Block Design for Interactions

	Source	**Sum of squares**	**Degrees of freedom**
(Residual)	Interaction	$5\sum_{ab}\left[\overline{y}_{ab+}-\overline{y}_{+b+}-\overline{y}_{a++}+\overline{y}_{+++})\right]^2$	$(t-1)(b-1)$ $=(2-1)(3-1)$ $=2$
(Treatments)	Image	$15\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2$	$t-1$ $=2-1$ $=1$
(Blocks)	Colour	$10\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2$	$(b-1)$ $=3-1$ $=2$
(Total)	Treatments	$5\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})^2$	$tb-1$ $=(2)(3)-1$ $=5$

Recall that in the model for the randomized block design, we assumed that the treatment effects were the same for each block. In other words we assumed that there was no interaction between the treatments and the blocks. The interaction sum of squares in the above table corresponds to the residual sum of squares in

the randomized block ANOVA. If there is no interaction, the corresponding interaction mean square is an estimate of σ^2.

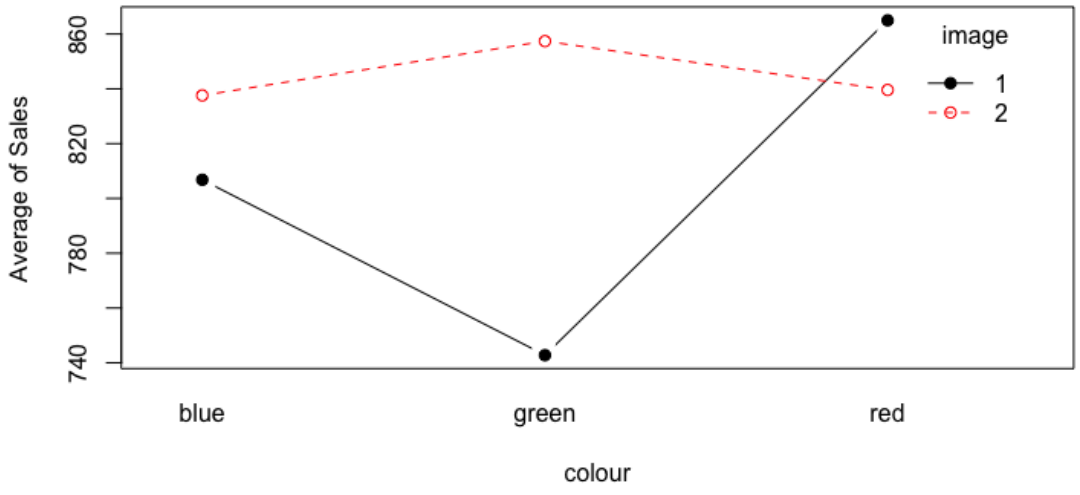
For the example, the ANOVA table with the treatment sum of squares partitioned is shown below. We can test the hypothesis of no interaction by comparing the Interaction mean square to the residual MS. If the hypothesis is true, then the corresponding ratio of estimators has an $F_{2,20}$ distribution. Here we have that $p - value = P(F_{2,20} \geq 3.55) = 0.048$.

Table 3.4: ANOVA Table for Constructed Block Design for Interactions

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (MS)**	**F**	**p-value**
Treatments	50548	5	10110	2.89	0.040
-Interaction	24817	2	12409	3.55	0.048
-Image	12000	1	12000	3.43	0.166
-Colour	13731	2	6865	1.96	0.079
Blocks	672733	4	168183	48.11	
Residual	69910	20	3496		
Total	793191	29			

Note that we have indented the degrees of freedom and sum of squares for the components of the treatment to indicate that these are sub-totals.

From the ANOVA, we conclude that there is evidence of interaction between colour and image. The interaction plot shows that changing the colour has little effect for image 2 but a large effect for image 1.



A Note About the Analysis:
In any experiment where the treatment structure depend on multiple factors, once an interaction is detected, you MUST LOOK AT THE FACTORS TOGETHER. It does not make sense to consider the factors separately (i.e. the so-called main effects) because the effect of one factor on the response variate depends on the level of the second. In the analysis, the first step is to look for evidence of a treatment effect. If there is such evidence, next we look for evidence of interaction. If there

- appears to be an interaction, then look at the interaction plot or the corresponding table of averages and interpret the treatment effects using that plot
- is no evidence of an interaction, then the factor effects can be examined separately. We can use the appropriate F test applied to the corresponding row of the ANOVA table with the treatment sum of squares partitioned as above. The main effects of each factor can be displayed by calculating the average response for each level of the factor.

Advantages of Using Treatment with Factorial Structure
In many situations where experiments are used, the original problem can be formulated in terms of assessing the effect of several factors on the response variate. There are two strategies:

- carry out a sequence of simple one factor experiments holding all other factors fixed
- use a single more complicated factorial plan in which all factors are varied within the same study

The first so-called "one-at-a-time" strategy is intuitively appealing because it seems that the effects of the factors will be confounded if you try to change them simultaneously. You need to be able to convince people that this simple strategy is not only inefficient but will not be able to detect interactions.

Suppose we have two factors A and B at two levels low and high. With the one-at-a-time approach, we start by fixing B at its low level and carrying out a simple experiment say with four replicates at each level of A. We can now assess the effect of factor A by comparing the treatment averages. Next, we fix A at its low level and investigate factor B with a similar eight run experiment. Now we can assess factor B. Note that you might only use 4 runs in the second experiment since, in the first, you already had four runs with A and B at their low levels. So we have at least 12 runs. We can summarize the average responses in a table. Note that each entry is the average of four observations.

	B low	**B high**
A low	$\overline{y}_{low\ low}$	$\overline{y}_{low\ high}$
A high	$\overline{y}_{high\ low}$	

There is no information in the two experiments about the interaction between A and B without the missing cell average. This is the primary selling point for the factorial strategy. To discover interactions we must vary the factors in a factorial structure.

Now suppose we carry out an eight run experiment with four treatments defined by the combinations of levels of A and B. Now we have observations in all four cells of the table and we can estimate the interaction. The factorial approach is more informative.

If the interaction is small, we can compare the main effects of A (and B) by comparing the high and low averages, each with four replicates. That is, we can estimate the main effects of the factors with the same precision as in the one-at-a-time approach using 8 instead of 12 runs. The factorial approach is more efficient.

Using R
To specify the main effects and an interaction term in the model, we use the * notation. In example 2, we write the model as

```
model <- lm( sales ~ block + colour + image + colour*image)
```

Then we apply the commands `anova()` and `summary()` as usual. We can produce an interaction plot with the function

```
interaction.plot(colour, image, sales, type='b',
    pch=c(19,1), col=1:2, lty=1:2, ylab="Average of Sales")
```

3.3 Exercises

1. A large organization is planning a massive retraining of its employees in the use of Office programs. As part of the planning, the HR department considers two factors that might impact the success of the training.

   - Factor 1: use of internal versus external trainers
   - Factor 2: use of interactive computer-aided materials versus non-interactive materials

   To investigate these two factors, a pilot experiment is organized in which 10 employees are randomly assigned to one of the four treatments. Two weeks after the training, each subject in the experiment is given a standard test.

   (a) What does it mean to say that the two factors interact?

(b) The HR director pointed out that the cost of the pilot could be substantially reduced if only three treatments were used.

- Treatment 1: internal trainers, non-interactive materials
- Treatment 2: internal trainers, interactive materials
- Treatment 3: external trainers, non-interactive materials

The effect of the use of interactive materials can be assessed by comparing treatment 2 to treatment 1. The effect of using external trainers can be assessed by comparing treatment 3 to treatment 1. Write a careful explanation to the Director explaining the drawbacks to this suggestion.

2. The manufacturer of a "frost-free" refrigerator found that in a high humidity, high temperature environment frost did build up inside the fridge. To remedy the problem, a new design was developed and four prototypes were built. In an experimental investigation, the four prototypes and four standard fridges were tested in two environments, one normal and one with high temperature and humidity. Frost build-up was measured after one week's operation. A lower score is better. The data are stored in the file *fac_exercise2.txt* and shown below.

design	**condition**	**response**	**design**	**condition**	**response**
new	extreme	1.35	old	extreme	2.56
new	extreme	1.63	old	extreme	2.51
new	extreme	1.43	old	extreme	2.22
new	extreme	1.57	old	extreme	2.35
new	normal	1.27	old	normal	1.45
new	normal	1.44	old	normal	1.67
new	normal	1.53	old	normal	1.34
new	normal	1.40	old	normal	1.47

Based on the standard model, the estimate of the residual standard deviation is $\hat{\sigma} = 0.133$. The table of treatment averages is shown below.

	New	**old**	**average**
normal	1.41	1.48	1.45
extreme	1.50	2.41	1.95
average	1.45	1.95	1.70

(a) Is there any evidence of differences among the treatments?

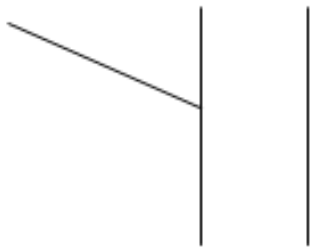
(b) Is there any evidence of interaction?

(c) Prepare an interaction plot.

(d) Use the interaction plot to argue that the new design is less sensitive to changes in environmental conditions (temperature and humidity). What limitations apply to this argument?

(e) Explain why we can not assess the individual effects of temperature and humidity on frost buildup in this investigation.

3. In the study of an optical illusion (called the Poggendorf effect), 15 subjects were given the same 9 diagrams in random order. Each diagram was a full page version



of the above picture. The subject was asked to predict where the straight diagonal line would meet the second parallel. The response variate was the distance (in mm) from the actual point of intersection to the predicted point. There were two factors: distance between the parallel lines (3 levels) and the angle of the diagonal (3 levels). The data are given in the file *fac_exercise3.txt* and shown below.

The treatments are labeled 1 to 9 according to the code

Width/angle	40°	30°	20°
4 cm	1	4	7
5 cm	2	5	8
6 cm	3	6	9

Based on the full model, the estimate of the residual standard deviation is 3.95.

subj-				treatment						
ject	1	2	3	4	5	6	7	8	9	avg
1	14.2	7.6	11.0	7.3	9.0	16.4	15.4	19.1	22.2	13.58
2	7.3	10.4	10.4	15.0	12.5	12.9	4.6	12.6	16.4	11.34
3	6.5	13.1	15.7	9.7	14.3	18.8	10.8	20.7	15.1	13.86
4	10.8	9.4	16.5	-1.1	4.8	11.7	14.5	9.7	15.3	10.18
5	6.7	11.3	15.6	8.5	9.4	16.0	13.2	14.3	17.2	12.47
6	4.9	5.4	21.0	14.0	7.9	14.3	13.4	13.8	14.1	12.09
7	10.9	8.0	13.7	13.7	8.8	12.1	8.1	19.3	17.7	12.48
8	12.5	11.6	11.9	14.1	11.9	13.7	9.6	7.6	10.1	11.44
9	7.1	13.6	10.3	11.8	13.4	8.7	6.8	7.2	19.1	10.89
10	6.2	12.8	15.2	13.7	15.0	14.5	15.6	22.6	17.4	14.78
11	17.5	11.9	3.5	12.0	11.2	13.7	8.9	14.4	14.4	11.94
12	6.5	11.6	11.8	11.4	9.8	9.9	10.3	8.2	18.7	10.91
13	19.6	15.3	8.8	6.1	14.1	14.0	17.3	8.4	20.0	13.73
14	11.0	13.3	12.4	12.1	0.8	5.0	8.2	10.6	8.6	9.11
15	17.6	10.1	16.3	3.5	9.6	8.9	13.5	10.5	15.5	11.72
avg	10.62	11.03	12.94	10.12	10.17	12.71	11.35	13.27	16.12	12.03

(a) Construct the ANOVA table including the partition of the treatment sum of squares into main effects and interaction components.

(b) Is there any evidence of interaction? Prepare an interaction plot to help interpret your answer.

(c) What can you say about the effects of changing the angle and width of the diagram in light of your answer to part 3b?

4. For Example 2 in Chapter 3, verify algebraically that the treatment sum of squares (without the multiplier 5) can be partitioned as

$$\begin{aligned} & \sum_{ab}(\overline{y}_{ab+} - \overline{y}_{+++})^2 \\ = \; & 2\sum_{a}(\overline{y}_{a++} - \overline{y}_{+++})^2 + 3\sum_{b}(\overline{y}_{+b+} - \overline{y}_{+++})^2 + \sum_{ab}(\overline{y}_{ab+} - \overline{y}_{a++} - \overline{y}_{+b+} + \overline{y}_{+++})^2 \end{aligned}$$

5. A company that makes products for dentists carried out an experiment with the objective to better understand the factors that affect the pain patients feel after a ceramic inlay is glued to a prepared tooth. The trial involved two factors, the type of glue (formulation A,B,C) and the addition of a short term desensitizer under the inlay. Based on the results of the questionnaire given to each patient, post-operative pain was assessed on a scale of 1-10. The experiment used blocking in that 12 dentists were involved in the trial and the dentists carried out all six treatments on a set of their patients, randomly assigned to the six treatments. The data are given in the file *fac_exercise5.txt* and are shown below along with some numerical summaries.

	treatment						
dentist	**1**	**2**	**3**	**4**	**5**	**6**	**average**
1	1.0	4.4	2.4	3.9	4.2	3.1	3.17
2	2.2	4.4	4.4	3.5	5.1	3.1	3.78
3	4.2	1.9	2.1	5.3	3.0	3.9	3.40
4	3.0	3.4	3.4	2.8	2.4	4.7	3.28
5	4.7	6.2	5.1	4.3	5.5	6.5	5.38
6	5.4	8.9	5.3	4.8	5.1	4.9	5.73
7	3.1	5.1	2.6	2.4	2.1	2.9	3.03
8	2.5	3.4	4.2	2.3	3.9	2.3	3.10
9	3.5	2.9	1.6	3.0	2.1	2.2	2.55
10	4.6	4.7	4.2	2.6	3.5	3.8	3.90
11	3.9	6.6	3.6	2.4	4.0	4.9	4.23
12	3.5	4.4	3.0	3.6	2.9	5.0	3.73
average	3.47	4.69	3.49	3.41	3.65	3.94	3.77

The first two treatments are brand A, the second two brand B and the third two brand C with the first treatment of each brand having the desensitizer absent. The estimate of the residual standard deviation is $\hat{\sigma} = 1.01$.

(a) Write out a model to describe the repeated application of the Plan. Briefly explain each term in the model.

(b) Is there any evidence of a difference among the treatments?

(c) For practice, find a 95% confidence interval for the effect $\tau_1 - \tau_2$.

(d) Partition the treatment sum of squares into components for assessing the interaction and the effects of the two factors.

(e) Is there any evidence of interaction?

(f) Construct an interaction plot.

(g) Does it make sense to assess the factor effects separately here? Explain.

6. Experiments with factorial structure of the treatments are widely used to improve manufacturing processes. For example, in a casting operation, the defect rate in the current process is 5%. To reduce the rate of defects, the process engineer has a large number of process parameters (factors) that he can adjust. To keep matters simple, suppose that there are only two factors, pouring temperature (T) of the iron and the level of an innoculant (I) that is added to the iron as it is poured.

   Engineers are often taught that in such situations, the best strategy is to vary one factor at a time. The purpose of this question is to convince you that this is not the case.

(a) Suppose that the engineer decides to investigate T at two levels (above and below the current process setting) holding I fixed. He plans an experiment with 16 runs, 8 at each level of temperature. The response variate is the defect rate for a run. Build a model for this plan and write down the estimate of changing T and the standard deviation of the corresponding estimator.

The engineer then plans to use the best level of T and repeat the plan in part 6a to look at I.

(b) The statistician recommends a 16 run factorial experiment using all four treatments (4 replicates per treatment). Build a model for this plan. Show that if there is no interaction that this plan is superior for estimating the effects of changing I and T.

(c) Show using a numerical example (a table of averages, for example) that if there is interaction, the plan in a) may not get to the optimal combination of the two factors.

(d) Briefly summarize why the single factorial plan is better than the two one-at-a-time plans.

Chapter 4

Sample Survey Issues

In the second half of the course, we consider the planning and analysis of simple sample surveys. For the most part, we follow the book "Sampling: Design and Analysis" by S.L. Lohr. We will cover most of the material in Chapters 1-4. There are multiple copies of this book in the Davis Centre Library. See also the reference list attached to the course outline.

In this chapter, we deal with

- the language of sample surveys
- examples of sampling protocols
- classification of error (the difference between the estimate and the attribute)
- assessment of error

Sample surveys are widely used to estimate attributes of interest in a specified target population.

The survey can be one-time only and informal (e.g. internet polls on the many websites) or regular and highly complex (e.g. the Canadian Labour Force survey that estimates unemployment rates across Canada on a month to month basis. For details, see search for labour force survey on Statistics Canada's website www.statcan.gc.ca).

Surveys are used to estimate attributes of human populations as in the above examples and also any other collection of objects such as financial records.

A census is an investigation of a population where we try to examine every unit. The reasons for using a sample survey rather than a census of the target population to learn about attributes are

- cost
- timeliness
- ethical issues relating to efficient use of resources

- the improved quality of the estimates available from a carefully conducted survey rather than a sloppy census.

We use some specialized language to describe survey methodology within the PPDAC framework. Problem Plan, Data, Analysis and Conclusion. See Stat 231 for a complete description of the PPDAC framework. For formal surveys, we concentrate on the sampling protocol. Note that we often select units in clusters to implement a sampling protocol as illustrated in the example below.

4.1 Example

In the Labour Force Survey, the target population is defined as:

> "The target population is the non-institutionalized population 15 years of age and over. The survey is conducted nationwide, in both the provinces and the territories. Excluded from the survey's coverage are: persons living on reserves and other Aboriginal settlements in the provinces; full-time members of the Canadian Armed Forces, the institutionalized population, and households in extremely remote areas with very low population density. These groups together represent an exclusion of less than 2% of the Canadian population aged 15 and over."

The sampling protocol does not choose units (people who meet the inclusion criteria) directly. Instead, a sample of households is selected and then variates are measured on every appropriate unit in the selected households.

> "The LFS uses a probability sample that is based on a stratified multi-stage design. Each province is divided into large geographic stratum. The first stage of sampling consists of selecting smaller geographic areas, called clusters, from within each stratum. The second stage of sampling consists of selecting dwellings from within each selected cluster. "

We call the households the **sampling units**. The **frame** is the list of sampling units on which the sampling protocol operates. The frame defines the study population. Developing a good frame (one which covers the target population) is often one of the most expensive components of conducting the survey. Formal surveys have a frame. Note that informal surveys such as internet polls do not use a frame since the units are self-selecting. In this instance, the study population is only vaguely specified.

In the Labour force Survey, there are separate frames for each stage of the sampling. One frame is a list of clusters (small geographic areas) within each large

geographic stratum. The second frame is a list of households within each selected cluster.

4.2 Sampling Protocols

There are many sampling protocols that can be used to select the sample from the study population.

A **probability sampling** protocol uses a probability distribution to select the sample from the frame. More formally, if the frame is denoted by $U = \{1, 2, \ldots, N\}$, then a probability sampling protocol assigns a probability to every subset S of U and the sample is selected according to this distribution. We will look at ways to implement such a protocol later.

Example: Suppose an auditor has a file of 1220 records and plans to select a sample of 20 records to examine the quality of the file. The auditor decides to use *simple random sampling* (SRS), a protocol in which all samples of size 20 have the same probability of being selected. Here we write the frame as $U = \{1, 2, \ldots, 1220\}$ and, for any subset S of U, we have

$$P(S) = \begin{cases} \frac{1}{\binom{1220}{20}} & \text{if } S \text{ has size } 20 \\ 0 & \text{otherwise} \end{cases}$$

Example: For the labour force survey, the sampling protocol is described as:

> "The LFS uses a probability sample that is based on a stratified multi-stage design. Each province is divided into large geographic stratum. The first stage of sampling consists of selecting smaller geographic areas, called clusters, from within each stratum. The second stage of sampling consists of selecting dwellings from within each selected cluster."

There are many non-probability sampling protocols. Some examples are:

- **convenience sampling** – "take what you can get", e.g. a survey of people in a mall by a marketing firm
- **self-selection sampling** – units choose themselves, usually with little control, e.g. many internet polls
- **quota sampling** – units are selected so that some attributes of the sample match known attributes in the target population, e.g. in a marketing survey, each interviewer is directed to find a sample whose attributes match the local population in terms of age, income profile and gender.

- **judgment sampling** – units are selected so that the samplers think that the sample will be representative of the target population (i.e. match the target population with respect to the attributes of interest).

We concentrate on formal surveys that use probability sampling protocols since we can use mathematical tools to assess, at least partially, the error that occurs in drawing conclusions about the target population from the sample. This is a major advantage of these sampling protocols.

4.3 Errors

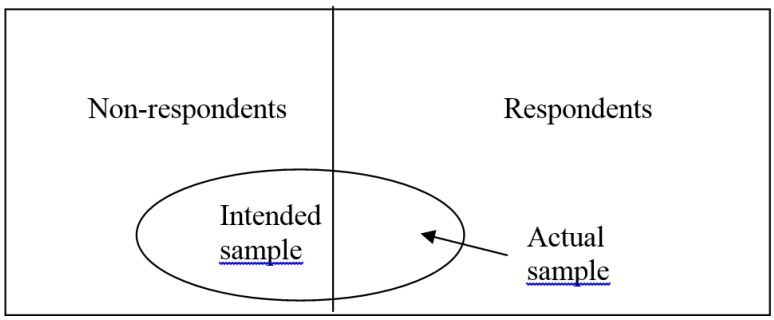
In applying PPDAC to estimate population attributes, we can classify errors as:

- **Study error:** the difference in attributes of interest between the target and study population.
- **Sample error:** the difference in attributes of interest between the study population and the sample.
- **Measurement error:** the difference in the attributes of interest due to the difference between the true and measured values of the variates on the units in the sample.

In the context of sample surveys, study error is called **frame error**. The attributes of the units listed in the frame may not match those of the target population.

For surveys of human populations, an important component of sample error is **non-response error**. Suppose the attributes of interest in the respondent and non-respondent populations are different. Then the sample attributes may not match those in the frame because one or more units in the sample may have refused to provide data.

If we divide the frame into those units that would respond and those that would not, we can see the effect of non-response error.



The actual sample is different from the intended sample and the attributes in the actual sample may not match those in the frame.

Measurement error may occur because of systematic differences in interviewers, people in the sample may lie, forget or modify their answers to please the interviewer. Interviewers may influence the responses by using a different protocol for asking the questions. Measurement error may also occur if the question we pose does not match the question used to define the response variate in the target population.

We have all seen statements such as "19 times out of 20, a sample of this size is accurate to within 3 percentage points" at the bottom of the conclusions from a survey. This confidence interval captures the uncertainty due to a component of the sample and measurement error. We use the probability model that generated the sample to describe how the sample attributes would behave if we were to repeat the same sampling protocol over and over. The confidence interval does not capture uncertainty due to frame error, non-response, systematic errors in the sampling protocol and measurement system etc. We can control these latter sources of error only through good planning and execution of the survey.

4.4 Questionnaire Design

Here is a brief set of considerations in designing the instrument (the questionnaire) for the survey of a human population. This is a very complex subject. There are many books and papers written on questionnaire design. If you are involved in an important survey, hire an expert.

The proper design of the questionnaire and a good plan for its administration can substantially reduce non-response error, recall error, error due misunderstanding the question and so on.

The following list is adapted from Lohr pp 10-15.

- Decide what you want to find out (understand the Problem)
- Keep the questions clear and simple
- Use specific instead of general questions
- Decide whether to use open-ended or closed questions
- Ask only one concept in each question
- Use forced choice rather than agree/disagree questions
- Avoid leading questions and contexts

- Relate each question to your objective – what will you do with the data?
- Keep the questionnaire short.
- Explain the purpose of the survey
- Ensure confidentiality
- Pay attention to question-order effects
- Test your questions before the survey
- Plan to report the actual questions used

Chapter 5

Probability Sampling

Formal surveys use probability sampling, a protocol that selects units for the sample based on a probability model on subsets of the frame. The major advantage of probability sampling is that the sampling protocol produces a statistical model that we can use to assess sample error, i.e. to generate confidence intervals and hypothesis tests for model parameters that represent attributes of interest in the study population. If we execute the protocol as planned, then we know that the model is appropriate. In this chapter, we first examine several probability sampling protocols and then look at simple random sampling (SRS) in detail.

Denote the **frame** by the set $U = \{1, 2, \ldots, N\}$ so that there are N units in the frame. Then a **probability sampling protocol** specifies the probability that the sample is s for any subset of $s \subset U$. We consider protocols where the sample size n is fixed so that the only subsets with positive probability have n units.

Here are some common sampling protocols explained in terms of an example.
Example: Suppose $N = 10000$ and $n = 100$.

- **Simple random sampling**: all $\binom{10000}{100}$ samples of size 100 have the same probability.
- **Stratified random sampling**: divide the frame into sub-frames called strata. For example,

  $$U_1 = \{1, \ldots, 10000\}, \ldots, U_{10} = \{9001, \ldots, 10000\}.$$

  For each **stratum**, select a simple random sample of size 10. There are $\binom{1000}{10}^{10}$ possible samples, each with the same probability.
- **Cluster sampling**: Divide the frame into clusters, for example

  $$C_1 = \{1, \ldots, 10\}, C_2 = \{11, \ldots, 20\}, \ldots, C_{1000} = \{9991, \ldots, 10000\}.$$

Select 10 clusters using simple random sampling. The sample is the 100 units in the 10 selected clusters. There are $\binom{10000}{100}$ possible samples, each with the same probability.

- **Systematic sampling**: Define clusters with $n = 100$ units per cluster

  $C_1 = \{1, 101, \ldots, 9901\}, C_2 = \{2, 102, \ldots, 9902\}, \ldots, C_{100} = \{100, 200, \ldots, 10000\}$.

  Use simple random sampling to select one cluster as the sample. We call this protocol systematic because we can select the sample by choosing the first unit from $\{1, \ldots, 100\}$ at random and then taking every subsequent 100^{th} unit. There are 100 possible samples, each with the same probability.

- **Two-stage sampling**: Select the sample in two stages. For example,

  - Stage 1: Select two strata (here called **primary units**) from the 10 described in stratified sampling above using simple random sampling. There are $\binom{10}{2}$ possible samples of primary units.
  - Stage 2: Select 50 units from each of the two selected primary units using simple random sampling.

  There are $\binom{10}{2}\binom{1000}{50}^2$ possible samples, each with the same probability. One advantage of two stage sampling is that we only need to build a frame at the second stage for those primary units selected in the first stage.

The **inclusion probability** p_i for any unit i in the frame is the probability that the unit is included in the sample. In the above example, you can show that $p_i = \frac{1}{100}$ for each unit i for each of the described sampling protocols – see the exercises.

You should be able to provide definitions of these sampling protocols in general. Note that complex surveys such as the labour force survey use multi-stage sampling with stratification in the primary stage and cluster sampling (the clusters are households) in the ultimate stage. Also note that we use SRS within each of the above protocols so we need to understand the properties of this most important protocol.

5.1 Simple Random Sampling (SRS)

For simple random sampling, we select n units from a frame of N units so that each sample of size n has the same probability of selection (recall, the probability equals 0 if the size is not equal to n). Since there are $\binom{N}{n}$ possible samples, each

has probability $\frac{1}{\binom{N}{n}}$. The inclusion probability for unit i in the frame is

$$p_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!((N-1)-(n-1))!}}{\frac{N!}{n!(N-n)!}} = \frac{(N-1)!n!(N-n)!}{N!(n-1)!(N-n)!} = \frac{n}{N}.$$

Note that **the numerator is the number of samples that contain the particular unit**. To select a sample containing unit i, we select the remaining units from the remaining pool of $N-1$ units in the frame. Technically, this protocol is often called **simple random sampling without replacement** since we do not allow the same unit to be included in the sample more than once. We use the shortened form of the name.

As we saw in the above example, there are many sampling protocols with the same inclusion probabilities as SRS. Therefore we **cannot** define simple random sampling by saying that every unit has the same chance of being included in the sample.

Let y_i be the value of the response variate for unit i. Suppose that we are interested in estimating the average response in the target population. We denote this average in the frame (study population) by μ and the sample average by $\hat{\mu} = \sum_{i \in s} y_i/n$, where s is the selected sample. Similarly, we denote the standard deviation in the frame by σ and the sample standard deviation by

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

where $\hat{r}_i = y_i - \hat{\mu}$ is the estimated residual for the i^{th} unit.

We do not build a response model here as we did in the earlier part of the course. Instead, we use the probability mechanism that generated the sample to look at the properties of the estimates if we were to repeat the sampling over and over. That is, we define the estimators

$$\begin{aligned} \tilde{\mu} &= \frac{\sum_{i \in S} y_i}{n} \\ \tilde{\sigma} &= \sqrt{\frac{\sum_{i \in S}(y_i - \tilde{\mu})^2}{n-1}} \end{aligned}$$

where S is a random subset with $P(S = s) = \frac{1}{\binom{N}{n}}$ for every subset $s \subset U$ of size n. It is convenient to re-express the estimators in terms of random variables rather than a random subset. Define indicator variables

$$I_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i = 1, \ldots, N.$$

Then, we can write

$$\tilde{\mu} = \frac{\sum_{i \in S} y_i}{n} = \frac{\sum_{i \in U} I_i y_i}{n}$$

and similarly,

$$\tilde{\sigma} = \sqrt{\frac{\sum_{i \in U} I_i (y_i - \tilde{\mu})^2}{n-1}}$$

in terms of the indicator random variables $I_1, \ldots, I_N$. Note that the sums are over the entire frame U.

We cannot calculate the exact distribution of the estimator $\tilde{\mu}$. However we can find many of its properties. We have the following important results:

For simple random sampling of n units from a frame of N units, we have

$$E(\tilde{\mu}) = \mu, \qquad stdev\,(\tilde{\mu}) = \sqrt{1 - \frac{n}{N}} \left(\frac{\sigma}{\sqrt{n}} \right)$$

We call $\tilde{\mu}$ an **unbiased estimator** of μ since $E(\tilde{\mu}) = \mu$. To prove this statement, note that for the indicator variable we have $P(I_i = 1) = \frac{n}{N}$ so

$$E\left[I_i\right] = 0 \times (1 - \frac{n}{N}) + 1 \times \frac{n}{N} = \frac{n}{N},$$

and hence we have

$$E(\tilde{\mu}) = E\left[\frac{\sum_{i \in U} I_i y_i}{n} \right] = \frac{\sum_{i \in U} E(I_i) y_i}{n} = \frac{\sum_{i \in U} \left(\frac{n}{N}\right) y_i}{n} = \mu.$$

To prove the formula for $stdev(\tilde{\mu})$, we need the result from Stat 230 that, for any linear combination of dependent random variables $V_1, \ldots, V_n$, we have

$$Var\left[\sum_i a_i V_i \right] = \sum_i a_i^2 Var(V_i) + \sum_{i \neq j} a_i a_j Cov(V_i, V_j).$$

where $Cov(V_i, V_j) = E[V_i V_j] - E(V_i)E(V_j)$.
Applying this result, we have

$$Var(\tilde{\mu}) = \frac{1}{n^2} \left[\sum_{i \in U} y_i^2 Var(I_i) + \sum_{i \neq j, i, j \in U} y_i y_j Cov(I_i, I_j) \right].$$

Since I_i is an indicator random variable, we have

$$\begin{aligned} Var(I_i) &= E(I_i^2) - E(I_i)^2 \\ &= P(I_i = 1) - P(I_i = 1)^2 \\ &= \frac{n}{N}\left(1 - \frac{n}{N}\right). \end{aligned}$$

To find the covariance of I_i and I_j, we need to find $E[I_iI_j]$. Since the product is zero unless $I_i = 1$ and $I_j = 1$, we have

$$E(I_iI_j) = P\,(\text{ units } i \text{ and } j \text{ are both in the sample}) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

so the covariance is

$$\begin{aligned} Cov(I_i, I_j) &= E[I_iI_j] - E(I_i)E(I_j) \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \\ &= -\frac{n}{N}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}. \end{aligned}$$

Combining the above results we get

$$\begin{aligned} Var(\tilde{\mu}) &= \frac{1}{n^2}\left[\frac{n}{N}(1-\frac{n}{N})\sum_{i\in U} y_i^2 - \frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i\neq j,\, i,j\in U} y_iy_j\right] \\ &= \left(\frac{1}{n}\right)\left(1-\frac{n}{N}\right)\frac{1}{N}\left[\sum_{i\in U} y_i^2 - \frac{\sum_{i\neq j,i,j\in U} y_iy_j}{N-1}\right] \end{aligned}$$

We can simplify the expression inside the braces with a bit of algebra.

$$\begin{aligned} \sum_{i\in U} y_i^2 - \frac{\sum_{i\neq j,i,j\in U} y_iy_j}{N-1} &= \frac{1}{N-1}\left[(N-1)\sum_{i\in U} y_i^2 - \sum_{i\neq j,i,j\in U} y_iy_j\right] \\ &= \frac{1}{N-1}\left[N\sum_{i\in U} y_i^2 - \left(\sum_{i\in U} y_i^2 - \sum_{i\neq j,i,j\in U} y_iy_j\right)\right] \\ &= \frac{1}{N-1}\left[N\sum_{i\in U} y_i^2 - \left(\sum_{i\in U} y_i\right)^2\right] \\ &= \frac{1}{N-1}\left[N\sum_{i\in U} y_i^2 - N^2\mu^2\right] \\ &= \frac{N}{N-1}\left[\sum_{i\in U} y_i^2 - N\mu^2\right] \\ &= N\sigma^2, \end{aligned}$$

so, at last, we have

$$\begin{aligned} Var(\tilde{\mu}) &= \left(\frac{1}{n}\right)\left(1-\frac{n}{N}\right)\frac{1}{N}\left[\sum_{i\in U} y_i^2 - \frac{\sum_{i\neq j, i,j\in U} y_i y_j}{N-1}\right] \\ &= \left(\frac{1}{n}\right)\left(1-\frac{n}{N}\right)\frac{1}{N}\left[N\sigma^2\right] \\ &= \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} \end{aligned}$$

and so, at last, we have

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

as required.

The formula for the standard deviation of an average is the basic result in the sampling part of the course. We use the result repeatedly so it is worthwhile learning it.

The factor $1-\frac{n}{N} = 1-f$ is called the **finite population correction factor (fpc)** where $f = \frac{n}{N}$ is the **sampling fraction**, the proportion of the population units included in the sample. In most cases f does not matter, as it is very small. The standard deviation of the the estimator corresponding to the sample average, using the model generated by simple random sampling, is the usual standard deviation for an average, $\frac{\sigma}{\sqrt{n}}$, multiplied by the square root of the finite population correction factor. The factor arises because the terms in the sum that defines the estimator are dependent. If the sampling fraction is appreciable, then there can be a significant reduction in the standard deviation. **For many applications, the sampling fraction is negligible and we ignore the fpc.**

We can also show that $E\left(\tilde{\sigma}^2\right) = \sigma^2$. See the exercises.

To use the above results, we need one more fact, a version of the **Central Limit Theorem** for the average of a sequence of dependent random variables. Here we state one of the consequences of the theorem avoiding all technicalities.

If N, n and $N-n$ are suitably large, then

$$\frac{(\tilde{\mu}-\mu)}{\sqrt{1-f}\frac{\tilde{\sigma}}{\sqrt{n}}} \sim G(0,\,1), \text{ approximately.}$$

We use this result to build confidence intervals for various attributes of interest.

Example 1
An auditor has a file of stated counts and prices for $N = 1256$ items stored in a warehouse of a producer of small automotive parts. The total stated value was \$4,311,712. The auditor planned to use SRS to select a sample of item numbers and then physically count the actual number of instances of those selected items. The purposes of the sampling were to assess:

i) the true total value of the inventory

ii) the average dollar error per item

iii) the proportion of items with counts in error

The auditor selected a simple random sample of $n = 50$ items and re-counted (actually a pair of co-op students did the counts). The first 10 lines of the data are shown below and the complete data set is available in the file *inventory.txt*

The average actual value of the sampled items was \$2895.29 with corresponding sample standard deviation \$1997.22. There were 11 items with count errors in the sample. The sample average dollar error (actual value – stated value) was \$2.33 and the sample standard deviation of the dollar errors was \$41.93.

Table 5.1: Part of the Sampled Data

item	stated.number	item.price	stated.value	actual.number	actual.value
1	1335	0.61	814.35	1335	814.35
25	1192	1.64	1954.88	1192	1954.88
39	1294	1.50	1941.00	1294	1941.00
53	1269	2.04	2588.76	1269	2588.76
56	1427	3.24	4623.48	1419	4597.56
121	1529	2.81	4296.49	1529	4296.49
207	1446	2.48	3586.08	1446	3586.08
212	1106	1.13	1249.78	1106	1249.78
223	847	4.95	4192.65	847	4192.65
225	1016	10.27	10434.32	1016	10434.32

i) To estimate the total actual value τ, we start by estimating the population average μ and then use the fact that

$$\begin{aligned} \mu &= \frac{\tau}{N} \\ \Leftrightarrow \tau &= N\mu. \end{aligned}$$

If y_i is the actual value of item number i, then we estimate μ using the sample average $\hat{\mu} = 2895.29$. To assess the precision of this estimate, we find an approximate 95% confidence interval based on the approximation to the distribution of $\tilde{\mu}$ described above. The form of the interval is the same as usual.

$$\hat{\mu} \pm c \times \text{standard deviation (estimate)}$$

where the standard deviation is the estimate of the standard deviation of the estimator $\tilde{\mu}$. Here the standard error is

$$\sqrt{1 - \frac{n}{N}} \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{1 - \frac{50}{1256}} \frac{1997.22}{\sqrt{50}} = 276.77.$$

For a 95% confidence interval, we use $c = 1.96$ from the standard Gaussian (last row of the t tables). Substituting, the confidence interval for μ is 2895.29 ± 542.47.

The average actual value is poorly estimated.

We are interested in the total actual value $\tau = N\mu$ and we can get a 95% confidence interval for τ by multiplying the above interval by $N = 1256$. The interval is

$$3636484 \pm 681341$$

This interval is very wide meaning we have estimated the total value of the inventory very imprecisely. Compared to the stated value of \$4,311,712, it is difficult to assess whether or not there are material errors in the inventory (as the stated total lies in the 95% confidence interval). One possibility is to increase the sample size, but see below.

ii) We use the same methodology to estimate the average error. The sample average error is $\hat{\mu}_{error} = 2.33$ with corresponding sample standard deviation 41.93. A 95% confidence interval for the average error is

$$\hat{\mu}_{error} \pm c \times \text{estimated standard deviation, with } c = 1.96 \text{ as before}$$

where the standard error is $\sqrt{1 - \frac{50}{1256}} \frac{41.93}{\sqrt{50}} = 5.81$. Hence the 95% confidence interval for the average error is 2.33 ± 11.39.

The average error more precisely estimated than is the average actual value. We can exploit this result and the fact that we know the total stated value to get a better estimate of the total actual value. Since the error was defined as the actual value minus the stated value, we estimate the true total value

as the total stated value ($4,311,712) plus the total dollar error. Recall that

$$\begin{aligned} \text{error} &= \text{actual} - \text{stated} \\ \Leftrightarrow \text{actual} &= \text{stated} + \text{error} \end{aligned}$$

The estimate of the total error is

$$\hat{\tau}_{error} = N\hat{\mu}_{error} = 1256 \times 2.33 = \$2926.48$$

so the 95% confidence interval for τ_{error} is $1256 \times (2.33 \pm 11.39)$ or 2926 ± 14306.

A 95% confidence interval for the total true value of the inventory is then

$$\underbrace{4311712}_{\text{stated}} + \underbrace{2926}_{\text{error}} \pm \underbrace{14306}_{\text{CI half width}} \quad \text{or } 4314638 \pm 14306.$$

The confidence limits are about 0.3% of the estimated total so the difference between the stated and actual value of the inventory is likely immaterial. Note by exploiting the known information (the total stated value) and the fact that we can estimate the average error with much smaller standard error than the average actual value, we can get a much more precise estimate of the total actual value. This procedure is called **difference estimation**.

iii) To estimate the proportion of items with counts in error, we define the **indicator variables**

$$z_i = \begin{cases} 1 & \text{if the } i^{th} \text{ item is in error} \\ 0 & \text{otherwise} \end{cases}$$

The attribute of interest is $\pi = \sum_{i \in U} z_i/N$, the population average of the binary variate. We can use the same theory as above. The sample average is $\hat{\pi} = \frac{11}{50}$ and the sample standard deviation is

$$\begin{aligned} \sqrt{\frac{\sum_{i \in s}(z_i - \overline{z})^2}{49}} &= \sqrt{\frac{\sum_{i \in s} z_i^2 - 50\overline{z}^2}{49}} \\ &= \sqrt{\frac{\sum_{i \in s} z_i - 50\hat{\pi}^2}{49}} \\ &= \sqrt{\frac{50\hat{\pi}(1-\hat{\pi})}{49}} \\ &= 0.418 \end{aligned}$$

Note for a binary response variate, the sample standard deviation is a function of the sample average proportion. An approximate 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\sqrt{1 - \frac{50}{1256}}\frac{0.418}{\sqrt{50}} \quad \text{or} \quad 0.22 \pm 0.11$$

There is considerable evidence that more than 10% of the item counts are in error, though the average error is likely to be small and the total error immaterial.

Notes for example 1

- We use the same theory to get estimates and approximate confidence intervals for averages, totals and proportions. The form of the interval for these attributes is always

$$\text{estimate} \pm c \times \text{standard deviation (estimate)}$$

- In the above example, the finite population correction factor $1 - f = 1 - \frac{50}{1256} = 0.96$ played a very small role (especially as it enters the calculations as $\sqrt{1-f} = 0.98$) and we could have safely ignored it.

- For binary response, the sample standard deviation is $\sqrt{\frac{n\hat{\pi}(1-\hat{\pi})}{n-1}} \approx \sqrt{\hat{\pi}(1-\hat{\pi})}$. We usually ignore the factor $\sqrt{\frac{n}{n-1}}$ when we apply this formula.

Example 2

To assess the quality of a shipment of 2000 cartons of headlights (packed 12 to a carton), a manufacturing organization decides to select a sample of 30 cartons using SRS for inspection. The attribute of interest is the proportion of headlights that are defective. A headlight is declared defective if it fails to pass any one of a large number of tests. The data (number of defective items per sampled carton) are shown below.

0	1	1	4	0	0	0	0	2	3
0	0	3	2	0	0	0	1	1	0
0	0	2	2	0	1	0	1	0	0

The sample average and standard deviation are

$$\hat{\mu} = 0.80 \qquad \text{and} \qquad \hat{\sigma} = 1.13.$$

Here we are using cluster sampling with the clusters defined as the cartons. If μ is the average number of defectives per carton, then the proportion of defective items in the population is

$$\pi = \frac{2000\mu}{12(2000)} = \frac{\mu}{12}. \tag{5.1}$$

A 95% confidence interval for μ is

$$\hat{\mu} \pm 1.96\sqrt{1 - \frac{30}{2000}}\frac{\hat{\sigma}}{\sqrt{30}} \text{ or } 0.80 \pm 0.40.$$

Hence using line (5.1), a 95% confidence interval for π is

$$0.067 \pm 0.033 \text{ or } 6.7\% \pm 3.3\%.$$

The proportion of defective headlights is poorly estimated but is significantly larger than zero.
Note how we adapt the results from SRS to apply to cluster sampling.

5.2 Sample Size Determination

We can use the same theory to answer the most common question in Statistics.

"How large a sample do I need?"

The obvious answer is "What is your objective?" It may take some effort to elicit a specific response but with some guidance, you can determine the target population, the attributes of interest, a possible frame and the required precision for the estimates. Since we will select only one sample, we base the sample size determination on the attribute of primary interest. We suppose that we can state the precision in terms of the length of a confidence interval for this attribute. See the Exercises for another formulation of the precision in terms of relative error.

Problem: Suppose that we are interested in estimating a population average μ and we want a confidence interval for μ of length 2ℓ, i.e. the confidence interval should be $\hat{\mu} \pm \ell$. Using SRS, we have

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

and solving for the sample size

$$n = \left(\frac{1}{N} + \frac{\ell^2}{c^2\hat{\sigma}^2}\right)^{-1}$$

To determine the sample size, n, we need to specify the confidence level to find c and, with more difficulty, to guess the value of $\hat{\sigma}$. If the $\frac{\ell^2}{c^2\hat{\sigma}^2}$ term on line (5.2) is much larger than the $\frac{1}{N}$ term, then we can omit the $\frac{1}{N}$ term, and take the required sample size to be approximately $\frac{c^2\hat{\sigma}^2}{\ell^2}$. The answer is very sensitive to the value of σ. In other words, we often do now know enough to give a good answer to this

question.

One way to get an idea of $\hat{\sigma}$ is to carry out a small pilot survey. We get an estimate of $\hat{\sigma}$ to help determine the sample size in the main survey and we also can use the pilot study to test the questionnaire and the rest of the proposed methodology. Sometimes we can use the results of previous surveys with similar response variates on the same population to get an idea of the value of $\hat{\sigma}$.

Example 1
In the audit example, suppose that the above description was a pilot survey and the overall goal was to estimate the average error with 95% confidence within plus or minus one dollar. How many more items do we need to include in the sample? Here we have

$$\begin{aligned} \ell &= 1 \\ c &= 1.96 \\ N &= 1256 \text{ and} \\ \hat{\sigma} &= 41.93 \end{aligned}$$

from the initial survey. To achieve the required precision, we have

$$n = \frac{1}{\frac{1}{1256} + \frac{1}{1.96^2 \times 41.93^2}} = 1059.$$

Here, because $\hat{\sigma} = 41.93$ is so large, we are forced to examine an extra 1009 items to achieve the desired precision. Since this is most of the frame, we would likely recommend a complete census.

Example 2
A polling firm has been hired to conduct a cross-Canada survey to solicit opinions from adults on a number of issues. The primary question has a Yes/No answer and the sample size is selected based on estimating the proportion of adult Canadians π who would answer Yes to the question. The client asks for a confidence interval of length 5 percentage points (0.05) with 99% confidence so we have

$$\begin{aligned} \ell &= 0.025 \\ c &= 2.57. \end{aligned}$$

The estimated standard deviation will be

$$\sqrt{\frac{n\hat{\pi}(1-\hat{\pi})}{n-1}} \approx \sqrt{\hat{\pi}(1-\hat{\pi})}.$$

The required sample size is

$$\begin{aligned} n &= \frac{1}{\frac{1}{N} + \frac{0.025^2}{2.57^2 \times \hat{\pi} \times (1-\hat{\pi})}} \\ &\approx \frac{2.57^2 \hat{\pi} \times (1-\hat{\pi})}{0.025^2} \\ &= 10568 \times \hat{\pi}(1-\hat{\pi}) \end{aligned}$$

where we ignore the term $\frac{1}{N}$ since it is so small. Here the required sample size is bounded because the function $\hat{\pi}(1-\hat{\pi})$ has maximum value of 0.25 when $\hat{\pi} = 0.5$. We know that if we choose $n = 10568 \times 0.25 = 2642$, we will meet the requirements. If we have a better idea of $\hat{\pi}$ from a pilot survey or elsewhere, we may be able to reduce the sample size from this upper bound.

Note that these sample size determinations do **not** take frame error, non- response error and other such errors into account.

5.3 When and How to Implement SRS

Here we briefly look at when SRS should be used and how to implement the sampling protocol.

- To implement SRS, we need a frame, U, for the target population of interest. If the frame consists of a list of items or people, we can assign each unit a unique number from 1 to N and then use available software to select a sample of n units using simple random sampling. In R, the command

  ```
  s <- sample(u,n)
  ```

  selects a random sample of size `n` from the vector `u` and stores the result in `s`.

- We must be able to examine the selected units. For example, if the units are water heaters packed in cartons stored in large stacks, we can select the sample of identifiers using SRS but we are unlikely to find someone willing to sort through the cartons to find the selected units.

- SRS is the simplest probability sampling protocol. Because of the difficulty of completing a frame, we may use cluster or multi-stage sampling instead. With cluster sampling the frame can be the list of clusters. With multi-stage sampling we can build the frame as we go.

- For many populations, it is more efficient (e.g. shorter confidence interval with a smaller sample size) to stratify the population and use stratified random sampling. See Chapter 7.

5.3.1 An In-Class Exercise in SRS

During a STAT class, suppose we took an approximate SRS of the students present to estimate the average number, μ, of Canadian provinces visited by the students in their lifetimes. We agreed that a visit to a given province must consist of at least one day spent in that province.

We then collected the following 29 observations of our response variate of interest.

# of provinces visited	1	2	3	4	5	6	7	8	9	10	Total
# of students	6	11	4	5	1	0	0	1	0	1	29

The sample summary statistics are

$$\begin{aligned} \hat{\mu} &= 2.86 \\ \hat{\sigma} &= 2.05 \end{aligned}$$

We construct a 95% confidence interval for μ as follows. There are $N = 182$ students currently enrolled in Stat 332. The standard error is

$$\sqrt{1 - \frac{n}{N}} \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{1 - \frac{29}{182}} \frac{(2.05)}{\sqrt{29}} = 0.349.$$

For 95% confidence with a $G(0, 1)$ distribution, we need $c = 1.96$. Therefore our approximate 95% confidence interval for the average number of provinces visited is

$$\hat{\mu} \pm c(\text{standard error}) = 2.86 \pm (1.96)(0.349) = 2.86 \pm 0.684 = [2.176, 3.544].$$

We have high confidence that the true average number of provinces visited lies in this interval.

Follow-up Questions:

1. What errors could be involved in our setup?
   (a) Sample error - not all students come to the lecture at 8:30 AM any longer
   (b) Measurement error - students lied, or the instructor keyed the data in incorrectly
2. What could be done to minimize the effects of these errors?
   (a) To address sample error, make attendance at lectures mandatory, with some marks assigned to it, or conduct the survey on-line rather than in-class.

(b) To address units lying, give the answers secretly instead of in front of the class.

(c) To address instructor keying in the data incorrectly, show data as it is entered and have the units correct if needed.

5.4 Exercises

1. Consider the sampling protocols defined in Example 1.

   (a) Show that the inclusion probability for each unit in the frame is $\frac{1}{100}$ for every protocol.

   (b) On a final examination, a student once defined simple random sampling as follows: "simple random sampling is a method of selecting units from a population so that every unit has the same chance of selection". Is this a correct answer?

   (c) Show that the estimator corresponding to the sample average $\hat{\mu} = \frac{\sum_{i \in s} y_i}{n}$ is unbiased for μ for each of the protocols.

2. Consider the estimate $\hat{\sigma} = \sqrt{\frac{\sum_{i \in s}(y_i - \overline{y})^2}{n-1}}$ and the corresponding estimator $\tilde{\sigma}$.

   (a) For SRS, show that $\tilde{\sigma}^2$ is an unbiased estimator for σ^2.

   Hint: Use the fact that $\sum_{i \in s}(y_i - \overline{y})^2 = \sum_{i \in s} y_i^2 - n\overline{y}^2$.

   (b) Is $\tilde{\sigma}$ unbiased for σ?

3. To estimate the total number of male song sparrows in a 10 km by 10 km square (`http://www.birdsontario.org/atlas/atlasmain.html`) for a breeding bird atlas, a simple random sample of 50 one hectare plots (a hectare is 100m by 100m) is selected. Using a GPS system, your intrepid instructor visits each of the selected plots (after dawn but before 9:00 am between May 24 and July 6) and counts the number of singing male song sparrows detected in a 10 minute period. The data are summarized below.

# of sparrows	0	1	2	3	4
# of plots	28	13	5	3	1

   (a) Find a 95% confidence interval for τ, the total number of male song sparrows in the square.

   (b) Suppose that I wanted to estimate the total number of male song sparrows to within 1000 with 95% confidence. How many additional plots are needed?

4. Suppose we want to estimate a population average so that the relative precision is specified. That is, we want to find the sample size required (SRS) so that the length of the confidence interval 2ℓ divided by the sample average is pre-determined.

   (a) For a given confidence level and required precision $p\%$, find a formula for the required sample size.

   (b) What knowledge of the population attributes do we need to make this formula usable?

5. One cheap (but poor) way to check the quality of a batch of items is called **acceptance sampling**. Suppose that there are $N = 1000$ items in a shipment and you cannot tolerate more than 1% defective (your first mistake – why should you tolerate any defective items from your supplier). You decide to select and inspect a sample of 20 items and accept the shipment if you find zero defectives. If you find 1 or more defective items, you inspect the complete shipment.

   (a) How would you select the sample?

   (b) Calculate the probability $p(\pi)$ that you accept the shipment as a function of π, the percentage of defective items in the shipment.

   (c) Graph $p(\pi)$ for $0 \leq \pi \leq 10\%$

   (d) Given the results in part 5c, you decide to increase the sample size so that there is only a 5% chance of accepting a shipment with 1% defective. What sample size do you recommend?

Chapter 6

Ratio and Regression Estimation with SRS

In this chapter, we consider two related problems:

- estimating a ratio such as the proportion or average response of a subpopulation (domain) with unknown size and
- improving the sample average as an estimate of the frame average by using explanatory variates.

6.1 Estimating a Ratio

In Chapter 5, we looked at assessing the estimates of the frame average (or total) when the sampling protocol is SRS and the estimate is the sample average. Here we consider estimating a ratio. The distinguishing feature is that both the numerator and denominator will change if we were to repeat the sampling protocol over and over.

Consider again the inventory example from Chapter 5. Suppose we want to estimate the average size of the dollar error in those files that are in error. Using the following notation.

- Let y_i be the dollar error in the i^{th} account. (Note that in Chapter 5, y_i instead denoted the total dollar amount for the i^{th} account.)
- Let μ be the average dollar error per file. (Note that in Chapter 5, μ instead denoted the average dollar amount per file.)
- As before, π is the proportion of files in error.
- As before, z_i is an indicator variable which equals 1 if the i^{th} file is in error and zero otherwise.

Then we can write the desired attribute as

$$\theta = \frac{\sum_{i\in U} y_i z_i}{\sum_{i\in U} z_i} = \frac{\frac{1}{N}\sum_{i\in U} y_i z_i}{\frac{1}{N}\sum_{i\in U} z_i} = \frac{\mu}{\pi}$$

Note that

$$\sum_{i\in U} y_i z_i = \sum_{i\in U} y_i.$$

We use the estimate $\hat{\theta} = \hat{\mu}/\hat{\pi}$ with corresponding estimator $\tilde{\theta} = \tilde{\mu}/\tilde{\pi}$. To assess the estimate and produce confidence intervals for θ, we find the (approximate) distribution of $\tilde{\theta}$ by finding its mean and variance and then using a Gaussian approximation.

To derive the approximation, we use Taylor's theorem for a function of two variables. Recall that we can expand $f(x, y)$ about the point (x_0, y_0) to get a linear approximation

$$f(x,y) \approx f(x_0,y_0) + \frac{\partial f(x_0,y_0)}{\partial x}(x - x_0) + \frac{\partial f(x_0,y_0)}{\partial y}(y - y_0)$$

The linear function on the right has the same value and first partial derivatives as $f(x, y)$ at the point (x_0, y_0).

If $f(x, y) = \frac{x}{y}$, then we have

$$\frac{\partial f}{\partial x} = \frac{1}{y} \quad \text{and so} \quad \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} = \frac{1}{y_0},$$
$$\frac{\partial f}{\partial y} = -\frac{x}{y^2} \quad \text{and so} \quad \left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} = -\frac{x_0}{y_0^2}$$

and thus

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

Replacing (x, y) by the random variables $(\tilde{\mu}, \tilde{\pi})$ and (x_0, y_0) by (μ, π), we have

$$\frac{\tilde{\mu}}{\tilde{\pi}} \approx \frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu} - \mu) - \frac{\mu}{\pi^2}(\tilde{\pi} - \pi).$$

For large sample sizes, the approximation is reasonable since we expect $(\tilde{\mu}, \tilde{\pi})$ to

be close to (μ, π). Hence we have

$$\begin{aligned} E\left[\tilde{\theta}\right] &= E\left[\frac{\tilde{\mu}}{\tilde{\pi}}\right] \\ &\approx E\left[\frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu} - \mu) - \frac{\mu}{\pi^2}(\tilde{\pi} - \pi)\right] \\ &= \frac{\mu}{\pi} + \frac{1}{\pi}\underbrace{E[\tilde{\mu} - \mu]}_{=0} - \frac{\mu}{\pi^2}\underbrace{E[\tilde{\pi} - \pi]}_{=0} \\ &= \frac{\mu}{\pi}. \end{aligned}$$

Now note that rearranging the approximation for $\tilde{\theta}$ gives

$$\begin{aligned} \tilde{\theta} &\approx \frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu} - \mu) - \frac{\mu}{\pi^2}(\tilde{\pi} - \pi) \\ &= \frac{\mu}{\pi} - \frac{\mu}{\pi} + \frac{\mu}{\pi} + \left(\frac{1}{\pi}\right)\left(\tilde{\mu} - \left(\frac{\mu}{\pi}\right)\tilde{\pi}\right) \\ &= \frac{\mu}{\pi} + \left(\frac{1}{\pi}\right)(\tilde{\mu} - \theta\tilde{\pi}), \end{aligned}$$

and so we obtain

$$Var(\tilde{\theta}) \approx \frac{1}{\pi^2} Var[\tilde{\mu} - \theta\tilde{\pi}].$$

The estimator $\tilde{\theta}$ is approximately unbiased (but see Exercise 1).

We can write the variance in several forms. Notice that the estimate corresponding to $\tilde{\mu} - \theta\tilde{\pi}$ can be written as

$$\begin{aligned} \hat{\mu} - \theta\hat{\pi} &= \frac{\sum_{i\in s} y_i}{n} - \theta\left(\frac{\sum_{i\in s} z_i}{n}\right) \\ &= \frac{\sum_{i\in s}(y_i - \theta z_i)}{n}, \end{aligned}$$

which is the sample average of $r_1, \ldots, r_n$ where $r_i = y_i - \theta z_i$.

A Brief Explanation of the r_is: We are in the process of estimating, θ, the average dollar error for items in error. If the dollar error is the variate of interest then $r_i = y_i - \theta z_i$ is the i^{th} **residual**. The indicator variable z_i behaves like a covariate.
Using the basic formula for the variance of an average with SRS,

$$Var[\tilde{\mu} - \theta\tilde{\pi}] = (1 - f)\frac{\sigma_r^2}{n}$$

where σ_r^2 is the variance of the r_is and $f = \frac{n}{N}$ is the **sampling fraction** as usual.

We can estimate this variance by the corresponding sample variance

$$\begin{aligned}
\left(\frac{1-f}{n}\right)\frac{\sum_{i\in s}(r_i-\overline{r})^2}{n-1} &= \left(\frac{1-f}{n}\right)\frac{\sum_{i\in s}[y_i-\theta z_i-(\overline{y}-\theta\overline{z})]^2}{n-1} \\
&= \left(\frac{1-f}{n}\right)\frac{\sum_{i\in s}[y_i-\overline{y}-\theta(z_i-\overline{z})]^2}{n-1} \\
&\approx \left(\frac{1-f}{n}\right)\frac{\sum_{i\in s}[y_i-\overline{y}-\hat{\theta}(z_i-\overline{z})]^2}{n-1} \qquad (6.1) \\
&= \left(\frac{1-f}{n}\right)\frac{\sum_{i\in s}(y_i-\hat{\theta}z_i)^2}{n-1}
\end{aligned}$$

where we replace θ by its estimate

$$\hat{\theta} = \frac{\hat{\mu}}{\hat{\pi}} = \frac{\frac{1}{n}\sum_{i=1}^n y_i}{\frac{1}{n}\sum_{i=1}^n z_i} = \frac{\overline{y}}{\overline{z}}$$

on line (6.1) and note that this implies $\hat{\theta}\overline{z} = \overline{y}$. The estimate of the variance of the estimator $\tilde{\theta}$ is then

$$\widehat{Var}(\tilde{\theta}) = \frac{1}{\hat{\pi}^2}\frac{(1-f)}{n}\frac{\sum_{i\in s}(y_i-\hat{\theta}z_i)^2}{n-1}$$

where the last term is the sample variance of the estimated residuals $y_1-\hat{\theta}z_1, \ldots, y_n-\hat{\theta}z_n$.

To construct a confidence interval for θ, for large values of n and N, the estimator is approximately Gaussian so the confidence interval has the standard form

$$\text{estimate} \pm c\ \hat{\text{std}}\text{ev}(\tilde{\theta}).$$

where c is chosen for a $G(0,1)$ distribution based on the desired confidence level. In the example (see the data file *inventory.txt*), we have

$$\hat{\theta} = \frac{\hat{\mu}}{\hat{\pi}} = \frac{2.33}{0.22} = 10.57.$$

To find the estimate of the standard deviation, first calculate the sample standard deviation of $\hat{r}_1 = y_1 - \hat{\theta}z_1, \ldots, \hat{r}_n = y_n - \hat{\theta}z_n$ in R by creating the vector

```
r <- y - theta.hat * z
```

From the sample data, we get the sample standard deviation 41.69. Then multiply by the factor $\left(\frac{1}{\hat{\pi}}\right)\sqrt{\frac{1-f}{n}}$. In the example, we have the standard error 26.26. A 95% confidence interval for θ is 10.57 ± 51.48.

Remarks:

- We can estimate the average error in accounts with errors very imprecisely.

- Also note that this confidence interval is wider than that for μ, the average error, because we also have uncertainty about the proportion of files in error.

- We can use the same approach via Taylor's theorem to estimate any other function of variate averages in which we have interest.

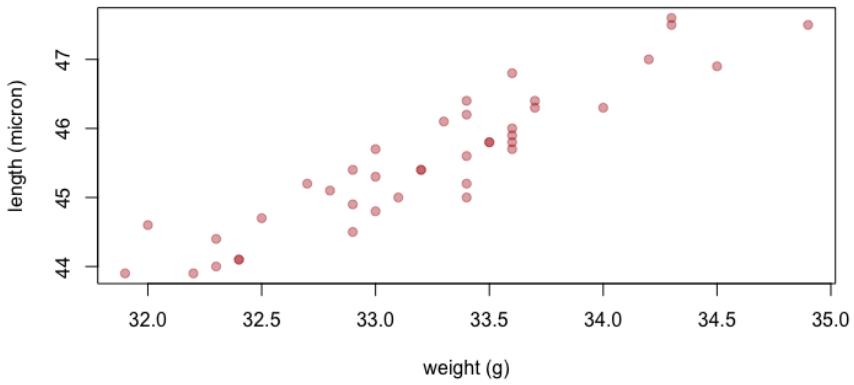
6.2 Ratio Estimation of the Average

Suppose the purpose of the survey is to estimate the study population average $\mu(y)$ for some variate y. Note the change in notation to explicitly include y in the definition of the attribute. In many surveys, there are other (explanatory) variates that can be measured on each unit in the sample and for which we have complete knowledge of their attributes in the population.

For example, in the inventory survey, we know the stated value and the stated number of items for each file in the population and hence we can calculate population attributes for these variates. In many surveys of human populations, the demographics (gender ratio, age distribution, etc.) of the population are known, perhaps from a census. When we get the sample, we can determine the values of the response variate y and the explanatory variates, say gender and age for each person in the sample. The idea of the methods discussed here is to adjust the sample average $\hat{\mu}(y)$ based on differences between the sample and (known) population attributes of the explanatory variates. For simplicity, we consider only one explanatory variate.

Example

In the assessment of a lot of $N = 10000$ incoming molded parts, a company selects a sample of $n = 40$ parts to check the average length of a critical dimension. From previous experience, they know that the dimension is related to part weight so they measure the weight of each part in the sample and also the weight of the entire shipment. The sample data are included in the file *molded.txt* and are plotted below. The plot shows a strong correlation between the length and weight.



The average and standard deviations for the two variates are:

	x (weight)	y (length)
sample average	33.24	45.56
sample st. dev.	0.691	1.005

The population average weight is $\mu(x) = 33.10$ grams determined as the total weight (measured all at once) divided by the number of pieces $N = 10000$.

The ratio estimate of $\mu(y)$ is

$$\hat{\mu}(y)_{ratio} = \frac{\hat{\mu}(y)}{\hat{\mu}(x)}\mu(x) = \hat{\theta}\mu(x)$$

where $\hat{\mu}(x), \hat{\mu}(y)$ are the sample averages for x and y respectively and $\theta = \frac{\mu(y)}{\mu(x)}$.

The sample is collected haphazardly since it is too expensive to create a frame. We develop the estimators and their properties assuming SRS – this corresponds to assuming that the haphazard sampling protocol mirrors SRS if the protocol is repeated over and over.

We use the results on the estimation of a ratio θ from Section 6.1 to derive an approximation for the mean and standard deviation of $\tilde{\mu}(y)_{ratio}$. Applying the results from Section 6.1 with

$$\tilde{\mu} = \tilde{\mu}(y), \quad \tilde{\pi} = \tilde{\mu}(x), \quad y_i = y_i, \quad \text{and} \quad z_i = x_i$$

gives

$$E\left[\tilde{\mu}(y)_{ratio}\right] = E\left[\tilde{\theta}\mu(x)\right] = E\left[\tilde{\theta}\right]\mu(x) \approx \theta\mu(x) = \mu(y)$$

which makes the ratio estimator approximately unbiased. Then the variance of the ratio estimator is approximately

$$\begin{aligned} Var(\tilde{\mu}(y)_{ratio}) &= Var\left[\frac{\tilde{\mu}(y)}{\tilde{\mu}(x)}\mu(x)\right] \\ &= Var\left[\tilde{\theta}\mu(x)\right] \\ &= \mu(x)^2 Var[\tilde{\theta}] \\ &= \mu(x)^2\left(\frac{1}{\mu(x)^2}\right) Var[\tilde{\mu}(y) - \theta\tilde{\mu}(x)] \\ &= Var[\tilde{\mu}(y) - \theta\tilde{\mu}(x)]. \end{aligned}$$

Using the results on the estimation of a ratio, we estimate the variance of $\tilde{\mu}(y)_{ratio}$ by

$$\widehat{Var}(\tilde{\mu}(y)_{ratio}) \approx \left(\frac{1-f}{n}\right)\frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i)^2 = \left(\frac{1-f}{n}\right)\frac{\sum_{i\in s}\hat{r}_i^2}{n-1}$$

where $\hat{r}_i = y_i - \hat{\theta}x_i$ as before.

In the example we have $\hat{\theta} = 1.371$ and $\frac{1}{n-1}\sum_{i\in s}(y_i - \hat{\theta}x_i)^2 = 0.147$ so the ratio estimate of the population average is $\hat{\mu}(y)_{ratio} = \frac{45.56}{33.24} \times 33.10 = 45.37$ with corresponding standard error (ignoring the fpc) of

$$\sqrt{\frac{0.147}{40}} = 0.061$$

and thus confidence interval half-width of $(1.96)(0.061) = 0.12$. An approximate 95% confidence interval for the population average length based on the ratio estimate is 45.37 ± 0.12 microns.

Here the ratio estimate is more precise than the sample average $\hat{\mu}(y) = 45.56$ since this estimate (with $\hat{\sigma}(y) = 1.005357$) gives a confidence interval (ignoring the fpc)

$$\hat{\mu}(y) \pm 1.96\frac{\hat{\sigma}(y)}{\sqrt{40}} \text{ or } 45.56 \pm 1.96\left(\frac{1.005357}{\sqrt{40}}\right) \text{ or } 45.56 \pm 0.31.$$

Comparing sample average and ratio estimator

We can compare the estimated variance of the ratio estimator, $\tilde{\mu}(y)_{ratio}$, and the estimator based on the sample average, $\tilde{\mu}(y)$, the estimator based on the sample average. Consider

$$\widehat{Var}[\tilde{\mu}(y)_{ratio}] = (1-f)\frac{1}{n}\left[\frac{\sum_{i\in s}(y_i - \hat{\theta}x_i)^2}{n-1}\right]$$

versus

$$\widehat{Var}[\tilde{\mu}(y)] = (1-f)\,\frac{1}{n}\left[\frac{\sum_{i\in s}(y_i-\overline{y})^2}{n-1}\right]$$

The ratio estimate is more precise (i.e. gives a shorter confidence interval) than the sample average if

$$\sum_{i\in s}(y_i-\hat{\theta}x_i)^2 < \sum_{i\in s}(y_i-\overline{y})^2$$

The expression on the left is the residual sum of squares if we "fit" a line through the origin to the sample scatterplot. The expression on the right is the total sum of squares. Qualitatively, the ratio estimate is more precise if a line through the origin explains some of the variation in the response variate. This gain in precision is the major advantage of the ratio estimate.

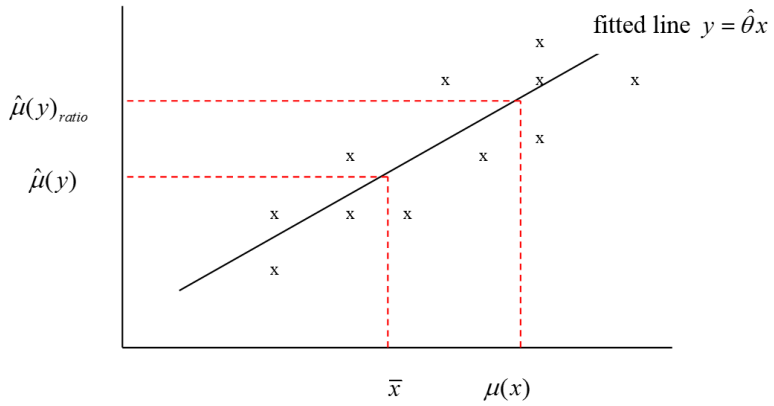
Notes on Ratio Estimation of the Average

1. To apply ratio estimate effectively, we need

   - to measure the explanatory variate x_i for each unit i in the sample
   - to know $\mu(x)$, the population average of the explanatory variate
   - a relationship of the form $y = \beta x + noise$, a straight line through the origin, between x and y in the study population. The smaller the noise, the greater the benefit in using the ratio estimate.

     In the example, we have $y =$ length and $x =$ weight, so it is reasonable to assume that a line through the origin will model our data well.

2. If we think of ratio estimation in terms of fitting a line to the scatterplot, then the estimate is an adjustment based on the fact that the sample average $\overline{x}$ is different than the population average $\mu(x)$.



   In this case, the sample average $\overline{x}$ is smaller than the population average $\mu(x)$ so we adjust the estimate of $\mu(y)$ upward using the relationship between

y and x. The closer $\overline{x}$ is to $\mu(x)$, the smaller is the adjustment. We can also see the adjustment by rewriting the ratio estimate as

$$\hat{\mu}(y)_{ratio} = \frac{\mu(x)}{\hat{\mu}(x)}\hat{\mu}(y)$$

3. If we fit a response model to the above data (e.g. $Y_i = \beta x_i + R_i,\ R_i \sim G(0,\sigma)$) then we estimate the slope using

$$\hat{\beta} \ = \ \frac{\sum_{i\in s} x_i y_i}{\sum_{i\in s} x_i^2}. \tag{6.2}$$

This suggests another estimate $\hat{\beta}\mu(x)$ for $\mu(y)$. Since $\hat{\beta} = \frac{\frac{\sum_{i\in s} x_i y_i}{n}}{\frac{\sum_{i\in s} x_i^2}{n}}$ can be written as the ratio of two averages, we can derive the variance of $\tilde{\beta}$ as we did for $\tilde{\theta}$ and hence find the variance of $\tilde{\mu}(y) = \tilde{\beta}\mu(x)$.

Remark: Recall from Stat 231 the standard formula

$$\hat{\beta} \ = \ \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \overline{x})y_i}{\sum_i (x_i - \overline{x})x_i}. \tag{6.3}$$

Also recall the remark about "redefining" the observed x_i as $x_i' = x_i - c$, where c is a constant value that makes $\sum_{i=1}^n x_i'$ close to zero. (Often we take $c = \overline{x}$, which makes $\sum_{i=1}^n x_i'$ exactly zero.) The intercept α changes if we redefine x, but not β. This remark explains why we lose no generality by using the formula on line (6.2) above instead of the more general formula on line (6.3).

You may wonder how the precision of this estimate compares to that of the ratio estimate.

With ratio estimation as described above, we estimate the slope using $\hat{\theta} = \frac{\overline{y}}{\overline{x}}$. Suppose that variation increases with x_i, so that we start with a response model

$$Y_i = \beta x_i + R_i,\ R_i \sim G(0, \sigma\sqrt{x_i}),$$

then we can divide by $\sqrt{x_i}$ to get the model

$$\frac{Y_i}{\sqrt{x_i}} = \beta\sqrt{x_i} + R_i^*,\ R_i^* \sim G(0,\sigma)$$

with constant standard deviation. You can easily verify that the least squares estimate of β in this model is $\hat{\beta} = \frac{\overline{y}}{\overline{x}}$.

If there is constant variation about the line, we expect the estimator based on $\hat{\beta}$ to be superior. If the variation increases as x increases, then we expect the ratio estimate to be better. In either case, because we are exploiting structure in the study population, the estimates will be superior to the sample average.

4. Suppose the response variate y is binary and the goal is to estimate the population proportion π. If we have a continuous explanatory variate x, we need more complex models (and subsequent analysis) to exploit the relationship between the variates in the study population. If the explanatory variate is binary or categorical we can use post-stratification (see the Chapter 7) to improve the precision of the estimation of π.

6.3 Regression Estimation of the Average

Once we discussed ratio estimation in terms of fitting a line through the origin to the data in the sample, you will have considered what happens if the line does not go through the origin. Here we look at using information on the explanatory variate if the relationship between y and x is linear with constant variation about the line. In other words, we have the conditions necessary for fitting the response model

$$Y_i = \alpha + \beta(x_i - \overline{x}) + R_i, R_i \sim G(0, \sigma)$$

to the data in the sample. Note that $\overline{x} = \hat{\mu}(x)$ is the sample average for the explanatory variate.

To produce the regression estimate $\hat{\mu}(y)_{reg}$, we

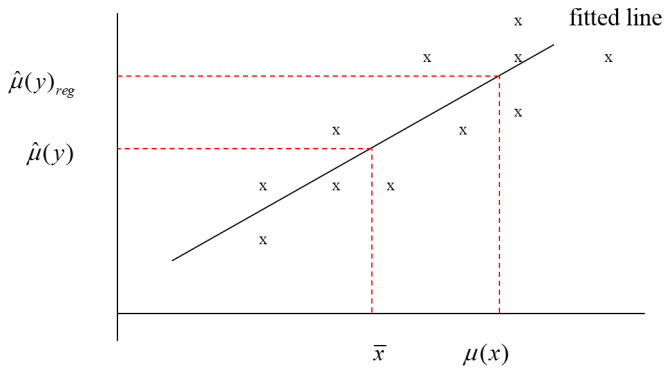
- fit a model using least squares to estimate α and β to get

$$\begin{aligned} \hat{\alpha} &= \overline{y} = \hat{\mu}(y) \\ \hat{\beta} &= \frac{\sum_{i \in s}(x_i - \overline{x})y_i}{\sum_{i \in s}(x_i - \overline{x})^2}. \end{aligned}$$

- substitute the known mean $\mu(x)$ into the fitted line

$$\hat{\mu}(y)_{reg} = \hat{\mu}(y) + \hat{\beta}[\mu(x) - \hat{\mu}(x)]$$

We can view $\hat{\mu}(y)_{reg}$ as an adjustment to the sample average $\hat{\mu}(y)$ as we did with the ratio estimate. Suppose that β is positive so that in the study population larger values of x correspond to larger values of y. If the sample average $\hat{\mu}(x)$ of the explanatory variate is less than the known population average $\mu(x)$, we adjust the estimate of $\mu(y)$ upward. The adjustment is shown on the following plot.



The properties of the estimator $\tilde{\mu}(y)_{reg} = \tilde{\mu}(y) + \tilde{\beta}[\mu(x) - \tilde{\mu}(x)]$ are complicated because of the three random components. We can simplify the argument with the following hand-wave. Rewrite the estimator as

$$\tilde{\mu}(y)_{reg} - \mu(y) = [\tilde{\mu}(y) - \mu(y)] + \beta[\mu(x) - \tilde{\mu}(x)] + [\tilde{\beta} - \beta][\mu(x) - \tilde{\mu}(x)]$$

In large samples, we expect each of the terms within the brackets [ ] to be small. The right-most term, a product of two small quantities, is an order smaller than the other two terms. Hence we can say that

$$\tilde{\mu}(y)_{reg} - \mu(y) \approx [\tilde{\mu}(y) - \mu(y)] + \beta[\mu(x) - \tilde{\mu}(x)]$$

and we have

$$\begin{aligned} E[\tilde{\mu}(y)_{reg} - \mu(y)] &\approx E[\tilde{\mu}(y) - \mu(y)] + \beta E[\mu(x) - \tilde{\mu}(x)] \\ &= 0. \end{aligned}$$

That is, the regression estimate is approximately unbiased.
We can estimate $Var(\tilde{\mu}(y)_{reg})$ by noting that

$$\hat{\mu}(y)_{reg} = \frac{\sum_{i \in s}[y_i - \beta(x_i - \mu(x))]}{n}$$

is the sample average of $r_1, \ldots, r_n$ where

$$\begin{aligned} r_i &= y_i - \beta(x_i - \mu(x)) \text{ and} \\ \overline{r} &= \overline{y} - \beta(\overline{x} - \mu(x)). \end{aligned}$$

Using the basic result for the variance of an average with SRS, we have

$$\begin{aligned} Var(\tilde{\mu}(y)_{reg}) &= (1-f)\frac{\sigma_r^2}{n} \\ &= (1-f)\frac{1}{n}\left[\frac{\sum_{i \in U}[r_i - \overline{r}]^2}{N-1}\right] \\ &= (1-f)\frac{1}{n}\left[\frac{\sum_{i \in U}[y_i - \overline{y} - \beta(x_i - \overline{x})]^2}{N-1}\right] \end{aligned}$$

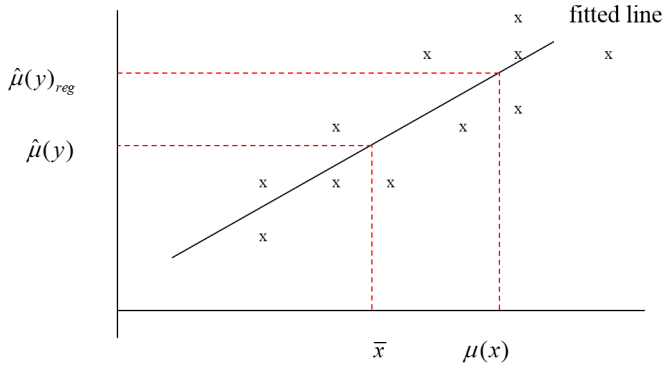
which can be estimated by

$$\widehat{Var}(\tilde{\mu}(y)_{reg}) = \left(\frac{1-f}{n}\right) \frac{\sum_{i\in s}\left[y_i - \overline{y} - \hat{\beta}(x_i - \overline{x})\right]^2}{n-1}$$

where we have replaced β by the estimate $\hat{\beta}$. The last factor is the sample variance of the estimated residuals from the least squares fit of the line to the sample data.

Example

The volume of usable wood y in a Douglas fir is related to the basal area, x, the cross-sectional area of the tree measured at breast height. Volume is expensive to measure because it requires that the tree be destroyed. To estimate the total volume in a section of forest that was to be sold, a sample of 25 trees was selected by dividing the section into small subsections. A SRS of 25 sub-sections was selected and then a tree was selected at random within the sub-section. We will treat this protocol as if it were SRS. The selected trees were sacrificed and the basal area and volume were measured. The data and fitted line are plotted below.



The equation of the fitted line is $y = 6.17 + 1.51(x - 1.31)$ and the residual sum of squares is 2.268.

A second much larger (and cheaper) survey was carried out to estimate the total number of trees $N = 56800$ and the average basal area $\mu(x) = 1.40$. We assume that the errors in estimating $\mu(x)$ and N are negligible.

The regression estimate is $\hat{\mu}(y)_{reg} = 6.17 + 1.51(1.40 - 1.31) = 6.31$ and the standard error of the estimate is

$$\widehat{stdev}(\tilde{\mu}(y)_{reg}) \approx \sqrt{\frac{1}{25}\frac{2.268}{24}} = 0.061.$$

An approximate 95% confidence interval for $\mu(y)$ based on the regression estimator is

$$6.31 \pm (1.96)(0.061) = 6.31 \pm 0.12.$$

An 95% confidence interval for the total volume, $\tau(y) = 56800\mu(y)$, is then $358,408 \pm 6,816$.

The estimate for $\mu(y)$ based on the sample average $\hat{\mu}(y)$ gives a 95% confidence interval 6.17 ± 0.22 since the sample standard deviation of the 25 measured volumes is 0.111. The regression estimate is more precise in this case.

Comparing Estimators

We can compare the precision of the sample average, the ratio estimate and the regression estimate by looking at the sum of squares of the estimated residuals under the three fits

1. Sample average: $\sum_{i \in s}(y_i - \overline{y})^2$
2. Ratio estimate: $\sum_{i \in s}(y_i - \hat{\theta}x_i)^2$ where $\hat{\theta} = \frac{\overline{y}}{\overline{x}}$
3. Regression estimate: $\sum_{i \in s}(y_i - \overline{y} - \hat{\beta}(x_i - \overline{x}))^2$ where $\hat{\beta}$ is the estimated slope

The regression estimate has the smallest residual sum of squares in all cases. We do not always use the regression estimate because of bias and other technical considerations. The major reason for using ratio and regression estimates is the gain in precision.

Notes on Regression Estimation of the Average

1. To use the regression estimate effectively, we need
   - a continuous response variate y and a continuous explanatory variate x
   - knowledge of the study population average of the explanatory variate
   - a linear relation between y and x with smaller residual variation leading to a more precise estimate.
2. A special simple case of regression estimation is to use the difference $d_i = y_i - x_i$ as the response variate and then estimate the population average by

$$\hat{\mu}(y)_{diff} = \hat{\mu}(d) + \mu(x)$$

This estimate is more precise than the sample average if the variation in the differences $d_1, \ldots, d_n$ is less than the variation in $y_1, \ldots, y_n$. We used a difference estimate in the inventory example to estimate the total true value of the files.

3. Regression estimation can be extended to multiple explanatory variates and non-linear relationships. We use least squares to estimate the relationship between the response and explanatory variates in the sample and then adjust the sample average using the fitted model and the sample averages for the explanatory variates. Note this adjustment accounts for differences in the sample attributes of the explanatory variates and the known population attributes.

6.4 Exercises

1. Consider the sampling protocols defined in Example 1. Find the quadratic expansion of $f(x, y) = \frac{y}{x}$ about the point $(\mu(x), \mu(y))$ to estimate the bias in the estimator $\tilde{\theta} = \frac{\tilde{\mu}(y)}{\tilde{\mu}(x)}$. Note that the general form of the expansion is

$$\begin{aligned} f(x,y) \approx\ & f(x_0,y_0) + \frac{\partial f(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial f(x_0,y_0)}{\partial y}(y-y_0) \\ & + \frac{\partial^2 f(x_0,y_0)}{\partial x^2}\frac{(x-x_0)^2}{2} \\ & + \frac{\partial^2 f(x_0,y_0)}{\partial x \partial y}(x-x_0)(y-y_0) \\ & + \frac{\partial^2 f(x_0,y_0)}{\partial y^2}\frac{(y-y_0)^2}{2} \end{aligned}$$

This quadratic function has the same value, first and second derivatives at the point (x_0, y_0) as does $f(x, y)$. You can easily check this statement by differentiating the right side of the expression.

2. In order to count the number of small items in a large container, a shipping company selects a sample of 25 items and weighs them. They then weigh the whole shipment (excluding the container). Assume that there is small error in weighing and act as if SRS is used (it is not, the sampling is haphazard). Let y_i be the weight of the i^{th} item in the population and let the total known weight be τ.

   (a) Show that an estimate of the population size is

$$\hat{N} = \frac{\tau}{\sum_{i \in s} \frac{y_i}{25}}.$$

(b) Find the (approximate) mean and standard deviation of the corresponding estimator $\tilde{N}$. You can ignore the unknown sampling fraction $f = \frac{n}{N}$.

(c) In the example, the sample average mass is 75.45 g, the sample standard deviation is 0.163 g and the total mass is 154.2 kg. Find a 95% confidence interval for the total number of items in the container.

3. Briefly describe when you would use the ratio or regression estimate instead of the sample average to estimate the population average.

4. Many bird species have specialized habitat. We can exploit this knowledge when we are trying to estimate population totals or density. For example, wood thrush are a forest dwelling bird that live in the hardwood forests of eastern North America. Suppose we wanted to estimate the number of wood thrush pairs nesting within the region of Waterloo, an area of highly fragmented forest patches. Using aerial photography, we know that there are 1783 such patches (minimum size 3 ha) with an average size 13.4 ha. A simple random sample of 50 woodlots is selected and the number of nesting pairs y_i is counted in each woodlot by counting the number of singing males. The area x_i of each sampled woodlot is also recorded. The data are available in the file *thrush.txt.* Find a 95% confidence intervals for the total number of thrushes based on the

   (a) sample average $\overline{y}$

   (b) ratio estimate

   (c) regression estimate

5. The City of Waterloo wants to estimate the average amount of water per house $\mu(y)$ that is used to water lawns and gardens in the month of July. A SRS of 50 houses is selected and special metering units are installed to measure the volume of water y from external taps. The total volume of water x is measured by the regular meter. From water records, it is known that the average total water consumption per house is $\mu(x) = 15.6$ cubic metres. The data are stored in the file *water.txt.*

   (a) Prepare a scatterplot of y versus x.

   (b) Estimate $\mu(y)$ using the sample average, the ratio estimate and the regression estimate.

   (c) Find 95% confidence intervals based on each estimate.

   (d) Which estimation procedure is preferable here? Why?

Chapter 7

Stratified Random Sampling

In the previous chapter, we looked at ways to use an explanatory variate with known attributes to improve on the sample average as an estimate of the study population average with SRS. The basic idea was to exploit a structural relationship between the response and explanatory variates with known attributes in the population.

In this chapter, we change both the sampling protocol and the estimate to get a procedure that usually produces a better estimate of the study population average. The idea is to divide the study population in sub-populations, called **strata**, and sample independently using SRS from each **stratum**. Then combine the estimates of each stratum average to get an estimate of the population average.

Some examples of possible strata are:

- Provinces and large urban centers in national opinion surveys
- Small and large accounts in auditing a population of accounts
- Home faculties in a survey of UW students
- Sites in a survey of employees in a multi-site company

In many examples, we have questions about the strata averages as well as the overall population average. Stratified sampling gives information about these averages and often an improved estimate of the overall average.

7.1 Stratified Random Sampling

Suppose that we divide the population U into H mutually exclusive strata $U_1, \ldots, U_H$ with sizes $N_1, \ldots, N_H$ so that $N = N_1 + \cdots + N_H$. For the variate of interest, we denote the stratum averages and standard deviations by $\mu_h, \sigma_h, h = 1, \ldots, H$.

With this notation, we write

$$\mu = \frac{N_1\mu_1 + \cdots + N_H\mu_H}{N} = W_1\mu_1 + \cdots + W_H\mu_H,$$

the **weighted average** of the stratum averages. We call $W_h = \frac{N_h}{N}$ the **stratum weight**, the proportion of the total units found in that stratum.

Now suppose for each stratum, we independently select a sample of size n_h from stratum h using SRS and calculate the sample average $\hat{\mu}_h$. We can combine these estimates to get the **stratified estimate of the population average**

$$\hat{\mu}_{strat} = W_1\hat{\mu}_1 + \cdots + W_H\hat{\mu}_H.$$

The corresponding estimator is $\tilde{\mu}_{strat} = W_1\tilde{\mu}_1 + \cdots + W_H\tilde{\mu}_H$. Since we used SRS within each stratum, we have

$$E(\tilde{\mu}_{strat}) = W_1E(\tilde{\mu}_1) + \cdots + W_HE(\tilde{\mu}_H) = W_1\mu_1 + \cdots + W_H\mu_H = \mu$$

and

$$\begin{aligned} Var(\tilde{\mu}_{strat}) &= W_1^2(1-f_1)\frac{\sigma_1^2}{n_1} + \cdots + W_H^2(1-f_H)\frac{\sigma_H^2}{n_H} \\ &= W_1^2\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\sigma_1^2 + \cdots + W_H^2\left(\frac{1}{n_H} - \frac{1}{N_H}\right)\sigma_H^2 \end{aligned}$$

where $f_h = \frac{n_h}{N_h}$ is the sampling fraction and σ_h is the standard deviation of the response variate for stratum h. We can estimate the variance by

$$\widehat{Var}(\tilde{\mu}_{strat}) = W_1^2(1-f_1)\frac{\hat{\sigma}_1^2}{n_1} + \cdots + W_H^2(1-f_H)\frac{\hat{\sigma}_H^2}{n_H}$$

where $\hat{\sigma}_h^2 = \frac{\sum_{j\in s_h}(y_{hj}-\hat{\mu}_h)^2}{n_h-1}$ is the sample standard deviation within stratum h and y_{hj} is the observed response variate for the j^{th} unit in the sample from stratum h.

Example 1
Let μ be the average concentration of N_a and let π be the proportion of contaminated wells (see below). Find 95% CIs for μ and π.
To estimate average water quality and the proportion of wells with contamination, a survey of residential wells was carried out in the rural part of the region of Waterloo. The population of 13345 wells was identified from assessment records. Three strata were created.

Stratum	Size	Weight	Sample Size
farms with animals	2365	0.177	150
farms without animals	1297	0.097	100
houses	9683	0.726	250

A random sample of wells was selected from each stratum and the water was tested for a large number of characteristics. Here we look at only two:

- y: sodium (Na) concentration (mg/L)
- u: the water was contaminated by coliform bacteria ($u = 1$) or not ($u = 0$)

The data are summarized below.

Stratum	**Average Na**	**St Dev Na**	**% contaminated**
farms with animals	237.3	41.45	17.2
farms without animals	245.6	37.62	11.4
houses	220.1	51.23	13.2

The estimate of the population average Na concentration μ is

$$\hat{\mu}_{strat} = 0.177(237.3) + 0.097(245.6) + 0.726(220.1) = 225.6\text{mg/L}$$

The estimated variance of the estimator $\tilde{\mu}_{strat}$ is

$$\begin{aligned} & \widehat{Var}(\tilde{\mu}_{strat}) \\ = \; & (.177)^2 \left(1 - \frac{150}{2365}\right) \frac{41.45^2}{150} + (.097)^2 \left(1 - \frac{100}{1297}\right) \frac{37.62^2}{100} + (.726)^2 \left(1 - \frac{250}{9683}\right) \frac{51.23^2}{250} \\ = \; & 5.845 \end{aligned}$$

and so the standard error is $\sqrt{5.845} = 2.418$ mg/L. An approximate 95% confidence interval for μ is

$$\begin{aligned} & \hat{\mu} \pm c\hat{\sigma} \text{ mg/L.} \\ = \; & 225.6 \pm (1.96)(2.418) \text{ mg/L.} \\ = \; & 225.6 \pm 4.7 \text{ mg/L.} \end{aligned}$$

For the binary response variate u, the stratified estimate of the proportion of contaminated wells, $\pi = W_1\pi_1 + W_2\pi_2 + W_3\pi_3$, is

$$\hat{\pi}_{strat} = 0.177(0.172) + 0.097(0.114) + 0.726(0.132) = 0.137$$

or 13.7%. The estimated variance of the associated estimator is

$$\begin{aligned} & \hat{Var}(\tilde{\pi}_{strat}) \\ = \; & (0.177)^2 \left(1 - \frac{150}{2365}\right) \frac{[0.172(1-0.172)]}{150} + (0.097)^2 \left(1 - \frac{100}{1297}\right) \frac{[0.114(1-0.114)]}{100} \\ & + (0.726)^2 \left(1 - \frac{250}{9683}\right) \frac{[0.132(1-0.132)]}{250} \\ = \; & 0.000272 \end{aligned}$$

and so the standard error is $\sqrt{0.000272} = 0.0165$. Since the response variate, u, is binary we used the approximation

$$Var(\tilde{\pi}_h) = \approx (1 - f_h) \frac{1}{n_h} \pi_h (1 - \pi_h)$$

in each stratum. The 95% confidence interval for π is $0.137 \pm (1.96)(0.0165)$ or 0.137 ± 0.032 or $13.7\% \pm 3.2\%$.

Note that you need to be careful moving from percentages to proportions. The formulae are expressed in terms of proportions.

We can compare the strata means and proportions – see Exercise 1.
There are several questions of interest:

1. When is stratified sampling more efficient than SRS?
2. How should we allocate the total sample among the strata?
3. Can we combine ratio/regression estimation with stratified sampling?

The answer to the last question is the easiest. We can estimate the strata averages in the best way possible, e.g. ratio or regression estimation if appropriate, and then combine the estimates using the stratum weights. Since each stratum is sampled independently, the estimated variance is the sum of the squared strata weights times the variance of the stratum estimates. Note that these variances are calculated using the formulae for ratio or regression estimates.

7.2 Comparison to SRS

To examine the efficiency of stratified sampling, we need to consider the sampling weights $w_i = \frac{n_i}{n}$, the stratum weights $W_i = \frac{N_i}{N}$ and the relative sizes of $\sigma_1, \ldots, \sigma_H$ versus σ. Looking at the variance of $\tilde{\mu}_{strat}$

$$\begin{aligned} Var(\tilde{\mu}_{strat}) &= W_1^2(1 - f_1)\frac{\sigma_1^2}{n_1} + \cdots + W_H^2(1 - f_H)\frac{\sigma_H^2}{n_H} \\ &\approx \left(\frac{1}{n}\right)\left(\frac{W_1^2}{w_1}\sigma_1^2 + \cdots + \frac{W_H^2}{w_H}\sigma_H^2\right) \end{aligned}$$

we see that if we were to give a very small sample weight to a stratum with high stratum weight and $\sigma_h^2 > \sigma^2$, then $Var(\tilde{\mu}_{strat}) > Var(\tilde{\mu})$. In other words, there is no uniform result; it is possible to have larger variance with stratified sampling. However, this is a contrived situation and in most cases, if we construct the strata with care, stratified sampling will be much better than SRS.

To confirm this point, consider **proportional allocation** where, except for rounding, we have $w_h = W_h$ or, in other words, n_h is proportional to N_h. We also ignore the finite population corrections. Substituting $W_h = w_h = n_h/n$, we have

$$\begin{aligned} & Var(\tilde{\mu}_{strat}) \\ = \; & W_1 \left(\frac{n_1}{n}\right)\left(\frac{\sigma_1^2}{n_1}\right) + \cdots + W_H \left(\frac{n_1}{n}\right)\left(\frac{\sigma_1^2}{n_1}\right) \\ = \; & \frac{W_1\sigma_1^2 + \cdots + W_H\sigma_H^2}{n} \end{aligned}$$

and the variance of the stratified estimator will be less than the variance of the sample average from SRS if

$$W_1\sigma_1^2 + \cdots + W_H\sigma_H^2 < \sigma^2$$

The left side is the weighted average of the within strata variances. If we form the strata so that these variances are small, i.e. form the strata so that there is greater consistency within strata compared to the whole population, then the weighted average will be less than the overall variance.

Another way to make the same point is to use the ANOVA partition of the total sum of squares into two components, **within** and **between** strata. Suppose that y_{hj} is the response variate for the j^{th} unit in stratum h. By subtracting and adding μ_h, squaring and simplifying, we can partition the total sum of squares

$$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj} - \mu)^2 = \sum_{h=1}^{H}\sum_{j=1}^{N_h}((y_{hj} - \mu_h) - (\mu - \mu_h))^2$$

as the sum of

Between strata:	$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(\mu_h - \mu)^2 = \sum_{h=1}^{H} N_h(\mu_h - \mu)^2$
Within strata:	$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj} - \mu_h)^2 = \sum_{h=1}^{H}(N_h - 1)\sigma_h^2$
Total:	$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj} - \mu)^2 = (N-1)\sigma^2$

where the sums are over the whole population. Here is a proof that the cross-term

is zero.

$$\begin{aligned}
& \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\mu_h)(\mu-\mu_h) \\
= & \sum_{h=1}^{H}\left[(\mu-\mu_h)\underbrace{\sum_{j=1}^{N_h}(y_{hj}-\mu_h)}_{=0 \text{ in each stratum } h, \text{ as follows:}}\right] \\
& \sum_{j=1}^{N_h}(y_{hj}-\mu_h) \\
= & \sum_{j=1}^{N_h}y_{hj}-\sum_{j=1}^{N_h}\mu_h \\
= & N_h\left(\frac{\sum_{j=1}^{N_h}y_{hj}}{N_h}\right)-N_h\mu_h \\
= & N_h\mu_h-N_h\mu_h \\
= & 0.
\end{aligned}$$

We have

$$\begin{aligned}
\sigma^2 &= \frac{\sum_h\sum_j(y_{hj}-\mu)^2}{N-1} \\
&= \frac{\sum_h N_h(\mu_h-\mu)^2}{N-1}+\frac{\sum_h(N_h-1)\sigma_h^2}{N-1} \\
&\approx \sum_h W_h(\mu_h-\mu)^2+\sum_h W_h\sigma_h^2
\end{aligned}$$

the difference in variance for stratified versus sample average estimator is proportional to $\sum_{h=1}^{H} W_h(\mu_h-\mu)^2$. For proportional allocation (and for any other allocation), we should **make the stratum means as different as possible** in order to achieve the greatest gain over SRS.

7.3 Optimal Allocation

Suppose that at the Plan stage, we decide that we can afford to select a sample of size n. How should we divide the sampling effort among the strata if the objective is to minimize the variance of the resulting estimator? This is the **allocation problem**. We want to determine $n_1,\ldots,n_H$ so that $n_1+\cdots+n_H=n$ and $Var(\tilde{\mu}_{strat})$ is minimized. We treat $n_1,\ldots,n_H$ as continuous variables and use a

Lagrange multiplier. That is, we find the critical point of the function

$$\begin{aligned} & f(n_1, \ldots, n_H, \lambda) \\ = \; & W_1^2(1-f_1)\frac{\sigma_1^2}{n_1} + \cdots + W_H^2(1-f_H)\frac{\sigma_H^2}{n_H} + \lambda(n_1 + \cdots + n_H - n) \\ = \; & W_1^2\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\sigma_1^2 + \cdots + W_H^2\left(\frac{1}{n_H} - \frac{1}{N_H}\right)\sigma_H^2 + \lambda(n_1 + \cdots + n_H - n). \end{aligned}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial f}{\partial n_h} &= -\frac{W_h^2\sigma_h^2}{n_h^2} + \lambda, \; h = 1, \ldots, H \qquad (7.1) \\ \frac{\partial f}{\partial \lambda} &= n_1 + \cdots + n_H - n. \qquad (7.2) \end{aligned}$$

Setting these to zero, we get

$$n_h = \frac{W_h\sigma_h}{\sqrt{\lambda}}$$

and solving for λ,

$$\begin{aligned} n &= \frac{W_1\sigma_1 + \cdots + W_H\sigma_H}{\sqrt{\lambda}}, \text{ or} \\ \sqrt{\lambda} &= \frac{W_1\sigma_1 + \cdots + W_H\sigma_H}{n}. \end{aligned}$$

Hence we have, for optimal allocation

$$n_h = \left(\frac{W_h\sigma_h}{W_1\sigma_1 + \cdots + W_H\sigma_H}\right) n \qquad (7.3)$$

or more simply n_h is proportional to $W_h\sigma_h$. We allocate more sampling effort to those strata that have higher weight or larger within stratum standard deviation. If we ignore the *fpc*, then for the **optimal allocation**, we get

$$\begin{aligned} Var(\tilde{\mu}_{strat}) &= W_1^2\frac{\sigma_1^2}{n_1} + \cdots + W_H^2\frac{\sigma_H^2}{n_H} \\ &= W_1^2\frac{\sigma_1^2}{\left(\frac{nW_1\sigma_1}{W_1\sigma_1+\cdots+W_H\sigma_H}\right)} + \cdots + W_H^2\frac{\sigma_H^2}{\left(\frac{nW_H\sigma_H}{W_1\sigma_1+\cdots+W_H\sigma_H}\right)}, \text{ using line (7.3)} \\ &= \left(\frac{W_1\sigma_1 + \cdots + W_H\sigma_H}{n}\right)(W_1\sigma_1 + \cdots + W_H\sigma_H) \\ &= \frac{(W_1\sigma_1 + \cdots + W_H\sigma_H)^2}{n} \end{aligned}$$

Note that **proportional allocation** is optimal if σ_h is the same for each stratum. In detail, if $\sigma_h = \sigma$ holds for all h, then optimal allocation gives

$$\begin{aligned} n_h &= \left(\frac{W_h \sigma_h}{W_1 \sigma_1 + \cdots + W_H \sigma_H} \right) n \\ &= \left(\frac{W_h \sigma}{W_1 \sigma + \cdots + W_H \sigma} \right) n \\ &= \left(\frac{W_h}{\underbrace{W_1 + \cdots + W_H}_{=1}} \right) n \\ &= W_h n, \text{ i.e. proportional allocation.} \end{aligned}$$

In order to use the optimal allocation, we need to know (unlikely) or have an estimate of the within-stratum standard deviations, perhaps from a pilot survey, before selecting the sample. If we do so and decide to use optimal allocation, then we can use the preceding formula to select the total sample size n to achieve a confidence limit with a predetermined length. That is, for a given level of confidence, the approximate confidence interval has length

$$2\ell = \frac{2c}{\sqrt{n}}(W_1 \sigma_1 + \cdots + W_H \sigma_H)$$

so we select a sample with total size

$$n = \frac{c^2}{\ell^2}(W_1 \sigma_1 + \cdots + W_H \sigma_H)^2$$

where c is a value from the $G(0,1)$ tables determined by the level of confidence.

7.4 Forming the Strata

Stratified sampling can produce large increases in precision (i.e. shorter confidence intervals) compared to SRS for the same total sample size. Put in another way, for a given level of precision, we can use a smaller sample size with stratified sampling. However, stratification adds complexity; for example, we need to identify the stratum for each unit in the frame before we begin.

The first consideration is the purpose of the survey. In many cases such as the labour force survey, we want to estimate the rate of unemployment in each province so it is natural to stratify by province. Since we are interested in the provincial rates, we need to ensure that each province gets a large enough sample to estimate the within-stratum rate so here the allocation problem is very different.

If we are interested only in the overall frame average or total, we form the strata so that the averages are likely to be very different. If we have complete knowledge of some explanatory variate that we believe to be related to the response variate, we can use the values of the explanatory variate to form the strata.

Proportional allocation is popular because we do not need to know the within strata standard deviations and we can be almost sure to do better than with SRS. We can use estimates from a pilot study or an earlier version of the survey to optimally allocate the sample across the strata.

In some cases, a particular stratum may be so important that we do a complete census. For example, in many applications in auditing, accounts are stratified on the basis of stated value. The likelihood of large errors is greater in larger accounts so every account in the stratum of the largest accounts is included in the sample.

7.5 Post Stratification

We now return to an issue discussed in Chapter 6. Suppose there is a discrete explanatory variate such as gender or age class and we know the proportion of the population that falls in each class. This corresponds to knowing the mean for a continuous explanatory variate that we might use in a ratio or regression estimate.

We cannot use the discrete variate to form strata since we do not know the value of the variate for every unit in the frame. Instead, we know the population proportions or weights $W_1, \ldots, W_H$ for the H classes.

We select a sample of size n using SRS from the frame and once we have the sample, we can determine the class for each unit. Thus we can observe $n_1, \ldots, n_H$ units in each class. The sample sizes are not controlled and if we were to repeat the sampling they would change. A natural estimate of the population average is

$$\hat{\mu}_{post} = W_1\hat{\mu}_1 + \cdots + W_H\hat{\mu}_H.$$

We call this the **post-stratification estimate** because we do not establish the stratum for each unit in the sample until after it is selected. The estimate looks like the stratified estimate – the estimators are different because the denominator of $\tilde{\mu}_h$ is random for the post-stratification estimator.

To determine the properties of this procedure, we have a small aside.

Aside:

Suppose X and Y are two discrete random variables with joint probability function $P(X = x, Y = y)$. Then, we can write

$$\begin{aligned} E(X) &= \sum_y \sum_x xP(X = x, Y = y) \\ &= \sum_y \left[\sum_x xP\left(X = x \mid Y = y\right) \right] P(Y = y) \end{aligned}$$

The expression in [ ] is the conditional expected value of X for a given value $Y = y$ and is written $E(X|Y = y)$. Note that $E(X|Y = y)$ is a function of y **only** since we have summed over all values of x. With this notation we have

$$E(X) = \sum_y E(X|Y = y)P(Y = y)$$

The right side is the expected value of the function $E(X|Y = y)$ so we write

$$E(X) = E\left[E(X|Y = y)\right]$$

In words, we can calculate the expected value of X in two steps. First, find the conditional expectation for each value of y and second, find the expected value of the conditional expectation over the distribution of Y.

We use the result from the Aside to find $E(\tilde{\mu}_{post})$. Consider the two random variables $\tilde{\mu}_h, \tilde{n}_h$. Then we have

$$E(\tilde{\mu}_h) = E\left[E(\tilde{\mu}_h|\tilde{n}_h = n_h)\right]$$

As long as $n_h \neq 0$, we know using the results from SRS that $E(\tilde{\mu}_h|\tilde{n}_h = n_h) = \mu_h$. If we ignore the event $n_h = 0$ (which happens with small probability in large samples) we have a good approximation

$$\begin{aligned} E(\tilde{\mu}_h) &= E\left[E(\tilde{\mu}_h|\tilde{n}_h = n_h)\right] \\ &\approx E(\mu_h) \\ &= \mu_h. \end{aligned}$$

Hence we have

$$\begin{aligned} E(\tilde{\mu}_{post}) &= E(W_1\tilde{\mu}_1 + \cdots + W_H\tilde{\mu}_H) \\ &= W_1\mu_1 + \cdots + W_H\mu_H \\ &= \mu. \end{aligned}$$

The post stratified estimate is unbiased (almost).
To find the variance, we need a second result.

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 \\ &= E[E(X^2|Y=y)] - E[E(X|Y=y)]^2 \\ &= E[E(X^2|Y=y) - E(X|Y=y)^2] + E[E(X|Y=y)^2] - E[E(X|Y=y)]^2. \end{aligned}$$

We now interpret the two pieces. The expression inside the first [ ] is $Var(X|Y = y)$ so the first term is $E[Var(X|Y = y)]$. The second and third terms are $Var[E(X|Y = y)]$ so we have the result

$$Var(X) = E[Var(X|Y=y)] + Var[E(X|Y=y)]$$

To find $Var(\tilde{\mu}_{post})$, we condition on $\tilde{n}_1 = n_1, \ldots, \tilde{n}_H = n_H$.
Since $E(\tilde{\mu}|\tilde{n}_1 = n_1, \ldots, \tilde{n}_H = n_H) = \mu$ for all values of $n_1, \ldots, n_H$, we have

$$\begin{aligned} Var[E(\tilde{\mu}_{post}|\tilde{n}_1 = n_1, \ldots, \tilde{n}_H = n_H)] &= Var[\mu] \\ &= 0. \end{aligned}$$

Also, from SRS, we have

$$\begin{aligned} & Var(\tilde{\mu}_{post}|\tilde{n}_1 = n_1, \ldots, \tilde{n}_H = n_H) \\ = \; & W_1^2 \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\sigma_1^2 + \cdots + W_H^2 \left(\frac{1}{n_H} - \frac{1}{N_H}\right)\sigma_H^2 \end{aligned}$$

and so

$$\begin{aligned} & E[Var(\tilde{\mu}_{post}|_{\tilde{n}_1=n_1,\ldots,\tilde{n}_H=n_H})] \\ = \; & W_1^2 \left(E\left[\frac{1}{\tilde{n}_1}\right] - \frac{1}{N_1}\right)\sigma_1^2 + \cdots + W_H^2 \left(E\left[\frac{1}{\tilde{n}_H}\right] - \frac{1}{N_H}\right)\sigma_H^2. \end{aligned}$$

Combining the two pieces we get

$$Var(\tilde{\mu}_{post}) = W_1^2 \left(E\left[\frac{1}{\tilde{n}_1}\right] - \frac{1}{N_1}\right)\sigma_1^2 + \cdots + W_H^2 \left(E\left[\frac{1}{\tilde{n}_H}\right] - \frac{1}{N_H}\right)\sigma_H^2.$$

We approximate this variance by

$$\hat{Var}(\tilde{\mu}_{post}) = W_1^2 \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\hat{\sigma}_1^2 + \cdots + W_H^2 \left(\frac{1}{n_H} - \frac{1}{N_H}\right)\hat{\sigma}_H^2$$

that is identical to the variance of the stratified estimator for the observed allocation $n_1, \ldots, n_H$.

Example
A market research organization interviews a randomly selected sample of 300 households in a community to estimate the average amount of money spent on online rental of TV shows and movies in the previous week. From census data, they know the distribution of household size in the community but this information is not available for each unit in the frame. They post stratify the data as follows.

Household size	Population weight (census)	Sample size	Sample weight	Sample average	Sample standard deviation
1	0.232	87	0.290	13.45	3.67
2	0.381	109	0.363	20.22	4.56
3	0.193	54	0.180	25.67	5.89
4	0.123	27	0.090	28.21	5.23
> 4	0.071	23	0.077	28.10	6.77
Total	1.000	300	1.000		

The estimated average is

$$\begin{aligned} & \hat{\mu}_{post} \\ = \; & (0.232)(13.45) + (0.381)(20.22) + (0.193)(25.67) + (0.123)(28.21) + (0.071)(28.10) \\ = \; & 21.24 \end{aligned}$$

and the estimated standard deviation of the corresponding estimator (ignoring fpc's) is $\sqrt{0.085327283} = 0.292$. Here is a detailed computation of the estimated variance.

$$\begin{aligned} & \widehat{Var}(\tilde{\mu}_{post}) \\ = \; & W_1^2 \left(\frac{\sigma_1^2}{n_1}\right) + W_2^2 \left(\frac{\sigma_2^2}{n_2}\right) + W_3^2 \left(\frac{\sigma_3^2}{n_3}\right) + W_4^2 \left(\frac{\sigma_4^2}{n_4}\right) + W_5^2 \left(\frac{\sigma_5^2}{n_5}\right) \\ = \; & (0.232)^2 \left(\frac{(3.67)^2}{87}\right) + (0.381)^2 \left(\frac{(4.56)^2}{109}\right) + (0.193)^2 \left(\frac{(5.89)^2}{54}\right) \\ & + (0.123)^2 \left(\frac{(5.23)^2}{27}\right) + (0.071)^2 \left(\frac{(6.77)^2}{23}\right) \\ = \; & 0.085327283. \end{aligned}$$

An approximate 95% confidence interval for the population average amount spent is 21.24 ± 0.57.

Remarks:

1. Note that the sample average of

   $$\begin{aligned} & (0.290)(13.45) + (0.363)(20.22) + (0.180)(25.67) + (0.090)(28.21) + (0.077)(28.10) \\ = \; & 20.56 \end{aligned}$$

   has been adjusted upward because of the over-representation of households of size 1 in the sample.

2. In the above example, we ignored non-response. The company telephoned many more than 300 households to get the required number of completions. Non-response is a major source of error when sampling human populations. The confidence intervals that we have constructed do **not** take this error into account. There are many analytic methods and sampling strategies to deal with this important issue.

7.6 Exercises

1. In many surveys, there is interest in estimating strata averages or differences in strata averages.

   (a) In general, for SRS, write down the distribution for the estimators $\tilde{\mu}_h$ and $\tilde{\mu}_h - \tilde{\mu}_k$.

   (b) In the well survey, find a 95% confidence interval for the proportion of wells in farms with animals that are contaminated.

   (c) In the well survey, find a 95% confidence interval for the average Na difference between the two types of farm wells.

2. Suppose that the purpose of the survey is to estimate a population proportion π. If there are H strata,

   (a) Write down the stratified estimate of π and the variance of the corresponding estimator.

   (b) What is the variance of $\tilde{\pi}_{strat}$ for proportional allocation?

   (c) How should the strata be formed so that the stratified sampling protocol is superior to SRS?

3. Suppose the well survey was to be re-done with the same overall sample size 500. How would you recommend allocating the sample to the strata if

   (a) Estimating the average Na level was the primary goal.

   (b) Estimating the proportion of contaminated wells was the primary goal.

   (c) For each case, compare the predicted standard deviations of $\tilde{\mu}_{strat}$ and $\tilde{\pi}_{strat}$ to what occurred in the current survey.

4. Consider the difference of the variances of $\tilde{\mu}_{strat}$ under proportional and optimal allocation for a sample of size n. Ignore the fpc.

   (a) Show that this difference can be written as $\left(\frac{1}{n}\right)\sum_h(\sigma_h - \overline{\sigma})^2 W_h$ where $\overline{\sigma} = \sum_h \sigma_h W_h$ is the weighted average standard deviation over the H strata.

   (b) When will the gain be large with optimal allocation relative to proportional allocation?

5. In an informal sample of math students at UW, 100 people were asked their opinion (on a 5 point scale) about the core courses and their value. One particular statement was (with scores):

"All mathematics students are required to take Stat 231." strongly agree – 1 ; agree – 2 ; neutral – 3 ; disagree – 4 ; strongly disagree – 5

The sample results, broken down by year are shown below. Estimate the average score for all math students and find an approximate 95% confidence interval for the population average. Note that SRS was not used here so were are making assumptions about the estimators that may be unwarranted. There are about 3300 students in the faculty.

Year	**Sample size**	**Population weight**	**Average score**	**Standard deviation**
1	39	0.31	2.8	1.22
2	23	0.24	3.5	1.09
3	26	0.23	3.2	1.03
4	12	0.22	3.1	0.87

Chapter 8

Non-Response in Surveys

We introduced the idea of non-response error in Chapter 4. Non-response can occur when the units in the survey are people or groups of people such as households. There is a huge literature on the effects of non-response, dealing with non-response and preventing non-response.

There are two types of non-response. Unit non-response occurs when the unit selected (i.e. a person) cannot be located or refuses to become part of the sample. Item non-response occurs when the answer to a particular question (i.e. an item) is missing. For example, a respondent may answer all questions except for those related to his/her income. This chapter addresses issues around unit non-response.

8.1 Defining Response Rates

In this section, we look at the definition of the response rate. You might think that this should be simple but it is not.

Example 1: Three organizations conducted a telephone survey using the same questionnaire at roughly the same time with the same frame (of course with different samples). They reported the following response rates:

- Organization 1: 49.4%
- Organization 2: 69.2%
- Organization 3: 81.3%

These rates vary considerably. How is this possible? In the absence of more information we can speculate. Maybe the organization with the highest response rate did many more follow-up calls than the others. Perhaps their interviewers were better trained and more convincing on the phone. Maybe the organization with a high response rate called during the daytime, in the evening and on week-ends

whereas the others only called during business hours (a big mistake as prime calling time is in the early evening). Any of these reasons might explain the variation in the response rates. However, the most likely answer is that the organizations used different ways of calculating response rates. We will come back to this example but first some definitions.

The American Association Public Opinion Research (AAPOR) is the largest US-based organization for survey professionals. AAPOR has defined different response rates that are widely used. We introduce these definitions for a survey in which the interviewers use a sample of telephone numbers. When a number is called and answered, the first step is to determine if the person answering is eligible to be included in the survey. For example, if the interviewer reaches a fax machine, that number is ineligible. The response rate definitions divide the sample into three components:

- eligible responders
- eligible non-responders
- non-responders with unknown eligibility

We can divide the number of eligible responders into those who complete the interview and those who do not (partial response). We denote the number in the first and second groups as I and P respectively. We partition the number of eligible non-responders into refusals (R), non-contacts (NC) and other eligible non-responders (O). Telephone numbers in the sample with unknown eligibility are divided into whether the household is occupied (UH) and "unknown other" (UO). We summarize these numbers in the table below.

For a survey the response rate is the number of interviews conducted divided by the number of units contacted. However, there are various ways to count. For instance, to determine the number of interviews conducted should we include only completed interviews or do partial interviews also count. Similarly, to determine the number of units contacted it is clear we should include all units that either were interviewed or refused, but what about units with unknown eligibility? We could simply ignore the unknown eligibility units (RR5 and RR6 below), include them all (RR1 and RR2), or use an estimate of the number of units with unknown eligibility that are in fact eligible (RR3 and RR4). With this latter approach, the most common procedure is to compute the proportion of eligible households among households with known eligibility assuming that the proportion is the same of households with unknown eligibility. That is, we calculate

$$e = \frac{\#\text{ of known eligibles}}{\#\text{ of known eligibles + known ineligibles}}$$

Table 8.1: Notation for Partition of the Sample into Respondents and Non-respondents

Component	**Notation**	**Number of**
Eligible responders	I	complete interviews
	P	partial interviews
Eligible non-responders	R	refusals or break-offs
	NC	eligible non-contacts
	O	other eligible non-respondents
non-responders with unknown eligibility	UH	households with unknown occupancy and eligibility
	UO	unknown eligibility due to other reasons

and estimate the number of eligibles as $e \times (UH + UO)$.
Some of different AAPOR response rates are then

$$
\begin{aligned}
RR1 &= \frac{I}{(I+P)+(R+NC+O)+(UH+UO)} \\
RR2 &= \frac{I+P}{(I+P)+(R+NC+O)+(UH+UO)} \\
RR3 &= \frac{I}{(I+P)+(R+NC+O)+e(UH+UO)} \\
RR4 &= \frac{I+P}{(I+P)+(R+NC+O)+e(UH+UO)} \\
RR6 &= \frac{I+P}{(I+P)+(R+NC+O)}
\end{aligned}
$$

Remarks:

1. Response rates RR1 and RR2 are identical except in how partial responses are treated in the numerator.

2. Analogously, the same applies for RR3/RR4 and RR5/RR6 [RR5 is not shown]).

3. The rates RR2, RR4 and RR6 are identical except in how households of unknown eligibility are treated. RR2 treats all such households as eligible, RR4 estimates the number of eligibles as described above and RR6 assumes none of these households are eligible.

4. The AAPOR specification allows the survey researcher to estimate the number of eligibles in any way they please as long as they report how estimation was done. We often do not have information on all of the subcomponents in the formulas. The important point is how households of unknown eligibility are treated and, to a lesser extent, how incompletes are treated.

Example 2: The Survey Research Centre (SRC) at the University of Waterloo conducted a telephone survey. The family at a sampled telephone number was eligible if there was a teenager that held a (part-time) job over the previous two months. The sample is broken down by disposition codes are shown in the Disposition Table. Disposition tables are prepared by all survey organizations.

Dispositions		
Interview 1.0		
1.1	Complete	507
1.2	Final - Complete Teen	27
	sub-total	**534**
Eligible, Non-Interview 2.0		
2.11	Refusal- Household	48
2.112	Refusal - Teen	75
	sub-total	**123**
Unknown Eligibility 3.0		
3.1	Refusal- Unknown Eligibility	310
3.12	No answer/ answering machine	114
	sub-total	424
Not Eligible 4.0		
4.1	Ineligible - No Teens	860
4.2	NIS/fax/business	504
4.2	Person incompetent	7
4.33	Language Barrier	93
4.4	Cell phone/ Number Change	112
4.7	Ineligible- No Working Teens	677
4.71	Ineligible - no 2 month teens	19
	sub-total	**2272**
Total		**3353**

We first estimate eligibility among households with unknown eligibility. If you call a phone number and get a fax line, this is included in disposition code 4.2. You can always identify a fax line, so for the purpose of estimating e, we exclude the 504 fax lines from the calculation. We have

$$e = \frac{\text{known eligibles}}{\text{known eligibles and ineligibles}} = \frac{(534 + 123)}{(534 + 123) + 2272 - 504} = 27.09\%$$

and hence we estimate that there are $0.2709 \times 424 \approx 115$ eligible among the 424 unknowns. Using this estimate, we are now ready to compute response rates. Possibilities include

- Minimal response rate: $RR1 = \frac{534}{534+123+424} = 49.4\%$
- Response rate with estimated eligibles: $RR4 = \frac{534}{534+123+115} = 69.2\%$
- Maximal response rate (cheating): $RR6 = \frac{534}{534+123} = 81.3\%$

Because we have no partial responses, these are the only three response rates that can be computed. The difference between the response rates is large. These three response rates are the ones given in Example 1 to represent three different organizations. These were not separate organizations after all - just different ways of computing a response rate based on the same survey. Knowing a response rate by itself is of limited value. We also need to know how it was calculated. Survey companies understandably have a tendency to report the response rate that makes them look best often without explanation of how the rate was calculated. In general, the maximal response rate RR6 should not be used because it assumes that all households of unknown eligibility are ineligible. This assumption is not realistic.

AAPOR expressively forbids using the word "response rate" for convenience sampling and recommends calling it "participation rate" instead. This practice has not been widely adopted, however. Even scientific reports usually refer to a participation rate of a convenience sample as the response rate. This is problematic if it leads people to believe that high participation rates of convenience samples imply a high quality survey. A high response rate in a probability sample is desirable because the probability sample is the basis for inference. Convenience samples do not have such a basis for inference. It is not clear why high participation rates are important.

Response rates have been declining. A response rate of 80% is considered to be very good; however this is almost never achieved in practice these days. People get flooded with surveys. They cannot selectively respond to important surveys because they often have difficulties distinguishing between surveys based on probability samples which can be used for inference and those conducted with convenience samples. Marketing surveys are almost never probability surveys.

In other cases surveys are abused as a promotional tool. For example, after a few legitimate questions, the survey turns into an infomercial for specific products or political ideology. Some surveys contain lead-in questions of the type "Don't you agree that ...". Those surveys are then sometimes used to make claims "In a survey of 10,000 Canadians, 85% believed [...]". With readily available

web survey technology, surveys have also become more abundant. Whatever the reason, response rates are lower than a decade to two ago.

8.2 Response Rates are often Over-interpreted

Why do we care about the response rate? If non-responders would have answered differently than responders, estimates based on the respondent data will be biased. If the response rate is very high, i.e. if there are few non-responders, then the potential for bias is low.

The potential-for-bias argument is a valid theoretical argument. However, an influential paper (Groves, 2006) showed in practice the nonresponse rate is a poor predictor of nonresponse bias. The paper was based on an empirical analysis of surveys with response rates ranging from 15% to 75%. Therefore, the results may not apply to response rates outside of this range (extrapolation).

What does this mean in practice? A survey with a response rate of 80% remains desirable. When comparing two surveys it is not obvious that a survey with a response rate of 40% is preferable to a survey with a response rate of 20%. The opportunity for bias is large in both surveys. Response rate is only one consideration. The Groves (2006) result implies in practice other factors may affect bias more than nonresponse. It is difficult to quantify how carefully survey questions are designed and how well a survey was pretested, yet those elements are probably more important to overall survey quality then the response rate achieved.

In the absence of more concrete guidance from researchers, survey practitioners tend to mistakenly judge the quality of a survey based on what is easy to calculate: the response rate.

8.3 Non-response Bias

Non-respondents may be different than respondents with respect to a variable of interest. For example, if we ask about income and people with very high income are less likely to respond, then our estimate of average income based on responders only will likely underestimate average income in the population. The estimate has **non-response bias**.

We can derive an expression for the non-response bias as follows. Denote the population size by N, the number of people in a population who would respond to a survey by N_R and the number of respondents who would not respond in a

survey by N_M so $N = N_R + N_M$. Then, the true population mean is

$$\mu = \frac{N_R}{N}\mu_R + \frac{N_M}{N}\mu_M,$$

where μ_R and μ_M are the means in the populations of responders and non-responders respectively. Assuming we use SRS to get a sample of respondents, then $E[\tilde{\mu}] = \mu_R$ and the nonresponse bias in the sample average is

$$\begin{aligned} E[\tilde{\mu} - \mu] &= \mu_R - \underbrace{\left(\frac{N_R}{N}\mu_R + \frac{N_M}{N}\mu_M\right)}_{E(\mu)=\mu} \\ &= \left(\frac{N_R + N_M}{N}\right)\mu_R - \left(\frac{N_R}{N}\mu_R + \frac{N_M}{N}\mu_M\right) \\ &= \frac{N_M}{N}(\mu_R - \mu_M). \end{aligned}$$

Example 3: Suppose you want to estimate the proportion, π, of cottages without a flush toilet in cottage country in some clearly delineated rural area in Ontario. Suppose there are 2000 cottages in the area of interest. Further, suppose you could (in principle) determine the flush toilet status of 1600 of these cottages (either by knocking at the door; or by peeking through the windows of unoccupied cottages). Suppose, the remaining 400 cottages are unoccupied and/or too inaccessible so you cannot easily determine the whether they have a flush toilet or not.

Further suppose among the 1600 cottages for which the flush toilet status can be determined 10% do not have a toilet, but among the 400 cottages of unknown flush toilet status 25% do not have a toilet. The estimate of cottages without a flush toilet based on respondents for a sample would be about 0.10 (subject to variation). We can compute the bias as

$$\text{Bias} = \frac{400}{2000}\,(0.10 - 0.25) = -0.03$$

The bias is -3%. Therefore, an unbiased estimate is 13%.

As clever as this example may look, it is useless in practice since we do not have an estimate of μ_M (neither from the population nor the sample) and therefore we cannot estimate the bias. Occasionally, however, we can obtain an estimate of μ_M. The next section describes one such approach.

8.4 Correcting for Nonresponse using Two Phase Sampling

Non-response is a problem because the subpopulation of non-responders may be different than respondents with respect to the population parameter we want to

estimate. We cannot adjust for bias because we do not know how non-respondents would answer the survey. One proposed solution is to survey a subsample of non-respondents. This idea is crazy! Non-respondents are people who do not respond to the survey so how can we survey them? We either approach them in a different way (phone survey instead of web survey), incentivize them more (double the incentive) or differently (if we previously offered a Starbucks coupon as an incentive to participate in the survey, we might now offer cash), or we might try calling them at different times (maybe they were on vacation). We clearly do not expect a high response rate among non-respondents but even a 5% response rate gives us some information.

Suppose we want to estimate a population mean. There are two steps:

- Phase I: use SRS with sample size n (regular survey)
- Phase II: use SRS to select a subsample of size m from the non-respondents from Phase I (follow-up survey)

We have effectively split the population into two subpopulations, responders and non-responders and are getting estimates for each. This is (post-)stratified sampling with two strata, where the stratum sizes N_R and N_M are unknown. However, the fraction of respondents $\frac{N_R}{N}$ can be estimated by $\frac{n_R}{n}$ and the fraction of non-respondents $\frac{N_M}{N}$ by $\frac{n_M}{n}$. Since

$$\mu = \frac{N_R}{N}\mu_R + \frac{N_M}{N}\mu_M,$$

therefore we can estimate μ as

$$\hat{\mu} = \frac{n_R}{n}\hat{\mu}_R + \frac{n_M}{n}\hat{\mu}_M,$$

where $\hat{\mu}_M$ is the estimate of the population mean of the population of non-respondents in Phase II, and $\hat{\mu}_R$ is the estimate of the population mean of the population of responders from Phase I.

Example 4: In Ontario, legal immigrants have to wait three months until they qualify for the Ontario Health Insurance Plan (OHIP). In the meantime, they or their employer have to purchase health insurance. However, not all doctors are willing to accept patients with insurances other than OHIP because it increases paperwork. The purpose of the survey is to estimate, the proportion of family physicians practicing in the Region of Waterloo who would answer "yes" to the question "Do you accept patients who do not have OHIP?"

Using simple random sampling (SRS), we selected $n = 100$ family physicians from the $N = 300$ in the population and asked the question. However, physicians are

busy and in Phase I, only 30 of the 100 physicians responded. There is also a concern that physicians who do not accept OHIP are embarrassed and are less likely to respond. If this is true, then the proportion of responders answering "yes" would overestimate the corresponding population proportion. Therefore we conducted a follow-up survey offering \$ 50 for the response. (From experience, offering less than \$ 50 for a survey of physicians is pointless.) From the 70 non-responding physicians we selected a subsample of $m = 20$ at random. Of these 20 physicians in Phase II, 15 responded. The results are shown below.

Table 8.2: Summary of Responses to the OHIP physician survey

	Sample Size	Number of Respondents	Number answering "yes"
Phase I	100	30	20
Phase II	20	15	3

Here,

$$\begin{aligned} N &= 300 \\ \underbrace{100}_{n} &= \underbrace{30}_{n_R} + \underbrace{70}_{n_M} \\ \hat{\pi}_R &= \frac{20}{30} = 0.667 \\ m &= 20 \\ \hat{\pi}_M &= \frac{3}{15} = 0.2. \end{aligned}$$

This is two-phase sampling from a population of size $N = 300$, where each phase contributes a stratum to a (post-)stratified design. Let N_R be the total number of responders in the population and let N_M be the total number of non-responders. From the observed data, we estimate

$$N_R = \left(\frac{n_R}{n}\right) N = \left(\frac{30}{100}\right)(300) = (3)(30) = 90,$$

and

$$N_M = \left(\frac{n_M}{n}\right) N = \left(\frac{70}{100}\right)(300) = (7)(30) = 210.$$

We organize the observed data in this (post-stratified) table.

h	Stratum	Weight (estimated)	N_h (estimated)	Sample Size n_h	$\hat{\pi}_h$	$\hat{\sigma}_h^2 = \hat{\pi}_h(1-\hat{\pi}_h)$
1	Responders	0.3	90	30	$\frac{20}{30} = \frac{2}{3}$	$\left(\frac{2}{3}\right)\left(1-\frac{2}{3}\right) = \frac{2}{9}$
2	Non-Responders	0.7	210	15	$\frac{3}{15} = \frac{1}{5}$	$\left(\frac{1}{5}\right)\left(1-\frac{1}{5}\right) = \frac{4}{25}$

Remark: The entries in the Sample Size n_h column are the **counts of observed values**, i.e. the counts of units that actually responded in each phase, **not** the number of surveys that we initially sent out.

The estimate of π is

$$\hat{\pi} = \frac{n_R\hat{\pi}_R + n_M\hat{\pi}_M}{n} = \frac{30\left(\frac{2}{3}\right) + 70\left(\frac{1}{5}\right)}{100} = 0.34.$$

The estimated percentage of Phase I physicians accepting patients with insurances other than OHIP, 67%, is much larger than the Phase II estimate, 20%. Because the stratum of the responders is much smaller than the stratum of non-responders, the estimate for the population is closer to 20% than to 67%. Our initial hunch was right: physicians were reluctant to admit that they do not take patients with insurances other than OHIP. In this case it was very important to employ two-phase sampling.

The estimate of the total number of family physicians in the Region of Waterloo who would answer "yes" is

$$N\hat{\pi} = (300)(0.34) = 102.$$

It is more difficult to estimate the standard deviation of $\tilde{\pi}$ because the numbers of respondents and non-respondents are variable if we imagine repeating the survey over and over.
We can use the results from the section on post-stratification in Chapter 7 to get a standard error.
In our example, this gives

$$\begin{aligned}
& \widehat{Var}(\tilde{\pi}) \\
= \; & W_1^2\left(1-\frac{n_1}{N_1}\right)\frac{\hat{\sigma}_1^2}{n_1} + W_2^2\left(1-\frac{n_2}{N_2}\right)\frac{\hat{\sigma}_2^2}{n_2} \\
= \; & \left(\frac{3}{10}\right)^2\left(1-\frac{30}{90}\right)\left(\frac{\frac{2}{9}}{30}\right) + \left(\frac{7}{10}\right)^2\left(1-\frac{15}{210}\right)\left(\frac{\frac{4}{25}}{15}\right) \\
= \; & \left(\frac{3}{10}\right)^2\left(\frac{2}{3}\right)\left(\frac{1}{135}\right) + \left(\frac{7}{10}\right)^2\left(\frac{13}{14}\right)\left(\frac{4}{375}\right) \\
= \; & 0.005297777,
\end{aligned}$$

so that the estimated standard deviation is $\sqrt{0.005297777} = 0.072785835$.
Inclusion Probabilities: In this two-phase sampling, not every unit has the same probability of being included in the final sample because we surveyed only a subsample of the Phase I non-respondents. Any respondent is included with probability $\frac{100}{300} = \frac{1}{3}$.

- Let A be the event that a non-responder is included in the Phase I sample.
- Let B be the event that a non-responder is included in the Phase II sample.

Any non-respondent is then included with probability

$$P(B \cap A) = P(B|A)P(A) = \left(\frac{20}{70}\right)\left(\frac{100}{300}\right) = \left(\frac{2}{7}\right)\left(\frac{1}{3}\right) = 0.095.$$

In estimating the total we must take into account that we only surveyed a subsample of the non-responders. We estimate the total as:

$$N\hat{\pi} = \frac{N}{n}\left(n_R\hat{\pi}_R + n_M\hat{\pi}_M\right).$$

Example 4 (continued): With $N = 300$ and, from above, $\hat{\pi} = 0.34$, the estimate for the total is $N\hat{\pi} = (300)(0.34) = 102$. We estimate that 102 of the 300 physicians accept patients with an insurance other than OHIP.
What assumptions does the two-phase sampling make?

- The method assumes there is no non-response bias in the second phase. This is reasonable if the response rate in Phase II is very high. In the example, the response rate is $\frac{15}{20} = 75\%$ in the second phase. However, in practice this is unrealistic as non-respondents are difficult people to get a hold of to begin with.
- Alternatively, one can hope non response bias in Phase II is very small. This may be true because both respondents and non-respondents in Phase II were non-respondents in Phase I and it is more difficult to imagine a reason why they should be different.
- Alternatively, one can employ a third phase.
- This example illustrates why it is often more important to devote resources to follow up on non-respondents than to select a larger (Phase I) sample.
- Follow-up on non-respondents reduces bias whereas a larger sample size reduces variability.

8.5 Exercises

1. Drawing on the similarities between the AAPOR formulas, what is the formula for RR5 and why?
2. The survey report for the Parent and Teens survey contains the following comment: "The UW SRC considers respondents that appear incompetent or to have a significant language barrier to be not eligible for the survey (under 4.0) whereas AAPOR designates these records as eligible, non-interview (under 2.0)". What would be the response rates RR1, RR4, RR6 if these were considered eligible?

References

- Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." Public Opinion Quarterly 70:646–75.

Chapter 9

Exercises - Solutions

9.1 Chapter 1 Solutions

1. Look at the description of the investigation to compare two versions of the same Web page (`http://www.dack.com/web/flashVhtml/`). Use the notion of study error to criticize the conclusion that the Flash version is inferior to the HTML version.

   Solution: The target population is all web users who may wish to purchase items from the store. The study population appears to be employees from the author's workplace many of whom are CS experts. There is a strong possibility of study error since the study population is likely to be much more proficient using either version than the target population

2. In a marketing investigation, a large supermarket chain wants to look at the effect of shelf placement for three brands of coffee. There are three shelf positions top, middle and bottom. For a given treatment, the response variate is the total sales ($) over a one week period. There are 24 stores available.

   (a) Explain why there are 6 different treatments.

   Solution: The treatments are vertical arrangements of three letters A, B, C corresponding to the three brands. There are $3! = 6$ such arrangements.

   (b) Describe an experimental plan that does not involve blocking be sure to discuss randomization and replication.

   Solution: Divide the 24 stores into six groups of four at random. Then assign a different treatment to each group. Note that there are four units (replication) within each group.

   (c) Repeat part 2b but include blocking in your plan.

Solution: Divide the stores into four blocks of six based on some explanatory variate such as previous weekly average sales. Within each block this explanatory variate should be as constant as possible but it can vary between the blocks. Then, within each of the four blocks, assign the six treatments at random to the six stores. Note that each treatment is applied to four stores, one per block (replication).

(d) Compare the two plans.

Solution: Blocking is useful to avoid confounding and to increase the precision of the treatment comparisons. It may cost a small amount to gather the information needed to form the blocks but it is almost always worthwhile.

3. To compare a distance education option versus a traditional lecture version of a course, a common examination is administered and the grades of the students are compared. The data are summarized below.

style	number of students	average	st dev
lecture	47	71.3	10.2
distance ed	36	68.7	11.3

(a) Is there any evidence of that performance is worse with distance education version? Be sure write down the model you use.

Solution: To model the data, we have

$$Y_{ij} = \mu_i + R_{ij}, \;\; R_{ij} \sim G(0, \sigma), \;\; i = 1, 2; \; j = 1, \ldots, n_i$$

where $n_1 = 47, n_2 = 36$. The estimates of the model parameters are

$$\hat{\mu}_1 = 71.3, \;\; \hat{\mu}_2 = 68.7, \; \hat{\sigma} = \sqrt{\frac{46 \times 10.2^2 + 35 \times 11.3^2}{46 + 35}} = 10.7$$

See Question 6a for the formula for $\hat{\sigma}$. To answer the question, we use a test of the hypothesis that $\theta = \mu_2 - \mu_1 = 0$. The steps are

- Suppose that the hypothesis is true i.e. $\theta = 0$
- The estimate of the parameter is $\hat{\theta} = \hat{\mu}_2 - \hat{\mu}_1 = 68.7 - 71.3 = -2.6$. The corresponding estimator is $\tilde{\theta} \sim G(\theta, \sigma\sqrt{\frac{1}{36} + \frac{1}{47}})$
- The discrepancy measure is $d = \frac{|\hat{\theta} - 0|}{\hat{\sigma}\sqrt{\frac{1}{36} + \frac{1}{47}}} = 1.098$
- The significance level is $SL = P(|t_{81}| > 1.098) \approx 0.30$
- There is no evidence against the hypothesis.

We conclude there is **no evidence** of a difference in student performance with the two teaching methods.

(b) This is not an experimental plan. Why? Does this observation have any impact on your conclusion in part 3a?

Solution: There may have been systematic differences in the two groups of students (because we did not randomly assign students to the two classes) that could explain the lack of effect. Perhaps the DE students were weaker to begin with and might have done much worse with the lecture style.

4. To compare the bias in two measurement systems I and II, a sample of 10 parts is selected from production over one day. Each part is measured on both measurement systems in a random order. The data are shown below and stored in the file *ed_exercise4.txt* in row/column format.

part	system I	system II
1	3.2	3.1
2	5.6	5.8
3	1.2	1.2
4	3.8	3.6
5	7.2	7.7
6	4.2	4.2
7	3.4	3.6
8	5.2	5.6
9	2.8	3.1
10	2.7	2.8

(a) Write a model for the repeated application of the Plan that treats parts as blocks. What parameter in the model can you use to estimate the relative bias of the two systems?

Solution: To model the repeated execution of the Plan we have

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}, \;\; R_{ij} \sim G(0, \sigma), \quad i = 1, 2; \; j = 1, \ldots, 10$$

where $\tau_1 + \tau_2 = 0$, $\sum_j \beta_j = 0$. The combination $\mu + \tau_i + \beta_j$ represents the average measurement if we repeatedly measure part j with system i. The relative bias is then $\theta = (\mu + \tau_2 + \beta_j) - (\mu + \tau_1 + \beta_j) = \tau_2 - \tau_1$.

(b) Estimate the relative bias and give a 95% confidence interval.

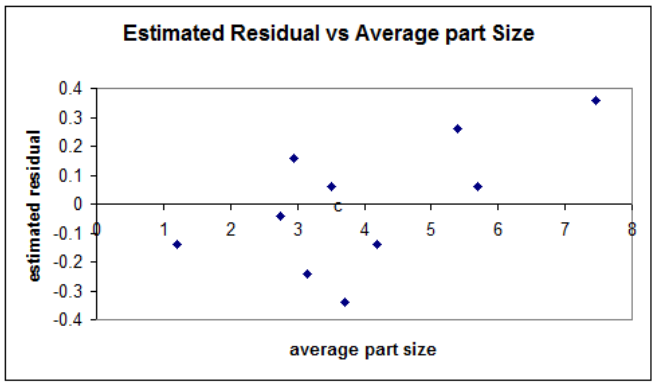
Solution: We can thus analyze the 10 differences $d_j = y_{2j} - y_{1j}$ with the corresponding model $D_j \sim G(\theta, \sigma_d) \quad j = 1, \ldots, 10$. The estimate of θ is the average difference $\hat{\theta} = -0.14$ and the estimate of σ_d is the standard deviation of the 10 differences $\hat{\sigma}_d = 0.222$ with 9 degrees of freedom. The estimator corresponding to $\hat{\theta}$ has standard deviation $\frac{\sigma_d}{\sqrt{10}}$. The 95% confidence interval is

$$\hat{\theta} \pm c \frac{\hat{\sigma}_d}{\sqrt{10}}$$

where $c = 2.262$. Substituting we get the confidence interval $(-0.299, 0.019)$. Note that there is no evidence of a difference in bias between the two systems.

(c) Plot the estimated residuals against the average measured value for each part. Does it appear that the bias changes with part size?

Solution:



The plot shows the estimated residuals for the 10 differences. There is some indication that the residuals are larger for larger part size that indicates an increase in relative bias.

(d) Consider an alternate plan in which 20 parts are selected and randomly divided into two groups of 10. Each group is measured by one system only. Why would it be more difficult to detect a bias with this plan? Note that if the measurement is destructive we must use a plan such as this.

Solution: The variability within each group of 10 parts would be large because it includes part to part differences. Larger variability (larger σ) makes it more difficult to detect a difference in bias since, for example, the confidence interval will be wider than with the plan in 4a above.

5. Suppose in an experimental plan to compare two treatments without the use of blocking, the investigator has funds to measure $2r$ units. The plan is **balanced** if r units are allocated to each treatment.

(a) For what allocation, will the standard deviation of the estimator of the treatment difference be smallest?

Solution: Suppose r_1 units are assigned to treatment 1 and r_2 to treatment 2. The standard deviation of the estimator is $\sigma\sqrt{\frac{1}{r_1}+\frac{1}{r_2}} = \sigma\sqrt{\frac{2r}{r_1 r_2}}$. We minimize the standard deviation by maximizing the denominator. Using the AM-GM mean inequality, we have

$(r_1 r_2)^{\frac{1}{2}} \leq \frac{r_1+r_2}{2} = r$ with equality if and only if $r_1 = r_2$. Hence the standard deviation is minimized if the Plan is balanced.

(b) What is the impact of this choice on the confidence interval for the treatment difference?

Solution: This choice makes the confidence interval have shortest possible length, assuming that we get the same estimate of σ from the various Plans.

6. Let's do some calculus.

(a) For the unblocked model, find the least squares estimates of μ, τ_1, τ_2 that minimize

$$W(\mu, \tau_1, \tau_2) = \sum_{i,j} (y_{ij} - \mu - \tau_i)^2$$

subject to the constraint $\tau_1 + \tau_2 = 0$.

Solution: In the sum we assume that $i = 1, 2$, $j = 1, \ldots, r$. To minimize W subject to the constraint, we use the method of Lagrange with a single multiplier. That is, we minimize the function

$$\begin{aligned} V(\mu, \tau_1, \tau_2, \lambda) &= W(\mu, \tau_1, \tau_2) + \lambda(\tau_1 + \tau_2) \\ &= \sum_{i,j} (y_{ij} - \mu - \tau_i)^2 + \lambda(\tau_1 + \tau_2) \end{aligned}$$

with respect to $\mu, \tau_1, \tau_2, \lambda$. The four partial derivatives are

$$\begin{aligned} \frac{\partial V}{\partial \mu} &= -2 \sum_{i,j} (y_{ij} - \mu - \tau_i) \\ \frac{\partial V}{\partial \tau_i} &= -2 \sum_{j} (y_{ij} - \mu - \tau_i) + \lambda, \ i = 1, 2 \\ \frac{\partial V}{\partial \lambda} &= \tau_1 + \tau_2 \end{aligned}$$

We set these derivatives to 0 and solve to find the estimates $\hat{\mu}, \hat{\tau}_1, \hat{\tau}_2, \hat{\lambda}$. Don't forget the constraint$\tau_1+\tau_2 = 0$. The first equation gives $-2\sum_{i,j}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$ or simplifying $y_{++} - 2r\hat{\mu} - r\underbrace{(\hat{\tau}_1 + \hat{\tau}_2)}_{=0} = 0$ or $\hat{\mu} = \overline{y}_{++}$. The second set of equations become $-2\sum_j (y_{ij} - \hat{\mu} - \hat{\tau}_i) + \hat{\lambda} = 0, \ i = 1, 2$ or, simplifying, $\widehat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++} - \frac{\hat{\lambda}}{2r}$. When we substitute this solution into the final equation (the constraint), we get

$$\widehat{\tau}_1 + \widehat{\tau}_2 = \left(\overline{y}_{1+} - \overline{y}_{++}\right) + \left(\overline{y}_{2+} - \overline{y}_{++}\right) - \frac{\lambda}{r} = 0$$

or $\hat{\lambda} = 0$ since $(\overline{y}_{1+} - \overline{y}_{++}) + (\overline{y}_{2+} - \overline{y}_{++}) = 0$ because $\overline{y}_{++} = \frac{\overline{y}_{1+} + \overline{y}_{2+}}{2}$. Hence we have $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}, \ i = 1, 2$ as required.

(b) For the model with b blocks, find the least squares estimates of $\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b$ that minimize

$$W(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b) = \sum_{i,j} (y_{ij} - \mu - \tau_i - \beta_j)^2$$

subject to the constraints $\sum_i \tau_i = 0$, $\sum_j \beta_j = 0$.

Solution: We minimize the scary function

$$\begin{aligned} & V(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b, \lambda, \delta) \\ = \; & W(\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b) + \lambda(\tau_1 + \tau_2) + \delta(\beta_1 + \cdots + \beta_b) \\ = \; & \sum_{i,j} (y_{ij} - \mu - \tau_i - \beta_j)^2 + \lambda(\tau_1 + \tau_2) + \delta(\beta_1 + \cdots + \beta 2_b) \end{aligned}$$

by setting the partial derivatives with respect to $\mu, \tau_1, \tau_2, \beta_1, \ldots, \beta_b, \lambda, \delta$ equal to 0.

The partial derivatives are

$$\begin{aligned} \frac{\partial V}{\partial \mu} &= -2 \sum_{i,j} (y_{ij} - \mu - \tau_i - \beta_j \\ \frac{\partial V}{\partial \tau_i} &= -2 \sum_{j} (y_{ij} - \mu - \tau_i - \beta_j) + \lambda, \; i = 1, 2 \\ \frac{\partial V}{\partial \beta_j} &= -2 \sum_{i} (y_{ij} - \mu - \tau_i - \beta_j) + \delta, \; j = 1, \ldots, b \\ \frac{\partial V}{\partial \lambda} &= \tau_1 + \tau_2 \\ \frac{\partial V}{\partial \delta} &= \sum_j \beta_j \end{aligned}$$

Solving the corresponding equations and using the constraints, the first equation gives

$$-2 \sum_{i,j} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0 \;\; \text{or} \;\; \hat{\mu} = \overline{y}_{++}$$

Setting the second equations to 0 gives $0 = -2 \sum_j (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) + \hat{\lambda}$ or $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++} - \frac{\hat{\lambda}}{b}$. Substitution in the constraint gives $\hat{\lambda} = 0$ so that $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$. Again we use the fact that the sum of the deviations from an average is 0. Similarly we have (by symmetry if you like a more formal answer) $\hat{\beta}_j = \overline{y}_{+j} - \overline{y}_{++}$.

7. Alternate forms for $\hat{\sigma}$ - these are useful for calculation purposes.

(a) In the unblocked plan, show that the estimate of the residual standard deviation can be written as

$$\hat{\sigma} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

where n_i, s_i are the sample size and sample standard deviation for treatment i respectively. Interpret this formula in words.

Solution: The estimate of σ from least squares is

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{\sum_i \sum_j (\overline{y}_{ij} - \overline{y}_{i+})^2}{r_1+r_2-2}} \\ &= \sqrt{\frac{\sum_j (\overline{y}_{1j} - \overline{y}_{1+})^2 + \sum_j (\overline{y}_{2j} - \overline{y}_{2+})^2}{r_1+r_2-2}} \\ &= \sqrt{\frac{(r_1-1)s_1^2 + (r_2-1)s_2^2}{r_1+r_2-2}}.\end{aligned}$$

Note that expression under the square root is the weighted average (weights determined by the degrees of freedom) of the squares of the sample standard deviations within each treatment.

(b) In the plan for comparing two treatments with n blocks, show that

$$\hat{\sigma} = \frac{1}{\sqrt{2}} \sqrt{\frac{\sum_j (d_j - \overline{d}_+)^2}{n-1}}$$

where $d_j = y_{1j} - y_{2j}$ is the difference of the response variates within block j. Hint: For two numbers a_1, a_2, show that $stdev(a_1, a_2) = \frac{|a_1-a_2|}{\sqrt{2}}$ and then, with $a_i = y_{ij} - \overline{y}_{i+}$, use this result in each block.

Solution: Note that $a_1 - \frac{a_1+a_2}{2} = \frac{a_1-a_2}{2}$, $a_2 - \frac{a_1+a_2}{2} = -\frac{a_1-a_2}{2}$ so the standard deviation of a_1, a_2 is

$$\sqrt{\frac{(a_1-\overline{a})^2 + (a_2-\overline{a})^2}{2-1}} = \sqrt{2\frac{(a_1-a_2)^2}{4}} = \frac{|a_1-a_2|}{\sqrt{2}}$$

For the plan with blocking we have

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{\sum_{i,j}(y_{ij} - \overline{y}_{i+} - \overline{y}_{+j} + \overline{y}_{++})^2}{2n - (n+2+1-2)}} \\ &= \sqrt{\frac{\sum_j ((y_{1j} - \overline{y}_{1+}) - (\overline{y}_{+j} - \overline{y}_{++}) + (y_{2j} - \overline{y}_{2+}) - (\overline{y}_{+j} - \overline{y}_{++}))^2}{n+1}}.\end{aligned}$$

The numerator (inside the sum) is of the form $(a_1 - \overline{a})^2 + (a_2 - \overline{a})^2 = \frac{(a_1-a_2)^2}{2}$ so we have

$$\begin{aligned} \hat{\sigma} &= \sqrt{\frac{\sum_i((y_{1j} - \overline{y}_{2j}) - (\overline{y}_{1+} - \overline{y}_{2+}))^2}{2(n-1)}} \\ &= \frac{1}{\sqrt{2}}\sqrt{\frac{\sum_j(d_j - \overline{d}_+)^2}{n-1}}, \end{aligned}$$

as required.

8. A technical assistance center (TAC) provides advice to mechanics who are trying to repair a car. The mechanic who cannot solve a problem telephones the TAC and is connected with a highly trained technician who tries to diagnose the problem and provide advice for a solution. To reduce cost, the management of the TAC decides to investigate a change. In the current system, the mechanic calls the center and the technician collects and enters information into a database about the vehicle identification, mileage and the dealership. Then the technician asks the mechanic about the problem. In the proposed system, the mechanic will enter the information about the vehicle, mileage and dealership using the keys on the telephone. Then the technician will be contacted and can start immediately to ask about the problem. The TAC management estimates that the technicians time costs about twice that of the mechanic. The computer system can automatically measure the time from the start of the call and connection to the technician. Carefully explain how you would design an experimental plan to investigate the cost saving available from the proposed system.

Solution: There are many different solutions.

The target population is all calls to the TAC in the future. The response variate is the cost associated with the call. We are interested in comparing the average cost per call with the new and old systems.

The study population is a set of calls over a finite time, say 2 weeks into the future. We need to define a set of these calls to be in the sample – assuming both systems are simultaneously available, we can take all calls and randomly assign them to the new and old system. We can block by technician (i.e. make sure that each technician gets half of his or her calls with each system). To calculate the cost of a call, we need to monitor the times of both the mechanic and technician until the technician starts to ask about the problem. If there are differential overhead costs (different by system or call) associated with each call, we should add these to get the total cost. We need to record the technician, system and call cost for each call in the sample.

9. To investigate a new packaging plan, the producer of a consumer product placed the current packaging and the new packaging side by side in 15 stores for one week and recorded the sales (in $) of each type of package. Different stores charged different amounts for the product but kept the price of each packaging style the same. The total sales to the nearest dollar are shown below. The average weekly sales from the past year were also recorded. The data are stored in the file *ed_exercise9.txt* and given below

store	old package	new package	average weekly sales
1	1406	1351	2888
2	1134	1135	2416
3	1124	1607	2684
4	497	484	816
5	621	1084	1688
6	309	726	1256
7	1172	1732	3008
8	1599	1045	2560
9	693	913	1464
10	800	533	1220
11	965	868	1724
12	445	816	1240
13	1357	1418	3076
14	868	1033	1904
15	899	1389	2112

(a) Write a brief report for your manager that includes the results of a formal analysis of the data and a conclusion. The key issue is to estimate the change in sales if the new packaging is adopted.

Solution: Comparison of Packaging Methods (Executive Summary)

To assess a new packaging plan, we placed the current and new packaging side by side in 15 stores chosen in our sale's area. We monitored the sales of each type of packaging with the following results.

Packaging type	Average Weekly Sales
Old	$925.93
New	$1075.60

The new package produced an increased average sale of about $150 per store. Based on a formal statistical analysis, we are highly confident that adopting the new packaging will increase sales on average by anywhere from -$28 to $328. This interval is very wide because there was considerable variability in the results from one store to the next.

We recommend further consideration of the new packaging. If there are no large cost disadvantages, the data suggest that we should proceed.

Here is a detailed explanation of where the above confidence interval comes from.

- Let μ be the average sales.
- Let τ_1 be the treatment effect from the old packaging.
- Let τ_2 be the treatment effect from the new packaging.

We show that a 95% confidence interval for $\theta = \tau_2 - \tau_1$ is $[-28, 328]$.

From the given data, we compute

$$\begin{aligned}
\hat{\mu} &= 1000.766667 \\
\overline{y}_{1+} &= 925.9333333 \\
\overline{y}_{2+} &= 1075.6 \\
\hat{\tau}_1 &= 925.9333333 - 1000.766667 = -74.83333333 \\
\hat{\tau}_2 &= 1075.6 - 1000.766667 = 74.83333333 \\
\hat{\theta} &= 74.83333333 - (-74.83333333) = 149.6666667 \\
\hat{\sigma}_D &= 321.4966489.
\end{aligned}$$

There are $30-(1+2+15)+(1+1) = 14$ degrees of freedom. Therefore, for 95% confidence, we select $c = 2.145$. The standard error of the differences is

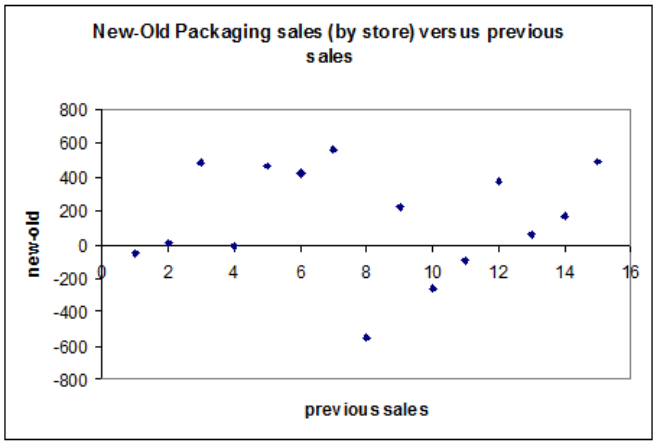
$$\frac{\hat{\sigma}_D}{\sqrt{15}} = \frac{321.4966489}{\sqrt{15}} = 83.0100778.$$

Hence a 95% confidence interval for θ is

$$149.6666667 \pm (2.145)(83.0100778), \text{ or } 149.6666667 \pm 178.0389098.$$

(b) Technical issue: Does the treatment effect depend on the past average sales?

Solution: The easiest way to look at this issue is to plot the block differences (new - old packaging) against the past average sales. From the chart on the next page, there is no pattern showing a different effect for stores with large or small past sales.



(c) The original report contained a 95% confidence interval to describe the change (on average) due to the new packaging. A manager reading the report asked the analyst for a non-technical explanation of the interval. What would you have written?

Solution: The confidence interval provides a range of plausible values for the average change in sales over all stores if the two packaging plans had been applied. If we repeated the study over and over, in 95% of the cases, the interval we construct will contain this overall average change.

(d) The company adopted the new packaging for all stores. After a brief increase in sales, they were disappointed in the results. In a review of the trial and results, the analyst pointed out that it might have been better to put only a single packaging in each store. Explain.

Solution: One problem is that during the investigation both types of packaging were available to the customers. This is different that practice once the new packaging was adopted. It would be better to use a single packaging in each store to better reflect this reality.

10. A manufacturing organization is planning to train a large number of employees in problem solving, continuous improvement and the use of simple statistical methods. After considerable research, cost comparisons etc, the training manager has narrowed the choice of training method to two. Method 1 involves external consultants who will deliver classroom training combined with role playing and hands-on exercises. Method 2 uses individual on-line training with machine generated feedback. Before committing to either approach, the manager decides to compare the two methods with an experiment. She is especially interested in knowing if the classroom training is more effective since it is more costly. She randomly assigns 24 employees into two groups of 12. She then contracts to have each group use one of the two methods. Each employee is given a test immediately after the training is

over (score is y1) and then another test two weeks later (score is y2). Carry out an analysis of the data given in the file *ed_exercise10.txt* and write a brief report on your findings.

Solution: The Analysis – not necessary as part of the solution.

There is no blocking with this Plan so that the model (for either response variate) is

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad R_{ij} \sim G(0, \sigma) \text{ independent}$$

where we have the constraint $\tau_1 + \tau_2 = 0$ on the treatment effects. To compare the two treatments, we can construct a one-sided hypothesis test for $\theta = \tau_1 - \tau_2 = 0$. We have the estimates and the estimated standard deviation for the two response variates.

Response	$\hat{\tau}_1 = \overline{y}_{1+} - \overline{y}_{++}$	$\hat{\tau}_2 = \overline{y}_{2+} - \overline{y}_{++}$	$\hat{\sigma}$ (df)	$est\ st\ dev(\tilde{\tau}_1 - \tilde{\tau}_2) = \hat{\sigma}\sqrt{\frac{1}{12} + \frac{1}{12}}$
$y1$	0.333	-0.333	5.22 (22)	2.13
$y2$	2.25	-2.25	4.90 (22)	2.00

We are interested in seeing if $\theta > 0$ so we get the discrepancy measures $d = \frac{\hat{\theta}-0}{est\ st\ dev(\hat{\theta})}$ which is 0.31 for y1 and 2.25 for y2. Calculating $P(t_{22} \geq d)$, the corresponding p-values are 0.38 and 0.02.

The Report

We compared two training methods, one involving classroom instruction and one that used on-line training. We assesses the participant's knowledge immediately after the testing and again after two weeks. The average scores are reported in the table seen below.

Assessment time	**Classroom method**	**On-line method**
immediate	73.6	72.9
after two weeks	62.8	58.3

There was no evidence of a difference in scores immediately after the test but the classroom training produced statistically higher scores for the later assessment when the average difference was 4.5 points. It is not clear to us whether this difference is practically important or whether it will persist over a longer period. [you could also say something about the drop in scores over the two week period].

11. For some parameter θ, suppose we have an estimator $\tilde{\theta} \sim G(\theta, k\sigma)$ and an estimate of σ with q degrees of freedom. Prove that a particular value of θ, say $\theta = \theta_0$, falls within the $100(1-p)\%$ confidence interval if and only if the p-value for the hypothesis $\theta = \theta_0$ is greater than or equal to p.

    Solution: The form of the confidence interval is $\hat{\theta} \pm c\hat{\sigma}k$ where the c is defined by $P(|t| \leq c) = 1 - p$ or $P(|t| \geq c) = p$. The value θ_0 is in the confidence interval iff $-ck\hat{\sigma} \leq \hat{\theta} - \theta_0 \leq ck\hat{\sigma}$ or equivalently $d = \frac{|\hat{\theta}-\theta_0|}{k\hat{\sigma}} \leq c$. To test the hypothesis $\theta = \theta_0$, the significance level is $SL = P(|t| \geq d)$. Hence $d \leq c$, iff $SL \geq p$.

12. Suppose that we have eight subjects (this is a deliberately chosen small number to avoid a lot of computation) and want to compare two treatments. We randomly split the subjects into two groups and assign each group a treatment. The data are shown below. Here we look at an alternate model and analysis based on the random assignment of the treatments.

Treatment A	**Treatment B**
6.3	5.9
5.4	6.6
5.7	6.7
5.8	6.1

    The Model: For every subject we suppose that there is a fixed value for the response variate for each treatment. That is, for each subject we have a pair (y_a, y_b) corresponding to the two values of the response variate. In the investigation, we get to see exactly one of the two elements of the pair depending on which treatment is applied according to the random allocation.

    (a) How many different ways can the groups be formed?

    Solution: There are $\binom{8}{4} = 70$ ways to pick the three subjects who get treatment A

    The Analysis: If there is no treatment difference, then we hypothesize that $y_a = y_b$ for every subject. We have observed a difference in treatment average $|5.80 - 6.325| = 0.525$.

    (b) Using this as a discrepancy measure, what is the probability that we would see such a large difference if our hypothesis is true?

    Solution: With a little work we can list all 70 subsets of size 4 from the 8 observed values and calculate the discrepancy measure for each. The values range from 0.025 to 3.425. There are 16 subsets which give a discrepancy that is greater than or equal to the observed value 0.525 so the chance (due to the random assignment) of observing a discrepancy this large is $16/70 = 0.23$.

(c) What conclusion do you draw based on the significance level found in part 12b?

Solution: If there is no treatment difference, we would see such a large difference in the averages about 23% of the time due to the random assignment so there is no evidence of a treatment difference.

We can extend the model to allow for a treatment difference. Suppose that $y_a = y_b + \theta$ for every subject so that suppose that θ represents the treatment effect.

(d) How would you test the hypothesis that $\theta = \theta_0$?

Solution: If we subtract θ_0 from each the treatment A values, then we can test the hypothesis $\theta = 0$ as above using the new data.

(e) How would you find a confidence interval for θ? [use a generalization of the result in question 3 - you do not need to prove this generalization which in fact is true.]

Solution: Based on the result in Question 3, we find all values of θ_0 with SL greater than 1 minus the confidence level. We apply part 12d over and over with different values of θ_0 until we converge on the required set of values.

(f) Suppose that we had 50 subjects in each group. How can you carry out the analyses described above?

Solution: There are $\begin{pmatrix} 100 \\ 50 \end{pmatrix}$ different ways of assigning treatment A that is far too many to list. One possibility is to randomly select 1000 of these combinations and compare the observed discrepancy to those calculated for the random sample of 1000 values. The proportion of times that the observed value exceeds the values in the set of 1000 is a good estimate of the significance level.

Note that all of the above is based on a model derived solely from the random assignment of the treatments. We will exploit the same idea when we look at the analysis of surveys.

9.2 Chapter 2 Solutions

1. Practice with the F distribution.

(a) Suppose $U \sim F_{6,24}$. Find c so that $P(U \geq c) = 0.05$.
Solution: From the tables, we have $c = 2.51$.

(b) Estimate $P(U \geq 3)$.
Solution: From the tables, interpolating we have $P(U \geq 3) \approx 0.03$

(c) What is the distribution of $\frac{1}{U}$?

Solution: Since $U = \frac{K_6^2}{K_{24}^2}$, we have $1/U \sim F_{24,6}$

(d) Find d so that $P(U \leq d) = 0.05$.

Solution: Using the result from c), $P(U \leq d) = P(1/U \geq 1/d)$and from the tables we get $1/d \approx 3.85$ so $d \approx 0.26$

(e) Show that if $W \sim t_k$ then W^2 has an F-distribution. What are the degrees of freedom?

Solution: $W^2 \sim \frac{G(0,1)^2}{K_k^2} = \frac{K_1^2}{K_k^2} = F_{1,k}$. Recall that $K_k = \sqrt{Z_1^2 + \cdots + Z_k^2}$ where $Z_i \sim G(0,1)$.

2. In a small investigation, three treatments A,B,C were compared by assigning 6 units at random to each treatment. The data are in the file *crd_rbd_exercise2.txt.*

	A	**B**	**C**
	11.82	15.46	15.43
	13.03	12.90	14.28
	10.78	14.88	14.76
	14.31	13.75	12.07
	14.21	18.59	13.80
	8.56	13.80	13.46
average	12.12	14.90	13.97
st. dev.	2.22	2.02	1.16

(a) Calculate the ANOVA table.

Solution: The treatment sum of squares $6\sum_i(\overline{y}_{i+} - \overline{y}_{++})^2$is 12 times the square of the sample standard deviation of the treatment averages. Hence the treatment sum of squares is 24.00. The total sum of squares is 17 times the square of the sample standard deviation of the 18 response variate values or 75.77. We calculate the residual sum of squares by subtraction. The completed table is shown below.

(b) Is there any evidence of a difference among the treatments?

Solution: To test the hypothesis $\tau_1 = \tau_2 = \tau_3 = 0$, the discrepancy is 3.48 and the p-value is $P(F_{2,15} \geq 3.48) = 0.057$ so there is weak evidence of a difference among the treatments.

Source	**Sum of squares**	**Degrees of freedom**	**Mean square**	**Ratio to residual ms**
Treatments	24.00	2	12.00	3.48
Residual	51.76	15	3.45	
Total	75.765	17		

(c) Treatment A was a control. Is there any evidence that the average effect of treatments B and C exceeds the effect of the control? [Use a one-sided test here - why?]

Solution: From the ANOVA table, we have $\hat{\sigma} = \sqrt{3.45} = 1.86$. Consider the contrast $\theta = \frac{\tau_2+\tau_3}{2} - \tau_1$ to compare the average of treatments 2 and 3 to the control. We want a one-sided test to see if there is evidence that $\theta > 0$. The estimate of θ is $\hat{\theta} = \frac{\overline{y}_{2+}+\overline{y}_{2+}}{2} - \overline{y}_{1+} = 2.315$ and the corresponding estimator has standard deviation $\sigma\sqrt{\frac{1}{24}+\frac{1}{24}+\frac{1}{6}} = \frac{\sigma}{2}$. To test the hypothesis, suppose that $\theta = 0$. The (one-sided) discrepancy is $d = \frac{\hat{\theta}-0}{\hat{\sigma}/2} = 2.49$ and the p-value is $P(t_{15} \geq 2.49) = 0.013$ so there is strong evidence against the hypothesis. There is strong evidence that the average treatment effect of treatments 2 and 3 exceeds the effect of the control.

(d) Find a 95% confidence interval for the difference between the effects of B and C.

Solution: The difference is $\theta = \tau_2 - \tau_3$ with estimate $\hat{\theta} = \overline{y}_{2+} - \overline{y}_{3+} = 0.93$ and corresponding estimator $\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\frac{1}{6}+\frac{1}{6}}\right)$. The 95% confidence interval for the difference in effects is $\hat{\theta} \pm c\hat{\sigma}\sqrt{\frac{1}{3}}$ or 0.93 ± 2.28.

3. Suppose we have a balanced design with t treatments and r units per treatment. If we represent the data by $y_{ij}, i = 1, \ldots t, j = 1, \ldots, r$, show algebraically that the sum of the treatments sum of squares and the residual sum of squares is the total sum of squares. That is, show that

$$\sum_i r(\overline{y}_{i+} - \overline{y}_{++})^2 + \sum_{i,j}(y_{ij} - \overline{y}_{i+})^2 = \sum_{i,j}(y_{ij} - \overline{y}_{++})^2.$$

Explain this expression in terms of the variation in the data.

Solution: We start with the right side and add and subtract the treatment average $\overline{y}_{i+}$. That is,

$$\begin{aligned}
\sum_{i,j}(y_{ij} - \overline{y}_{++})^2 &= \sum_{i,j}[(y_{ij} - \overline{y}_{i+}) + (\overline{y}_{i+} - \overline{y}_{++})]^2 \\
&= \sum_{i,j}\left[(y_{ij} - \overline{y}_{i+})^2 + 2(y_{ij} - \overline{y}_{i+})(\overline{y}_{i+} - \overline{y}_{++}) + (\overline{y}_{i+} - \overline{y}_{++})^2\right] \\
&= \sum_{i,j}(y_{ij} - \overline{y}_{i+})^2 + 2\sum_i(\overline{y}_{i+} - \overline{y}_{++})\underbrace{\sum_j(y_{ij} - \overline{y}_{i+})}_{=0} + r\sum_i(\overline{y}_{i+} - \overline{y}_{++})^2 \\
&= \sum_{i,j}(y_{ij} - \overline{y}_{i+})^2 + r\sum_i(\overline{y}_{i+} - \overline{y}_{++})^2,
\end{aligned}$$

as required since $\sum_j (y_{ij} - \overline{y}_{i+}) = 0$ for every i. The left side is the total variation. The right side is the sum of the variation within and among treatments.

4. Consider the model for the balanced completely randomized design

$$Y_{ij} = \mu + \tau_i + R_{ij}, R_{ij} \sim G(0, \sigma), i = 1, \ldots, t, j = 1, \ldots, r$$

where $\sum_i \tau_i = 0$. Show that $E\left[\frac{\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2}{t-1}\right] = \sigma^2 + r\frac{\sum_i \tau_i^2}{t-1}$. [Hint: First show that the numerator can be written as $\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2 = \sum_i r\overline{Y}_{i+}^2 - rt\overline{Y}_{++}^2$ and then exploit the fact that for any random variable U, $E[U^2] = Var(U) + E(U)^2$.]

Solution: First we expand the numerator.

$$\begin{aligned}
\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2 &= r\sum_i \overline{Y}_{i+}^2 - 2r\overline{Y}_{++}\sum_i \overline{Y}_{i+} + rt\overline{Y}_{++}^2 \\
&= r\sum_i \overline{Y}_{i+}^2 - 2rt\overline{Y}_{++}^2 + rt\overline{Y}_{++}^2 \\
&= r\sum_i \overline{Y}_{i+}^2 - rt\overline{Y}_{++}^2.
\end{aligned}$$

Hence we have $E\left[\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2\right] = rE\left[\sum_i \overline{Y}_{i+}^2\right] - rtE\left[\overline{Y}_{++}^2\right]$. Since $\overline{Y}_{i+} \sim G\left(\mu + \tau_i, \frac{\sigma}{\sqrt{r}}\right), \overline{Y}_{++} \sim G\left(\mu, \frac{\sigma}{\sqrt{rt}}\right)$, we have

$$\begin{aligned}
r\sum_i E\left[\overline{Y}_{i+}^2\right] - rtE\left[\overline{Y}_{++}^2\right] &= r\sum_i \left[(\mu + \tau_i)^2 + \frac{\sigma^2}{r}\right] - rt\left(\mu^2 + \frac{\sigma^2}{rt}\right) \\
&= rt\mu^2 + 2r\mu\underbrace{\sum_i \tau_i}_{=0} + r\sum_i \tau_i^2 + t\sigma^2 - rt\mu^2 - \sigma^2 \\
&= (t-1)\sigma^2 + r\sum_i \tau_i^2.
\end{aligned}$$

Finally, dividing by $t-1$, we have $E\left[\frac{\sum_i r(\overline{Y}_{i+} - \overline{Y}_{++})^2}{t-1}\right] = \sigma^2 + r\frac{\sum_i \tau_i^2}{t-1}$ as required.

5. For the imbalanced design with model

$$Y_{ij} = \mu + \tau_i + R_{ij}, R_{ij} \sim G(0, \sigma), i = 1, \ldots, t, j = 1, \ldots, r_i$$

and constraint $\sum_i r_i\tau_i = 0$, show that

$$\hat{\mu} = \overline{y}_{++}, \ \hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}, \ \hat{\sigma} = \sqrt{\frac{\sum_i \sum_{j=1}^{r_i} (y_{ij} - \overline{y}_{i+})^2}{\sum_i (r_i - 1)}}.$$

Solution: To find the least squares estimates, subject to the constraint, we find the critical points of the function $V(\mu, \tau_1, \ldots, \tau_1, \lambda) = \sum_i \sum_j (y_{ij} - \mu - \tau_i)^2 + \lambda \sum_i r_i \tau_i$ with respect to $\mu, \tau_1, \ldots, \tau_t, \lambda$. Setting to 0 the partial derivative with respect to μ gives $\frac{\partial V}{\partial \mu} = -2 \sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$ or $\sum_{i=1}^{t} \sum_{j=1}^{r_i} y_{ij} - \hat{\mu} \sum_i r_i - \sum_i r_i \hat{\tau}_i = 0$. Since the estimates satisfy the constraint, we get $\hat{\mu} = \overline{y}_{++}$. Setting to 0 the partial derivatives with respect to τ_i for each i and λ gives

$$\begin{aligned} \frac{\partial V}{\partial \tau_i} &= -2 \sum_j (y_{ij} - \hat{\mu} - \hat{\tau}_i) + \hat{\lambda} r_i = 0, \quad i = 1, \ldots, t \\ \frac{\partial V}{\partial \lambda} &= \sum_i r_i \hat{\tau}_i = 0 \end{aligned}$$

We solve the first equation for $\hat{\tau}_i$ in terms of $\hat{\lambda}$ and substitute into the second equation.

$$\begin{aligned} \hat{\tau}_i &= \overline{y}_{i+} - \overline{y}_{++} - \hat{\lambda}/2 \\ \sum_i r_i (\overline{y}_{i+} - \overline{y}_{++} - \hat{\lambda}/2) &= 0. \end{aligned}$$

Solving the second equation, we get $\hat{\lambda} = 0$ so that $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$ as required. The estimate of σ follows from the substitution of the estimates to get the minimum value of the least squares function. The degrees of freedom follow the normal rule.

6. In an experiment to compare five treatments A,B,C,D,E, 8 units were randomly assigned to each treatment. A partial ANOVA table for the data is shown below, along with the treatment averages.

Table 7: ANOVA Table for Completely Randomized Design

Source	**Sum of squares**	**Degrees of freedom**	**Mean square (ms)**	**Ratio to residual ms**
Treatments	3.505	4	0.876	0.571
Residual	53.665	35	1.533	
Total	57.170	39		

treatment	**A**	**B**	**C**	**D**	**E**
average	10.06	10.41	9.95	9.86	9.49

(a) Complete the ANOVA table.

Solution: The completed table is shown above.

(b) Is there any evidence of a difference among the treatments?

Solution: Suppose there are no differences. Then the discrepancy measure is 0.571, far less than 1. Hence the p-value is greater than 0.5 so there is no evidence of a treatment difference.

(c) Having completed part 6b, a novice statistician looked at the table of averages and noticed that treatment B had the highest average and treatment E the lowest. She decided to see if treatment B was significantly greater, on average, than treatment E since those averages were farthest apart. Carry out the test of hypothesis.

Solution: The estimate of the contrast $\theta = \tau_B - \tau_E$ is $\hat{\theta} = 0.98$ and the corresponding estimator has standard deviation $\sigma\sqrt{\frac{1}{8} + \frac{1}{8}}$. To test the hypothesis that $\theta = 0$, the one-sided discrepancy measure is $d = \frac{\hat{\theta}-0}{\hat{\sigma}/2} = 1.58$ and the corresponding p-value is $P(t_{35} \geq 1.58) = 0.06$.

(d) Explain why the test in part 6c is misleading.

Solution: Note that the p-value is unexpectedly small because we did not compare any two treatments but compared the treatments with the highest and lowest observed averages.

7. In an industrial study, there were two factors, feed rate and temperature with two levels each (called low and high). For each of the four treatments, 12 parts were produced. The response variate of interest was surface finish, a measure of how smooth the part is. The data are given below. Note that the lower case letter in the definition of the treatment indicates the factor was at its low level. A partial ANOVA table is also given below. The data are given in the file *crd_rdb_exercise7.txt*.

	ft	fT	Ft	FT
	0.53	0.76	0.46	0.58
	0.35	0.75	0.62	0.4
	0.37	0.31	0.82	0.66
	0.47	0.66	0.67	0.51
	0.99	0.74	0.94	0.45
	0.41	0.73	0.4	0.73
	0.04	0.61	0.53	0.93
	0.32	0.56	0.79	0.89
	0.36	0.84	0.33	0.46
	0.62	0.62	0.7	0.36
	0.31	0.67	0.51	0.65
	0.34	0.68	0.74	0.79
average	0.426	0.661	0.626	0.618

Table 8: ANOVA Table for Completely Randomized Design

Source	Sum of squares	Degrees of freedom	Mean square (ms)	Ratio to residual ms
Treatments	0.4054	3	0.135	3.88
Residual	1.5311	44	0.035	
Total	1.9360	47		

(a) Is there any evidence of a difference among the treatments?

Solution: The ANOVA table is completed above. If there are no treatment differences (i.e. $\tau_i = 0$ for all i), then the ratio of the treatment and residual mean square 3.88 is the discrepancy measure and the p-value is $P(F_{3,44} \geq 3.88) = 0.015$. There is strong evidence of a difference among the treatments.

(b) Consider the following questions:

- Does increasing the level of the feed rate effect increase the average surface finish?
- Does changing the level of temperature effect the average surface finish?
- Is the effect of changing the temperature the same for both levels of feed rate?

Construct appropriate contrasts and carry out the necessary hypothesis test to examine each of the questions.

Solution: Label the treatment effects $\tau_1, \tau_2, \tau_3, \tau_4$in the same order as the above table. Then the contrasts corresponding to the questions are:

$$\theta_1 = \frac{\tau_3 + \tau_4}{2} - \frac{\tau_1 + \tau_2}{2},\ \theta_2 = \frac{\tau_2 + \tau_4}{2} - \frac{\tau_1 + \tau_3}{2},\ \theta_3 = (\tau_4 - \tau_3) - (\tau_2 - \tau_1)$$

with corresponding estimates $\hat{\theta}_1 = 0.079,\ \hat{\theta}_2 = 0.114, \hat{\theta}_3 = -0.243$ and estimators

$$\begin{aligned}
\tilde{\theta}_1 &\sim G\left(\theta_1, \sigma\sqrt{\frac{1}{48} + \frac{1}{48} + \frac{1}{48} + \frac{1}{48}}\right) \\
\tilde{\theta}_2 &\sim G\left(\theta_2, \sigma\sqrt{\frac{1}{48} + \frac{1}{48} + \frac{1}{48} + \frac{1}{48}}\right) \\
\tilde{\theta}_3 &\sim G\left(\theta_3, \sigma\sqrt{\frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12}}\right)
\end{aligned}$$

The estimate of σ is $\hat{\sigma} = 0.187$.

For the first question, suppose that $\theta_1 = 0$. Here we need a one-sided discrepancy $d = \frac{\hat{\theta}_1 - 0}{\hat{\sigma}\sqrt{\frac{1}{12}}} = 3.46$ and $P(t_{44} \geq 1.463) = 0.075$ so there is weak evidence that $\theta_1 > 0$. In words, there is weak evidence that increasing the feed rate increases the average surface finish.

For the second question, suppose that $\theta_2 = 0$. Here we need a two-sided discrepancy $d = \frac{|\hat{\theta}_2 - 0|}{\hat{\sigma}\sqrt{\frac{1}{12}}} =$ and $P(|t_{44}| \geq 2.112) = 0.04$ so there is some evidence that $\theta_2 \neq 0$.

In words, there is some evidence that changing the temperature changes the average surface finish.

For the third question, suppose that $\theta_3 = 0$. Here we need a two-sided discrepancy $d = \frac{|\hat{\theta}_3 - 0|}{\hat{\sigma}\sqrt{\frac{1}{3}}} = 2.251$ and $P(|t_{44}| \geq 2.251) = 0.029$ so there is some evidence that $\theta_3 \neq 0$. In words, there is some evidence that the effect of changing temperature depends on the feed rate.

(c) It was decided to run the process at the settings FT. Find a 95% confidence interval for the average surface finish at this setting.

Solution: The estimate of the average is the treatment average 0.618 with associated standard error $\hat{\sigma}/\sqrt{12} = 0.054$. From the t tables we also have $P(|t_{44}| \leq 2.015) = 0.95$ so the confidence interval is 0.618 ± 0.109

(d) The treatments were not assigned to the units at random. Instead, the factor levels were set and 12 parts were run off. Briefly discuss the consequences of this decision as it affects your answers to parts 7a, 7b and 7c.

Solution: There are two dangers. First there may be confounding (i.e. some other explanatory variate was changing systematically over the 48 runs and hence the differences we attributed to changes in feed rate and temperature may be due to this unknown explanatory variate. We have lost the protection that random assignment of the treatments provides against confounding. Second, because we carried out the runs very quickly, there may be study error – the short period over which we carried out the investigation may not represent the process over a longer term.

8. In an investigation to understand the impact of packaging on sales of a consumer product, a large number of new designs was considered and tested with small focus groups. As a result, four new designs (here called B,C,D and E) were considered as possibilities for further testing. Forty stores, all having close to the same historical sales, were available and it was decided to test each of the 5 designs (four new plus the current, denoted A) in eight stores each. The designs were assigned to the stores at random and the total

sales for one week was selected as the response variate. The major questions of interest are:

- Are there significant differences among the designs?
- Are the new designs better on average than the current design?
- Design E is predicted to be the best. What average sales can be expected from this design?

Write an executive summary to address these questions. Be sure to include one table or graph that can be used to support your conclusions. Also include an appendix that provides the technical back-up for your conclusions. The data are stored in the file *crd_rdb_exercise8.txt.*

A	B	C	D	E
525	499	500	512	535
518	498	480	490	525
523	525	515	527	570
470	502	473	506	529
492	537	493	505	508
540	527	484	496	519
506	516	527	530	529
517	543	488	500	523

Solution:

Summary

To conclude our study of new packaging for product number xxxx, we field tested 4 new formats and the current packaging in 40 stores for a one week period. We measured the total sales for each store. The average sales ($) for each format are shown below.

A	B	C	D	E
511	518	495	508	530

Based on statistical analysis, there is very strong evidence that different designs produce different sales. On average we cannot conclude that the new designs are superior but as expected, design E had the highest sales averaging $530 over the eight stores in which it was tested.

Based on performance, we recommend design E. We do caution that the investigation was carried out in a group of similar sized stores (based on past sales) and we do not know if we will see the same superior performance for design E in other sized stores.

Technical Appendix

We use the standard model $Y_{ij} = \mu + \tau_i + R_{ij}$, $R_{ij} \sim G(0, \sigma)$, $i = 1, \ldots, 5$; $j = 1, \ldots, 8$.

To assess the 5 designs, the ANOVA table is

Source	DF	SS	MS	F	P
design	4	5261	1315	4.00	0.009
residual	35	11511	329		
Total	39	16772			

and the table of treatment averages is

A	B	C	D	E
511.375	518.375	495.00	508.25	529.75

To test the hypothesis that $\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0$, the discrepancy is the ratio of the design to residual mean square which is 4.00 in this case. Using the F distribution, we have $SL = P(F_{4,35} \geq 4.00) \leq 0.01$ so there is strong evidence of a difference among the designs.

To compare the new designs (on average) to the current design A, consider the contrast $\theta = \frac{\tau_2+\tau_3+\tau_4+\tau_5}{4} - \tau_1$ with corresponding estimate $\hat{\theta} = \frac{\hat{\tau}_2+\hat{\tau}_3+\hat{\tau}_4+\hat{\tau}_5}{4} - \hat{\tau}_1 = 1.469$ and estimator $\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\frac{1}{128} + \frac{1}{128} + \frac{1}{128} + \frac{1}{128} + \frac{1}{8}}\right)$. Suppose that $\theta = 0$. Because of the form of the question, we need a one-sided discrepancy $d = \frac{\hat{\theta}-0}{\hat{\sigma}\sqrt{5/32}} = 0.205$. The p-value is $P(t_{35} \geq .205) = 0.42$ so there is no evidence that, on average, the new designs produce higher average sales.

We see that design E has the highest average sales of about \$530.

9. One of the assumptions that we have made since Stat 231 is that $\hat{\sigma}$, the square root of the sum of squares of the estimated residuals divided by the degrees of freedom, is an estimate of the standard deviation σ.

   (a) For any set of numbers $u_1, \ldots, u_n$ with average $\overline{u}$, show that $\sum_i (u_i - \overline{u})^2 = \sum_i u_i^2 - n\overline{u}^2$.
   Solution:

$$\begin{aligned}
\sum_i (u_i - \overline{u})^2 &= \sum_i (u_i^2 - 2u_i\overline{u} + \overline{u}^2) \\
&= \sum_i u_i^2 - 2\overline{u}\sum_i u_i + \sum_i \overline{u}^2 \\
&= \sum_i u_i^2 - 2\overline{u}(n\overline{u}) + n\overline{u}^2 \\
&= \sum_i u_i^2 - n\overline{u}^2
\end{aligned}$$

as required.

(b) For the balanced completely randomized design, show that $E(\tilde{\sigma}^2) = \sigma^2$. [Hint: use the result from part 9a and the fact that for any random variable U, we have $Var(U) = E[U^2] - (E(U))^2]$

Solution: We can write $\tilde{\sigma}^2 = \frac{\sum_{i,j}(y_{ij}-\overline{y}_{i+})^2}{t(n-1)}$. Consider the components for each treatment in the sum separately i.e. consider $\frac{\sum_j(y_{ij}-\overline{y}_{i+})^2}{n-1} = \frac{\sum_j y_{ij}^2 - n\overline{y}_{i+}^2}{n-1}$ from the result in a). Now we look at the corresponding random variables. Note that

$$\begin{aligned} Y_{ij} &\sim G(\mu+\tau_i, \sigma), \\ \overline{Y}_{ij} &\sim G\left(\mu+\tau_i, \frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

and, for any random variable W, we have

$$Var(W) = E[W^2] - E(W)^2.$$

Hence we have

$$\begin{aligned} E[Y_{ij}^2] &= \sigma^2 + (\mu+\tau_i)^2, \\ E[\overline{Y}_{i+}^2] &= \frac{\sigma^2}{n} + (\mu+\tau_i)^2. \end{aligned}$$

Substituting we get

$$\begin{aligned} E\left[\frac{\sum_j Y_{ij}^2 - n\overline{Y}_{i+}^2}{n-1}\right] &= \frac{\sum_j E(Y_{ij}^2) - nE(\overline{Y}_{i+}^2)}{n-1} \\ &= \frac{n[\sigma^2 + (\mu+\tau_i)^2] - n\left[\frac{\sigma^2}{n} + (\mu+\tau_i)^2\right]}{n-1} \\ &= \sigma^2. \end{aligned}$$

Since each component of the sum has expectation σ^2, so does the average. That is $E(\tilde{\sigma}^2) = \sigma^2$.

(c) Suppose W is any positive random variable such that $E(W^2) = \lambda^2$. Prove that $E(W) \leq \lambda$ with equality if and only if W is constant.

Solution: $0 \leq Var(W) = E(W^2) - E(W)^2$ so $E(W)^2 \leq E(W^2) = \lambda^2$. Since W is positive, we have the required result. Note that equality can occur iff $0 = Var(W)$ i.e. W is constant.

(d) Is $E(\tilde{\sigma}) = \sigma$?

Solution: Combining the results from 9a and 9b we see that $E(\tilde{\sigma}) < \sigma$. In words we have an unbiased estimate of σ^2 and a biased estimate of σ.

10. Suppose you were asked to help plan an investigation to compare five types of road paint with respect to their durability. The response variate is the time from application of a test stripe on a highway until the paint has lost 90% of its "brightness". The rest of the planning team has no formal training in statistical planning.

    (a) Explain the notion of blocking in this context.

    Solution: To form a block, we find a set of five units (pieces of pavement where we can apply a test stripe) that are similar with respect to explanatory variates other than paint type that may affect durability. For example, we want all the units in a block to experience the same traffic load.

    (b) Why would you recommend blocking?

    Solution: Blocking will prevent confounding by the explanatory variates that are held fixed. For example, by ensuring that the units within a block all experience the same traffic load, we know that we differences in durability among the paint types within the block cannot be attributed to traffic load. Blocking also increases the precision of the treatment comparisons because the residual standard deviation is reduced by holding, for example, the traffic load fixed.

    (c) How might blocking be implemented?

    Solution: We can implement blocking by selecting sections of road that experience different traffic loads, have different pavement types and ages etc. Within each section, we specify five units, strips of pavement where we can apply the five paint types. A section is a block.

11. A seed company has produced three new varieties of corn using genetic modification. They plan to field test these varieties in south-western Ontario using their current best seller as a control. The trial is for one growing season only. The response variate is the yield, measured in kilograms per hectare. The company has available 60 two hectare test plots scattered throughout the region.

    (a) How would you recommend that blocking could be used in this trial?

    Solution: Divide each of the two hectare plots into four subplots. These subplots will be subject to the same weather and should be the same basic soil type since they are close together. Each set of four subplots is a block.

    (b) Discuss two reasons for blocking using this context.

    Solution:

i. Blocking eliminates the possibility of confounding by explanatory variates such as the weather because the weather is the same for each subplot within a block.

ii. Blocking also helps to improve the precision of the comparison of variety averages. The residual standard deviation will be smaller because it does not include the effects of changing weather. If we used a completely randomized design, the variation of the yield within each treatment would be partially due to changes in the weather from plot to plot. Blocking eliminates this variation.

12. In another packaging trial, an advertising firm wants to compare two features, colour and image to see if changing either factor impacts sales. There are 6 treatments (3 colours and two images). 30 stores are arranged in blocks of size 6 based on geographic location and within each block, the six treatments are assigned at random. The total sale over a two week period is the response variate. The data are shown below and given in the file *crd_rdb_exercise12.txt*.

	treatment						
block	**1**	**2**	**3**	**4**	**5**	**6**	**average**
1	1014	980	894	958	822	938	934.3
2	872	762	795	828	832	792	813.5
3	991	1133	941	1048	868	1054	1005.8
4	827	807	891	768	700	833	804.3
5	621	516	513	586	492	670	566.3
average	865.0	839.6	806.8	837.6	742.8	857.4	824.9

(a) Complete the ANOVA table for this investigation.

Solution: We can calculate the treatment and block sum of squares using the averages given in the table. For example, the treatment sum of squares is

$$5\sum_i(\overline{y}_{i+}-\overline{y}_{++})^2 = 5\times 5\times var(865.0,\ 839.6,\ \ldots,857.4) = 50548.3$$

where var is the sample variance (also the function in excel). Similarly the block sum of squares is

$$6\sum_j(\overline{y}_{+j}-\overline{y}_{++})^2 = 6\times 4\times var(934.3,\ldots,566.3) = 672728.6$$

The total sum of squares is $\sum_{i,j}(y_{ij}-\overline{y}_{++})^2 = 29var(1014,\ldots,670) = 793191.5$. We can find the residual sum of squares by subtraction. The completed table is

Source	Sum of squares	Degrees of freedom	Mean square (ms)	Ratio to residual ms
Blocks	672728.6	4		
Treatments	50548.3	5	10110	2.89
Residual	69914.6	20	3496	
Total	793191.5	29		

(b) Is there any evidence of a difference among the treatments?

Solution: The ratio of the treatment to residual mean square is the discrepancy measure and the significance level is $P(F_{5,20} \geq 2.89) = 0.04$. There is some evidence of differences among the treatments.

The factor levels for the treatments are shown below.

treatment	colour	image
1	red	1
2	red	2
3	blue	1
4	blue	2
5	green	1
6	green	2

(c) Find a 95% confidence interval for the contrast that compares the two images. What do you conclude?

Solution:

From the ANOVA table, we have $\hat{\sigma} = \sqrt{3496} = 59.1$ with 20 degrees of freedom. We are interested in the contrast $\theta = \frac{\tau_2+\tau_4+\tau_6}{3} - \frac{\tau_1+\tau_3+\tau_5}{3}$ with estimate $\hat{\theta} = \frac{839.6+837.6+857.4}{3} - \frac{865.0+806.8+742.8}{3} = 40.0$. The corresponding estimator $\tilde{\theta}$ has standard deviation

$$\sigma\sqrt{\frac{1}{3^25} + \frac{1}{3^25} + \frac{1}{3^25} + \frac{1}{3^25} + \frac{1}{3^25} + \frac{1}{3^25}} = \sigma\sqrt{\frac{2}{15}}.$$

Using t_{20}, the 95% confidence interval for the contrast θ is $40.0 \pm 2.086 \times 59.1 \times \sqrt{\frac{2}{15}}$ or 40.0 ± 45.0. We conclude that the effect of changing images averaged across colours is poorly estimated since the confidence interval is so wide. The effect may or may not be positive.

(d) Is there any evidence that there is a difference in average sales for the two images if the colour is red? Is blue? Is green? What do you conclude?

Solution: Here we are looking at the three contrasts $\tau_2 - \tau_1$, $\tau_4 - \tau_3$, $\tau_6 - \tau_5$. Each contrast is estimated by the difference of the corresponding treatment means and in each case the corresponding estimator has standard deviation $\sigma\sqrt{\frac{2}{5}}$. The estimates, discrepancy measure (two-sided)

and significance level for testing the hypothesis that each contrast is 0 are

Contrast	Estimate	Discrepancy	Sig. Level
$\tau_2 - \tau_1$	-25.4	0.68	0.50
$\tau_4 - \tau_3$	31.0	0.83	0.42
$\tau_6 - \tau_5$	114.6	3.07	0.006

There is evidence that changing the image has a strong effect if the colour is green but not otherwise.

13. I once read a textbook on experimental design that strongly recommended that you should check the significance of the block effects using the ANOVA and an appropriate F test.

    (a) Using the data from Example 1 in this chapter, show how this can be done.

    Solution: From the ANOVA, the task (blocks) mean square is 38.2 and the ratio to the residual mean square is 24.0. Since the model is symmetric in the block and treatment effects, we can argue that if all the block effects are 0, then the block mean square is an estimate of the residual standard deviation σ^2 and that the ratio of the corresponding estimators follows an $F_{5,15}$ distribution. In the example the significance level $P(F_{5,15} \geq 24.0) < 0.01$ so there is strong evidence of a block effect.

    (b) The text then recommended that, if the block effects were significantly different, your investigation was poorly planned and you should start over. What do you think of this advice?

    Solution: This is terrible advice. The blocks are chosen to be different with respect to some explanatory variate that likely affects the response. We want the replication (different blocks) to cover the study population. We expect there to be significant block effects in a well planned experiment.

14. Starting with the randomized block model,

    (a) Show that $\overline{Y}_{i+} \sim G\left(\mu + \tau_i, \frac{\sigma}{\sqrt{b}}\right)$

    Solution:

    (b) Find the standard deviation of the estimator corresponding to the difference of two treatment effects $\tau_1 - \tau_2$.

    Solution: The model is $Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}, R_{ij} \sim G(0, \sigma), i = 1, \ldots, t; j = 1, \ldots, b$.

We have

$$\begin{aligned}\overline{Y}_{i+} &= \frac{Y_{i1}+\cdots+Y_{ib}}{b} \\ &= \frac{\mu+\tau_i+\beta_1+R_{i1}+\cdots+\mu+\tau_i+\beta_b+R_{ib}}{b} \\ &= \mu+\tau_i+\frac{R_{i1}+\cdots+R_{ib}}{b},\end{aligned}$$

since $\sum_j \beta_j = 0$. Hence we have $\overline{Y}_{i+} \sim G\left(\mu+\tau_i,\ \frac{\sigma}{\sqrt{b}}\right)$.
The estimate of $\tau_1-\tau_2$ is $(\overline{y}_{1+}-\overline{y}_{++})-(\overline{y}_{2+}-\overline{y}_{++}) = \overline{y}_{1+}-\overline{y}_{2+}$ so the corresponding estimator $\overline{Y}_{1+}-\overline{Y}_{2+}$ has standard deviation $\sigma\sqrt{\frac{1}{b}+\frac{1}{b}}$.

9.3 Chapter 3 Solutions

1. A large organization is planning a massive retraining of its employees in the use of Office programs. As part of the planning, the HR department considers two factors that might impact the success of the training.

   - Factor 1: use of internal versus external trainers
   - Factor 2: use of interactive computer-aided materials versus non-interactive materials

   To investigate these two factors, a pilot experiment is organized in which 10 employees are randomly assigned to one of the four treatments. Two weeks after the training, each subject in the experiment is given a standard test.

   (a) What does it mean to say that the two factors interact?

   Solution: The factors interact if the effect on test scores of changing from internal to external trainers depends on whether or not computer-aided materials are used

   (b) The HR director pointed out that the cost of the pilot could be substantially reduced if only three treatments were used.

   - Treatment 1: internal trainers, non-interactive materials
   - Treatment 2: internal trainers, interactive materials
   - Treatment 3: external trainers, non-interactive materials

   The effect of the use of interactive materials can be assessed by comparing treatment 2 to treatment 1. The effect of using external trainers can be assessed by comparing treatment 3 to treatment 1. Write a careful explanation to the Director explaining the drawbacks to this suggestion.

Solution: If we use your plan, we will not see if there is interaction between the two factors and we may miss a large opportunity to improve the training program. The average test scores could look like

	Computer aided	**standard**
internal	35	32
external		36

We will not know the average score in the missing cell and it would be a mistake to assume that it is around 39 (i.e. that the effect of changing trainers is the same for both types of materials). We may make the wrong decision unless we use all four treatments in the experiment.

2. The manufacturer of a "frost-free" refrigerator found that in a high humidity, high temperature environment frost did build up inside the fridge. To remedy the problem, a new design was developed and four prototypes were built. In an experimental investigation, the four prototypes and four standard fridges were tested in two environments, one normal and one with high temperature and humidity. Frost build-up was measured after one week's operation. A lower score is better. The data are stored in the file *fac_exercise2.txt* and shown below.

design	**condition**	**response**	**design**	**condition**	**response**
new	extreme	1.35	old	extreme	2.56
new	extreme	1.63	old	extreme	2.51
new	extreme	1.43	old	extreme	2.22
new	extreme	1.57	old	extreme	2.35
new	normal	1.27	old	normal	1.45
new	normal	1.44	old	normal	1.67
new	normal	1.53	old	normal	1.34
new	normal	1.40	old	normal	1.47

Based on the standard model, the estimate of the residual standard deviation is $\hat{\sigma} = 0.133$. The table of treatment averages is shown below.

	New	**old**	**average**
normal	1.41	1.48	1.45
extreme	1.50	2.41	1.95
average	1.45	1.95	1.70

(a) Is there any evidence of differences among the treatments?

Solution: The ANOVA table is

```
Source      DF         SS         MS     ratio    p-value
treatment    3    2.71012    0.90337     50.90      0.000
```

residual	12	0.21297	0.01775
Total	15	2.92309	

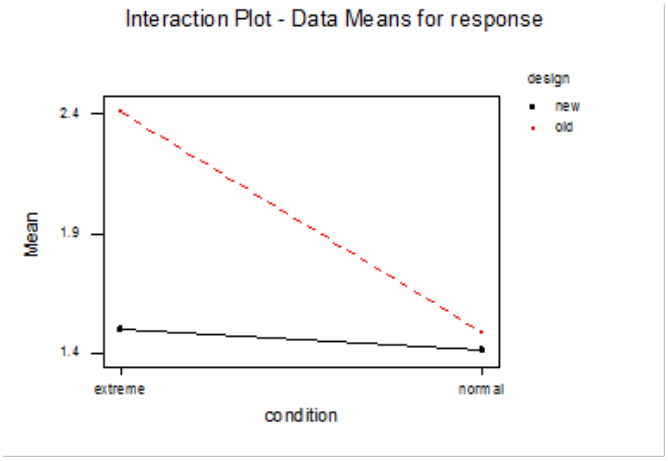
Hence there is very strong evidence of a difference among the treatments.

(b) Is there any evidence of interaction?

Solution: To assess the interaction, consider the contrast $\theta = (\tau_2 - \tau_1) - (\tau_4 - \tau_3)$ with corresponding estimate $\hat{\theta} = (1.48 - 1.41) - (2.41 - 1.50) = -0.84$ and estimator $\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}}\right)$. To test the hypothesis of no interaction i.e. $\theta = 0$, the discrepancy measure is $d = \frac{|\hat{\theta} - 0|}{\hat{\sigma}\sqrt{1}} = 6.31$ so there is very strong evidence of an interaction between the two factors.

(c) Prepare an interaction plot.

Solution: The interaction plot is shown below.



(d) Use the interaction plot to argue that the new design is less sensitive to changes in environmental conditions (temperature and humidity). What limitations apply to this argument?

Solution: For the new design, the average response does not change very much as we change conditions from normal to extreme. If the same behaviour is exhibited for intermediate conditions (interpolation), then the new design is much less insensitive to changing environmental conditions.

(e) Explain why we can not assess the individual effects of temperature and humidity on frost buildup in this investigation.

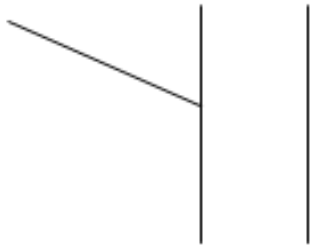
Solution: In the investigation the effect of temperature and humidity are confounded because they both vary together. We consider the only the two conditions:

i. both humidity and temperature extreme

ii. both humidity and temperature normal

The differences (in frost buildup) we see at the two conditions may be due to either or both of temperature and humidity. To look at the effect of temperature and humidity individually we need an experiment with factorial structure in temperature and humidity.

3. In the study of an optical illusion (called the Poggendorf effect), 15 subjects were given the same 9 diagrams in random order. Each diagram was a full page version



of the above picture. The subject was asked to predict where the straight diagonal line would meet the second parallel. The response variate was the distance (in mm) from the actual point of intersection to the predicted point. There were two factors: distance between the parallel lines (3 levels) and the angle of the diagonal (3 levels). The data are given in the file *fac_exercise3.txt* and shown below.

The treatments are labeled 1 to 9 according to the code

Width/angle	40°	30°	20°
4 cm	1	4	7
5 cm	2	5	8
6 cm	3	6	9

Based on the full model, the estimate of the residual standard deviation is 3.95.

				treatment						
subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**average**
1	14.2	7.6	11.0	7.3	9.0	16.4	15.4	19.1	22.2	13.58
2	7.3	10.4	10.4	15.0	12.5	12.9	4.6	12.6	16.4	11.34
3	6.5	13.1	15.7	9.7	14.3	18.8	10.8	20.7	15.1	13.86
4	10.8	9.4	16.5	-1.1	4.8	11.7	14.5	9.7	15.3	10.18
5	6.7	11.3	15.6	8.5	9.4	16.0	13.2	14.3	17.2	12.47
6	4.9	5.4	21.0	14.0	7.9	14.3	13.4	13.8	14.1	12.09
7	10.9	8.0	13.7	13.7	8.8	12.1	8.1	19.3	17.7	12.48
8	12.5	11.6	11.9	14.1	11.9	13.7	9.6	7.6	10.1	11.44
9	7.1	13.6	10.3	11.8	13.4	8.7	6.8	7.2	19.1	10.89
10	6.2	12.8	15.2	13.7	15.0	14.5	15.6	22.6	17.4	14.78
11	17.5	11.9	3.5	12.0	11.2	13.7	8.9	14.4	14.4	11.94
12	6.5	11.6	11.8	11.4	9.8	9.9	10.3	8.2	18.7	10.91
13	19.6	15.3	8.8	6.1	14.1	14.0	17.3	8.4	20.0	13.73
14	11.0	13.3	12.4	12.1	0.8	5.0	8.2	10.6	8.6	9.11
15	17.6	10.1	16.3	3.5	9.6	8.9	13.5	10.5	15.5	11.72
average	10.62	11.03	12.94	10.12	10.17	12.71	11.35	13.27	16.12	12.03

(a) Construct the ANOVA table including the partition of the treatment sum of squares into main effects and interaction components.

Solution: This experimental plan is **blocked**. There are $t = (3)(3) = 9$ treatments and $b = 15$ blocks. Within each block each unit receives each of the nine treatments.

We keep as many decimal places as possible throughout the computations to reduce rounding errors.

The ANOVA table is

Source	Sum of Squares	Degrees of Freedom	Mean Square	Ratio
Treatments	451.8757037	8	56.48446296	
-Interaction	30.29185185	4	7.572962963	0.48524457
-Angle	167.0463704	2	83.52318519	
-Width	254.5374815	2	127.2687407	
Blocks	287.9441481	14		
Residual	1747.926519	112	15.60648677	
Total	2487.74637	134		

(b) Is there any evidence of interaction? Prepare an interaction plot to help interpret your answer.

Solution: From the ANOVA table in part 3a, we have the observed discrepancy measure 0.48524457. Therefore our p-value is

$$p = P(F_{4,112} \geq 0.48524457) = 0.746530219.$$

Therefore we have **no evidence of an interaction**.

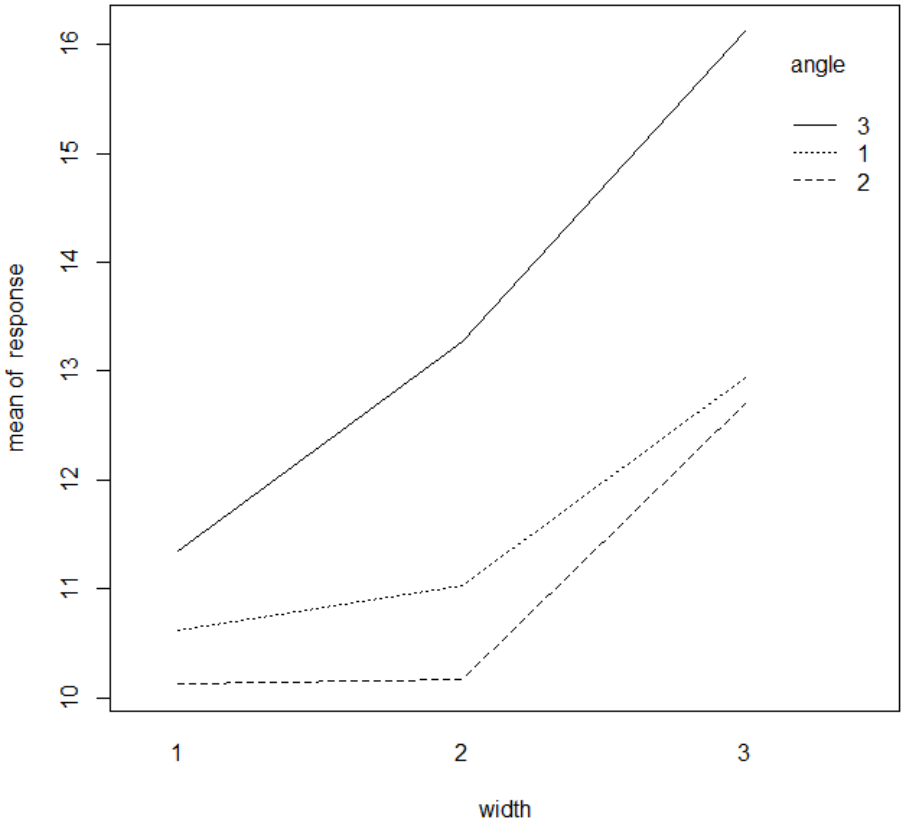
The *R*-code to produce the following interaction plot is stored in the file *fac_exercise3.R* and is reproduced below:

```
a<-read.table('fac_exercise3.txt',header=T)
attach(a)

b<-lm(response~subject+width+angle+width*angle)

interaction.plot(width,angle,response)
```

The interaction plot is:



(c) What can you say about the effects of changing the angle and width of the diagram in light of your answer to part 3b?

Solution: If there is no interaction then we can assess the effects of changing the width (or angle) by comparing the appropriate averages of three treatments. In the following tables, we see how the average response changes with changing angle and width.

angle	n	average response
40°	45	11.52888889
30°	45	10.99777778
20°	45	13.57777778

width	n	average response
4 cm	45	10.69555556
5 cm	45	11.48666667
6 cm	45	13.92222222

4. For Example 2 in Chapter 3, verify algebraically that the treatment sum of squares (without the multiplier 5) can be partitioned as

$$\begin{aligned} & \sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})^2 \\ = \; & 2\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2+3\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2+\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{a++}-\overline{y}_{+b+}+\overline{y}_{+++})^2 \end{aligned}$$

Solution:

$$\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})^2 = 2\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2+3\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2+\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{a++}-\overline{y}_{+b+}+\overline{y}_{+++})^2$$

Note that $i = 1, 2, 3$; $k = 1, 2$. Start from the last term on the right side. We have

$$\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{a++}-\overline{y}_{+b+}+\overline{y}_{+++})^2 = \sum_{ab}[(\overline{y}_{ab+}-\overline{y}_{+++})-(\overline{y}_{a++}-\overline{y}_{+++})-(\overline{y}_{+b+}-\overline{y}_{+++})]^2$$

Now we expand the sum of squares on the right side. Note that there are 6 terms in the expansion that we deal with separately.

The first term is $\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})^2$

The second term is $\sum_{ab}(\overline{y}_{a++}-\overline{y}_{+++})^2 = 2\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2$

The third term is $\sum_{ab}(\overline{y}_{+b+}-\overline{y}_{+++})^2 = 3\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2$

The fourth term is $-2\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})(\overline{y}_{a++}-\overline{y}_{+++}) = -4\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2$

The fifth term is $-2\sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})(\overline{y}_{+b+}-\overline{y}_{+++}) = -6\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2$

The final term is $2\sum_{ab}(\overline{y}_{a++}-\overline{y}_{+++})(\overline{y}_{+b+}-\overline{y}_{+++}) = 2\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++}) = 0$.

Combining the terms we have the result that

$$\begin{aligned} & \sum_{ab}(\overline{y}_{ab+}-\overline{y}_{a++}-\overline{y}_{+b+}+\overline{y}_{+++})^2 \\ = \; & \sum_{ab}(\overline{y}_{ab+}-\overline{y}_{+++})^2-2\sum_{a}(\overline{y}_{a++}-\overline{y}_{+++})^2-3\sum_{b}(\overline{y}_{+b+}-\overline{y}_{+++})^2 \end{aligned}$$

as required.

5. A company that makes products for dentists carried out an experiment with the objective to better understand the factors that affect the pain patients feel after a ceramic inlay is glued to a prepared tooth. The trial involved two factors, the type of glue (formulation A,B,C) and the addition of a short term desensitizer under the inlay. Based on the results of the questionnaire given

to each patient, post-operative pain was assessed on a scale of 1-10. The experiment used blocking in that 12 dentists were involved in the trial and the dentists carried out all six treatments on a set of their patients, randomly assigned to the six treatments. The data are given in the file *fac_exercise5.txt* and are shown below along with some numerical summaries.

	treatment						
dentist	**1**	**2**	**3**	**4**	**5**	**6**	**average**
1	1.0	4.4	2.4	3.9	4.2	3.1	3.17
2	2.2	4.4	4.4	3.5	5.1	3.1	3.78
3	4.2	1.9	2.1	5.3	3.0	3.9	3.40
4	3.0	3.4	3.4	2.8	2.4	4.7	3.28
5	4.7	6.2	5.1	4.3	5.5	6.5	5.38
6	5.4	8.9	5.3	4.8	5.1	4.9	5.73
7	3.1	5.1	2.6	2.4	2.1	2.9	3.03
8	2.5	3.4	4.2	2.3	3.9	2.3	3.10
9	3.5	2.9	1.6	3.0	2.1	2.2	2.55
10	4.6	4.7	4.2	2.6	3.5	3.8	3.90
11	3.9	6.6	3.6	2.4	4.0	4.9	4.23
12	3.5	4.4	3.0	3.6	2.9	5.0	3.73
average	3.47	4.69	3.49	3.41	3.65	3.94	3.77

The first two treatments are brand A, the second two brand B and the third two brand C with the first treatment of each brand having the desensitizer absent. The estimate of the residual standard deviation is $\hat{\sigma} = 1.01$.

(a) Write out a model to describe the repeated application of the Plan. Briefly explain each term in the model.

Solution: This experimental plan is **blocked**. There are $t = (2)(3) = 6$ treatments and $b = 12$ blocks. Within each block one unit receives each of the six treatments.

We keep as many decimal places as possible throughout the computations to reduce rounding errors.

The ANOVA table is

Source	Sum of Squares	Degrees of Freedom	Mean Square	Ratio
Treatments	14.32166667	5	2.864333333	2.838272828
-Interaction	5.446944444	2	2.723472222	2.698693311
-Desensitizer	4.108888889	1	4.108888889	
-Brand	4.765833333	2	2.382916667	
Blocks	59.44833333	11		
Residual	55.505	55	1.009181818	
Total	129.275	71		

(b) Is there any evidence of a difference among the treatments?

Solution: From the ANOVA table in part 5a, we have the observed discrepancy measure 2.838272828. Therefore our p-value is

$$p = P(F_{5,55} \geq 2.838272828) = 0.023765683.$$

Therefore we have **evidence of a difference among the treatments**.

(c) For practice, find a 95% confidence interval for the effect $\tau_1 - \tau_2$.

Solution: From the ANOVA table in part 5a, we have $\hat{\sigma} = \sqrt{1.009181818} = 1.004580419$ with 55 degrees of freedom. The estimate of $\theta = \tau_2 - \tau_1$ is

$$\begin{aligned} \hat{\theta} &= \hat{\tau}_2 - \hat{\tau}_1 \\ &= \overline{y}_{2+} - \overline{y}_{1+} \\ &= 4.691666667 - 3.466666667 \\ &= 1.225. \end{aligned}$$

The corresponding estimator is

$$\tilde{\theta} \sim G\left(\theta, \sigma\sqrt{\frac{1}{12} + \frac{1}{12}}\right) = G\left(\theta, \sigma\sqrt{\frac{1}{6}}\right).$$

From $P(t_{55} \geq c)0.025$, we get $c = 2.004044783$. Hence an approximate 95% confidence interval is

$$\hat{\theta} \pm (2.004044783)\hat{\sigma}\sqrt{\frac{1}{6}} \text{ or } 1.225 \pm 0.821895317.$$

(d) Partition the treatment sum of squares into components for assessing the interaction and the effects of the two factors.

Solution: From the ANOVA table in part 5a, we have the observed discrepancy measure 2.698693311. Therefore our p-value is

$$p = P(F_{2,55} \geq 2.698693311) = 0.076203499.$$

Therefore we have **weak evidence against the null hypothesis of no interaction**.

(e) Is there any evidence of interaction?

Solution: The R-code to produce the following interaction plot is reproduced below:

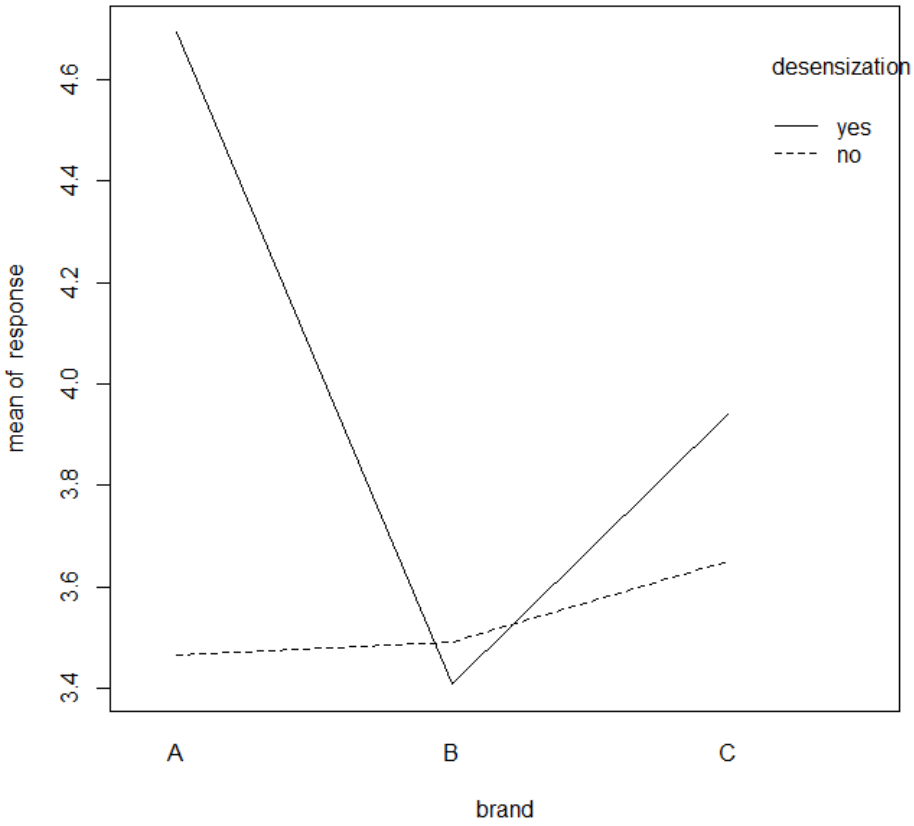
```
a<-read.table('fac_exercise5.txt',header=T)
attach(a)

b<-lm(response~dentist+brand+desensization+brand*desensization)

interaction.plot(brand,desensization,response)
```

(f) Construct an interaction plot.

Solution: The interaction plot is:



(g) Does it make sense to assess the factor effects separately here? Explain.

Solution: Since there is weak evidence of interaction, it is better to interpret the effects of the factors together. From the plot, we see that when the desensitizer is absent, there is little difference among the three glues but that when the desensitizer is present, brand A has a much higher average pain score.

6. Experiments with factorial structure of the treatments are widely used to improve manufacturing processes. For example, in a casting operation, the defect rate in the current process is 5%. To reduce the rate of defects, the process engineer has a large number of process parameters (factors) that he can adjust. To keep matters simple, suppose that there are only two factors, pouring temperature (T) of the iron and the level of an innoculant (I) that is added to the iron as it is poured.

   Engineers are often taught that in such situations, the best strategy is to vary one factor at a time. The purpose of this question is to convince you that this is not the case.

   (a) Suppose that the engineer decides to investigate T at two levels (above and below the current process setting) holding I fixed. He plans an experiment with 16 runs, 8 at each level of temperature. The response variate is the defect rate for a run. Build a model for this plan and write down the estimate of changing T and the standard deviation of the corresponding estimator.

   The engineer then plans to use the best level of T and repeat the plan in part 6a to look at I.

   Solution: The model is $Y_{ij} = \mu + \tau_i + R_{ij},\ R_{ij} \sim G(0, \sigma)\ \ i = 1, 2;\ j = 1, \ldots, 8$. To estimate the effect of changing temperature, we estimate $\theta = \tau_1 - \tau_2$ by $\hat{\theta} = \overline{y}_{1+} - \overline{y}_{2+}$ where the corresponding estimator $\tilde{\theta} = \overline{Y}_{1+} - \overline{Y}_{2+} \sim G(\theta, \sigma\sqrt{\frac{1}{8} + \frac{1}{8}})$ with standard deviation $\frac{\sigma}{2}$.

   Now to estimate the effect of changing the second factor I, conduct another 8 runs at one of the levels of T used in the first investigation (it does not matter which one since there is no interaction). Again we estimate the effect as the difference of averages and the corresponding estimator has standard deviation $\frac{\sigma}{2}$. In total, we use 24 runs to estimate the two effects.

   (b) The statistician recommends a 16 run factorial experiment using all four treatments (4 replicates per treatment). Build a model for this plan. Show that if there is no interaction that this plan is superior for estimating the effects of changing I and T.

   Solution: The model is $Y_{ij} = \mu + \tau_i + R_{ij},\ R_{ij} \sim G(0, \sigma)\ \ i = 1, \ldots, 4;\ j = 1, \ldots, 4$. If there is no interaction, then we can estimate the effect of changing the temperature using the contrast $\theta = \frac{(\tau_1 - \tau_2) - (\tau_3 - \tau_4)}{2}$, the average of the effects for each level of I. The corresponding estimator is $\tilde{\theta} = \frac{(\overline{Y}_{11+} - \overline{Y}_{12+}) - (\overline{Y}_{21+} - \overline{Y}_{22+})}{2} \sim G\left(\theta, \sigma\sqrt{\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16}}\right)$ with standard deviation $\frac{\sigma}{2}$. Similarly, the estimator for the effect of changing innoculant also has standard deviation $\frac{\sigma}{2}$. Hence with 16 runs, we can estimate the effects of both factors simultaneously with

the same precision as with a plan that uses 24 runs and changes the factors one-at-a-time.

(c) Show using a numerical example (a table of averages, for example) that if there is interaction, the plan in a) may not get to the optimal combination of the two factors.

Solution: If there is interaction, then we might get a table of averages as shown below if we first looked at changes in T, picked the best value and then investigated changes in I.

	Temperature level 1	**Temperature level 2**
Innoculantlevel 1	5.2%	4.3%
Innoculantlevel 2		4.7%

With this strategy, we pick temperature level 2 and innoculant level 1 to minimize the scrap rate. But if there is interaction, we cannot predict what happens at the fourth treatment combination. If the rate for this treatment is 3.2% then we would miss this opportunity by varying the factors one-at-a-time.

(d) Briefly summarize why the single factorial plan is better than the two one-at-a-time plans.

Solution: The approach of varying both factors simultaneously is better whether or not interaction is present. If there is no interaction, it is more efficient to vary the factors simultaneously. If there is interaction, we cannot determine that this is so without varying both factors together.

9.4 Chapter 4 Solutions

9.5 Chapter 5 Solutions

1. Consider the sampling protocols defined in Example 1.

(a) Show that the inclusion probability for each unit in the frame is $\frac{1}{100}$ for every protocol.

Solution: For each protocol, the model is uniform. That is, the chance of any possible sample is equal. To find the inclusion probability, we need only count the number of samples that contain a particular unit. Once the unit is in the sample, we count the ways of selecting the remaining 99 units.

SRS: there are $\begin{pmatrix} 9999 \\ 99 \end{pmatrix}$ ways to select the other units so $p_i = \dfrac{\begin{pmatrix} 9999 \\ 99 \end{pmatrix}}{\begin{pmatrix} 10000 \\ 100 \end{pmatrix}} = \frac{1}{100}$.

Stratified sampling: there are $\begin{pmatrix} 999 \\ 9 \end{pmatrix}\begin{pmatrix} 1000 \\ 10 \end{pmatrix}^9$ ways to select the other units so $p_i = \dfrac{\begin{pmatrix} 999 \\ 9 \end{pmatrix}\begin{pmatrix} 1000 \\ 10 \end{pmatrix}^9}{\begin{pmatrix} 1000 \\ 10 \end{pmatrix}^{10}} = \frac{1}{100}$.

Cluster sampling: there are $\begin{pmatrix} 999 \\ 9 \end{pmatrix}$ ways to choose the remaining clusters so $p_i = \dfrac{\begin{pmatrix} 999 \\ 9 \end{pmatrix}}{\begin{pmatrix} 1000 \\ 10 \end{pmatrix}} = \frac{1}{100}$.

Systematic sampling: there is only one way to select the sample so $p_i = \frac{1}{100}$.

Two stage sampling: there are $\begin{pmatrix} 9 \\ 1 \end{pmatrix}$ ways to select the second primary unit. Then the other 99 secondary units can be selected in $\begin{pmatrix} 999 \\ 49 \end{pmatrix}\begin{pmatrix} 1000 \\ 50 \end{pmatrix}$ ways so $p_i = \dfrac{\begin{pmatrix} 9 \\ 1 \end{pmatrix}\begin{pmatrix} 999 \\ 49 \end{pmatrix}\begin{pmatrix} 1000 \\ 50 \end{pmatrix}}{\begin{pmatrix} 10 \\ 2 \end{pmatrix}\begin{pmatrix} 1000 \\ 50 \end{pmatrix}^2} = \frac{1}{100}$.

(b) On a final examination, a student once defined simple random sampling as follows: "simple random sampling is a method of selecting units from a population so that every unit has the same chance of selection". Is this a correct answer?

Solution: No, because there are many sampling protocols that satisfy this definition as shown in part 1a.

(c) Show that the estimator corresponding to the sample average $\hat{\mu} = \frac{\sum_{i \in s} y_i}{n}$ is unbiased for μ for each of the protocols.

Solution: Let

$$I_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}, i = 1, \ldots, N$$

so that $E(I_i) = p_i$. Then we can write

$$\begin{aligned} \tilde{\mu} &= \frac{\sum_{i\in U} y_i I_i}{n} \text{ and} \\ E(\tilde{\mu}) &= \frac{\sum_{i\in U} y_i E(I_i)}{n} \\ &= \frac{\sum_{i\in U} y_i p_i}{n} \\ &= \mu \end{aligned}$$

since $p_i = \frac{n}{N}$ for all five protocols.

2. Consider the estimate $\hat{\sigma} = \sqrt{\frac{\sum_{i\in s}(y_i-\overline{y})^2}{n-1}}$ and the corresponding estimator $\tilde{\sigma}$.

   (a) For SRS, show that $\tilde{\sigma}^2$ is an unbiased estimator for σ^2. [Hint: Use the fact that $\sum_{i\in s}(y_i - \overline{y})^2 = \sum_{i\in s} y_i^2 - n\overline{y}^2$].
   Solution: Using the hint, we can write

$$(n-1)\tilde{\sigma}^2 = \sum_{i\in U} y_i^2 I_i - n\tilde{\mu}^2$$

   where $E(I_i) = \frac{n}{N}$ and

$$\begin{aligned} E(\tilde{\mu}^2) &= Var(\tilde{\mu}) + E(\tilde{\mu})^2 \\ &= (1-f)\frac{\sigma^2}{n} + \mu^2. \end{aligned}$$

   Combining the results we have

$$\begin{aligned} E(\tilde{\sigma}^2) &= \frac{1}{n-1}\left(\sum_{i\in U} y_i^2 \frac{n}{N} - n\left[\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} - \mu^2\right]\right) \\ &= \frac{1}{n-1}\left(\frac{n}{N}\left[\sum_{i\in U} y_i^2 - N\mu^2\right] - \left(1-\frac{n}{N}\right)\sigma^2\right) \\ &= \frac{1}{n-1}\left(\frac{n}{N}(N-1)\sigma^2 - \left(1-\frac{n}{N}\right)\sigma^2\right) \\ &= \sigma^2 \end{aligned}$$

   (b) Is $\tilde{\sigma}$ unbiased for σ?
   Solution: No. $Var(\tilde{\sigma}) = E(\tilde{\sigma}^2) - E(\tilde{\sigma})^2$. So by part 2a, $E(\tilde{\sigma}) = \sqrt{\sigma^2 - Var(\tilde{\sigma})}$
   Unless $n = N$, i.e. the sample consists of the entire population, $Var(\tilde{\sigma}) > 0$. So $E(\tilde{\sigma}) < \sigma$ and the estimator is biased.

3. To estimate the total number of male song sparrows in a 10 km by 10 km square (`http://www.birdsontario.org/atlas/atlasmain.html`) for a breeding bird atlas, a simple random sample of 50 one hectare plots (a hectare is 100m by 100m) is selected. Using a GPS system, your intrepid instructor visits each of the selected plots (after dawn but before 9:00 am between May 24 and July 6) and counts the number of singing male song sparrows detected in a 10 minute period. The data are summarized below.

# of sparrows	0	1	2	3	4
# of plots	28	13	5	3	1

(a) Find a 95% confidence interval for τ, the total number of male song sparrows in the square.

Solution: Observe that there are $100^2 = 10000$ plots of size 100 m^2 in a square of size (10 km)^2. Thus, our frame for this problem is $U = \{1, \ldots, 10000\}$.

Define

- Y_i = number of song sparrows in plot $i, i = 1, \ldots, 10000$.
- μ = average number of song sparrows in a plot.
- σ = standard deviation of each Y_i.

From the sample data, we have

$$\begin{aligned}\hat{\mu} &= \frac{(28)(0)+(13)(1)+(5)(2)+(3)(3)+(1)(4)}{50}\\ &= \frac{36}{50}\\ &= 0.72, \text{ and}\\ \hat{\sigma} &= \sqrt{\frac{28(0-0.72)^2+13(1-0.72)^2+5(2-0.72)^2+3(3-0.72)^2+1(4-0.72)^2}{50-1}}\\ &= 1.011.\end{aligned}$$

For 95% confidence with a $G(0, 10)$ distribution, we choose $c = 1.96$. Therefore an approximate 95% confidence interval for μ is

$$\begin{aligned}&\hat{\mu} \pm c\sqrt{1-\frac{n}{N}}\frac{\hat{\sigma}}{\sqrt{n}}\\ =\ & 0.72 \pm (1.96)\sqrt{1-\frac{50}{10000}}\frac{(1.011)}{\sqrt{50}}\\ =\ & 0.72 \pm 0.28.\end{aligned}$$

As $\mu = \frac{\tau}{10000} \Leftrightarrow \tau = 10000\mu$, therefore an approximate 95% confidence interval for τ is

$$10000(0.72 \pm 0.28) = 7200 \pm 2800.$$

(b) Suppose that I wanted to estimate the total number of male song sparrows to within 1000 with 95% confidence. How many additional plots are needed?

Solution: We want an approximate 95% confidence interval for τ with half-width 1000. As $\mu = \frac{\tau}{10000}$, this corresponds with an approximate 95% confidence interval for μ with half-width $\frac{1000}{10000} = 0.1$.

From now on, work with the approximate 95% confidence interval for μ, to keep the numbers more manageable. Then we have

$$\begin{aligned} \ell &= 0.1, \text{ as above,} \\ c &= 1.96, \text{ as before, and} \\ \hat{\sigma} &= 1.011 \text{ so that, applying the formula from class gives} \\ n &= \left(\frac{1}{N} + \frac{\ell^2}{c^2\hat{\sigma}^2}\right)^{-1} \\ &= \left(\frac{1}{10000} + \frac{(0.1)^2}{(1.96)^2(1.011)^2}\right)^{-1} \\ &= 378. \end{aligned}$$

We need 378 plots in total. As we have already sampled 50 plots, we therefore need an additional 328 plots.

4. Suppose we want to estimate a population average so that the relative precision is specified. That is, we want to find the sample size required (SRS) so that the length of the confidence interval 2ℓ divided by the sample average is pre-determined.

(a) For a given confidence level and required precision p%, find a formula for the required sample size.

Solution: In general, the length of a confidence interval for μ is $2c\sqrt{\frac{1}{n} - \frac{1}{N}}\,\hat{\sigma}$. We want to find the sample size n so that

$$\frac{2c\sqrt{\frac{1}{n} - \frac{1}{N}}\hat{\sigma}}{\hat{\mu}} = \frac{p}{100}.$$

Solving for n we have the ugly formula

$$n = \frac{1}{\frac{1}{N} + \left(\frac{p\hat{\mu}}{200c\hat{\sigma}}\right)^2}.$$

(b) What knowledge of the population attributes do we need to make this formula usable?

Solution: We need an estimate of the so-called coefficient of variation $\frac{\sigma}{\mu}$.

5. One cheap (but poor) way to check the quality of a batch of items is called **acceptance sampling**. Suppose that there are $N = 1000$ items in a shipment and you cannot tolerate more than 1% defective (your first mistake – why should you tolerate any defective items from your supplier). You decide to select and inspect a sample of 20 items and accept the shipment if you find zero defectives. If you find 1 or more defective items, you inspect the complete shipment.

(a) How would you select the sample?

Solution: It would be nice to use SRS but it is likely too expensive unless the items are already numbered and it is easy to locate an idea with a specified label. Usually haphazard (for small items) or systematic sampling is used in this context.

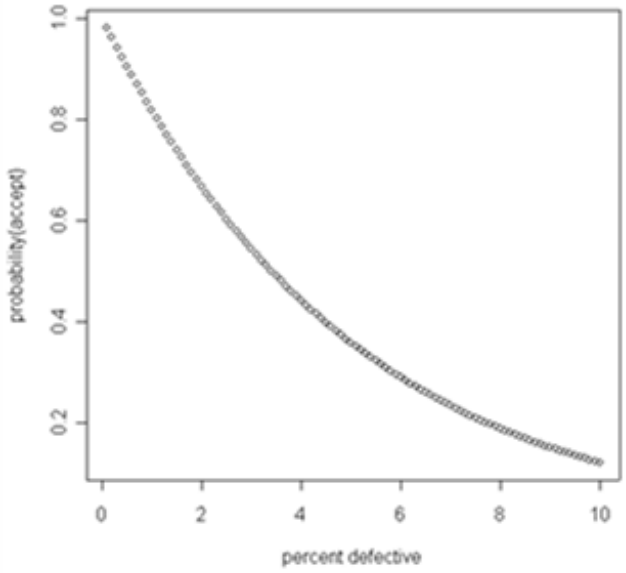
(b) Calculate the probability $p(\pi)$ that you accept the shipment as a function of π, the percentage of defective items in the shipment.

Solution: Since we are sampling such a small fraction of the shipment, we can approximate the number of defective items in the sample by a binomial random variable with $n = 20$ and the probability of a defective item π. Then we have

$$P(\text{ accept shipment}) = (1-\pi)^{20}$$

(c) Graph $p(\pi)$ for $0 \leq \pi \leq 10\%$

Solution: Using R, we have the following graph.



(d) Given the results in part 5c, you decide to increase the sample size so that there is only a 5% chance of accepting a shipment with 1% defective. What sample size do you recommend?

Solution: Suppose the sample size is n. Assuming that we can use the binomial approximation, we have $P($ accept shipment$)=(1\text{-}\pi)^n$. We want to find n so that $(1-0.01)^n = 0.05$ so $n = 298$. This is so large that the binomial approximation may breakdown and, on a practical basis, is completely unreasonable. I recommend you tell your supplier to ensure that there are no defective items in the shipments. Sampling inspection is not useful here.

9.6 Chapter 6 Solutions

1. Consider the sampling protocols defined in Example 1. Find the quadratic expansion of $f(x, y) = \frac{y}{x}$ about the point $(\mu(x), \mu(y))$ to estimate the bias

in the estimator $\tilde{\theta} = \frac{\tilde{\mu}(y)}{\tilde{\mu}(x)}$. Note that the general form of the expansion is

$$\begin{aligned} f(x,y) &\approx f(x_0,y_0) + \frac{\partial f(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial f(x_0,y_0)}{\partial y}(y-y_0) \\ &+ \frac{\partial^2 f(x_0,y_0)}{\partial x^2}\frac{(x-x_0)^2}{2} \\ &+ \frac{\partial^2 f(x_0,y_0)}{\partial x \partial y}(x-x_0)(y-y_0) \\ &+ \frac{\partial^2 f(x_0,y_0)}{\partial y^2}\frac{(y-y_0)^2}{2} \end{aligned}$$

This quadratic function has the same value, first and second derivatives at the point (x_0, y_0) as does $f(x, y)$. You can easily check this statement by differentiating the right side of the expression.

Solution: To use the expansion, we have

$$\begin{aligned} \frac{\partial f}{\partial x} &= -\frac{\mu(y)}{\mu(x)^2} \\ \frac{\partial f}{\partial y} &= \frac{1}{\mu(x)} \\ \frac{\partial^2 f}{\partial x^2} &= \frac{2\mu(y)}{\mu(x)^3} \\ \frac{\partial^2 f}{\partial x \partial y} &= -\frac{1}{\mu(x)^2} \\ \frac{\partial^2 f}{\partial y^2} &= 0, \end{aligned}$$

so we can write

$$\begin{aligned} &\tilde{\theta} \\ \approx\ &\theta - \frac{\mu(y)}{\mu(x)^2}\left[\tilde{\mu}(x) - \mu(x)\right] \\ &+ \frac{1}{\mu(x)}\left[\tilde{\mu}(y) - \mu(y)\right] \\ &+ \frac{2\mu(y)}{\mu(x)^3}\frac{\left[\tilde{\mu}(x) - \mu(x)\right]^2}{2} - \\ &\frac{1}{\mu(x)^2}\left[\tilde{\mu}(x) - \mu(x)\right]\left[\tilde{\mu}(y) - \mu(y)\right], \end{aligned}$$

and

$$\begin{aligned} E(\tilde{\theta}) &\approx \theta + \frac{\mu(y)}{\mu(x)^3} Var(\tilde{\mu}(x)) - \frac{1}{\mu(x)^2} Cov(\tilde{\mu}(x), \tilde{\mu}(y)) \\ &= \theta + \frac{1}{\mu(x)^2}[\theta\, Var(\tilde{\mu}(x)) - Cov(\tilde{\mu}(x), \tilde{\mu}(y)]. \end{aligned}$$

The approximate bias is given by the second term. We know $Var(\tilde{\mu}(x)) = (1-f)\frac{\sigma^2(x)}{n}$ and with a bit of effort, we can show $Cov(\tilde{\mu}(x), \tilde{\mu}(y)) = (1-f)\frac{Cov(x,y)}{n}$ where $Cov(x,y)$ is the population covariance. The key point is to notice that the bias has a factor $\frac{1}{n}$ and will be small if the sample size is large.

2. Chapter 6, Exercise 2

In order to count the number of small items in a large container, a shipping company selects a sample of 25 items and weighs them. They then weigh the whole shipment (excluding the container). Assume that there is small error in weighing and act as if SRS is used (it is not, the sampling is haphazard). Let y_i be the mass of the i^{th} item in the population and let the total known mass be τ.

(a) Show that an estimate of the population size is

$$\hat{N} = \frac{\tau}{\sum_{i \in s} \frac{y_i}{25}}.$$

Solution: Note that

$$\begin{aligned} \tau &= N\mu(y) \\ \Leftrightarrow N &= \frac{\tau}{\mu(y)} \end{aligned} \tag{9.1}$$

so we can construct an estimate of N using our knowledge of estimating $\mu(y)$. The sample average and population average should be close, so we have the estimate

$$\hat{N} = \frac{\tau}{\hat{\mu}(y)} = \frac{\tau}{\sum_{i \in s} \frac{y_i}{25}}, \text{ with corresponding estimator} \tag{9.2}$$

$$\tilde{N} = \frac{\tau}{\tilde{\mu}(y)} \tag{9.3}$$

(b) Find the (approximate) mean and standard deviation of the corresponding estimator $\tilde{N}$. You can ignore the unknown sampling fraction $f = \frac{n}{N}$.

Solution: Because of the shape of line (9.3), we consider expanding the function $f(y) = \frac{1}{y}$. The linear approximation about $\mu(y)$ is

$$\frac{1}{y} \approx \frac{1}{\mu(y)} - \frac{1}{\mu(y)^2}(y - \mu(y))$$

and hence we have

$$\begin{aligned}
\frac{1}{\tilde{\mu}(y)} &\approx \frac{1}{\mu(y)} - \frac{1}{\mu(y)^2}(\tilde{\mu}(y) - \mu(y)) \text{ and so} \\
E\left[\frac{1}{\tilde{\mu}(y)}\right] &\approx \frac{1}{\mu(y)} \\
Var\left[\frac{1}{\tilde{\mu}(y)}\right] &\approx \left(\frac{1}{\mu(y)^4}\right) Var(\tilde{\mu}(y)).
\end{aligned}$$

Since $\tilde{N} = \frac{\tau}{\tilde{\mu}(y)}$, therefore we have

$$\begin{aligned}
E(\tilde{N}) &= E\left[\frac{\tau}{\tilde{\mu}(y)}\right] \\
&= \tau E\left[\frac{1}{\tilde{\mu}(y)}\right] \\
&\approx \frac{\tau}{\mu(y)} \\
&= N, \text{ and} \\
Var(\tilde{N}) &= Var\left[\frac{\tau}{\tilde{\mu}(y)}\right] \\
&= \tau^2 Var\left[\frac{1}{\tilde{\mu}(y)}\right] \\
&= \left(\frac{\tau^2}{\mu(y)^4}\right) Var(\tilde{\mu}(y)) \\
&= \left(\frac{\tau}{\mu(y)}\right)^2 \left(\frac{1}{\mu(y)^2}\right) Var(\tilde{\mu}(y)) \\
&\approx \left(\frac{N^2}{\mu(y)^2}\right) Var(\tilde{\mu}(y)) \\
&= \left(\frac{N^2}{\mu(y)^2}\right) \left(\frac{1-f}{n}\right) \sigma(y)^2.
\end{aligned}$$

(c) In the example, the sample average mass is 75.45 g, the sample standard deviation is 0.163 g and the total mass is 154.2 kg. Find a 95% confidence interval for the total number of items in the container.

Solution: To find the confidence interval we have

$$\tilde{N} \sim G\left(N,\ N\sqrt{\frac{1-f}{n}}\frac{\sigma(y)}{\mu(y)}\right), \text{ approximately.}$$

Note that the mean and standard deviation both depend on the unknown N which is different from the usual situation. Instead we work

with $\frac{\tilde{N}}{N} \sim G\left(1,\ \sqrt{\frac{1-f}{n}}\frac{\sigma(y)}{\mu(y)}\right)$ and hence we want

$$P\left(\left|\frac{\frac{\tilde{N}}{N-1}}{\sqrt{\frac{1-f}{n}}\frac{\sigma(y)}{\mu(y)}}\right| \leq 1.96\right) = 0.95.$$

Re-arranging the inequality and substituting $\hat{N}$, $\hat{\mu}(y)$, $\hat{\sigma}(y)$ for $\tilde{N}$, $\mu(y)$, $\sigma(y)$, we get the confidence interval

$$\left(\frac{\hat{N}}{1+(1.96)\sqrt{\frac{1-f}{n}}\frac{\hat{\sigma}(y)}{\hat{\mu}(y)}}, \frac{\hat{N}}{1-(1.96)\sqrt{\frac{1-f}{n}}\frac{\hat{\sigma}(y)}{\hat{\mu}(y)}}\right)$$

In the example (recording everything in kg), we have

$$\begin{aligned}
\hat{\mu}(y) &= 0.07545 \\
\tau &= 154.2 \\
\hat{N} &= \frac{\tau}{\hat{\mu}(y)} \\
&= \frac{154.2}{0.07545} \\
&= 2044 \\
\hat{\sigma}(y) &= 0.000163
\end{aligned}$$

and the 95% confidence interval is (2035, 2053).

3. Briefly describe when you would use the ratio or regression estimate instead of the sample average to estimate the population average.

   Solution: We need an explanatory variate with known population average that can be measured on each unit in the sample. If the response variate is approximately proportional to the explanatory variate, then the ratio estimate is more precise than the sample average. If the response variate is approximately linear in the explanatory variate, then the regression estimate is more precise.

4. Many bird species have specialized habitat. We can exploit this knowledge when we are trying to estimate population totals or density. For example, wood thrush are a forest dwelling bird that live in the hardwood forests of eastern North America. Suppose we wanted to estimate the number of wood thrush pairs nesting within the region of Waterloo, an area of highly fragmented forest patches. Using aerial photography, we know that there are 1783 such patches (minimum size 3 ha) with an average size 13.4 ha. A simple random sample of 50 woodlots is selected and the number of nesting

pairs y_i is counted in each woodlot by counting the number of singing males. The area x_i of each sampled woodlot is also recorded. The data are available in the file *thrush.txt*. Find a 95% confidence intervals for the total number of thrushes based on the

(a) sample average $\overline{y}$

Solution: We need the following summaries of the data:

$$\begin{aligned}
\hat{\mu}(x) &= 11.72 \\
\hat{\mu}(y) &= 1.62 \\
\hat{\theta} &= \frac{\hat{\mu}(y)}{\hat{\mu}(x)} \\
&= 0.138 \\
\hat{\alpha} &= \hat{\mu}(y) \\
&= 1.62 \\
\hat{\beta} &= 0.228 \\
\hat{\sigma}(y) &= 1.34 \\
\frac{\sum_{i\in s}(y_i - \hat{\theta}\, x_i)^2}{n-1} &= 0.399 \\
\frac{\sum_{i\in s}[y_i - \hat{\alpha} - \hat{\beta}(\, x_i - \hat{\mu}(x))]^2}{n-1} &= 0.142
\end{aligned}$$

The sample average is 1.62 with associated estimated standard deviation $\sqrt{\frac{1-f}{n}\hat{\sigma}(y)^2} = 0.187$ so an approximate 95% confidence interval for the average number of thrushes per woodlot is 1.62 ± 0.37 and the interval for the total number of thrushes is $1783(1.62 \pm 0.37) = 2888 \pm 653$.

(b) ratio estimate

Solution: The ratio estimate is $\hat{\theta}\mu(x) = 1.85$ with associated estimated standard deviation $\sqrt{\frac{1-f}{n}\frac{\sum_{i\in s}(y_i - \hat{\theta}\, x_i)^2}{n-1}} = 0.088$ so an approximate 95% confidence interval for the population average is 1.85 ± 0.17 and for the population total is $1783(1.85 \pm 0.17) = 3299 \pm 308$.

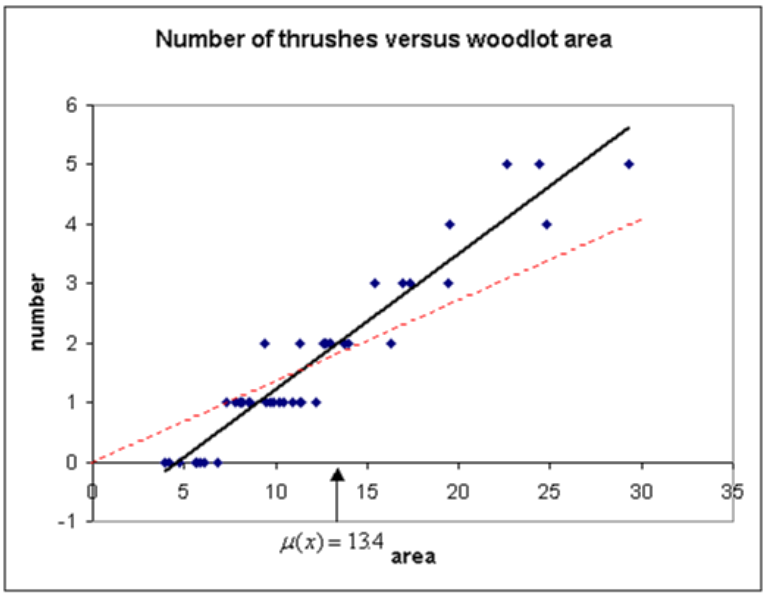
(c) regression estimate

Solution: The regression estimate is $\hat{\mu}(y) + \hat{\beta}(\mu(x) - \hat{\mu}(x)) = 2.00$ with associated estimated standard deviation

$$\sqrt{\frac{1-f}{n}\frac{\sum_{i\in s}\left[y_i - \hat{\alpha} - \hat{\beta}\,(x_i - \hat{\mu}(x))\right]^2}{n-1}} = 0.053.$$

so an approximate 95% confidence interval for the population average is 2.00 ± 0.10 and for the population total is $1783(2.00 \pm 0.10) = 3566 \pm 184$

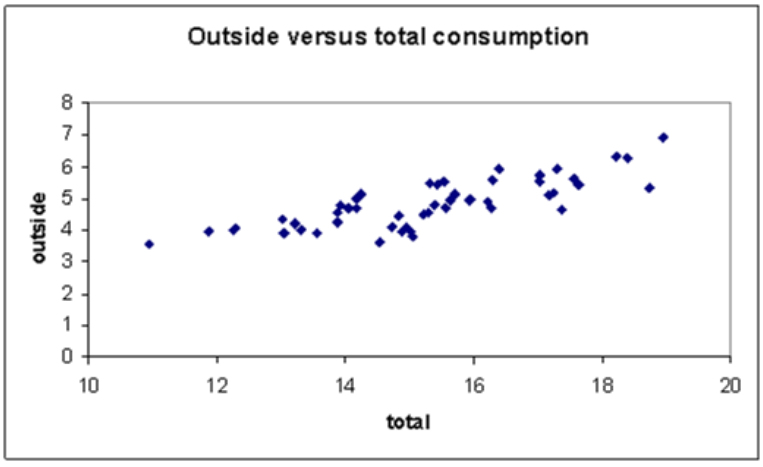
The following scatterplot shows the fitted regression line (solid), the "fitted" line through the origin (dotted) and $\mu(x) = 13.4$.



5. The City of Waterloo wants to estimate the average amount of water per house $\mu(y)$ that is used to water lawns and gardens in the month of July. A SRS of 50 houses is selected and special metering units are installed to measure the volume of water y from external taps. The total volume of water x is measured by the regular meter. From water records, it is known that the average total water consumption per house is $\mu(x) = 15.6$ cubic metres. The data are stored in the file *water.txt*.

   (a) Prepare a scatterplot of y versus x.

   Solution:



(b) Estimate $\mu(y)$ using the sample average, the ratio estimate and the regression estimate.

Solution: We need to calculate the sample average for y and x, and the sample standard deviation for y. We also need to find the residual sum of squares if we fit a line of the form $y = \hat{\theta}\, x$ with $\hat{\theta} = \frac{\hat{\mu}(y)}{\hat{\mu}(x)}$ and the least squares line $y = \hat{\mu}(y) + \hat{\beta}\,(x - \hat{\mu}(x))$. These calculations can be done in excel – note that the fitted regression line can be found by adding a trend line to the scatterplot and asking for the equation. We get

$$\begin{aligned} \hat{\mu}(x) &= 15.293 \\ \hat{\mu}(y) &= 4.819 \\ \hat{\sigma}(y) &= 0.761 \\ \hat{\theta} &= 0.315 \\ \hat{\beta} &= 0.331 \end{aligned}$$

The three estimates are:

$$\begin{aligned} \hat{\mu}(y) &= 4.189 \\ \hat{\mu}(y)_{ratio} &= \hat{\theta}\mu(x) \\ &= 4.914 \\ \hat{\mu}(y)_{reg} &= \hat{\mu}(y) + \hat{\beta}\,[\mu(x) - \hat{\mu}(x)] \\ &= 4.920. \end{aligned}$$

We also have the estimated residual mean square for the ratio and regression fit - 0.215 (ratio) and 0.214 (regression).

(c) Find 95% confidence intervals based on each estimate.

Solution: The 95% confidence intervals have the form: estimate $\pm$ $1.96\frac{\hat{\sigma}}{\sqrt{n}}$ where $\hat{\sigma}$ is calculated from the appropriate residual mean square and we ignore the fpc. The three intervals are 4.82 ± 0.21, 4.91 ± 0.13, 4.92 ± 0.13.

(d) Which estimation procedure is preferable here? Why?

Solution: Here the ratio and regression estimates are more precise because we exploit the relationship between y and x to get the estimates. The residual variation after fitting the models is much less than the overall variation

9.7 Chapter 7 Solutions

1. In many surveys, there is interest in estimating strata averages or differences in strata averages.

(a) In general, for SRS, write down the distribution for the estimators $\tilde{\mu}_h$ and $\tilde{\mu}_h - \tilde{\mu}_k$.

Solution: Assuming relatively large sample sizes within the strata we have approximately

$$\begin{aligned} \tilde{\mu}_h &\sim G\left(\mu_h, (1-f_h)^{\frac{1}{2}} \frac{\sigma_h}{\sqrt{n_h}}\right) \\ \tilde{\mu}_h - \tilde{\mu}_k &\sim G\left(\mu_h - \mu_k, \sqrt{\frac{(1-f_h)\sigma_h^2}{n_h} + \frac{(1-f_k)\sigma_k^2}{n_k}}\right) \end{aligned}$$

(b) In the well survey, find a 95% confidence interval for the proportion of wells in farms with animals that are contaminated.

Solution: The estimate of the proportion contaminated is $\hat{\pi}_1 = .172$ with associated estimated standard deviation $\sqrt{\frac{1-f}{n}\hat{\pi}_1(1-\hat{\pi}_1)} = 0.030$ so the 95% confidence interval is 0.172 ± 0.058.

(c) In the well survey, find a 95% confidence interval for the average Na difference between the two types of farm wells.

Solution: For farms with animals, we have $\hat{\mu}_1 = 237.3$ with associated estimated standard deviation $\sqrt{\frac{(1-f_1)\hat{\sigma}_1^2}{n_1}} = 3.275$.

For farms without animals, we have $\hat{\mu}_2 = 245.6$ with associated estimated standard deviation $\sqrt{\frac{(1-f_2)\hat{\sigma}_2^2}{n_2}} = 3.614$ and hence we have $\hat{\mu}_2 - \hat{\mu}_1 = 8.30$ with associated estimated standard deviation $\sqrt{3.275^2 + 3.614^2} = 4.877$. Hence a 95% confidence interval for $\mu_2 - \mu_1$ is 8.30 ± 9.60. There is no evidence of a difference in average Na levels between the two groups of farms.

2. Suppose that the purpose of the survey is to estimate a population proportion π. If there are H strata,

(a) Write down the stratified estimate of π and the variance of the corresponding estimator.

Solution: Since

$$\pi = W_1\pi_1 + \cdots + W_H\pi_H,$$

we have

$$\hat{\pi} = W_1\hat{\pi}_1 + \cdots + W_H\hat{\pi}_H,$$

and

$$\begin{aligned} & Var(\tilde{\pi}_{strat}) \\ = \; & W_1^2 Var(\tilde{\pi}_1) + \cdots + W_H^2 Var(\tilde{\pi}_H) \\ = \; & W_1^2 \left(\frac{1-f_1}{n_1}\right)\left(\frac{n_1}{n_1-1}\right)\pi_1(1-\pi_1) + \cdots + W_H^2 \left(\frac{1-f_H}{n_H}\right)\left(\frac{n_H}{n_H-1}\right)\pi_H(1-\pi_H) \\ \approx \; & W_1^2 \left(\frac{1-f_1}{n_1}\right)\pi_1(1-\pi_1) + \cdots + W_H^2 \left(\frac{1-f_H}{n_H}\right)\pi_H(1-\pi_H), \end{aligned} \tag{9.4}$$

ignoring the $\left(\frac{n_h}{n_h-1}\right)$ factors.

(b) What is the variance of $\tilde{\pi}_{strat}$ for proportional allocation?

Solution: For proportional allocation, take $W_h = \frac{n_h}{n} \Leftrightarrow n_h = nW_h$. Then the expression on line (9.4) becomes

$$\begin{aligned} & W_1^2 \left(\frac{1-f_1}{nW_1}\right)\pi_1(1-\pi_1) + \cdots + W_H^2 \left(\frac{1-f_H}{nW_H}\right)\pi_H(1-\pi_H) \\ = \; & \frac{1}{n}\left[W_1\left(1-\frac{n_1}{N_1}\right)\pi_1(1-\pi_1) + \cdots + W_H\left(1-\frac{n_H}{N_H}\right)\pi_H(1-\pi_H)\right] \\ = \; & \frac{1}{n}\left[W_1\left(1-\frac{n\left(\frac{N_1}{N}\right)}{N_1}\right)\pi_1(1-\pi_1) + \cdots + W_H\left(1-\frac{n\left(\frac{N_H}{N}\right)}{N_H}\right)\pi_H(1-\pi_H)\right] \\ = \; & \frac{1}{n}\left[W_1\left(1-\frac{n}{N}\right)\pi_1(1-\pi_1) + \cdots + W_H\left(1-\frac{n}{N}\right)\pi_H(1-\pi_H)\right] \\ = \; & \frac{1}{n}\left[W_1\left(1-f\right)\pi_1(1-\pi_1) + \cdots + W_H\left(1-f\right)\pi_H(1-\pi_H)\right], \text{ letting } f = \frac{n}{N} \\ = \; & \left(\frac{1-f}{n}\right)\left[W_1\pi_1(1-\pi_1) + \cdots + W_H\pi_H(1-\pi_H).\right] \end{aligned} \tag{9.5}$$

(c) How should the strata be formed so that the stratified sampling protocol is superior to SRS?

Solution: Choose the strata with π_h close to 0 or close to 1.

3. Suppose the well survey was to be re-done with the same overall sample size 500. How would you recommend allocating the sample to the strata if

(a) Estimating the average Na level was the primary goal.

Solution: For optimal allocation, we have n_h is proportional to $W_h\sigma_h$. If we assume that the standard deviations do not change markedly, we use the estimates from the current survey to allocate the sample. We have

stratum	Weight	St Dev
1	0.177	41.45
2	0.097	37.62
3	0.726	51.23

so the optimal sample sizes are A: 76, 38 and 386. more weight is given to stratum three because it is larger and has higher estimated standard deviation.

(b) Estimating the proportion of contaminated wells was the primary goal.

Solution: For optimal allocation, we have n_h is proportional to $W_h\sqrt{\pi_h(1-\pi_h)}$ and we use the current estimates to get

stratum	Weight	sd
1	0.177	0.37738
2	0.097	0.317811
3	0.726	0.338491

So the optimal sample sizes are: 97, 45 and 358.

(c) For each case, compare the predicted standard deviations of $\tilde{\mu}_{strat}$ and $\tilde{\pi}_{strat}$ to what occurred in the current survey.

Solution: We have

$$\begin{aligned} Var(\tilde{\mu}_{strat}) &= W_1^2\frac{(1-f_1)}{n_1}\sigma_1^2 + \cdots + W_H^2\frac{(1-f_H)}{n_H}\sigma_H^2 \\ Var(\tilde{\pi}_{strat}) &= W_1^2\frac{(1-f_1)}{n_1}\pi_1(1-\pi_1) + \cdots + W_H^2\frac{(1-f_H)}{n_H}\pi_H(1-\pi_H) \end{aligned}$$

Using the current estimates and the two new allocations, we get the estimated standard deviations

allocation	**current**	**A**	**B**
$\tilde{\mu}_{strat}$	2.42	2.11	2.13
$\tilde{\pi}_{strat}$	0.016	0.015	0.015

The estimator of the proportion is much less sensitive to changes in the allocation.

4. Consider the difference of the variances of $\tilde{\mu}_{strat}$ under proportional and optimal allocation for a sample of size n. Ignore the fpc.

(a) Show that this difference can be written as $\left(\frac{1}{n}\right)\sum_h(\sigma_h-\overline{\sigma})^2W_h$ where $\overline{\sigma}=\sum_h\sigma_hW_h$ is the weighted average standard deviation over the H strata.

Solution: From the course notes, we have

$$Var_{prop}(\tilde{\mu}_{strat}) = \frac{1}{n}\sum_{h=1}^{H} W_h\sigma_h^2, \text{ and} \tag{9.6}$$

$$Var_{opt}(\tilde{\mu}_{strat}) = \frac{1}{n}\left(\sum_{g=1}^{H} W_g\sigma_g\right)^2. \tag{9.7}$$

Then subtracting line (9.7) from line (9.6) gives

$$\begin{aligned}
& Var_{prop}(\tilde{\mu}_{strat}) - Var_{opt}(\tilde{\mu}_{strat}) \\
= & \frac{1}{n}\left[\sum_{h=1}^{H} W_h\sigma_h^2 - \left(\sum_{g=1}^{H} W_g\sigma_g\right)^2\right] \\
= & \frac{1}{n}\left[\sum_{h=1}^{H} W_h\sigma_h^2 - 2\left(\sum_{g=1}^{H} W_g\sigma_g\right)^2 + \left(\sum_{g=1}^{H} W_g\sigma_g\right)^2\right] \\
= & \frac{1}{n}\left[\sum_{h=1}^{H} W_h\sigma_h^2 - 2A\left(\sum_{h=1}^{H} W_h\sigma_h\right) + A^2\underbrace{\sum_{h=1}^{H} W_h}_{=1}\right], \text{ letting } A = \sum_{g=1}^{H} W_g\sigma_g = \overline{\sigma} \\
= & \frac{1}{n}\sum_{h=1}^{H} W_h\left(\sigma_h^2 - 2\sigma_h A + A^2\right) \\
= & \frac{1}{n}\sum_{h=1}^{H} W_h\left(\sigma_h - A\right)^2 \\
= & \frac{1}{n}\sum_{h=1}^{H} W_h\left(\sigma_h - \overline{\sigma}\right)^2, \qquad (9.8)
\end{aligned}$$

as required, so we are finished.

(b) When will the gain be large with optimal allocation relative to proportional allocation?

Solution: The gain will be largest when the standard deviations vary widely.

Note that, because the expression on line (9.8) is non-negative for any legal choice of the parameters involved, we will always have that

$$\begin{aligned}
Var_{prop}(\tilde{\mu}_{strat}) - Var_{opt}(\tilde{\mu}_{strat}) &\geq 0 \\
\Leftrightarrow Var_{prop}(\tilde{\mu}_{strat}) &\geq Var_{opt}(\tilde{\mu}_{strat}), \qquad (9.9)
\end{aligned}$$

which confirms our choice of optimal allocation as giving the smaller variance in this comparison.

5. In an informal sample of math students at UW, 100 people were asked their opinion (on a 5 point scale) about the core courses and their value. One particular statement was (with scores):

> "All mathematics students are required to take Stat 231." strongly agree – 1 ; agree – 2 ; neutral – 3 ; disagree – 4 ; strongly disagree – 5

The sample results, broken down by year are shown below. Estimate the average score for all math students and find an approximate 95% confidence interval for the population average. Note that SRS was not used here so were are making assumptions about the estimators that may be unwarranted. There are about 3300 students in the faculty.

Year	Sample size	Population weight	Average score	Standard deviation
1	39	0.31	2.8	1.22
2	23	0.24	3.5	1.09
3	26	0.23	3.2	1.03
4	12	0.22	3.1	0.87

Solution: We can estimate the average score as if we had stratified the sampling beforehand.

$$\hat{\mu}_{post} = (0.31)(2.8) + (0.24)(3.5) + (0.23)(3.2) + (0.22)(3.1) = 3.126.$$

Multiplying the population weights by $N = 3300$ yields these stratum sizes:

Year (h)	$N_h = \text{weight} \times N$
1	$(0.31)(3300) = 1023$
2	$(0.24)(3300) = 792$
3	$(0.23)(3300) = 759$
4	$(0.22)(3300) = 726$

The approximate estimated variance of $\tilde{\mu}_{post}$ is

$$\begin{aligned}
&\left(\frac{1-f_1}{n_1}\right) W_1^2\hat{\sigma}_1^2 + \left(\frac{1-f_2}{n_2}\right) W_2^2\hat{\sigma}_2^2 \\
&+ \left(\frac{1-f_3}{n_3}\right) W_3^2\hat{\sigma}_3^2 + \left(\frac{1-f_4}{n_4}\right) W_4^2\hat{\sigma}_4^2 \\
= \;& \left(\frac{1-\frac{39}{1023}}{39}\right)(0.31)^2(1.22)^2 + \left(\frac{1-\frac{23}{792}}{23}\right)(0.24)^2(1.09)^2 \\
&+ \left(\frac{1-\frac{26}{759}}{26}\right)(0.23)^2(1.03)^2 + \left(\frac{1-\frac{12}{729}}{12}\right)(0.22)^2(0.87)^2 \\
= \;& 0.011503916,
\end{aligned}$$

and hence the estimated standard error is

$$\sqrt{0.011503916} = 0.107256309 \approx 0.107.$$

The approximate 95% confidence interval is $3.13 \pm (1.96)(0.107)$ or 3.13 ± 0.21.

9.8 Chapter 8 Solutions

1. Drawing on the similarities between the AAPOR formulas, what is the formula for RR5 and why?

   Solution: Use the formula for RR6 and remove the P in the numerator. All persons/households of unknown eligibility are counted as ineligible. So, $RR5 = \frac{I}{(I+P)+(R+NC+O)}$

2. The survey report for the Parent and Teens survey contains the following comment: "The UW SRC considers respondents that appear incompetent or to have a significant language barrier to be not eligible for the survey (under 4.0) whereas AAPOR designates these records as eligible, non-interview (under 2.0)". What would be the response rates RR1, RR4, RR6 if these were considered eligible?

   Solution: The response rate would go down because the number of non-responding eligibles is increased. From the table, there are 100 such households increasing the number of ineligibles to 223. This also changes the percentage of eligible among the households with known eligibility:

$$\frac{\text{known eligibles}}{\text{known eligibles and ineligibles}} = \frac{(534+223)}{(534+223)+2172-504} = 31.2\%$$

$$e(U) = 0.312 * 424 \approx 132$$

Minimal response rate: $RR1 = \frac{534}{534+223+424} = 45.2\%$

Response rate with estimated eligibles: $RR4 = \frac{534}{534+223+132} = 60.1\%$

Maximal response rate (cheating): $RR6 = \frac{534}{534+223} = 70.5\%$

t-table (right tail)

For each row (degrees of freedom k) and column (right tail probability α), the table entry e satisfies $\Pr(t_k \geq e) = \alpha$. Note that the t-distribution is symmetric about 0.

degrees of freedom	right tail probability 0.25	0.10	0.05	0.025	0.01
1	1.000	3.078	6.314	12.706	31.821
2	0.816	1.886	2.920	4.303	6.965
3	0.765	1.638	2.353	3.182	4.541
4	0.741	1.533	2.132	2.776	3.747
5	0.727	1.476	2.015	2.571	3.365
6	0.718	1.440	1.943	2.447	3.143
7	0.711	1.415	1.895	2.365	2.998
8	0.706	1.397	1.860	2.306	2.896
9	0.703	1.383	1.833	2.262	2.821
10	0.700	1.372	1.812	2.228	2.764
11	0.697	1.363	1.796	2.201	2.718
12	0.695	1.356	1.782	2.179	2.681
13	0.694	1.350	1.771	2.160	2.650
14	0.692	1.345	1.761	2.145	2.624
15	0.691	1.341	1.753	2.131	2.602
16	0.690	1.337	1.746	2.120	2.583
17	0.689	1.333	1.740	2.110	2.567
18	0.688	1.330	1.734	2.101	2.552
19	0.688	1.328	1.729	2.093	2.539
20	0.687	1.325	1.725	2.086	2.528
21	0.686	1.323	1.721	2.080	2.518
22	0.686	1.321	1.717	2.074	2.508
23	0.685	1.319	1.714	2.069	2.500
24	0.685	1.318	1.711	2.064	2.492
25	0.684	1.316	1.708	2.060	2.485
26	0.684	1.315	1.706	2.056	2.479
27	0.684	1.314	1.703	2.052	2.473
28	0.683	1.313	1.701	2.048	2.467
29	0.683	1.311	1.699	2.045	2.462
30	0.683	1.310	1.697	2.042	2.457
35	0.682	1.306	1.690	2.030	2.438
40	0.681	1.303	1.684	2.021	2.423
45	0.680	1.301	1.679	2.014	2.412
50	0.679	1.299	1.676	2.009	2.403
gaussian	0.675	1.282	1.646	1.962	2.330

F-table (right tail) $\alpha = 0.10$

For each row (denominator degrees of freedom k) and column (numerator degrees of freedom j), the table entry e satisfies $P(F(j,k) \geq e) = \alpha$.

denominator degrees of freedom	numerator degrees of freedom											
	1	2	3	4	5	6	7	8	9	10	20	30
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.74	62.26
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.44	9.46
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.18	5.17
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.84	3.82
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.21	3.17
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.84	2.80
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.59	2.56
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.42	2.38
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.30	2.25
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.20	2.16
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.12	2.08
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.06	2.01
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.01	1.96
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	1.96	1.91
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.92	1.87
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.89	1.84
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.86	1.81
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.84	1.78
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.81	1.76
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.79	1.74
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.78	1.72
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.76	1.70
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.74	1.69
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.73	1.67
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.72	1.66
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.67	1.61
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.61	1.54
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.57	1.50
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.49	1.42

F-table (right tail) $\alpha = 0.05$

For each row (denominator degrees of freedom k) and column (numerator degrees of freedom j), the table entry e satisfies $P(F(j,k) \geq e) = \alpha$.

denominator degrees of freedom	numerator degrees of freedom											
	1	2	3	4	5	6	7	8	9	10	20	30
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	248.02	250.10
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.65	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.54	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.46	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.39	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.33	2.25
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.28	2.19
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.23	2.15
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.19	2.11
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.16	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.10	2.01
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.07	1.98
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.05	1.96
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.03	1.94
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.01	1.92
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.74
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.78	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.68	1.57

F-table (right tail) $\alpha = 0.01$

For each row (denominator degrees of freedom k) and column (numerator degrees of freedom j), the table entry e satisfies $P(F(j,k) \geq e) = \alpha$.

denominator degrees of freedom	numerator degrees of freedom											
	1	2	3	4	5	6	7	8	9	10	20	30
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6209	6260
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.45	99.47
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	26.69	26.50
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.02	13.84
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.55	9.38
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.40	7.23
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.16	5.99
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.36	5.20
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.81	4.65
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.41	4.25
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.10	3.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	3.86	3.70
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.66	3.51
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.51	3.35
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.37	3.21
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.26	3.10
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.16	3.00
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.08	2.92
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.00	2.84
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	2.94	2.78
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	2.88	2.72
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.83	2.67
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	2.78	2.62
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.74	2.58
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.70	2.54
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.55	2.39
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.37	2.20
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.27	2.10
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.07	1.89