# Statistics 502 Lecture Notes

Peter D. Hoff

# Contents

# List of Figures

# Chapter 1

# Principles of experimental design

## 1.1 Induction

Much of our scientific knowledge about processes and systems is based on *induction*: reasoning from the specific to the general.

**Example (survey):** Do you favor increasing the gas tax for public transportation?

- Specific cases: 200 people called for a telephone survey
- Inferential goal: get information on the opinion of the **entire** city.

**Example (Women's Health Initiative):** Does hormone replacement improve health status in post-menopausal women?

- Specific cases: Health status monitored in 16,608 women over a 5-year period. Some took hormones, others did not.
- Inferential goal : Determine if hormones improve the health of women not in the study.

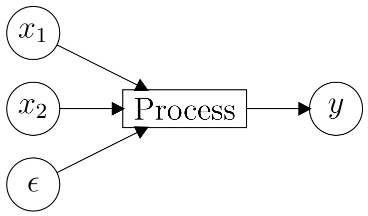


Figure 1.1: Model of a variable process

## 1.2 Model of a process or system

We are interested in how the *inputs* of a process affect an *output*. Input variables consist of

**controllable factors** $x_1$: measured and determined by scientist.

**uncontrollable factors** $x_2$: measured but not determined by scientist.

**noise factors** $\epsilon$: unmeasured, uncontrolled factors, often called experimental variability or "error".

For any interesting process, there are inputs such that:

$$\text{variability in input} \rightarrow \text{variability in output}$$

If variability in an input factor $x$ leads to variability in output $y$, we say $x$ is a *source of variation*. In this class we will discuss methods of designing and analyzing experiments to determine important sources of variation.

## 1.3 Experiments and observational studies

Information on how *inputs* affect *output* can be gained from:

- Observational studies: Input and output variables are observed from a pre-existing population. It may be hard to say what is input and what is output.

- Controlled experiments: One or more input variables are controlled and manipulated by the experimenter to determine their *effect* on the output.

**Example (Women's Health Initiative, WHI):**

- Population: Healthy, post-menopausal women in the U.S.
- Input variables:
  1. estrogen treatment, yes/no
  2. demographic variables (age, race, diet, etc.)
  3. unmeasured variables (?)
- Output variables:
  1. coronary heart disease (eg. MI)
  2. invasive breast cancer
  3. other health related outcomes
- Scientific question: How does estrogen treatment affect health outcomes?

**Observational Study:**

1. **Observational population:** 93,676 women enlisted starting in 1991, tracked over eight years on average. Data consists of $x$= input variables, $y$=health outcomes, gathered concurrently on existing populations.
2. **Results:** good health/low rates of CHD generally associated with estrogen treatment.
3. **Conclusion:** Estrogen treatment is positively associated with health outcomes, such as prevalence of CHD.

**Experimental Study (WHI randomized controlled trial):**

1. **Experimental population:**

   373,092 women determined to be eligible

   $\hookrightarrow$ 18,845 provided consent to be in experiment

   $\hookrightarrow$ 16,608 included in the experiment

$$\text{16,608 women randomized to either} \begin{cases} x = 1 & \text{(estrogen treatment)} \\ x = 0 & \text{(no estrogen treatment)} \end{cases}$$

Women were of different ages and were treated at different clinics. Women were *blocked* together by age and clinic, and then treatments were randomly assigned within each age×treatment *block*. This type of random allocation is called a *randomized block design*.

| | | age group | | |
|---|---|---|---|---|
| | | 1 (50-59) | 2 (60-69) | 3 (70-79) |
| clinic | 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ |

$$\begin{aligned} n_{i,j} &= \text{\# of women in study, in clinic } i \text{ and in age group } j \\ &= \text{\# of women in } \textit{block } i, j \end{aligned}$$

**Randomization scheme:** For each block, 50% of the women in that block were randomly assigned to treatment ($x = 1$) and the remaining assigned to control ($x = 0$).

**Question:** Why did they *randomize* within a block?

2. **Results:** JAMA, July 17 2002. Also see the NLHBI press release . Women on treatment had a **higher incidence rate** of

   - CHD
   - breast cancer
   - stroke
   - pulmonary embolism

   and a **lower incidence rate** of

   - colorectal cancer
   - hip fracture

**3. Conclusion:** Estrogen treatment is not a viable preventative measure for CHD in the general population. That is, our *inductive inference* is

**(specific)** higher CHD rate in treatment population than control

**suggests**

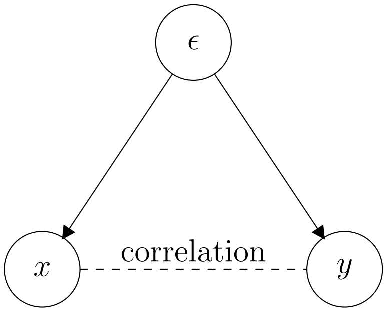
**(general)** if everyone in the population were treated, the incidence rate of CHD would increase.

**Question:** Why the different conclusions between the two studies? Consider the following possible explanation: Let

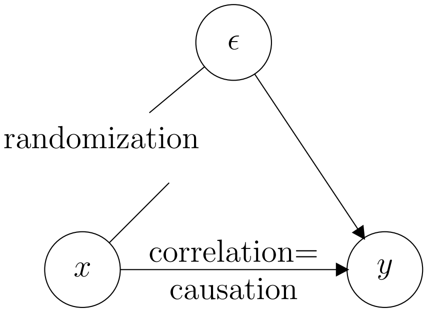
$x =$ estrogen treatment

$\epsilon =$ "health consciousness" (not directly measured)

$y =$ health outcomes



Association between $x$ and $y$ may be due to an unmeasured variable $\epsilon$.



Randomization breaks the association between $\epsilon$ and $x$.

> Observational studies can suggest good experiments to run, but can't definitively show *causation*.
>
> Randomization can eliminate correlation between $x$ and $y$ due to a different cause $\epsilon$, aka a *confounder*.

"No causation without randomization"

## 1.4 Steps in designing an experiment

1. Identify research hypotheses to be tested.

2. Choose a set of *experimental units*, which are the units to which treatments will be randomized.

3. Choose a *response/output* variable.

4. Determine potential *sources of variation* in response:

   (a) factors of interest

   (b) nuisance factors

5. Decide which variables to measure and control:

   (a) treatment variables

   (b) potential large sources of variation in the units (blocking variables)

6. Decide on the experimental procedure and how treatments are to be randomly assigned.

The order of these steps may vary due to constraints such as budgets, ethics, time, etc..

### Three principles of Experimental Design

**1.** *Replication*: Repetition of an experiment.

Replicates are runs of an experiment or sets of experimental units that have the same values of the control variables.

More replication → more precise inference

Let

$y_{A,i}$ = response of the $i$th unit assigned to treatment $A$

$y_{B,i}$ = response of the $i$th unit assigned to treatment $B$

$i = 1, \ldots, n$.

Then $\bar{y}_A \neq \bar{y}_B$ provides evidence that treatment affects response, i.e. treatment is a source of variation, with the amount of evidence increasing with $n$.

**2.** *Randomization*: Random assignment of treatments to experimental units. This removes potential for systematic bias on the part of the researcher, and removes any preexperimental source of bias. Makes confounding the effect of treatment with an unobserved variable unlikely (but not impossible).

**3.** *Blocking*: Randomization within blocks of homogeneous experimental units. The goal is to evenly distribute treatments across large potential sources of variation.

### Example (Crop study):

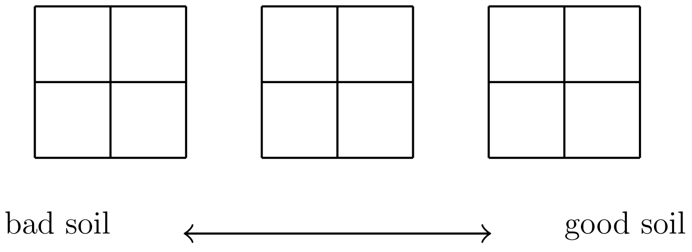
Hypothesis: Tomato type ($A$ versus $B$) affects tomato yield.

Experimental units: three plots of land, each to be divided into a $2 \times 2$ grid.

Outcome: Tomato yield.

Factor of interest: Tomato type, $A$ or $B$.

Nuisance factor : Soil quality.



Questions for discussion:

- What are the benefits of this design?
- What other designs might work?
- What other designs wouldn't work?
- Should the plots be divided up further? If so, how should treatments then be assigned?

# Chapter 2

# Test statistics and randomization distributions

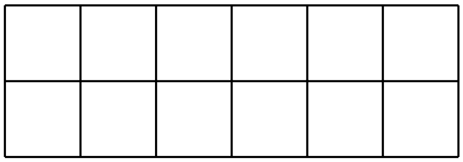
**Example: Wheat yield**

Question: Is one fertilizer better than another, in terms of yield?

Outcome variable: Wheat yield.

Factor of interest: Fertilizer type, $A$ or $B$. One *factor* having two *levels*.

Experimental material: One plot of land to be divided into 2 rows of 6 subplots each.



**1. Design question:** How should we assign treatments/factor levels to the plots? We want to avoid confounding a treatment effect with another potential source of variation.

**2. Potential sources of variation:** Fertilizer, soil, sun, water, etc.

**3. Implementation:** If we assign treatments **randomly**, we can avoid any pre-experimental bias in results: 12 playing cards, 6 red, 6 black were shuffled and dealt:

$$\begin{array}{rcl} \text{1st card black} & \rightarrow & \text{1st plot gets } B \\ \text{2nd card red} & \rightarrow & \text{2nd plot gets } A \\ \text{3rd card black} & \rightarrow & \text{3rd plot gets } B \\ & \vdots & \end{array}$$

This is the first design we will study, a *completely randomized design.*

**4. Results:**

| | | | | | |
|---|---|---|---|---|---|
| $B$<br>26.9 | $A$<br>11.4 | $B$<br>26.6 | $A$<br>23.7 | $B$<br>25.3 | $B$<br>28.5 |
| $B$<br>14.2 | $A$<br>17.9 | $A$<br>16.5 | $A$<br>21.1 | $B$<br>24.3 | $A$<br>19.6 |

How much evidence is there that fertilizer type is a source of yield variation? Evidence about differences between two populations is generally measured by comparing summary statistics across the two sample populations. (Recall, a *statistic* is any computable function of known, observed data).

## 2.1 Summaries of sample populations

**Distribution:**

- Empirical distribution: $\hat{\Pr}(a, b] = \#(a < y_i \leq b)/n$
- Empirical CDF (cumulative distribution function)

$$\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$$

- Histograms
- Kernel density estimates

Note that these summaries more or less retain all the information in the data except the unit labels.

**Location:**

- sample mean or average : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- sample median : A/the value $y_{.5}$ such that

  $$\#(y_i \leq y_{.5})/n \geq 1/2 \qquad \#(y_i \geq y_{.5})/n \geq 1/2$$

  To find the median, sort the data in increasing order, and call these values $y_{(1)}, \ldots, y_{(n)}$. If there are no ties, then

  if $n$ is odd, then $y_{(\frac{n+1}{2})}$ is the median;

  if $n$ is even, then all numbers between $y_{(\frac{n}{2})}$ and $y_{(\frac{n}{2}+1)}$ are medians.

**Scale:**

- sample variance and standard deviation:

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s = \sqrt{s^2}$$

- interquantile range:

  $[y_{.25}, y_{.75}]$ (interquartile range)

  $[y_{.025}, y_{.975}]$ (95% interval)

**Example: Wheat yield**

All of these sample summaries are easily obtained in R:

```
> yA<-c(11.4, 23.7, 17.9, 16.5, 21.1, 19.6)
> yB<-c(26.9, 26.6, 25.3, 28.5, 14.2, 24.3)


> mean(yA)
[1] 18.36667
> mean(yB)
[1] 24.3

> median(yA)
[1] 18.75
> median(yB)
[1] 25.95
```

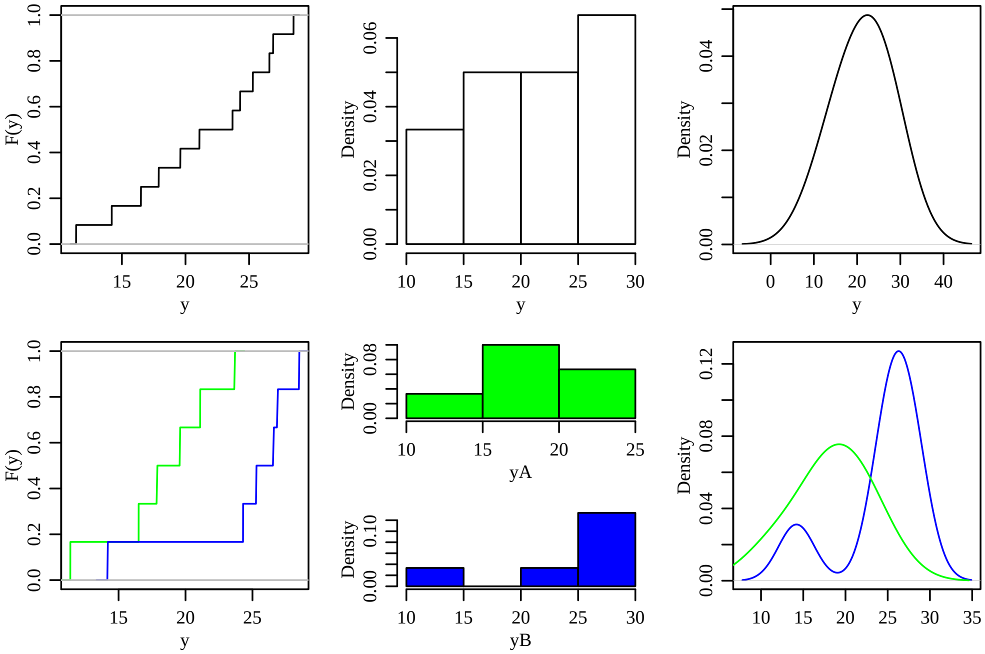


Figure 2.1: Wheat yield distributions

```
> sd(yA)
[1] 4.234934
> sd(yB)
[1] 5.151699

> quantile(yA,prob=c(.25,.75))
   25%    75%
16.850 20.725
> quantile(yB,prob=c(.25,.75))
   25%    75%
24.550 26.825
```

So there is a difference in yield for **these** wheat fields.
Would you recommend $B$ over $A$ for future plantings?
Do you think these results generalize to a **larger population**?

## 2.2 Hypothesis testing via randomization

**Questions:**

- Could the observed differences be due to fertilizer type?
- Could the observed differences be due to plot-to-plot variation?

**Hypothesis tests:**

- $\mathrm{H}_0$ (null hypothesis): Fertilizer type does not affect yield.
- $\mathrm{H}_1$ (alternative hypothesis): Fertilizer type does affect yield.

A *statistical hypothesis test* evaluates the plausibility of $\mathrm{H}_0$ in light of the data.

Suppose we are interested in mean wheat yields. We can evaluate $\mathrm{H}_0$ by answering the following questions:

- Is a mean difference of 5.93 plausible/probable if $\mathrm{H}_0$ is true?
- Is a mean difference of 5.93 large compared to experimental noise?

To answer the above, we need to compare

$\{|\bar{y}_B - \bar{y}_A| = 5.93\}$, the **observed difference** in the experiment to

values of $|\bar{y}_B - \bar{y}_A|$ that **could have been observed if** $\mathrm{H}_0$ **were true**.

Hypothetical values of $|\bar{y}_B - \bar{y}_A|$ that could have been observed under $\mathrm{H}_0$ are referred to as samples from the *null distribution*.

**Finding a null distribution:** Let

$$g(\mathbf{Y}_A, \mathbf{Y}_B) = g(\{Y_{1,A}, \ldots, Y_{6,A}\}, \{Y_{1,B}, \ldots, Y_{6,B}\}) = |\bar{Y}_B - \bar{Y}_A|.$$

This is a function of the outcome of the experiment. It is a *statistic*. Since we will use it to perform a hypothesis test, we will call it a *test statistic*.

Observed test statistic: $g(11.4, 23.7, \ldots, 14.2, 24.3) = 5.93 = g_{\mathrm{obs}}$

Hypothesis testing procedure: Compare $g_{\mathrm{obs}}$ to $g(\mathbf{Y_A}, \mathbf{Y_B})$ for values of $\mathbf{Y_A}$ and $\mathbf{Y_B}$ that **could have** been observed, if $\mathrm{H}_0$ were true.

Recall the design of the experiment:

1. Cards were shuffled and dealt $B, R, B, R, \ldots$ and fertilizer types planted in subplots:

| | | | | | |
|---|---|---|---|---|---|
| $B$ | $A$ | $B$ | $A$ | $B$ | $B$ |
| $B$ | $A$ | $A$ | $A$ | $B$ | $A$ |

2. Crops were grown and wheat yields obtained:

| | | | | | |
|---|---|---|---|---|---|
| $B$<br>26.9 | $A$<br>11.4 | $B$<br>26.6 | $A$<br>23.7 | $B$<br>25.3 | $B$<br>28.5 |
| $B$<br>14.2 | $A$<br>17.9 | $A$<br>16.5 | $A$<br>21.1 | $B$<br>24.3 | $A$<br>19.6 |

Now imagine **re-doing** the experiment in a universe where "$\mathrm{H}_0$: no treatment effect" is true:

1. Cards are shuffled and dealt $B, R, B, B, \ldots$ and wheat types planted in subplots:

| | | | | | |
|---|---|---|---|---|---|
| $B$ | $A$ | $B$ | $B$ | $A$ | $A$ |
| $A$ | $B$ | $B$ | $A$ | $A$ | $B$ |

2. Crops are grown and wheat yields obtained:

| | | | | | |
|---|---|---|---|---|---|
| $B$<br>26.9 | $A$<br>11.4 | $B$<br>26.6 | $B$<br>23.7 | $A$<br>25.3 | $A$<br>28.5 |
| $A$<br>14.2 | $B$<br>17.9 | $B$<br>16.5 | $A$<br>21.1 | $A$<br>24.3 | $B$<br>19.6 |

Under this hypothetical treatment assignment,

$$\begin{aligned} (\mathbf{Y}_A, \mathbf{Y}_B) &= \{11.4, 25.3, \ldots, 21.1, 19.6\} \\ |\bar{Y}_B - \bar{Y}_A| &= 1.07 \end{aligned}$$

This represents an outcome of the experiment in a universe where

- The treatment assignment is $B, A, B, B, A, A, A, B, B, A, A, B$;
- $\text{H}_0$ is true.

**IDEA:** To consider what types of outcomes we would see in universes where $\text{H}_0$ is true, compute $g(\mathbf{Y}_A, \mathbf{Y}_B)$ under every possible treatment assignment and assuming $\text{H}_0$ is true.

Under our randomization scheme, there were

$$\frac{12!}{6!6!} = \binom{12}{6} = 924$$

**equally likely** ways the treatments could have been assigned. For each one of these, we can calculate the value of the test statistic that would've been observed under $\text{H}_0$:

$$\{g_1, g_2, \ldots, g_{924}\}$$

This enumerates **all potential** pre-randomization outcomes of our test statistic, assuming **no treatment effect**. Along with the fact that each treatment

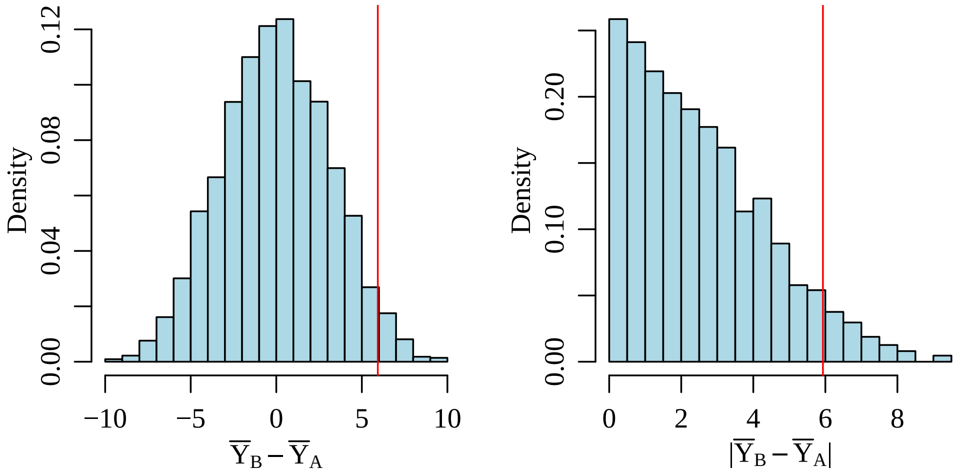


Figure 2.2: Approximate randomization distribution for the wheat example

assignment is equally likely, these value give a *null distribution*, a probability distribution of possible experimental results, if $\mathrm{H}_0$ is true.

$$F(x|\mathrm{H}_0) = \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \leq x|\mathrm{H}_0) = \frac{\#\{g_k \leq x\}}{924}$$

This distribution is sometimes called the **randomization distribution**, because it is obtained by the randomization scheme of the experiment.

**Comparing data to the null distribution:**

Is there any contradiction between $\mathrm{H}_0$ and our data?

$$\Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq 5.93|\mathrm{H}_0) = 0.056$$

According to this calculation, the probability of observing a mean difference of 5.93 or more is unlikely under the null hypothesis. This probability calculation is called a *p-value*. Generically, a *p*-value is

"The probability, under the null hypothesis, of obtaining a result as or more extreme than the observed result."

**The basic idea:**

$$\begin{array}{rcl} \text{small } p\text{-value} & \rightarrow & \text{evidence against H}_0 \\ \text{large } p\text{-value} & \rightarrow & \text{no evidence against H}_0 \end{array}$$

**Approximating a randomization distribution:**
We don't want to have to enumerate all $\binom{n_A+n_B}{n_A}$ possible treatment assignments. Instead, repeat the following $S$ times for some large number $S$:

(a) randomly simulate a treatment assignment from the population of possible treatment assignments, under the randomization scheme.

(b) compute the value of the test statistic, given the simulated treatment assignment and under $\mathrm{H}_0$.

The *empirical distribution* of $\{g_1, \ldots, g_S\}$ **approximates** the null distribution:

$$\frac{\#(g_s \geq g_{\text{obs}})}{S} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq g_{\text{obs}} | \mathrm{H}_0)$$

The approximation improves if $S$ is increased.

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c(”B”, ”A”, ”B”, ”A”, ”B”, ”B”, ”B”, ”A”, ”A”, ”A”, ”B”, ”A”)

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim==”B”]) - mean(y[xsim==”A”] ) )
}
```

## 2.3 Essential nature of a hypothesis test

Given $\mathrm{H}_0$, $\mathrm{H}_1$ and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant *test statistic* $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between $\mathrm{H}_0$ and $\mathrm{H}_1$ in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a *null distribution* : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under $\mathrm{H}_0$. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that **could have happened under** $\mathrm{H}_0$.

3. Compute the *p-value*: The probability under $\mathrm{H}_0$ of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y})|\mathrm{H}_0)$$

If the $p$-value is small ⇒ evidence against $\mathrm{H}_0$

If the $p$-value is large ⇒ not evidence against $\mathrm{H}_0$

Even if we follow these guidelines, we must be careful in our specification of $\mathrm{H}_0$, $\mathrm{H}_1$ and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

**Questions:**

- Is a small $p$-value evidence in favor of $\mathrm{H}_1$?
- Is a large $p$-value evidence in favor of $\mathrm{H}_0$?
- What does the $p$-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out.

## 2.4 Sensitivity to the alternative hypothesis

In the previous section we said that the test statistic $g(\mathbf{y})$ should be able to "differentiate between $\mathrm{H}_0$ and $\mathrm{H}_1$ in ways that are scientifically relevant." What does this mean?

Suppose our data consist of samples $\mathbf{y}_A$ and $\mathbf{y}_B$ from two populations $A$ and $B$. Previously we used $g(\mathbf{y}_A, \mathbf{y}_B) = |\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$. Let's consider two different test statistics:

**$t$-statistic:**

$$\begin{aligned} g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\ s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2 \end{aligned}$$

This is a scaled version of our previous test statistic, in which we compare the difference in sample means to a pooled version of the sample standard deviation and the sample size. Note that this statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in $n_A$ and $n_B$;
- decreasing in $s_p$.

A more complete motivation for using this statistic will be given in the next chapter.

**Kolmogorov-Smirnov statistic:**

$$g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = \max_{y \in \mathbb{R}} |\hat{F}_B(y) - \hat{F}_A(y)|$$

This is just the size of the largest gap between the two sample CDFs.

**Comparing the test statistics:**

Suppose we perform a CRD and obtain samples $\mathbf{y}_A$ and $\mathbf{y}_B$ like those in Figure 2.3. For these data,

- $n_A = n_B = 40$
- $\bar{y}_A = 10.05$, $\bar{y}_B = 9.70$.
- $s_A = 0.87$, $s_B = 2.07$

The main difference between the two samples seems to be in their variances and not in their means. Now let's consider evaluating

$$\text{H}_0\text{: treatment does not affect response}$$

using our two new test statistics. We can approximate the null distributions of $g_t(\mathbf{Y}_A, \mathbf{Y}_B)$ and $g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B)$ by randomly reassigning the treatments but leaving the responses fixed:

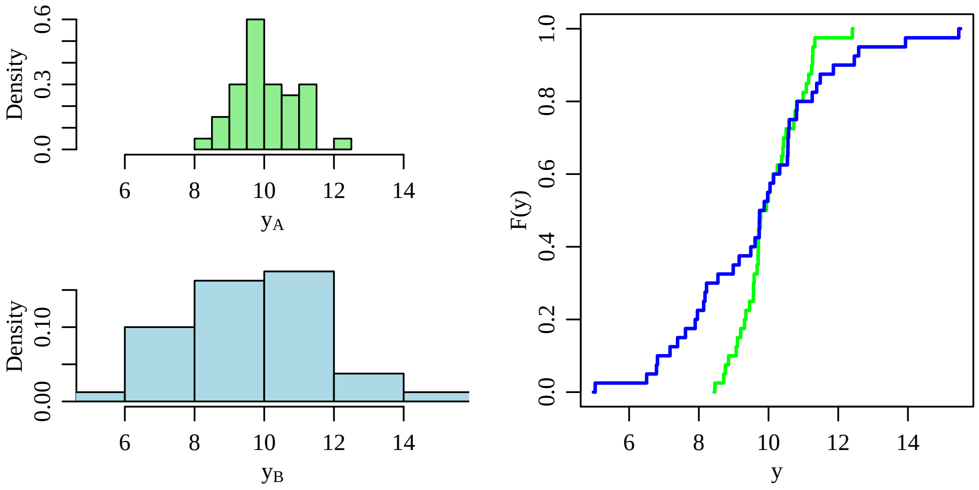


Figure 2.3: Histograms and empirical CDFs of the first two hypothetical samples.

```
Gsim<−NULL
for(s in 1:5000)
{
  xsim<−sample(x)
  yAsim<−y[xsim==”A”] ; yBsim<−y[xsim==”B”]
  g1<− g.tstat(yAsim,yBsim)
  g2<− g.ks(yAsim,yBsim)
  Gsim<−rbind(Gsim,c(g1,g2))
}
```

These calculations give:

$$t\text{-statistic}: \quad g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.00 \ , \ \Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.00) = 0.321$$

$$\text{KS-statistic:} \ g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.30 \ , \ \Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.30) = 0.043$$

The hypothesis test based on the $t$-statistic does not indicate strong evidence against $\text{H}_0$, whereas the test based on the KS-statistic does. The reason is that the $t$-statistic is **only sensitive to differences in means**. In particular, if $\bar{y}_A = \bar{y}_B$ then the $t$-statistic is zero, its minimum value. In contrast, the KS-statistic is **sensitive to any differences in the sample distributions**. Now let's consider a second dataset, shown in Figure 2.5, for which

- $n_A = n_B = 40$

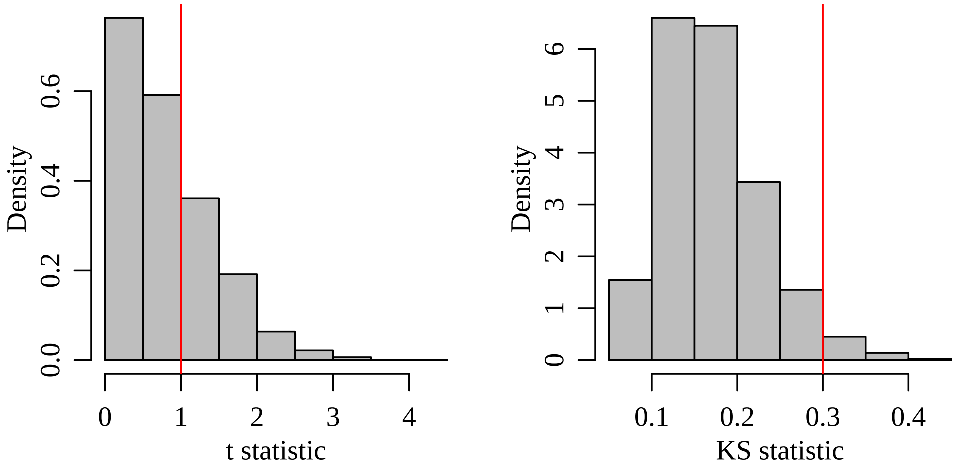


Figure 2.4: Randomization distributions for the $t$ and $KS$ statistics for the first example.

- $\bar{y}_A = 10.11$, $\bar{y}_B = 10.73$.
- $s_A = 1.75$, $s_B = 1.85$

The difference in sample means is about twice as large as in the previous example, and the sample standard deviations are pretty similar. The $B$-samples are slightly larger than the $A$-samples on average. Is there evidence that this is caused by treatment? Again, we evaluate $\mathrm{H}_0$ using the randomization distributions of our two test statistics.

$$t\text{-statistic}: \quad g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.54 \ , \ \Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.54) = 0.122$$

$$\text{KS-statistic:} \ g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.25 \ , \ \Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.25) = 0.106$$

This time the two test statistics indicate similar evidence against $\mathrm{H}_0$. This is because the difference in the two sample distributions could primarily be summarized as the difference between the sample means, which the $t$-statistic can identify.

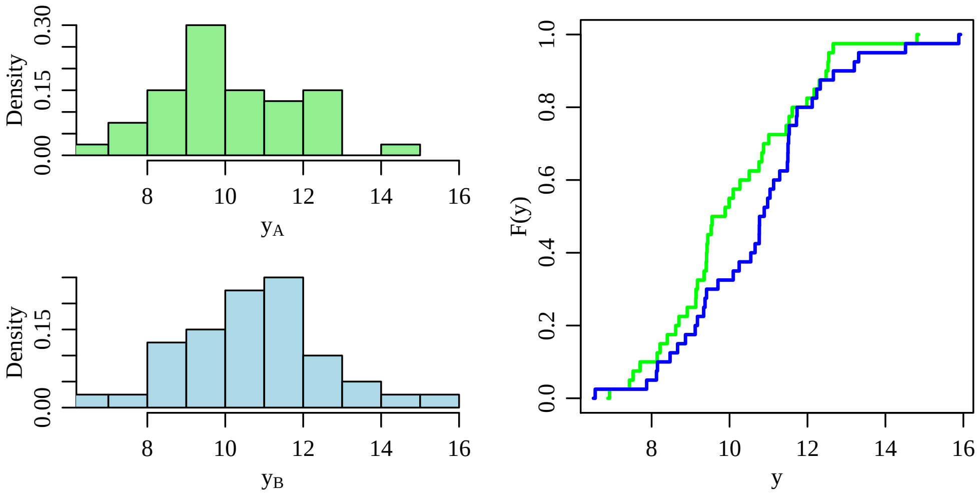


Figure 2.5: Histograms and empirical CDFs of the second two hypothetical samples.

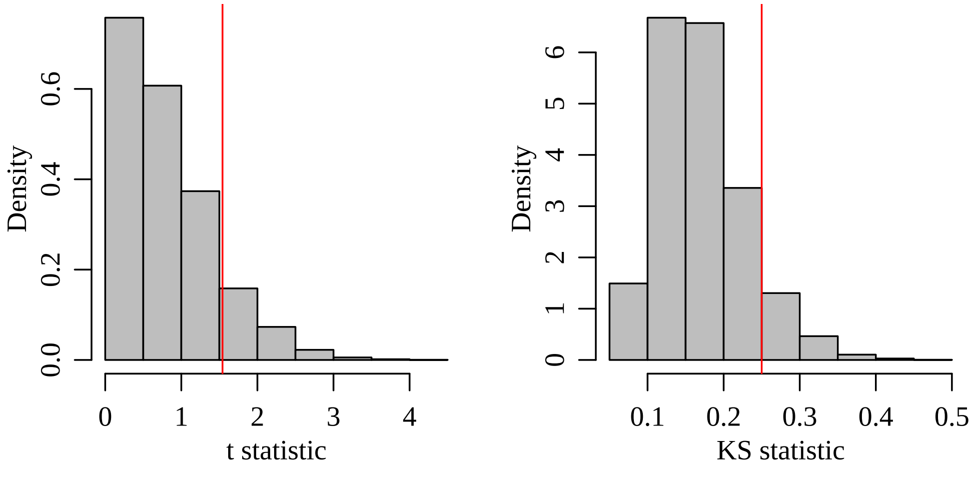


Figure 2.6: Randomization distributions for the $t$ and $KS$ statistics for the second example.

**Discussion:** These last two examples suggest we should abandon $g_t$ in favor of $g_{KS}$ if we are interested in comparing the following hypothesis:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment does affect response} \end{aligned}$$

This is because, as we found, $g_t$ is not sensitive all violations of $\mathrm{H}_0$, it is only sensitive to violations of $\mathrm{H}_0$ where there is a difference in means. However, in many situations we are actually interested in comparing the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment increases responses or decreases responses} \end{aligned}$$

In this case $\mathrm{H}_0$ and $\mathrm{H}_1$ are not complementary, and we are only interested in evidence against $\mathrm{H}_0$ of a certain type, i.e. evidence that is consistent with $\mathrm{H}_1$. In this situation we may want to use a statistic like $g_t$.

## 2.5 Basic decision theory

**Task:** Accept or reject $\mathrm{H}_0$ based on data.

| | truth | |
|---|---|---|
| action | $\mathrm{H}_0$ true | $\mathrm{H}_0$ false |
| accept $\mathrm{H}_0$ | correct decision | type II error |
| reject $\mathrm{H}_0$ | type I error | correct decision |

As we discussed,

- the $p$-value can measure of evidence against $\mathrm{H}_0$;
- the smaller the $p$-value, the larger the evidence against $\mathrm{H}_0$.

**Decision procedure:**

1. Compute the $p$-value by comparing observed test statistic to the null distribution.
2. Reject $\mathrm{H}_0$ if the $p$-value $\leq \alpha$, otherwise accept $\mathrm{H}_0$.

This procedure is called a *level-$\alpha$ test.* It controls the pre-experimental probability of a *type I error*, or for a series of experiments, controls the *type I error rate.*

$$\begin{aligned}
\Pr(\text{type I error}|\mathrm{H}_0) &= \Pr(\text{reject } \mathrm{H}_0 \;|\mathrm{H}_0) \\
&= \Pr(p\text{-value} \leq \alpha|\mathrm{H}_0) \\
&= \alpha
\end{aligned}$$

**Single Experiment Interpretation:** If you use a level-$\alpha$ test for your experiment where $\mathrm{H}_0$ is true, then **before you run the experiment** there is probability $\alpha$ that you will erroneously reject $\mathrm{H}_0$.

**Many Experiments Interpretation:** If level-$\alpha$ tests are used in a large population of experiments, then $\mathrm{H}_0$ will be declared false in $(100 \times \alpha)\%$ of the experiments in which $\mathrm{H}_0$ is true.

$$\begin{aligned}
\Pr(\mathrm{H}_0 \text{ rejected}|\mathrm{H}_0 \text{ true}) &= \alpha \\
\Pr(\mathrm{H}_0 \text{ accepted}|\mathrm{H}_0 \text{ true}) &= 1-\alpha \\
\Pr(\mathrm{H}_0 \text{ rejected}|\mathrm{H}_0 \text{ false}) &= ? \\
\Pr(\mathrm{H}_0 \text{ accepted}|\mathrm{H}_0 \text{ false}) &= ?
\end{aligned}$$

$\Pr(\mathrm{H}_0 \text{ rejected}|\mathrm{H}_0 \text{ false})$ is the *power*. Typically we need to be more specific than "$\mathrm{H}_0$ false" in order to calculate the power. We need to specify **how** it is false.

# Chapter 3

# Tests based on population models

## 3.1 Relating samples to populations

If the experiment is

- complicated,
- non- or partially randomized, or
- includes nuisance factors

then a null distribution based on randomization may be difficult to obtain. An alternative approach to hypothesis testing is based on formulating a sampling model.

Consider the following model for our wheat yield experiment:

- There is a large/infinite population of plots of similar size/shape/composition as the plots in our experiment.
- When $A$ is applied to these plots, the distribution of plot yields can be represented by a probability distribution $p_A$ with

$$\text{expectation} = \mathrm{E}[Y_A] = \textstyle\int y p_A(y) dy = \mu_A,$$
$$\text{variance} = \mathrm{Var}[Y_A] = \mathrm{E}[(Y_A - \mu_A)^2] = \sigma_A^2.$$

- When $B$ is applied to these plots, the distribution of plot yields can be represented by a probability distribution $p_B$ with

$$\begin{aligned}&\text{expectation} = \mathrm{E}[Y_B] = \textstyle\int y p_B(y)dy = \mu_B,\\ &\text{variance} = \mathrm{Var}[Y_B] = \mathrm{E}[(Y_B - \mu_B)^2] = \sigma_B^2.\end{aligned}$$

- The plots that received $A$ in our experiment can be viewed as independent samples from $p_A$, and likewise for plots receiving $B$.

$$\begin{aligned}Y_{1,A},\ldots,Y_{n_A,A} &\sim \text{ i.i.d. } p_A\\ Y_{1,B},\ldots,Y_{n_B,B} &\sim \text{ i.i.d. } p_B\end{aligned}$$

**Recall from intro stats:**

$$\begin{aligned}\mathrm{E}[\bar{Y}_A] &= \mathrm{E}[\frac{1}{n_A}\sum_{i=1}^{n_A} Y_{i,A}]\\ &= \frac{1}{n_A}\sum_{i=1}^{n_A}\mathrm{E}[Y_{i,A}]\\ &= \frac{1}{n_A}\sum_{i=1}^{n_A}\mu_A = \mu_A\end{aligned}$$

We say that $\bar{Y}_A$ is an *unbiased estimator* of $\mu_A$. Furthermore, if $Y_{1,A},\ldots,Y_{n_A,A}$ are independent samples from population $A$, then

$$\begin{aligned}\mathrm{Var}[\bar{Y}_A] &= \mathrm{Var}[\frac{1}{n_A}\sum Y_{i,A}]\\ &= \frac{1}{n_A^2}\sum \mathrm{Var}[Y_{i,A}]\\ &= \frac{1}{n_A^2}\sum \sigma_A^2 = \sigma_A^2/n_A.\end{aligned}$$

This means that as $n_A \to \infty$,

$$\mathrm{Var}[\bar{Y}_A] = \mathrm{E}[(Y_A - \mu_A)^2] \to 0$$

which, with unbiasedness, implies

$$\bar{Y}_A \to \mu_A$$

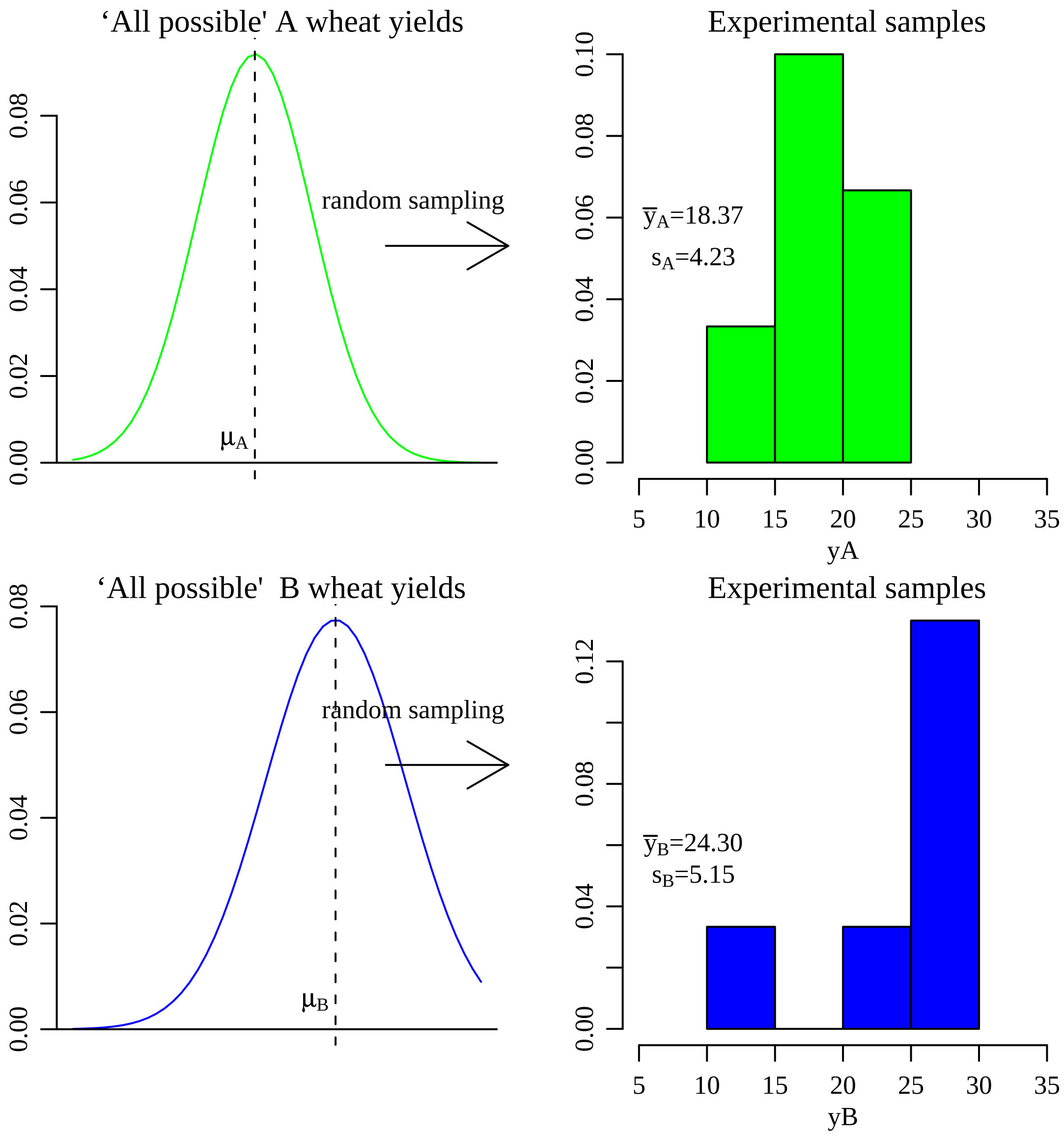


Figure 3.1: The population model

and we say that $\bar{Y}_A$ is a *consistent estimator* for $\mu_A$. Several of our other sample characteristics are also consistent for the corresponding population characteristics. As $n \to \infty$,

$$\begin{aligned} \bar{Y}_A &\to \mu_A \\ s_A^2 &\to \sigma_A^2 \\ \frac{\#\{Y_{i,A} \leq x\}}{n_A} = \hat{F}_A(x) &\to F_A(x) = \int_{-\infty}^{x} p_A(y)dy \end{aligned}$$

**Back to hypothesis testing:**

We can formulate null and alternative hypotheses in terms of population quantities. For example, if $\mu_B > \mu_A$, we would recommend $B$ over $A$, and vice versa.

- $\mathrm{H}_0 : \mu_A = \mu_B$
- $\mathrm{H}_1 : \mu_A \neq \mu_B$

The experiment is performed and it is observed that

$$18.37 = \bar{y}_A < \bar{y}_B = 24.3$$

This is **some** evidence that $\mu_A > \mu_B$. How much evidence is it? Should we reject the null hypothesis? Consider evaluating evidence against $\mathrm{H}_0$ with our $t$-statistic:

$$g_t(\mathbf{y}_A, \mathbf{y}_B) = \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}$$

To decide whether to reject $\mathrm{H}_0$ or not, we need to know the distribution of $g(\mathbf{Y}_A, \mathbf{Y}_B)$ under $\mathrm{H}_0$. Consider the following setup:

**Assume:**

$$\begin{aligned} &Y_{1,A}, \ldots, Y_{n_A,A} \sim \text{i.i.d. } p_A \\ &Y_{1,B}, \ldots, Y_{n_B,B} \sim \text{i.i.d. } p_B \end{aligned}$$

**Evaluate:** $\mathrm{H}_0 : \mu_A = \mu_B$ versus $\mathrm{H}_1 : \mu_A \neq \mu_B$ , i.e., whether or not

$$\int y p_A(y)dy = \int y p_B(y)dy.$$

To make this evaluation and obtain a $p$-value, we need the distribution of $g(\mathbf{Y}_A, \mathbf{Y}_B)$ under $\mu_A = \mu_B$. This will involve assumptions about/approximations to $p_A$ and $p_B$.

## 3.2 The normal distribution

The normal distribution is useful because in many cases,

- our data are approximately normally distributed, and/or
- our sample means are approximately normally distributed.

These are both due to the central limit theorem. Letting $P(\mu, \sigma)$ denote a population with mean $\mu$ and variance $\sigma^2$, then

$$\left.\begin{array}{ccc} X_1 & \sim & P_1(\mu_1, \sigma_1^2) \\ X_2 & \sim & P_2(\mu_2, \sigma_2^2) \\ & \vdots & \\ X_m & \sim & P_m(\mu_m, \sigma_m^2) \end{array}\right\} \Rightarrow \sum_{j=1}^{m} X_i \dot\sim \text{ normal} \left( \sum \mu_j, \sum \sigma_j^2 \right).$$

Sums of varying quantities are approximately normally distributed.

### Normally distributed data

Consider crop yields from plots of land:

$$Y_i = a_1 \times \text{seed}_i + a_2 \times \text{soil}_i + a_3 \times \text{water}_i + a_4 \times \text{sun}_i + \cdots$$

The empirical distribution of crop yields from a population of fields with varying quantities of seed, soil, water, sun, etc. will be approximately normal $(\mu, \sigma)$, where $\mu$ and $\sigma$ depend on the effects $a_1, a_2, a_3, a_4, \ldots$ and the variability of seed, soil, water, sun, etc..

Additive effects $\Rightarrow$ normally distributed **data**

### Normally distributed means

Consider the following scenario:

Experiment 1: sample $y_1^{(1)}, \ldots, y_n^{(1)} \sim$ i.i.d. $p$ and compute $\bar{y}^{(1)}$;

Experiment 2: sample $y_1^{(2)}, \ldots, y_n^{(2)} \sim$ i.i.d. $p$ and compute $\bar{y}^{(2)}$;

$\vdots$

Experiment $m$: sample $y_1^{(m)}, \ldots, y_n^{(m)} \sim$ i.i.d. $p$ and compute $\bar{y}^{(m)}$.

A histogram of $\{\bar{y}^{(1)}, \ldots, \bar{y}^{(m)}\}$ will look approximately normally distributed with

$$\begin{aligned} \text{sample mean } \{y^{(1)}, \ldots, y^{(m)}\} &\approx \mu \\ \text{sample variance } \{y^{(1)}, \ldots, y^{(m)}\} &\approx \sigma^2/n \end{aligned}$$

i.e. the *sampling distribution* of the **mean** is approximately normal$(\mu, \sigma^2/n)$, even if the sampling distribution of the **data** are not normal.

**Basic properties of the normal distribution:**

- $Y \sim$ normal$(\mu, \sigma^2) \Rightarrow aY + b \sim$normal$(a\mu + b, a^2\sigma^2)$.
- $Y_1 \sim$ normal$(\mu_1, \sigma_1^2)$, $Y_2 \sim$ normal$(\mu_2, \sigma_2^2), Y_1, Y_2$ independent $\Rightarrow Y_1 + Y_2 \sim$ normal$(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- if $Y_1, \ldots, Y_n \sim$ i.i.d. normal$(\mu, \sigma^2)$, then $\bar{Y}$ is statistically independent of $s^2$.

**How does this help with hypothesis testing?**

Consider testing H$_0$ : $\mu_A = \mu_B$ (treatment doesn't affect mean). Then regardless of distribution of data, under H$_0$ :

$$\begin{aligned} \bar{Y}_A &\;\dot{\sim}\; \text{normal}(\mu, \sigma_A^2/n_A) \\ \bar{Y}_B &\;\dot{\sim}\; \text{normal}(\mu, \sigma_B^2/n_B) \\ \bar{Y}_B - \bar{Y}_A &\;\dot{\sim}\; \text{normal}(0, \sigma_{AB}^2) \end{aligned}$$

where $\sigma_{AB}^2 = \sigma_A^2/n_A + \sigma_B^2/n_B$. So if we knew the variances, we'd have a null distribution.

## 3.3 Introduction to the $t$-test

Consider a simple one-sample hypothesis test:

$Y_1, \ldots, Y_n \sim$ i.i.d. $P$, with mean $\mu$ and variance $\sigma^2$.

$\text{H}_0 : \mu = \mu_0$

$\text{H}_1 : \mu \neq \mu_0$

**Examples:**

Physical therapy

- $Y_i$ = muscle strength after treatment - muscle score before.
- $\text{H}_0 :\ \text{E}[Y_i] = 0$

Physics

- $Y_i$ = boiling point of a sample of an unknown liquid.
- $\text{H}_0 :\ \text{E}[Y_i] = 100°$ C.

To test $\text{H}_0$, we need a test statistic and its distribution under $\text{H}_0$.
$|\bar{y} - \mu_0|$ might make a good test statistic:

- it is sensitive to deviations from $\text{H}_0$.
- its sampling distribution is approximately known:

  $\text{E}[\bar{Y}] = \mu$

  $\text{Var}[\bar{Y}] = \sigma^2/n$

  $\bar{Y}$ is approximately normal.

Under $\text{H}_0$

$$(\bar{Y} - \mu_0) \sim \text{normal}(0, \sigma^2/n),$$

but we can't use this as a null distribution because $\sigma^2$ is unknown. What if we scale $(\bar{Y} - \mu_0)$? Then

$$f(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

is approximately standard normal and we write $f(\mathbf{Y}) \sim \text{normal}(0, 1)$. Since this distribution contains no unknown parameters we could potentially use it as a null distribution. However, having observed the data $\mathbf{y}$, is $f(\mathbf{y})$ a statistic?

- $\bar{y}$ is computable from the data and $n$ is known;

- $\mu_0$ is our hypothesized value, a fixed number that we have chosen.
- $\sigma$ is not determined by us and is unknown.

The solution to this problem is to approximate the **population** variance $\sigma^2$ with the **sample** variance $s^2$.

**One-sample t-statistic:**

$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

For a given value of $\mu_0$ this is a statistic. What is the null distribution of $t(\mathbf{Y})$?

$$s \approx \sigma \quad \text{so} \quad \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \approx \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

If $Y_1, \ldots, Y_n \sim$ i.i.d. normal$(\mu_0, \sigma^2)$ then $\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}$ is normal$(0,1)$, and so it would seem that $t(\mathbf{Y})$ is approximately distributed as a standard normal distribution under $\mathrm{H}_0 : \mu = \mu_0$. However, if the approximation $s \approx \sigma$ is poor, like when $n$ is small, we need to take account of our uncertainty in the estimate of $\sigma$.

**The $\chi^2$ distribution**

$$\begin{aligned} Z_1, \ldots, Z_n \sim \text{i.i.d. normal}(0,1) \quad &\Rightarrow \quad \sum Z_i^2 \sim \chi^2_n, \text{chi-squared dist with } n \text{ degrees of freedom} \\ & \quad\quad \sum (Z_i - \bar{Z})^2 \sim \chi^2_{n-1} \end{aligned}$$

$$\begin{aligned} Y_1, \ldots, Y_n \sim \text{i.i.d. normal}(\mu, \sigma) \quad &\Rightarrow \quad (Y_1 - \mu)/\sigma, \ldots, (Y_n - \mu)/\sigma \sim \text{i.i.d. normal}(0,1) \\ &\Rightarrow \quad \frac{1}{\sigma^2} \sum (Y_i - \mu)^2 \sim \chi^2_n \\ & \quad\quad \frac{1}{\sigma^2} \sum (Y_i - \bar{Y})^2 \sim \chi^2_{n-1} \end{aligned}$$

**Some intuition:** Which vector do you expect to be bigger: $(Z_1, \ldots, Z_n)$ or $(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$? $(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$ is a vector of length $n$ but lies in an $n-1$ dimensional space. In fact, it is has a **singular multivariate normal** distribution, with a covariance matrix of rank $(n-1)$.

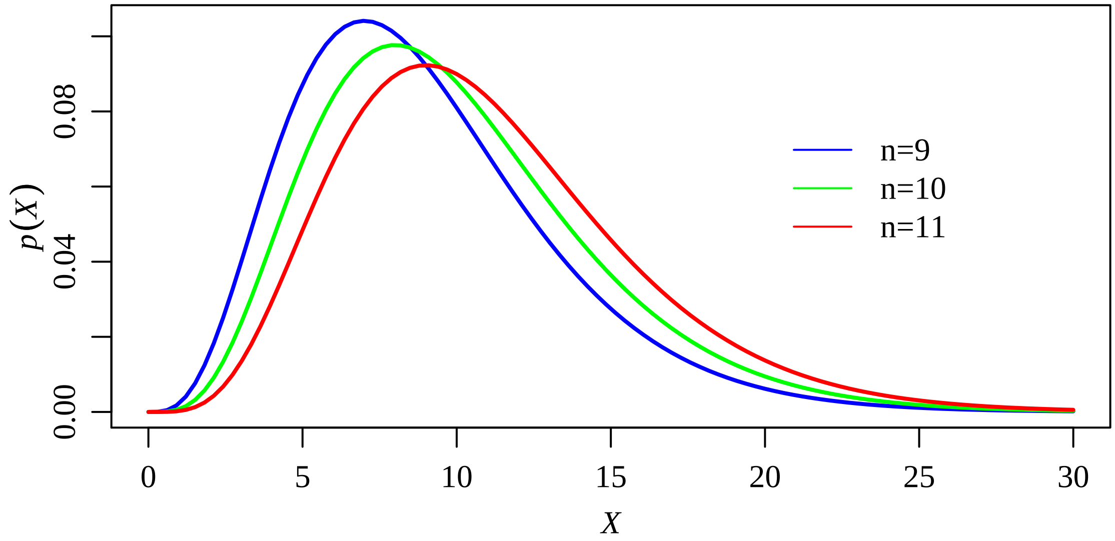


Figure 3.2: $\chi^2$ distributions

Getting back to the problem at hand,

$$\frac{n-1}{\sigma^2}s^2 = \frac{n-1}{\sigma^2}\frac{1}{n-1}\sum(Y_i - \bar{Y})^2 \sim \chi^2_{n-1},$$

which is a known distribution we can look up on a table or with a computer.

**The $t$-distribution**

If

- $Z \sim$ normal (0,1) ;
- $X \sim \chi^2_m$;
- $Z, X$ statistically independent,

then

$$\frac{Z}{\sqrt{X/m}} \sim t_m, \text{the } t\text{-distribution with } m \text{ degrees of freedom}$$

How does this help us? Recall that if $Y_1, \ldots, Y_n \sim$ i.i.d. normal$(\mu, \sigma^2)$,

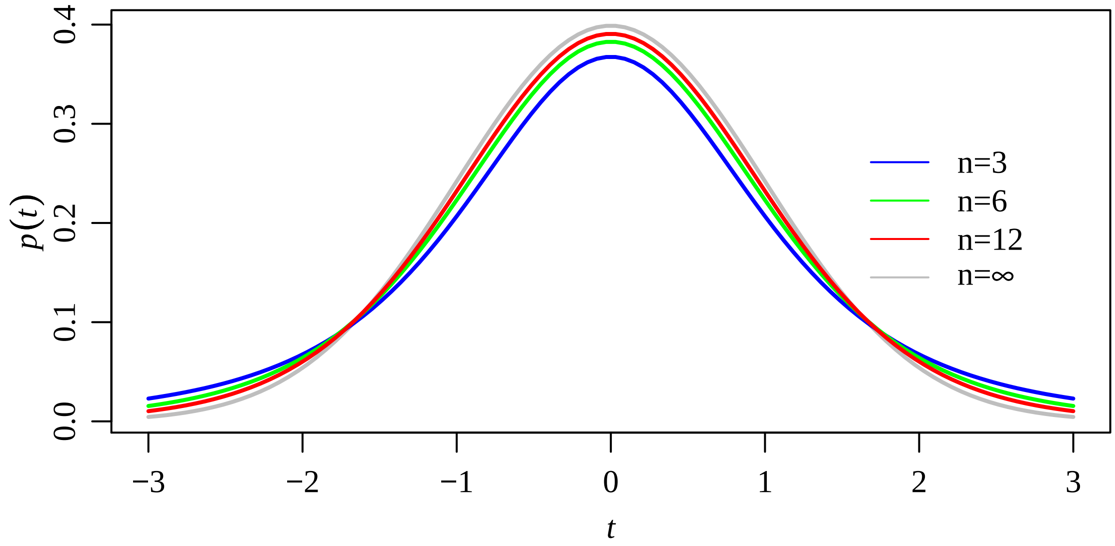


Figure 3.3: $t$-distributions

- $\sqrt{n}(\bar{Y}-\mu)/\sigma$ ~normal(0,1)
- $\frac{n-1}{\sigma^2}s^2 \sim \chi^2_{n-1}$
- $\bar{Y}, s^2$ are independent.

Let $Z = \sqrt{n}(\bar{Y}-\mu)/\sigma$, $X = \frac{n-1}{\sigma^2}s^2$. Then

$$\begin{aligned}\frac{Z}{\sqrt{X/(n-1)}} &= \frac{\sqrt{n}(\bar{Y}-\mu)/\sigma}{\sqrt{\frac{n-1}{\sigma^2}s^2/(n-1)}} \\ &= \frac{\bar{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}\end{aligned}$$

This is still not a statistic because $\mu$ is unknown. However, under a specific hypothesis like $\mathrm{H}_0 : \mu = \mu_0$, it is a statistic:

$$t(\mathbf{Y}) = \frac{\bar{Y}-\mu_0}{s/\sqrt{n}} \sim t_{n-1} \quad \text{if } \mathrm{E}[Y] = \mu_0$$

It is called the *$t$-statistic.*

Some questions for discussion:

- What does $X/m$ converge to as $m \to \infty$?
- What happens to the distribution of $t(\mathbf{Y})$ when $n \to \infty$? Why?
- Consider the situation where the data are not normally distributed.
  - What is the distribution of $\sqrt{n}(\bar{Y} - \mu)/\sigma$ for large and small $n$?
  - What is the distribution of $\frac{n-1}{\sigma^2}s^2$ for large and small $n$?
  - Are $\sqrt{n}(\bar{Y} - \mu)$ and $s^2$ independent for small $n$? What about for large $n$?

**Two-sided, one-sample t-test:**

1. Sampling model: $Y_1, \ldots, Y_n \sim$ i.i.d. normal$(\mu, \sigma^2)$
2. Null hypothesis: $\mathrm{H}_0 : \mu = \mu_0$
3. Alternative hypothesis: $\mathrm{H}_1 : \mu \neq \mu_0$
4. Test statistic: $t(\mathbf{Y}) = \sqrt{n}(\bar{Y} - \mu_0)/s$
   - Pre-experiment we think of this as a random variable, an unknown quantity that is to be randomly sampled from a population.
   - Post-experiment this is a fixed number.
5. Null distribution: Under the normal sampling model and $\mathrm{H}_0$, the *sampling distribution* of $t(\mathbf{Y})$ is the $t$-distribution with $n-1$ degrees of freedom:
   $$t(\mathbf{Y}) \sim t_{n-1}$$
   If $\mathrm{H}_0$ is not true, then $t(\mathbf{Y})$ does not have a $t$-distribution. If the data are normal but the mean is not $\mu_0$, then $t(\mathbf{Y})$ has a *non-central $t$-distribution*, which we will use later to calculate power, or the type II error rate. If the data are not normal then the distribution of $t(\mathbf{Y})$ is not a $t$-distribution.

6. $p$-value: Let $\mathbf{y}$ be the observed data.

$$\begin{aligned} p\text{-value} &= \Pr(|t(\mathbf{Y})| \geq |t(\mathbf{y})| | \mathrm{H}_0) \\ &= \Pr(|T_{n-1}| \geq |t(\mathbf{y})|) \\ &= 2 \times \Pr(T_{n-1} \geq |t(\mathbf{y})|) \\ &= \mathtt{2 * (1 - pt(tobs, n - 1))} \\ &= \mathtt{t.test(y, mu = mu0)} \end{aligned}$$

7. Level-$\alpha$ decision procedure: Reject $\mathrm{H}_0$ if

   - $p$-value $\leq \alpha$ or equivalently
   - $|t(\mathbf{y})| \geq t_{(n-1),1-\alpha/2}$ (for $\alpha = .05$, $t_{(n-1),1-\alpha/2} \approx 2$ ).

   The value $t_{(n-1),1-\alpha/2} \approx 2$ is called the *critical value* value for this test. In general, the critical value is the value of the test statistic above which we would reject $\mathrm{H}_0$.

**Question:** Suppose our procedure is to reject $\mathrm{H}_0$ only when $t(\mathbf{y}) \geq t_{(n-1),1-\alpha}$. Is this a level-$\alpha$ test?

## 3.4 Two sample tests

Recall the wheat example:

| | | | | | |
|---|---|---|---|---|---|
| *B*<br>26.9 | *A*<br>11.4 | *B*<br>26.6 | *A*<br>23.7 | *B*<br>25.3 | *B*<br>28.5 |
| *B*<br>14.2 | *A*<br>17.9 | *A*<br>16.5 | *A*<br>21.1 | *B*<br>24.3 | *A*<br>19.6 |

**Sampling model:**

$$\begin{aligned} Y_{1A}, \ldots, Y_{n_A A} &\sim \text{i.i.d. normal}(\mu_A, \sigma^2) \\ Y_{1B}, \ldots, Y_{n_B B} &\sim \text{i.i.d. normal}(\mu_B, \sigma^2). \end{aligned}$$

In addition to normality we assume for now that both variances are equal.

**Hypotheses:** $\mathrm{H}_0 : \mu_A = \mu_B$; $\quad \mathrm{H}_A$: $\mu_A \neq \mu_B$

Recall that

$$\bar{Y}_B - \bar{Y}_A \sim N\left(\mu_B - \mu_A, \sigma^2\left[\frac{1}{n_A} + \frac{1}{n_B}\right]\right).$$

Hence if $H_0$ is true then

$$\bar{Y}_B - \bar{Y}_A \sim N\left(0, \sigma^2\left[\frac{1}{n_A} + \frac{1}{n_B}\right]\right).$$

How should we estimate $\sigma^2$ ?

$$\begin{aligned} s_p^2 &= \frac{\sum_{i=1}^{n_A}(y_{i,A} - \bar{y}_A)^2 + \sum_{i=1}^{n_B}(y_{i,B} - \bar{y}_B)^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2 \end{aligned}$$

This gives us the following *two-sample* $t$-statistic:

$$t(\mathbf{Y}_A, \mathbf{Y}_B) = \frac{\bar{Y}_B - \bar{Y}_A}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A + n_B - 2}$$

**Self-check exercises:**

1. Show that $(n_A + n_B - 2)s_p^2/\sigma^2 \sim \chi^2_{n_A+n_B-2}$ (recall how the $\chi^2$ distribution was defined).
2. Show that the two-sample $t$-statistic has a $t$-distribution with $n_A + n_B - 2$ d.f.

**Numerical Example (wheat again)**

Suppose we want to have a type-I error rate of $\alpha = 0.05$.

Decision procedure:

- Level $\alpha$ test of $\mathrm{H}_0 : \mu_A = \mu_B$, with $\alpha = 0.05$
- Reject $\mathrm{H}_0$ if $p$-value $< 0.05$
- Reject $\mathrm{H}_0$ if $|t(\boldsymbol{y}_A, \boldsymbol{y}_B)| > t_{10,.975} = 2.23$

Data:

- $\bar{y}_A = 18.36$, $s_A^2 = 17.93$, $n_A = 6$
- $\bar{y}_B = 24.30$, $s_B^2 = 26.54$, $n_B = 6$

$t$-statistic:

- $s_p^2 = 22.24$, $s_p = 4.72$
- $t(y_A, y_B) = 5.93/(4.72\sqrt{1/6 + 1/6}) = 2.18$

Inference:

- Hence the $p$-value= $\Pr(|T_{10}| \geq 2.18) = 0.054$
- Hence $\mathrm{H}_0$: $\mu_A = \mu_B$ is **not rejected** at level $\alpha = 0.05$

```
> t.test(y[x==”A”],y[x==”B”],var.equal=TRUE)

        Two Sample t−test

data:  y[x == ”A”] and y[x == ”B”]
t = −2.1793, df = 10, p−value = 0.05431
alternative hypothesis:true difference in means is not equal to 0
95 percent confidence interval:
 −11.999621   0.132954
sample estimates:
mean of x mean of y
 18.36667  24.30000
```

Always keep in mind where the $p$-value comes from: See Figure 3.4.

**Comparison to the randomization test:**
Recall that we have already compared the two-sample $t$-statistic to its *randomization distribution*. A sample from the randomization distribution were obtained as follows:

1. Sample a treatment assignment according to the randomization scheme.
2. Compute the value of $t(\mathbf{Y}_A, \mathbf{Y}_B)$ under this treatment assignment and assuming the null hypothesis.

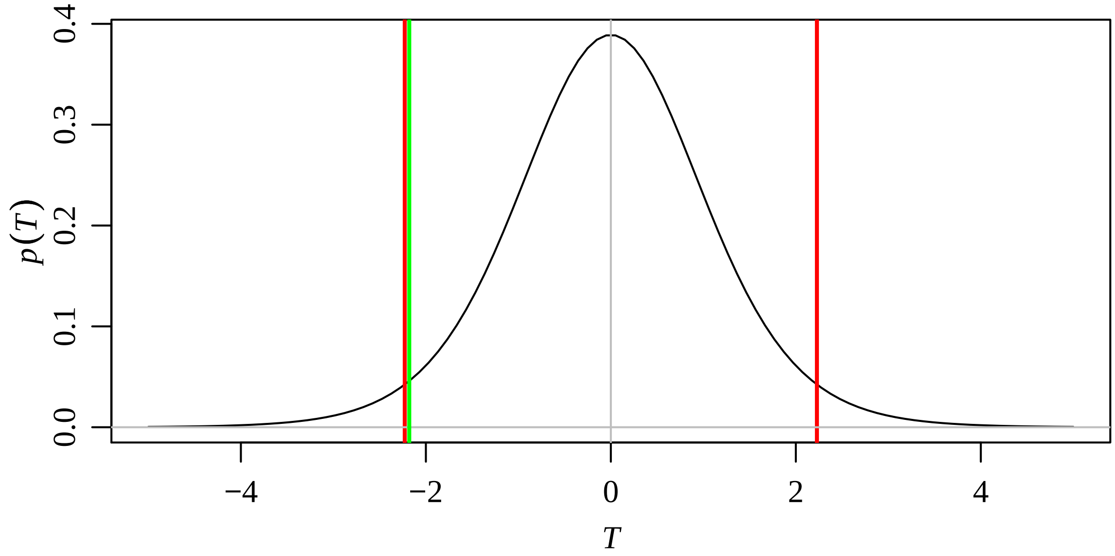


Figure 3.4: The $t$-distribution under $\mathrm{H}_0$ for the wheat example

The *randomization distribution* of the $t$-**statistic** is then approximated by the empirical distribution of

$$t^{(1)}, \ldots, t^{(S)}$$

To obtain the $p$-value, we compare $t_{\text{obs}} = t(\boldsymbol{y}_A, \boldsymbol{y}_B)$ to the empirical distribution of $t^{(1)}, \ldots, t^{(S)}$:

$$p\text{-value} = \frac{\#(|t^{(s)}| \geq |t_{\text{obs}}|)}{S}$$

```
t.stat.obs<-t.test( y[x==”A”],y[x==”B”], var.equal=T)$stat
t.stat.sim<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  tmp<-t.test(y[xsim==”B”],y[xsim==”A”],var.equal=T)
  t.stat.sim[s]<-tmp$stat
}

mean( abs(t.stat.sim) >= abs(t.stat.obs) )
```

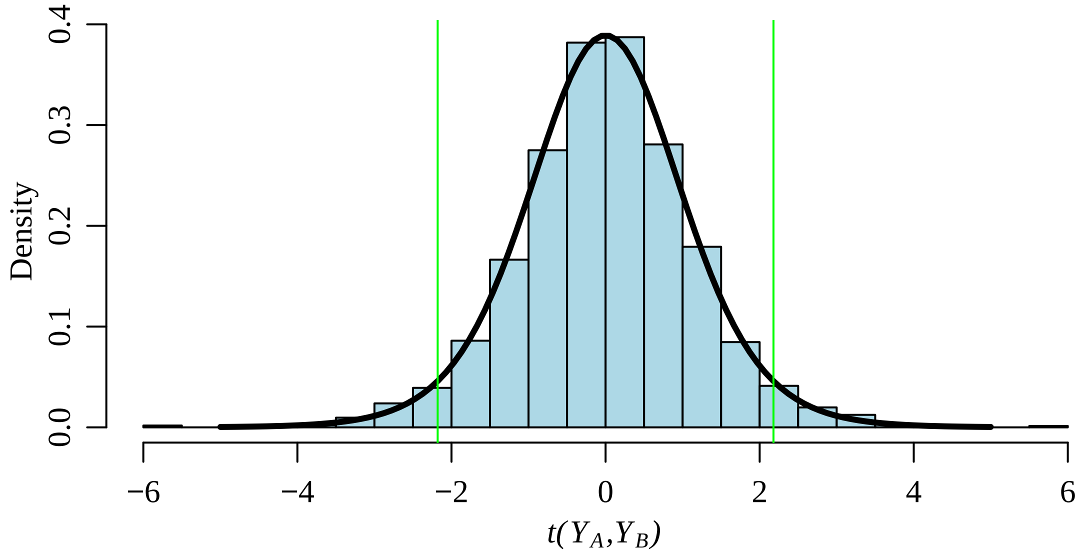


Figure 3.5: Randomization and $t$-distributions for the $t$-statistic under $\mathrm{H}_0$

When I ran this, I got

$$\frac{\#(|t^{(s)}| \geq 2.18)}{S} = 0.058 \approx 0.054 = \Pr(|T_{n_A+n_B-2}| \geq 2.18)$$

Is this surprising? These two $p$-values were obtained via two completely different ways of looking at the problem!

**Assumptions:** Under $\mathrm{H}_0$,

- Randomization Test:
  1. Treatments are randomly assigned
- $t$-test:
  1. Data are independent samples
  2. Each population is normally distributed
  3. The two populations have the same variance

**Imagined Universes:**

- Randomization Test: **Numerical responses** remain **fixed**, we imagine only alternative treatment assignments.

- $t$-test: **Treatment assignments** remain **fixed**, we imagine an alternative sample of experimental units and/or conditions, giving **different numerical responses**.

**Inferential Context / Type of generalization**

- Randomization Test: inference is specific to our particular experimental units and conditions.
- $t$-test: under our assumptions, inference claims to be **generalizable** to other units / conditions, i.e. to a larger population.

Yet the numerical results are often nearly identical.

**Keep the following concepts clear:**

**$t$-statistic** : a scaled difference in sample means, computed from the data.

**$t$-distribution** : the probability distribution of a normal random variable divided by the square-root of a $\chi^2$ random variable.

**$t$-test** : a comparison of a $t$-statistic to a $t$-distribution

**randomization distribution** : the probability distribution of a test statistic under random treatment reassignments and $\mathrm{H}_0$

**randomization test** : a comparison of a test statistic to its randomization distribution

**randomization test with the $t$-statistic** : a comparison of the $t$-statistic to its randomization distribution

**Some history:**

de Moivre (1733): Approximating binomial distributions $T = \sum_{i=1}^{n} Y_i,\ \ Y_i \in \{0, 1\}$.

Laplace (1800s): Used normal distribution to model measurement error in experiments.

Gauss (1800s) : Justified least squares estimation by assuming normally distributed errors.

Gosset/Student (1908): Derived the $t$-distribution.

Gosset/Student (1925): “testing the significance”

Fisher (1925): “level of significance”

Fisher (1920s?): Fisher’s exact test.

Fisher (1935): “It seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied.”

Box and Andersen (1955): Approximation of randomization null distributions (and a long discussion).

Rodgers (1999): “If Fisher’s ANOVA had been invented 30 years later or computers had been available 30 years sooner, our statistical procedures would probably be less tied to theoretical distributions as what they are today”

## 3.5 Checking assumptions

For our $t$-statistic

$$t(\mathbf{Y}_A, \mathbf{Y}_B) = \frac{\bar{Y}_B - \bar{Y}_A}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

we showed that if $Y_{1,A}, \ldots, Y_{n_A,A}$ and $Y_{1,B}, \ldots, Y_{n_B,B}$ are independent samples from $p_A$ and $p_B$ respectively, and

(a) $\mu_A = \mu_B$

(b) $\sigma_A^2 = \sigma_B^2$

(c) $p_A$ and $p_B$ are normal distributions

then

$$t(\mathbf{Y}_A, \mathbf{Y}_B) \sim t_{n_A+n_B-2}$$

So our null distribution really assumes conditions (a), (b) and (c). Thus if we perform a level-$\alpha$ test and reject $\mathrm{H}_0$, we are really just rejecting that (a), (b), (c) are all true.

$$\text{reject } \mathrm{H}_0 \quad \Rightarrow \quad \text{one of (a), (b) or (c) is not true}$$

For this reason, we will often want to check if conditions (b) and (c) are plausibly met. If

(b) is met

(c) is met

$\mathrm{H}_0$ is rejected, then

this is evidence that $\mathrm{H}_0$ is rejected **because** $\mu_A \neq \mu_B$.

### 3.5.1 Checking normality

Normality can be checked by a *normal probability plot*

**Idea:** Order observations within each group:

$$y_{(1),A} \leq y_{(2),A} \leq \cdots \leq y_{(n_A),A}$$

compare these **sample quantiles** to the quantiles of a **standard normal distribution**:

$$z_{\frac{1}{n_A}-\frac{1}{2}} \leq z_{\frac{2}{n_A}-\frac{1}{2}} \leq \cdots \leq z_{\frac{n_A}{n_A}-\frac{1}{2}}$$

here $\Pr(Z \leq z_{\frac{k}{n_A}-\frac{1}{2}}) = \frac{k}{n_A} - \frac{1}{2}$. The $-\frac{1}{2}$ is a continuity correction.

If data are normal, the relationship should be approximately linear. Thus we plot the pairs

$$(z_{\frac{1}{n_A}-\frac{1}{2}}, y_{(1),A}), \ldots, (z_{\frac{n_A}{n_A}-\frac{1}{2}}, y_{(n_A),A}).$$

Normality may be checked roughly by fitting straight lines to the probability plots and examining their slopes

### 3.5.2 Unequal variances

For now, we will use the following rule of thumb:

If $1/4 < s_A^2/s_B^2 < 4$, we won't worry too much about unequal variances.

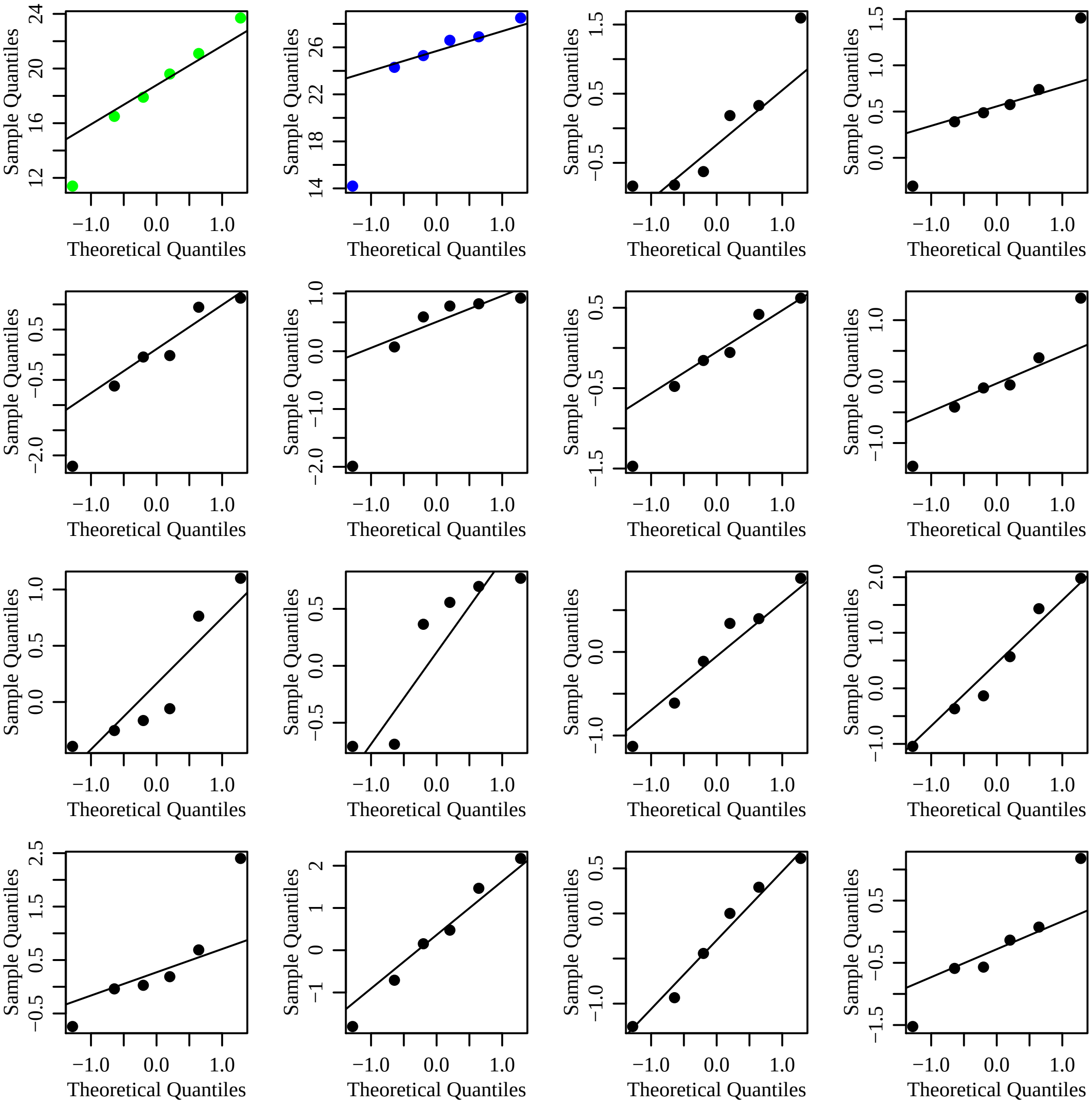


Figure 3.6: Normal scores plots.

This may not sound very convincing. In later sections, we will show how to perform formal hypothesis tests for equal variances. However, this won't completely solve the problem. If variances do seem unequal we have a variety of options available:

- use the randomization null distribution;
- transform the data to stabilize the variances (to be covered later);
- use a modified $t$-test that allows unequal variance.

The modified $t$-statistic is

$$t_w(\boldsymbol{y}_A, \boldsymbol{y}_B) = \frac{\bar{y}_B - \bar{y}_A}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

This statistic looks pretty reasonable, and for large $n_A$ and $n_B$ its null distribution will indeed be a normal$(0,1)$ distribution. However, the exact null distribution is only **approximately** a $t$-distribution, even if the data are actually normally distributed. The $t$-distribution we compare $t_w$ to is a $t_{\nu_w}$-distribution, where the degrees of freedom $\nu_w$ are given by

$$\nu_w = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{1}{n_A-1}\left(\frac{s_A^2}{n_A}\right)^2 + \frac{1}{n_B-1}\left(\frac{s_B^2}{n_B}\right)^2}.$$

This is known as *Welch's approximation*; it may not give an integer as the degrees of freedom.

This $t$-distribution is *not*, in fact the exact sampling distribution of $t_{\text{diff}}(\mathbf{y_A}, \mathbf{y_B})$ under the null hypothesis that $\mu_A = \mu_B$, and $\sigma_A^2 \neq \sigma_B^2$. This is because the null distribution depends on the ratio of the unknown variances, $\sigma_A^2$ and $\sigma_B^2$. This difficulty is known as the **Behrens-Fisher problem**.

### Which two-sample t-test to use?

- If the sample sizes are the same ($n_A = n_B$) then the test statistics $t_w(\boldsymbol{y}_A, \boldsymbol{y}_B)$ and $t(\boldsymbol{y}_A, \boldsymbol{y}_B)$ are the same; however the degrees of freedom used in the null distribution will be different unless the sample standard deviations are the same.

- If $n_A > n_B$, but $\sigma_A^2 < \sigma_B^2$, and $\mu_A = \mu_B$ then the two sample test based on comparing $t(\boldsymbol{y}_A, \boldsymbol{y}_B)$ to a $t$-distribution on $n_A + n_B - 2$ d.f. will reject more than 5% of the time.
  - If the null hypothesis that both the means and variances are equal, i.e.

    H$_0$: $\mu_A = \mu_B$ **and** $\sigma_A^2 = \sigma_B^2$

    is *scientifically relevant*, then we are computing a valid $p$-value, and this higher rejection rate is a good thing! Since when the variances are unequal the null hypothesis is false.
  - If however, the hypothesis that is most scientifically relevant is H$_0$: $\mu_A = \mu_B$

    without placing any restrictions on the variances, then the higher rejection rate in the test that assumes the variances are the same could be very misleading, since $p$-values may be smaller than they are under the correct null distribution (in which $\sigma_A^2 \neq \sigma_B^2$).

    Likewise we will underestimate the probability of type I error.
- If $n_A > n_B$ and $\sigma_A^2 > \sigma_B^2$, then the $p$-values obtained from the test using $t(\boldsymbol{y}_A, \boldsymbol{y}_B)$ will tend to be **conservative** (= larger) than those obtained with $t_w(\boldsymbol{y}_A, \boldsymbol{y}_B)$.

In short: one should be careful about applying the test based on $t(\mathbf{y_A}, \mathbf{y_B})$ if the sample standard deviations appear very different, *and* it is not reasonable to assume equal means and variances under the null hypothesis.

# Chapter 4

# Confidence intervals and power

## 4.1 Confidence intervals via hypothesis tests

Recall that

- $\mathrm{H}_0 : \mathrm{E}[Y] = \mu_0$ is **rejected** if

$$\sqrt{n}|(\bar{y} - \mu_0)/s| \geq t_{1-\alpha/2}$$

- $\mathrm{H}_0 : \mathrm{E}[Y] = \mu_0$ is **not rejected** if

$$\begin{aligned} \sqrt{n}|(\bar{y} - \mu_0)/s| &\leq t_{1-\alpha/2} \\ |\bar{y} - \mu_0| &\leq \frac{1}{\sqrt{n}} s \times t_{1-\alpha/2} \\ \bar{y} - \frac{s}{\sqrt{n}} \times t_{1-\alpha/2} \leq \mu_0 &\leq \bar{y} + \frac{s}{\sqrt{n}} \times t_{1-\alpha/2} \end{aligned}$$

If $\mu_0$ satisfies this last line, then it is in the *acceptance region*. Otherwise it is in the *rejection region*. In other words, "plausible" values of $\mu$ are in the interval

$$\bar{y} \pm \frac{s}{\sqrt{n}} \times t_{1-\alpha/2}$$

We say this interval is a "$100 \times (1 - \alpha)\%$ *confidence interval*" for $\mu$. This interval contains only those values of $\mu$ that are **not rejected** by this level-$\alpha$ test.

### Main property of a confidence interval

Suppose you are going to

1. gather data;
2. compute a $100 \times (1-\alpha)\%$ confidence interval.

Further suppose $\mathrm{H}_0 : \mathrm{E}[Y] = \mu_0$ is true. What is the probability that $\mu_0$ will be in your to-be-sampled (random) interval? In other words, what is the probability that the **random interval will contain the true value**?

$$\begin{aligned}
\Pr(\mu_0 \text{ in interval}|\mathrm{E}[Y]=\mu_0) &= 1-\Pr(\mu_0 \text{not in interval}|\mathrm{E}[Y]=\mu_0) \\
&= 1-\Pr(\text{reject } \mathrm{H}_0|\mathrm{E}[Y]=\mu_0) \\
&= 1-\Pr(\text{reject } \mathrm{H}_0|\mathrm{H}_0 \text{ is true}) \\
&= 1-\alpha
\end{aligned}$$

The quantity $1-\alpha$ is called the *coverage probability.* It is

- the pre-experimental probability that your confidence interval will cover the true value;
- the large sample fraction of experiments in which the confidence interval covers the true mean.

### Confidence interval for a difference between treatments

Recall that we may construct a 95% confidence interval by finding those null hypotheses that would not be rejected at the 0.05 level.

**Sampling model:**

$$\begin{aligned}
Y_{1,A}, \ldots, Y_{n_A,A} &\sim \text{ i.i.d. normal}(\mu_A, \sigma^2) \\
Y_{1,B}, \ldots, Y_{n_B,B} &\sim \text{ i.i.d. normal}(\mu_B, \sigma^2).
\end{aligned}$$

Consider evaluating whether $\delta$ is a reasonable value for the difference in means:

$$\mathrm{H}_0\text{: } \mu_B - \mu_A = \delta$$
$$\mathrm{H}_1 : \mu_B - \mu_A \neq \delta$$

Under $\text{H}_0$, you should be able to show that

$$\frac{(\bar{Y}_B - \bar{Y}_A) - \delta}{s_p\sqrt{1/n_A + 1/n_B}} \sim t_{n_A+n_B-2}$$

Thus a given difference $\delta$ is *accepted* at level $\alpha$ if

$$\begin{aligned}
\frac{|\bar{y}_B - \bar{y}_A - \delta|}{s_p\sqrt{1/n_A + 1/n_B}} &\leq \; t_c \\
(\bar{y}_B - \bar{y}_A) - s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}t_c \leq \; &\delta \; \leq (\bar{y}_B - \bar{y}_A) + s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}t_c
\end{aligned}$$

where $t_c = t_{1-\alpha/2,n_A+n_B-2}$ is the *critical value.*

**Wheat example:**

- $\bar{y}_B - \bar{y}_A = 5.93$
- $s_p = 4.72$, $s_p\sqrt{1/n_A + 1/n_B} = 2.72$
- $t_{.975,10} = 2.23$

A 95% C.I. for $\mu_B - \mu_A$ is

$$\begin{aligned}
5.93 \;&\pm\; 2.72 \times 2.23 \\
5.93 \;&\pm\; 6.07 = (-0.13, 11.99)
\end{aligned}$$

**Questions:**

- What does the fact that 0 is in the interval say about $\text{H}_0 : \mu_A = \mu_B$?
- What is the interpretation of this interval?
- Could we have constructed an interval via a randomization test?

## 4.2 Power and Sample Size Determination

Suppose that we are designing a study where we intend to gather data on two groups, and will decide if there is a difference between the groups based on the data. Further suppose our decision procedure will be determined by a level-$\alpha$ two-sample $t$-test.

**Two sample experiment and $t$-test:**

- $\mathrm{H}_0$: $\mu_A = \mu_B$ $\mathrm{H}_1$: $\mu_A \neq \mu_B$
- Randomize treatments to the two groups via a CRD.
- Gather data.
- Perform a level $\alpha$ hypothesis test: reject $\mathrm{H}_0$ if

$$|t_{\mathrm{obs}}| \geq t_{1-\alpha/2, n_A+n_B-2}.$$

Recall, if $\alpha = 0.05$ and $n_A, n_B$ are large then $t_{1-\alpha/2,n_A+n_B-2} \approx 2$.

We know that the type I error rate is $\alpha = 0.05$, or more precisely:

$$\Pr(\text{type I error}|\mathrm{H}_0 \text{ true}) = \Pr(\text{reject } \mathrm{H}_0|\mathrm{H}_0 \text{ true}) = 0.05$$

What about

$$\begin{aligned} \Pr(\text{type II error}|\mathrm{H}_0 \text{ false}) &= \Pr(\text{accept } \mathrm{H}_0|\mathrm{H}_0 \text{ false}) \\ &= 1 - \Pr(\text{reject } \mathrm{H}_0|\mathrm{H}_0 \text{ false}) \end{aligned}$$

This is not yet a well-defined problem: there are many different ways in which the null hypothesis may be false, e.g. $\mu_B - \mu_A = 0.0001$ and $\mu_B - \mu_A = 10{,}000$ are both instances of the alternative hypothesis. However, clearly we have

$$\Pr(\text{reject } \mathrm{H}_0|\mu_B - \mu_A = .0001) < \Pr(\text{reject } \mathrm{H}_0|\mu_B - \mu_A = 10{,}000)$$

if the treatment doesn't affect the variance.

To make the question concerning Type II error-rate better defined we need to be able to refer to a *specific* alternative hypothesis. For example, in the case of the two-sample test, for a specific difference $\delta$ we may want to calculate:

$$1 - \Pr(\text{type II error}|\mu_B - \mu_A = \delta) = \Pr(\text{reject } \mathrm{H}_0|\mu_B - \mu_A = \delta).$$

We define the *power* of a two-sample $t$-test test **under a specific alternative** to be:

$$\begin{aligned} \text{Power}(\delta) &= \Pr(\text{reject } \mathrm{H}_0 \mid \mu_B - \mu_A = \delta) \\ &= \Pr\left(|t(\boldsymbol{Y}_A, \boldsymbol{Y}_B)| \geq t_{1-\alpha/2,n_A+n_B-2} \middle| \mu_B - \mu_A = \delta\right). \end{aligned}$$

Remember, the "critical" value $t_{1-\alpha/2,n_A+n_B-2}$ above which we reject the null hypothesis was computed from the null distribution.

However, now we want to work out the probability of getting a value of the $t$-statistic greater than this critical value, **when a specific alternative hypothesis is true**. Thus we need to compute the distribution of our $t$-statistic under the specific alternative hypothesis.

If we suppose $Y_{1,A}, \ldots, Y_{n_A,A} \sim$ i.i.d. normal$(\mu_A, \sigma^2)$ and $Y_{1,B}, \ldots, Y_{n_B,B} \sim$ i.i.d. normal$(\mu_B, \sigma^2)$, where $\mu_B - \mu_A = \delta$ then to calculate the power we need to know the distribution of

$$t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) = \frac{\bar{Y}_B - \bar{Y}_A}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}.$$

We know that if $\mu_B - \mu_A = \delta$ then

$$\frac{\bar{Y}_B - \bar{Y}_A - \delta}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A+n_B-2}$$

but unfortunately

$$t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) = \frac{\bar{Y}_B - \bar{Y}_A - \delta}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} + \frac{\delta}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}. \tag{4.1}$$

The first part in the above equation has a $t$-distribution, which is centered around zero. The second part moves the $t$-statistic away from zero by an amount that depends on the pooled sample variance. For this reason, we call the distribution of the $t$-statistic under $\mu_B - \mu_A = \delta$ the **non-central** $t$-distribution. In this case, we write

$$t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) \sim t^*_{n_A+n_B-2}\underbrace{\left(\frac{\delta}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}\right)}_{\substack{\text{non-centrality}\\ \text{parameter.}}}$$

Note that this distribution is more complicated than just a $t$-distribution plus a constant "shift" away from zero. For the $t$-statistic, the amount of the shift depends on the (random) pooled sample variance.

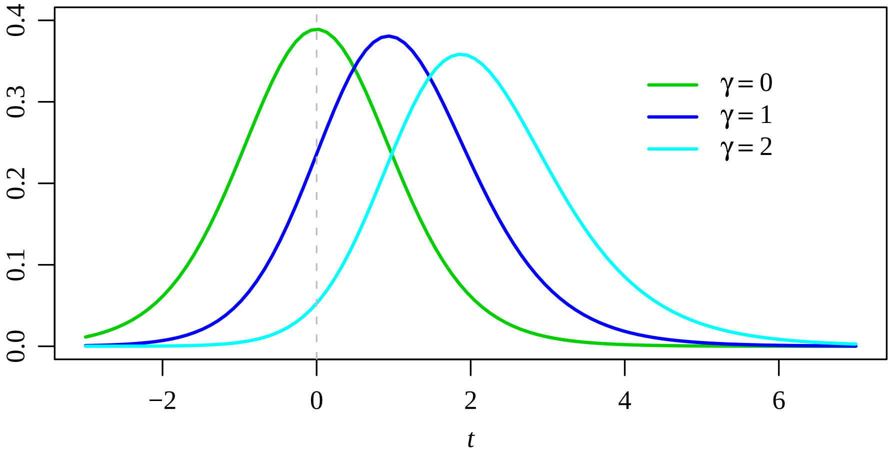


Figure 4.1: A $t_{10}$ distribution and two non-central $t_{10}$-distributions.

### 4.2.1 The non-central $t$-distribution

A noncentral t-distributed random variable can be represented as

$$T = \frac{Z + \gamma}{\sqrt{X/\nu}}$$

where

- $\gamma$ is a constant;
- $Z$ is standard normal;
- $X$ is $\chi^2$ with $\nu$ degrees of freedom, independent of $Z$.

The quantity $\gamma$ is called the noncentrality parameter.

**Exercise:** Using the above representation, show that the distribution of the $t$-statistic is a non-central $t$ distribution, assuming the data are normal and the variance is the same in both groups.

For a non-central $t$-distribution,

- the mean is not zero;

- the distribution is not symmetric.

It can be shown that

$$\mathrm{E}[t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) | \mu_B - \mu_A = \delta] = \frac{\delta}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \times \frac{\sqrt{\frac{\nu}{2}}\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}$$

where $\nu = n_A + n_B - 2$, the degrees of freedom, and $\Gamma(x)$ is the *gamma function*, a generalization of the factorial:

- $\Gamma(n+1) = n!$ if $n$ is an integer
- $\Gamma(r+1) = r\Gamma(r)$
- $\Gamma(1) = 1$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Anyway, you can show that for large $\nu$,

$$\begin{aligned}\sqrt{\frac{\nu}{2}}\Gamma(\frac{\nu-1}{2})/\Gamma(\frac{\nu}{2}) &\approx 1 \text{ so} \\ \mathrm{E}[t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) | \mu_B - \mu_A = \delta] &\approx \frac{\delta}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}\end{aligned}$$

This isn't really such a big surprise, because we know that:

$$\bar{Y}_B - \bar{Y}_A \sim \text{normal}(\delta, \sigma^2[1/n_A + 1/n_B]).$$

Hence

$$\frac{\bar{Y}_B - \bar{Y}_A}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim \text{normal}\left(\frac{\delta}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}, 1\right).$$

We also know that for large values of $n_A, n_B$, we have $s \approx \sigma$, so the noncentral $t$-distribution will (for large enough $n_A$, $n_B$) look approximately normal with

- mean $\delta/(\sigma\sqrt{(1/n_A) + (1/n_B)})$;
- standard deviation 1.

Another way to get the same result is to refer back to the expression for the $t$-statistic given in 4.1:

$$\begin{aligned} t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) &= \frac{\bar{Y}_B - \bar{Y}_A - \delta}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} + \frac{\delta}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \\ &= \quad a_{n_A,n_b} \quad + \quad b_{n_A,n_B} \end{aligned}$$

The first term $a_{n_A,n_B}$ has a $t$-distribution, and becomes standard normal as $n_A, n_B \to \infty$. As for $b_{n_A,n_B}$, since $s_p^2 \to \sigma^2$ as $n_A$ or $n_B \to \infty$, we have

$$\frac{1}{b_{n_A,n_B}} \frac{\delta}{s_p\sqrt{1/n_A + 1/n_B}} \to 1 \text{ as } n_A, n_B \to \infty.$$

### 4.2.2 Computing the Power of a test

Recall our level-$\alpha$ testing procedure using the $t$-test:

1. Sample data, compute $t_{\text{obs}} = t(\boldsymbol{Y}_A, \boldsymbol{Y}_B)$ .
2. Compute the $p$-value, $\Pr(|T_{n_A+n_B-2}| > |t_{\text{obs}}|)$.
3. Reject $\text{H}_0$ if the $p$-value $\leq \alpha \Leftrightarrow |t_{\text{obs}}| \geq t_{1-\alpha/2,n_A+n_B-2}$.

For this procedure, we have shown that

$$\begin{aligned} \Pr(\text{reject } \text{H}_0 | \mu_B - \mu_A = 0) &= \Pr(\ p\text{-value} \leq \alpha | \mu_B - \mu_A = 0) \\ &= \Pr(|t(\boldsymbol{Y}_A, \boldsymbol{Y}_B)| \geq t_{1-\alpha/2,n_A+n_B-2} | \text{H}_0) \\ &= \Pr(|T_{n_A+n_B-2}| \geq t_{1-\alpha/2,n_A+n_B-2}) \\ &= \alpha \end{aligned}$$

But what is the probability of rejection under $\text{H}_\delta : \mu_B - \mu_A = \delta$? Hopefully this is bigger than $\alpha$! Let $t_c = t_{1-\alpha/2,n_A+n_B-2}$, the $1 - \alpha/2$ quantile of a $t$-distribution with $n_A + n_B - 2$ degrees of freedom.

$$\begin{aligned} \Pr(\text{reject } \text{H}_0 \mid \mu_B - \mu_A = \delta) &= \Pr\left(|t(\boldsymbol{Y}_A, \boldsymbol{Y}_B)| > t_c \mid \ \mu_B - \mu_A = \delta\right) \\ &= \Pr\left(|T^*| > t_c\right) \\ &= \Pr\left(T^* > t_c\right) + \Pr\left(T^* < -t_c\right) \\ &= \left[1 - \Pr\left(T^* < t_c\right)\right] \quad + \Pr\left(T^* < -t_c\right) \end{aligned}$$

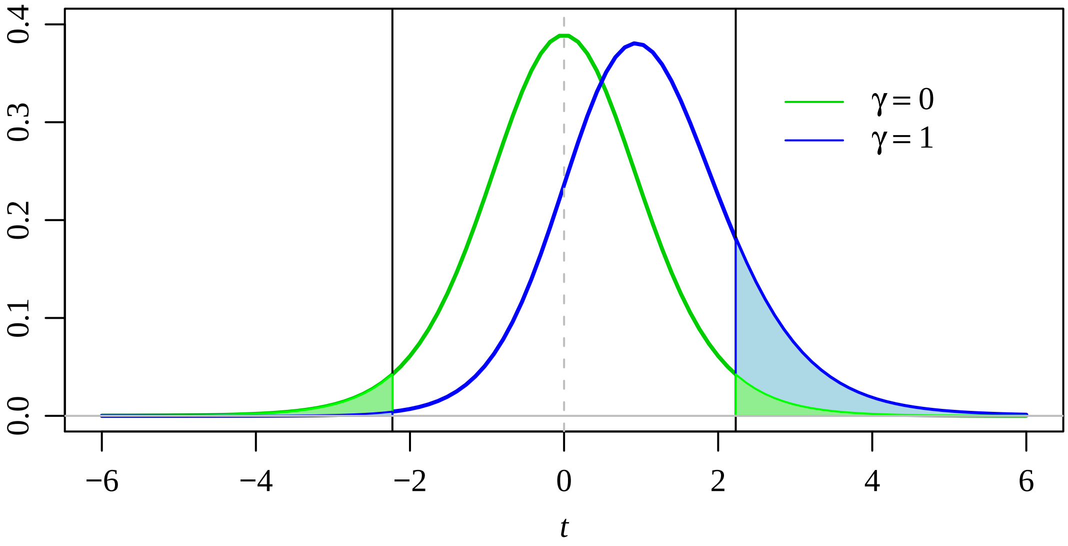


Figure 4.2: Critical regions and the non-central $t$-distribution

where $T^*$ has the non-central $t$-distribution with $= n_A + n_B - 2$ degrees of freedom and non-centrality parameter

$$\gamma = \frac{\delta}{\sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}.$$

We will want to make this calculation in order to see if our sample size is sufficient to have a reasonable chance of rejecting the null hypothesis. If we have a rough idea of $\delta$ and $\sigma^2$ we can evaluate the power using this formula.

```
t.crit <- qt( 1-alpha/2 , nA + nB - 2 )

t.gamma<- delta/( sigma*sqrt(1/nA + 1/nB ))

t.power <- 1- pt( t.crit , nA+nB-2 , ncp=t.gamma ) +
              pt(-t.crit , nA+nB-2 , ncp=t.gamma )
```

When you do these calculations you should think of Figure 4.2. Letting $T^*$ and $T$ be non-central and central $t$-distributed random variables respectively, make sure you can relate the following probabilities to the figure:

- $\Pr(T^* > t_c)$

- $\Pr(T^* < -t_c)$
- $\Pr(T > t_c)$
- $\Pr(T < -t_c)$

Note that if the power $\Pr(|T^*| > t_c)$ is large, then one of $\Pr(T^* > t_c)$ or $\Pr(T^* < -t_c)$ will be very close to zero.

### Approximating the power

Recall that for large $n_A, n_B$,

$$t(\boldsymbol{Y}_A, \boldsymbol{Y}_B) \dot{\sim} \text{normal}(\gamma, 1)$$

The normal approximation to the power is thus given by

$$\Pr(|X| > t_c) = [1 - \Pr(X < t_c)] + \Pr(X < -t_c)$$

where $X \sim$normal$(\gamma, 1)$. This can be computed in R as

```
t.norm.power <- 1- pnorm( t.crit , mean=t.gamma ) +
                   pnorm(-t.crit , mean=t.gamma )
```

This will be a reasonable approximation for large $n_A, n_B$. It may be an over-estimate or under-estimate of the power obtained from the $t$-distribution.

Finally, keep in mind that in our calculations we have assumed that the variances of the two populations are equal.

**Example (selecting a sample size):** Suppose the wheat researchers wish to redo the experiment using a larger sample size. How big should their sample size be if they want to have a good chance of rejecting the null hypothesis $\mu_B - \mu_A = 0$ at level $\alpha = 0.05$, if the true difference in means is $\mu_B - \mu_A = 5$ or more?

> $\mu_B - \mu_A = 5$
>
> $\sigma^2$ is unknown: We'll assume the pooled sample variance from the first experiment is a good approximation: $\sigma^2 = 22.24$.

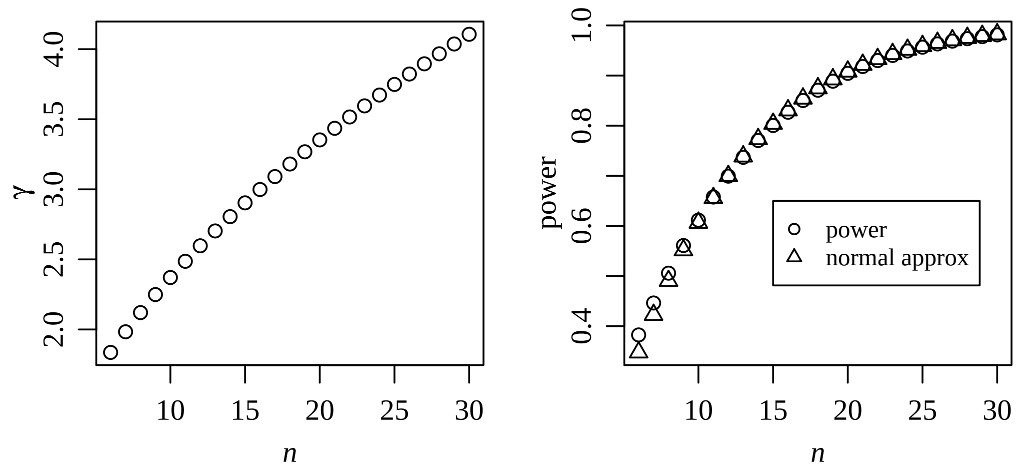


Figure 4.3: $\gamma$ and power versus sample size, and the normal approximation to the power.

Under these conditions, if $n_A = n_B = n$, then

$$\begin{aligned} \gamma &= \frac{\mu_B - \mu_A}{\sigma\sqrt{1/n_A + 1/n_B}} \\ &= \frac{5}{4.72\sqrt{2/n}} = .75\sqrt{n} \end{aligned}$$

What is the probability we'll reject $\mathrm{H}_0$ at level $\alpha = 0.05$ for a given sample size?

```
delta<-5 ; s2<- (  (nA-1)*var(yA) + (nB-1)*var(yB) )/(nA-1+nB-1)

alpha<-0.05 ; n<-seq(6,30)

t.crit<- qt(1-alpha/2,2*n-2)

t.gamma<-delta/sqrt(s2*(1/n+1/n))

t.power<-1-pt( t.crit,2*n-2,ncp=t.gamma)+
           pt(-t.crit,2*n-2,ncp=t.gamma)

t.normal.power<- 1- pnorm( t.crit, mean=t.gamma ) +
                    pnorm(-t.crit, mean=t.gamma )
```

So we see that if the true mean difference were $\mu_B - \mu_A = 5$, then the original study only had about a 40% chance of rejecting $\text{H}_0$. To have an 80% chance or greater, the researchers would need a sample size of 15 for each group.

Note that the true power depends on the unknown true mean difference and true variance (assuming these are equal in the two groups). Even though our *power calculations* were done under potentially inaccurate values of $\mu_B - \mu_A$ and $\sigma^2$, they still give us a sense of the power under various parameter values:

- How is the power affected if the mean difference is bigger? smaller?
- How is the power affected if the variance is bigger? smaller?

**Example (power as a function of the effect):** Suppose a chemical company wants to know if a new procedure $B$ will yield more product than the current procedure $A$. Running experiments comparing $A$ to $B$ are expensive and they are only budgeted to run an experiment with at most 10 observations in each group.

Is running the experiment worthwhile? To assess this we can calculate the power under $n_A = n_B = 10$ for a variety of values of $\mu_B - \mu_A$ and $\sigma$. The first panel plots power as a function of the mean difference for three different values of $\sigma$. From this plot, we can see that if the mean difference is 1 and the variance is 1, then we have almost a 60% chance of rejecting the null hypothesis, although we only have about a 23% chance of doing so if the variance is 9 ($\sigma = 3$).

Because the power varies as the ratio of effect size to the standard deviation, it is often useful to plot power in terms of this ratio. The *scaled effect size* $\theta$, where

$$\theta = (\mu_B - \mu_A)/\sigma,$$

represents the size of the treatment effect scaled by the experimental variability (the standard deviation). The noncentrality parameter is then

$$\gamma = \theta/\sqrt{1/n_A + 1/n_B}.$$

With $n_A = n_B = 10$, we have $\gamma = 2.24 \times \theta$. A plot of power versus $\theta$ for a level-0.05 test appears in the first panel of Figure 4.4. From this we see that $\text{H}_0$ will be rejected with probability 80% or more only if $|\theta|$ is bigger than about 1.33. In other words, for a sample size of 10 in each group, the effect must be at least 1.33 times as big as the standard deviation in order to have an 80% chance of rejecting $\text{H}_0$.

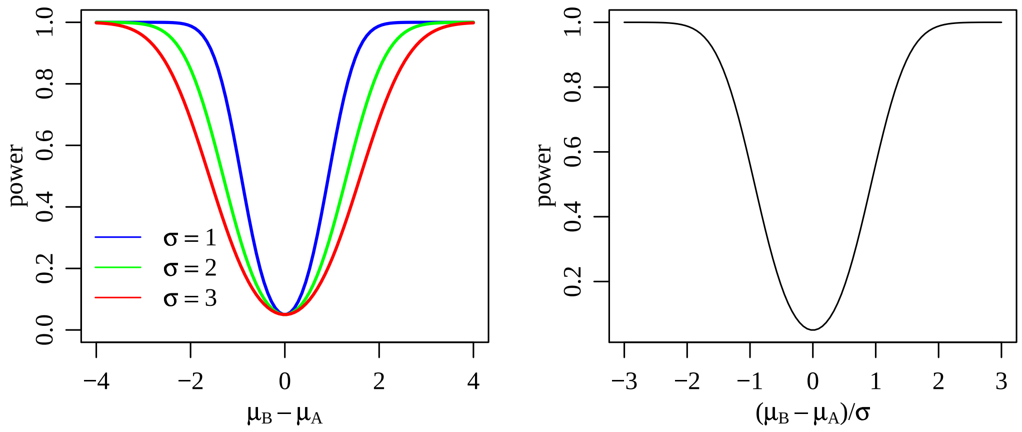


Figure 4.4: Null and alternative distributions for another wheat example, and power versus sample size.

### Increasing power

As we've seen by the normal approximation to the power, for a fixed type I error rate the power is a function of the noncentrality parameter $\gamma$

$$\gamma = \frac{\mu_B - \mu_A}{\sigma\sqrt{1/n_A + 1/n_B}},$$

so clearly power is

- increasing in $|\mu_B - \mu_A|$;
- increasing in $n_A$ and $n_B$;
- decreasing in $\sigma^2$.

The first of these we do not generally control with our experiment (indeed, it is the unknown quantity we are trying to learn about). The second of these, sample size, we clearly do control. The last of these, the variance, seems like something that might be beyond our control. However, the experimental variance can often be reduced by dividing up the experimental material into more homogeneous subgroups of experimental units. This design technique, known as blocking, will be discussed in an upcoming chapter.

# Chapter 5

# Introduction to ANOVA

**Example (Response times):**

Background: Psychologists are interested in how learning methods affect short-term memory.

Hypothesis: Different learning methods may result in different recall time.

Treatments: 5 different learning methods (*A, B, C, D, E*).

Experimental design: (CRD) 20 male undergraduate students were randomly assigned to one of the 5 treatments, so that there are 4 students assigned to each treatment. After a learning period, the students were given cues and asked to recall a set of words. Mean recall time for each student was recorded in seconds.

Results:

| Treatment | sample mean | sample sd |
|---|---|---|
| $A$ | 6.88 | 0.76 |
| $B$ | 5.41 | 0.84 |
| $C$ | 6.59 | 1.37 |
| $D$ | 5.46 | 0.41 |
| $E$ | 5.64 | 0.83 |

Question: Is treatment a source of variation?

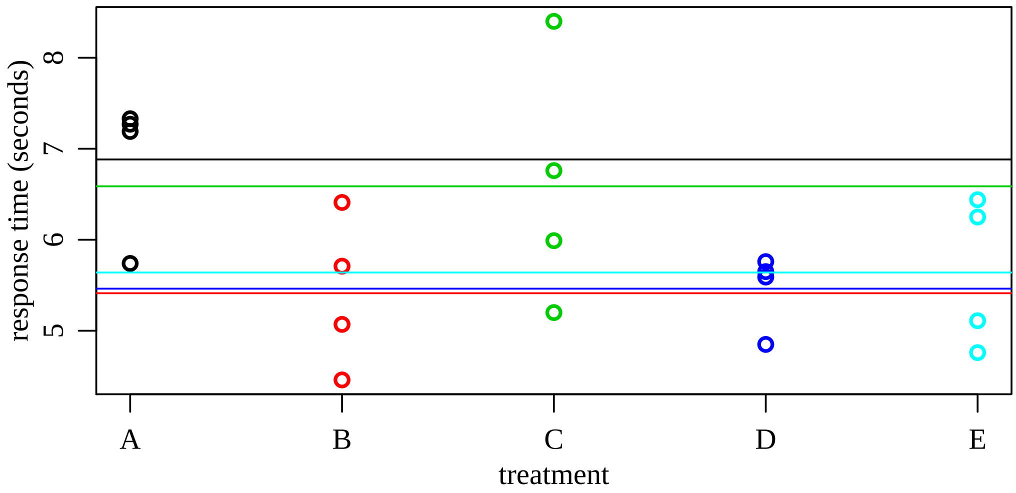


Figure 5.1: Response time data

**Possible data analysis method:** Perform $t$-tests of

$$\mathrm{H}_{0i_1i_2} : \mu_{i_1} = \mu_{i_2} \quad \text{versus} \quad \mathrm{H}_{1i_1i_2} : \mu_{i_1} \neq \mu_{i_2}$$

for each of the $\binom{5}{2} = 10$ possible *pairwise comparisons*. Reject a hypothesis if the associated $p$-value $\leq \alpha$.

**Problem:** If there is no treatment effect at all, then

$$\begin{aligned}
\Pr(\text{reject } \mathrm{H}_{0i_1i_2} | \mathrm{H}_{0i_1i_2} \text{ true}) &= \alpha \\
\Pr(\text{reject any } \mathrm{H}_{0i_1i_2} | \text{ all } \mathrm{H}_{0i_1i+2} \text{ true }) &= 1 - \Pr(\text{ accept all } \mathrm{H}_{0i_1i_2} | \text{ all } \mathrm{H}_{0i_1i_2} \text{ true }) \\
&\approx 1 - \prod_{i<j} (1-\alpha)
\end{aligned}$$

If $\alpha = 0.05$, then

$$\Pr(\text{reject one or more } \mathrm{H}_{0i_1i_2} | \text{ all } \mathrm{H}_{0i_1i_2} \text{ true }) \approx 1 - .95^{10} = 0.40$$

So, even though the *pairwise error rate* is 0.05 the *experiment-wise error rate* is about 0.40. This issue is called the problem of *multiple comparisons* and will be discussed further in Chapter 6. For now, we will discuss a method of testing the **global hypothesis** of **no variation due to treatment**:

$$\mathrm{H}_0 : \mu_{i_1} = \mu_{i_2} \;\; \text{for all } i_1, i_2 \quad \text{versus} \quad \mathrm{H}_1 : \mu_{i_1} \neq \mu_{i_2} \text{ for some } i_1, i_2$$

To do this, we will compare

**treatment variability:** variability across treatments, to

**experimental variability:** variability among experimental units.

First we need to have a way of quantifying these things.

## 5.1 A model for treatment variation

### Data:

$$\begin{aligned} y_{ij} &= \text{measurement from the } j\text{th replicate under the } i\text{th treatment.} \\ i &= 1, \ldots, m \text{ indexes treatments} \\ j &= 1, \ldots, n \text{ indexes observations or replicates.} \end{aligned}$$

### Treatment means model:

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ \mathrm{E}[\epsilon_{ij}] &= 0 \\ \mathrm{Var}[\epsilon_{ij}] &= \sigma^2 \end{aligned}$$

$\mu_i$ is the $i$th treatment mean,

$\epsilon_{ij}$'s represent *within treatment variation*, also known as error or noise.

### Treatment effects model:

$$\begin{aligned} y_{ij} &= \mu + \tau_i + \epsilon_{ij} \\ \mathrm{E}[\epsilon_{ij}] &= 0 \\ \mathrm{Var}[\epsilon_{ij}] &= \sigma^2 \end{aligned}$$

$\mu$ is the *grand mean*;

$\tau_1, \ldots, \tau_m$ are the *treatment effects*, representing *between treatment variation*

$\epsilon_{ij}$'s still represent *within treatment variation.*

In this model, we typically restrict $\sum \tau_i = 0$, otherwise the model is overparameterized.

The **treatment means** and **treatment effects** models represent two *parameterizations* of the same model:

$$\mu_i = \mu + \tau_i \Leftrightarrow \tau_i = \mu_i - \mu$$

**Null (or reduced) model:**

$$\begin{aligned} y_{ij} &= \mu + \epsilon_{ij} \\ \mathrm{E}[\epsilon_{ij}] &= 0 \\ \mathrm{Var}[\epsilon_{ij}] &= \sigma^2 \end{aligned}$$

This is a special case of the above two models with

- $\mu = \mu_1 = \cdots \mu_m$ , or equivalently
- $\tau_i = 0$ for all $i$.

In this model, there is **no variation due to treatment**.

### 5.1.1 Model Fitting

What are good estimates of the parameters? One criteria used to evaluate different values of $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_m\}$ is the *least squares criterion*:

$$\mathrm{SSE}(\boldsymbol{\mu}) = \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \mu_i)^2$$

The value of $\boldsymbol{\mu}$ that minimizes $\mathrm{SSE}(\boldsymbol{\mu})$ is called the *least-squares estimate*, and will be denoted $\hat{\boldsymbol{\mu}}$.

**Estimating treatment means:** SSE is the sum of a bunch of quadratic terms, and so is a convex function of $\boldsymbol{\mu}$. The global minimizer $\hat{\boldsymbol{\mu}}$ satisfies

$\nabla\text{SSE}(\hat{\boldsymbol{\mu}}) = \mathbf{0}$. Taking derivatives, we see that

$$\begin{aligned}
\frac{\partial}{\partial\mu_i}\text{SSE}(\boldsymbol{\mu}) &= \frac{\partial}{\partial\mu_i}\sum_{j=1}^{n}(y_{ij}-\mu_i)^2 \\
&= -2\sum(y_{ij}-\mu_i) \\
&= -2n(\bar{y}_{i\cdot}-\mu_i), \text{ so} \\
\nabla\text{SSE}(\boldsymbol{\mu}) &= -2n(\bar{\boldsymbol{y}}-\boldsymbol{\mu})
\end{aligned}$$

Therefore, the global minimum occurs at $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}} = \{\bar{y}_{1\cdot}, \ldots, \bar{y}_{t\cdot}\}$.

Interestingly, $\text{SSE}(\hat{\boldsymbol{\mu}})$ provides a measure of experimental variability:

$$\text{SSE}(\hat{\boldsymbol{\mu}}) = \sum\sum(y_{ij}-\hat{\mu}_i)^2 = \sum\sum(y_{ij}-\bar{y}_{i\cdot})^2$$

Recall $s_i^2 = \sum(y_{ij}-\bar{y}_{i\cdot})^2/(n-1)$ estimates $\sigma^2$ using data from group $i$. If we have more than one group, we want to pool our estimates to be more precise:

$$\begin{aligned}
s^2 &= \frac{(n-1)s_1^2 + \cdots + (n-1)s_m^2}{(n-1)+\cdots+(n-1)} \\
&= \frac{\sum(y_{1j}-\bar{y}_{1\cdot})^2 + \cdots + \sum(y_{1j}-\bar{y}_{1\cdot})^2}{m(n-1)} \\
&= \frac{\sum\sum(y_{ij}-\hat{\mu}_i)^2}{m(n-1)} \\
&= \frac{\text{SSE}(\hat{\boldsymbol{\mu}})}{m(n-1)} \equiv \text{MSE}
\end{aligned}$$

The values $\hat{\boldsymbol{\mu}}$ and $s^2$ have various interpretations depending on the assumptions we are willing to make. Consider the following assumptions:

A0: Data are independently sampled from their respective populations

A1: Populations have the same variance

A2: Populations are normally distributed

Then

A0 → $\hat{\boldsymbol{\mu}}$ is an unbiased estimator of $\boldsymbol{\mu}$

A0+A1 → $s^2$ is an unbiased estimator of $\sigma^2$

A0+A1+A2 →

$(\hat{\boldsymbol{\mu}}, s^2)$ are the *minimum variance unbiased estimators* of $(\boldsymbol{\mu}, \sigma^2)$
$(\hat{\boldsymbol{\mu}}, \frac{n-1}{n}s^2)$ are the *maximum likelihood estimators* of $(\boldsymbol{\mu}, \sigma^2)$

**Within-treatment variability:**

SSE$(\hat{\boldsymbol{\mu}}) \equiv$ SSE is a measure of within treatment variability:

$$\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2$$

It is the sum of squared deviation of observation values from their group mean.

$$\text{MSE} = \text{SSE}/m(n-1) = \frac{1}{m}(s_1^2 + \cdots + s_m^2)$$

Clearly MSE is a measure of average (or mean) within-treatment variability.

**Between-treatment variability:**

The analogous measure of between treatment variability is

$$\text{SST} = \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = n\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

where

$$\begin{aligned}\bar{y}_{\cdot\cdot} &= \frac{1}{mn}\sum\sum y_{ij}\\ &= \frac{1}{m}(\bar{y}_1 + \cdots + \bar{y}_m)\end{aligned}$$

is the *grand mean* of the sample. We call SST the *treatment sum of squares.* We also define MST = SST$/(m-1)$ as the *treatment mean squares* or *mean squares (due to) treatment.* Notice that MST is simply $n$ times the **sample variance** of the **sample means**:

$$\text{MST} = n \times \left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right]$$

### 5.1.2 Testing hypothesis with MSE and MST

Consider evaluating the null hypothesis of no treatment effect:

$$\mathrm{H}_0 : \{\mu_1, \ldots, \mu_m\} \text{ all equal} \quad \text{versus} \quad \mathrm{H}_1 : \{\mu_1, \ldots, \mu_m\} \text{ not all equal}$$

Note that

$$\{\mu_1, \ldots, \mu_m\} \text{ not all equal } \Leftrightarrow \sum_{i=1}^{m} (\mu_i - \bar{\mu})^2 > 0$$

Probabilistically,

$$\sum_{i=1}^{m} (\mu_i - \bar{\mu})^2 > 0 \Rightarrow \text{ a large } \sum_{i=1}^{m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \text{ will probably be observed.}$$

Inductively,

$$\text{a large } \sum_{i=1^m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \text{ observed } \Rightarrow \sum_{i=1^m} (\mu_i - \bar{\mu})^2 > 0 \text{ is plausible}$$

So a large value of SST or MST gives evidence that there are differences between the true treatment means. But how large is large? We need to know what values of MST to expect under $\mathrm{H}_0$.

**MST under the null:**

Suppose $\mathrm{H}_0 : \mu_1 = \cdots = \mu_m = \mu$ is true . Then

$$\begin{aligned} \mathrm{E}[(\bar{Y}_{i\cdot}] = \mu \quad &\rightarrow \quad \mathrm{E}[\sqrt{n}\bar{Y}_{i\cdot}] = \sqrt{n}\mu \\ \mathrm{Var}[\bar{Y}_{i\cdot}] = \sigma^2/n \quad &\rightarrow \quad \mathrm{Var}[\sqrt{n}\bar{Y}_{i\cdot}] = \sigma^2 \end{aligned}$$

So under the null $\sqrt{n}\bar{Y}_{1\cdot}, \ldots, \sqrt{n}\bar{Y}_{m\cdot}$ are $m$ independent random variables having the same mean and variance. Recall that if $X_1, \ldots, X_n \sim$ i.i.d. $P$, then an unbiased estimate of the variance of population $P$ is given by the sample variance $\sum (X_i - \bar{X})^2/(n-1)$. Therefore,

$$\frac{\sum (\sqrt{n}\bar{Y}_i - \sqrt{n}\bar{Y})^2}{m-1} \text{ is an unbiased estimate of } \mathrm{Var}[\sqrt{n}\bar{Y}_i] = \sigma^2.$$

Notice that

$$\begin{aligned}
\frac{\sum(\sqrt{n}\bar{Y}_i - \sqrt{n}\bar{Y})^2}{m-1} &= \frac{n\sum(\bar{Y}_i - \bar{Y})^2}{m-1} \\
&= \frac{\text{SST}}{m-1} \\
&= \text{MST},
\end{aligned}$$

so $\text{E}[\text{MST}|\text{H}_0] = \sigma^2$.

**MST under an alternative:**
We can show that under a given value of $\boldsymbol{\mu}$,

$$\begin{aligned}
\text{E}[\text{MST}|\boldsymbol{\mu}] &= \sigma^2 + \frac{n\sum_{i=1}^m(\mu_i - \bar{\mu})^2}{m-1} \\
&= \sigma^2 + n\frac{\sum_{i=1}^m \tau_i^2}{m-1} \\
&\equiv \sigma^2 + nv_\tau^2
\end{aligned}$$

So $\text{E}[\text{MST}|\boldsymbol{\mu}] \geq \sigma^2$, with equality only if there is **no variability in treatment means**, i.e. $v_\tau^2 = 0$.

**Expected value of MSE:**
$\text{MSE} = \frac{1}{m}\sum_{i=1}^m s_i^2$, so

$$\begin{aligned}
\text{E}[\text{MSE}] &= \frac{1}{m}\sum \text{E}[s_i^2] \\
&= \frac{1}{m}\sum \sigma^2 = \sigma^2
\end{aligned}$$

Let's summarize our potential estimators of $\sigma^2$:

If $\text{H}_0$ is true:

- $\text{E}[\text{MSE}|\text{H}_0] = \sigma^2$
- $\text{E}[\text{MST}|\text{H}_0] = \sigma^2$

If $\text{H}_0$ is false:

- $\text{E}[\text{MSE}|\text{H}_1] = \sigma^2$

- $\mathrm{E}[(\mathrm{MST}|\mathrm{H}_1] = \sigma^2 + nv_\tau^2$

This should give us an idea for a test statistic:

If $\mathrm{H}_0$ is true:

- $\mathrm{MSE} \approx \sigma^2$
- $\mathrm{MST} \approx \sigma^2$

If $\mathrm{H}_0$ is false

- $\mathrm{MSE} \approx \sigma^2$
- $\mathrm{MST} \approx \sigma^2 + nv_\tau^2 > \sigma^2$

So

under $\mathrm{H}_0$, MST/MSE should be around 1,

under $\mathrm{H}_0^c$, MST/MSE should be bigger than 1.

Thus the *test statistic* $F(\boldsymbol{Y}) = \mathrm{MST}/\mathrm{MSE}$ is sensitive to deviations from the null, and can be used to measure evidence against $\mathrm{H}_0$. Now all we need is a null distribution.

**Example (response times):**

```
ybar.t<-tapply(y,x,mean)
s2.t<-tapply(y,x,var)

SSE<- sum( (n-1)*s2.t )
SST<- n*sum( (ybar.t-mean(y))^2 )

MSE<-SSE/(m*(n-1))
MST<-SST/(m-1)



> SSE
[1] 12.0379
> SST
[1] 7.55032
```

```
> MSE
[1] 0.8025267
> MST
[1] 1.88758
```

It is customary to summarize these calculations in a table:

| Source of variation | Sums of Squares | Mean Squares | $F$-ratio |
|---|---|---|---|
| Treatment | 7.55 | 1.89 | 2.35 |
| Noise | 12.04 | 0.80 | |

The $F$-ratio deviates a bit from 1. Possible explanations:

- $H_1$: The $F$-value is a result of actual differences between treatments.
- $H_0$: The $F$-value is a result of a chance assignment of treatments, eg. the slow students were randomized to $A$, the fast students to $D$.

To evaluate $H_0$ we look at how likely such a treatment assignment is:

**Randomization test:**

```
F.obs<-anova(lm(y~as.factor(x)))$F[1]

> F.obs
[1] 2.352046

$
set.seed(1)
F.null<-NULL
for(nsim in 1:1000)
{
  x.sim<-sample(x)
  F.null<-c(F.null, anova(lm(y~as.factor(x.sim)))$F[1] )
}
$
> mean(F.null>=F.obs)
[1] 0.102
```

The observed **between-group variation** is larger than the observed **within-group variation**, but not larger than the types of $F$-statistics we'd expect to get if the null hypothesis were true.

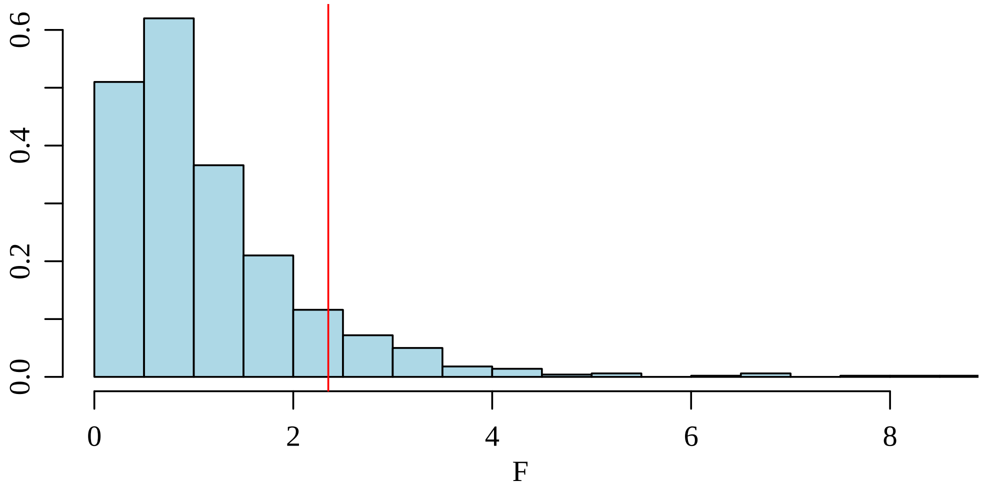


Figure 5.2: Randomization distribution of the $F$-statistic

## 5.2 Partitioning sums of squares

Informally, we can think about the total variation in a datasets as follows:

$$\begin{aligned} \text{Total variability} &= \text{variability due to treatment} &+& \text{variability due to other things} \\ &= \text{between treatment variability} &+& \text{within treatment variability} \end{aligned}$$

This can in fact be made formal:

$$\text{SSTotal} \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \text{SST} + \text{SSE}$$

Proof:

$$\begin{aligned} \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2 &= \sum_{i}\sum_{j}[(y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})]^2 \\ &= \sum_{i}\sum_{j}\left\{(y_{ij} - \bar{y}_{i\cdot})^2 + 2(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right\} \\ &= \sum_{i}\sum_{j}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i}\sum_{j}2(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + \sum_{i}\sum_{j}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\ &= (1) + (2) + (3) \end{aligned}$$

$$\begin{array}{rcl} (1) & = & \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 = \mathrm{SSE} \\ (3) & = & \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = n \sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \mathrm{SST} \\ (2) & = & 2 \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = 2 \sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \sum_j (y_{ij} - \bar{y}_{i\cdot}) \end{array}$$

but note that

$$\begin{aligned} \sum_j (y_{ij} - \bar{y}_{i\cdot}) &= \left( \sum_j y_{ij} \right) - n\bar{y}_{i,\cdot} \\ &= n\bar{y}_{i,\cdot} - n\bar{y}_{i,\cdot} = 0 \end{aligned}$$

for all $j$. Therefore (2) = 0 and we have

$$\begin{array}{rcl} \text{total sum of squared deviations from grand mean} & = & \text{between trt sums of squares +} \\ & & \text{within trt sums of squares} \end{array}$$

or more succinctly,

$$\mathrm{SSTotal} = \mathrm{SST} + \mathrm{SSE}.$$

**Putting it all together:**

$$\begin{array}{lcr} \mathrm{H}_0 : \mu_{i_1} = \mu_{i_2} \text{ for all } i_1, i_2 & \Rightarrow & \textbf{Reduced model} : y_{ij} = \mu + \epsilon_{ij} \\ \mathrm{H}_1 : \mu_{i_1} \neq \mu_{i_2} \text{ for some } i_1 \neq i_2 & \Rightarrow & \textbf{Full model} : y_{ij} = \mu_i + \epsilon_{ij} \end{array}$$

- A *fitted value* or *predicted value* of an observation $y_{ij}$ is denoted $\hat{y}_{ij}$ and represents the modeled value of $y_{ij}$, without the noise.
- A *residual* $\hat{\epsilon}_{ij}$ is the observed value minus the fitted value, $\hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij}$.

If we believe $\mathrm{H}_1$,

- our estimate of $\mu_i$ is $\hat{\mu}_i = \bar{y}_{i\cdot}$
- the fitted value of $y_{ij}$ is $\hat{y}_{ij} = \hat{\mu}_i = \bar{y}_{i\cdot}$
- the residual for $(ij)$ is $\hat{\epsilon}_{ij} = (y_{ij} - \hat{y}_{ij}) = (y_{ij} - \bar{y}_{i\cdot})$.
- the model lack-of-fit is measured by the "sum of squared errors":

$$\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 = \mathrm{SSE}_F$$

If we believe $\mathrm{H}_0$,

- our estimate of $\mu$ is $\hat{\mu} = \bar{y}_{\cdot\cdot}$
- the fitted value of $y_{ij}$ is $\hat{y}_{ij} = \hat{\mu} = \bar{y}_{\cdot\cdot}$
- the residual for $(ij)$ is $\hat{\epsilon}_{ij} = (y_{ij} - \hat{y}_{ij}) = (y_{ij} - \bar{y}_{\cdot\cdot})$.
- the model lack-of-fit in this case is

$$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \mathrm{SSE}_R = \mathrm{SSTotal}$$

The **improvement in model fit** by including treatment parameters is

| | | | | |
|---|---|---|---|---|
| error in reduced model | - | error in full model | = | improvement in model fit |
| SSTotal | - | SSE | = | SST |

**The main idea:** Variance can be **partitioned** into parts representing different *sources*. The variance explained by different sources can be compared and analyzed. This gives rise to the *ANOVA table*.

### 5.2.1 The ANOVA table

**ANOVA = Analysis of Variance**

| Source | Degrees of Freedom | Sums of Squares | Mean Squares | F-ratio |
|---|---|---|---|---|
| Treatment | $m-1$ | SST | MST=SST$/(m-1)$ | MST/MSE |
| Noise | $m(n-1)$ | SSE | MSE=SSE$/m(n-1)$ | |
| Total | $mn-1$ | SSTotal | | |

**Reaction time example:**

| Source of variation | Degrees of Freedom | Sums of Squares | Mean Squares | F-ratio |
|---|---|---|---|---|
| Treatment | 4 | 7.55 | 1.89 | 2.352 |
| Noise | 15 | 12.04 | 0.80 | |
| Total | 19 | 19.59 | | |

**Uses of the ANOVA table:**

- No model assumptions → table gives a descriptive decomposition of variation.
- Assuming $\text{Var}[Y_{ij}] = \sigma^2$ in all groups →
  - $\text{E}[\text{MSE}] = \sigma^2$
  - $\text{E}[\text{MST}] = \sigma^2 + nv_\tau^2$, where $v_\tau^2$ = variance of true group means.
- Assuming treatments were randomly assigned → hypothesis tests, $p$-values.
- Assuming data are random samples from normal population → hypothesis tests, $p$-values, confidence intervals for parameters, power calculations.

### 5.2.2 Understanding Degrees of Freedom:

Consider an experiment with $m$ treatments/groups :

$$\left.\begin{array}{cc} \underline{\text{ Data}} & \underline{\text{Group means}} \\ y_{11}\cdots,y_{1n} & \bar{y}_{1\cdot} \\ y_{21}\cdots,y_{2n} & \bar{y}_{2\cdot} \\ \vdots & \vdots \\ y_{m1}\cdots,y_{mn} & \bar{y}_{m\cdot} \end{array}\right\} \bar{y} = \frac{n\bar{y}_1 + \cdots + n\bar{y}_m}{mn} = \frac{\bar{y}_1 + \cdots + \bar{y}_m}{m}$$

We can “decompose” each observation as follows:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

This leads to

$$\begin{array}{ccccc} (y_{ij} - \bar{y}) & = & (\bar{y}_i - \bar{y}) & + & (y_{ij} - \bar{y}_i) \\ \text{total variation} & = & \text{between group variation} & + & \text{within group variation} \end{array}$$

All data can be decomposed this way, leading to the decomposition of the data vector of length $m \times n$ into two parts, as shown in Table 5.1. How do we interpret the degrees of freedom? We’ve heard of degrees of freedom before, in the definition of a $\chi^2$ random variable:

| Total | | Treatment | | Error |
|---|---|---|---|---|
| $y_{11} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{1.} - \bar{y}_{..})$ | $+$ | $(y_{11} - \bar{y}_{1.})$ |
| $y_{12} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{1.} - \bar{y}_{..})$ | $+$ | $(y_{12} - \bar{y}_{1.})$ |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| $y_{1n} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{1.} - \bar{y}_{..})$ | $+$ | $(y_{1n} - \bar{y}_{1.})$ |
| $y_{21} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{2.} - \bar{y}_{..})$ | $+$ | $(y_{21} - \bar{y}_{2.})$ |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| $y_{2n} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{2.} - \bar{y}_{..})$ | $+$ | $(y_{2n} - \bar{y}_{2.})$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ |
| $y_{m1} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{m.} - \bar{y}_{..})$ | $+$ | $(y_{m1} - \bar{y}_{m.})$ |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| $y_{mn} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{m.} - \bar{y}_{..})$ | $+$ | $(y_{mn} - \bar{y}_{m.})$ |
| | | | | |
| SSTotal | $=$ | SSTrt | $+$ | SSE |
| $mn - 1$ | $=$ | $m - 1$ | $+$ | $m(n - 1)$ |

Table 5.1: ANOVA decomposition

dof = number of statistically independent elements in a vector

In the ANOVA table, the dof have a geometric interpretation:

dof = number of components of a vector that can vary independently

To understand this latest definition, consider $x_1, x_2, x_3$ and $\bar{x} = (x_1 + x_2 + x_3)/3$:

$$\begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

How many degrees of freedom does the vector $(c_1, c_2, c_3)^T$ have? How many components can vary independently, if we know the elements are equal to some numbers minus the average of the numbers?

$$\begin{aligned} c_1 + c_2 + c_3 &= x_1 - \bar{x} + x_2 - \bar{x} - x_3 - \bar{x} \\ &= (x_1 + x_2 + x_3) - 3\bar{x} \\ &= 3\bar{x} - 3\bar{x} \\ &= 0 \end{aligned}$$

Thus we must have $c_1 + c_2 = -c_3$, and so $c_1, c_2, c_3$ can't all be independently varied, only two at a time can be arbitrarily changed. This vector thus lies in a two-dimensional subspace of $\mathbb{R}^3$, and has 2 degrees of freedom. In general,

$$\begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_m - \bar{x} \end{pmatrix}$$

is an $m$-dimensional vector in an $m - 1$ dimensional subspace, having $m - 1$ degrees of freedom.

**Exercise:** Return to the vector decomposition of the data and obtain the degrees of freedom of each component. Note that

- dof = dimension of the space the vector lies in
- SS = squared length of the vector

We will soon see the relationship between the geometric interpretation of degrees of freedom and the interpretation involving $\chi^2$ random variables.

### 5.2.3 More sums of squares geometry

Consider the following vectors:

$$\mathbf{y} = \begin{pmatrix} \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} y_{m1} \\ \vdots \\ y_{mn} \end{pmatrix} \end{pmatrix} \quad \bar{\mathbf{y}}_{\text{trt}} = \begin{pmatrix} \begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{1\cdot} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \bar{y}_{m\cdot} \\ \vdots \\ \bar{y}_{m\cdot} \end{pmatrix} \end{pmatrix} \quad \bar{\mathbf{y}}_{\cdot\cdot} = \mathbf{1}\bar{y}_{\cdot\cdot} = \begin{pmatrix} \begin{pmatrix} \bar{y}_{\cdot\cdot} \\ \vdots \\ \bar{y}_{\cdot\cdot} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \bar{y}_{\cdot\cdot} \\ \vdots \\ \bar{y}_{\cdot\cdot} \end{pmatrix} \end{pmatrix}$$

We can express our decomposition of the data as follows:

$$\begin{array}{ccccc} (\mathbf{y} - \bar{\mathbf{y}}_{\cdot\cdot}) & = & (\bar{\mathbf{y}}_{\text{trt}} - \bar{\mathbf{y}}_{\cdot\cdot}) & + & (\mathbf{y} - \bar{\mathbf{y}}_{\text{trt}}) \\ \mathbf{a} & = & \mathbf{b} & + & \mathbf{c} \end{array}$$

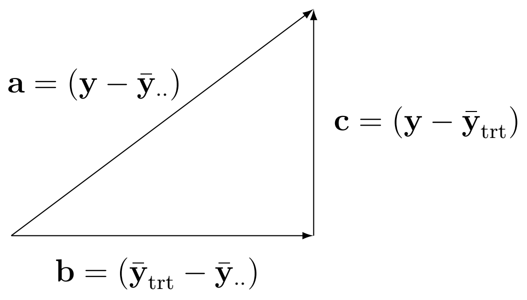
This is just vector addition/subtraction on vectors of length $mn$. Recall that two vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal/perpendicular/at right angles if

$$\mathbf{u} \cdot \mathbf{v} \equiv \sum_{i=1}^{m} u_i v_i = 0$$

We have already shown that $\mathbf{b}$ and $\mathbf{c}$ are orthogonal:

$$\begin{aligned} \mathbf{b} \cdot \mathbf{c} &= \sum_{k=1}^{mn} b_l c_l \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{i\cdot}) \\ &= \sum_{i=1}^{m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot}) \\ &= \sum_{i=1}^{m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \times 0 \\ &= 0 \end{aligned}$$

So the vector $\mathbf{a}$ is the vector sum of two orthogonal vectors. We can draw this as follows:


$\mathbf{a} = (\mathbf{y} - \bar{\mathbf{y}}_{\cdot\cdot})$
$\mathbf{c} = (\mathbf{y} - \bar{\mathbf{y}}_{\text{trt}})$
$\mathbf{b} = (\bar{\mathbf{y}}_{\text{trt}} - \bar{\mathbf{y}}_{\cdot\cdot})$

Now recall

- $||\mathbf{a}||^2 = ||\mathbf{y} - \bar{\mathbf{y}}_{\cdot\cdot}||^2 = \text{SSTotal}$
- $||\mathbf{b}||^2 = ||\bar{\mathbf{y}}_{\text{trt}} - \bar{\mathbf{y}}_{\cdot\cdot}||^2 = \text{SST}$
- $||\mathbf{c}||^2 = ||\mathbf{y} - \bar{\mathbf{y}}_{\text{trt}}||^2 = \text{SSE}$

What do we know about right triangles?

$$\begin{array}{ccccc} ||\mathbf{a}||^2 & = & ||\mathbf{b}||^2 & + & ||\mathbf{c}||^2 \\ \text{SSTotal} & = & \text{SST} & + & \text{SSE} \end{array}$$

So the ANOVA decomposition is an application of Pythagoras' Theorem.

One final observation: recall that

- dof $(\bar{\mathbf{y}}_{\text{trt}} - \bar{\mathbf{y}}_{\cdot\cdot}) = m - 1$
- dof $(\mathbf{y} - \bar{\mathbf{y}}_{\text{trt}}) = m(n-1)$
- $(\bar{\mathbf{y}}_{\text{trt}} - \bar{\mathbf{y}}_{\cdot\cdot})$ and $(\mathbf{y} - \bar{\mathbf{y}}_{\text{trt}})$ are orthogonal.

The last lines means the degrees of freedom must add, so

$$\text{dof}(\mathbf{y} - \bar{\mathbf{y}}_{\cdot\cdot}) = (m-1) + m(n-1) = mn - 1$$

## 5.3 Unbalanced Designs

If the number of *replications* is constant for all *levels* of the treatment/factor then the design is called *balanced*. Otherwise it is said to be *unbalanced*.

**Unbalanced data:**

- $y_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$
- let $N = \sum_{i=1}^{m} n_i$ be the total sample size.

How does the ANOVA decomposition work in this case? What are the parameter estimates for the full and reduced model?

**Null model:**

$$y_{ij} = \mu + \epsilon_{ij} \quad \mathrm{Var}[\epsilon_{ij}] = \sigma^2$$

You should be able to show that the least-squares estimators are

- $\hat{\mu} = \bar{y}_{\cdot\cdot} = \frac{1}{N} \sum y_{ij}$
- $s^2 = \frac{1}{N-1} \sum (y_{ij} - \bar{y}_{\cdot\cdot})^2$

**Full model:**

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \tau_i + \epsilon_{ij} \\ \mathrm{Var}[\epsilon_{ij}] &= \sigma^2 \end{aligned}$$

How should we estimate these parameters? When $n_i = n$ for all $i$, we had

Treatment means parameterization: $\hat{\mu}_i = \bar{y}_{i\cdot}$

Treatment effects parameterization: $\hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\tau}_i = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$

which meant that $\bar{y}_{\cdot\cdot} = \frac{1}{m} \bar{y}_{i\cdot}$. Similarly, we had

$$s^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2}{\sum_{i=1}^{m} (n-1)}$$

which implied that $s^2 = \frac{1}{m}\sum s_i^2$. However, if $n_{i_1} \neq n_{i_2}$, then in general

$$\frac{1}{m}\sum_{i=1}^{m} \bar{y}_{i\cdot} \neq \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}, \text{ and } \frac{1}{m}\sum_{i=1}^{m} s_i^2 \neq \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}{\sum_{i=1}^{m}(n_i-1)}$$

What should the parameter estimates be? With a bit of calculus you can show that the least squares estimates of $\mu_i$ or $(\mu, \tau_i)$ are

- $\hat{\mu}_i = \bar{y}_i$
- $\hat{\tau}_i = \bar{y}_i - \bar{y}_{\cdot\cdot}$, $\hat{\mu} = \bar{y}_{\cdot\cdot}$

We no longer have $\hat{\mu} = \frac{1}{m}\sum \mu_i$, or $\sum \hat{\tau}_i = 0$, but we do have

$$\begin{aligned}
\frac{\sum_{i=1}^{m} n_i \bar{y}_i}{\sum n_i} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}/N \\
&= \bar{y}_{\cdot\cdot}, \text{ so} \\
\sum n_i \hat{\mu}_i / \sum n_i &= \hat{\mu}, \text{ and} \\
\sum n_i \hat{\tau}_i &= 0.
\end{aligned}$$

So $\hat{\mu}$ is a *weighted average* of the $\hat{\mu}_i$'s, and a weighted average of the $\hat{\tau}_i$'s is zero. Similarly,

$$s^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}{\sum_{i=1}^{m}(n_i-1)} = \frac{\sum_{i=1}^{m}(n_i-1)s_i^2}{\sum_{i=1}^{m}(n_i-1)}$$

so $s^2$ is a weighted average of the $s_i^2$'s.

### 5.3.1 Sums of squares and degrees of freedom

The vector decomposition is shown in table 5.2. Let **a**, **b** and **c** be the three vectors in the table. We define the sums of squares as the squared lengths of these vectors:

- SSTotal $= ||\mathbf{a}||^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$
- SSTrt $= ||\mathbf{b}||^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{m} n_i(y_{i\cdot} - \bar{y}_{\cdot\cdot})^2$
- SSE $= ||\mathbf{c}||^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$

| Total | | Treatment | | Error |
|---|---|---|---|---|
| $y_{11} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{1.} - \bar{y}_{..})$ | $+$ | $(y_{11} - \bar{y}_{1.})$ |
| $y_{12} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{1.} - \bar{y}_{..})$ | $+$ | $(y_{12} - \bar{y}_{1.})$ |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| $y_{1n_1} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{1.} - \bar{y}_{..})$ | $+$ | $(y_{1n_1} - \bar{y}_{1.})$ |
| $y_{21} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{2.} - \bar{y}_{..})$ | $+$ | $(y_{21} - \bar{y}_{2.})$ |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| $y_{2n_2} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{2.} - \bar{y}_{..})$ | $+$ | $(y_{2n_2} - \bar{y}_{2.})$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ |
| $y_{m1} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{m.} - \bar{y}_{..})$ | $+$ | $(y_{m1} - \bar{y}_{m.})$ |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| . | $=$ | . | $+$ | . |
| $y_{mn_m} - \bar{y}_{..}$ | $=$ | $(\bar{y}_{m.} - \bar{y}_{..})$ | $+$ | $(y_{mn_m} - \bar{y}_{t.})$ |
| SSTotal | $=$ | SSTrt | $+$ | SSE |
| $N - 1$ | $=$ | $m - 1$ | $+$ | $\sum_{i=1}^{m}(n_i - 1)$ |

Table 5.2: ANOVA decomposition, unbalanced case

Lets see if things add in a nice way. First, lets check orthogonality:

$$\begin{aligned}
\mathbf{b}\cdot\mathbf{c} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})(y_{ij}-\bar{y}_{i\cdot}) \\
&= \sum_{i=1}^{m}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot}) \\
&= \sum_{i=1}^{m}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\times 0 = 0
\end{aligned}$$

Note: **c** must be orthogonal to **any** vector that is constant within a group. So we have

$$\left.\begin{array}{c}\mathbf{a}=\mathbf{b}+\mathbf{c}\\ \mathbf{b}\cdot\mathbf{c}=0\end{array}\right\}\rightarrow ||\mathbf{a}||^2=||\mathbf{b}||^2+||\mathbf{c}||^2$$

and so SSTotal = SST + SSE as before. What about degrees of freedom?

- dof($\mathbf{c}$) $= \sum_{i=1}^{m}(n_i-1) = N-m$ should be clear
- dof($\mathbf{a}$) $= N-1$ should be clear
- dof($\mathbf{b}$) = ?

Note that in general,

$$\frac{1}{m}\sum_{i=1}^{m}\bar{y}_{i\cdot}\neq\bar{y}_{\cdot\cdot},$$

But in the vector **b** we have $n_i$ copies of $(\bar{y}_{\cdot}-\bar{y}_{\cdot\cdot})$, and

$$\frac{1}{\sum n_i}\sum n_i\bar{y}_{i\cdot}=\bar{y}_{\cdot\cdot}$$

and so the vector **b does** sum to zero. Another way of looking at it is that the vector **b** is made up of $m$ numbers, which don't sum to zero, but their **weighted average** sums to zero, and so the degrees of freedom are $m-1$.

### 5.3.2 ANOVA table for unbalanced data:

| Source | Deg. of Freedom | Sum of Squares | Mean Square | F-Ratio |
|---|---|---|---|---|
| Treatment | $m-1$ | SST | MST=$\frac{\text{SST}}{m-1}$ | MST/MSE |
| Noise | $N-m$ | SSE | MSE = $\frac{\text{SSE}}{N-m}$ | |
| Total | $N-1$ | SSTotal | | |

Now suppose the following model is correct:

$$y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{Var}[\epsilon_{ij}] = \sigma^2$$

Does MSE still estimate $\sigma^2$?

$$\begin{aligned}
\text{MSE} &= \text{SSE}/(N-m) \\
&= \frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\cdot})^2 + \cdots \sum_{j=1}^{n_m}(y_{mj}-\bar{y}_{m\cdot})^2}{(n_1-1)+\cdots+(n_m-1)} \\
&= \frac{(n_1-1)s_1^2+\cdots+(n_m-1)s_m^2}{(n_1-1)+\cdots+(n_m-1)}
\end{aligned}$$

So MSE is a weighted average of a bunch of unbiased estimates of $\sigma^2$, so it is still unbiased.

Is the $F$-statistic still sensitive to deviations from $\text{H}_0$? Note that a group with more observations contributes more to the grand mean, but it also contributes more terms to the SST. One can show

- $\text{E[MSE]} = \sigma^2$
- $\text{E[MST]} = \sigma^2 + \frac{N}{m-1}v_\tau^2$ , where
  - $v_\tau^2 = \frac{\sum_{i=1}^m n_i\tau_i^2}{N}$
  - $\tau_i = \mu_i - \bar{\mu}$
  - $\mu = \frac{\sum_{i=1}^m n_i\mu_i}{\sum n_i}$.

So yes, MST/MSE will still be sensitive to deviations from the null, but the groups with larger sample sizes have a bigger impact on the power.

## 5.4 Normal sampling theory for ANOVA

**Example (Blood coagulation):** From a large population of farm animals, 24 animals were selected and randomly assigned to one of four different diets $A, B, C, D$. However, due to resource constraints, $r_A = 4, r_B = 6, r_C = 6, r_D = 8$.

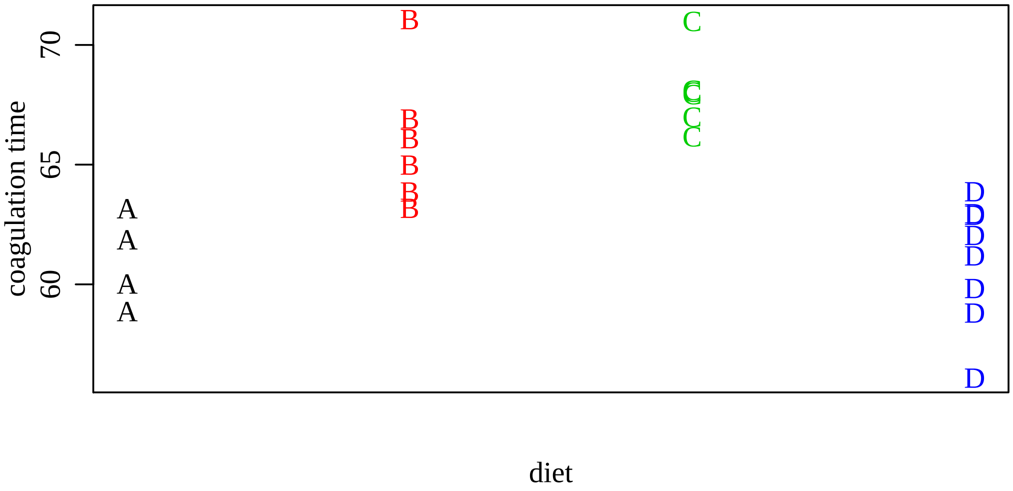


Figure 5.3: Coagulation data

**Questions:**

- Does diet have an effect on coagulation time?
- If a given diet were assigned to all the animals in the population, what would the distribution of coagulation times be?
- If there is a diet effect, how do the mean coagulation times differ?

The first question we can address with a randomization test. For the second and third we need a **sampling model**:

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ \epsilon_{11} \ldots \epsilon_{mn_m} &\sim \text{i.i.d. normal}(0, \sigma^2) \end{aligned}$$

This model implies

- independence of errors
- constant variance
- normally distributed data

Another way to write it is as follows:

$$y_{A1}, \ldots, y_{A4} \sim \text{i.i.d. normal}(\mu_A, \sigma^2)$$

$$y_{B1}, \ldots, y_{B6} \sim \text{i.i.d. normal}(\mu_B, \sigma^2)$$

$$y_{C1}, \ldots, y_{C6} \sim \text{i.i.d. normal}(\mu_C, \sigma^2)$$

$$y_{D1}, \ldots, y_{D8} \sim \text{i.i.d. normal}(\mu_D, \sigma^2)$$

So we are viewing the 4 samples under $A$ as a *random sample* from the population of coagulation times that would be present if all animals got $A$ (and similarly for samples under $B$, $C$ and $D$).

```
> anova(lm(ctime~diet))
Analysis of Variance Table

Response: ctime
          Df Sum Sq Mean Sq F value
diet       3  228.0    76.0  13.571
Residuals 20  112.0     5.6
```

The $F$-statistic is large, but how unlikely is it under $H_0$?

Two viewpoints on null distributions:

- The 24 animals we selected are fixed. Other possible outcomes of the experiment correspond to **different assignments of the treatments**.
- The 24 animals we selected are random samples from a larger population of animals. Other possible outcomes of the experiment correspond to **different experimental units**.

Randomization tests correspond to the first viewpoint, sampling theory to the second.

### 5.4.1 Sampling distribution of the $F$-statistic

Recall the $\chi^2$ distribution:

$$Y_1, \ldots, Y_n \sim \text{i.i.d. normal}(\mu, \sigma^2) \Rightarrow \frac{1}{\sigma^2}\sum(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$$

Also,

$$\left.\begin{array}{c} X_1 \sim \chi^2_{k_1} \\ X_2 \sim \chi^2_{k_2} \\ X_1, X_2 \text{ independent} \end{array}\right\} \Rightarrow X_1 + X_2 \sim \chi^2_{k_1+k_2}$$

**Distribution of SSE:**

$$\begin{array}{ccccccc} \frac{\sum\sum(Y_{ij}-\bar{Y}_{i\cdot})^2}{\sigma^2} & = & \frac{1}{\sigma^2}\sum(Y_{1j} - \bar{Y}_{1\cdot})^2 & + & \cdots & + & \frac{1}{\sigma^2}\sum(Y_{mj} - \bar{Y}_{m\cdot})^2 \\ & \sim & \chi^2_{n_1-1} & + & \cdots & + & \chi^2_{n_m-1} \\ & \sim & \chi^2_{N-m} & & & & \end{array}$$

So $\text{SSE}/\sigma^2 \sim \chi^2_{N-m}$.

**Distribution of SST under the null:** Under $\text{H}_0$,

$$\begin{aligned} \bar{Y}_i &\sim \text{normal}(\mu, \sigma^2/n_i) \\ \sqrt{n_i}\bar{Y}_i &\sim \text{normal}(n_i\mu, \sigma^2) \\ \frac{1}{\sigma^2}\text{SST} &= \frac{1}{\sigma^2}\sum n_i(\bar{Y}_i - \bar{Y}_{\cdot\cdot})^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{m}(\sqrt{n}_i\bar{Y}_i - \sqrt{n_i}\bar{Y}_{\cdot\cdot})^2 \sim \chi^2_{m-1} \end{aligned}$$

**Results so far:**

- $\text{SSE}/\sigma^2 \sim \chi^2_{N-m}$
- $\text{SST}/\sigma^2 \sim \chi^2_{m-1}$
- SSE, SST independent (why?)

**Introducing the $F$-distribution:** If

$$\left.\begin{array}{c} X_1 \sim \chi^2_{k_1} \\ X_2 \sim \chi^2_{k_2} \\ X_1 \perp X_2 \end{array}\right\} \Rightarrow \frac{X_1/k_1}{X_2/k_2} \sim F_{k_1,k_2}$$

$F_{k_1,k_2}$ is the "$F$-distribution with $k_1$ and $k_2$ degrees of freedom."

**Application:** Under $\mathrm{H}_0$

$$\frac{\left(\frac{\mathrm{SST}}{\sigma^2}\right)/(m-1)}{\left(\frac{\mathrm{SSE}}{\sigma^2}\right)/(N-m)} = \frac{MST}{MSE} \sim F_{m-1,N-m}$$

A large value of $F$ is evidence against $\mathrm{H}_0$, so reject $\mathrm{H}_0$ if $F > F_{\mathrm{crit}}$. How to determine $F_{\mathrm{crit}}$?

**Level-$\alpha$ testing procedure:**

1. gather data

2. construct ANOVA table

3. reject "$\mathrm{H}_0 : \mu_i = \mu$ for all $i$" if $F > F_{\mathrm{crit}}$

where $F_{\mathrm{crit}}$ is the $1-\alpha$ quantile of an $F_{t-1,N-t}$ distribution, available in R via `qf(1−alpha, dof.trt, dof.err)`. Under this procedure (and a host of assumptions),

$$\Pr(\text{reject } \mathrm{H}_0 | \mathrm{H}_0 \text{ true}) = \alpha$$

Plots of several different $F$-distributions appear in Figure 5.4. Study these plots until you understand the relationship between the shape of the curves and the degrees of freedom. Now let's get back to the data analysis:

```
> anova(lm(ctime~diet))
Analysis of Variance Table

Response: ctime
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3  228.0    76.0  13.571 4.658e-05 ***
Residuals 20  112.0     5.6
```

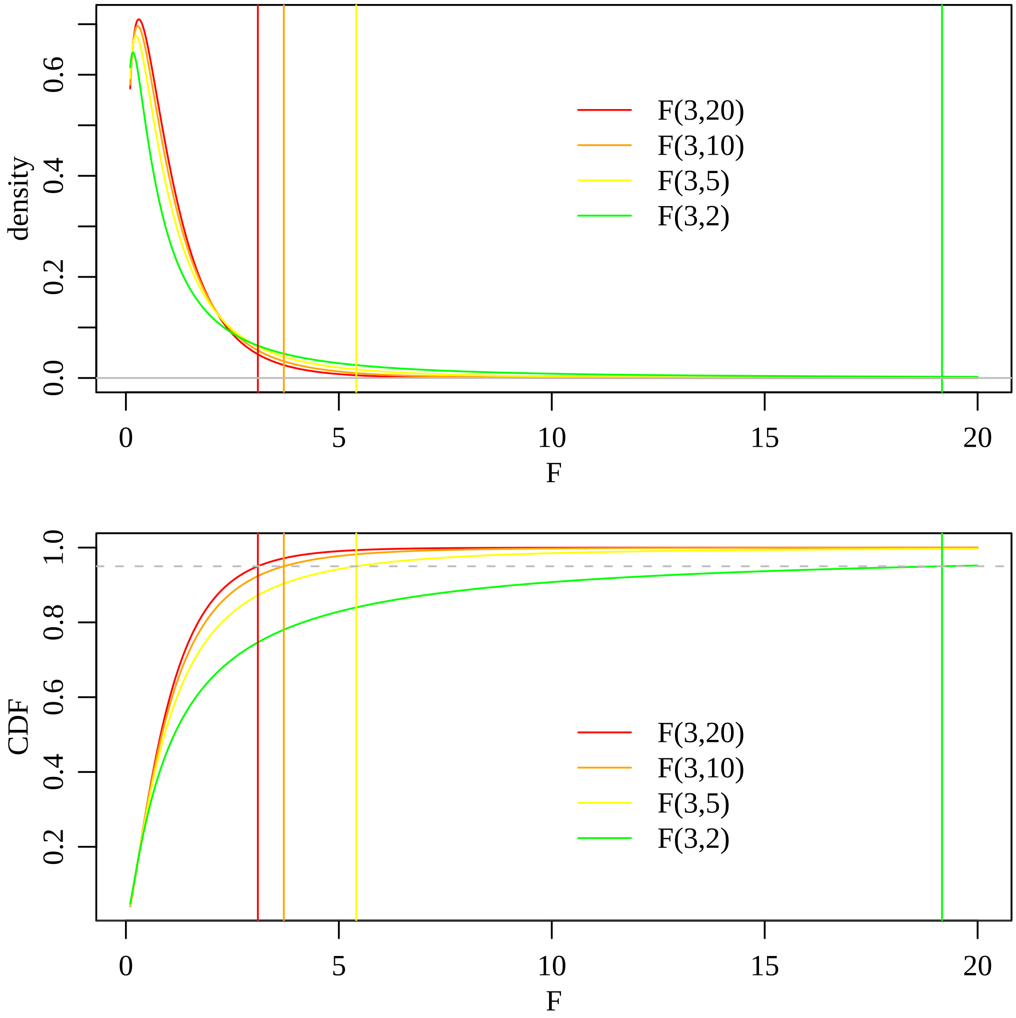


Figure 5.4: F-distributions

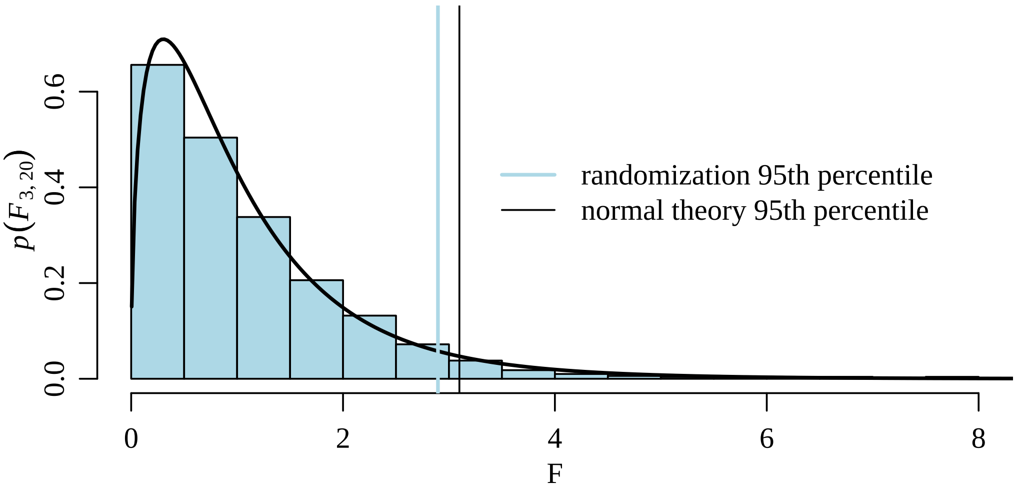


Figure 5.5: Normal-theory and randomization distributions of the $F$-statistic

```
Fobs<-anova(lm(ctime~diet))$F[1]
Fsim<-NULL
for(nsim in 1:1000) {
diet.sim<-sample(diet)
Fsim<-c(Fsim, anova(lm(ctime~diet.sim))$F[1] )
                         }

> mean(Fsim>=Fobs)
[1] 0

> 1-pf(Fobs,3,20)
[1] 4.658471e-05
```

### 5.4.2 Comparing group means

If $H_0$ is rejected, there is evidence that some population means are different from others. We can explore this further by making treatment comparisons. If $H_0$ is rejected we

- estimate $\mu_i$ with $\bar{y}_i$
- estimate $\sigma_i^2$ with
  - $s_i^2$ : if variances are very unequal, this might be a better estimate.

- MSE : if variances are close and $n_i$ is small, this is generally a better estimate.

Standard practice: Unless there is strong evidence to the contrary, we typically assume $\text{Var}[Y_{ij}] = \text{Var}[Y_{kl}] = \sigma^2$, and use $s^2 \equiv \text{MSE}$ to estimate $\sigma^2$. In this case,

$$\begin{aligned} \text{Var}[\hat{\mu}_i] = \text{Var}[\bar{Y}_{i\cdot}] &= \sigma^2/n_i \\ &\approx s^2/n_i \end{aligned}$$

The "standard error of the mean" = $\text{SE}[\hat{\mu}_i] = \sqrt{s^2/n_i}$ , is an **estimate** of $\text{SD}[\hat{\mu}_i] = \sqrt{\text{Var}[\bar{Y}_{i\cdot}]} = \sigma/\sqrt{n_i}$. The standard error is a very useful quantity.

**Standard error:** The usual definition of the standard error of an estimator $\hat{\theta}$ of a parameter $\theta$ is an **estimate** of its sampling standard deviation:

$$\begin{aligned} \hat{\theta} &= \hat{\theta}(\boldsymbol{Y}) \\ \text{Var}[\hat{\theta}] &= \gamma^2 \\ \widehat{\text{Var}[\hat{\theta}]} &= \hat{\gamma}^2 \\ \text{SE}[\hat{\theta}] &= \hat{\gamma} \end{aligned}$$

where $\hat{\gamma}^2$ is an estimate of $\gamma^2$. For example,

$$\begin{aligned} \hat{\mu}_i &= \bar{Y}_{i\cdot} \\ \text{Var}[\hat{\mu}_i] &= \sigma^2/n_i \\ \widehat{\text{Var}[\hat{\mu}_i]} &= \hat{\sigma}^2/n_i = s^2/n_i \\ \text{SE}[\hat{\mu}_i] &= s/\sqrt{n_i} \end{aligned}$$

**Confidence intervals for treatment means:** Obtaining confidence intervals is very similar to the one-sample case. The only difference is that we use data from all of the groups to estimate the variance. As a result, the degrees of freedom changes.

$$\begin{aligned} \frac{\bar{Y}_{i\cdot} - \mu_i}{\text{SE}[\bar{Y}_{i\cdot}]} = \frac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{\text{MSE}/n_i}} &= \frac{\bar{Y}_{i,\cdot} - \mu_i}{s/\sqrt{n_i}} \\ &\sim t_{N-m} \end{aligned}$$

Note that degrees of freedom are those associated with MSE, NOT $n_i - 1$. As a result,

$$\bar{Y}_i \pm \text{SE}[\bar{Y}_{i\cdot}] \times t_{1-\alpha/2,N-m}$$

is a $100 \times (1-\alpha)\%$ confidence interval for $\mu_i$.

**A handy rule of thumb:** If $\hat{\theta}$ is an estimator of $\theta$, then in many situations

$$\hat{\theta} \pm 2 \times \text{SE}[\hat{\theta}]$$

is an approximate 95% CI for $\theta$.

**Coagulation Example:**

$$\bar{Y}_i \pm \text{SE}[\bar{Y}_{i\cdot}] \times t_{1-\alpha/2,N-m}$$

For 95% confidence intervals,

- $t_{1-\alpha,N-m} = t_{.975,20} =$ qt(.975,20) $\approx 2.1$
- $\text{SE}[\bar{Y}_i] = \sqrt{s^2/n_i} = \sqrt{5.6/n_i} = 2.37/\sqrt{n_i}$

| Diet | $\mu_{\text{diet}}$ | $n_i$ | $\text{SE}[\hat{\mu}_{\text{diet}}]$ | 95% CI |
|---|---|---|---|---|
| C | 68 | 6 | 0.97 | (65.9,70.0) |
| B | 66 | 6 | 0.97 | (63.9,68.0) |
| A | 61 | 4 | 1.18 | (58.5,63.5) |
| D | 61 | 8 | 0.84 | (59.2,62.8) |

### 5.4.3 Power calculations for the F-test

Recall, the **power** of a test is the probability of rejecting $H_0$ when it is false. Of course, this depends on **how** the null hypothesis is false.

$$\begin{aligned}\text{Power}(\boldsymbol{\mu}, \sigma^2, n) &= \Pr(\text{reject } H_0 | \mu, \sigma^2, n) \\ &= \Pr(F_{\text{obs}} > F_{1-\alpha,m-1,N-m} | \mu, \sigma^2, n)\end{aligned}$$

**The noncentral F-distribution:**

$$\begin{array}{rcl} Y_{11}\ldots,Y_{1n} & \sim & \text{i.i.d. normal}(\mu_1,\sigma^2) \\ \vdots & \vdots & \vdots \\ Y_{m1}\ldots,Y_{mn} & \sim & \text{i.i.d. normal}(\mu_t,\sigma^2) \end{array}$$

Under this sampling model, $F = F(\boldsymbol{Y})$ has a *non-central F distribution* with

- degrees of freedom $n-1, N-m$
- noncentrality parameter $\lambda$

$$\lambda = n\sum \tau_i^2/\sigma^2$$

where $\tau_i = \mu_i - \bar{\mu}$ is the $i$th treatment effect.

In many texts, power is expressed as a function of the quantity $\Phi$:

$$\Phi = \sqrt{\frac{n\sum\tau_i^2}{\sigma^2 m}} = \sqrt{\frac{\sum\tau_i^2/m}{\sigma^2/n}} = \sqrt{\lambda/m}$$

Lets try to understand what $\Phi$ represents:

$$\begin{array}{rcl} \Phi^2 & = & \dfrac{\sum\tau_i^2/m}{\sigma^2/n} \\ & = & \dfrac{\text{treatment variation}}{\text{experimental uncertainty}} \\ & = & \text{treatment variation} \times \text{experimental precision} \end{array}$$

Note that "treatment variation" means "average squared treatment effect size". We can gain some more intuition by rewriting $\lambda$ as follows:

$$\begin{array}{rcl} \lambda & = & n\sum\tau_i^2/\sigma^2 \\ & = & n \times m \times \left(\dfrac{\sum\tau_i^2}{m}\right)\dfrac{1}{\sigma^2} \\ & = & N \times \dfrac{\text{between-treatment variation}}{\text{within-treatment variation}} \end{array}$$

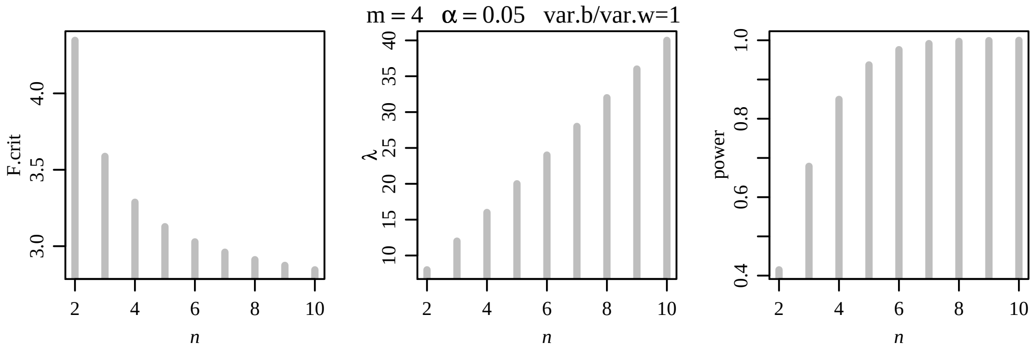


Figure 5.6: Power as a function of $n$ for $m = 4$, $\alpha = 0.05$ and $\bar{\tau}^2/\sigma^2 = 1$

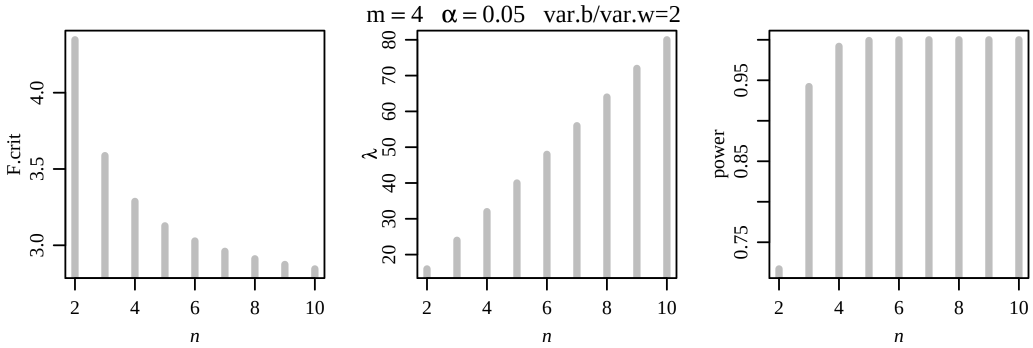


Figure 5.7: Power as a function of $n$ for $m = 4$, $\alpha = 0.05$ and $\bar{\tau}^2/\sigma^2 = 2$

So presumably power is increasing in $\Phi$, and $\lambda$.

$$\begin{aligned}
\text{Power}(\boldsymbol{\mu}, \sigma^2, n) &= \Pr(\text{reject H}_0 | \mu, \sigma^2, n) \\
&= \Pr(F_{\text{obs}} > F_{1-\alpha, m-1, N-m} | \mu, \sigma^2, n) \\
&= \text{1−pf( qf(1−alpha,m−1,N−m) , m−1 , N−m , ncp=lambda )}
\end{aligned}$$

## 5.5 Model diagnostics

Our model is

$$y_{ij} = \mu_j + \epsilon_{ij}.$$

We have shown that, if

A1: $\{\epsilon_{ij}\}$'s are independent;

A2: $\mathrm{Var}[\epsilon_{ij}] = \sigma^2$ for all $j$;

A3: $\{\epsilon_{ij}\}$'s are normally distributed.

then

$$F = \mathrm{MST}/\mathrm{MSE} \sim F_{m-1,N-m,\lambda}$$

where $\lambda$ is the noncentrality parameter. If in addition

H$_0$: $\mu_i = \mu$ for all $i = 1, \ldots, m$.

then the noncentrality parameter is zero and $F = \mathrm{MST}/\mathrm{MSE} \sim F_{m-1,N-m}$. We make these assumptions to

- do power calculations when designing a study,
- test hypotheses after having gathered the data, and
- make confidence intervals comparing the different treatments.

What should we do if the model assumptions are not correct?

We could have written

$$A0 : \{\epsilon_{ij}\} \sim \text{i.i.d. normal}(0, \sigma^2)$$

to describe all of $A1 - A3$. We don't do this because some violations of assumptions are more serious than others. Statistical folklore says the order of importance is $A1$, $A2$ then $A3$. We will discuss $A1$ when we talk about blocking. For now we will talk about $A2$ and $A3$.

### 5.5.1 Detecting violations with residuals

Violations of assumptions can be checked via *residual analysis*.

**Parameter estimates:**

$$\begin{aligned} y_{ij} &= \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot}) \\ &= \hat{\mu} + \hat{\tau}_i + \hat{\epsilon}_{ij} \end{aligned}$$

Our *fitted value* for any observation in group $i$ is $\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \hat{y}_{i\cdot}$

Our estimate of the error is $\hat{\epsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot}$.

$\hat{\epsilon}_{ij}$ is called the *residual* for observation $i, j$.

Assumptions about $\epsilon_{ij}$ can be checked by examining the values of $\hat{\epsilon}_{ij}$'s:

### 5.5.2 Checking normality assumptions:

Two standard graphical ways of assessing normality are with the following:

- Histogram:

  Make a histogram of $\hat{\epsilon}_{ij}$'s. This should look approximately bell-shaped if the (super)population is really normal **and** there are enough observations. If there are enough observations, graphically compare the histogram to a $N(0, s^2)$ distribution.

  In small samples, the histograms need not look particularly bell-shaped.

- Normal probability, or qq-plot:

  If $\epsilon_{ij} \sim N(0, \sigma^2)$ then the ordered residuals $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(mn)})$ should correspond linearly with quantiles of a standard normal distribution.

How non-normal can a sample from a normal population look? You can always check yourself by simulating data in R. See Figure **??**

**Example (Hermit Crab Data):** Is there variability in hermit crab population across six different coastline sites? A researchers sampled the population in 25 randomly sampled transects in each of the six sites.

**Data:** $y_{ij}$ = population total in transect $j$ of site $i$.

**Model:** $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Note that the data are *counts* so they cannot be exactly normally distributed.

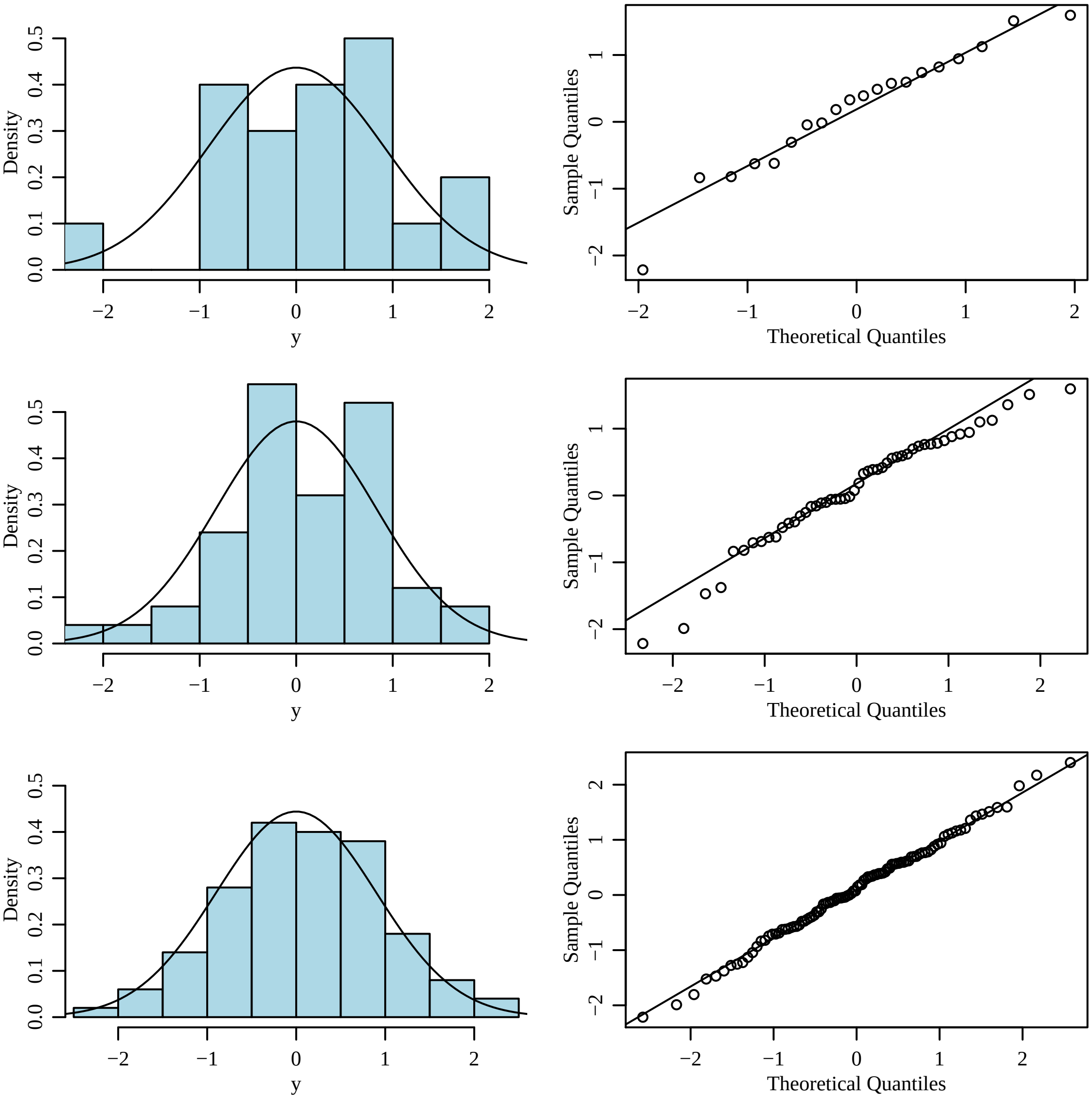


Figure 5.8: Normal scores plots of normal samples, with $n \in \{20, 50, 100\}$

**Data description:** See Figure 5.9.

| site | sample mean | sample median | sample std dev |
|---|---|---|---|
| 1 | 33.80 | 17 | 50.39 |
| 2 | 68.72 | 10 | 125.35 |
| 3 | 50.64 | 5 | 107.44 |
| 4 | 9.24 | 2 | 17.39 |
| 5 | 10.00 | 2 | 19.84 |
| 6 | 12.64 | 4 | 23.01 |

**ANOVA:**

```
> anova(lm(crab[,2]~as.factor(crab[,1])))
Analysis of Variance Table

Response: crab[, 2]
                      Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(crab[, 1])   5  76695   15339  2.9669 0.01401 *
Residuals            144 744493    5170
```

**Residuals:**

$$\hat{\epsilon}_{ij} = y_{ij} - \hat{\mu}_i = y_{ij} - \bar{y}_{i\cdot}$$

Residual diagnostic plots are in Figure 5.10. The data are clearly not normally distributed.

### 5.5.3 Checking variance assumptions

The null distribution in the F-test is based on $\text{Var}[\epsilon_{ij}] = \sigma^2$ for all groups $i$.

(a) tabulate *residual variance* in each treatment:

| Trt | Sample var. of $(\hat{\epsilon}_{i1}, \ldots, \hat{\epsilon}_{in})$ |
|---|---|
| 1 | $s_1^2$ |
| $\vdots$ | $\vdots$ |
| $m$ | $s_m^2$ |

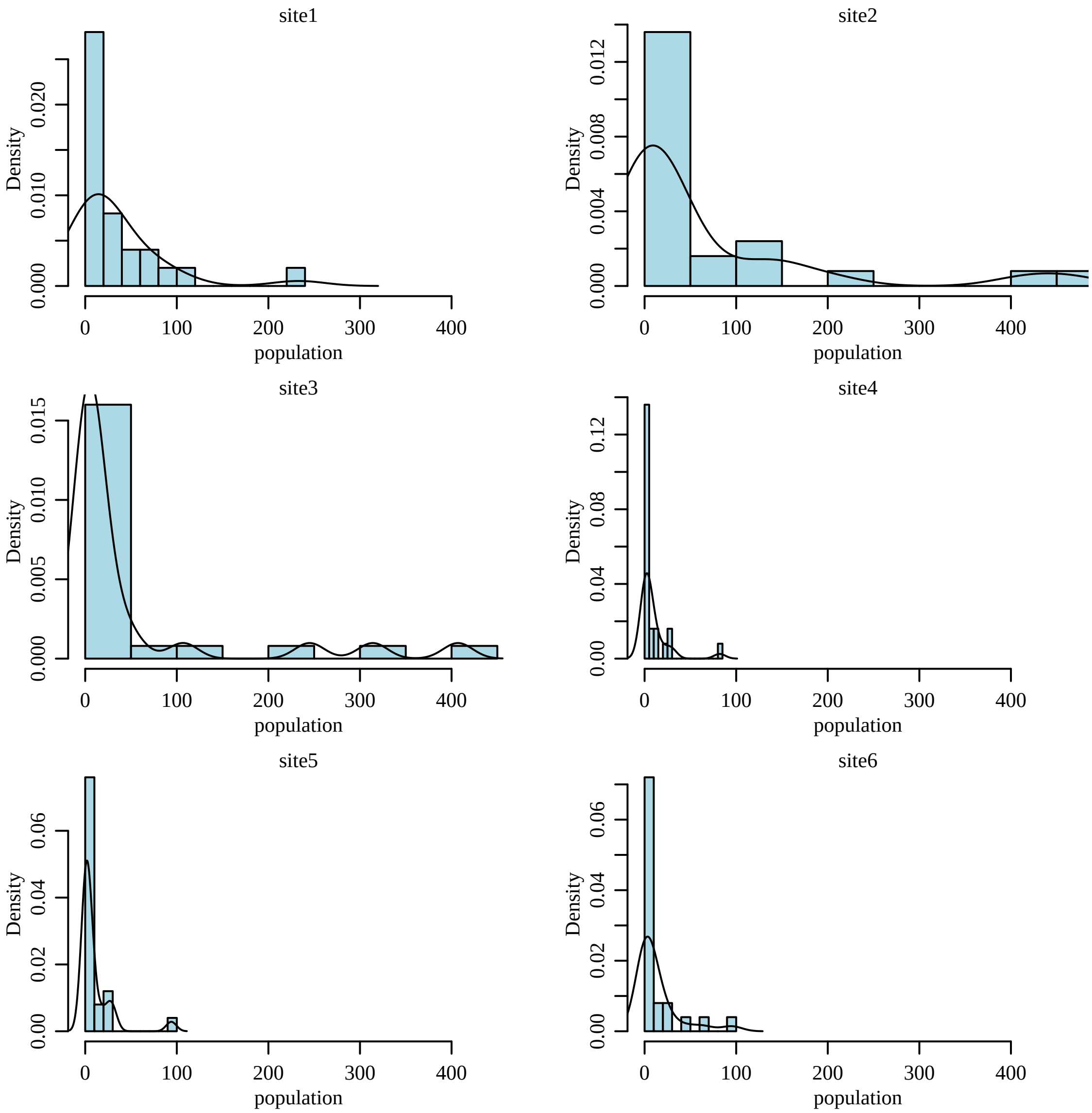


Figure 5.9: Crab data

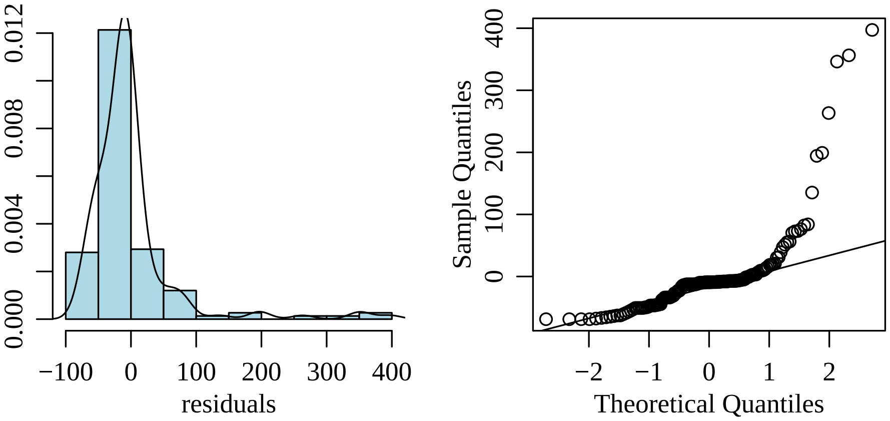


Figure 5.10: Crab residuals

**Rule of thumb:**

If $s^2_{\text{largest}}/s^2_{\text{smallest}} < 4$: don't worry;

If $s^2_{\text{largest}}/s^2_{\text{smallest}} > 7$: worry: need to account for the non-constant variance, especially if sample sizes are different. Consider a randomization test, or converting the data to ranks for the overall test of treatment effects.

(b) Residual vs. fitted value plots

For many types of data there is a *mean-variance relationship*: typically, groups with **large** means tend to have **large** variances. This is especially true when the underlying distributions are skewed.

To check this, plot $\hat{\epsilon}_{ij}$ (residual) vs. $\hat{y}_{ij} = \bar{y}_{i\cdot}$ (fitted value).

**Example: Crab data**

| Site | Sample mean | Sample standard deviation |
|---|---|---|
| 4 | 9.24 | 17.39 |
| 5 | 10.00 | 19.84 |
| 6 | 12.64 | 23.01 |
| 1 | 33.80 | 50.39 |
| 3 | 50.64 | 107.44 |
| 2 | 68.72 | 125.35 |

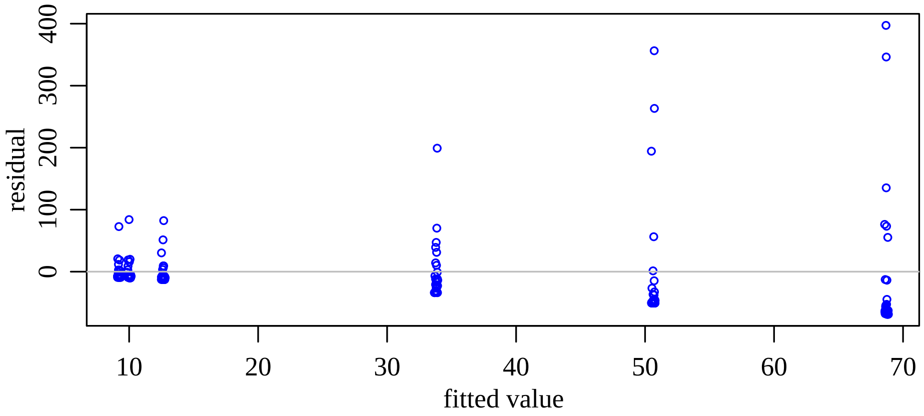


Figure 5.11: Fitted values versus residuals

Here we have $s^2_{\text{largest}}/s^2_{\text{smallest}} \approx 50$. This is a problem.

(c) Statistical tests of equality of variance:

**Levene's Test:** Let $d_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}_i$ is the sample median from group $i$.

These differences will be 'large' in group $i$, if group $i$ has a 'large' variance; and 'small' in group $i$, if group $i$ has a 'small' variance.

We compute:

$$F_0 = \frac{n\sum_{i=1}^{t}(d_{i\cdot} - \bar{d})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n}(d_{ij} - \bar{d}_{i\cdot})^2/(t(n-1))}$$

which is the ratio of the between group variability of the $d_{ij}$ to the within group variability of the $d_{ij}$.

**Reject** $\text{H}_0$: $\text{Var}[\epsilon_{ij}] = \sigma^2$ for all $i, j$ if $F_0 > F_{t-1,t(n-1),1-\alpha}$

**Crab data:**

$$F_0 = \frac{14,229}{4,860} = 2.93 > F_{5,144,0.95} = 2.28$$

hence we reject the null hypothesis of equal variances at the 0.05 level.

**See also**

- the F Max test
- Bartlett's test;
- the normal-theory test of equality for two variances (although this depends on normality).

**Crab data:** So the assumptions that validate the use of the $F$-test are violated. Now what?

### 5.5.4 Variance stabilizing transformations

Recall that one justification of the normal model,

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

was that if the noise $\epsilon_{ij} = X_{ij1} + X_{ij2} + \cdots$ was the result of the addition of unobserved **additive, independent** effects then by the central limit theorem $\epsilon_{ij}$ will be approximately normal.

However, suppose the effects are **multiplicative**, so that in fact:

$$Y_{ij} = \mu_i \times \epsilon_{ij} = \mu_i \times (X_{ij1} \times X_{ij2} \times \cdots)$$

In this case, the $Y_{ij}$ will not be normal, and the variances will **not** be constant:

$$\mathrm{Var}[Y_{ij}] = \mu_i^2 \mathrm{Var}[X_{ij1} \times X_{ij2} \times \cdots]$$

**Log transformation:**

$$\begin{aligned}
\log Y_{ij} &= \log \mu_i + (\log X_{ij1} + \log X_{ij2} + \cdots) \\
\mathrm{Var}[\log Y_{ij}] &= \mathrm{Var}[\log \mu_i + \log X_{ij1} + \log X_{ij2} + \cdots] \\
&= \mathrm{Var}[\log X_{ij1} + \log X_{ij2} + \cdots] \\
&= \sigma^2_{\log y}
\end{aligned}$$

So that the variance of the log-data does not depend on the mean $\mu_i$. Also note that by the central limit theorem the errors should be approximately normally distributed.

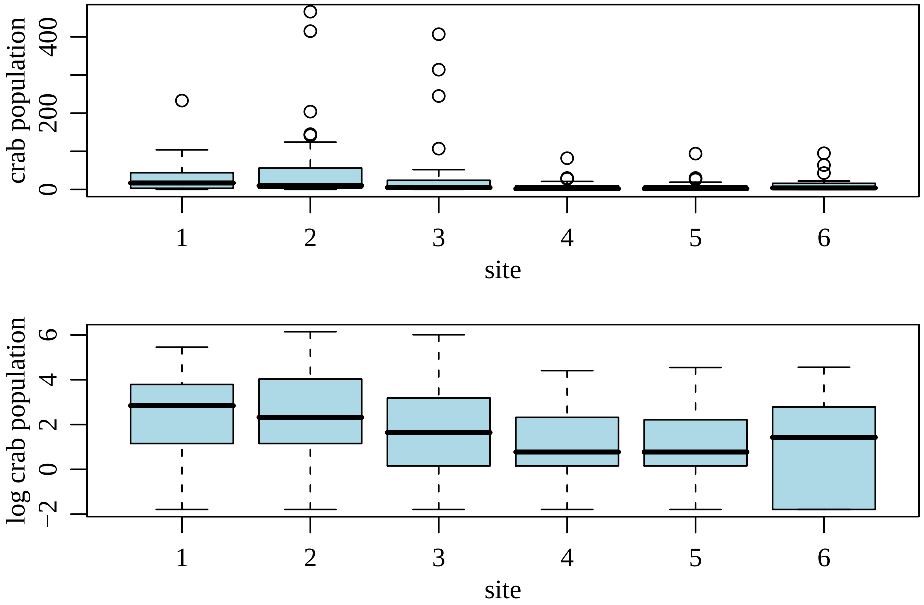


Figure 5.12: Data and log data

**Crab data:** Let $Y_{ij} = \log(Y_{ij}^{\text{raw}} + 1/6)$

| Site | Sample mean | Sample standard deviation |
|---|---|---|
| 6 | 0.82 | 2.21 |
| 4 | 0.91 | 1.87 |
| 5 | 1.01 | 1.74 |
| 3 | 1.75 | 2.41 |
| 1 | 2.16 | 2.27 |
| 2 | 2.30 | 2.44 |

```
> anova(lm(log(crab[,2]+1/6)~as.factor(crab[,1])))
Analysis of Variance Table

Response: log(crab[, 2] + 1/6)
                     Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(crab[, 1])   5  54.73   10.95  2.3226 0.04604 *
Residuals            144 678.60    4.71
> anova(lm(log(crab[,2]+1/6)~as.factor(crab[,1])))
Analysis of Variance Table
```

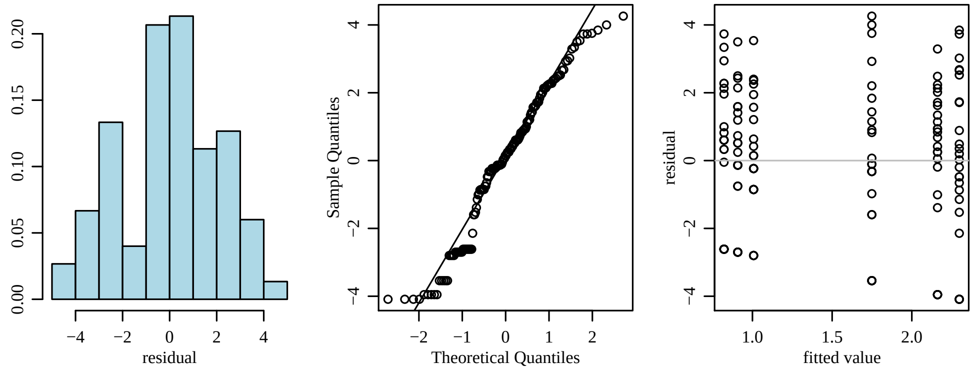


Figure 5.13: Diagnostics after the log transformation

```
Response: log(crab[, 2] + 1/6)
                        Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(crab[, 1])     5  54.73   10.95  2.3226 0.04604 *
Residuals              144 678.60    4.71
```

**Other transformations:** For data having multiplicative effects , we showed that

$$\sigma_i \propto \mu_i,$$

and taking the log stabilized the variances. In general, we may observe: $\sigma_i \propto \mu_i^\alpha$ i.e. the standard deviation of a group depends on the group mean.

The goal of a variance stabilizing transformation is to find a transformation of $y_{ij}$ to $y_{ij}^*$ such that $\sigma_{y_{ij}^*} \propto (\mu_i^*)^0 = 1$, i.e. the standard deviation doesn't depend on the mean.

Consider the class of *power transformations*, transformations of the form $Y_{ij}^* = Y_{ij}^\lambda$. Based on a Taylor series expansion of $g_\lambda(Y) = Y^\lambda$ around $\mu_i$, we have

$$\begin{aligned} Y_{ij}^* &= g_\lambda(Y_{ij}) \\ &\approx \mu_i^\lambda + (Y_{ij} - \mu_i)\lambda\mu_i^{\lambda-1} \\ \mathrm{E}[Y_{ij}^*] &\approx \mu_i^\lambda \\ \mathrm{Var}[Y_{ij}^*] &\approx \mathrm{E}[(Y_{ij} - \mu_i)^2](\lambda\mu_i^{\lambda-1})^2 \\ \mathrm{SD}[Y_{ij}^*] &\propto \mu_i^\alpha\mu_i^{\lambda-1} = \mu_i^{\alpha+\lambda-1} \end{aligned}$$

So if we observe $\sigma_i \propto \mu_i^\alpha$, then $\sigma_i^* \overset{\sim}{\propto} \mu_i^{\alpha+\lambda-1}$. So if we take $\lambda = 1-\alpha$ then we will have stabilized the variances to some extent. Of course, we typically don't know $\alpha$, but we could try to estimate it from data.

**Estimation of $\alpha$:**

$$\begin{aligned} \sigma_i \propto \mu_i^\alpha \Leftrightarrow \sigma_i &= c\mu_i^\alpha \\ \log \sigma_i &= \log c + \alpha \times \log \mu_i, \\ \text{so} \quad \log s_i &\approx \log c + \alpha \log \bar{y}_{i\cdot} \end{aligned}$$

Thus we may use the following procedure:

(1) Plot $\log s_i$ vs. $\log \bar{y}_{i\cdot}$

(2) Fit a least squares line: `lm(` $\log s_i \sim \log \bar{y}_{i\cdot}$`)`

(3) The **slope** $\hat{\alpha}$ of the line is an estimate of $\alpha$.

(4) Analyze $y_{ij}^* = y_{ij}^{1-\hat{\alpha}}$.

Here are some common transformations:

| mean-var relation | $\alpha$ | $\lambda = 1-\alpha$ | transform | $y_{ij}^*$ |
|---|---|---|---|---|
| $\sigma_y \propto$ const. | 0 | 1 | no transform! | $y_{ij}$ |
| $\sigma_y \propto \mu_i^{1/2}$ | 1/2 | 1/2 | square root | $y_{ij}^{1/2} = \sqrt{y_{ij}}$ |
| $\sigma_y \propto \mu_i^{3/4}$ | 3/4 | 1/4 | quarter power | $y_{ij}^{1/4}$ |
| $\sigma_y \propto \mu_i$ | 1 | 0 | log | $\log y_{ij}$ |
| $\sigma_y \propto \mu_i^{3/2}$ | 3/2 | -1/2 | reciproc. sqr. root | $y_{ij}^{-1/2}$ |
| $\sigma_y \propto \mu_i^2$ | 2 | -1 | reciprocal | $1/y_{ij}$ |

Note that

> all the mean-variance relationships here are examples of power-laws. Not all mean-variance relations are of this form.
>
> $\alpha = 1$ is the multiplicative model discussed previously.

### More about the log transform

How did $\alpha = 1$ give us $y^*_{ij} = \log y_{ij}$, shouldn't it be $y_{ij}^{1-\alpha} = y_{ij}^{\lambda} = y_{ij}^{0} = 1$ in there?

Everything will make sense if we define for any $\lambda \neq 0$:

$$y^{*(\lambda)} = \frac{y^\lambda - 1}{\lambda} \propto y^\lambda + c.$$

For $\lambda = 0$, it's natural to define the transformation as:

$$\begin{aligned} y^{*(0)} = \lim_{\lambda \to 0} y^{*(\lambda)} &= \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} \\ &= \left.\frac{y^\lambda \ln y}{1}\right|_{\lambda=0} = \ln y \end{aligned}$$

Note that for a given $\lambda \neq 0$ it will not change the results of the ANOVA on the transformed data if we transform using:

$$y^* = y^\lambda \quad \text{or} \quad y^{*(\lambda)} = \frac{y^\lambda - 1}{\lambda} = ay^\lambda + b.$$

**To summarize the procedure:** If the data present strong evidence of nonconstant variance,

(1) Plot $\log s_i$ vs. $\log \bar{y}_{i\cdot}$. **If** the relationship looks linear, then

(2) Fit a least squares line: `lm(` $\log s_i \sim \log \bar{y}_{i\cdot}$`)`

(3) The **slope** $\hat{\alpha}$ of the line is an estimate of $\alpha$.

(4) Analyze $y^*_{ij} = y_{ij}^{1-\hat{\alpha}}$.

This procedure is called the "Box-Cox" transformation (George Box and David Cox first proposed the method in a paper in 1964). If the relationship between $\log s_i$ and $\log \bar{y}_i$ does not look linear, then a Box-Cox transformation will probably not help.

## A few words of caution:

- For variance stabilization via the Box-Cox procedure to do much good, the linear relationship between means and variances should be quite tight.

- Remember the rule of thumb which says not to worry if the ratio of the largest to smallest variance is less than 4, i.e. don't use a transform unless there are drastic differences in variances.

- Don't get too carried away with the Box-Cox transformations. If $\hat{\alpha} = 0.53$ don't analyze $y^*_{ij} = y_{ij}^{0.47}$, just use $y_{ij}^{0.5}$. Remember, $\hat{\alpha}$ is just an estimate anyway (See Box and Cox 'Rebuttal', JASA 1982.)

- Remember to make sure that you describe the units of the transformed data, and make sure that readers of your analysis will be able to understand that the model is additive in the transformed data, but not in the original data. Also always include a descriptive analysis of the untransformed data, along with the p-value for the transformed data.

- Try to think about whether the associated non-linear model for $y_{ij}$ makes sense.

- Don't assume that the transformation is a magical fix: remember to look at residuals and diagnostics **after** you do the transform. If things haven't improved much, don't transform.

- Remember that the mean-variance relationship might not be cured by a transform in the Box-Cox class. For example, if the response is a binomial proportion (= proportion of successes out of $n$), we have mean $= p$, s.d. $= \sqrt{p(1-p)}$; the variance stabilizing transformation in this case is $y^* = \arcsin\sqrt{y}$.

- Keep in mind that statisticians disagree on the usefulness of transformations: some regard them as a 'hack' more than a 'cure':

- It can be argued that if the scientist who collected the data had a good reason for using certain units, then one should not just transform the data in order to bang it into an ANOVA-shaped hole. (Given enough time and thought we could instead build a non-linear model for the original data.)

- **The sad truth:** as always you will need to exercise **judgment** while performing your analysis.

These warnings apply whenever you might reach for a transform, whether in an ANOVA context, or a linear regression context.

**Example (Crab data):** Looking at the plot of means vs. sd.s suggests $\alpha \approx 1$, implying a log-transformation. However, the zeros in our data lead to problems, since $\log(0) = -\infty$.

Instead we can use $y^*_{ij} = \log(y_{ij} + 1/6)$. For the transformed data this gives us a ratio of the largest to smallest standard deviation of approximately 2 which is acceptable based on the rule of 4. Additionally, the residual diagnostic plots (Figure 5.13) are much improved
This table needs to be fixed: Third column needs to be sd(log(y)).

| site | sample sd | sample mean | log(sample sd) | log(sample mean) |
|---|---|---|---|---|
| 4 | 17.39 | 9.24 | 2.86 | 2.22 |
| 5 | 19.84 | 10.00 | 2.99 | 2.30 |
| 6 | 23.01 | 12.64 | 3.14 | 2.54 |
| 1 | 50.39 | 33.80 | 3.92 | 3.52 |
| 3 | 107.44 | 50.64 | 4.68 | 3.92 |
| 2 | 125.35 | 68.72 | 4.83 | 4.23 |

```
lm(formula = log_sd ~ log_mean)
Coefficients:
(Intercept)     log_mean
     0.6652       0.9839
```

## 5.6 Treatment Comparisons

Recall the coagulation time data from the beginning of the chapter: Four different diets were assigned to a population of 24 animals, with $n_1 = 4$, $n_2 = 6$, $n_3 = 6$ and $n_4 = 8$.

```
> anova(lm(ctime~diet))
Analysis of Variance Table

```

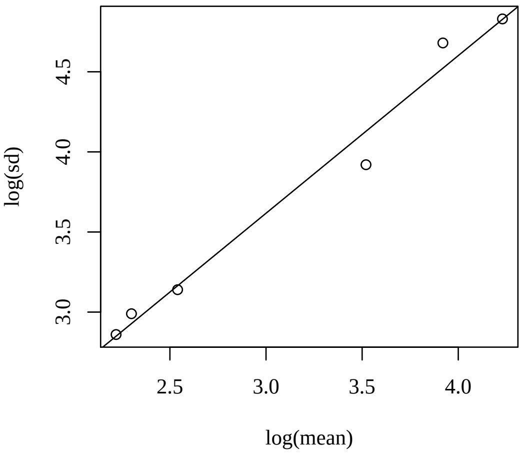


Figure 5.14: Mean-variance relationship of the transformed data

```
Response: ctime
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3  228.0    76.0  13.571 4.658e-05 ***
Residuals 20  112.0     5.6
```

We conclude from the $F$-test that there are substantial differences between the population treatment means. How do we decide what those differences are?

### 5.6.1 Contrasts

Differences between sets of means can be evaluated by estimating *contrasts*. A *contrast* is a linear function of the means such that the coefficients sum to zero:

$$C = C(\boldsymbol{\mu}, \boldsymbol{k}) = \sum_{i=1}^{m} k_i \mu_i \ , \text{ where } \sum_{i=1}^{m} k_i = 0$$

**Examples:**

- diet 1 vs diet 2 : $C = \mu_1 - \mu_2$

- diet 1 vs diets 2,3 and 4 : $C = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$
- diets 1 and 2 vs diets 3 and 4 : $C = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$

Contrasts are **functions of the unknown parameters**. We can estimate them, and obtain standard errors for them. This leads to confidence intervals and hypothesis tests.

**Parameter estimates:** Let

$$\hat{C} = \sum_{i=1}^{m} k_i \mu_i = \sum_{i=1}^{m} k_i \bar{y}_{i\cdot}$$

Then $\mathrm{E}[\hat{C}] = C$, so $\hat{C}$ is a *unbiased estimator* of $C$.

**Standard errors:**

$$\begin{aligned} \mathrm{Var}[\hat{C}] &= \sum_{i=1}^{m} \mathrm{Var}[k_i \bar{y}_{i\cdot}] \\ &= \sum_{i=1}^{m} k_i^2 \sigma^2 / n_i \\ &= \sigma^2 \sum_{i=1}^{m} k_i^2 / n_i \end{aligned}$$

So an estimate of $\mathrm{Var}[\hat{C}]$ is

$$s_C^2 = s^2 \sum_{i=1}^{m} k_i^2 / n_i$$

The *standard error* of a contrast is an estimate of its *standard deviation*

$$\mathrm{SE}[\hat{C}] = s\sqrt{\sum \frac{k_i^2}{n_i}}$$

**t-distributions for contrasts:** Consider

$$\frac{\hat{C}}{\mathrm{SE}[\hat{C}]} = \frac{\sum_{i=1}^{m} k_i \bar{y}_{i\cdot}}{s\sqrt{\sum k_i^2/n_i}}$$

If the data are normally distributed , then under $\mathrm{H}_0 : C = \sum k_i \mu_i = 0$,

$$\frac{\hat{C}}{\mathrm{SE}[\hat{C}]} \sim t_{N-m}$$

**Exercise:** Prove this result

**Hypothesis test:**

- $\mathrm{H}_0 : C = 0$ versus $\mathrm{H}_1 : C \neq 0$.
- Level-$\alpha$ test : Reject $\mathrm{H}_0$ if $|\hat{C}/\mathrm{SE}[\hat{C}]| > t_{1-\alpha/2,N-m}$.
- $p$-value: $\Pr(|t_{N-m}| > |\hat{C}(\mathbf{y})/\mathrm{SE}[\hat{C}(\mathbf{y})]|) =$ `2*(1-pt( abs(c_hat/se_c_hat),N-m))`

**Example:** Recall in the coagulation example $\hat{\mu}_1 = 61, \hat{\mu}_2 = 66$, and their 95% confidence intervals were (58.5,63.5) and (63.9,68.0). Let $C = \mu_A - \mu_B$.

**Hypothesis test:** $\mathrm{H}_0 : C = 0$.

$$\begin{aligned}\frac{\hat{C}}{\mathrm{SE}[\hat{C}]} &= \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{s\sqrt{1/6 + 1/4}} \\ &= \frac{-5}{1.53} \\ &= 3.27\end{aligned}$$

$p$-value = `2*(1-pt(3.27,20))` = 0.004.

**Confidence intervals:** Based on the normality assumptions, a $100 \times (1 - \alpha)\%$ confidence interval for $C$ is given by

$$\hat{C} \pm t_{1-\alpha/2,N-m} \times \mathrm{SE}[\hat{C}]$$

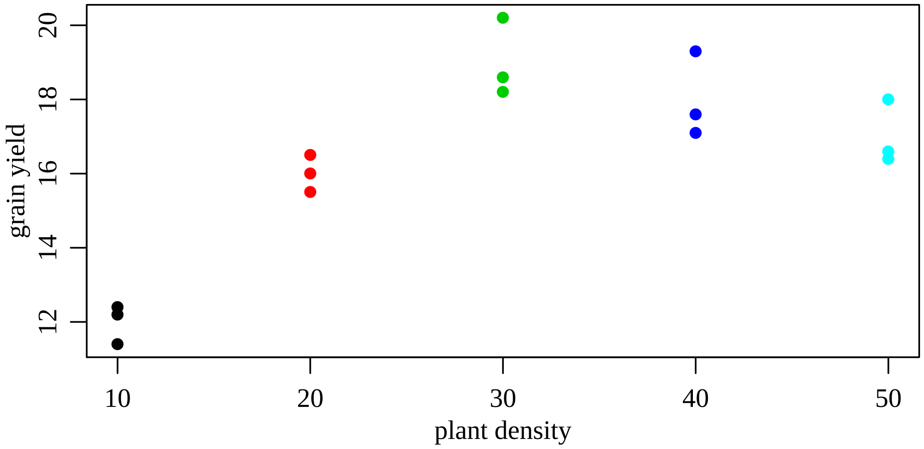


Figure 5.15: Yield-density data

### 5.6.2 Orthogonal Contrasts

What use are contrasts beyond just comparing two means? Consider the data in Figure 5.15, which show the results of a CRD for an experiment on the effects of planting density on crop yield in which there were three fields randomly assigned to each of 5 planting densities.

```
> anova(lm(y~as.factor(x))
             Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(x)  4 87.600  21.900  29.278 1.690e-05 ***
Residuals    10  7.480   0.748
```

There is strong evidence of an effect of planting density. How should we summarize the effect? In this experiment, the treatment levels have an ordering to them (this is not always the case). Consider the following $m - 1 = 4$ contrasts:

| Contrast | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ |
|---|---|---|---|---|---|
| $C_1$ | -2 | -1 | 0 | 1 | 2 |
| $C_2$ | 2 | -1 | -2 | -1 | 2 |
| $C_3$ | -1 | 2 | 0 | -2 | 1 |
| $C_4$ | 1 | -4 | 6 | -4 | 1 |

Note:

- These are all actually contrasts (the coefficients sum to zero).
- They are **orthogonal** : $C_i \cdot C_j = 0$.

What are these contrasts representing? Would would make them large?

- If all $\mu_i$'s are the same, then they will all be close to zero. This is the "sum to zero" part, i.e. $C_i \cdot \mathbf{1} = 0$ for each contrast.
- $C_1$ will be big if $\mu_1, \dots, \mu_5$ are monotonically increasing or monotonically decreasing. This contrast is measuring the *linear component* of the relationship between density and yield.
- $C_2$ is big if "there is a bump", up or down, in the middle of the treatments, i.e. $C_2$ is measuring the *quadratic component* of the relationship.
- Similarly, $C_3$ and $C_4$ are measuring the cubic and quartic parts of the relationship between density and yield.

The orthogonality bit is important: Suppose the true relationship between density and yield is linear, eg, $\mu_i = \alpha + \beta \times \text{density}_i$. Then $C_2(\boldsymbol{\mu}) = C_3(\boldsymbol{\mu}) = C_4(\boldsymbol{\mu}) = 0$.

You can produce these contrast coefficients in R:

```
> t(contr.poly(5))
           [,1]       [,2]          [,3]       [,4]      [,5]
.L -0.6324555 -0.3162278  0.000000e+00  0.3162278 0.6324555
.Q  0.5345225 -0.2672612 -5.345225e-01 -0.2672612 0.5345225
.C -0.3162278  0.6324555 -4.095972e-16 -0.6324555 0.3162278
^4  0.1195229 -0.4780914  7.171372e-01 -0.4780914 0.1195229
```

Here each contrast has been normalized so that $\sum k_i^2 = 1$. This doesn't change the orthogonality.
Estimating the contrasts is easy:

```
> trt.means<-tapply(y,x,mean)
> trt.means
10 20 30 40 50
12 16 19 18 17

> c.hat<-trt.means%*%contr.poly(5)
> c.hat
           .L        .Q        .C      ^4
[1,] 3.794733 -3.741657 0.3162278 0.83666
```

```
> 3*c.hat^2
       .L .Q  .C  ^4
[1,] 43.2 42 0.3 2.1

> sum(3*c.hat^2)
[1] 87.6
```

Coincidence? I think not. Recall, we represented treatment variation as a vector that lived in $m-1$ dimensional space. This variation can be further decomposed into $m-1$ orthogonal parts:

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Linear trt | 1 | 43.20 | 43.20 | 57.75 |
| Quad trt | 1 | 42.00 | 42.00 | 56.15 |
| Cubic trt | 1 | 0.30 | 0.30 | 0.40 |
| Quart trt | 1 | 2.10 | 2.10 | 2.81 |
| Total trt | 4 | 87.60 | 21.90 | 29.28 |
| Error | 10 | 7.48 | 0.75 | |
| Total | 14 | 95.08 | | |

The useful idea behind orthogonal contrasts is that the treatment variation can be decomposed into orthogonal parts. As you might expect, under $\mathrm{H}_{0r} : C_r = 0$, the $F$-statistic corresponding to the $r$th contrast has an $F$-distribution with 1 and $N-m$ degrees of freedom (assuming normality, constant variance, etc.). For the planting density data, we find strong evidence of linear and quadratic components to the relationship between density and yield.

### 5.6.3 Multiple Comparisons

An experiment with $m$ treatment levels has $\binom{m}{2}$ pairwise comparisons, i.e. contrasts of the form $C = \mu_i - \mu_j$. Should we perform hypothesis tests for all comparisons?

The more hypotheses we test, the higher the probability that at least one of them will be rejected, regardless of their validity.

**Two levels of error:** Define the hypotheses

- $\mathrm{H}_0 : \mu_i = \mu_j$ for all $i, j$

- $\mathrm{H}_{0ij} : \mu_i = \mu_j$

We can associate error rates to both of these types of hypotheses

- Experiment-wise type I error rate: Pr(reject $\mathrm{H}_0|\mathrm{H}_0$ is true ).
- Comparison-wise type I error rate : Pr(reject $\mathrm{H}_{0ij}|\mathrm{H}_{0ij}$ is true ).

Consider the following procedure:

1. Gather data
2. Compute all pairwise contrasts and their $t$-statistics
3. Reject each $\mathrm{H}_{0ij}$ for which $|t_{ij}| > t_{1-\alpha_C/2,N-m}$

Letting $t_{\text{crit}} = t_{1-\alpha_C/2,N-m}$, the **comparison-wise type I error rate** is of course

$$P(|t_{ij}| > t_{\text{crit}}|\mathrm{H}_{0ij}) = \alpha_C.$$

The **experiment-wise type I error rate** is the probability that we say differences between treatments exist when no differences exist:

$$P(|t_{ij}| > t_{\text{crit}} \text{ for some } i,j \ |\mathrm{H}_0) \geq \alpha_C$$

with equality only if there are two treatments total. The fact that the experiment-wise error rate is larger than the comparison-wise rate is called the issue of *multiple comparisons*. What is the experiment-wise rate in this analysis procedure?

$$\begin{aligned}
\Pr(\text{one or more } \mathrm{H}_{0ij} \text{ rejected } |\mathrm{H}_0) &= 1 - \Pr(\text{none of } \mathrm{H}_{0ij} \text{ rejected } |\mathrm{H}_0) \\
&\lesssim 1 - \prod_{i,j} \Pr(\mathrm{H}_{0ij} \text{ not rejected } |\mathrm{H}_0) \\
&= 1 - (1-\alpha_C)^{\binom{m}{2}}
\end{aligned}$$

We can approximate this with a Taylor series expansion: Let $f(x) = 1-(1-\alpha_C)^x$. Then $f'(0) = -\log(1-\alpha_C)$ and

$$\begin{aligned}
f(x) &\approx f(0) + xf'(0) \\
1-(1-\alpha_C)^x &\approx (1-1) - x\log(1-\alpha_C) \\
&\approx x\alpha_C \text{ for small } \alpha_C.
\end{aligned}$$

So

$$\Pr(\text{one or more } \mathrm{H}_{0ij} \text{ rejected } |\mathrm{H}_0) \lesssim \binom{m}{2}\alpha_C$$

If we are worried about possible dependence among the tests, perhaps a better way to derive this bound is to recall that

$$\begin{aligned}\Pr(A_1 \cup A_2) &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2) \\ &\leq \Pr(A_1) + \Pr(A_2) \text{ and that} \\ \Pr(A_1 \cup \cdots \cup A_k) &\leq \Pr(A_1) + \cdots + \Pr(A_k)\end{aligned}$$

(subadditivity). Therefore

$$\begin{aligned}P(\text{one or more } \mathrm{H}_{0ij} \text{ rejected } |\mathrm{H}_0) &\leq \sum_{i,j} P(\mathrm{H}_{0ij} \text{ rejected } |\mathrm{H}_0) \\ &= \binom{m}{2}\alpha_C\end{aligned}$$

**Bonferroni error control:**
The Bonferroni procedure for controlling experiment-wise type I error rate is as follows:

1. Compute pairwise $t$-statistics on all $\binom{m}{2}$ pairs.
2. Reject $\mathrm{H}_{0ij}$ if $|t_{ij}| > t_{1-\alpha_C/2,N-m}$

where $\alpha_C = \alpha_E/\binom{m}{2}$.

- the experiment-wise error rate is less than $\alpha_E$
- the comparison-wise error rate is $\alpha_C$.

So for example if $\alpha_E = 0.05$ and $m = 5$, then $\alpha_c = 0.005$.

**Fisher's Protected LSD (Fisher's least-significant-difference method):**
Another approach to controlling type I error makes use of the $F$-test.

1. Perform ANOVA and compute the $F$-statistic.
2. If $F(\mathbf{y}) < F_{1-\alpha_E,m-1,N-m}$ then don't reject $\mathrm{H}_0$ and stop.

3. If $F(\mathbf{y}) > F_{1-\alpha_E,m-1,N-m}$ then reject $\mathrm{H}_0$ **and** reject all $\mathrm{H}_{0ij}$ for which $|\hat{C}_{ij}/\mathrm{SE}[\hat{C}_{ij}]| > t_{1-\alpha_C/2,N-m}$.

Under this procedure,

- Experiment-wise type I error rate is $\alpha_E$
- Comparison-wise type I error rate is

$$\begin{aligned}\Pr(\text{reject } \mathrm{H}_{0ij}\ |\mathrm{H}_{0ij}) &= \Pr(F > F_{\text{crit}} \text{ and } |t_{ij}| > t_{\text{crit}}|\mathrm{H}_{0ij}) \\ &= \Pr(F > F_{\text{crit}}|\mathrm{H}_{0ij}) \times \Pr(|t_{ij}| > t_{\text{crit}}|F > F_{\text{crit}}, \mathrm{H}_{0ij})\end{aligned}$$

  - If $\mathrm{H}_0$ is also true, this is less than $\alpha_E$, and is generally between $\alpha_C \times \alpha_E$ and $\alpha_E$, depending on how many treatments there are. For example, if there only a few treatments, then $F > F_{\text{crit}}$ suggests $|t_{ij}| > t_{\text{crit}}$.
  - If $\mathrm{H}_0$ is not true, we don't know what this is, but some simulation work suggests this is approximately $\alpha_C \times$power, again depending on how many treatments there are and the true treatment variation.

### 5.6.4 False Discovery Rate procedures

### 5.6.5 Nonparametric tests

# Chapter 6

# Factorial Designs

**Example (Insecticide):** In developing methods of pest control, researchers are interested in the efficacy of different **types** of poison and different **delivery** methods.

- Treatments:
  - Type $\in \{I, II, III\}$
  - Delivery $\in \{A, B, C, D\}$
- Response: time to death, in minutes.

**Possible experimental design:** Perform two separate CRD experiments, one testing for the effects of Type and the other for Delivery. But,

which Delivery to use for the experiment testing for Type effects?

which Type to use for the experiment testing for Delivery effects?

To compare different Type×Delivery combinations, we need to do experiments under all 12 *treatment combinations*.

**Experimental design:** 48 insects randomly assigned to treatments: 4 to each treatment combination, i.e.

4 assigned to $(I, A)$,

4 assigned to $(I, B)$,

$\vdots$

It might be helpful to visualize the design as follows:

| | **Delivery** | | | |
|---|---|---|---|---|
| **Type** | A | B | C | D |
| 1 | $\boldsymbol{y}_{I,A}$ | $\boldsymbol{y}_{I,B}$ | $\boldsymbol{y}_{I,C}$ | $\boldsymbol{y}_{I,D}$ |
| 2 | $\boldsymbol{y}_{II,A}$ | $\boldsymbol{y}_{II,B}$ | $\boldsymbol{y}_{II,C}$ | $\boldsymbol{y}_{II,D}$ |
| 3 | $\boldsymbol{y}_{III,A}$ | $\boldsymbol{y}_{III,B}$ | $\boldsymbol{y}_{III,C}$ | $\boldsymbol{y}_{III,D}$ |

This type of design is called a *factorial design.* Specifically, this design is a $3 \times 4$ *two-factor* design with 4 *replications* per treatment combination.

**Factors:** Categories of treatments

**Levels of a factor:** the different treatments in a category

So in this case, Type and Delivery are both factors. There are 3 levels of Type and 4 levels of Delivery.

## 6.1 Data analysis:

Let's first look at a series of plots:

**Marginal Plots:** Based on these marginal plots, it looks like $(III, A)$ would be the most effective combination. But are the effects of Type consistent across levels of Delivery?

**Conditional Plots:** Type III looks best across delivery types. But the difference between types I and II seems to depend on delivery. For example, for delivery methods $B$ or $D$ there doesn't seem to be much of a difference between I and II.

**Cell Plots:** Another way of looking at the data is to just view it as a CRD with $3 \times 4 = 12$ different groups. Sometimes each group is called a *cell.*

Notice that there seems to be a bit of a mean-variance relationship. Let's take care of this before we go any further: Plotting means versus standard deviations on both the raw and log scale gives the relationship in Figure 6.4. Computing the least squares line gives

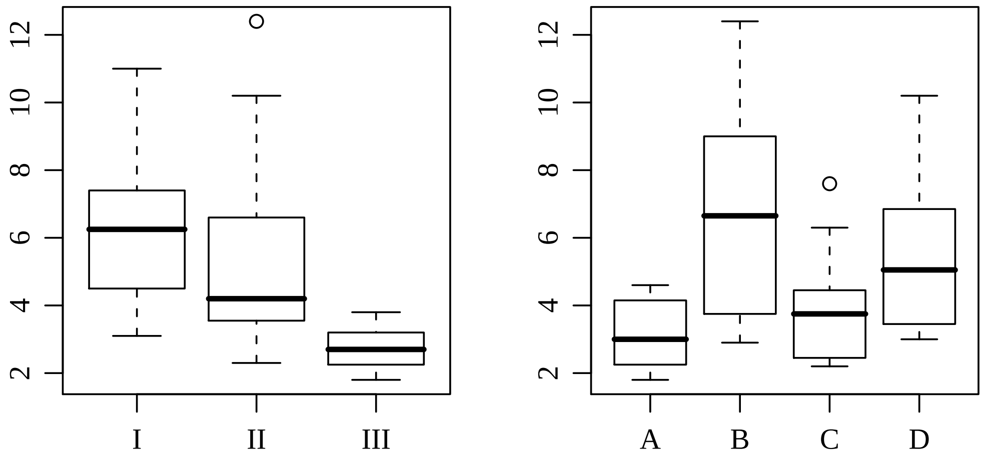


Figure 6.1: Marginal Plots.

```
> lm(log(sds)~log(means))
Coefficients:
(Intercept)    log(means)
     -3.203         1.977
```

So $\sigma_i \approx c\mu_i^2$. This suggests a reciprocal transformation, i.e. analyzing $y_{i,j} =$ 1/ time to death. Does this reduce the relationship between means and variances? Figure 6.5 shows the mean-variance relationship for the transformed data. The transformation seems to have improved things. We select this as our data scale and start from scratch, beginning with plots (Figure 6.6).

**Possible analysis methods:** Let's first try to analyze these data using our existing tools:

- Two one-factor ANOVAS: Just looking at Type, for example, the experiment is a one-factor ANOVA with 3 treatment levels and 16 reps per treatment. Conversely, looking at Delivery, the experiment is a one factor ANOVA with 4 treatment levels and 12 reps per treatment.

  ```
  > dat$y<-1/dat$y
  > anova(lm(dat$y~dat$type))

              Df  Sum Sq Mean Sq F value    Pr(>F)
  dat$type     2 0.34877 0.17439  25.621 3.728e-08 ***
  ```

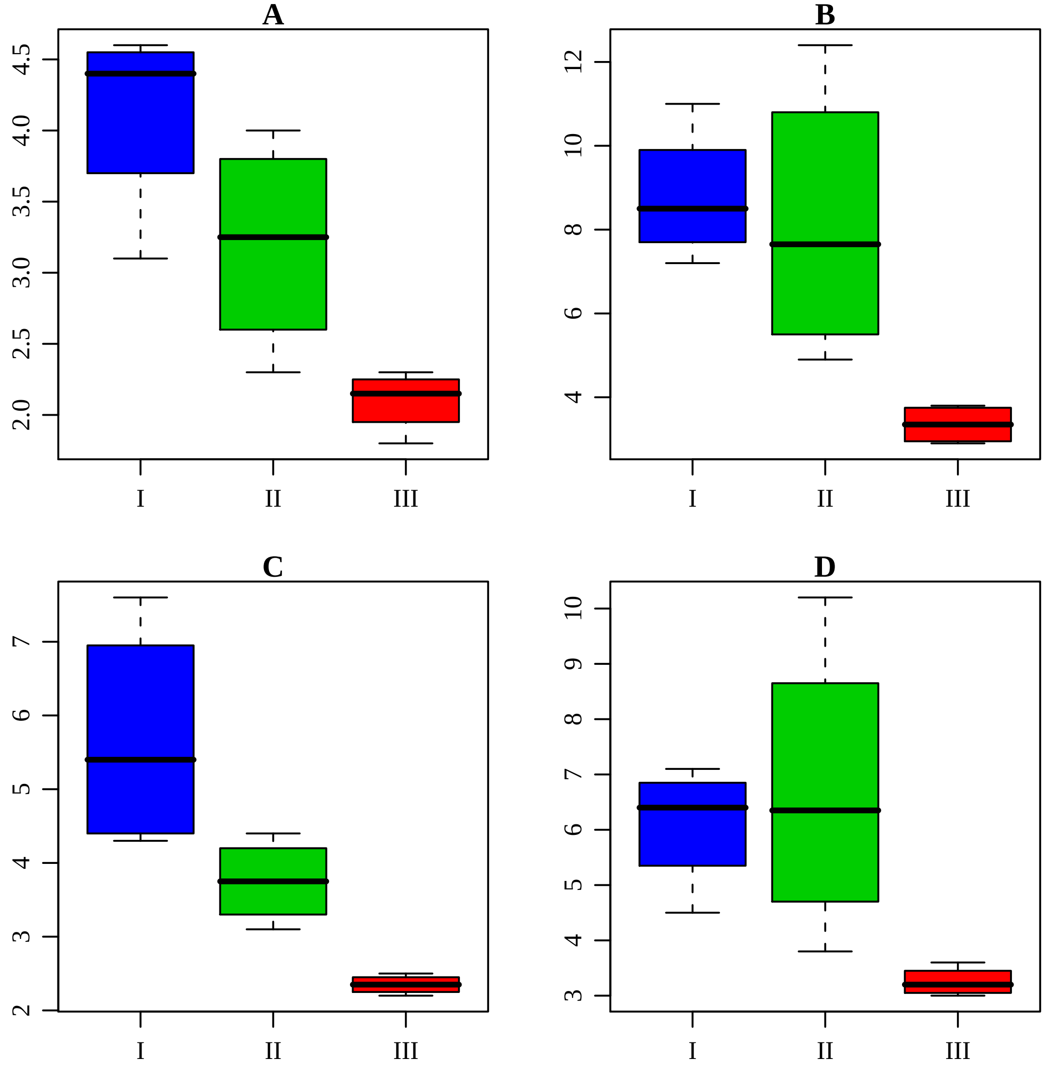


Figure 6.2: Conditional Plots.

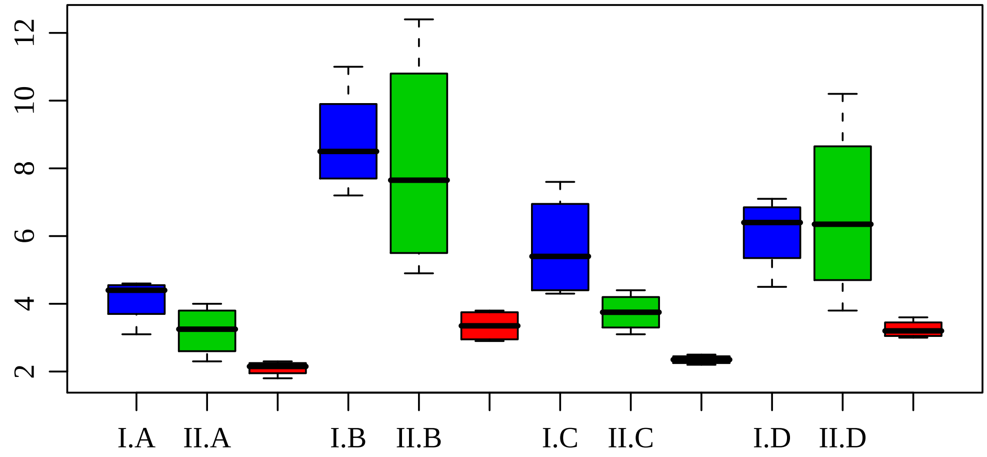


Figure 6.3: Cell plots.

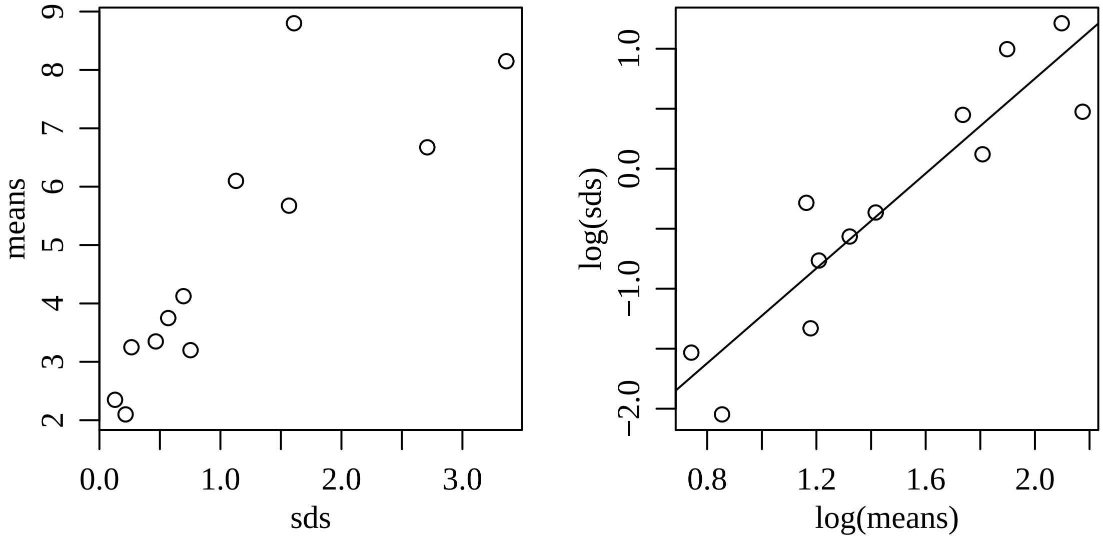


Figure 6.4: Mean-variance relationship.

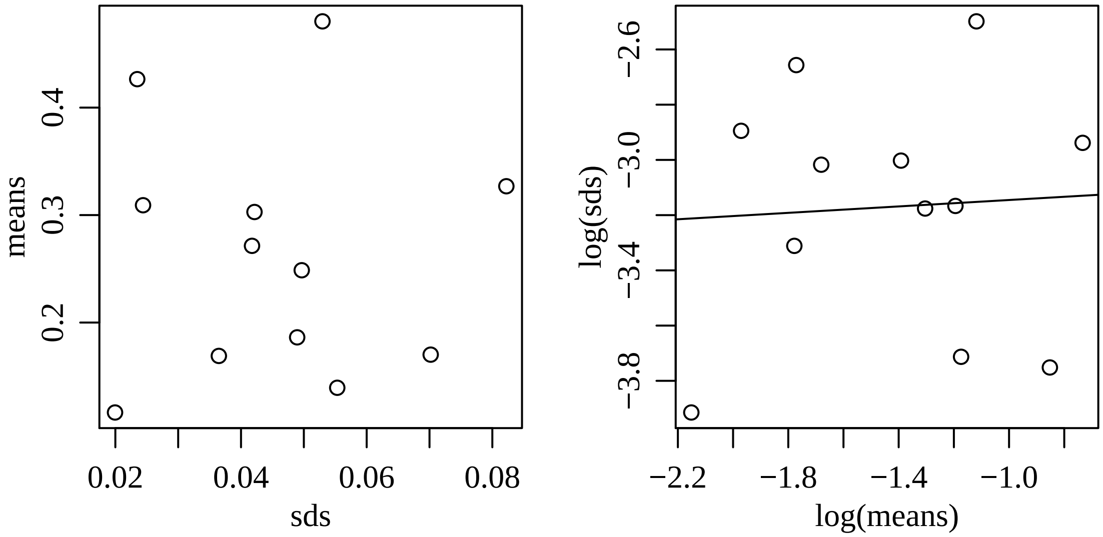


Figure 6.5: Mean-variance relationship for transformed data.

```
Residuals 45 0.30628 0.00681
---

> anova(lm(dat$y~dat$delivery))

               Df  Sum Sq Mean Sq F value    Pr(>F)
dat$delivery    3 0.20414 0.06805  6.6401 0.0008496 ***
Residuals      44 0.45091 0.01025
---
```

- A one-factor ANOVA: There are $3 \times 4 = 12$ different treatments:

```
> anova(lm(dat$y~dat$type:dat$delivery))

                          Df  Sum Sq Mean Sq F value    Pr(>F)
dat$type:dat$delivery     11 0.56862 0.05169  21.531 1.289e-12 ***
Residuals                 36 0.08643 0.00240
---
```

Why might these methods be insufficient?

- What are the SSE, MSE representing in the first two ANOVAS? Why are they bigger than the value in the in the third ANOVA?

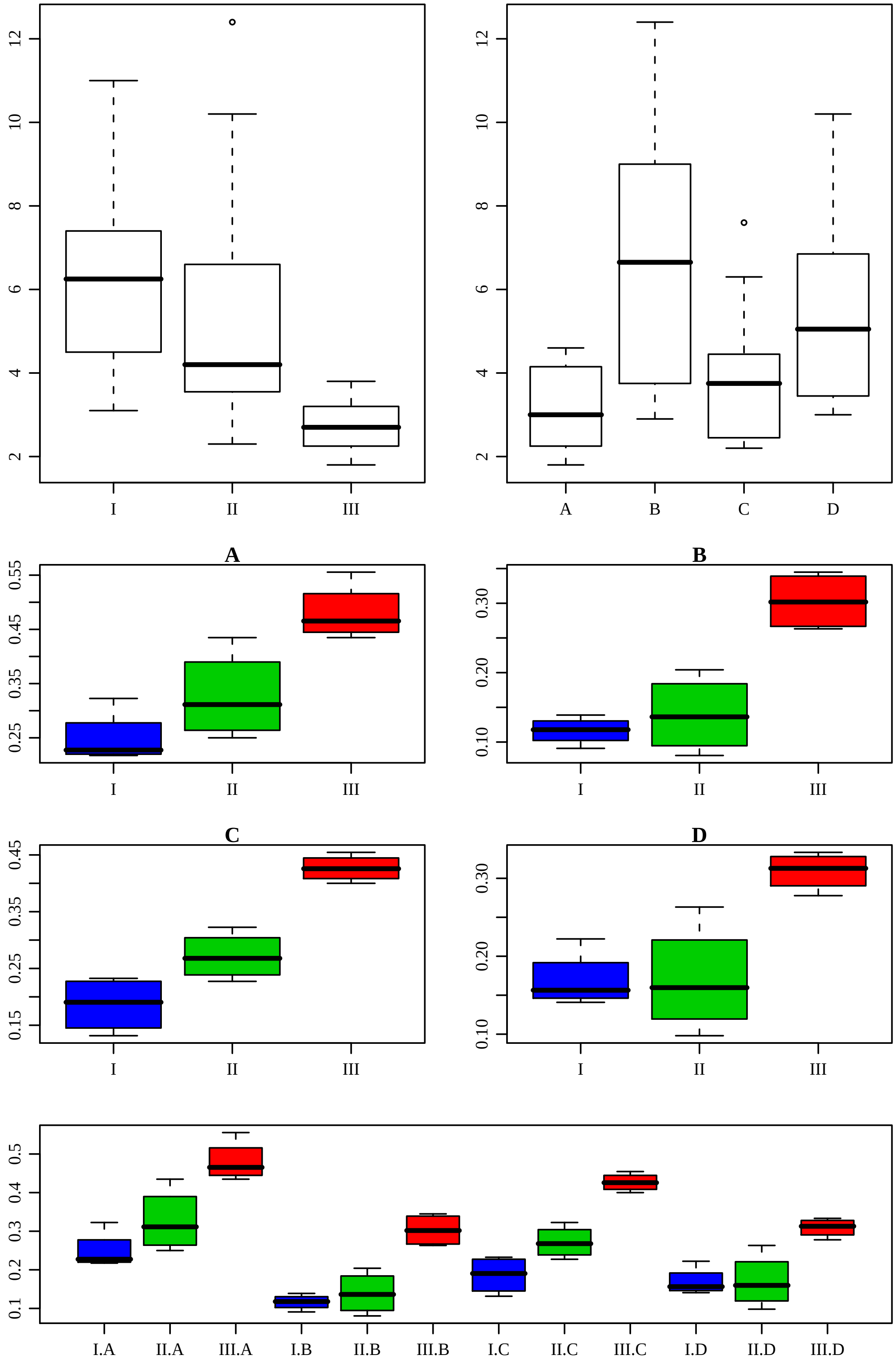


Figure 6.6: Plots of transformed poison data

- In the third ANOVA, can we assess the effects of Type and Delivery separately?
- Can you think of a situation where the $F$-stats in the first and second ANOVAs would be "small", but the $F$-stat in the third ANOVA "big"?

Basically, the first and second ANOVAs may mischaracterize the data and sources of variation. The third ANOVA is "valid," but we'd like a more specific result: we'd like to know which factors are sources of variation, and the relative magnitude of their effects. Also, if the effects of one factor are consistent across levels of the other, maybe we don't need to have a separate parameter for each of the 12 treatment combinations, i.e. a simpler model may suffice.

## 6.2 Additive effects model

$$Y_{i,j,k} = \mu + a_i + b_j + \epsilon_{i,j,k}, \qquad i = 1, \dots, m_1,\ j = 1, \dots, m_2,\ k = 1, \dots, n$$

$\mu$ = overall mean;

$a_1, \dots, a_{m_1}$ = **additive** effects of factor 1;

$b_1, \dots, b_{m_2}$ = **additive** effects of factor 2.

**Notes:**

1. Side conditions: As with with the treatment effects model in the one-factor case, we only need $m_1 - 1$ parameters to differentiate between $m_1$ means, so we usually
   - restrict $a_1 = 0$, $b_1 = 0$ (set-to-zero side conditions ), OR
   - restrict $\sum a_i = 0$, $\sum b_j = 0$ (sum-to-zero side conditions).
2. The additive model is a *reduced model* : There are $m_1 \times m_2$ groups or treatment combinations, and a *full model* fits a different population mean separately to each treatment combination, requiring $m_1 \times m_2$ parameters. In contrast, the additive model only has

| | |
|---|---|
| 1 | parameter for $\mu$ |
| $m_1 - 1$ | parameters for $a_i$'s |
| $m_2 - 1$ | parameters for $b_j$'s |
| $m_1 + m_2 - 1$ | parameters total. |

**Parameter estimation and ANOVA decomposition:**

$$\begin{array}{rcccccccc} y_{ijk} & = & \bar{y}_{\cdot\cdot\cdot} & + & (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) & + & (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) & + & (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) \\ & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j & + & \hat{\epsilon}_{ijk} \end{array}$$

These are the least-squares parameter estimates, under the *sum-to-zero* side conditions:

$$\sum \hat{a}_i = \sum (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) = n\bar{y}_{\cdot\cdot\cdot} - n\bar{y}_{\cdot\cdot\cdot} = 0$$

To obtain the **set-to-zero** side conditions, add $\hat{a}_1$ and $\hat{b}_1$ to $\hat{\mu}$, subtract $\hat{a}_1$ from the $\hat{a}_i$'s, and subtract $\hat{b}_1$ from the $\hat{b}_j$'s. Note that this does not change the fitted value in each group:

$$\begin{array}{rcccccc} \text{fitted}(y_{ijk}) & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j \\ & = & (\hat{\mu} + \hat{a}_1 + \hat{b}_1) & + & (\hat{a}_i - \hat{a}_1) & + & (\hat{b}_j - \hat{b}_1) \\ & = & \hat{\mu}^* & + & \hat{a}_i^* & + & \hat{b}_j^* \end{array}$$

As you might have guessed, we can write this decomposition out as vectors of length $m_1 \times m_2 \times n$:

$$\begin{array}{ccccccc} \boldsymbol{y} - \bar{y}_{\cdot\cdot\cdot} & = & \hat{\boldsymbol{a}} & + & \hat{\boldsymbol{b}} & + & \hat{\boldsymbol{\epsilon}} \\ v_T & = & v_1 & + & v_2 & + & v_e \end{array}$$

The columns represent

$v_T$ variation of the data around the grand mean;

$v_1$ variation of factor 1 means around the grand mean;

$v_2$ variation of factor 2 means around the grand mean;

$v_e$ variation of the data around fitted the values.

You should be able to show that these vectors are orthogonal, and so

$$\begin{array}{ccccccc} \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2 & = & \sum_i \sum_j \sum_k \hat{a}_i^2 & + & \sum_i \sum_j \sum_k \hat{b}_i^2 & + & \sum_i \sum_j \sum_k \hat{\epsilon}_i^2 \\ \text{SSTotal} & = & \text{SSA} & + & \text{SSB} & + & \text{SSE} \end{array}$$

**Degrees of Freedom:**

- $\hat{\boldsymbol{a}}$ contains $m_1$ different numbers but sums to zero → $m_1 - 1$ dof
- $\hat{\boldsymbol{b}}$ contains $m_2$ different numbers but sums to zero → $m_2 - 1$ dof

### ANOVA table

| Source | SS | df | MS | F |
|---|---|---|---|---|
| A | SSA | $m_1 - 1$ | SSA/df$_\mathrm{A}$ | MSA/MSE |
| B | SSB | $m_2 - 1$ | SSB/df$_\mathrm{B}$ | MSB/MSE |
| Error | SSE | $(m_1 - 1)(m_2 - 1) + m_1 m_2 (n - 1)$ | SSE/df$_\mathrm{E}$ | |
| Total | SSTotal | $m_1 m_2 n - 1$ | | |

```
> anova(lm(dat$y~dat$type+dat$delivery))
Analysis of Variance Table

Response: dat$y
             Df  Sum Sq Mean Sq F value
dat$type      2 0.34877 0.17439  71.708
dat$delivery  3 0.20414 0.06805  27.982
Residuals    42 0.10214 0.00243
```

This ANOVA has decomposed the variance in the data into the variance of **additive** Type effects, **additive** Delivery effects, and residuals. Does this adequately represent what is going on in the data? What do we mean by additive? Assuming the model is correct, we have:

$$\begin{aligned} \mathrm{E}[Y|\text{type=I, delivery=A}] &= \mu + a_1 + b_1 \\ \mathrm{E}[Y|\text{type=II, delivery=A}] &= \mu + a_2 + b_1 \end{aligned}$$

This says that the difference between Type I and Type II is $a_1 - a_2$ regardless of Delivery. Does this look right based on the plots? Consider the following table:

| | Effect of Type I vs II, given Delivery | |
|---|---|---|
| Delivery | full model | additive model |
| A | $\mu_{IA} - \mu_{IIA}$ | $(\mu + a_1 + b_1) - (\mu + a_2 + b_1) = a_1 - a_2$ |
| B | $\mu_{IB} - \mu_{IIB}$ | $(\mu + a_1 + b_2) - (\mu + a_2 + b_2) = a_1 - a_2$ |
| C | $\mu_{IC} - \mu_{IIC}$ | $(\mu + a_1 + b_3) - (\mu + a_2 + b_3) = a_1 - a_2$ |
| D | $\mu_{ID} - \mu_{IID}$ | $(\mu + a_1 + b_4) - (\mu + a_2 + b_4) = a_1 - a_2$ |

- The full model allows differences between Types to vary across levels of Delivery

- The reduced/additive model says differences are constant across levels of Delivery.

Therefore, the reduced model is appropriate if

$$(\mu_{IA} - \mu_{IIA}) = (\mu_{IB} - \mu_{IIB}) = (\mu_{IC} - \mu_{IIC}) = (\mu_{ID} - \mu_{IID})$$

How can we test for this? Consider the following **parameterization** of the full model:

**Interaction model:**

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}$$

$\mu$ = overall mean;

$a_1, \ldots, a_{m_1}$ = **additive** effects of factor 1;

$b_1, \ldots, b_{m_2}$ = **additive** effects of factor 2.

$(ab)_{ij}$ = **interaction terms** = **deviations from additivity**.

The *interaction term* is a correction for non-additivity of the factor effects. This is a full model: It fits a separate mean for each treatment combination:

$$E(Y_{ijk}) = \mu_{ij} = \mu + a_i + b_j + (ab)_{ij}$$

**Parameter estimation and ANOVA decomposition:**

$$\begin{array}{rcccccccccc} y_{ijk} & = & \bar{y}_{\cdot\cdot\cdot} & + & (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) & + & (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) & + & (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) & + & (y_{ijk} - \bar{y}_{ij\cdot}) \\ & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j & + & \widehat{(ab)}_{ij} & + & \hat{\epsilon}_{ijk} \end{array}$$

Note that the interaction term is equal to the fitted value under the full model ($y_{ij}$) minus the fitted value under the additive model ($\bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$). Deciding between the additive/reduced model and the interaction/full model is tantamount to deciding if the variance explained by the $\widehat{(ab)}_{ij}$'s is large or not, i.e. whether or not the full model is close to the additive model.

## 6.3 Evaluating additivity:

Recall the additive model

$$y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$$

- $i = 1, \dots, m_1$ indexes the levels of factor 1
- $j = 1, \dots, m_2$ indexes the levels of factor 1
- $k = 1, \dots, n$ indexes the replications for each of the $m_1 \times m_2$ treatment combinations.

Parameter estimates are obtained via the following decomposition:

$$\begin{array}{ccccccccc} y_{ijk} & = & \bar{y}_{\cdot\cdot\cdot} & + & (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) & + & (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) & + & (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) \\ & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j & + & \hat{\epsilon}_{ijk} \end{array}$$

**Fitted value:**

$$\begin{aligned} \hat{y}_{ijk} &= \hat{\mu} + \hat{a}_i + \hat{b}_j \\ &= \bar{y}_{\cdot\cdot\cdot} + (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) \\ &= \bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot} \end{aligned}$$

Note that in this model the fitted value in one cell depends on data from the others.

**Residual:**

$$\begin{aligned} \hat{\epsilon}_{ijk} &= y_{ijk} - \hat{y}_{ijk} \\ &= (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) \end{aligned}$$

As we discussed, this model is "adequate" if the differences across levels of one factor don't depend on the level of the other factor, i.e.

- $\bar{y}_{1j\cdot} - \bar{y}_{2j\cdot} \approx \hat{a}_1 - \hat{a}_2$ for all $j = 1, \dots, m_2$, and
- $\bar{y}_{i1\cdot} - \bar{y}_{i2\cdot} \approx \hat{b}_1 - \hat{b}_2$ for all $i = 1, \dots, m_1$.

This adequacy can be assessed by looking differences between the sample means from each cell and the *fitted* values of these averages under the additive model.

$$\begin{aligned} \bar{y}_{ij\cdot} - (\hat{\mu} + \hat{a}_i + \hat{b}_j) &= \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot} \\ &\equiv \widehat{(ab)}_{ij} \end{aligned}$$

The term $\widehat{(ab)}_{ij}$ measures the deviation of the cell means to the estimated additive model. It is called an *interaction.* It does not measure how factor 1 "interacts" with factor 2. It measures how much the data deviate from the additive effects model.

**Interactions and the full model:** The interaction terms also can be derived by taking the additive decomposition above one step further: The residual in the additive model can be written:

$$\begin{aligned}\hat{\epsilon}^A_{ijk} &= y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot} \\ &= (y_{ijk} - \bar{y}_{ij\cdot}) + (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) \\ &= \hat{\epsilon}^I_{ijk} + \hat{(ab)}_{ij}\end{aligned}$$

This suggests the following *full model*, or *interaction model*

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}$$

with parameter estimates obtained from the following decomposition:

$$\begin{array}{ccccccccccc} y_{ijk} &=& \bar{y}_{\cdot\cdot\cdot} &+& (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) &+& (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) &+& (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) &+& (y_{ijk} - \bar{y}_{ij\cdot}) \\ &=& \hat{\mu} &+& \hat{a}_i &+& \hat{b}_j &+& (\hat{ab})_{ij} &+& \hat{\epsilon}_{ijk}\end{array}$$

**Fitted value:**

$$\begin{aligned}\hat{y}_{ijk} &= \hat{\mu} + \hat{a}_i + \hat{b}_j + (\hat{ab})_{ij} \\ &= \bar{y}_{ij\cdot} = \hat{\mu}_{ij}\end{aligned}$$

> This is a *full model* for the treatment means: The estimate of the mean in each cell depends only on data from that cell. Contrast this to additive model.

**Residual:**

$$\begin{aligned}\hat{\epsilon}_{ijk} &= y_{ijk} - \hat{y}_{ijk} \\ &= y_{ijk} - \bar{y}_{ij\cdot}\end{aligned}$$

Thus the full model ANOVA decomposition partitions the variability among the cell means $\bar{y}_{11\cdot}, \bar{y}_{12\cdot}, \ldots, \bar{y}_{m_1 m_2\cdot}$ into

- the overall mean $\hat{\mu}$
- additive treatment effects $\hat{a}_i$ , $\hat{b}_j$
- what is “leftover” : $\hat{(ab)}_{ij} = \bar{y}_{ij\cdot} - (\hat{\mu} + \hat{a}_i + \hat{b}_j)$

As you might expect, these different parts are orthogonal, resulting in the following orthogonal decomposition of the variance.

<table>
<tr><td colspan="2"></td><td colspan="2">var explained by add model</td><td>+</td><td colspan="3">error in add model</td></tr>
<tr><td colspan="2"></td><td colspan="4">var explained by full model</td><td>+</td><td>error in full model</td></tr>
<tr><td>Total SS</td><td>=</td><td>SSA +</td><td>SSB</td><td>+</td><td>SSAB</td><td>+</td><td>SSE</td></tr>
</table>

**Example: (Poison)** $3 \times 4$ two-factor CRD with 4 reps per treatment combination.

```
            Df  Sum Sq Mean Sq F value
pois$deliv   3 0.20414 0.06805  27.982
pois$type    2 0.34877 0.17439  71.708
Residuals   42 0.10214 0.00243


                      Df  Sum Sq Mean Sq F value
pois$deliv             3 0.20414 0.06805 28.3431
pois$type              2 0.34877 0.17439 72.6347
pois$deliv:pois$type   6 0.01571 0.00262  1.0904
Residuals             36 0.08643 0.00240
```

So notice

- $0.10214 = 0.01571 + 0.08643$, that is $\mathrm{SSE}_{\mathrm{add}} = \mathrm{SSAB}_{\mathrm{int}} + \mathrm{SSE}_{\mathrm{int}}$
- $42 = 6 + 36$, that is $\mathrm{dof}(\mathrm{SSE}_{\mathrm{add}}) = \mathrm{dof}(\mathrm{SSAB}_{\mathrm{int}}) + \mathrm{dof}(\mathrm{SSE}_{\mathrm{int}})$
- SSA, SSB, dof(A), dof(B) are unchanged in the two models
- $\mathrm{MSE}_{\mathrm{add}} \approx \mathrm{MSE}_{\mathrm{int}}$ , but degrees of freedom are larger in the additive model. Which do you think is a better estimate of the within-group variance?

**Expected sums of squares:**

- If $\mathrm{H}_0 : (ab)_{ij} = 0$ is true, then
  - $\mathrm{E[MSE]} = \sigma^2$
  - $\mathrm{E[MSAB]} = \sigma^2$
- If $\mathrm{H}_0 : (ab)_{ij} = 0$ is not true, then
  - $\mathrm{E[MSE]} = \sigma^2$

  - $\mathrm{E[MSAB]} = \sigma^2 + \mathrm{r}\tau^2_{\mathrm{AB}} > \sigma^2$.

This suggests

- An evaluation of the adequacy of the additive model can be assessed by comparing MSAB to MSE. Under $\mathrm{H}_0 : (ab)_{ij} = 0$ ,

  $$F_{\mathrm{AB}} = \mathrm{MSAB/MSE} \sim F_{(m_1-1)\times(m_2-1),m_1m_2(n-1)}$$

  Evidence against $\mathrm{H}_0$ can be evaluated by computing the $p$-value.

- If the additive model is adequate then $\mathrm{MSE_{int}}$ and MSAB are two independent estimates of roughly the same thing (why independent?). We may then want to combine them to improve our estimate of $\sigma^2$.

```
                            Df  Sum Sq Mean Sq F value    Pr(>F)
pois$deliv                   3 0.20414 0.06805 28.3431 1.376e-09 ***
pois$type                    2 0.34877 0.17439 72.6347 2.310e-13 ***
pois$deliv:pois$type         6 0.01571 0.00262  1.0904    0.3867
Residuals                   36 0.08643 0.00240
```

For these data, there is strong evidence of both treatment effects, and little evidence of non-additivity. We may want to use the additive model.

## 6.4 Inference for additive treatment effects

Consider a two-factor experiment in which it is determined that the effects of factor $F_1$ and $F_2$ are large. Now we want to compare means across levels of one of the factors.

Recall in the pesticide example we had 4 reps for each of 3 levels of Poison type and 4 levels of Delivery. So we have 4 × 4 = 16 observations for each level of Type.

**The wrong approach:** The two-sample t-test is

$$\frac{\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}}{s_{12}\sqrt{2/(4\times 4)}}$$

For the above example,

- $\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} = 0.047$

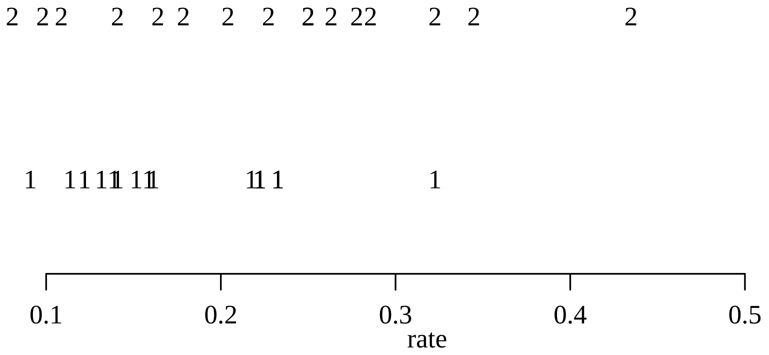


Figure 6.7: Comparison between types I and II, without respect to delivery.

- $s_{12} = 0.081$, $n \times m_2 = 4 \times 4 = 16$.
- t-statistic = -1.638, df=30, $p$-value=0.112.

Questions:

- What is $s_{12}^2$ estimating?
- What should we be comparing the factor level differences to?

If Delivery is a **known** source of variation, we should compare differences between levels of Poison type to variability **within** a treatment combination, i.e. $\sigma^2$. For the above example, $s_{12} = .081$, whereas $s_{\text{MSE}} = \sqrt{0.00240} \approx 0.05$, a ratio of about 1.65.

$$\frac{\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}}{s_{\text{MSE}}\sqrt{2/(4 \times 4)}} = 2.7, \qquad p\text{-value} \approx 0.01$$

**Testing additive effects** Let $\mu_{ij}$ be the population mean in cell $ij$. The relationship between the cell means model and the parameters in the interaction model are as follows:

$$\begin{aligned} \mu_{ij} &= \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot}) + (\mu_{ij} - \mu_{i\cdot} - \mu_{j\cdot} + \mu_{\cdot\cdot}) \\ &= \mu + a_i + b_j + (ab)_{ij} \end{aligned}$$

and so

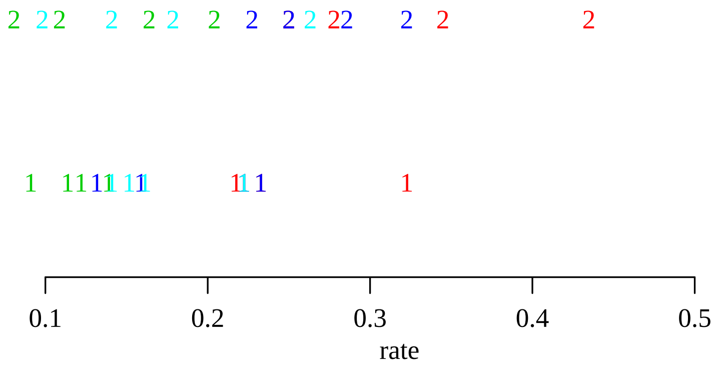


Figure 6.8: Comparison between types I and II, with delivery in color.

- $\sum_i a_i = 0$
- $\sum_j b_j = 0$
- $\sum_i (ab)_{ij} = 0$ for each $j$, $\sum_j (ab)_{ij} = 0$ for each $i$

Suppose we are interested in the difference between $F_1 = 1$ and $F_1 = 2$. With the above representation we can show that

$$\begin{aligned}
\frac{1}{m_2}\sum_j \mu_{1j} - \frac{1}{m_2}\sum_j \mu_{2j} &= \frac{1}{m_2}\sum_j (\mu_{1j} - \mu_{2j}) \\
&= \frac{1}{m_2}\sum_j ([\mu + a_1 + b_j + (ab)_{1j}] - [\mu + a_2 + b_j + (ab)_{2j}]) \\
&= \frac{1}{m_2}\left(\sum_j (a_1 - a_2) + \sum_j (ab)_{1j} - \sum_j (ab)_{2j}\right) \\
&= a_1 - a_2
\end{aligned}$$

and so the "effect of $F_1 = 1$ versus $F_1 = 2$" $= a_1 - a_2$ can be viewed as a *contrast* of cell means. Note that $a_1 - a_2$ is a *contrast* of treatment (population) means:

| | $F_2 = 1$ | $F_2 = 2$ | $F_2 = 3$ | $F_2 = 4$ | |
|---|---|---|---|---|---|
| $F_1 = 1$ | $\mu_{11}$ | $\mu_{12}$ | $\mu_{13}$ | $\mu_{14}$ | $4\bar{\mu}_{1\cdot}$ |
| $F_1 = 2$ | $\mu_{21}$ | $\mu_{22}$ | $\mu_{23}$ | $\mu_{24}$ | $4\bar{\mu}_{2\cdot}$ |
| $F_1 = 3$ | $\mu_{31}$ | $\mu_{32}$ | $\mu_{33}$ | $\mu_{34}$ | $4\bar{\mu}_{3\cdot}$ |
| | $3\bar{\mu}_{\cdot 1}$ | $3\bar{\mu}_{\cdot 2}$ | $3\bar{\mu}_{\cdot 3}$ | $3\bar{\mu}_{\cdot 4}$ | $12\bar{\mu}_{\cdot\cdot}$ |

So

$$\begin{aligned} a_1 - a_2 &= \bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot} \\ &= (\mu_{11} + \mu_{12} + \mu_{13} + \mu_{14})/4 - (\mu_{21} + \mu_{22} + \mu_{23} + \mu_{24})/4 \end{aligned}$$

Like any contrast, we can estimate/make inference for it using contrasts of sample means:

$$a_1 - a_2 = \hat{a}_1 - \hat{a}_2 = \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} \text{ is an unbiased estimate of } a_1 - a_2$$

Note that this estimate is the corresponding *contrast* among the $m_1 \times m_2$ sample means:

| | $F_2 = 1$ | $F_2 = 2$ | $F_2 = 3$ | $F_2 = 4$ | |
|---|---|---|---|---|---|
| $F_1 = 1$ | $\bar{y}_{11\cdot}$ | $\bar{y}_{12\cdot}$ | $\bar{y}_{13\cdot}$ | $\bar{y}_{14\cdot}$ | $4\bar{y}_{1\cdot\cdot}$ |
| $F_1 = 2$ | $\bar{y}_{21\cdot}$ | $\bar{y}_{22\cdot}$ | $\bar{y}_{23\cdot}$ | $\bar{y}_{24\cdot}$ | $4\bar{y}_{2\cdot\cdot}$ |
| $F_1 = 3$ | $\bar{y}_{31\cdot}$ | $\bar{y}_{32\cdot}$ | $\bar{y}_{33\cdot}$ | $\bar{y}_{34\cdot}$ | $4\bar{y}_{3\cdot\cdot}$ |
| | $3\bar{y}_{\cdot 1\cdot}$ | $3\bar{y}_{\cdot 2\cdot}$ | $3\bar{y}_{\cdot 3\cdot}$ | $3\bar{y}_{\cdot 4\cdot}$ | $12\bar{y}_{\cdot\cdot\cdot}$ |

So

$$\begin{aligned} \hat{a}_1 - \hat{a}_2 &= \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} \\ &= (\bar{y}_{11\cdot} + \bar{y}_{12\cdot} + \bar{y}_{13\cdot} + \bar{y}_{14\cdot})/4 - (\bar{y}_{21\cdot} + \bar{y}_{22\cdot} + \bar{y}_{23\cdot} + \bar{y}_{24\cdot})/4 \end{aligned}$$

Hypothesis tests and confidence intervals can be made using the standard assumptions:

- $\mathrm{E}[\hat{a}_1 - \hat{a}_2] = a_1 - a_2$
- Under the assumption of constant variance:

$$\begin{aligned} \mathrm{Var}[\hat{a}_1 - \hat{a}_2] &= \mathrm{Var}[\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}] \\ &= \mathrm{Var}[\bar{y}_{1\cdot\cdot}] + \mathrm{Var}[\bar{y}_{2\cdot\cdot}] \\ &= \sigma^2/(n \times m_2) + \sigma^2/(n \times m_2) \\ &= 2\sigma^2/(n \times m_2) \end{aligned}$$

- Under the assumption that the data are normally distributed

  $$\hat{a}_1 - \hat{a}_2 \sim \text{normal}(a_1 - a_2, \sigma\sqrt{2/(n \times m_2)})$$

  and is independent of MSE (why?)

- So

  $$\frac{(\hat{a}_1 - \hat{a}_2) - (a_1 - a_2)}{\sqrt{\text{MSE}\frac{2}{n\times m_2}}} = \frac{(\hat{a}_1 - \hat{a}_2) - (a_1 - a_2)}{\text{SE}[\hat{a}_1 - \hat{a}_2]} \sim t_\nu$$

  where $\nu$ are the degrees of freedom associated with our estimate of $\sigma^2$, i.e. the residual degrees of freedom in our model.

  - $\nu = m_1 m_2(n-1)$ under the full/interaction model
  - $\nu = m_1 m_2(n-1) + (m_1 - 1)(m_2 - 1)$ under the reduced/additive model

**Review:** Explain the degrees of freedom for the two models.

**t-test:** Reject $\text{H}_0 : a_1 = a_2$ if

$$\begin{aligned} \frac{\hat{a}_1 - \hat{a}_2}{\sqrt{\text{MSE}\frac{2}{n\times m_2}}} &> t_{1-\alpha_C/2,\nu} \\ |\hat{a}_1 - \hat{a}_2| &> \sqrt{\text{MSE}\frac{2}{n \times m_2}} \times t_{1-\alpha_C/2,\nu} \end{aligned}$$

So the quantity

$$\begin{aligned} \text{LSD}_1 &= t_{1-\alpha_C/2,\nu} \times SE(\hat{\alpha}_1 - \hat{\alpha}_2) \\ &= t_{1-\alpha_C/2,\nu} \times \sqrt{\text{MSE}\frac{2}{n \times m_2}} \end{aligned}$$

is a "yardstick" for comparing levels of factor 1. It is sometimes called the *least significant difference* for comparing levels of Factor 1. It is analogous to the LSD we used in the 1-factor ANOVA.

**Important note:** The LSD depends on which factor you are looking at: The comparison of levels of Factor 2 depends on

$$\begin{aligned} \text{Var}[\bar{y}_{\cdot 1 \cdot} - \bar{y}_{\cdot 2 \cdot}] &= \text{Var}[\bar{y}_{\cdot 1 \cdot}] + \text{Var}[\bar{y}_{\cdot 2 \cdot}] \\ &= \sigma^2/(n \times m_1) + \sigma^2/(n \times m_1) \\ &= 2\sigma^2/(n \times m_1) \end{aligned}$$

So the LSD for factor 2 differences is

$$\text{LSD}_2 = t_{1-\alpha_C/2,\nu} \times \sqrt{\text{MSE}\frac{2}{n \times m_1}}$$

```
                        Df  Sum Sq Mean Sq F value    Pr(>F)
pois$deliv               3 0.20414 0.06805 28.3431 1.376e-09 ***
pois$type                2 0.34877 0.17439 72.6347 2.310e-13 ***
pois$deliv:pois$type     6 0.01571 0.00262  1.0904    0.3867
Residuals               36 0.08643 0.00240
```

There is not very much evidence that the effects are not additive. Let's assume there is no interaction term. If we are correct then we will have increased the precision of our variance estimate.

```
           Df  Sum Sq Mean Sq F value    Pr(>F)
pois$deliv  3 0.20414 0.06805  27.982 4.192e-10 ***
pois$type   2 0.34877 0.17439  71.708 2.865e-14 ***
Residuals  42 0.10214 0.00243
```

So there is strong evidence against the hypothesis that the additive effects are zero for either factor. Which treatments within a factor are different from each other?

**effects of poison type:** At $\alpha_C = 0.05$,

$$\begin{aligned}
\text{LSD}_{\text{type}} &= t_{.975,42} \times \text{SE}[\hat{a}_1 - \hat{a}_2] \\
&= 2.018 \times \sqrt{.0024\frac{2}{4 \times 4}} \\
&= 2.018 \times 0.0173 = 0.035
\end{aligned}$$

| Type | Mean | LSD Grouping |
|---|---|---|
| I | 0.18 | 1 |
| II | 0.23 | 2 |
| III | 0.38 | 3 |

**effects of poison delivery:** At $\alpha_C = 0.05$,

$$\begin{aligned}
\text{LSD}_{\text{type}} &= t_{.975,42} \times \text{SE}[\hat{b}_1 - \hat{b}_2] \\
&= 2.018 \times \sqrt{.0024\frac{2}{4 \times 3}} \\
&= 2.018 \times 0.02 = 0.040
\end{aligned}$$

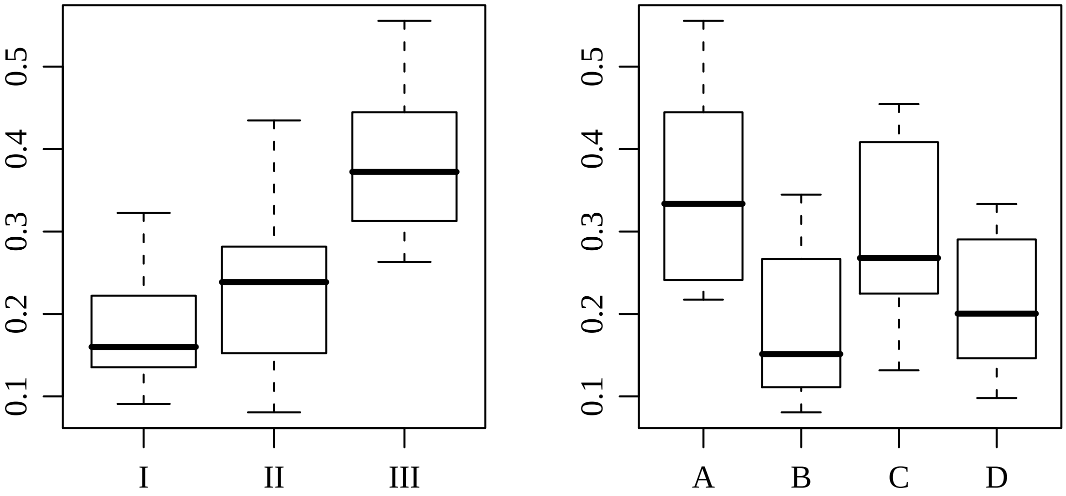


Figure 6.9: Marginal plots of the data.

| Type | Mean | LSD Grouping |
|---|---|---|
| B | 0.19 | 1 |
| D | 0.21 | 1 |
| C | 0.29 | 2 |
| A | 0.35 | 3 |

Note the differences between these comparisons and those from two-sample t-tests.

**Interpretation of estimated additive effects:** If the additive model is clearly wrong, can we still interpret additive effects? The full model is

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

A reparameterization of this model is the interaction model:

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}$$

where

- $\mu = \frac{1}{m_1 m_2} \sum_i \sum_j \mu_{ij} = \bar{\mu}_{\cdot\cdot}$

- $a_i = \frac{1}{m_2}\sum_j(\mu_{ij} - \bar{\mu}_{\cdot\cdot}) = \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot}$
- $b_j = \frac{1}{m_1}\sum_i(\mu_{ij} - \bar{\mu}_{\cdot\cdot}) = \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot\cdot}$
- $(ab)_{ij} = \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot} = \mu_{ij} - (\mu + a_i + b_j)$

The terms $\{a_1, \ldots, a_{m_1}\}$ , $\{b_1, \ldots, b_{m_2}\}$ are sometimes called "main effects". The additive model is

$$y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$$

Sometimes this is called the "main effects" model. If this model is correct, it implies that

$$\begin{aligned}(ab)_{ij} = 0 \;\; \forall i, j \;\; &\Leftrightarrow \\ \mu_{i_1 j} - \mu_{i_2 j} = a_{i1} - a_{i2} \;\; \forall i_1, i_2, j \;\; &\Leftrightarrow \\ \mu_{i j_1} - \mu_{i j_2} = b_{j_1} - b_{j_2} \;\; \forall i, j_1, j_2 &\end{aligned}$$

What are the **estimated additive effects** estimating, in either case?

$$\begin{aligned}\hat{a}_1 - \hat{a}_2 &= (\bar{y}_{1\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) - (\bar{y}_{2\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) \\ &= \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} \\ &= \frac{1}{m_2}\sum_{j=1}^{m_2}\bar{y}_{1j\cdot} - \frac{1}{m_2}\sum_{j=1}^{m_2}\bar{y}_{2j\cdot} \\ &= \frac{1}{m_2}\sum_{j=1}^{m_2}(\bar{y}_{1j\cdot} - \bar{y}_{2j\cdot}) \qquad (2)\end{aligned}$$

Now

$$E[\frac{1}{m_2}\sum_{j=1}^{m_2}(\bar{y}_{1j\cdot} - \bar{y}_{2j\cdot})] = \frac{1}{m_2}\sum_{j=1}^{m_2}(\mu_{1j} - \mu_{2j}) = a_1 - a_2,$$

so $\hat{a}_1 - \hat{a}_2$ is estimating $a_1 - a_2$ regardless if additivity is correct or not. Now, how do we **interpret** this effect?

- Regardless of additivity, $\hat{a}_1 - \hat{a}_2$ can be interpreted as the estimated difference in response between having factor 1 =1 and factor 1=2, **averaged over the experimental levels of factor 2**.

- If additivity is correct, $\hat{a}_1 - \hat{a}_2$ can further be interpreted as the estimated difference in response between having factor 1 =1 and factor 1=2, **for every level of factor 2 in the experiment**.

Statistical folklore suggests that if there is significant non-additivity, then you can't interpret main/additive effects. As we can see, this is not true: the additive effects have a very definite interpretation under the full model. In some cases (block designs, coming up), we may be interested in additive effects even if there is significant interaction.

**Dissecting the interaction:** Sometimes if there is an interaction, we might want to go in and compare individual cell means. Consider the following table of means from a 2 × 3 two-factor ANOVA.

| | | | |
|---|---|---|---|
| $\bar{y}_{11\cdot}$ | $\bar{y}_{12\cdot}$ | $\bar{y}_{13\cdot}$ | $\bar{y}_{14\cdot}$ |
| $\bar{y}_{21\cdot}$ | $\bar{y}_{22\cdot}$ | $\bar{y}_{23\cdot}$ | $\bar{y}_{24\cdot}$ |

A large interaction SS in the ANOVA table gives us evidence, for example, that $\mu_{1j} - \mu_{2j}$ varies across levels $j$ of factor 2. It may be useful to dissect this variability further, and understand how the non-additivity is manifested in the data: For example, consider the three plots in Figure 4.10. These all would give a large interaction SS, but imply very different things about the effects of the factors.

**Contrasts for examining interactions:** Suppose we want to compare the effect of (factor 1=1) to (factor 1=2) across levels of factor 2. This involves contrasts of the form:

$$C = (\mu_{1j} - \mu_{2j}) - (\mu_{1k} - \mu_{2k})$$

This contrast can be estimated with the *sample contrast*:

$$\hat{C} = (\bar{y}_{1j\cdot} - \bar{y}_{2j\cdot}) - (\bar{y}_{1k\cdot} - \bar{y}_{2k\cdot})$$

As usual, the standard error of this contrast is the estimate of its standard deviation:

$$\begin{aligned}\mathrm{Var}[\hat{C}] &= \sigma^2/r + \sigma^2/r + \sigma^2/r + \sigma^2/r = 4\sigma^2/r \\ \mathrm{SE}[\hat{C}] &= 2\sqrt{\mathrm{MSE/n}}\end{aligned}$$

Confidence intervals and t-tests for $C$ can be made in the usual way.

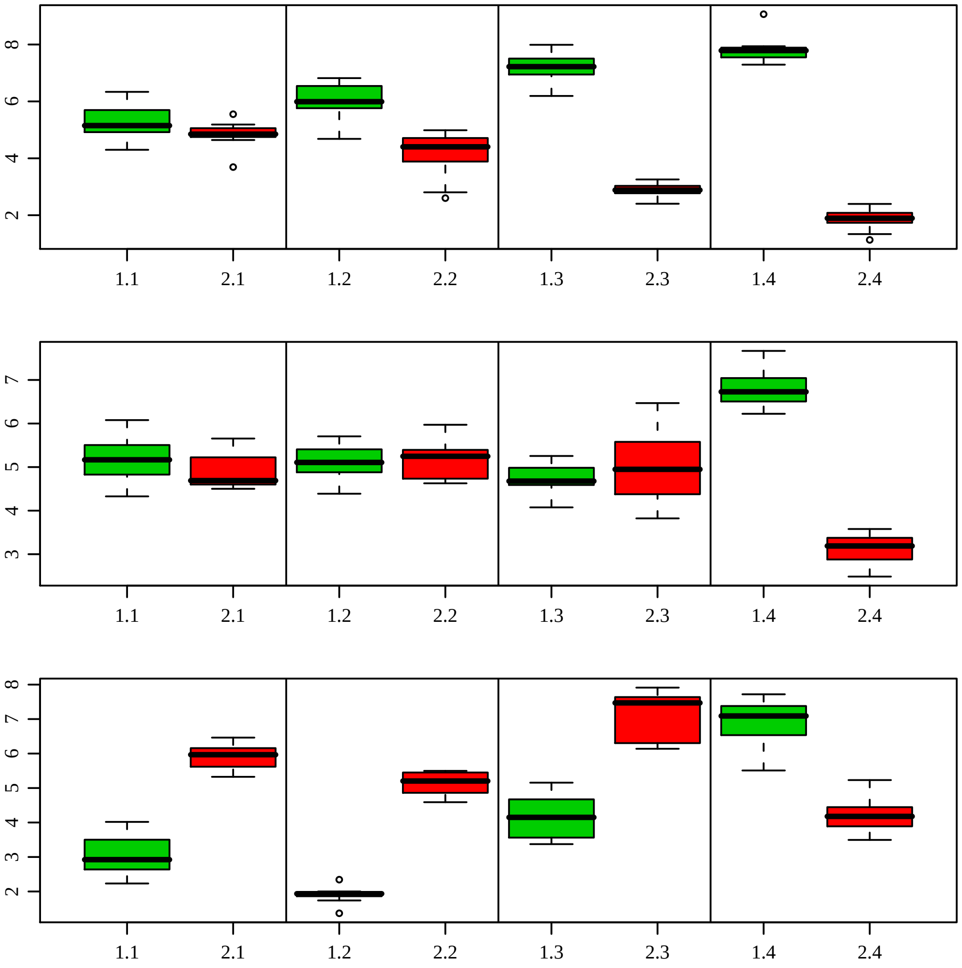


Figure 6.10: Three datasets exhibiting non-additive effects.

## 6.5 Randomized complete block designs

Recall our first design:

**CRD:** to assess effects of a single factor, say $F_1$, on response, we randomly allocate levels of $F_1$ to experimental units. Typically, one hopes the e.u.'s are *homogeneous* or nearly so:

- scientifically: If the units are nearly homogeneous, then any observed variability in response can be attributed to variability in factor levels.
- statistically: If the units are nearly homogeneous, then MSE will be small, confidence intervals will be precise and hypothesis tests powerful.

But what if the e.u.'s are not homogeneous?

**Example:** Let $F_2$ be a factor that is a large source of variation/heterogeneity, but is **not recorded** (age of animals in experiment, gender, field plot conditions, soil conditions).

**ANOVA**

| Source | SS | MS | F-ratio |
|---|---|---|---|
| F1 | SS1 | MS1 =SS1/$(m_1 - 1)$ | MS1/MSE |
| (F2+Error) | SS2+SSE | (SS2+SSE)/$N - m_1$ | |

If SS2 is **large**, F-stat for $F_1$ may be **small**.

If a factor

1. affects response
2. varies across experimental units

then it will increase the variance in response and also the experimental error variance/MSE if unaccounted for. If $F_2$ is a known, potentially large source of variation, we can control for it pre-experimentally with a **block design**.

**Blocking:** The stratification of experimental units into groups that are more homogeneous than the whole.

**Objective:** To have less variation among units within blocks than between blocks.

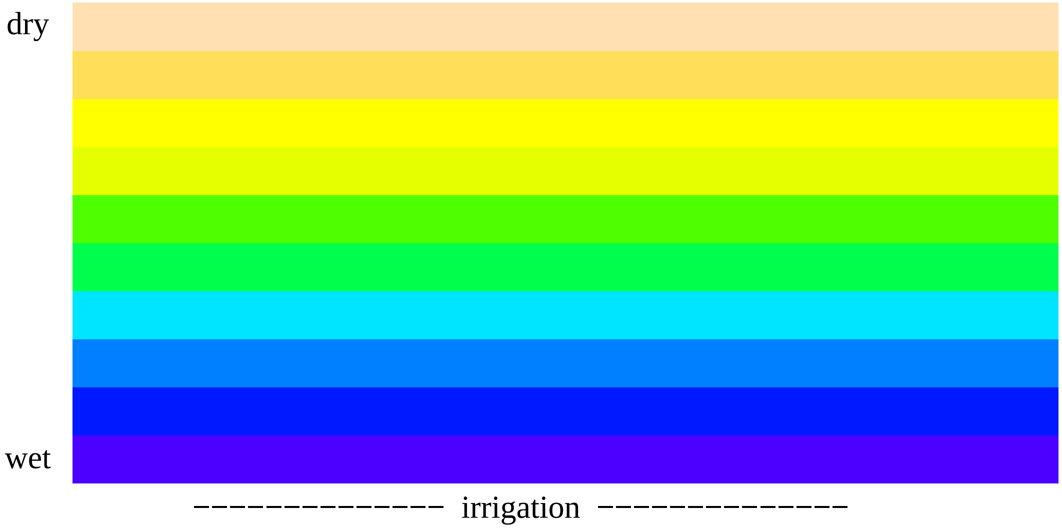


Figure 6.11: Experimental material in need of blocking.

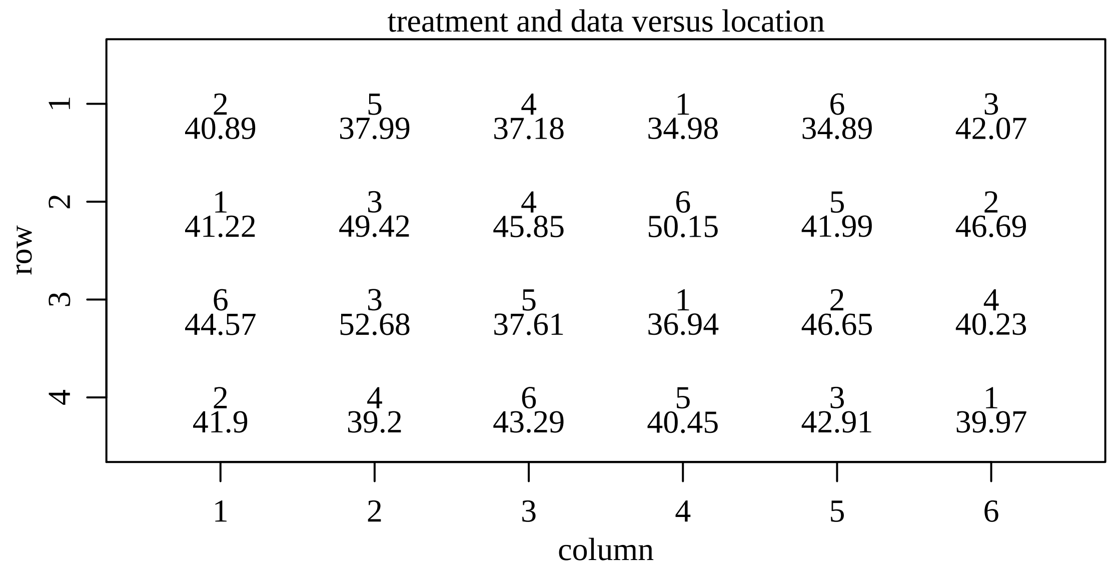
**Typical blocking criteria:**

- location
- physical characteristics
- time

**Example(Nitrogen fertilizer timing):** How does the timing of nitrogen additive affect nitrogen uptake?

- Treatment: Six different timing schedules $1, \ldots, 6$: Level 4 is "standard"
- Response: Nitrogen uptake (ppm$\times 10^{-2}$ )
- Experimental material: One irrigated field

Soil moisture is thought to be a source of variation in response.



| row \ column | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 2<br>40.89 | 5<br>37.99 | 4<br>37.18 | 1<br>34.98 | 6<br>34.89 | 3<br>42.07 |
| 2 | 1<br>41.22 | 3<br>49.42 | 4<br>45.85 | 6<br>50.15 | 5<br>41.99 | 2<br>46.69 |
| 3 | 6<br>44.57 | 3<br>52.68 | 5<br>37.61 | 1<br>36.94 | 2<br>46.65 | 4<br>40.23 |
| 4 | 2<br>41.9 | 4<br>39.2 | 6<br>43.29 | 5<br>40.45 | 3<br>42.91 | 1<br>39.97 |

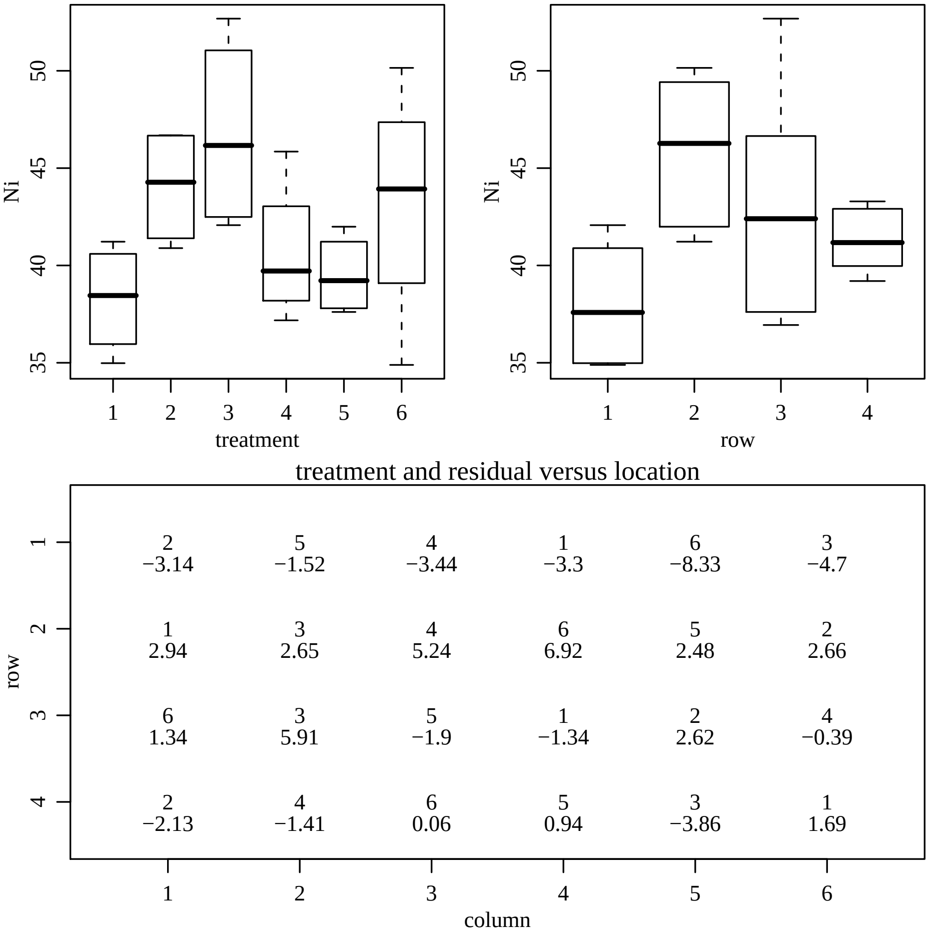
Figure 6.12: Results of the experiment

**Design:**

1. Field is divided into a $4 \times 6$ grid.
2. Within each row or *block*, each of the 6 treatments are randomly allocated.

1. The experimental units are *blocked* into presumably more homogeneous groups.
2. The blocks are *complete*, in that each treatment appears in each block.
3. The blocks are *balanced*, in that there are
   - $m_1 = 6$ observations for each level of block
   - $m_2 = 4$ observations for each level of trt
   - $n = 1$ observation for each trt$\times$block combination.

This design is a *randomized complete block design.*



treatment and residual versus location

| row | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 2<br>−3.14 | 5<br>−1.52 | 4<br>−3.44 | 1<br>−3.3 | 6<br>−8.33 | 3<br>−4.7 |
| 2 | 1<br>2.94 | 3<br>2.65 | 4<br>5.24 | 6<br>6.92 | 5<br>2.48 | 2<br>2.66 |
| 3 | 6<br>1.34 | 3<br>5.91 | 5<br>−1.9 | 1<br>−1.34 | 2<br>2.62 | 4<br>−0.39 |
| 4 | 2<br>−2.13 | 4<br>−1.41 | 6<br>0.06 | 5<br>0.94 | 3<br>−3.86 | 1<br>1.69 |

Figure 6.13: Marginal plots, and residuals without controlling for row.

**Analysis of the RCB design with one rep:** Analysis proceeds just as in the two-factor ANOVA:

$$\begin{array}{ccccccc} y_{ij} - \bar{y}_{\cdot\cdot} & = & (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) & + & (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) & + & (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}) \\ \text{SSTotal} & = & \text{SSTrt} & + & \text{SSB} & + & \text{SSE} \end{array}$$

**ANOVA table**

| Source | SS | dof | MS | F-ratio |
|---|---|---|---|---|
| Trt | SST | $m_1 - 1$ | SST$/(m_1 - 1)$ | MST/MSE |
| Block | SSB | $m_2 - 1$ | SSB$/(m_2 - 1)$ | (MSB/MSE) |
| Error | SSE | $(m_1 - 1)(m_2 - 1)$ | SSE$/(m_1 - 1)(m_2 - 1)$ | |
| Total | SSTotal | $m_1 m_2 - 1$ | | |

**ANOVA for nitrogen example:**

| Source | SS | dof | MS | F-ratio | $p$-value |
|---|---|---|---|---|---|
| Trt | 201.32 | 5 | 40.26 | 5.59 | 0.004 |
| Block | 197.00 | 3 | 65.67 | 9.12 | |
| Error | 108.01 | 15 | 7.20 | | |
| Total | 506.33 | 23 | | | |

**Discussion:** Consider the following three ANOVA decompositions:

```
#######
> anova(lm(c(y)~as.factor(c(trt))    ))
                  Df  Sum Sq Mean Sq F value  Pr(>F)
as.factor(c(trt))  5 201.316  40.263  2.3761 0.08024 .
Residuals         18 305.012  16.945
#######

#######
> anova(lm(c(y)~as.factor(c(trt)) + as.factor(c(rw))   ))
                  Df  Sum Sq Mean Sq F value   Pr(>F)
as.factor(c(trt))  5 201.316  40.263  5.5917 0.004191 **
as.factor(c(rw))   3 197.004  65.668  9.1198 0.001116 **
Residuals         15 108.008   7.201
#######

#######
> anova(lm(c(y)~as.factor(c(trt)):as.factor(c(rw))   ))
                                   Df Sum Sq Mean Sq F value Pr(>F)
```

```
as.factor(c(trt)):as.factor(c(rw)) 23  506.33    22.01
Residuals                           0    0.00
#######
```

Can we test for interaction? Do we care about interaction in this case, or just main effects? Suppose it were true that "in row 2, timing 6 is significantly better than timing 4, but in row 3, treatment 3 is better." Is this relevant in for recommending a timing treatment for other fields?

**Did blocking help?** Consider CRD as an alternative:

| | | | | | | |
|---|---|---|---|---|---|---|
| block 1 | 2 | 4 | 3 | 2 | 1 | 4 |
| block 2 | 5 | 5 | 3 | 4 | 1 | 4 |
| block 3 | 6 | 3 | 4 | 2 | 6 | 5 |
| block 4 | 1 | 2 | 6 | 2 | 5 | 6 |

- Advantages:
  - more possible treatment assignments, so power is increased in a randomization test.
  - If we don't estimate block effects, we'll have more dof for error.
- Disadvantages:
  - It is possible, (but unlikely) that some treatment level will get assigned many times to a "good" row, leading to post-experimental bias.
  - If "row" is a big source of variation, then ignoring it may lead to an overly large MSE.

Consider comparing the $F$-statistic from a CRD with that from an RCB: According to Cochran and Cox (1957)

$$\begin{aligned}
\text{MSE}_{\text{crd}} &= \frac{\text{SSB} + n(m-1)\text{MSE}_{\text{rcbd}}}{nm-1} \\
&= \text{MSB}\left(\frac{n-1}{nm-1}\right) + \text{MSE}_{\text{rcbd}}\left(\frac{n(m-1)}{nm-1}\right)
\end{aligned}$$

In general, the effectiveness of blocking is a function of $\text{MSE}_{\text{crd}}/\text{MSE}_{\text{rcb}}$. If this is large, it is worthwhile to block. For the nitrogen example, this ratio is about 2.

## 6.6 Unbalanced designs

**Example:** Observational study of 20 fatal accidents.

- Response: $y$ = speed in excess of speed limit
- Recorded sources of variation:
  1. $R$ =rainy (rainy/not rainy)
  2. $I$ =interstate (interstate/two-lane highway),

| | cell means | | sum | marginal means |
|---|---|---|---|---|
| | interstate | two-lane | | |
| rainy | 15 | 5 | 130 | 13 |
| | $n_{11} = 8$ | $n_{12} = 2$ | $n_{1\cdot} = 10$ | |
| not rainy | 20 | 10 | 120 | 12 |
| | $n_{21} = 2$ | $n_{22} = 8$ | $n_{1\cdot} = 10$ | |
| sum | 160 | 90 | 250 | |
| | $n_{\cdot 1} = 10$ | $n_{\cdot 2} = 10$ | $n_{\cdot\cdot} = 20$ | |
| marginal mean | 16 | 9 | | $\bar{y}_{\cdot\cdot\cdot} = 12.5$ |

Let's naively compute sums of squares based on the decomposition:

$$\begin{aligned} y_{ijk} &= \bar{y}_{\cdot\cdot\cdot} + (\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot}) + (y_{ijk} - \bar{y}_{ij\cdot}) \\ y_{ijk} &= \bar{y}_{\cdot\cdot\cdot} + (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) + (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}) + (y_{ijk} - \bar{y}_{ij\cdot}) \end{aligned}$$

$$\begin{aligned} \text{SSF} &= \sum_i \sum_j \sum_k (\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 = \sum_i \sum_j n_{ij} (\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 \\ &= 8(15 - 12.5)^2 + 2(5 - 12.5)^2 + 2(20 - 12.5)^2 + 8(10 - 12.5)^2 = 325 \\ \text{SS1} &= 10 \times (13 - 12.5)^2 + 10 \times (12 - 12.5)^2 = 5 \\ \text{SS2} &= 10 \times (16 - 12.5)^2 + 10 \times (10 - 12.5)^2 = 245 \end{aligned}$$

Previously we had SS1+SS2+SS12 =SSF, so it seems we should have

$$\text{SS12} = \text{SSF} - \text{SS1} - \text{SS2} = 325 - 5 - 245 = 75$$

This suggests some interaction. But look at the cell means:

- "no rain" - "rain" =5 regardless of interstate

- "interstate" - "two=lane" =10 regardless of rain

There is absolutely no interaction! The problem is that SS1, SS2, SS12 are not orthogonal (as computed this way) and so SSF $\neq$ SS1 + SS2 + SSR12. There are other things to be careful about too:

**Unbalanced marginal means:** Supposed someone asked, "are the accident speeds higher on rainy days or on clear days?"

- $\bar{y}_{\text{rain}} - \bar{y}_{\text{clear}} = 13 - 12 = 1$: Marginal means suggest speeds were slightly higher on average during rainy days than on clear days.
- Cell means show, for **both** road types, speeds were higher on **clear** days by 5 mph on average.

**Explanation:** Cell frequencies are varying.

- Marginal means for **rain** are dominated by **interstate** accidents
- Marginal means for **no rain** are dominated by **two-lane** accidents

We say that the marginal effects are **unbalanced**. How can we make sense of marginal effects in such a situation? How can we test for non-additivity?

**Least-squares means:** The solution to this paradox is to use "least squares means", which are based on on the cell means model:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

$$\hat{\mu}_{ij} = \frac{1}{n_{ij}} \sum_k y_{ijk}, \qquad \hat{\sigma}^2 = s^2 = \frac{1}{N - m_1 m_2} \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2$$

The least-squares marginal means are:

$$\begin{aligned} \hat{\mu}_{i\cdot} &= \frac{1}{m_2} \sum_{j=1}^{m_2} \hat{\mu}_{ij} \neq \bar{y}_{i\cdot\cdot} \\ \hat{\mu}_{\cdot j} &= \frac{1}{m_1} \sum_{i=1}^{m_1} \hat{\mu}_{ij} \neq \bar{y}_{\cdot j\cdot} \end{aligned}$$

**The idea:**

1. get estimates for each cell
2. average cell estimates to get marginal estimates.

| | 1 | 2 | $\cdots$ | $m_2$ | |
|---|---|---|---|---|---|
| 1 | $\bar{y}_{11\cdot}$ | $\bar{y}_{12\cdot}$ | $\cdots$ | $\bar{y}_{1m_2\cdot}$ | $\hat{\mu}_{1\cdot}$ |
| 2 | $\bar{y}_{21\cdot}$ | $\bar{y}_{22\cdot}$ | $\cdots$ | $\bar{y}_{2m_2\cdot}$ | $\hat{\mu}_{2\cdot}$ |
| | $\vdots$ | $\vdots$ | | | |
| $m_1$ | $\bar{y}_{m_11\cdot}$ | $\bar{y}_{m_12\cdot}$ | $\cdots$ | $\bar{y}_{m_1m_2\cdot}$ | $\hat{\mu}_{m_1\cdot}$ |
| | $\hat{\mu}_{\cdot1}$ | $\hat{\mu}_{\cdot2}$ | $\cdots$ | $\hat{\mu}_{\cdot m_2}$ | |

**Accident example:**

| | interstate | two-lane | marginal mean | LS mean |
|---|---|---|---|---|
| rainy | 15 | 5 | 13 | 10 |
| not rainy | 20 | 10 | 12 | 15 |
| marginal mean | 16 | 9 | | |
| LS mean | 17.5 | 7.5 | | |

Comparisons between means can be made in the standard way:

**Standard errors:** $\hat{\mu}_{i\cdot} = \frac{1}{m_2}\sum_j \bar{y}_{ij\cdot}$, so

$$\begin{aligned}
\mathrm{Var}[\mu_{i\cdot}] &= \frac{1}{m_2^2}\sum_j \sigma^2/n_{ij} \\
\mathrm{SE}[\hat{\mu}_{i\cdot}] &= \frac{1}{m_2}\sqrt{\sum_j \frac{\mathrm{MSE}}{n_{ij}}} \qquad \text{and similarly} \\
\mathrm{SE}[\hat{\mu}_{\cdot j}] &= \frac{1}{m_1}\sqrt{\sum_i \frac{\mathrm{MSE}}{n_{ij}}}
\end{aligned}$$

Now use the SE to obtain confidence intervals, t-tests, etc.

**Testing for interaction:** Consider

$$\begin{aligned}
H_0 : Y_{ijk} &= \mu + a_i + b_j + \epsilon_{ijk} \\
H_1 : Y_{ijk} &= \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}
\end{aligned}$$

How do we evaluate $H_0$ when the data are unbalanced? We'll first outline the procedure, then discuss why it works.

1. Compute $\text{SSE}_\text{F} = \min_{\mu,a,b,(ab)} \sum_i \sum_j \sum_k (y_{ijk} - [\mu + a_i + b_j + (ab)_{ij}])^2$
2. Compute $\text{SSE}_\text{A} = \min_{\mu,a,b} \sum_i \sum_j \sum_k (y_{ijk} - [\mu + a_i + b_j])^2$
3. Compute $\text{SS12} = \text{SSE}_\text{A} - \text{SSE}_\text{F}$. Note that this is always positive.

Allowing for interaction improves the fit, and reduces error variance. SSI measures the improvement in fit. If SSI is large, i.e. $\text{SSE}_\text{A}$ is much bigger than $\text{SSE}_\text{F}$, this suggests the additive model does not fit well and the interaction term should be included in the model.

**Testing:**

$$F = \frac{\text{MSI}}{\text{MSE}} = \frac{\text{SSI}/(m_1-1)(m_2-1)}{\text{SSE}_\text{F}/(N - m_1 m_2)}$$

Under $\text{H}_0$, $F \sim F_{(m_1-1)(m_2-1),N-m_1m_2}$, so a level-$\alpha$ test of $H_0$ is

$$\text{reject H}_0 \text{ if } F > F_{1-\alpha,(m_1-1)(m_2-1),N-m_1m_2}.$$

Note:

- SSI is the change in fit in going from the additive to the full model;
- $(m_1 - 1)(m_2 - 1)$ is the change in number of parameters in going from the additive to the full model.

**A painful example:** A small scale clinical trial was done to evaluate the effect of painkiller dosage on pain reduction for cancer patients in a variety of age groups.

- Factors of interest:
  - Dose $\in$ { Low, Medium, High }
  - Age $\in$ { 41-50, 51-60, 61-70, 71-80 }
- Response = Change in pain index $\in \{-5, 5\}$
- Design: CRD, each treatment level was randomly assigned to ten patients, **not** blocked by age.

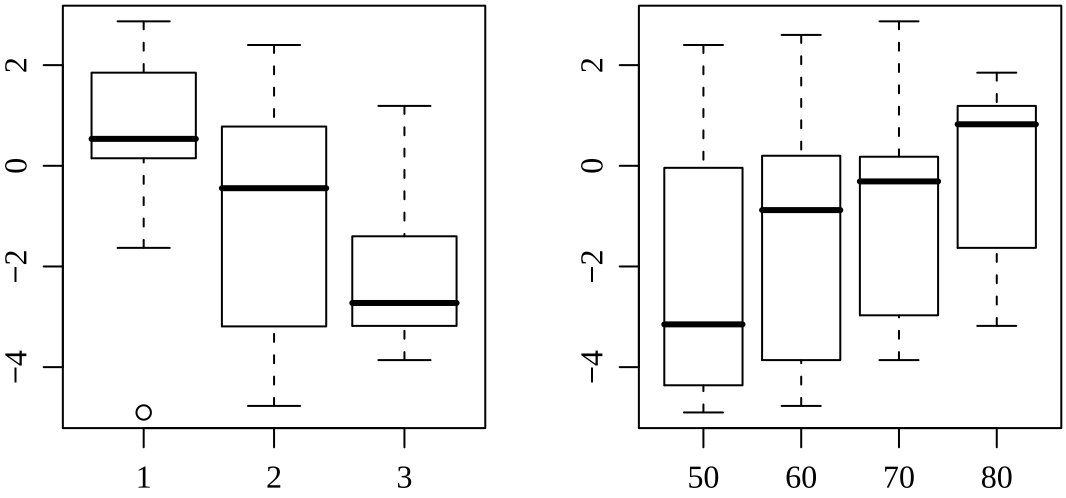


Figure 6.14: Marginal plots for pain data

```
> table(trt,ageg)
    ageg
trt  50 60 70 80
   1  1  2  3  4
   2  1  3  3  3
   3  2  1  4  3

> tapply(y,trt,mean)
     1      2      3
 0.381 -0.950 -2.131
>
> tapply(y,ageg,mean)
    50     60     70     80
-2.200 -1.265 -0.922 -0.139
```

Do these marginal plots and means misrepresent the data? To evaluate this possibility,

- compare the marginal plots in Figure 6.14 to the interaction plots in Figure 6.15;
- compute the LS means and compare to marginal means.

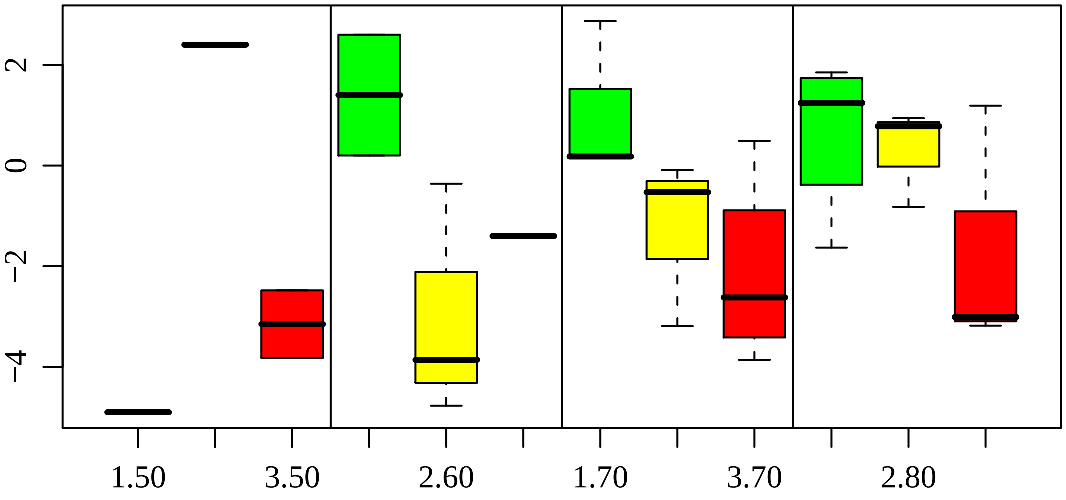


Figure 6.15: Interaction plots for pain data

```
> cellmeans<- tapply(y,list(trt,ageg),mean)

> cellmeans
     50        60        70        80
1 -4.90  1.400000  1.066667  0.677500
2  2.40 -2.996667 -1.270000  0.300000
3 -3.15 -1.400000 -2.152500 -1.666667

> trt_lsm<- apply(cellmeans,1,mean)
> age_lsm<- apply(cellmeans,2,mean)

> trt_lsm
         1          2          3
-0.4389583 -0.3916667 -2.0922917

> age_lsm
        50         60         70         80
-1.8833333 -0.9988889 -0.7852778 -0.2297222
```

What are the differences between LS means and marginal means? Not as extreme as in the accident example, but the differences can be explained by looking at the interaction plot, and the slight imbalance in the design:

- The youngest patients (ageg=50) were imbalanced towards the higher dose, so we might expect their marginal mean to be too low. Observe the change from the marginal mean = -2.2 to the LS mean = -1.883.

- The oldest patients (ageg=80) were imbalanced towards the lower dose, so we might expect their marginal mean to be too high. Observe the change from the marginal mean = -.14 to the LS mean = -.23 .

Let's look at the main effects in our model:

```
> trt_coef<- trt_lsm-mean(cellmeans)
> age_coef<- age_lsm-mean(cellmeans)

> trt_coef
         1          2          3
 0.5353472  0.5826389 -1.1179861

> age_coef
         50          60          70          80
-0.90902778 -0.02458333  0.18902778  0.74458333
```

What linear modeling commands in R will get you the same thing?

```
> options(contrasts=c("contr.sum","contr.poly"))
> fit_full<-lm( y~as.factor(ageg)*as.factor(trt))

> fit_full$coef[2:4]
as.factor(ageg)1 as.factor(ageg)2 as.factor(ageg)3
     -0.90902778      -0.02458333       0.18902778


> fit_full$coef[5:6]
as.factor(trt)1 as.factor(trt)2
      0.5353472       0.5826389
```

Note that the coefficients in the reduced/additive model are **not** the same:

```
> fit_add<-lm( y~as.factor(ageg)+as.factor(trt))

> fit_add$coef[2:4]
as.factor(ageg)1 as.factor(ageg)2 as.factor(ageg)3
      -0.7921041       -0.3593577        0.3049595

> fit_add$coef[5:6]
as.factor(trt)1 as.factor(trt)2
    1.208328354    -0.002085645
```

## 6.7 Non-orthogonal sums of squares:

Consider the following ANOVA table obtained from R:

```
> anova(  lm( y~as.factor(ageg)+as.factor(trt))  )
                Df  Sum Sq Mean Sq F value   Pr(>F)
as.factor(ageg)  3  13.355   4.452  0.9606 0.42737
as.factor(trt)   2  28.254  14.127  3.0482 0.06613 .
Residuals       24 111.230   4.635
```

It might be somewhat unsettling that R also produces the following table:

```
> anova(  lm( y~as.factor(trt)+as.factor(ageg))  )
                Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(trt)   2  31.588  15.794  3.4079 0.0498 *
as.factor(ageg)  3  10.021   3.340  0.7207 0.5494
Residuals       24 111.230   4.635
```

Where do these sums of squares come from? What do the $F$-tests represent? By typing "`?anova.lm`" in R we see that `anova()` computes

"`a sequential analysis of variance table for that fit. That is, the reductions in the residual sum of squares as each term of the formula is added in turn are given in as the rows of a table, plus the residual sum of squares.`"

**Sequential sums of squares:** Suppose in a linear model we have three **sets** of parameters, say $A$, $B$ and $C$. For example, let

- $A$ be the main effects of factor 1
- $B$ be the main effects of factor 2
- $C$ be their interaction.

Consider the following calculation:

**0.** Calculate SS0 = residual sum of squares from the model

$$(0) \quad y_{ijk} = \mu + \epsilon_{ijk}$$

**1.** Calculate SS1 = residual sum of squares from the model

$$(A) \quad y_{ijk} = \mu + a_i + \epsilon_{ijk}$$

**2.** Calculate SS2 = residual sum of squares from the model

$$(AB)\;\; y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$$

**3.** Calculate SS3 = residual sum of squares from the model

$$(ABC)\; y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}$$

We can assess the importance of A, B and C **sequentially** as follows:

- SSA = improvement in fit in going from (0) to (A) = SS0-SS1
- SSB—A = improvement in fit in going from (A) to (AB) = SS1-SS2
- SSC—AB = improvement in fit in going from (AB) to (ABC) = SS2-SS3

This is actually what R presents in an ANOVA table:

```
> ss0<-sum(   lm( y~1 )$res^2 )
> ss1<-sum(   lm( y~as.factor(ageg) )$res^2 )
> ss2<-sum(   lm( y~as.factor(ageg)+as.factor(trt) )$res^2 )
> ss3<

> s0-ss1
[1] 13.3554
>
> ss1-ss2
[1] 28.25390
>
> ss2-ss3
[1] 53.75015

> ss3
[1] 57.47955

> anova(  lm( y~as.factor(ageg)*as.factor(trt))   )
                                  Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(ageg)                    3 13.355   4.452  1.3941 0.27688
as.factor(trt)                     2 28.254  14.127  4.4239 0.02737 *
as.factor(ageg):as.factor(trt)     6 53.750   8.958  2.8054 0.04167 *
Residuals                         18 57.480   3.193
```

Why does order of the variables matter?

- In a balanced design, the parameters are orthogonal, and SSA = SSA—B , SSB = SSB—A and so on, so the order doesn't matter.
- In an unbalanced design, the estimates of one set of parameters depends on whether or not you are estimating the others, i.e. they are not orthogonal, and in general SSA $\neq$ SSA—B , SSB $\neq$ SSB—A.

I will try to draw a picture of this on the board.

**The bottom line:** For unbalanced designs, there is no "variance due to factor 1" or "variance due to factor 2". There is only "extra variance due to factor 1, beyond that explained by factor 2", and vice versa. This is essentially because of the non-orthogonality, and so the part of the variance that can be explained by factor 1 overlaps with the part that can be explained by factor 2. This will become more clear when you get to regression next quarter.

## 6.8 Analysis of covariance

**Example(Oxygen ventilation):** Researchers are interested in measuring the effects of exercise type on maximal oxygen uptake.

**Experimental units:** 12 healthy males between 20 and 35 who didn't exercise regularly.

**Design:** Six subjects randomly selected to a 12-week step aerobics program, the remaining to 12-weeks of outdoor running on flat terrain.

**Response:** Before and after the 12 week program, each subject's $O_2$ uptake was tested while on an inclined treadmill.

$$y = \text{change in } O_2 \text{ uptake}$$

**Initial analysis:** CRD with one two-level factor. The first thing to do is plot the data. The first panel of Figure 6.16 indicates a moderately large difference in the two sample populations. The second thing to do is a two-sample t-test:

$$t_{\text{obs}} = \left| \frac{\bar{y}_A - \bar{y}_B}{s_p\sqrt{2/6}} \right| = \left| \frac{7.705 - (-2.767)}{6.239 \times .577} \right| = 2.907, \quad \Pr(|t_{10}| \geq 2.907) = 0.0156$$

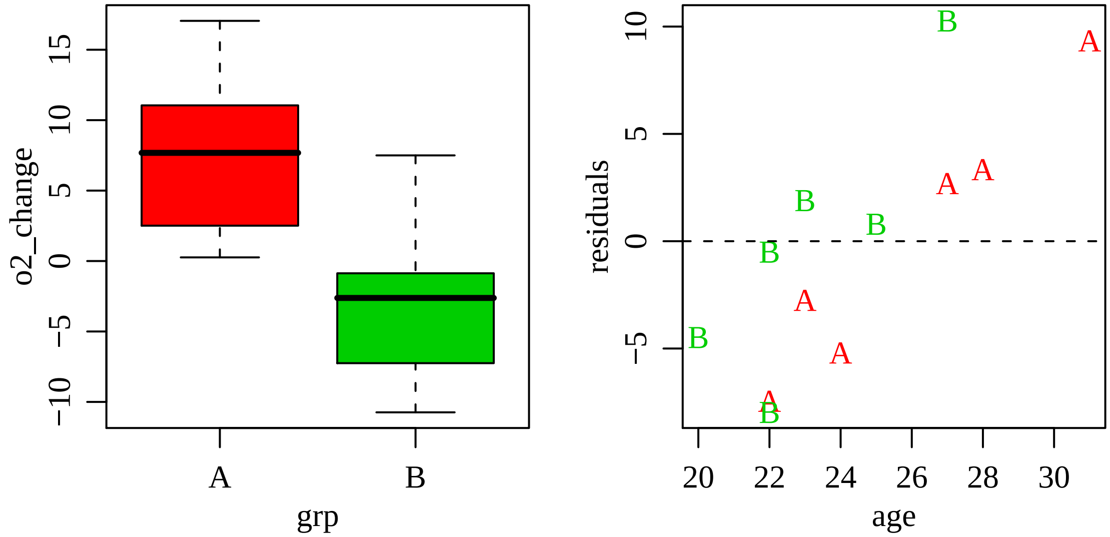


Figure 6.16: Oxygen uptake data

So we have reasonably strong evidence of a difference in treatment effects. However, a plot of residuals versus age of the subject indicates a couple of causes for concern with our inference:

1. The residual plot indicates that response increases with age (why?), regardless of treatment group.
2. Just due to chance, the subjects assigned to group $A$ were older than the ones assigned to $B$.

**Question:** Is the observed difference in $\bar{y}_A, \bar{y}_B$ due to

- treatment?
- age?
- both?

**A linear model for ANCOVA:** Let $y_{i,j}$ be the response of the $j$th subject in treatment $i$:

$$y_{i,j} = \mu + a_i + b \times x_{i,j} + \epsilon_{i,j}$$

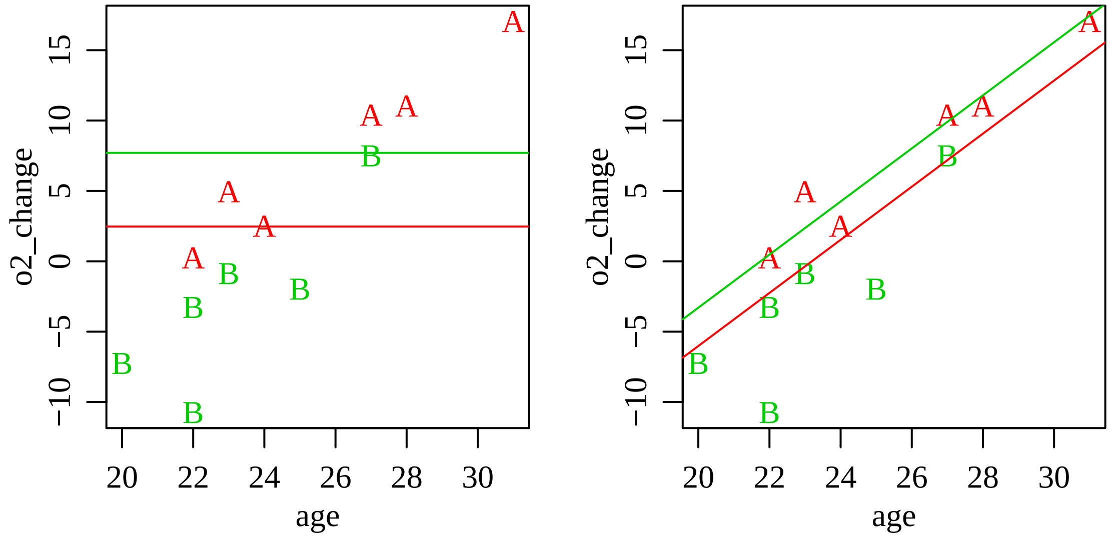


Figure 6.17: ANOVA and ANCOVA fits to the oxygen uptake data

This model gives a linear relationship between age and response for each group:

$$\begin{array}{rrcccccc} & & \text{intercept} & & \text{slope} & & \text{error} \\ \text{if } i = A, & Y_i = & (\mu + a_A) & + & b \times x_{i,j} & + & \epsilon_{i,j} \\ i = B, & Y_i = & (\mu + a_B) & + & b \times x_{i,j} & + & \epsilon_{i,j} \end{array}$$

Unbiased parameter estimates can be obtained by minimizing the residual sum of squares:

$$(\hat{\mu}, \hat{a}, \hat{b}) = \arg \min_{\mu,a,b} \sum_{i,j} (y_{i,j} - [\mu + a_i + b \times x_{i,j}])^2$$

The fitted model is shown in the second panel of Figure 6.17.

**Variance decomposition** Consider the following two ANOVAs:

```
> anova(lm(o2_change~grp))
          Df Sum Sq Mean Sq F value  Pr(>F)
grp        1 328.97  328.97  8.4503 0.01565 *
Residuals 10 389.30   38.93
```

```
> anova(lm(o2_change~grp+age))
          Df Sum Sq Mean Sq F value    Pr(>F)
grp        1 328.97  328.97  42.062 0.0001133 ***
age        1 318.91  318.91  40.776 0.0001274 ***
Residuals  9  70.39    7.82
```

The second one decomposes the variation in the data that is orthogonal to treatment (SSE from the first ANOVA) into a parts that can be ascribed to age (SS age in the second ANOVA), and everything else (SSE from second ANOVA). I will try to draw some triangles that describe this situation.

Now consider two other ANOVAs:

```
> anova(lm(o2_change~age))
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 576.09  576.09  40.519 8.187e-05 ***
Residuals 10 142.18   14.22


> anova(lm(o2_change~age+grp))
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 576.09  576.09 73.6594 1.257e-05 ***
grp        1  71.79   71.79  9.1788   0.01425 *
Residuals  9  70.39    7.82
```

Again, I will try to draw some triangles describing this situation.

**Blocking, ANCOVA and unbalanced designs:** Suppose we have a factor of interest (say $F1$) that we will randomly assign to experimental material, but it is known that there is some nuisance factor (say $F2$ ) that is suspected to be a large source of variation. Our options are:

- Block according to $F2$: This allows an even allocation of treatments across levels of $F2$. As a result, we can separately estimate the effects of $F1$ and $F2$.

- Do a CRD with respect to $F1$, but account for $F2$ somehow in the analysis:
  - For a standard two-factor ANOVA, $F2$ must be treated as a categorical variable ("old", "young").
  - ANCOVA allows for more precise control of variation due to $F2$.

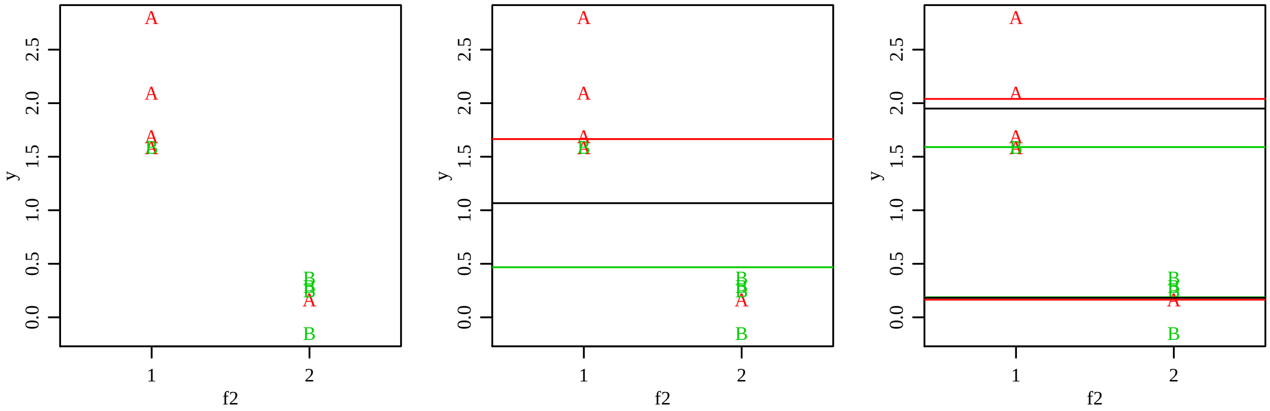


Figure 6.18: Unbalanced design: Controlling eliminates effect.

However, in either of these approaches the design is unlikely to be completely balanced , and so the effects of $F1$ can't be estimated separately from those of $F2$. This is primarily a problem for experiments with small sample sizes: The probability of design imbalance decreases as a function of sample size (as does the correlation between the parameter estimates) as long as $F1$ is randomly assigned.

## 6.9 Types of sums of squares

Figure 6.18 gives a simple example where an unbalanced design can lead to results that are difficult to interpret.

- Two factors:
  - F1 = treatment (A vs B)
  - F2 = block (location 1 versus 2)
- 10 experimental units
- Balanced CRD in F1, but not in F1×F2.

Looking at things marginally in F1, we have $\bar{y}_A > \bar{y}_B$, and there seems to be an effect of F1=$A$ versus $B$. This is highlighted in the second panel of the figure, which shows the difference between the sample marginal means and the grand mean seems large compared to error variance.

```
> anova(lm(y~f1+f2))
          Df Sum Sq Mean Sq F value   Pr(>F)
f1         1 3.5824  3.5824  21.577 0.002355 **
f2         1 4.2971  4.2971  25.882 0.001419 **
Residuals  7 1.1622  0.1660
```

However, notice the imbalance:

- 4 $A$ observations under F2=1, 1 under F2=2
- 1 $B$ observations under F2=1, 4 under F2=2

What happens if we explain the variability in $\boldsymbol{y}$ using F2 first, then F1? Look at the black lines in the third panel of the plot: These are the marginal means for F2. Within levels of F2, differences between levels of F1 are small.

```
> anova(lm(y~f2+f1))
          Df Sum Sq Mean Sq F value    Pr(>F)
f2         1 7.8064  7.8064 47.0185 0.0002405 ***
f1         1 0.0731  0.0731  0.4405 0.5281460
Residuals  7 1.1622  0.1660
```

The ANOVA quantifies this: There is variability in the data that can be explained by either F1 or F2. In this case,

- SSA > SSA|B
- SSB > SSB|A

Do these inequalities always hold? Consider the data in Figure 6.19. In this case F1=A is higher than F1=B for both values of F2. But there are more A observations in the low-mean value of F2 than the high-mean value. The second an third plots suggest

- Marginally, the difference between levels of F1 are small.
- Within each group, the difference between levels of F1 are larger.

Thus "controlling" for F2 **highlights** the differences between levels of F1. This is confirmed in the corresponding ANOVA tables:

```
> anova(lm(y~f1+f2))
          Df Sum Sq Mean Sq F value    Pr(>F)
f1         1 0.1030  0.1030  0.5317 0.4895636
f2         1 5.7851  5.7851 29.8772 0.0009399 ***
Residuals  7 1.3554  0.1936
```

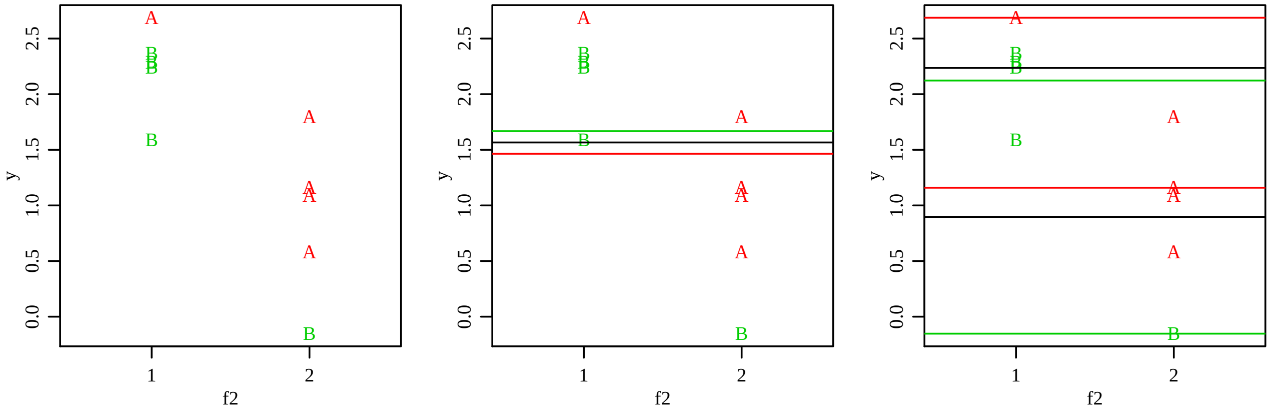


Figure 6.19: Unbalanced design: Controlling highlights effect.

```
> anova(lm(y~f2+f1))
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
f2            1  4.4804   4.4804  23.1391  0.001943 **
f1            1  1.4077   1.4077   7.2698  0.030806 *
Residuals     7  1.3554   0.1936
```

Which ANOVA table to use? Some software packages combine these anova tables to form ANOVAs based on "alternative" types of sums of squares. Consider a two-factor ANOVA in which we plan on decomposing variance into additive effects of F1, additive effects of F2, and their interaction.

**Type I SS:** Sequential, orthogonal decomposition of the variance.

**Type II SS:** Sum of squares for a factor is the improvement in fit from adding that factor, given inclusion of all other terms at that level or below.

**Type III SS:** Sum of squares for a factor is the improvement in fit from adding that factor, given inclusion of all other terms.

So for example:

- $\text{SSF1}_1 = \text{RSS}(0) - \text{RSS}(\text{F1})$ if F1 first in sequence
- $\text{SSF1}_1 = \text{RSS}(\text{F2}) - \text{RSS}(\text{F1} + \text{F2})$ if F1 second in sequence

- $\mathrm{SSF1}_2 = \mathrm{RSS(F2)} - \mathrm{RSS(F1 + F2)}$
- $\mathrm{SSF1}_3 = \mathrm{RSS(F2, F1 : F2)} - \mathrm{RSS(F1, F2, F1 : F2)}$

Type II sums of squares is very popular, and is to some extent the "default" for linear regression analysis. It seems natural to talk about the "variability due to a treatment, after controlling for other sources of variation." However, there are situations in which you might not want to "control" for other variables. Consider the following (real-life) scenario:

**Clinical trial:**

- Factor 1: Assigned dose of a drug (high or low)
- Factor 2: Dose reduction (yes or no)

Weaker patients experienced higher rates of drug side-effects than stronger patients, and side effects are worse under the high dose. As a result,

- Only the strongest $A$ patients remain in the non-reduced group;
- Only the weakest $B$ patients go into the reduced group.

"Controlling" or adjusting for dose reduction artificially inflates the perceived effect of treatment.

# Chapter 7

# Nested Designs

**Example(Potato):** Sulfur added to soil kills bacteria, but too much sulfur can damage crops. Researchers are interested in comparing two levels of sulfur additive (low, high) on the damage to two types of potatoes.

**Factors of interest:**

1. Potato type $\in \{A, B\}$
2. Sulfur additive $\in$ {low,high}

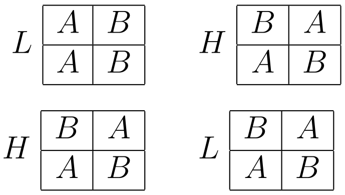
**Experimental material:** Four plots of land.

**Design constraints:**

- It is easy to plant different potato types within the same plot
- It is difficult to have different sulfur treatments in the same plot, due to leeching.

**Experimental Design:** A Split-plot design

1. Each sulfur additive was randomly assigned to two of the four plots.
2. Each plot was split into four subplots. Each potato type was randomly assigned to two subplots per plot.



**Randomization:**

Sulfur type was randomized to whole plots;

Potato type was randomized to subplots.

**Initial data analysis:** Sixteen responses, 4 treatment combinations.

- 8 responses for each potato type
- 8 responses for each sulfur type
- 4 responses for each potato×type combination

```
> fit.full<-lm(y~type*sulfur) ;  fit.add<-lm(y~type+sulfur)

> anova(fit.full)
              Df  Sum Sq Mean Sq F value    Pr(>F)
type           1 1.48840 1.48840 13.4459 0.003225 **
sulfur         1 0.54022 0.54022  4.8803 0.047354 *
type:sulfur    1 0.00360 0.00360  0.0325 0.859897
Residuals     12 1.32835 0.11070

> anova(fit.add)
          Df  Sum Sq Mean Sq F value   Pr(>F)
type       1 1.48840 1.48840 14.5270 0.00216 **
sulfur     1 0.54022 0.54022  5.2727 0.03893 *
Residuals 13 1.33195 0.10246
```

**Randomization test:** Consider comparing the observed outcome to the population of other outcomes that could've occurred under

- different treatment assignments;

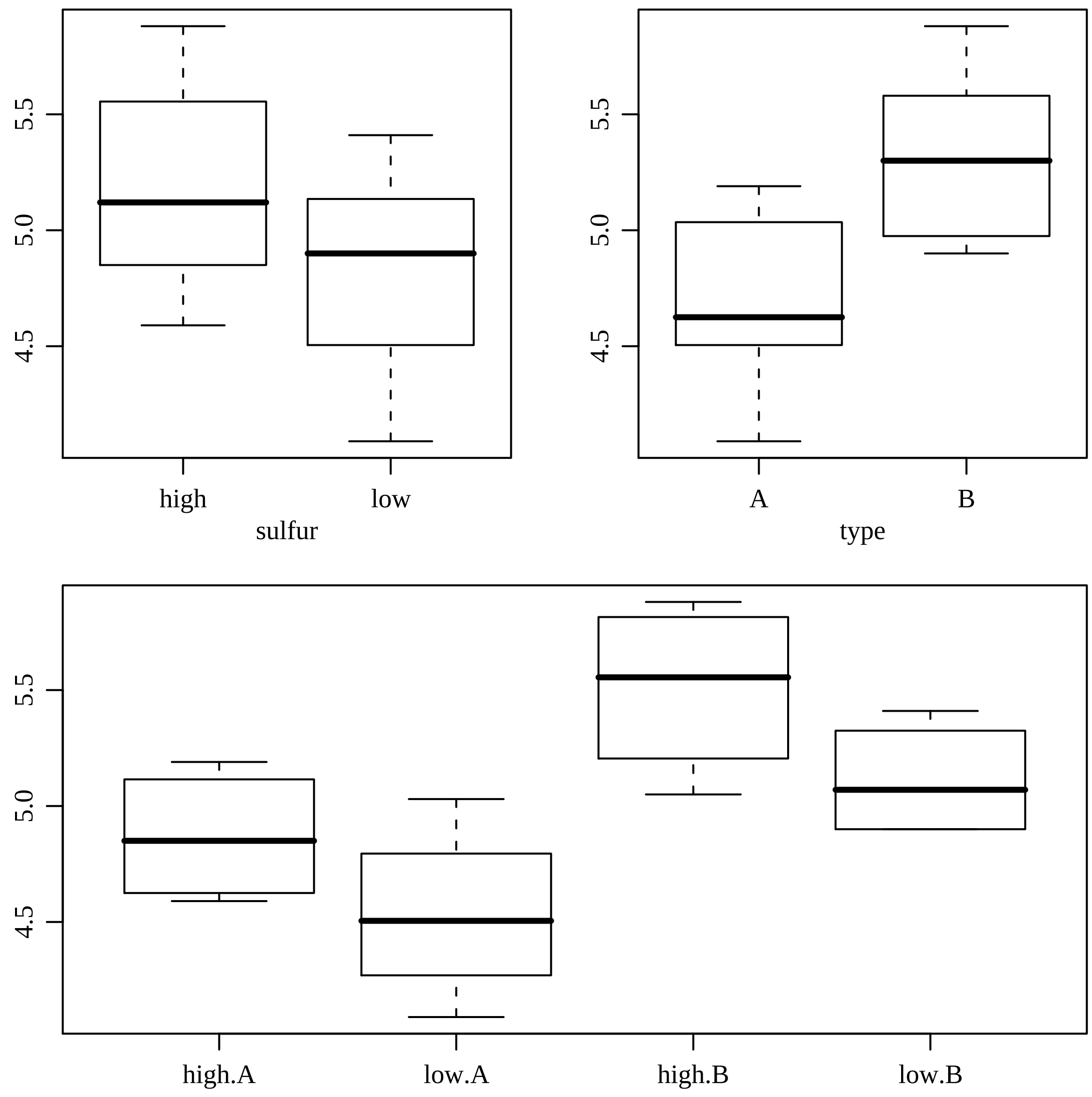


Figure 7.1: Potato data.

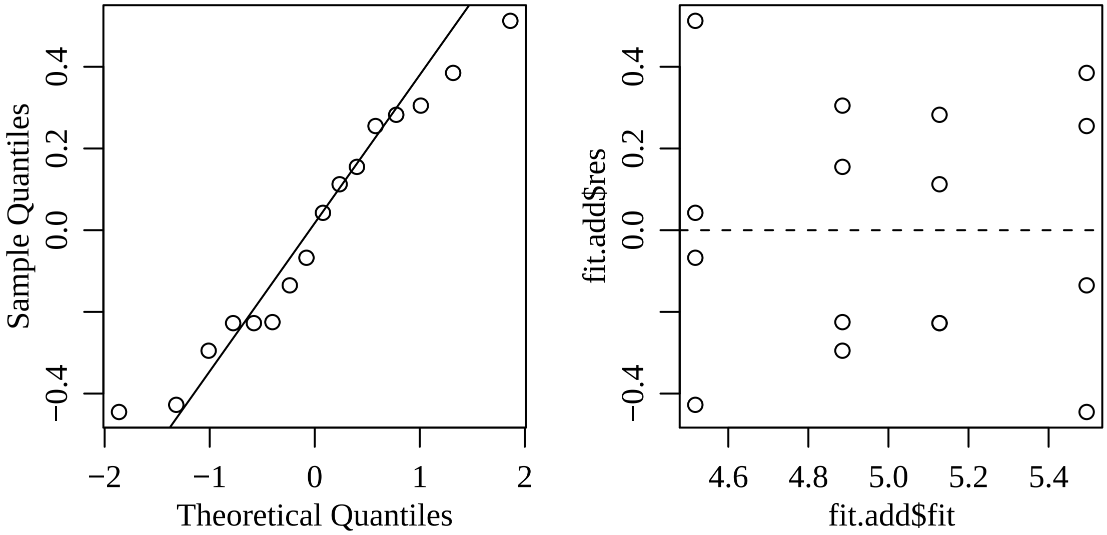


Figure 7.2: Diagnostic plots for potato ANOVA.

- no treatment effects.

| Treatment assignment | Field 1 | Field 2 | Field 3 | Field 4 |
|---|---|---|---|---|
| 1 | low<br>A A B B | low<br>A A B B | high<br>A A B B | high<br>A A B B |
| 2 | low<br>A B A B | low<br>A A B B | high<br>A A B B | high<br>A A B B |
| 3 | low<br>A B A B | low<br>A B A B | high<br>A A B B | high<br>A A B B |
| 4 | low<br>A B A B | low<br>A A B B | high<br>A B A B | high<br>A A B B |
| 5 | low<br>A B A B | high<br>A A B B | low<br>A A B B | high<br>A A B B |
| ⋮ | ⋮<br>⋮ ⋮ ⋮ ⋮ | ⋮<br>⋮ ⋮ ⋮ ⋮ | ⋮<br>⋮ ⋮ ⋮ ⋮ | ⋮<br>⋮ ⋮ ⋮ ⋮ |

Number of possible treatment assignments:

- $\binom{4}{2} = 6$ ways to assign sulfur

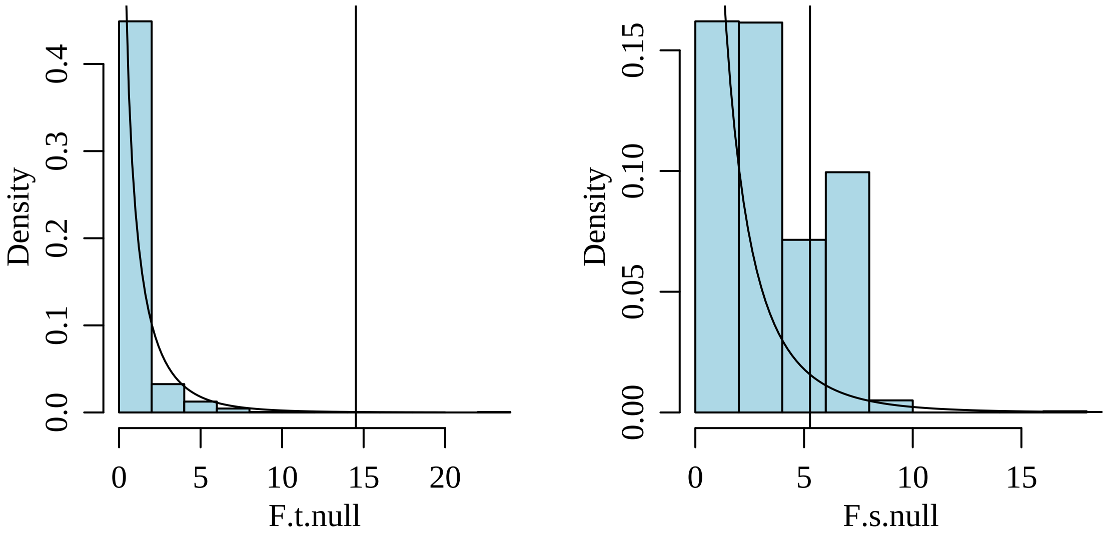


Figure 7.3: Potato data

- For each sulfur assignment there are $\binom{4}{2}^4 = 1296$ type assignments

So we have 7776 possible treatment assignments. It probably wouldn't be to hard to write code to go through all possible treatment assignments, but its very easy to obtain a Monte Carlo approximation to the null distribution:

```
F.t.null<-F.s.null<-NULL

 for(ns in 1:1000){
s.sim<-rep( sample(c(”low”,”low”,”high”,”high”)),rep(4,4))
t.sim<-c( sample(c(”A”,”A”,”B”,”B”)), sample(c(”A”,”A”,”B”,”B”)),
          sample(c(”A”,”A”,”B”,”B”)), sample(c(”A”,”A”,”B”,”B”))
)

 fit.sim<-anova( lm(y~as.factor(t.sim)+as.factor(s.sim)) )

F.t.null<-c(F.t.null,fit.sim[1,4])
F.s.null<-c(F.s.null,fit.sim[2,4])
                   }
```

```
> mean(F.t.null>=F.t.obs)
[1] 0.001
> mean(F.s.null>=F.s.obs)
[1] 0.352
```

What happened?

$$\begin{aligned} F^{\text{rand}}_{\text{type}} &\approx F_{1,13} \Rightarrow p^{\text{rand}}_{\text{type}} \approx p^{\text{anova1}}_{\text{type}} \\ F^{\text{rand}}_{\text{sulfur}} &\not\approx F_{1,13} \Rightarrow p^{\text{rand}}_{\text{sulfur}} \not\approx p^{\text{anova1}}_{\text{sulfur}} \end{aligned}$$

The $F$-distribution approximates a null randomization distribution if **treatments** are randomly assigned to **units**. But here, the sulfur treatment is being assigned to groups of four units.

- The precision of an effect is related to the number of **independent** treatment assignments made
- We have 16 assignments of **type**, but only 4 assignments of **sulfur**. It is difficult to tell the difference between sulfur effects and field effects. Our estimates of sulfur effects are **less precise** than those of type.

Note:

- From the point of view of **type** alone, the design is a RCB.
  - Each whole-plot(field) is a block. We have 2 observations of each type per block.
  - We compare MSType to the MSE from residuals left from a model with type effects, block effects and possibly interaction terms.
  - Degrees of freedom breakdown for the **sub-plot analysis**:

| Source | dof |
|---|---|
| block=whole plot | 3 |
| type | 1 |
| subplot error | 11 |
| subplot total | 15 |

- From the point of view of **sulfur** alone, the design is a CRD.
  - Each whole plot is an experimental unit.
  - We want to compare MSSulfur to the variation in whole plots, which are fields.
  - Degrees of freedom breakdown for the **whole-plot analysis**:

| Source | dof |
|---|---|
| sulfur | 1 |
| whole plot error | 2 |
| whole plot total | 3 |

- Recall, degrees of freedom quantify the precision of our estimates and the shape of the null distribution.

The basic idea is:

- We have different levels of experimental units (fields/whole-plots for sulfur, subfields/sub-plots for type).
- We compare the MS of a factor to the variation among the experimental units **of that factor's level**. In general, this variation is represented by the interaction between the experimental units label and all factors assigned at the given level.
- The level of an interaction is the level of the smallest experimental unit involved.

```
> anova(lm( y~type+type:sulfur+as.factor(field)  ) )
                 Df  Sum Sq Mean Sq F value    Pr(>F)
type              1 1.48840 1.48840 54.7005 2.326e-05 ***
as.factor(field)  3 1.59648 0.53216 19.5575 0.0001661 ***
type:sulfur       1 0.00360 0.00360  0.1323 0.7236270
Residuals        10 0.27210 0.02721
```

Thus there is strong evidence for type effects, and little evidence that the effects of type vary among levels of sulfur.

```
> anova(lm(y~sulfur+as.factor(field)))
                 Df  Sum Sq Mean Sq F value  Pr(>F)
sulfur            1 0.54022 0.54022  3.6748 0.07936 .
as.factor(field)  2 1.05625 0.52813  3.5925 0.05989 .
Residuals        12 1.76410 0.14701
```

The F-test here is not appropriate: We need to compare MSSulfur to the variation among whole-plots, after accounting for the effects of sulfur, i.e. MSField (note that including or excluding type here has no effect on this comparison).

```
MSS<-anova(lm(y~sulfur+as.factor(field)))[1,3]
MSWPE<-anova(lm(y~sulfur+as.factor(field)))[2,3]
F.sulfur<-MSS/MSWPE
```

```
> F.sulfur
[1] 1.022911
> 1-pf(F.sulfur,1,2)
[1] 0.4182903
```

This is more in line with the analysis using the randomization test.
The above calculations are somewhat tedious. In R there are several automagic ways of obtaining the correct $F$-test for this type of design. One way is with the `aov` command:

```
> fit1<-aov(y~type*sulfur + Error(factor(field)))
> summary(fit1)

Error: factor(field)
          Df  Sum Sq Mean Sq F value Pr(>F)
sulfur     1 0.54022 0.54022  1.0229 0.4183
Residuals  2 1.05625 0.52813

Error: Within
            Df  Sum Sq Mean Sq F value    Pr(>F)
type         1 1.48840 1.48840 54.7005 2.326e-05 ***
type:sulfur  1 0.00360 0.00360  0.1323    0.7236
Residuals   10 0.27210 0.02721

###

> fit2<-aov(y~type + sulfur + Error(factor(field)))
> summary(fit2)

Error: factor(field)
          Df  Sum Sq Mean Sq F value Pr(>F)
sulfur     1 0.54022 0.54022  1.0229 0.4183
Residuals  2 1.05625 0.52813

Error: Within
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       1 1.48840 1.48840  59.385 9.307e-06 ***
Residuals 11 0.27570 0.02506
```

The “Error(field)” option tells R that it should

- treat “fields” as sampled experimental units, i.e. subject to sampling heterogeneity;
- identify factors assigned to fields;

- test those factors via comparisons of MSFactor to MSField.

## 7.1 Mixed-effects approach

What went wrong with the normal sampling model approach? What is wrong with the following model?

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}$$

where

- $i$ indexes sulfur level, $i \in \{1, 2\}$ ;
- $j$ indexes type level, $j \in \{1, 2\}$ ;
- $k$ indexes reps, $k \in \{1, \ldots, 4\}$
- $\epsilon_{ijk}$ are i.i.d. normal

We checked the normality and constant variance assumptions for this model previously, and they seemed ok. What about independence? Figure 5.4 plots the residuals as a function of field. The figure indicates that **residuals** are more alike within a field than across fields, and so observations within a field are **positively correlated**. Statistical dependence of this sort is common to split-plot and other nested designs.
dependence within whole-plots

- affects the amount of information we have about factors applied at the whole-plot level: within a given plot, we can't tell the difference between plot effects and whole-plot factor effects.
- This doesn't affect the amount of information we have about factors applied at the sub-plot level: We can tell the difference between plot effects and sub-plot factor effects.

If residuals within a whole-plot are positively correlated, the most intuitively straightforward way to analyze such data (in my opinion) is with a hierarchical mixed-effects model:

$$y_{ijkl} = \mu + a_i + b_j + (ab)_{ij} + \gamma_{ik} + \epsilon_{ijkl}$$

where things are as before except

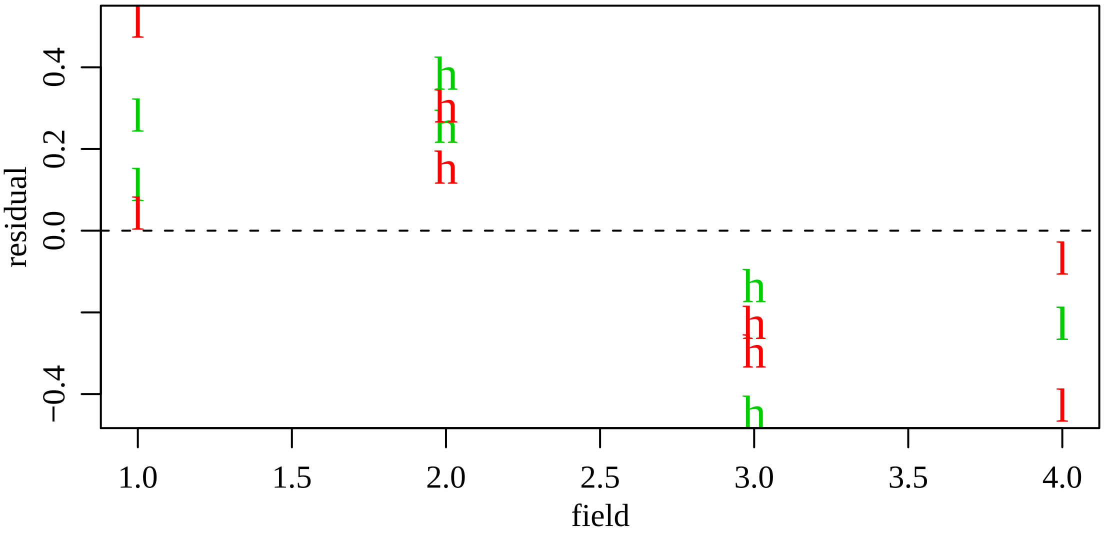


Figure 7.4: Potato data

- $\gamma_{ik}$ represents error variance at the whole plot level, i.e. variance in whole plot experimental units (fields or blocks in the above example). The index $k$ represents whole-plot reps $k = 1, \ldots, r_1$,

$$\{\gamma_{ik}\} \sim \text{normal}(0, \sigma_w^2)$$

- $\epsilon_{ijkl}$ represents error variance at the sub plot level, i.e. variance in sub plot experimental units The index $j$ represents sub-plot reps $l = 1, \ldots, r_2$,

$$\{\epsilon_{ijkl}\} \sim \text{normal}(0, \sigma_s^2)$$

Now every subplot within the same wholeplot has something in common, i.e. $\gamma_{ik}$. This models the **positive** correlation within whole plots:

$$\begin{aligned}
\text{Cov}(y_{i,j_1,k,l_1}, y_{i,j_2,k,l_2}) &= \text{E}[(y_{i,j_1,k,l_1} - \text{E}[y_{i,j_1,k,l_1}]) \times (y_{i,j_2,k,l_2} - \text{E}[y_{i,j_2,k,l_2}])] \\
&= \text{E}[(\gamma_{i,k} + \epsilon_{i,j_1,k,l_1}) \times (\gamma_{i,k} + \epsilon_{i,j_2,k,l_2})] \\
&= \text{E}[\gamma_{i,k}^2 + \gamma_{i,k} \times (\epsilon_{i,j_1,k,l_1} + \epsilon_{i,j_2,k,l_2}) + \epsilon_{i,j_1,k,l_1}\epsilon_{i,j_2,k,l_2}] \\
&= \text{E}[\gamma_{i,k}^2] + 0 + 0 = \sigma_w^2
\end{aligned}$$

This and more complicated random-effects models can be fit using the `lme` command in R. To use this command, you need the `nlme` package:

```
library(nlme)
fit.me<-lme(fixed=y~type+sulfur, random=~1|as.factor(field))

>summary(fit.me)

Fixed effects: y ~ type + sulfur
               Value Std.Error DF   t-value p-value
(Intercept)   4.8850 0.2599650 11 18.790991  0.0000
typeB         0.6100 0.0791575 11  7.706153  0.0000
sulfurlow    -0.3675 0.3633602  2 -1.011393  0.4183

> anova(fit.me)
            numDF denDF  F-value p-value
(Intercept)     1    11 759.2946  <.0001
type            1    11  59.3848  <.0001
sulfur          1     2   1.0229  0.4183
```

Notice that this gives the same results as

- the randomization test (approximately);
- the by-hand comparison of mean squares for factors to the appropriate MSE;
- the results from the `aov` command.

But the `lme` command allows for much more complex models to be fit.

## 7.2 Repeated measures analysis

**Sitka spruce data:** Longitudinal data on 79 spruce trees

- 54 grown in ozone-enriched chambers
- 25 grown in regular atmosphere (controls)

The size of each tree (roughly the volume) was measured at five time points: 152, 174, 201, 227 and 258 days after the beginning of experiment.

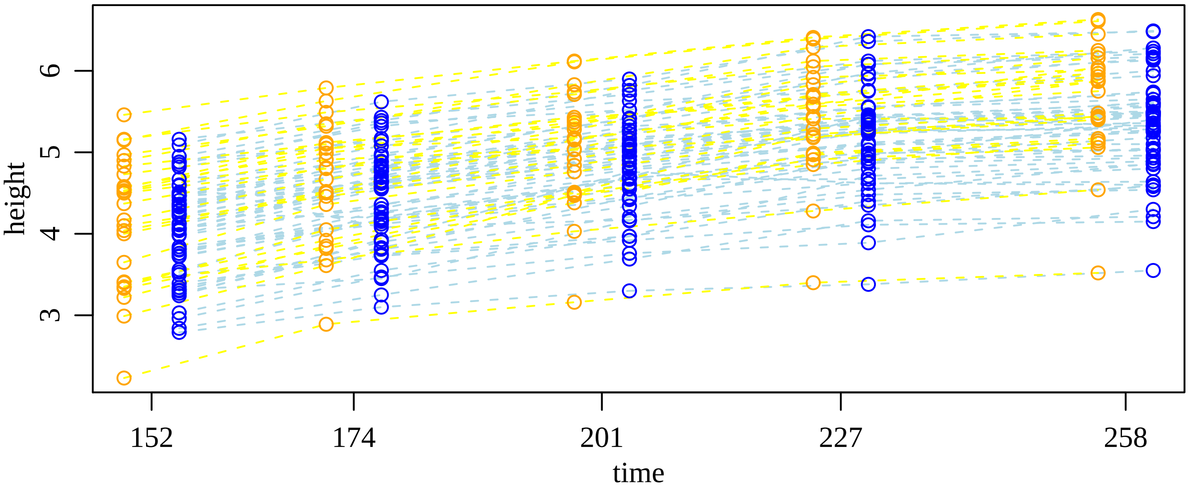


Figure 7.5: Sitka spruce data.

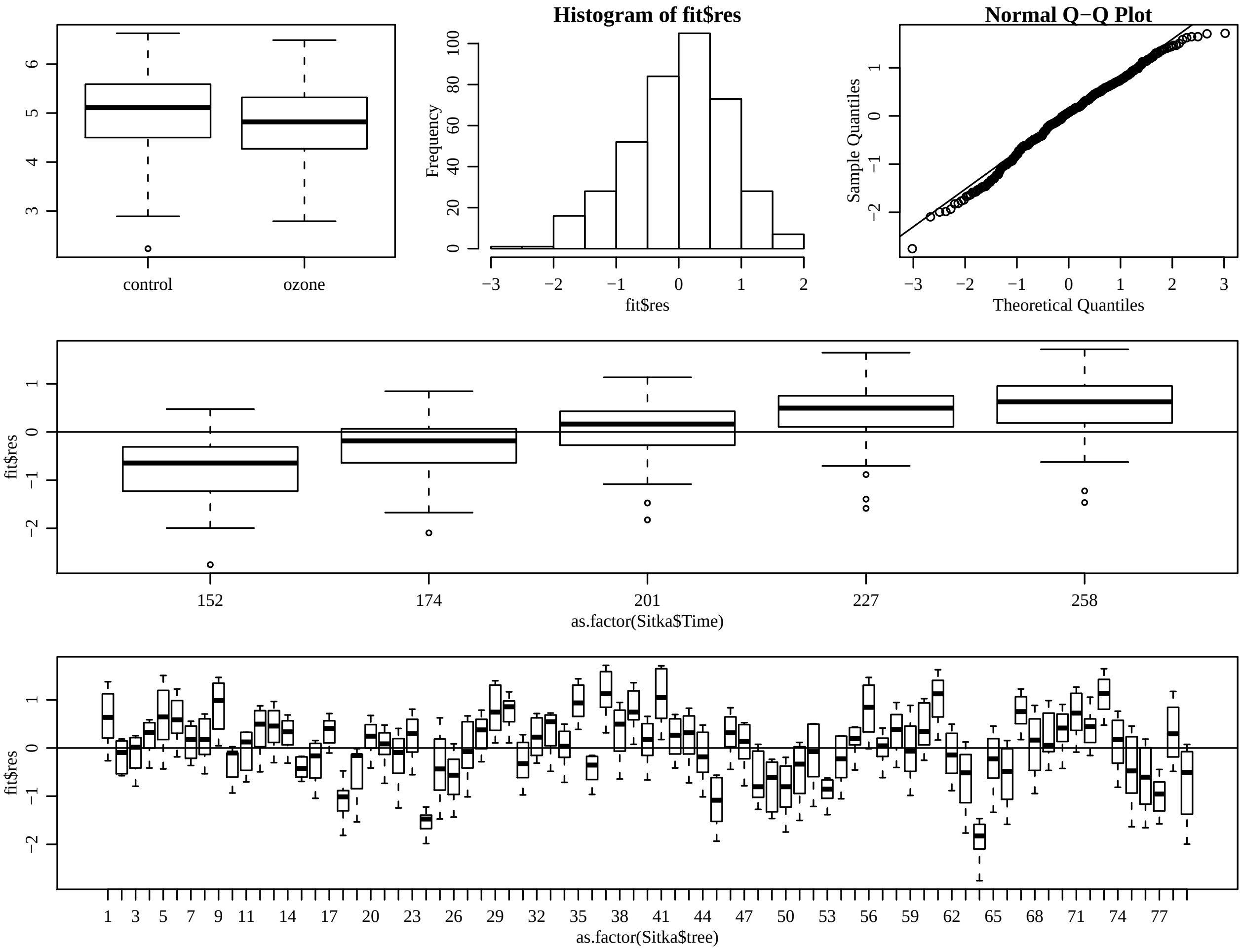
How should we analyze these data?

How should we evaluate evidence of a treatment (ozone) effect?

**Naive approach I:** A naive approach would be to ignore the study design and the temporal nature of the data.

- $5 \times 54 = 270$ ozone observations
- $5 \times 25 = 125$ control observations

```
> fit<-lm(Sitka$size~Sitka$treat)
> anova(fit)
              Df  Sum Sq Mean Sq F value  Pr(>F)
Sitka$treat    1   3.810   3.810  6.0561 0.01429 *
Residuals    393 247.222   0.629
```



**Naive approach II:** Clearly there is some effect of time. Let's now "account" for growth over time, using a simple ANCOVA:

$$y_{i,j,t} = \mu_0 + a_i + b \times t + c_i \times t + \epsilon_{i,j,t}$$

- for ozone $(i = 1)$ , $\mathrm{E}[y_{1,j,t}] = (\mu_0 + a_1) + (b + c_1) \times t$
- for control $(i = 2)$ , $\mathrm{E}[y_{2,j,t}] = (\mu_0 + a_2) + (b + c_2) \times t$

```
> fit<-lm(size~Time+treat+Time*treat,data=Sitka)
> anova(fit)

           Df  Sum Sq Mean Sq  F value    Pr(>F)
Time        1  89.564  89.564 222.9020 < 2.2e-16 ***
treat       1   3.810   3.810   9.4813  0.002222 **
Time:treat  1   0.551   0.551   1.3703  0.242480
Residuals 391 157.107   0.402
```

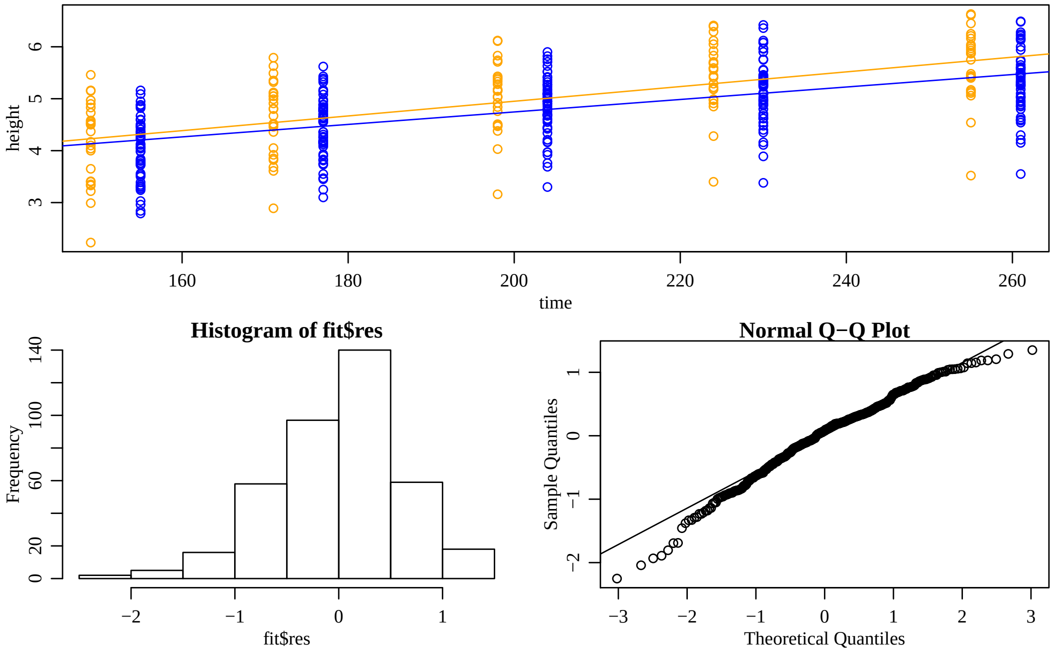


Figure 7.6: ANCOVA fit and residuals

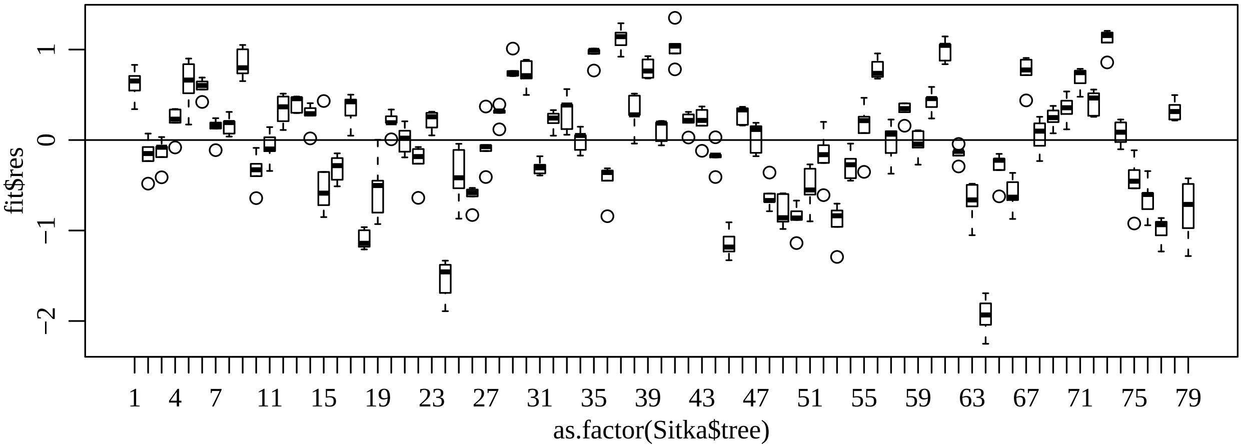


Figure 7.7: Within-tree dependence

**Conservative approach:** Any of our standard inference tools are suspect because we clearly do not have independent observations. The within-tree dependence indicates that we have

- less information than in 270 independent ozone and 125 independent control observations,
- but more information than in 54 independent ozone and 24 independent control observations.

So our "worst-case scenario" is analogous to having $n_1 = 54$ and $n_2 = 24$. One approach to analysis in such situations is to reduce the information to **one number per unit**, then compare numbers across treatments. For example, we could compute an average and fit a regression line for each tree $j$, giving

- $y_{i,j}^{\text{avg}}$, the average of the 5 observations for tree $i, j$
- $y_{i,j}^{\text{int}}$, the estimated intercept of the regr. line for tree $i, j$
- $y_{i,j}^{\text{int}}$, the estimated slope of the regr. line for tree $i, j$

We can then compare averages, intercepts and slopes across the two treatment groups. Note that, for each type of $y$, there is only one observation for each tree: we have eliminated the problem of dependent measurements.

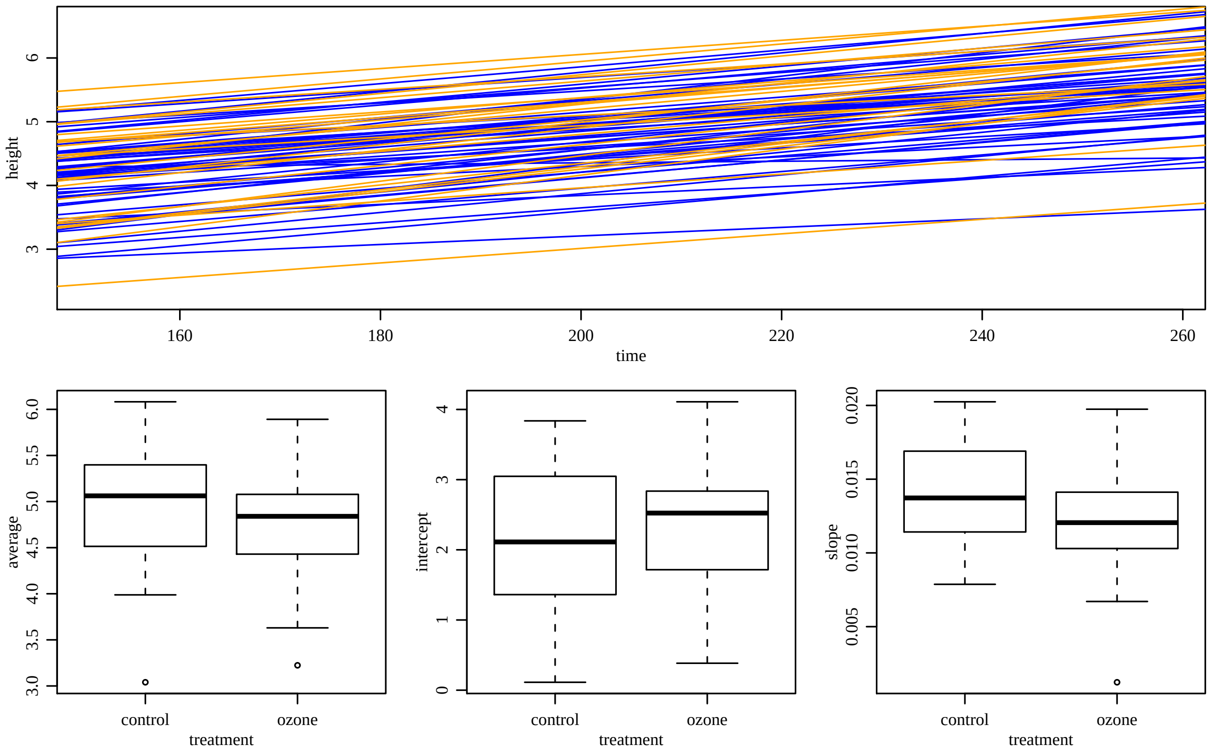


Figure 7.8: Reduction to tree-specific summary statistics

```
> anova(lm(y.avg~treat))
          Df  Sum Sq Mean Sq F value Pr(>F)
treat      1  0.7619  0.7619  2.0188 0.1594
Residuals 77 29.0617  0.3774

> anova(lm(y.int~treat))
          Df Sum Sq Mean Sq F value Pr(>F)
treat      1  0.840   0.840  1.0431 0.3103
Residuals 77 61.989   0.805

> anova(lm(y.slope~treat))
          Df     Sum Sq    Mean Sq F value   Pr(>F)
treat      1 0.00007815 0.00007815  7.6628 0.007058 **
Residuals 77 0.00078529 0.00001020
```

**Linear random effects models:** The last approach is extremely conservative, as it basically reduces all the information we have from a tree to one number. Of course, the observations from a single tree are not completely dependent, and so compressing the data in this way throws away potentially valuable information. To make use of all the information from a tree, we can use a random effects model which accounts for correlation of observations common to a given tree.

$$y_{i,j,t} = (a_1 + b_{1,j} + c_{1,i}) + (a_2 + b_{2,j} + c_{2,j}) \times t + \epsilon_{i,j,t}$$

where

- $(b_{1,j}, b_{2,j}), j = 1, 2$ are **fixed-effects**, measuring the heterogeneity of the average slope and intercept across the two levels of treatment;
- $(c_{1,1}, c_{2,1}), \ldots, (c_{1,n}, c_{2,n}) \sim$ i.i.d. multivariate normal$(\mathbf{0}, \Sigma)$ are **random effects** , inducing a within-tree correlation of observations.

```
> fit<-lme(fixed=size~treat+Time+treat*Time,
           random=~Time|tree,data=Sitka)
> anova(fit)
            numDF denDF  F-value p-value
(Intercept)     1   314 4967.056  <.0001
treat           1    77    2.118  0.1497
Time            1   314 1246.518  <.0001
treat:Time      1   314    7.663  0.0060
```