

Experimental design

Your text has a nice introduction to randomization, the placebo effect, and observational studies in Section 11.4, pp. 452-458. Some highlights:

- Randomization (randomly assigning treatments to experimental subjects) reduces the possibility of bias, conscious or unconscious.
- It is hard to anticipate or account for all possible sources of bias. Section 11.4.5.
- In a designed experiment, variables which may affect the response other than the treatments of interest, should be included in the design if possible (e.g. matching, blocking, etc.) Section 11.4.6.
- In observational studies, confounding variables need to be accounted for as best as possible. Section 11.4.7.

Basics of Experimental Design

Here we describe an experimental design to compare the effectiveness of four insecticides to eradicate beetles. The primary interest is determining which treatment is most effective, in the sense of providing the lowest typical survival time.

In a **completely randomized design** (CRD), the scientist might select a sample of genetically identical beetles for the experiment, and then randomly assign a predetermined number of beetles to the treatment groups (insecticides). The sample sizes for the groups need not be equal. A power analysis is often conducted to determine sample sizes for the treatments. For simplicity, assume that 48 beetles will be used in the experiment, with 12 beetles assigned to each group.

After assigning the beetles to the four groups, the insecticide is applied (uniformly to all experimental units or beetles), and the individual survival times recorded. A natural analysis of the data would be to compare the survival times using a one-way ANOVA.

There are several important controls that should be built into this experiment. The same strain of beetles should be used to ensure that the four treatment groups are alike as possible, so that differences in survival times are attributable to the insecticides, and not due to genetic differences among beetles. Other factors that may influence the survival time, say the concentration of the insecticide or the age of the beetles, would be held constant, or fixed by the experimenter, if possible. Thus, the same concentration would be used with the four insecticides.

In complex experiments, there are always potential influences that are not realized or thought to be unimportant that you do not or can not control. The **randomization** of beetles to groups ensures that there is no systematic dependence of the observed treatment differences on the uncontrolled influences. This is important in studies where genetic and environmental influences can not be easily controlled (as in humans, more so than in bugs or mice). The randomization of beetles to insecticides tends to diffuse or greatly reduce the effect of the uncontrolled influences on the comparison of insecticides, in the sense that these effects become part of the uncontrolled or error variation of the experiment.

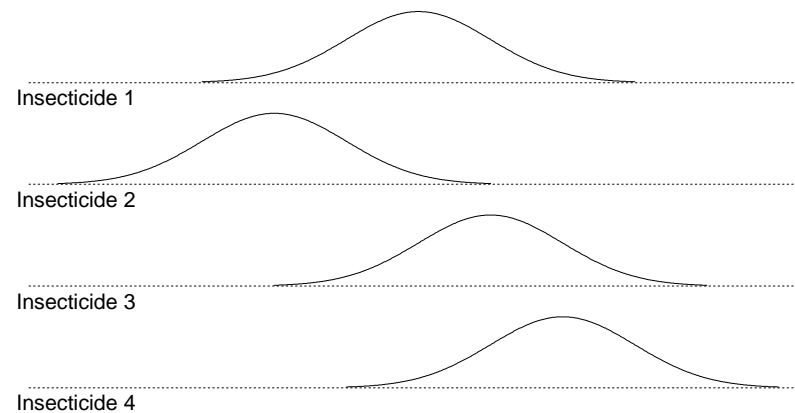
Randomization makes the “independent” in *iid* as close to reality as possible.

Suppose Y_{ij} is the response for the j^{th} experimental unit in the i^{th} treatment group, where $i = 1, 2, \dots, I$. The statistical model for a completely randomized **one-factor design** that leads to a one-way ANOVA is given by:

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

where μ_i is the (unknown) population mean for all potential responses to the i^{th} treatment, and e_{ij} is the residual or deviation of the response from the population mean. The responses within and across treatments are assumed to be independent, normal random variables with constant variance.

For the insecticide experiment, y_{ij} is the survival time for the j^{th} beetle given the i^{th} insecticide, where $i = 1, 2, 3, 4$ and $j = 1, 2, \dots, 12$. The random selection of beetles coupled with the randomization of beetles to groups ensures the independence assumptions. The assumed population distributions of responses for the $I = 4$ insecticides can be represented below:



Let $\mu = \frac{1}{I} \sum_i \mu_i$ be the grand mean, or average of the population means.

Let $\alpha_i = \mu_i - \mu$ be the i^{th} **treatment group effect**.

The treatment effects add to zero, $\alpha_1 + \alpha_2 + \cdots + \alpha_I = 0$, and measure the difference between the treatment population means and the grand mean.

Given this notation, the one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + e_{ij}.$$

The model specifies that the

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Residual}.$$

An hypothesis of interest is whether the population means are equal: $H_0 : \mu_1 = \dots = \mu_I$, which is equivalent to the hypothesis of no treatment effects: $H_0 : \alpha_1 = \dots = \alpha_I = 0$. If H_0 is true, then the one-way model is

$$Y_{ij} = \mu + e_{ij},$$

where μ is the common population mean. We discussed how to test H_0 and do multiple comparisons last time.

Most epidemiological studies are **observational studies** where the groups to be compared ideally consist of individuals that are similar on all characteristics that influence the response, except for the feature that defines the groups.

In a **designed experiment**, the groups to be compared are defined by treatments randomly assigned to individuals. If, in an observational study we can not define the groups to be homogeneous on important factors that might influence the response, then we should adjust for these factors in the analysis. This can be approximately accomplished by including characteristics into a regression model (linear or logistic), i.e. “adjusting for confounders.”

Paired Experiments and Randomized Block Experiment

A **randomized block design** (Section 12.3.3, pp. 500–503) is often used instead of a completely randomized design in studies where there is extraneous variation among the experimental units that may influence the response. A significant amount of the extraneous variation may be removed from the comparison of treatments by partitioning the experimental units into fairly **homogeneous subgroups** or **blocks**.

Say you are interested in comparing the effectiveness of four antibiotics for a bacterial infection. The recovery time after administering an antibiotic may be influenced by the patients general health, the extent of their infection, or their age. Randomly allocating experimental subjects to the treatments (and then comparing them using a one-way ANOVA) may produce one treatment having a “favorable” sample of patients with features that naturally lead to a speedy recovery. This sort of thing wasn’t a problem with the beetles.

Alternatively, if the characteristics that affect the recovery time are spread across treatments, then the variation within samples due to these uncontrolled features can dominate the effects of the treatment, leading to an inconclusive result.

A better way to design this experiment would be to **block** the subjects into groups of four patients who are alike as possible on factors other than the treatment that influence the recovery time. The four treatments are then randomly assigned to the patients (one per patient) within a block, and the recovery time measured. The blocking of patients usually produces a more sensitive comparison of treatments than does a completely randomized design because the variation in recovery times due to the blocks is eliminated from the comparison of treatments.

A randomized block design is a **paired experiment** when two treatments are compared. The usual analysis for a paired experiment is to difference the outcome within blocks and perform a simple two-sample test. In certain experiments, each experimental unit receives each treatment. The experimental units are “natural” blocks for the analysis and “serve as their own control.”

Example: Comparison of Treatments to Relieve Itching

(Example A, p. 501): Ten male volunteers between 20 and 30 years old were used as a study group to compare seven treatments (5 drugs, a placebo, and no drug) to relieve itching. Each subject was given a different treatment on seven study days. The time ordering of the treatments was randomized across days. Except on the no-drug day, the subjects were given the treatment intravenously, and then itching was induced on their forearms using an effective itch stimulus called cowage. The subjects recorded the duration of itching, in seconds. The data are given in the table below. From left to right the drugs are: papaverine, morphine, aminophylline, pentobarbitol, tripelenamine.

Patient	Nodrug	Placebo	Papv	Morp	Amino	Pento	Tripel
1	174	263	105	199	141	108	141
2	224	213	103	143	168	341	184
3	260	231	145	113	78	159	125
4	255	291	103	225	164	135	227
5	165	168	144	176	127	239	194
6	237	121	94	144	114	136	155
7	191	137	35	87	96	140	121
8	100	102	133	120	222	134	129
9	115	89	83	100	165	185	79
10	189	433	237	173	168	188	317

The volunteers in the study were treated as blocks in the analysis. At best, the volunteers might be considered a representative sample of males between the ages of 20 and 30. This limits the extent of inferences from the experiment. The scientists can not, without sound medical justification, extrapolate the results to children or to senior citizens.

The Analysis of a Randomized Block Design

Assume that you designed a randomized block experiment with I blocks and J treatments, where each treatment occurs once in each block. Let y_{ij} be the response for the j^{th} treatment within the i^{th} block. The model for the experiment is

$$Y_{ij} = \mu_{ij} + e_{ij},$$

where μ_{ij} is the population mean response for the j^{th} treatment in the i^{th} block and e_{ij} is the deviation of the response from the mean. The population means are assumed to satisfy the additive model

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

where μ is a grand mean, α_i is the effect for the i^{th} block, and β_j is the effect for the j^{th} treatment.

The responses are assumed to be independent across blocks, normally distributed and with constant variance. The randomized block model does not require the observations within a block to be independent, but does assume that the correlation between responses within a block is identical for each pair of treatments. This is a reasonable working assumption in many analyses.

The model is sometimes written as

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Block Effect} + \text{Residual}.$$

Given the data, let $\bar{y}_{i\cdot}$ be the i^{th} block sample mean (the average of the responses in the i^{th} block), $\bar{y}_{\cdot j}$ be the j^{th} treatment sample mean (the average of the responses on the j^{th} treatment), and $\bar{y}_{\cdot\cdot}$ be the average response of all IJ observations in the experiment.

An ANOVA table for the randomized block experiment partitions the Model SS into SS for Blocks and Treatments.

Source	df	SS	MS
Blocks	$I - 1$	$J \sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	
Treats	$J - 1$	$I \sum_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	
Error	$(I - 1)(J - 1)$	$\sum_{ij} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2$	
Total	$IJ - 1$	$\sum_{ij} (y_{ij} - \bar{y}_{\cdot\cdot})^2$	

A primary interest is testing whether the treatment effects are zero: $H_0 : \beta_1 = \dots = \beta_J = 0$. The treatment effects are zero if the population mean responses are identical for each treatment.

A formal test of no treatment effects is based on the p-value from the F -statistic $F_{obs} = \text{MS Treat}/\text{MS Error}$. The p -value is evaluated in the usual way (i.e. as an upper tail area from an F -distribution with $J - 1$ and $(I - 1)(J - 1)$ df.) This H_0 is rejected when the treatment averages $\bar{y}_{.j}$ vary significantly relative to the error variation.

A test for no block effects ($H_0 : \alpha_1 = \dots = \alpha_I = 0$) is often a secondary interest, because, if the experiment is designed well, the blocks will be, by construction, noticeably different. There are no block effects if the block population means are identical. A formal test of no block effects is based on the p-value from the the F -statistic $F_{obs} = \text{MS Blocks}/\text{MS Error}$. This H_0 is rejected when the block averages \bar{y}_i vary significantly relative to the error variation.

A Randomized Block Analysis of the Itching Data

The `anova` command is used to get the randomized block analysis. You will be shown the steps in Thursday's Lab, but I will mention a few important points.

- The data are comprised of three variables: **itchtime**, **person** (ranges from 1-10), and **treatment** (ranges from 1-7). A data file called `itch.txt` was created with these three variables to be read into R.
- In the ANOVA table, persons play the role of Blocks in this analysis.

```

> d <- read.table("c:/tim/PubH7400/itch.txt")
> d
  V1 V2 V3
1 174 1 1
2 263 1 2
3 105 1 3
4 199 1 4
5 141 1 5
6 108 1 6
7 141 1 7
8 224 2 1
9 213 2 2
10 103 2 3

et cetera...

> seconds <- d[,1]; person <- factor(d[,2]); treatment <- factor(d[,3])
> fit <- aov(seconds~person+treatment)
> summary(fit)
              Df Sum Sq Mean Sq F value    Pr(>F)
person          9 103280   11476   3.7078 0.001124 **
treatment       6  53013    8835   2.8548 0.017303 *
Residuals     54 167130    3095
---

```

The F -test for treatments yields a p -value of 0.017; there are significant differences among the treatments. The F -test for person (the blocks) yields a p -value of 0.001. Blocking significantly reduced the residual variability and was worthwhile.

What happens if we ignore the fact that we are taking repeated measures on each of the 10 subjects and *do not* block? This amounts to fitting the simple ANOVA model

$$Y_{ij} = \mu + \beta_j + e_{ij}$$

via `fit <- aov(seconds~treatment)`. The p -value jumps to 0.07 and we *fail to reject* the null hypothesis that there is no difference in itching time among the treatments.

From the original (blocked) analysis, we obtain Tukey intervals for all possible pairwise differences:

```

> TukeyHSD(fit,"treatment")
  Tukey multiple comparisons of means
    95% family-wise confidence level

$treatment
      diff      lwr      upr    p adj
2-1  13.8 -62.38544  89.985440 0.9977708
3-1 -72.8 -148.98544   3.385440 0.0699571
4-1 -43.0 -119.18544  33.185440 0.6006037
5-1 -46.7 -122.88544  29.485440 0.5039188
6-1 -14.5  -90.68544  61.685440 0.9970659
7-1 -23.8  -99.98544  52.385440 0.9609992
3-2 -86.6 -162.78544 -10.414560 0.0162806
4-2 -56.8 -132.98544  19.385440 0.2710014
5-2 -60.5 -136.68544  15.685440 0.2057246
6-2 -28.3 -104.48544  47.885440 0.9134888
7-2 -37.6 -113.78544  38.585440 0.7369272
4-3  29.8  -46.38544 105.985440 0.8919777
5-3  26.1  -50.08544 102.285440 0.9398221
6-3  58.3  -17.88544 134.485440 0.2430713
7-3  49.0  -27.18544 125.185440 0.4454107
5-4  -3.7  -79.88544  72.485440 0.9999990
6-4  28.5  -47.68544 104.685440 0.9107883
7-4  19.2  -56.98544  95.385440 0.9867157
6-5  32.2  -43.98544 108.385440 0.8515989
7-5  22.9  -53.28544  99.085440 0.9676471
7-6  -9.3  -85.48544  66.885440 0.9997652

```

With an overall FER of 5% we only reject $H_0 : \beta_3 = \beta_2$. That is, we reject that there is no difference in mean itching time between placebo and papaverine.

We *almost* reject that there is no difference between no drug and papaverine. If we had instead used Bonferroni here to simultaneously test $H_0 : \beta_3 = \beta_1$ and $H_0 : \beta_3 = \beta_2$, we reject (at the end of these slides).

A histogram of the residuals shows the normality assumption is fine.

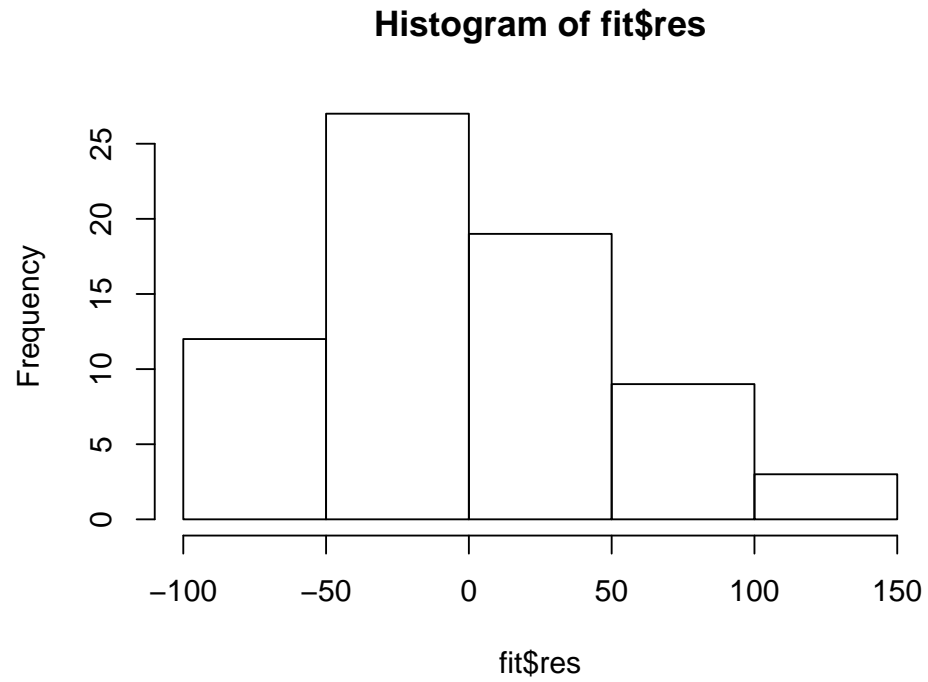


Figure 1: `> hist(fit$res)`.

We can get the estimated block effects $\hat{\alpha}_1, \dots, \hat{\alpha}_{10}$ and treatment effects $\hat{\beta}_1, \dots, \hat{\beta}_7$ as:

```
> model.tables(fit)
```

Tables of effects

treatment

1	2	3	4	5	6	7
26.71	40.51	-46.09	-16.29	-19.99	12.21	2.91

person

1	2	3	4	5	6	7	8	9	10
-2.71	32.29	-5.57	35.71	9.00	-21.29	-49.00	-30.00	-47.71	79.29

The estimate of the grand mean μ , although not produced above, is simply $\bar{y}_{..} = 164.2857$ from `mean(seconds)`.

Let's say that we really are only interested in two comparisons:
 $H_0 : \beta_3 = \beta_2$ and $H_0 : \beta_3 = \beta_1$. The R function `lm()` fits the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

but where $\alpha_1 = \beta_1 = 0$. So subject 1 and treatment 1 form the “baseline” group with mean μ . Once the model is fit, general linear model theory can be used to form CI's and test hypotheses on any linear function of the model parameters.

The parameters in the model are listed as

$$\boldsymbol{\theta} = (\mu, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)'$$

The hypotheses boil down to testing $H_0 : c_1 = 0$ and $H_0 : c_2 = 0$:

$$c_1 = \beta_3 - \beta_1 = \beta_3 = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]\boldsymbol{\theta}.$$

$$c_2 = \beta_3 - \beta_2 = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ -1 \ 1 \ 0 \ 0 \ 0 \ 0]\boldsymbol{\theta}.$$

```

> fit <- lm(seconds~person+treatment)
> fit

Call:
lm(formula = seconds ~ person + treatment)

Coefficients:
(Intercept)    person2    person3    person4    person5    person6
  188.286      35.000     -2.857     38.429     11.714    -18.571
  person7    person8    person9    person10  treatment2  treatment3
 -46.286    -27.286    -45.000     82.000     13.800    -72.800
 treatment4  treatment5  treatment6  treatment7
 -43.000    -46.700    -14.500    -23.800

> esticon(fit,c(0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0))
Confidence interval ( WALD ) level = 0.95
  beta0 Estimate Std.Error  t.value DF Pr(>|t|) Lower.CI Upper.CI
1     0     -72.8  24.87969 -2.926082 54 0.0050122 -122.6808 -22.91923
> esticon(fit,c(0,0,0,0,0,0,0,0,0,0,0,-1,1,0,0,0,0))
Confidence interval ( WALD ) level = 0.95
  beta0 Estimate Std.Error  t.value DF Pr(>|t|) Lower.CI Upper.CI
1     0     -86.6  24.87969 -3.480751 54 0.0009978 -136.4808 -36.71923

```

We multiply these p -values by the number of comparisons (2), to get 0.01 and 0.002. We reject both with an overall FER < 0.05 using Bonferroni.

We obtain the same conclusion when we *do not* include the blocking variable into the model and fit

$$Y_{ij} = \mu + \beta_j + e_{ij}.$$

```
> pairwise.t.test(seconds,treatment,p.adjust.method="none")
```

```
Pairwise comparisons using t tests with pooled SD
```

```
data: seconds and treatment
```

	1	2	3	4	5	6
2	0.6393	-	-	-	-	-
3	0.0156	0.0044	-	-	-	-
4	0.1472	0.0570	0.3130	-	-	-
5	0.1160	0.0431	0.3764	0.8999	-	-
6	0.6224	0.3378	0.0510	0.3344	0.2759	-
7	0.4197	0.2041	0.0994	0.5147	0.4374	0.7520

```
P value adjustment method: none
```

Here the two p -values are $0.0156 \times 2 = 0.031$ and $0.0044 \times 2 = 0.009$.

It's easier to reject when you focus your inference.