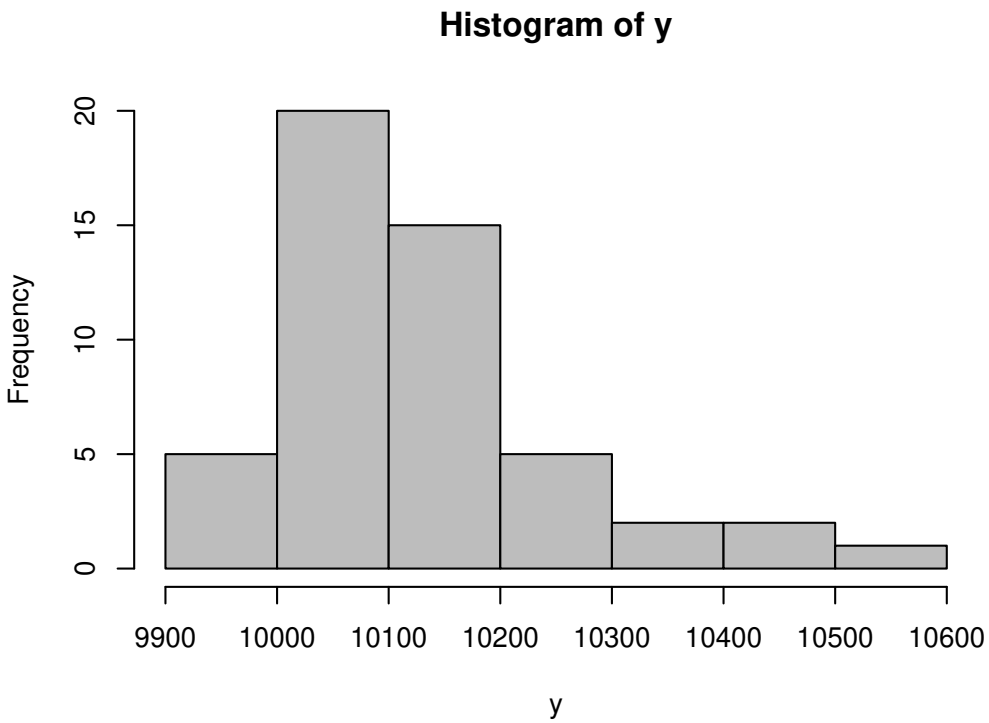
Review of transformations

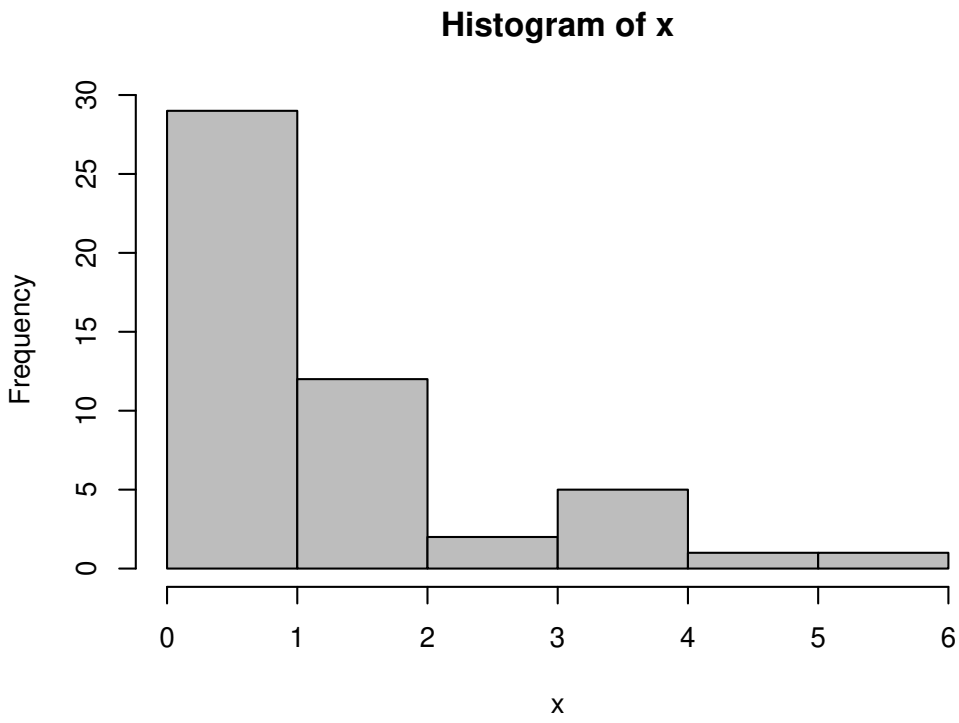
A data set gives the price of cars (measured in dollars), their mile-per-gallon rating, and other variables. You want to fit a multiple linear regression (MLR) model to predict price. The histogram of price is highly positively skewed.

Which of the following would be the first step in your analysis?

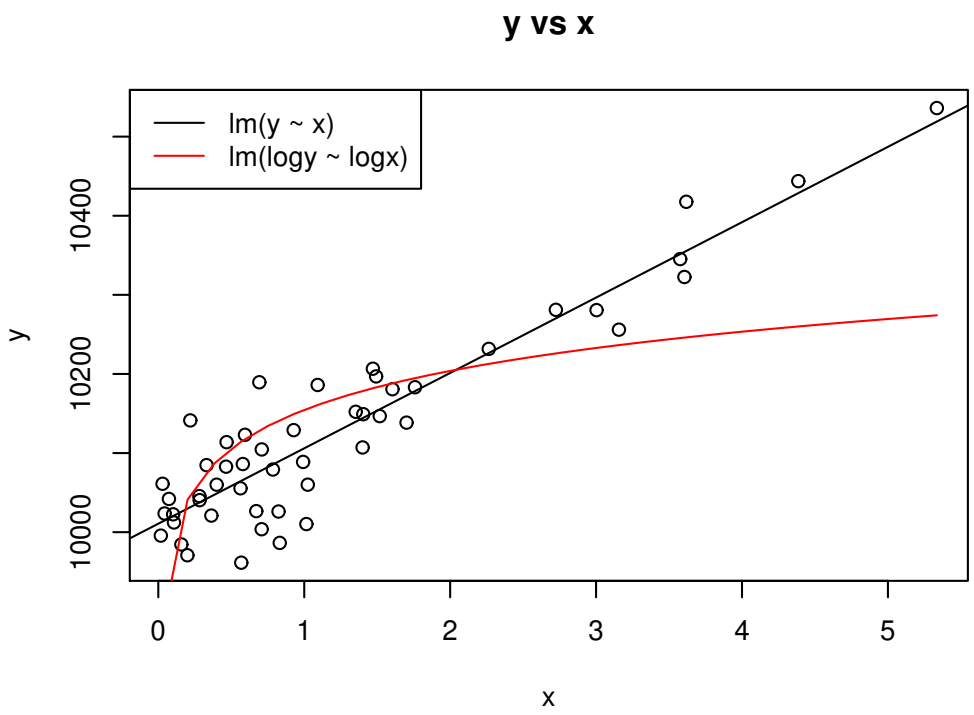
1. Fit a MLR model to price
2. Fit a MLR model to the data without the large outliers
3. Transform price to make its histogram look as normal as possible and then fit a MLR model to the transformed variable
4. Fit a Poisson regression model to price because it is non-negative integer-valued and its histogram is skewed

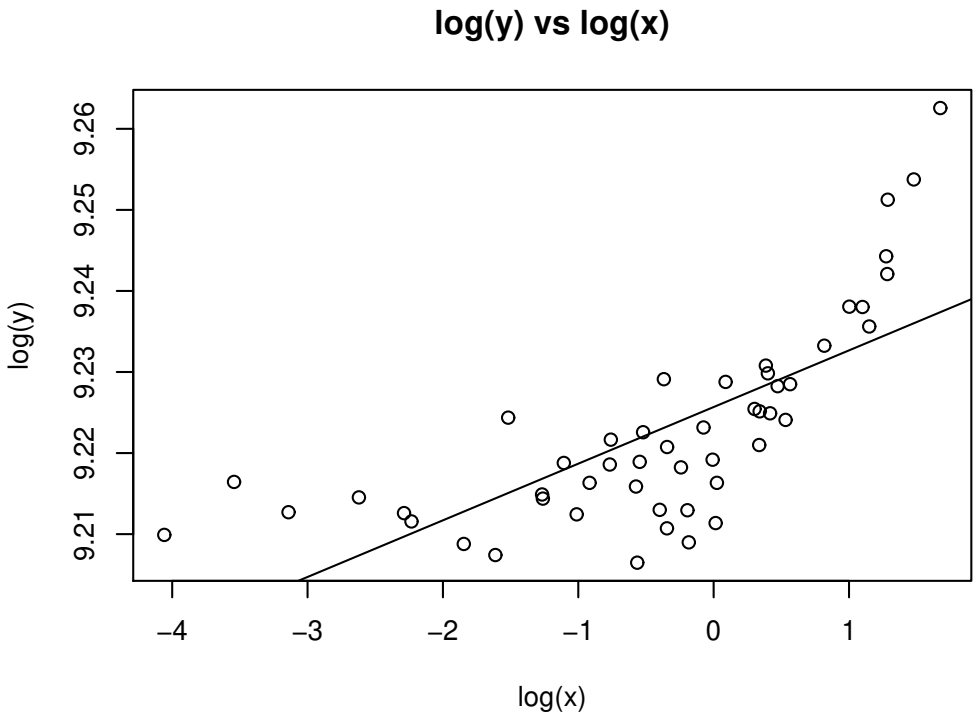
What if both X and Y are skewed?





Plots and fits of simulated data





Definitions

1. An **experimental unit** is the smallest unit to which a treatment can be applied
2. A **treatment** is the entire description of what can be applied to an experimental unit
3. An **observational unit** is the smallest unit on which a response will be measured; it is often be called a **plot** for brevity
4. **Treatment structure** means meaningful ways of dividing up the whole set of treatments
5. **Plot structure** means meaningful ways of dividing up the set of plots, ignoring the treatments
6. The **design** is the allocation of treatments to plots

Structure of a designed experiment

1. Set of treatments included in the study
2. Set of experimental units included in the study
3. Rules and procedures by which the treatments are assigned to the experimental units (or vice versa)
4. Measurements that are made on the experimental units after the treatments have been applied

Contributions of statistics to experimentation

Factorial experiments.

- Multifactor investigations permit analysis of a number of factors with the same precision as if the entire experiment had been devoted to the study of only one factor
- In addition, a factorial experiment provides information on interaction effects while classical one-factor-at-a-time approach requires a series of experiments to do so

Replication. Replication refers to the repetition of an experiment

- It permits sampling error to be assessed for testing presence of treatment effects and for establishing confidence intervals for the effects
- Replication also permits control over precision of the estimates or the power of the tests through manipulation of the sample size

Randomization. Treatments are assigned to experimental units at random

- This tends to average out between the treatments whatever systematic effects may be present, apparent or hidden, so that comparisons between treatments measure only the pure effects
- Randomization tends to eliminate influence of extraneous factors not under the direct control of the experimenter and precludes selection bias

Blocking.

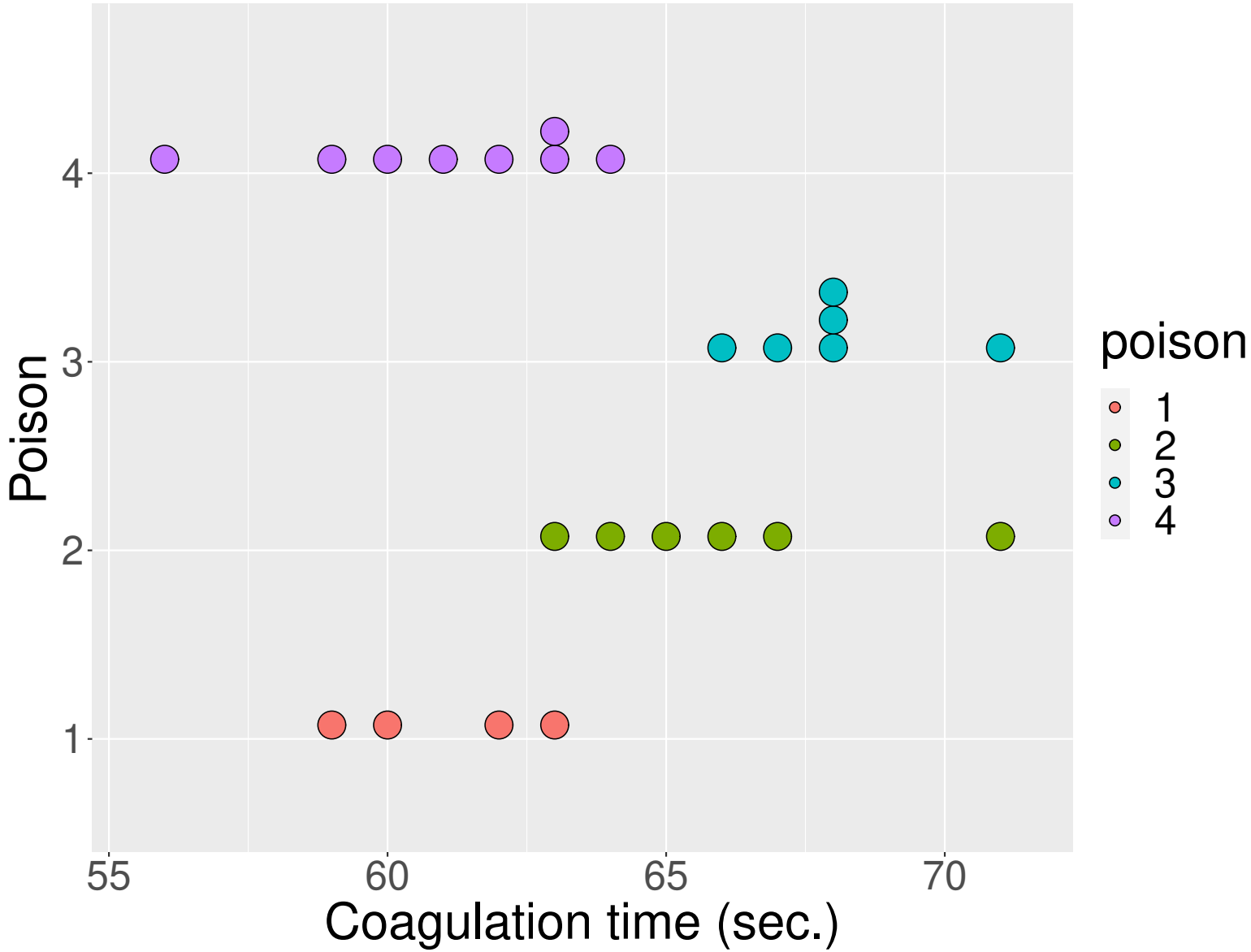
- Blocking is intended to reduce experimental errors and make the experiment more powerful by suitable restrictions on the randomization of treatments to experimental units
- For example, blocking by week will reduce experimental error variability if a time effect is present
- Another benefit of blocking is to increase range of validity for the conclusions of an experiment

Completely randomized design (CRD)

Table 1: Rat poison data

Poison	Blood coagulation times (sec.)							
1	62	60	63	59				
2	63	67	71	64	65	66		
3	68	66	71	67	68	68		
4	56	62	60	61	63	64	63	59

Are the population treatments means equal?



Two-sample t-test

Data: y_{jk} = kth observation from sample j ($j = 1, 2$; $k = 1, 2, \ldots, n_j$)

Null hypothesis: H_0: $\mu_1 = \mu_2$, where $\mu_j = E(y_{jk})$

Sample mean: $y_{j\cdot} = n_j^{-1} \sum_k y_{jk}$

Sample variance: $s_j^2 = (n_j - 1)^{-1} \sum_k (y_{jk} - y_{j\cdot})^2$

Pooled variance:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_j \sum_k (y_{jk} - y_{j\cdot})^2}{n_1 + n_2 - 2}$$

test statistic:

$$t = \frac{y_{1\cdot} - y_{2\cdot}}{s\sqrt{n_1^{-1} + n_2^{-1}}}$$

Two necessary ingredients for a multi-sample test

1. A test statistic
2. A null distribution for the test statistic

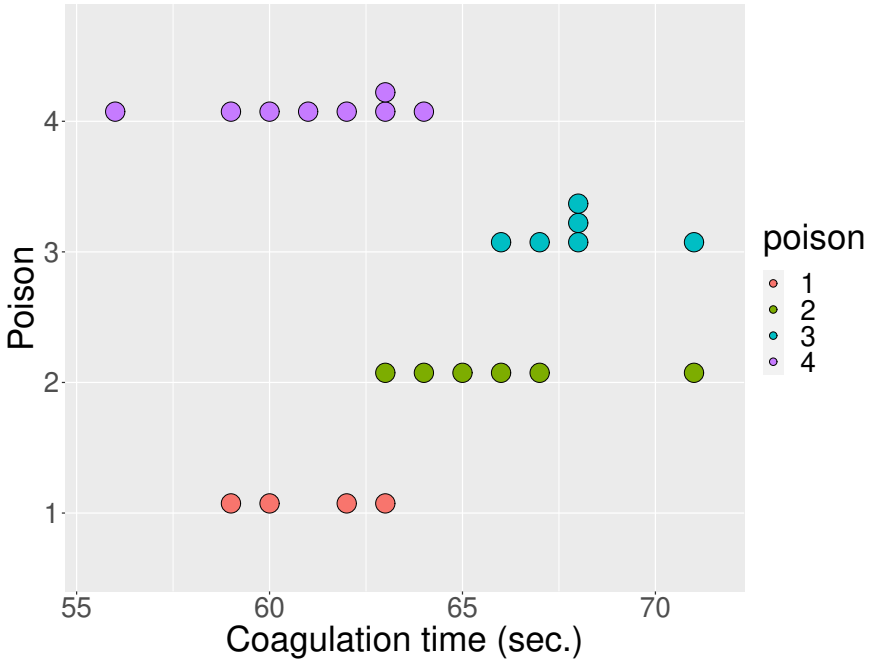
Homework 1
due 9 AM Tue Sep 27 in Canvas

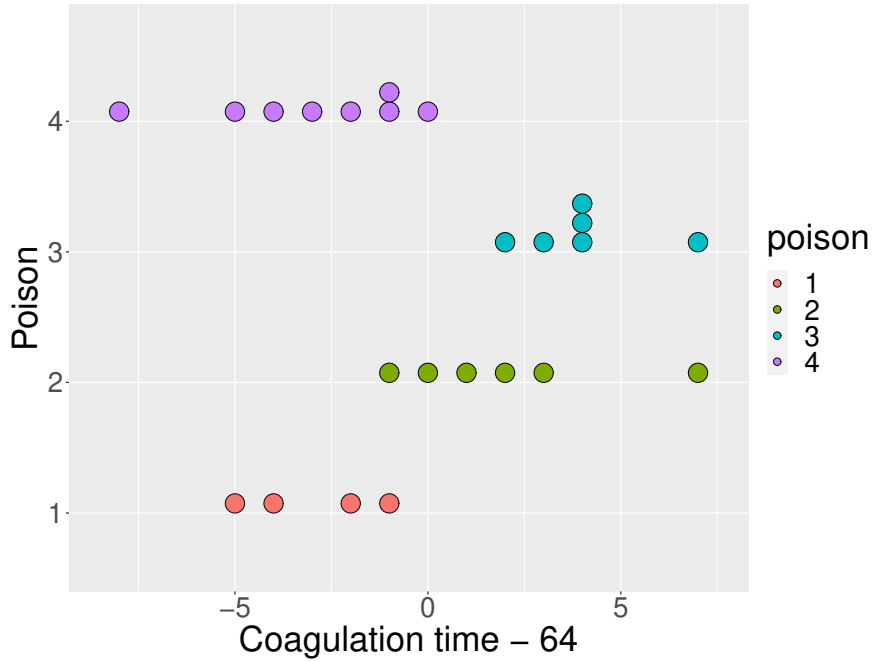
Problems are in hw1.pdf in Files folder on Canvas

Simplifying the data

$$\begin{bmatrix} 62 & 60 & 63 & 59 & & & & \\ 63 & 67 & 71 & 64 & 65 & 66 & & \\ 68 & 66 & 71 & 67 & 68 & 68 & & \\ 56 & 62 & 60 & 61 & 63 & 64 & 63 & 59 \end{bmatrix}$$

$$= 64 + \begin{bmatrix} -2 & -4 & -1 & -5 & & & & \\ -1 & 3 & 7 & 0 & 1 & 2 & & \\ 4 & 2 & 7 & 3 & 4 & 4 & & \\ -8 & -2 & -4 & -3 & -1 & 0 & -1 & -5 \end{bmatrix}$$


Poison
4
3
2
1
55
60
65
70
Coagulation time (sec.)
poison
1
2
3
4


Poison
4
3
2
1
−5
0
5
Coagulation time − 64
poison
1
2
3
4

Data decomposition

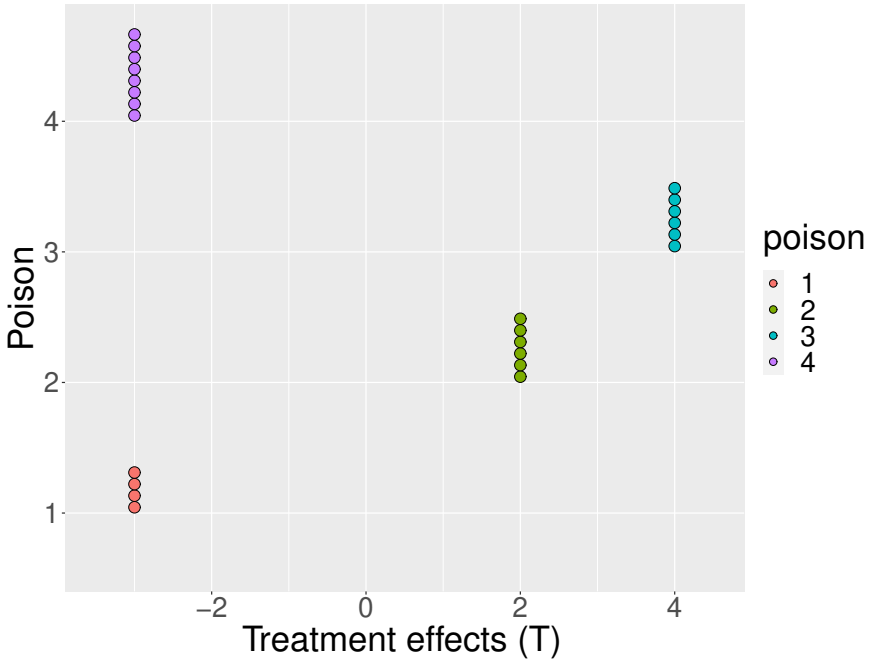
$$\begin{bmatrix} 62 & 60 & 63 & 59 & & & & \\ 63 & 67 & 71 & 64 & 65 & 66 & & \\ 68 & 66 & 71 & 67 & 68 & 68 & & \\ 56 & 62 & 60 & 61 & 63 & 64 & 63 & 59 \end{bmatrix} = 64 + \begin{bmatrix} -2 & -4 & -1 & -5 & & & & \\ -1 & 3 & 7 & 0 & 1 & 2 & & \\ 4 & 2 & 7 & 3 & 4 & 4 & & \\ -8 & -2 & -4 & -3 & -1 & 0 & -1 & -5 \end{bmatrix}$$

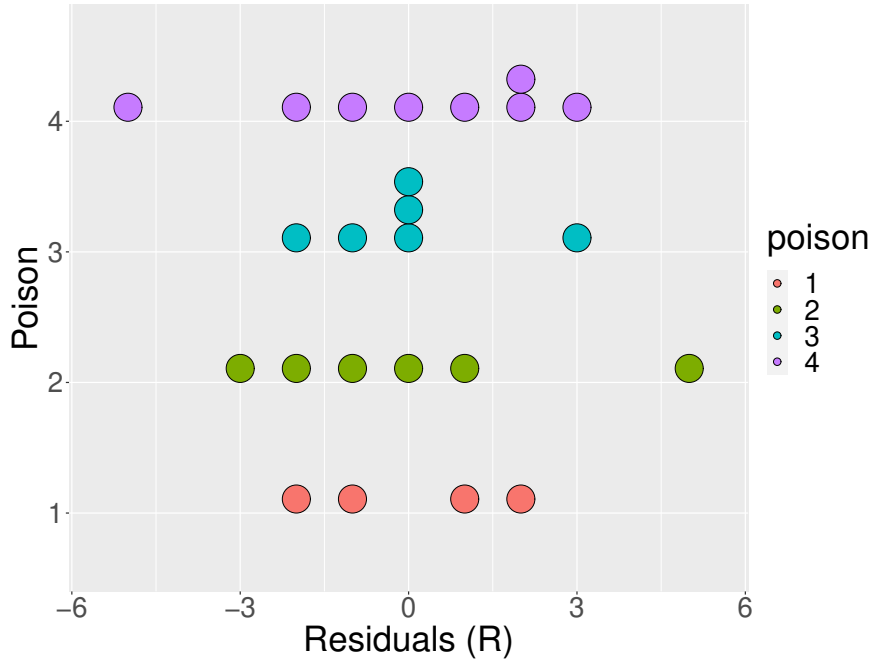
$$= 64 + \begin{bmatrix} -3 & -3 & -3 & -3 & & & & \\ 2 & 2 & 2 & 2 & 2 & 2 & & \\ 4 & 4 & 4 & 4 & 4 & 4 & & \\ -3 & -3 & -3 & -3 & -3 & -3 & -3 & -3 \end{bmatrix}$$

$$+ \begin{bmatrix} 1 & -1 & 2 & -2 & & & & \\ -3 & 1 & 5 & -2 & -1 & 0 & & \\ 0 & -2 & 3 & -1 & 0 & 0 & & \\ -5 & 1 & -1 & 0 & 2 & 3 & 2 & -2 \end{bmatrix}$$

$$= 64 + \mathbf{T} + \mathbf{R}$$

Idea: compare sizes of elements of T and R


Poison
4
3
2
1
−2
0
2
4
Treatment effects (T)
poison
1
2
3
4


Poison
4
3
2
1
−6
−3
0
3
6
Residuals (R)
poison
1
2
3
4

Notations

- t = number of treatments
- n_j = sample size of treatment j ($j = 1, 2, \ldots, t$)
- y_{jk} = kth observation for jth treatment ($k = 1, 2, \ldots, n_j$)

$$
\begin{aligned}
n &= n_1 + \ldots + n_t = \sum_{j=1}^{t} n_j \\
y_{j\cdot} &= \frac{y_{j1} + \ldots + y_{jn_j}}{n_j} = n_j^{-1} \sum_{k=1}^{n_j} y_{jk} \\
y_{\cdot\cdot} &= \frac{y_{11} + y_{12} + \ldots + y_{tn_t}}{n} = n^{-1} \sum_{j=1}^{t} \sum_{k=1}^{n_j} y_{jk} \\
&= \frac{n_1 y_{1\cdot} + \ldots + n_t y_{t\cdot}}{n_1 + \ldots + n_t} = n^{-1} \sum_{j=1}^{t} n_j y_{j\cdot}
\end{aligned}
$$

Data decomposition

$$y_{jk} \equiv y_{\cdot\cdot} + (y_{j\cdot} - y_{\cdot\cdot}) + (y_{jk} - y_{j\cdot})$$

$$\begin{bmatrix} 62 & 60 & 63 & 59 & & & & \\ 63 & 67 & 71 & 64 & 65 & 66 & & \\ 68 & 66 & 71 & 67 & 68 & 68 & & \\ 56 & 62 & 60 & 61 & 63 & 64 & 63 & 59 \end{bmatrix} = 64 + \begin{bmatrix} -3 & -3 & -3 & -3 & & & & \\ 2 & 2 & 2 & 2 & 2 & 2 & & \\ 4 & 4 & 4 & 4 & 4 & 4 & & \\ -3 & -3 & -3 & -3 & -3 & -3 & -3 & -3 \end{bmatrix}$$

$$+ \begin{bmatrix} 1 & -1 & 2 & -2 & & & & \\ -3 & 1 & 5 & -2 & -1 & 0 & & \\ 0 & -2 & 3 & -1 & 0 & 0 & & \\ -5 & 1 & -1 & 0 & 2 & 3 & 2 & -2 \end{bmatrix}$$

$$S_T = \sum_{j=1}^{t} \sum_{k=1}^{n_j} (y_{j\cdot} - y_{\cdot\cdot})^2 = \sum_{j=1}^{t} n_j (y_{j\cdot} - y_{\cdot\cdot})^2$$

$$S_R = \sum_{j} \sum_{k} (y_{jk} - y_{j\cdot})^2$$

- **Model:** $y_{jk} = \mu_j + \epsilon_{jk}$ ($k = 1, \ldots, n_j$; $j = 1, \ldots, t$), where μ_j is the mean for treatment j and ϵ_{jk} are independent with $E_{jk}(\epsilon) = 0$ and $V(\epsilon_{jk}) = \sigma^2$
- **Hypothesis:** H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$
- Let $t_j = (\mu_j - \mu)$, where μ is any number such that $\min_j \mu_j \leq \mu \leq \max_j \mu_j$
- Rewrite the model as

$$\begin{aligned} y_{jk} &= \mu + (\mu_j - \mu) + \epsilon_{jk} \\ &= \mu + t_j + \epsilon_{jk} \end{aligned}$$

- H_0 can be expressed in 3 equivalent forms:

$$H_0 : t_1 = t_2 = \ldots = t_t = 0$$

$$H_0 : \sum_{j=1}^{t} t_j^2 = 0$$

$$H_0 : \sum_{j=1}^{t} n_t t_j^2 = 0$$

- Choose $\mu = n^{-1} \sum_j n_j \mu_j$ and define $\hat{\mu}_j = y_{j\cdot}$.
- Then

$$\begin{aligned} \hat{\mu} &= n^{-1} \sum_j n_j \hat{\mu}_j = y_{\cdot\cdot} \\ \hat{t}_j &= \hat{\mu}_j - \hat{\mu} = y_{j\cdot} - y_{\cdot\cdot} \\ \hat{\epsilon}_{jk} &= y_{jk} - \hat{\mu}_j = y_{jk} - y_{j\cdot} \end{aligned}$$

and

$$\begin{aligned} y_{jk} &\equiv y_{\cdot\cdot} + (y_{j\cdot} - y_{\cdot\cdot}) + (y_{jk} - y_{j\cdot}) \\ &= \hat{\mu} + \hat{t}_j + \hat{\epsilon}_{jk} \end{aligned}$$

- Therefore

$$\begin{aligned} S_T &= \sum_j n_j (y_{j\cdot} - y_{\cdot\cdot})^2 = \sum_j n_j \hat{t}_j^2 \\ S_R &= \sum_j \sum_k (y_{jk} - y_{j\cdot})^2 = \sum_j \sum_k \hat{\epsilon}_{jk}^2 \end{aligned}$$

- Note: $\hat{\mu} \neq t^{-1} \sum_j \hat{\mu}_j$

ANOVA data decomposition

- Squaring both sides of the identity

$$y_{jk} - y_{\cdot\cdot} \equiv (y_{j\cdot} - y_{\cdot\cdot}) + (y_{jk} - y_{j\cdot})$$

gives (cross-product vanishes)

$$\begin{aligned}
\sum_j \sum_k (y_{jk} - y_{\cdot\cdot})^2 &\equiv \sum_j \sum_k (y_{j\cdot} - y_{\cdot\cdot})^2 + \sum_j \sum_k (y_{jk} - y_{j\cdot})^2 \\
&\equiv \sum_j n_j (y_{j\cdot} - y_{\cdot\cdot})^2 + \sum_j \sum_k (y_{jk} - y_{j\cdot})^2 \\
&\equiv S_T + S_R
\end{aligned}$$

- Let $S_D = \sum_j \sum_k (y_{jk} - y_{\cdot\cdot})^2$
- Then

$$S_D \equiv S_T + S_R$$

ANOVA Table for One-Factor Design

Source	Sum of squares	df	MS	F
Treatment	$S_T = \sum_j n_j(y_{j\cdot} - y_{\cdot\cdot})^2$	$\nu_T = t - 1$	S_T/ν_T	$\frac{S_T/\nu_T}{S_R/\nu_R}$
Residual	$S_R = \sum_j \sum_k (y_{jk} - y_{j\cdot})^2$	$\nu_R = n - t$	S_R/ν_R	
Total	$S_D = \sum_j \sum_k (y_{jk} - y_{\cdot\cdot})^2$	$\nu_D = n - 1$		

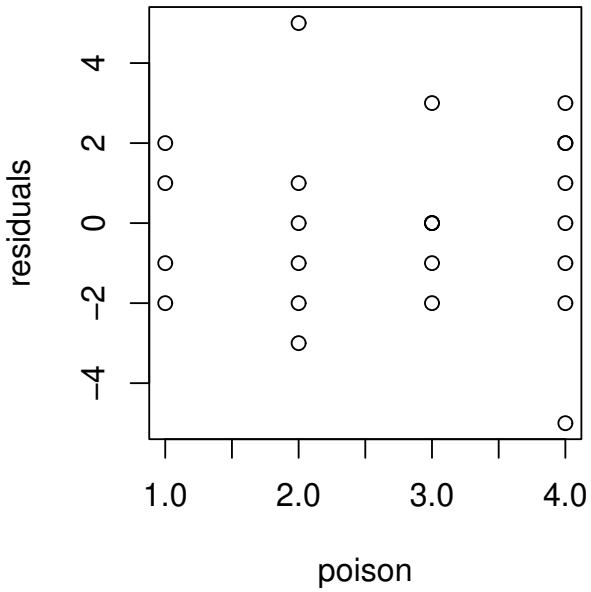
If $\epsilon_{jk} \sim N(0, \sigma^2)$, null distribution of F is F_{ν_T, ν_R}

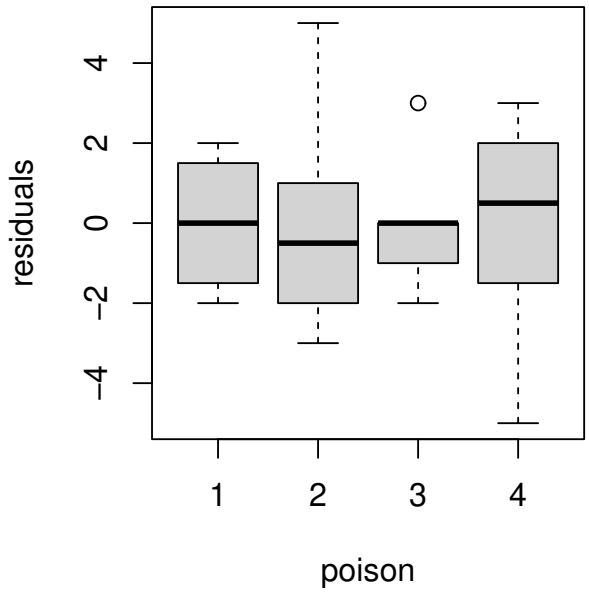
```
> time <- c(62,60,63,59, 63,67,71,64,65,66,68,66,71,67,68,68,56,
            62,60,61,63,64,63,59)
> poison <- as.factor(rep(1:4,c(4,6,6,8)))
> lm(time ~ poison)
          Df Sum Sq Mean Sq F value    Pr(>F)
poison     3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
```

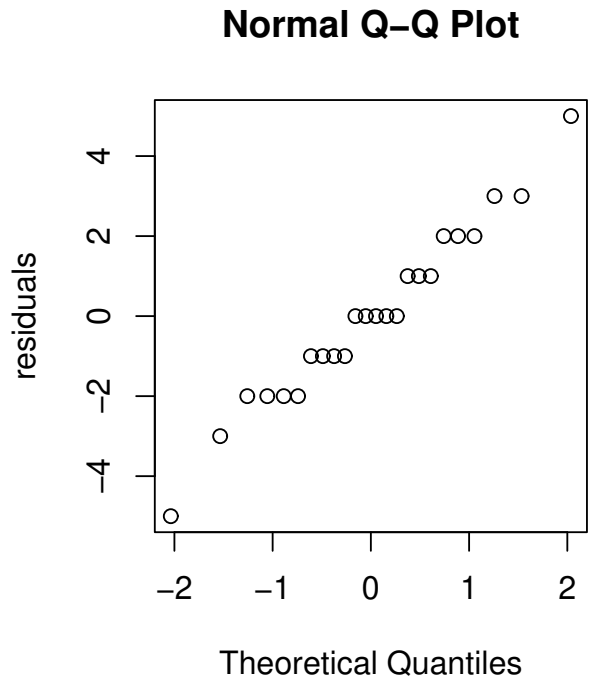
Properties of F-test

1. The ANOVA F-test is uniformly most powerful invariant, where the group of transformations include location and scale changes
2. The distribution of F depends only on t and n, not on the individual values of $n_1, n_2, \ldots, n_t$

Diagnostic plots for model $y_{jk} = \mu_j + \epsilon_{jk}$







R code for diagnostic plots

```
par(mfrow=c(1,3),pty="s",cex=1.05)
time <- c(62,60,63,59, 63,67,71,64,65,66,68,66,71,67,68,68,56,
            62,60,61,63,64,63,59)
poison <- as.factor(rep(1:4,c(4,6,6,8)))
model <- lm(time ~ poison)
resids <- model$res
plot(resids ~ as.numeric(poison),xlab="poison",ylab="residuals")
plot(resids ~ poison,xlab="poison",ylab="residuals")
qqnorm(resids,ylab="residuals")
```

Confidence intervals for linear combinations of means

- Given coefficients $c_1, c_2, \ldots, c_t$, let

$$l = \sum_j c_j \mu_j \qquad \hat{l} = \sum_j c_j \hat{\mu}_j$$

- Then

$$E(\hat{l}) = \sum_j c_j \mu_j \qquad V(\hat{l}) = \sigma^2 \sum_j c_j^2 / n_j$$

and

$$\frac{\sum_j c_j \hat{\mu}_j - \sum_j c_j \mu_j}{\sqrt{\hat{\sigma}^2 \sum_j n_j^{-1} c_j^2}} \sim t_{\nu_R}$$

- $100(1-\alpha)$% confidence interval for $\sum_j c_j \mu_j$ is

$$\sum_j c_j \hat{\mu}_j \pm t_{\alpha/2, \nu_R} \hat{\sigma} \sqrt{\sum_j c_j^2 / n_j}$$

Comparisonwise and experimentwise error rates

- The *comparisonwise error rate* is the probability of rejecting a particular hypothesis in a single test when the hypothesis is true
- The *experimentwise or familywise error rate* is the probability of rejecting one or more hypotheses when all hypotheses are true

Simultaneous confidence intervals

Confidence interval formulas for $\sum_i c_i \mu_i$ have the form

$$\sum_j c_j \hat{\mu}_j \pm \text{Multiplier} \times \hat{\sigma} \sqrt{\sum_j c_j^2 / n_j}$$

Bonferroni method

- Multiplier = $t_{\nu_R;\alpha/(2p)}$ for p comparisons
- Controls the experimentwise error rate

Scheffé method

- A *contrast* is a linear combination $\sum_j c_j \mu_j$ such that $\sum_j c_j = 0$
- It can be shown that

$$P\left\{ \frac{\left(\sum_j c_j \hat{\mu}_j - \sum_j c_j \mu_j\right)^2}{\sum_j c_j^2 / n_j} \leq (t-1)\hat{\sigma}^2 F_{t-1,\nu_R;\alpha} \quad \text{for all contrasts} \right\} = 1-\alpha$$

- Therefore the multiplier

$$\sqrt{(t-1)F_{t-1,\nu_R;\alpha}}$$

  can be used to construct simultaneous confidence intervals for all contrasts with experimentwise error rate α

- If one wants simultaneous confidence intervals for all linear combinations, replace $(t-1)$ by t, i.e., use the multiplier

$$\sqrt{tF_{t,\nu_R;\alpha}}$$

Tukey method

- If the n_i are all equal to n_0 (say), simultaneous intervals for all contrasts are

$$\sum_j c_j \hat{\mu}_j \pm \frac{q(t, \nu_R, \alpha)\hat{\sigma}}{2\sqrt{n_0}} \sum_j |c_j|$$

  where $q(t, \nu_R, \alpha)$ is the upper α quantile of the Studentized range statistic

- This method yields an experimentwise error rate less than or equal to α
- For pairwise contrasts, this simplifies to

$$\hat{\mu}_i - \hat{\mu}_j \pm \frac{q(t, \nu_R, \alpha)\hat{\sigma}}{\sqrt{n_0}}$$

- If the n_j are not all equal and pairwise comparisons are required, use

$$\hat{\mu}_i - \hat{\mu}_j \pm \frac{q(t, \nu_R, \alpha)\hat{\sigma}}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$q(k, \nu_R, 0.05)$

	k -->																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.969	26.976	32.819	37.082	40.408	43.119	45.397	47.357	49.071	50.592	51.957	53.194	54.323	55.361	56.320	57.212	58.044	58.824	59.558
2	6.085	8.331	9.798	10.881	11.734	12.435	13.027	13.539	13.988	14.389	14.749	15.076	15.375	15.650	15.905	16.143	16.365	16.573	16.769
3	4.501	5.910	6.825	7.502	8.037	8.478	8.852	9.177	9.462	9.717	9.946	10.155	10.346	10.522	10.686	10.838	10.980	11.114	11.240
4	3.926	5.040	5.757	6.287	6.706	7.053	7.347	7.602	7.826	8.027	8.208	8.373	8.524	8.664	8.793	8.914	9.027	9.133	9.233
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.801	6.995	7.167	7.323	7.466	7.596	7.716	7.828	7.932	8.030	8.122	8.208
6	3.460	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508	7.586
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.997	6.158	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097	7.169
8	3.261	4.041	4.529	4.886	5.167	5.399	5.596	5.767	5.918	6.053	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.801	6.869
9	3.199	3.948	4.415	4.755	5.024	5.244	5.432	5.595	5.738	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579	6.643
10	3.151	3.877	4.327	4.654	4.912	5.124	5.304	5.460	5.598	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405	6.467
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.486	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265	6.325
12	3.081	3.773	4.199	4.508	4.750	4.950	5.119	5.265	5.395	5.510	5.615	5.710	5.797	5.878	5.953	6.023	6.089	6.151	6.209
13	3.055	3.734	4.151	4.453	4.690	4.884	5.049	5.192	5.318	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055	6.112
14	3.033	3.701	4.111	4.407	4.639	4.829	4.990	5.130	5.253	5.364	5.463	5.554	5.637	5.714	5.785	5.852	5.915	5.973	6.029
15	3.014	3.673	4.076	4.367	4.595	4.782	4.940	5.077	5.198	5.306	5.403	5.492	5.574	5.649	5.719	5.785	5.846	5.904	5.958
16	2.998	3.649	4.046	4.333	4.557	4.741	4.896	5.031	5.150	5.256	5.352	5.439	5.519	5.593	5.662	5.726	5.786	5.843	5.896
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108	5.212	5.306	5.392	5.471	5.544	5.612	5.675	5.734	5.790	5.842
18	2.971	3.609	3.997	4.276	4.494	4.673	4.824	4.955	5.071	5.173	5.266	5.351	5.429	5.501	5.567	5.629	5.688	5.743	5.794
19	2.960	3.593	3.977	4.253	4.468	4.645	4.794	4.924	5.037	5.139	5.231	5.314	5.391	5.462	5.528	5.589	5.647	5.701	5.752
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.895	5.008	5.108	5.199	5.282	5.357	5.427	5.492	5.553	5.610	5.663	5.714
21	2.941	3.565	3.942	4.213	4.424	4.597	4.743	4.870	4.981	5.081	5.170	5.252	5.327	5.396	5.460	5.520	5.576	5.629	5.679
22	2.933	3.553	3.927	4.196	4.405	4.577	4.722	4.847	4.957	5.056	5.144	5.225	5.299	5.368	5.431	5.491	5.546	5.599	5.648
23	2.926	3.542	3.914	4.180	4.388	4.558	4.702	4.826	4.935	5.033	5.121	5.201	5.274	5.342	5.405	5.464	5.519	5.571	5.620
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545	5.594
25	2.913	3.523	3.890	4.153	4.358	4.526	4.667	4.789	4.897	4.993	5.079	5.158	5.230	5.297	5.359	5.417	5.471	5.522	5.570

www.real-statistics.com/statistics-tables/studentized-range-q-table/

$q(k, \nu_R, 0.05)$, cont'd.

df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
26	2.907	3.514	3.880	4.141	4.345	4.511	4.652	4.773	4.880	4.975	5.061	5.139	5.211	5.277	5.339	5.396	5.450	5.500	5.548
27	2.902	3.506	3.870	4.130	4.333	4.498	4.638	4.758	4.864	4.959	5.044	5.122	5.193	5.259	5.320	5.377	5.430	5.480	5.528
28	2.897	3.499	3.861	4.120	4.322	4.486	4.625	4.745	4.850	4.944	5.029	5.106	5.177	5.242	5.302	5.359	5.412	5.462	5.509
29	2.892	3.493	3.853	4.111	4.311	4.475	4.613	4.732	4.837	4.930	5.014	5.091	5.161	5.226	5.286	5.342	5.395	5.445	5.491
30	2.888	3.486	3.845	4.102	4.301	4.464	4.601	4.720	4.824	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429	5.475
31	2.884	3.481	3.838	4.094	4.292	4.454	4.591	4.709	4.812	4.905	4.988	5.064	5.134	5.198	5.257	5.313	5.365	5.414	5.460
32	2.881	3.475	3.832	4.086	4.284	4.445	4.581	4.698	4.802	4.894	4.976	5.052	5.121	5.185	5.244	5.299	5.351	5.400	5.445
33	2.877	3.470	3.825	4.079	4.276	4.436	4.572	4.689	4.791	4.883	4.965	5.040	5.109	5.173	5.232	5.287	5.338	5.386	5.432
34	2.874	3.465	3.820	4.072	4.268	4.428	4.563	4.680	4.782	4.873	4.955	5.030	5.098	5.161	5.220	5.275	5.326	5.374	5.420
35	2.871	3.461	3.814	4.066	4.261	4.421	4.555	4.671	4.773	4.863	4.945	5.020	5.088	5.151	5.209	5.264	5.315	5.362	5.408
36	2.868	3.457	3.809	4.060	4.255	4.414	4.547	4.663	4.764	4.855	4.936	5.010	5.078	5.141	5.199	5.253	5.304	5.352	5.397
37	2.865	3.453	3.804	4.054	4.249	4.407	4.540	4.655	4.756	4.846	4.927	5.001	5.069	5.131	5.189	5.243	5.294	5.341	5.386
38	2.863	3.449	3.799	4.049	4.243	4.400	4.533	4.648	4.749	4.838	4.919	4.993	5.060	5.122	5.180	5.234	5.284	5.331	5.376
39	2.861	3.445	3.795	4.044	4.237	4.394	4.527	4.641	4.741	4.831	4.911	4.985	5.052	5.114	5.171	5.225	5.275	5.322	5.367
40	2.858	3.442	3.791	4.039	4.232	4.388	4.521	4.634	4.735	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313	5.358
48	2.843	3.420	3.764	4.008	4.197	4.351	4.481	4.592	4.690	4.777	4.856	4.927	4.993	5.053	5.109	5.161	5.210	5.256	5.299
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199	5.241
80	2.814	3.377	3.711	3.947	4.129	4.277	4.402	4.509	4.603	4.686	4.761	4.829	4.892	4.949	5.003	5.052	5.099	5.142	5.183
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.043	5.086	5.126
240	2.786	3.335	3.659	3.887	4.063	4.205	4.324	4.427	4.517	4.596	4.668	4.733	4.792	4.847	4.897	4.944	4.988	5.030	5.069
inf	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974	5.012

$q(k, \nu_R, 0.10)$

	k -->																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8.929	13.437	16.358	18.488	20.150	21.504	22.642	23.621	24.477	25.237	25.918	26.536	27.100	27.618	28.097	28.542	28.958	29.347	29.713
2	4.129	5.733	6.772	7.538	8.139	8.633	9.049	9.409	9.725	10.006	10.259	10.488	10.698	10.891	11.070	11.237	11.392	11.538	11.676
3	3.328	4.467	5.199	5.738	6.162	6.511	6.806	7.062	7.287	7.487	7.667	7.831	7.982	8.120	8.248	8.368	8.479	8.584	8.683
4	3.015	3.976	4.586	5.035	5.388	5.679	5.926	6.139	6.327	6.494	6.645	6.783	6.909	7.025	7.132	7.233	7.326	7.414	7.497
5	2.850	3.717	4.264	4.664	4.979	5.238	5.458	5.648	5.816	5.965	6.100	6.223	6.336	6.439	6.536	6.626	6.710	6.788	6.863
6	2.748	3.558	4.065	4.435	4.726	4.966	5.168	5.344	5.499	5.637	5.762	5.875	5.979	6.075	6.164	6.247	6.325	6.398	6.466
7	2.679	3.451	3.931	4.280	4.555	4.780	4.971	5.137	5.283	5.413	5.530	5.637	5.735	5.826	5.910	5.988	6.061	6.130	6.195
8	2.630	3.374	3.834	4.169	4.431	4.646	4.829	4.987	5.126	5.250	5.362	5.464	5.558	5.644	5.724	5.799	5.869	5.935	5.997
9	2.592	3.316	3.761	4.084	4.337	4.545	4.721	4.873	5.007	5.126	5.234	5.333	5.423	5.506	5.583	5.655	5.722	5.786	5.845
10	2.563	3.270	3.704	4.018	4.264	4.465	4.636	4.783	4.913	5.029	5.134	5.229	5.316	5.397	5.472	5.542	5.607	5.668	5.726
11	2.540	3.234	3.658	3.965	4.205	4.401	4.567	4.711	4.838	4.951	5.053	5.145	5.231	5.309	5.382	5.450	5.514	5.573	5.630
12	2.521	3.204	3.621	3.921	4.156	4.349	4.511	4.652	4.776	4.886	4.986	5.076	5.160	5.236	5.308	5.374	5.436	5.495	5.550
13	2.504	3.179	3.589	3.885	4.116	4.304	4.464	4.602	4.724	4.832	4.930	5.019	5.100	5.175	5.245	5.310	5.371	5.429	5.483
14	2.491	3.158	3.563	3.854	4.081	4.267	4.424	4.560	4.679	4.786	4.882	4.969	5.050	5.124	5.192	5.256	5.316	5.372	5.426
15	2.479	3.140	3.540	3.828	4.052	4.235	4.390	4.524	4.641	4.746	4.841	4.927	5.006	5.079	5.146	5.209	5.268	5.324	5.376
16	2.469	3.124	3.520	3.804	4.026	4.207	4.360	4.492	4.608	4.712	4.805	4.890	4.968	5.040	5.106	5.169	5.227	5.282	5.333
17	2.460	3.110	3.503	3.784	4.003	4.182	4.334	4.464	4.579	4.681	4.774	4.857	4.934	5.005	5.071	5.133	5.190	5.244	5.295
18	2.452	3.098	3.487	3.766	3.984	4.161	4.310	4.440	4.553	4.654	4.746	4.829	4.905	4.975	5.040	5.101	5.158	5.211	5.262
19	2.445	3.087	3.474	3.751	3.966	4.142	4.290	4.418	4.530	4.630	4.721	4.803	4.878	4.948	5.012	5.072	5.129	5.182	5.232
20	2.439	3.077	3.462	3.736	3.950	4.124	4.271	4.398	4.510	4.609	4.699	4.780	4.855	4.923	4.987	5.047	5.103	5.155	5.205
21	2.433	3.069	3.451	3.724	3.936	4.109	4.255	4.380	4.491	4.590	4.678	4.759	4.833	4.901	4.965	5.024	5.079	5.131	5.180
22	2.428	3.061	3.441	3.712	3.923	4.095	4.239	4.364	4.474	4.572	4.660	4.740	4.814	4.882	4.944	5.003	5.058	5.109	5.158
23	2.424	3.054	3.432	3.701	3.911	4.082	4.226	4.350	4.459	4.556	4.644	4.723	4.796	4.863	4.926	4.984	5.038	5.089	5.138
24	2.420	3.047	3.423	3.692	3.900	4.070	4.213	4.336	4.445	4.541	4.628	4.707	4.780	4.847	4.909	4.966	5.020	5.071	5.119
25	2.416	3.041	3.416	3.683	3.890	4.059	4.201	4.324	4.432	4.528	4.614	4.693	4.765	4.831	4.893	4.950	5.004	5.055	5.102

$q(k, \nu_R, 0.10)$, cont'd.

df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
26	2.412	3.036	3.409	3.675	3.881	4.049	4.191	4.313	4.420	4.515	4.601	4.680	4.751	4.817	4.878	4.936	4.989	5.039	5.086
27	2.409	3.030	3.402	3.667	3.873	4.040	4.181	4.302	4.409	4.504	4.590	4.667	4.739	4.804	4.865	4.922	4.975	5.025	5.072
28	2.406	3.026	3.396	3.660	3.865	4.032	4.172	4.293	4.399	4.493	4.579	4.656	4.727	4.792	4.853	4.909	4.962	5.012	5.058
29	2.403	3.021	3.391	3.654	3.858	4.024	4.163	4.284	4.389	4.484	4.568	4.645	4.716	4.781	4.841	4.897	4.950	4.999	5.046
30	2.400	3.017	3.386	3.648	3.851	4.016	4.155	4.275	4.381	4.474	4.559	4.635	4.706	4.770	4.830	4.886	4.939	4.988	5.034
31	2.398	3.013	3.381	3.642	3.845	4.009	4.148	4.268	4.372	4.466	4.550	4.626	4.696	4.760	4.820	4.876	4.928	4.977	5.023
32	2.396	3.010	3.376	3.637	3.839	4.003	4.141	4.260	4.365	4.458	4.541	4.617	4.687	4.751	4.811	4.866	4.918	4.967	5.013
33	2.393	3.006	3.372	3.632	3.833	3.997	4.135	4.253	4.357	4.450	4.533	4.609	4.679	4.743	4.802	4.857	4.909	4.957	5.003
34	2.391	3.003	3.368	3.627	3.828	3.991	4.129	4.247	4.351	4.443	4.526	4.602	4.671	4.734	4.794	4.849	4.900	4.949	4.994
35	2.389	3.000	3.364	3.623	3.823	3.986	4.123	4.241	4.344	4.436	4.519	4.594	4.663	4.727	4.786	4.841	4.892	4.940	4.986
36	2.388	2.998	3.361	3.619	3.819	3.981	4.117	4.235	4.338	4.430	4.512	4.588	4.656	4.720	4.778	4.833	4.884	4.932	4.978
37	2.386	2.995	3.357	3.615	3.814	3.976	4.112	4.230	4.332	4.424	4.506	4.581	4.650	4.713	4.771	4.826	4.877	4.925	4.970
38	2.384	2.992	3.354	3.611	3.810	3.972	4.107	4.224	4.327	4.418	4.500	4.575	4.643	4.706	4.765	4.819	4.870	4.918	4.963
39	2.383	2.990	3.351	3.608	3.806	3.967	4.103	4.220	4.322	4.413	4.495	4.569	4.637	4.700	4.758	4.812	4.863	4.911	4.956
40	2.381	2.988	3.348	3.605	3.802	3.963	4.099	4.215	4.317	4.408	4.490	4.564	4.632	4.694	4.752	4.806	4.857	4.904	4.949
48	2.372	2.973	3.330	3.583	3.778	3.937	4.070	4.185	4.285	4.375	4.455	4.528	4.595	4.656	4.713	4.766	4.816	4.863	4.907
60	2.363	2.959	3.312	3.562	3.755	3.911	4.042	4.155	4.254	4.342	4.421	4.493	4.558	4.619	4.675	4.727	4.775	4.821	4.864
80	2.353	2.945	3.294	3.541	3.731	3.885	4.014	4.125	4.223	4.309	4.387	4.457	4.521	4.581	4.636	4.687	4.735	4.780	4.822
120	2.344	2.930	3.276	3.520	3.707	3.859	3.986	4.096	4.191	4.276	4.353	4.422	4.485	4.543	4.597	4.647	4.694	4.738	4.779
240	2.335	2.916	3.258	3.499	3.684	3.834	3.959	4.066	4.160	4.244	4.319	4.386	4.448	4.505	4.558	4.607	4.653	4.696	4.737
inf	2.326	2.902	3.240	3.478	3.661	3.808	3.931	4.037	4.129	4.211	4.285	4.351	4.412	4.468	4.519	4.568	4.612	4.654	4.694

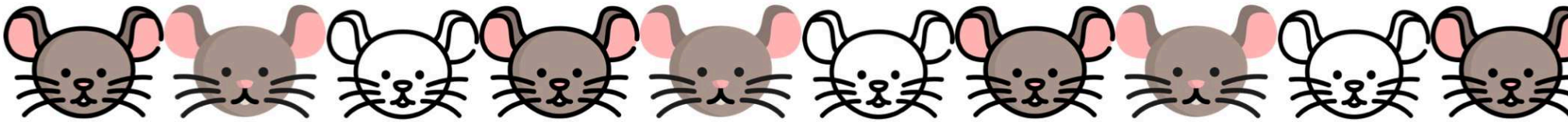
Multipliers for 95% simultaneous intervals in rat poison example

Bonferroni: $t_{20,0.025/6} = 2.927$ (all 6 pairwise differences in means)

Tukey: $q(4, 20, 0.05)/\sqrt{2} = 3.958/\sqrt{2} = 2.799$ (all 6 pairwise differences in means)

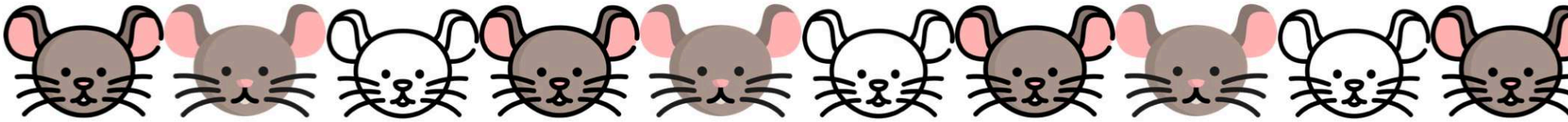
Scheff'e: $\sqrt{3F_{3,20;0.05}} = 3.049$ (all contrasts)

How to assign the treatments to the mice?

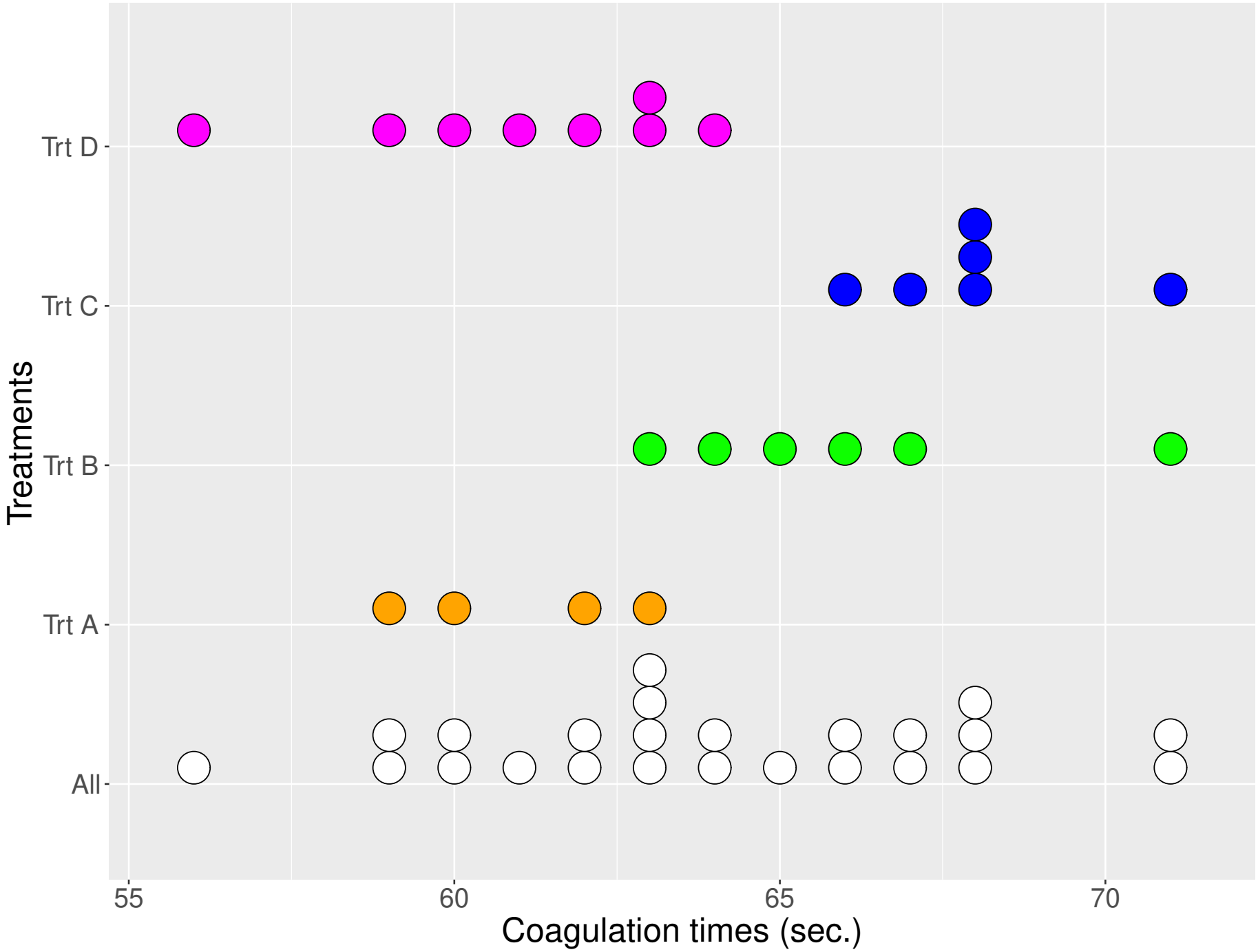


https://www.flaticon.com/free-icons/mouse

Which treatment assignment method is best?



1. Let an expert make the treatment assignments
2. Make all 24 doses available to all the mice and let them choose
3. Toss a 4-sided die that has probabilities (4/24, 6/24, 6/24, 8/24) of showing (A, B, C, D) along row of mice, skipping any treatments that are depleted
4. Put all 24 mice in a box; randomly pick one at a time, assigning 1st 4 mice to A, next 6 to B, etc.
5. Randomly draw tickets without replacement from a box of tickets containing 4 A, 6 B, 6 C, and 8 D and assign one ticket to each mouse


Trt D
Trt C
Trt B
Trt A
All
Treatments
55
60
65
70
Coagulation times (sec.)

Observed data

Poison	Blood coagulation times (sec.)							
A	62	60	63	59				
B	63	67	71	64	65	66		
C	68	66	71	67	68	68		
D	56	62	60	61	63	64	63	59

Two treatments, 2 samples per treatment

62	60	57	55
A	A	B	B

Normal theory t test:

$$\bar{y}_A = 61, \quad \bar{y}_B = 56, \quad s^2 = 2, \quad |t| = 3.536$$

$$\text{p-value} = P(|t_2| > 3.536) = 0.072$$

Random assignment by tossing a coin left to right, subject to 2 A's and 2 B's

62	60	57	55
A	A	B	B

Outcome				$\lvert t \rvert$	Probability
A	A	B	B	3.536	1/4
A	B	A	B	0.566	1/8
A	B	B	A	0	1/8
B	A	A	B	0	1/8
B	A	B	A	0.566	1/8
B	B	A	A	3.536	1/4

2-sided p-value = P($|t| \geq 3.536$) = 1/4 + 1/4 = 1/2

Random assignment by drawing without replacement tickets from a box with 2 A's and 2 B's

62	60	57	55
A	A	B	B

Outcome				$\lvert t \rvert$	Probability
A	A	B	B	3.536	1/6
A	B	A	B	0.566	1/6
A	B	B	A	0	1/6
B	A	A	B	0	1/6
B	A	B	A	0.566	1/6
B	B	A	A	3.536	1/6

2-sided p-value = P($|t| \geq 3.536$) = 1/6 + 1/6 = 1/3

Randomization test

- Under the null hypothesis that the treatments are the same, the same coagulation times would be obtained if the treatment labels are changed
- A randomization distribution is obtained by switching treatment labels
- Number of permutations of 24 numbers is $24! = 6.2 \times 10^{23}$
- For the poison data, the number treatment assignments by lottery is

$$\frac{24!}{4!\,6!\,6!\,8!} \approx 10^{12}$$

- It takes a C program 320 hours to do a complete enumeration of the 10^{12} combinations of which 40,875,222 produced a treatment SS greater than or equal to that observed, giving a randomization p-value of 3.305×10^{-5}
- The ANOVA F test gives a p-value of 4.658×10^{-5} (see slide 29)

Robustness of normal-theory F test

- Randomization method is always valid if random assignment is done
- Normal-theory method is correct and most powerful if the IIDN assumptions are satisfied
- Always randomize because we never know if the IIDN assumptions hold
- But randomization method has less power than normal-theory method
- And randomization method is highly compute-intensive
- In practice, we randomize but use normal-theory method to find p-values
- Justified because as the sample sizes tend to infinity, randomization distribution of F statistic converges to the F distribution

Two-factor experiment: Survival times (10 hrs) of mice

	Antidote (Q)			
Poison (P)	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Balanced two-way factorial: notations

- Denote the factors by P and Q, with levels $j = 1, 2, \ldots, p$; $k = 1, 2, \ldots, q$
- Let μ_{jk} denote the mean yield at level j of P and level k of Q

Notations (cont'd.)

$$\begin{aligned}
\mu_{\cdot\cdot} &= (pq)^{-1} \sum_j \sum_k \mu_{jk} \\
\mu_{j\cdot} &= q^{-1} \sum_k \mu_{jk} \\
\mu_{\cdot k} &= p^{-1} \sum_j \mu_{jk} \\
p_j &= \mu_{j\cdot} - \mu_{\cdot\cdot} \\
q_k &= \mu_{\cdot k} - \mu_{\cdot\cdot}
\end{aligned}$$

The definitions of p_j and q_k imply the **sum-to-zero** constraints

$$\begin{aligned}
\sum_j p_j &= \sum_j (\mu_{j\cdot} - \mu_{\cdot\cdot}) = 0 \\
\sum_k q_k &= \sum_k (\mu_{\cdot k} - \mu_{\cdot\cdot}) = 0
\end{aligned}$$

- **Additive model:** $\mu_{jk} = \mu_{\cdot\cdot} + p_j + q_k$
- In general, $\mu_{jk} - (\mu_{\cdot\cdot} + p_j + q_k) \neq 0$
- Define

$$\begin{aligned}
(pq)_{jk} &= \mu_{jk} - (\mu_{\cdot\cdot} + p_j + q_k) \\
&= \mu_{jk} - \mu_{\cdot\cdot} - (\mu_{j\cdot} - \mu_{\cdot\cdot}) - (\mu_{\cdot k} - \mu_{\cdot\cdot}) \\
&= \mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu_{\cdot\cdot} \\
&= (\mu_{jk} - \mu_{j\cdot}) - (\mu_{\cdot k} - \mu_{\cdot\cdot}) \\
&= (\mu_{jk} - \mu_{\cdot k}) - (\mu_{j\cdot} - \mu_{\cdot\cdot})
\end{aligned}$$

- This implies the **sum-to-zero** constraints

$$\sum_j (pq)_{jk} = \sum_j (\mu_{jk} - \mu_{\cdot\cdot} - p_j - q_k) = p(\mu_{\cdot k} - \mu_{\cdot\cdot} - q_k) = 0 \quad \text{for all } k$$

$$\sum_k (pq)_{jk} = \sum_k (\mu_{jk} - \mu_{\cdot\cdot} - p_j - q_k) = q(\mu_{j\cdot} - \mu_{\cdot\cdot} - p_j) = 0 \quad \text{for all } j$$

- Definition of $(pq)_{jk}$ implies $\mu_{jk} = \mu_{\cdot\cdot} + p_j + q_k + (pq)_{jk}$

ANOVA decomposition

- Model is $y_{jkl} = \mu_{jk} + \epsilon_{jkl}$ or, equivalently,

$$y_{jkl} = \mu + p_j + q_k + (pq)_{jk} + \epsilon_{jkl}, \quad j = 1, \ldots, p;\ k = 1, \ldots, q;\ l = 1, \ldots, r,$$

  where ϵ_{jkl} are i.i.d. $N(0, \sigma^2)$

- Data decomposition:

$$\begin{aligned} y_{jkl} &= \hat{\mu}_{jk} + \hat{\epsilon}_{jkl} \\ &= y_{jk\cdot} + (y_{jkl} - y_{jk\cdot}) \end{aligned}$$

  or, alternatively,

$$\begin{aligned} y_{jkl} &= \hat{\mu} + \hat{p}_j + \hat{q}_k + \widehat{(pq)}_{jk} + \hat{\epsilon}_{jkl} \\ &= y_{\cdots} + (y_{j\cdot\cdot} - y_{\cdots}) + (y_{\cdot k\cdot} - y_{\cdots}) + (y_{jk\cdot} - y_{j\cdot\cdot} - y_{\cdot k\cdot} + y_{\cdots}) \\ &\quad + (y_{jkl} - y_{jk\cdot}) \end{aligned}$$

- Sums of squares:

$$
\begin{aligned}
S_P &= qr \sum_j (y_{j\cdot\cdot} - y_{\cdot\cdot\cdot})^2 \\
S_Q &= pr \sum_k (y_{\cdot k\cdot} - y_{\cdot\cdot\cdot})^2 \\
S_{PQ} &= r \sum_j \sum_k (y_{jk\cdot} - y_{j\cdot\cdot} - y_{\cdot k\cdot} + y_{\cdot\cdot\cdot})^2 \\
S_R &= \sum_j \sum_k \sum_l (y_{jkl} - y_{jk\cdot})^2 \\
S_D &= \sum_j \sum_k \sum_l (y_{jkl} - y_{\cdot\cdot\cdot})^2
\end{aligned}
$$

Survival times (10 hrs) of mice in replicated 3×4

	Antidote (Q)			
Poison (P)	A	B	C	D
1	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
2	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
3	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

ANOVA table

Source	SS	df	MS	Ratio
Factor P	$S_P = 1.03301$	$p - 1 = 2$	$s_P^2 = 0.51651$	$s_P^2/s_R^2 = 23.22$
Factor Q	$S_Q = 0.92121$	$q - 1 = 3$	$s_Q^2 = 0.30707$	$s_Q^2/s_R^2 = 13.81$
Interaction	$S_{PQ} = 0.25014$	$(p-1)(q-1) = 6$	$s_{PQ}^2 = 0.04169$	$s_{PQ}^2/s_R^2 = 1.87$
Residual	$S_R = 0.80073$	$pq(r-1) = 36$	$s_R^2 = 0.02224$	
Total	$S_D = 3.00508$	$pqr - 1 = 47$		

```
y <- c(0.31,0.45,0.46,0.43,0.36,0.29,0.40,0.23,0.22,0.21,0.18,0.23,
0.82,1.10,0.88,0.72,0.92,0.61,0.49,1.24,0.30,0.37,0.38,0.29,0.43,0.45,
0.63,0.76,0.44,0.35,0.31,0.40,0.23,0.25,0.24,0.22,0.45,0.71,0.66,0.62,
0.56,1.02,0.71,0.38,0.30,0.36,0.31,0.33)
p <- factor(rep(rep(1:3,rep(4,3)),4))
q <- factor(rep(1:4,rep(12,4)))

anova(lm(y ~ p+q+p*q))
          Df  Sum Sq Mean Sq F value    Pr(>F)
p          2 1.03301 0.51651 23.2217 3.331e-07 ***
q          3 0.92121 0.30707 13.8056 3.777e-06 ***
p:q        6 0.25014 0.04169  1.8743    0.1123
Residuals 36 0.80072 0.02224

anova(lm(y ~ p+q))
          Df  Sum Sq Mean Sq F value    Pr(>F)
p          2 1.03301 0.51651  20.643 5.704e-07 ***
q          3 0.92121 0.30707  12.273 6.697e-06 ***
Residuals 42 1.05086 0.02502
```

Midterm exam 9:30–10:45am, Thu Nov 3
Rm 1800 Engineering Hall

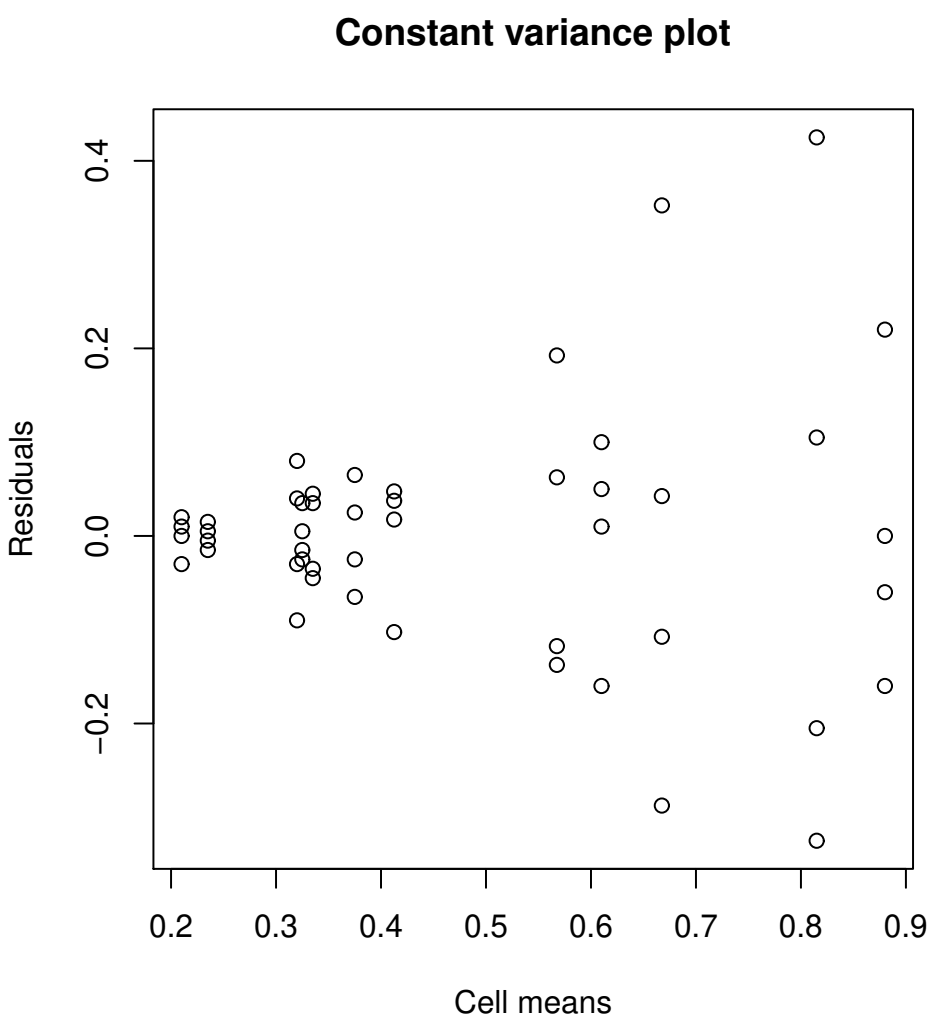
- This is an in-person (i.e., not online) exam
- Leave at least one seat between you and your neighbors
- Exam covers content in lectures and homeworks
- Exam is open-book, open-notes
- You may bring:
  - a handheld calculator
  - Standard normal, t, F, and q (Tukey) statistical tables
  - any books and notes (including paper copies of lecture slides and homework solutions)
- You are not allowed to use devices with communication capabilities, such as cell phones, tablets, and laptops, during the exam

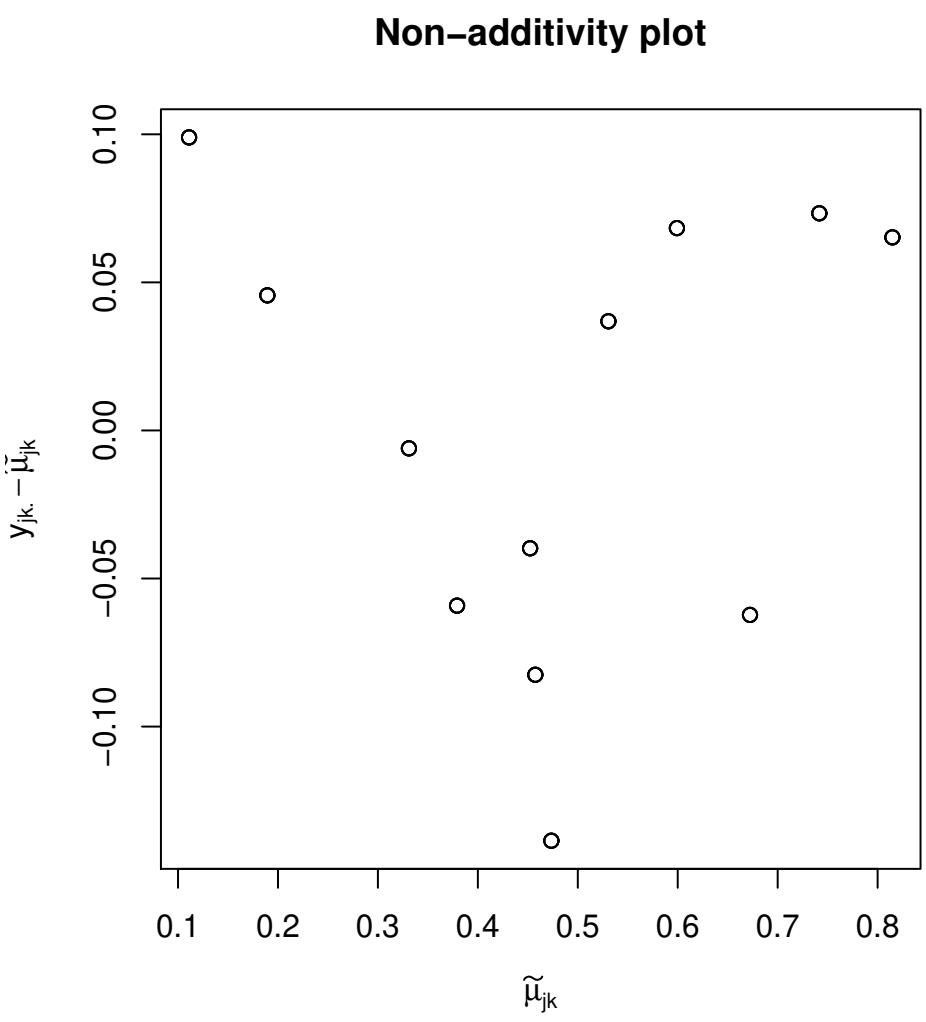
Model checking

1. Let $\hat{\mu}_{jk} = y_{jk\cdot}$ denote the estimated value of μ_{jk} under the full model
2. Plot $y_{jkl} - \hat{\mu}_{jk}$ versus $\hat{\mu}_{jk}$ to check for homogeneity of variance
3. Let $\tilde{\mu}_{jk} = y_{j\cdot\cdot} + y_{\cdot k\cdot} - y_{\cdot\cdot\cdot}$ denote the estimated value of μ_{jk} under the additive model
4. Plot $y_{jk\cdot} - \tilde{\mu}_{jk}$ versus $\tilde{\mu}_{jk}$: curvilinear shape indicates nonadditivity

```
additive <- lm(y ~ p+q)
full <- lm(y ~ p+q+p*q)
par(mfrow=c(1,2),pty="s")
plot(full$residuals ~ full$fitted,ylab="Residuals",xlab="Cell means")
title("Constant variance plot")
xlab <- expression(widetilde(mu)[jk])
ylab <- expression(y[jk.]-widetilde(mu)[jk])
plot(full$fitted-additive$fitted ~ additive$fitted,ylab=ylab,xlab=xlab)
title("Non-additivity plot")
```

Diagnostic plots





Power transformation

Let $\sigma_{jk}^2 = \text{var}(y_{jkl})$ and assume that $\sigma_{jk} \propto \mu_{jk}^{\beta}$ for some β

1. Compute sample mean y_{jk}. and standard deviation s_{jk} for each (j, k)-cell
2. Plot $\log s_{jk}$ versus $\log y_{jk}$. and fit a line to the plot
3. Let $\hat{\beta}$ be the slope of the line and estimate β with $\hat{\beta}$
4. Estimated power transformation is y^{λ}, where $\lambda = 1 - \hat{\beta}$
5. $\lambda = 0$ corresponds to the log transformation
6. Only works with positive observations

Justification

- Define

$$z = f(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0 \end{cases}$$

  so that $f'(y) = y^{\lambda - 1}$ for all λ

- Taylor expansion implies

$$z_{jkl} = f(y_{jkl}) \approx f(\mu_{jk}) + f'(\mu_{jk})(y_{jkl} - \mu_{jk})$$

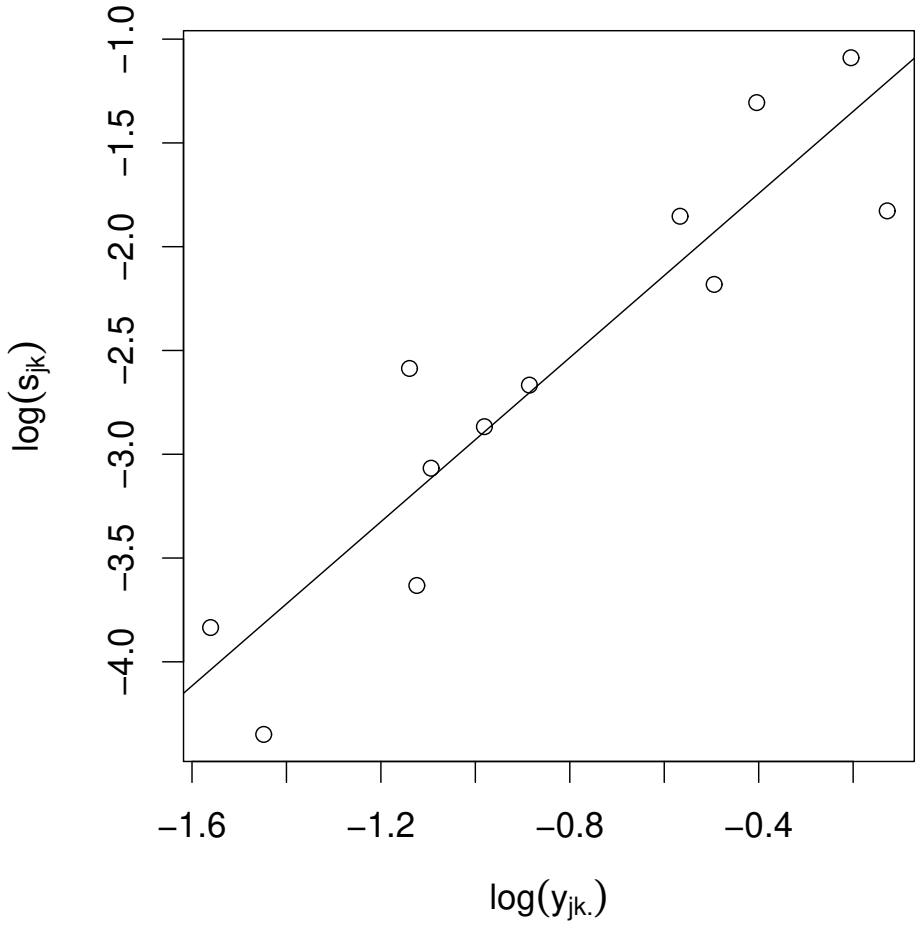
- Taking variance on both sides,

$$\begin{aligned} \text{Var}(z_{jkl}) &\approx (f'(\mu_{jk}))^2 \text{Var}(y_{jkl}) \\ &= \mu_{jk}^{2(\lambda-1)} \sigma_{jk}^2 \\ &\propto \mu_{jk}^{2(\lambda-1)} \mu_{jk}^{2\beta} \\ &= \mu_{jk}^{2(\lambda-1+\beta)} \end{aligned}$$

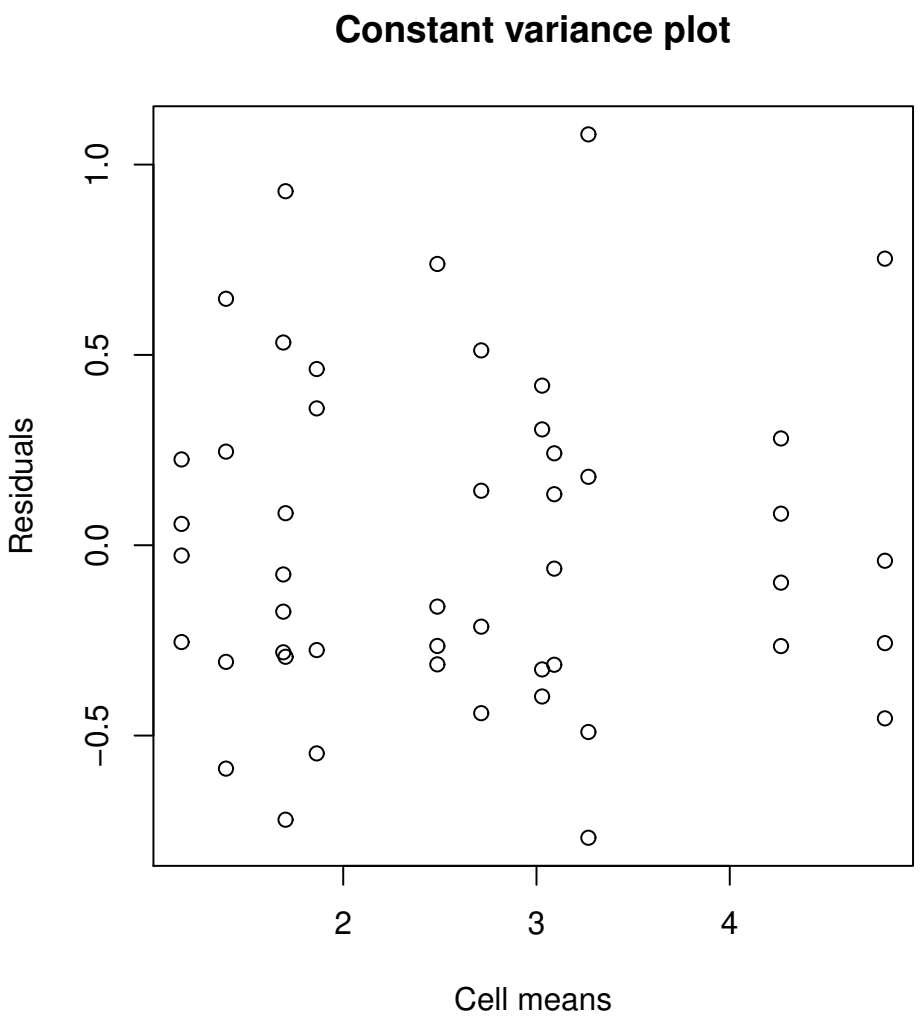
- $\text{Var}(z_{jkl})$ is constant if $\lambda - 1 + \beta = 0 \Rightarrow \lambda = 1 - \beta$

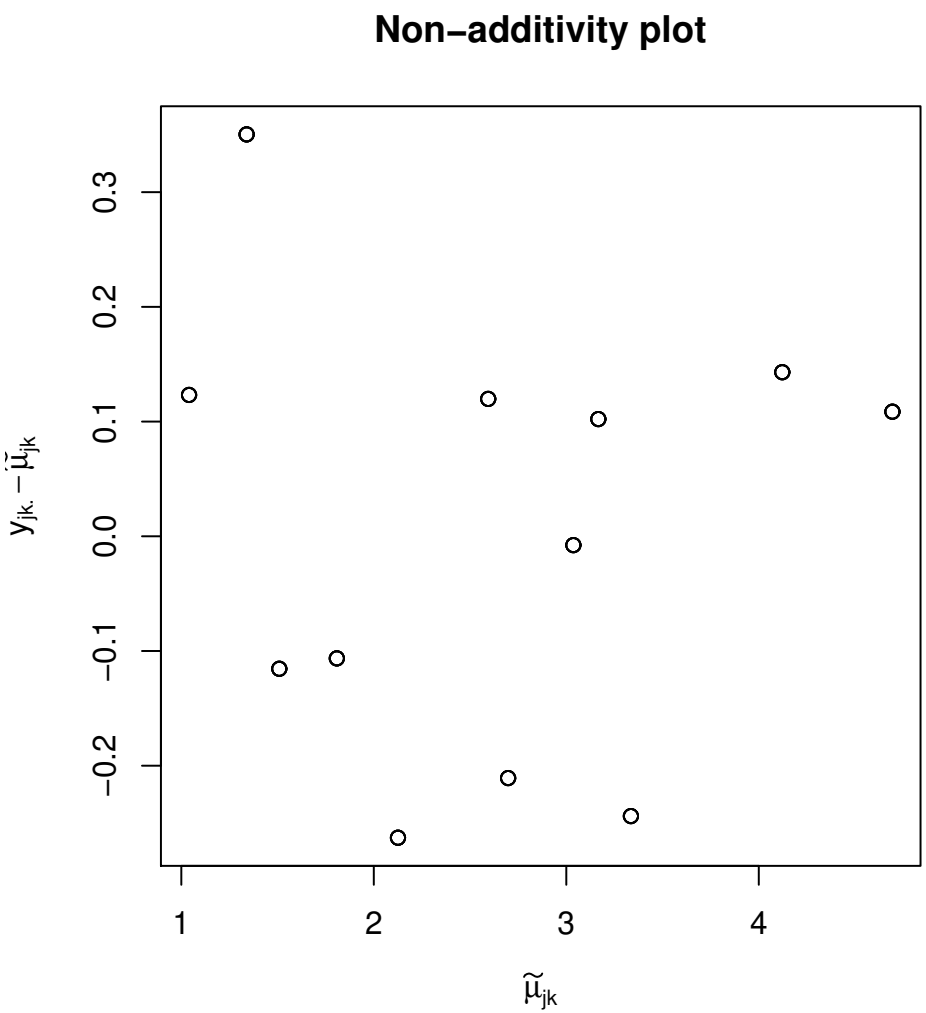
$$\hat{\lambda} = 1 - 1.9770405 \approx -1$$

```
cell.indx <- rep(1:12,rep(4,12))
cell.mean <- tapply(y,cell.indx,mean)
cell.var <- tapply(y,cell.indx,var)
yv <- log(sqrt(cell.var))
xv <- log(cell.mean)
plot(yv ~ xv,ylab=expression(log(s[jk])),
             xlab=expression(log(y[jk.])))
fitline <- lm(yv ~ xv)
abline(fitline)
print(fitline$coef)
(Intercept)          xv
 -0.9528323   1.9770405
```



Diagnostic plots after reciprocal transform





ANOVA before transformation

	Df	Sum Sq	Mean Sq	F value	P-value
Poison	2	1.03301	0.51651	23.2217	3.331e-07
Treatment	3	0.92121	0.30707	13.8056	3.777e-06
Interaction	6	0.25014	0.04169	1.8743	0.1123
Residuals	36	0.80072	0.02224		

ANOVA after reciprocal transformation

	Df	Sum Sq	Mean Sq	F value	P-value
Poison	2	34.877	17.4386	70.6302	5.17e-13
Treatment	3	20.414	6.8048	27.5610	2.48e-09
Interaction	6	1.571	0.2618	1.0603	0.4046
Residuals[a]	35	8.643	0.2469		

[a] Residual df reduced by 1 to account for estimation of λ

Interval estimation when interactions not significant

- Consider estimating $\mu_{j\cdot}$ and $\mu_{\cdot k}$
- Unbiased estimators are $\hat{\mu}_{j\cdot} = y_{j\cdot\cdot}$ and $\hat{\mu}_{\cdot k} = y_{\cdot k\cdot}$ with variances $\text{var}(y_{j\cdot\cdot}) = \sigma^2/(qr)$, $\text{var}(y_{\cdot k\cdot}) = \sigma^2/(pr)$, and estimated standard errors

$$s(y_{j\cdot\cdot}) = s_R/\sqrt{qr}, \quad s(y_{\cdot k\cdot}) = s_R/\sqrt{pr}$$

- $100(1-\alpha)$% confidence interval for $\mu_{j\cdot}$ is $y_{j\cdot\cdot} \pm t_{\nu_R;\alpha/2} \times s(y_{j\cdot\cdot})$
- $100(1-\alpha)$% confidence interval for $L = \sum_j c_j \mu_{j\cdot}$ is

$$\hat{L} \pm t_{\nu_R;\alpha/2} \times s_R \sqrt{\sum_j c_j^2/(qr)}$$

Contrasts of factor level means

- Let $L = \sum_j c_j \mu_j$. with $\sum_j c_j = 0$
- $\hat{L} = \sum_j c_j y_{j\cdot}$. with $s(\hat{L}) = s_R \sqrt{(qr)^{-1} \sum_j c_j^2}$
- $100(1-\alpha)$ intervals have the form $\hat{L} \pm T s(\hat{L})$
- **Tukey** multiplier (for all pairwise comparisons)

$$T = q(p, \nu_R, \alpha)/\sqrt{2}$$

- **Scheffé** multiplier (for all contrasts)

$$T = \sqrt{(p-1) F_{(p-1),\nu_R;\alpha}}$$

- **Bonferroni** multiplier (for any g comparisons)

$$T = t_{\nu_R;\alpha/(2g)}$$

Interval estimation
when interactions are significant

- Analysis equivalent to oneway ANOVA with pq levels
- If pairs of treatment means μ_{jk} are to be compared, either Tukey or Bonferroni may be used
- Scheffé method is directly applicable to contrasts of the form $L = \sum_j \sum_k c_{jk}\mu_{jk}$

Randomized block design

- One treatment factor with t levels and one block factor with b levels
- Treatments randomized within blocks
- No replicate observations
- Example: manufacture of penicillin (Box et al., 2005, p. 146)

Corn steep liquor	Treatment A	B	C	D	Mean
Block 1	89	88	97	94	92
Block 2	84	77	92	79	83
Block 3	81	87	87	85	85
Block 4	87	92	89	84	88
Block 5	79	81	80	88	82
Mean	84	85	89	86	86

Model: (no block-treatment interaction)

$$y_{ij} = \mu + b_i + t_j + \epsilon_{ij} \quad (i = 1, \ldots, b; j = 1, \ldots, t)$$

with ϵ_{ij} independent $N(0, \sigma^2)$ and $\sum_j t_j = 0$

Fixed block effects: $\sum_{i=1}^{b} b_i = 0$

Random block effects: $b_1, b_2, \ldots, b_b$ independent $N(0, \sigma_b^2)$

Data decomposition:

$$y_{ij} = y_{\cdot\cdot} + (y_{i\cdot} - y_{\cdot\cdot}) + (y_{\cdot j} - y_{\cdot\cdot}) + (y_{ij} - y_{i\cdot} - y_{\cdot j} + y_{\cdot\cdot})$$

ANOVA decomposition:

$$\sum_i \sum_j (y_{ij} - y_{\cdot\cdot})^2 = t \sum_i (y_{i\cdot} - y_{\cdot\cdot})^2 + b \sum_j (y_{\cdot j} - y_{\cdot\cdot})^2 + \sum_i \sum_j (y_{ij} - y_{i\cdot} - y_{\cdot j} + y_{\cdot\cdot})^2$$

ANOVA Table for Randomized Block Design

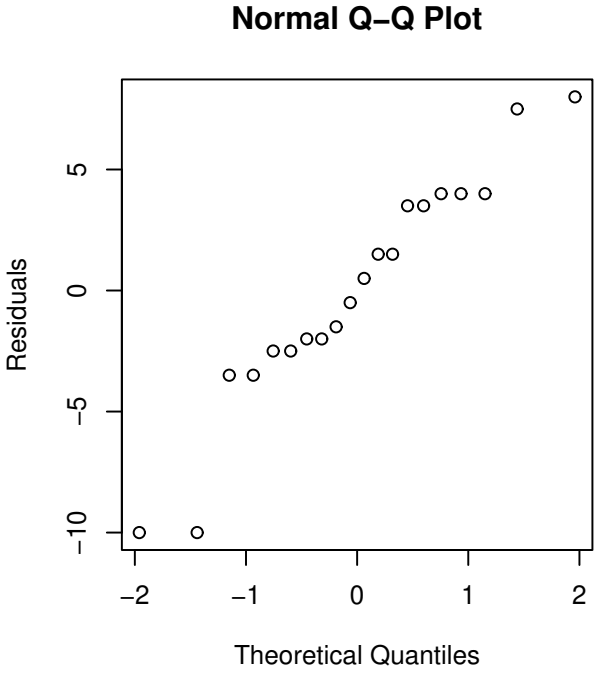
Source	SS	df	E(MS)
	Fixed block effects		
Blocks	$t\sum_i(y_{i\cdot}-y_{\cdot\cdot})^2$	$b-1$	$\sigma^2+t(b-1)^{-1}\sum_i b_i^2$
Treatments	$b\sum_j(y_{\cdot j}-y_{\cdot\cdot})^2$	$t-1$	$\sigma^2+b(t-1)^{-1}\sum_j t_j^2$
Error	$\sum_{ij}(y_{ij}-y_{i\cdot}-y_{\cdot j}+y_{\cdot\cdot})^2$	$(b-1)(t-1)$	σ^2
Total	$\sum_{ij}(y_{ij}-y_{\cdot\cdot})^2$	$bt-1$	
	Random block effects		
Blocks	$t\sum_i(y_{i\cdot}-y_{\cdot\cdot})^2$	$b-1$	$\sigma^2+t\sigma_b^2$
Treatments	$b\sum_j(y_{\cdot j}-y_{\cdot\cdot})^2$	$t-1$	$\sigma^2+b(t-1)^{-1}\sum_j t_j^2$
Error	$\sum_{ij}(y_{ij}-y_{i\cdot}-y_{\cdot j}+y_{\cdot\cdot})^2$	$(b-1)(t-1)$	σ^2
Total	$\sum_{ij}(y_{ij}-y_{\cdot\cdot})^2$	$bt-1$	

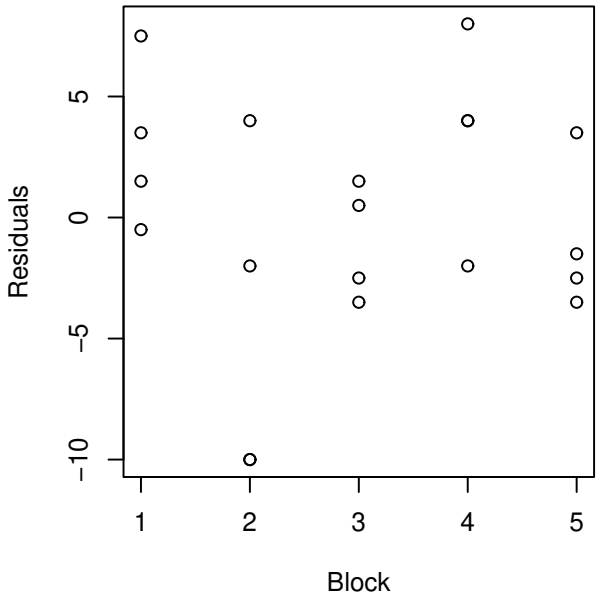
Implications of additive model

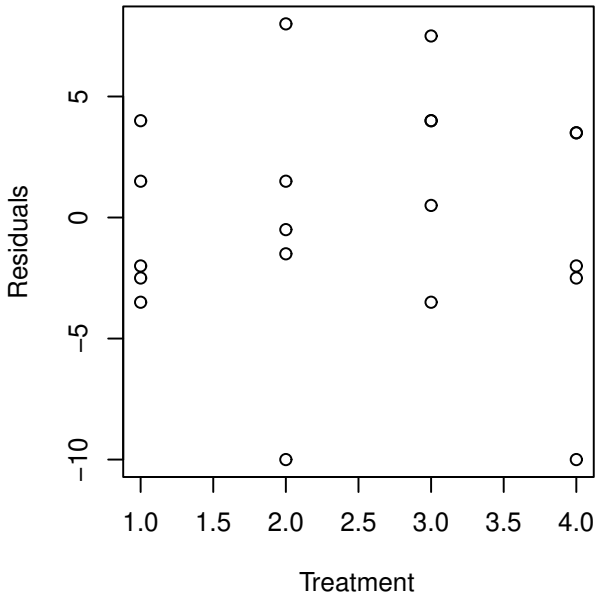
- Model implies that treatment and block effects are *additive*
- If block i leads to an increase of $b_i = 5$ units in the response, and treatment j to an increase of $t_j = 2$ units, then the combined effect is $5 + 2 = 7$
- The increase of $b_i = 5$ units in block i is the same for all treatments, and similarly the increase of $t_j = 2$ for treatment j is the same for all blocks
- If block and treatment effects were not additive, an *interaction* is said to occur between blocks and treatments Examples:
  1. Specific conditions in one block may make a certain treatment completely ineffective
  2. An additive model may still apply, but not in the scale used to measure the yields
- Interactions are of two types: *transformable interactions* and *nontransformable interactions*

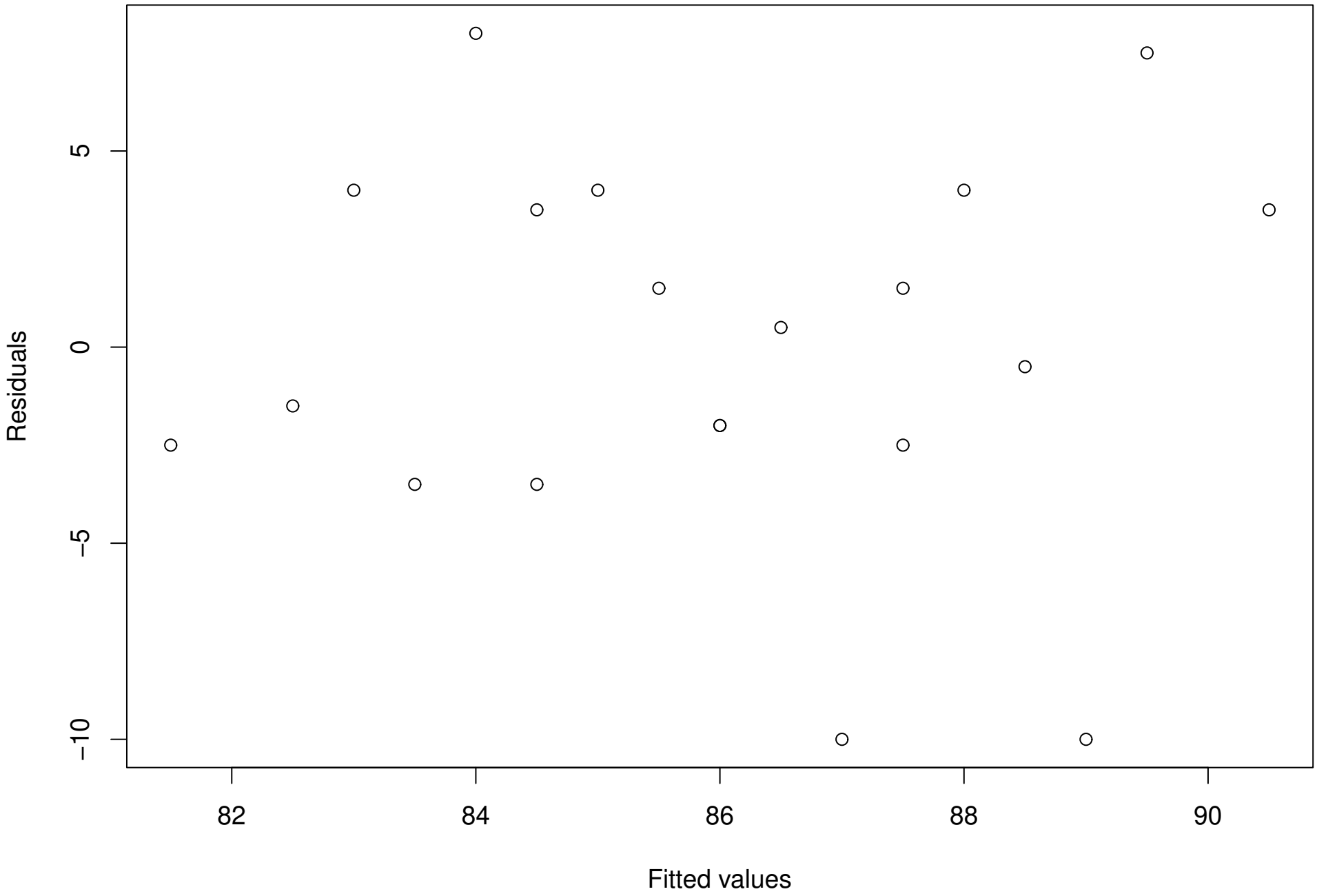
Diagnostic plots

- Plot dot diagram of all residuals to check for normality and outliers
- Plot dot diagrams of residuals separately for each block and each treatment to check for outliers and constancy of variance
- Plot residuals versus fitted values ($\hat{y}_{ij}$) to check for curvilinear relationship or funnel shape appearance (or both) which would suggest *nonadditivity* between blocks and treatments and which might be eliminated by a suitable transformation of the data


Normal Q–Q Plot
Residuals
Theoretical Quantiles
−10
−5
0
5
−2
−1
0
1
2


Residuals
Block
−10
−5
0
5
1
2
3
4
5


Residuals
Treatment
−10
−5
0
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Residuals
5
0
−5
−10
82
84
86
88
90
Fitted values

R code

```
y <- c(89, 88, 97, 94, 84, 77, 92, 79, 81, 87, 87, 85, 87, 92, 89,
       84, 79, 81, 80, 88)
b <- 5 ## number of blocks
k <- 4 ## number of treatments
block <- rep(1:b,rep(k,b))
treat <- rep(1:k,b))
model <- lm(y ~ as.factor(block)+as.factor(treat))
res <- model$residuals

qqnorm(res,ylab="Residuals")
plot(res ~ block,xlab="Block",ylab="Residuals")
plot(res ~ treat,xlab="Treatment",ylab="Residuals")
plot(res ~ model$fit,xlab="Fitted values",ylab="Residuals")
```

Test for transformable nonadditivity

1. Compute residuals $y_{ij} - \hat{y}_{ij}$ from additive model
2. Define $q_{ij} = \hat{y}_{ij}^2$
3. Treat q_{ij} as new data and fit additive model to them
4. Compute residuals $q_{ij} - \hat{q}_{ij}$
5. Compute

$$\begin{aligned} P &= \sum_{ij}(y_{ij} - \hat{y}_{ij})(q_{ij} - \hat{q}_{ij}) \\ Q &= \sum_{ij}(q_{ij} - \hat{q}_{ij})^2 \\ S_{na} &= P^2/Q \\ F &= \frac{S_{na}}{(S_R - S_{na})/\{(b-1)(t-1)-1\}} \end{aligned}$$

6. Construct Table 2 below

Table 2: ANOVA table for Tukey's test of transformable nonadditivity

Source	SS	d.f.	Ratio
Trans. nonadditivity	S_{na}	1	F
Remainder	$S_R - S_{na}$	$(b-1)(t-1)-1$	
Residual	S_R	$(b-1)(t-1)$	

7. Test seeks a correlation between the two sets of residuals
8. If F is too large (compared to F distribution with 1 and $(b-1)(t-1)-1$ degrees of freedom), a need for transformation of the data is suggested
9. One way to choose a suitable transformation is to do the analysis for a number of transformations and select one that yields a nonsignificant F value

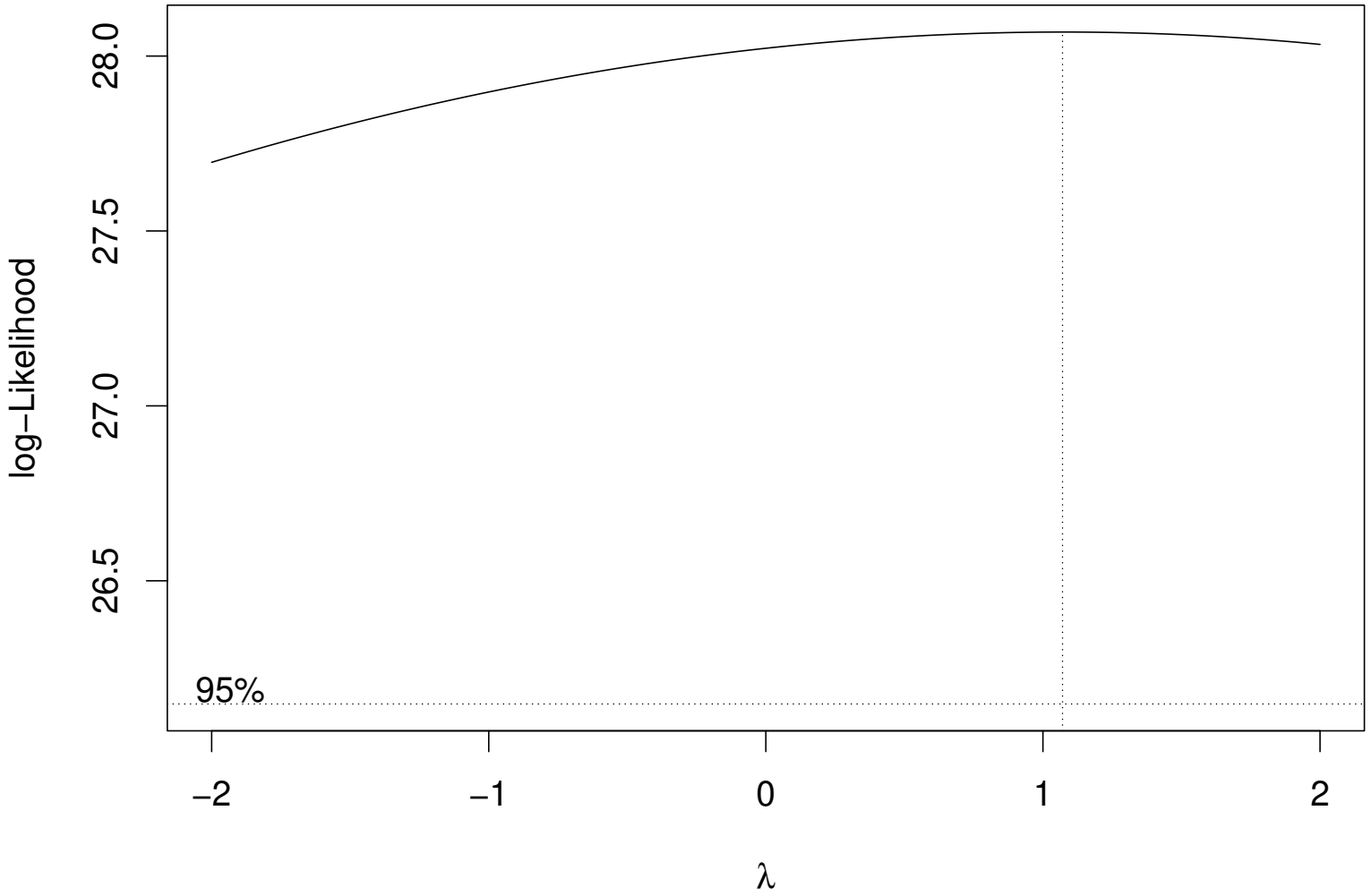
Tukey's test for penicillin data

Source	SS	d.f.	F
Trans. nonadditivity	S_{na} = 4.01	1	0.10
Remainder	$S_R - S_{na}$ = 440.99	$(b-1)(t-1)-1$ = 11	
Residual	S_R = 445.00	$(b-1)(t-1)$ = 12	

```
q <- model$fit^2
model2  <- lm(q ~ block+treat)
qres <- model2$res
P <- sum(res*qres)
Q <- sum(qres^2)
Sna <- P*P/Q
SR <- sum(res^2)
Fstat <- Sna/((SR-Sna)/((b-1)*(k-1)-1))
```

Box-Cox transformations

```
library(MASS)
bc <- boxcox(model)
```



Advantages of two-level factorial designs

1. They require relatively few runs per factor studied
2. Although they are unable to explore fully a wide region in the factor space, they can indicate major trends and so determine a promising direction for further experimentation
3. They can be augmented to form composite designs when more thorough local exploration is needed
4. The designs may be used as building blocks so that degree of complexity of the final design can match sophistication of the problem
5. Estimates of factor effects are more precise
6. Interactions can be estimated
7. Data analysis is simple
8. They form the basis for two-level fractional factorial designs

Factorial versus OFAT experiments (OFAT stands for "one-factor-at-a-time")

Two factors in three or four runs with 48 wafers

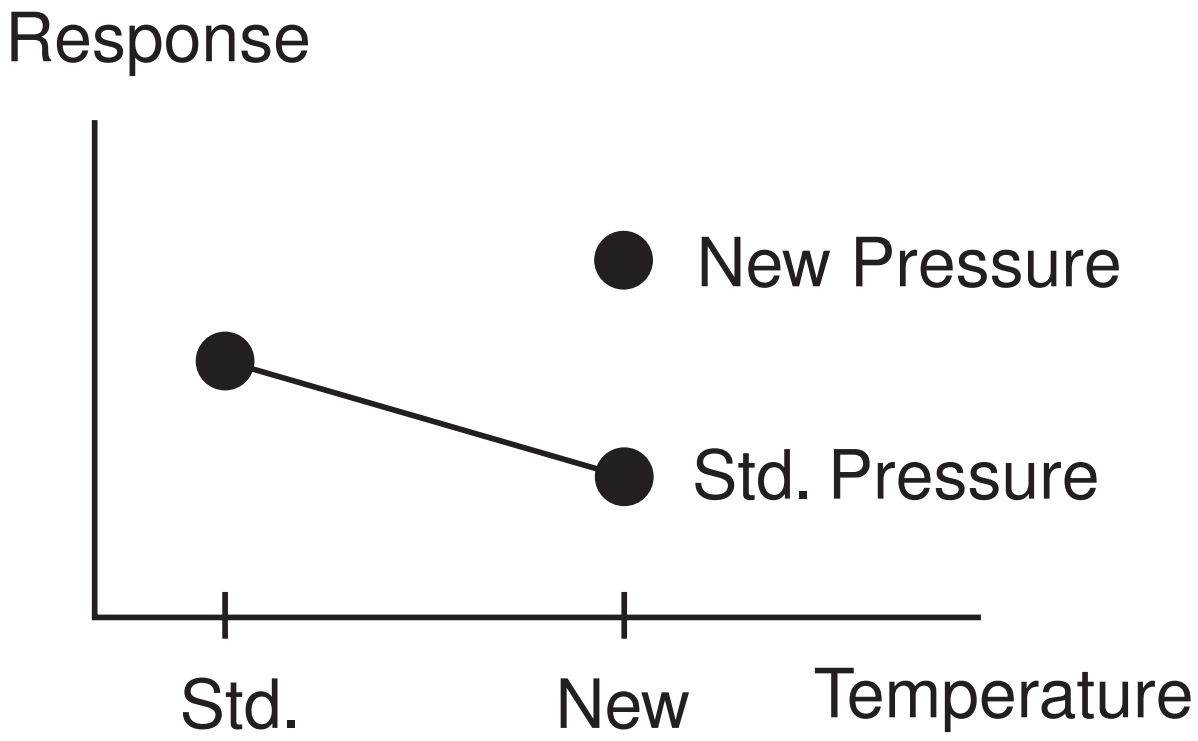
An engineer planned an experiment to compare pressure and temperature for a standard gas anneal process and a new gas anneal process using three experimental runs:

1. Standard pressure and standard temperature
2. Standard pressure and new temperature
3. New pressure and new temperature

The engineer planned to use one lot of 48 wafers, with 16 wafers for each run

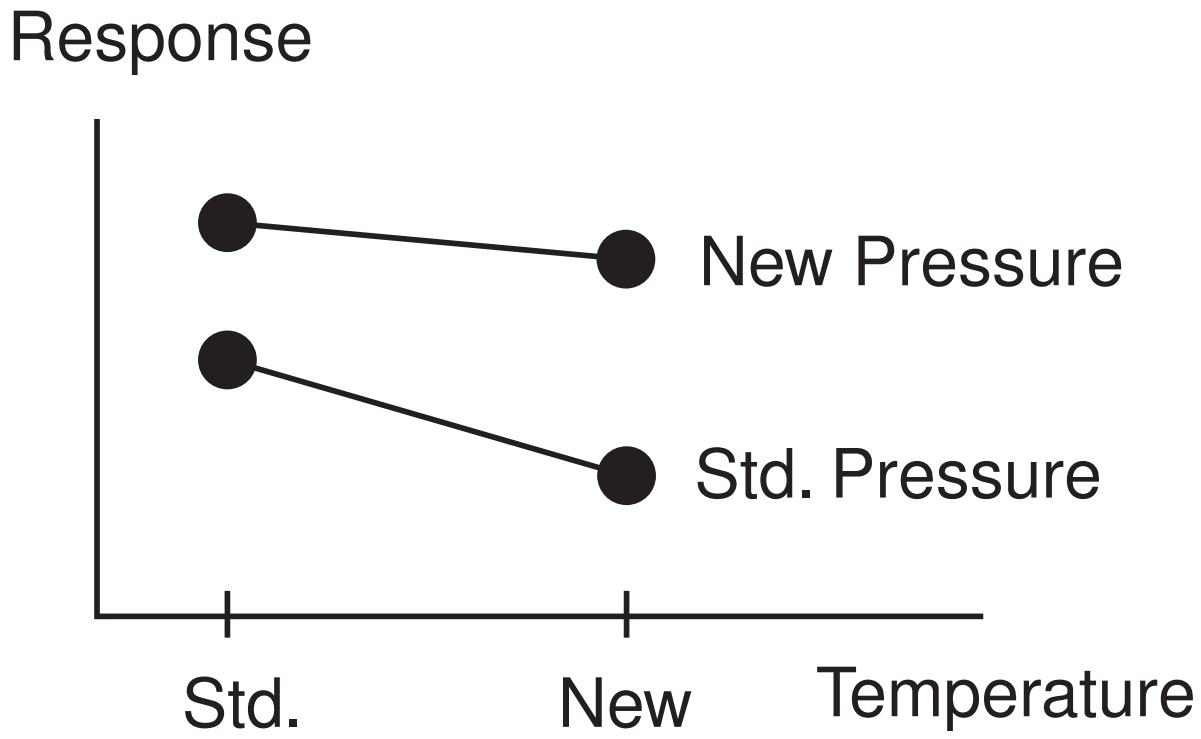
3 runs with 16 wafers each (OFAT)

	Temperature	
Pressure	Standard	New
Standard	16 wafers	16 wafers
New		16 wafers



2^2 design (4 runs with 12 wafers each)

	Temperature	
Pressure	Standard	New
Standard	12 wafers	12 wafers
New	12 wafers	12 wafers

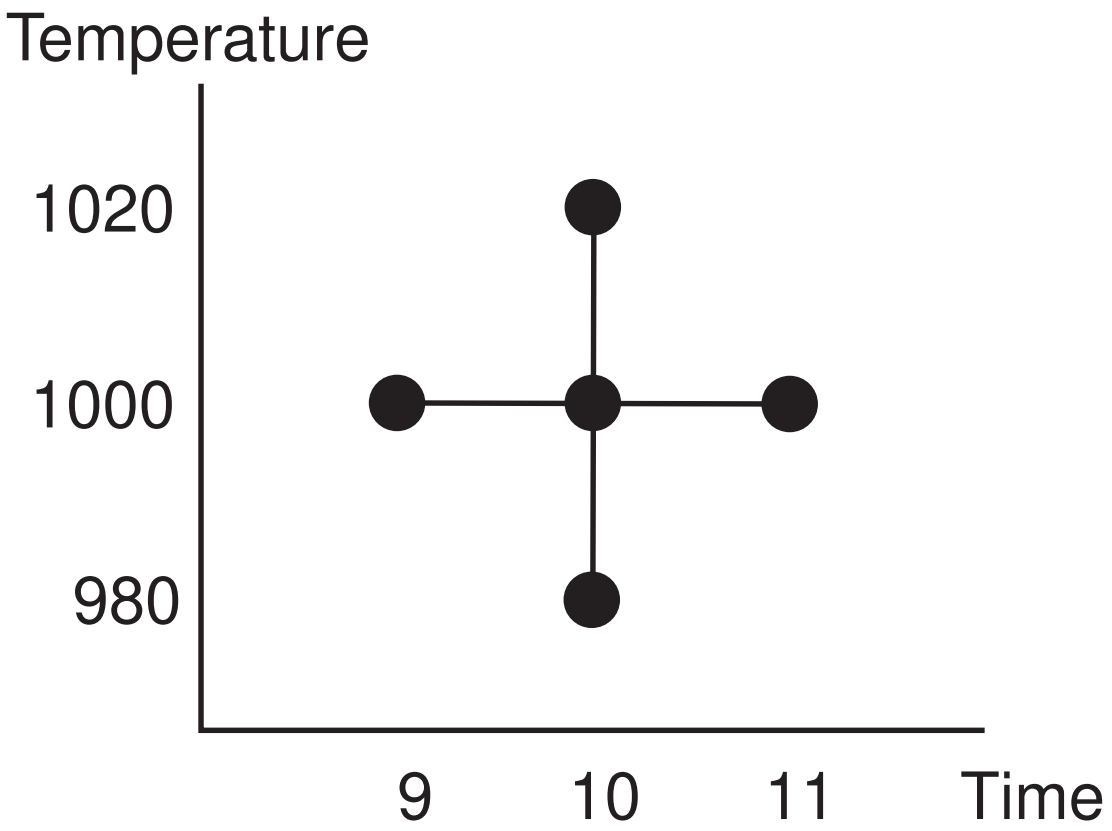


Why 2^2 design is better than OFAT

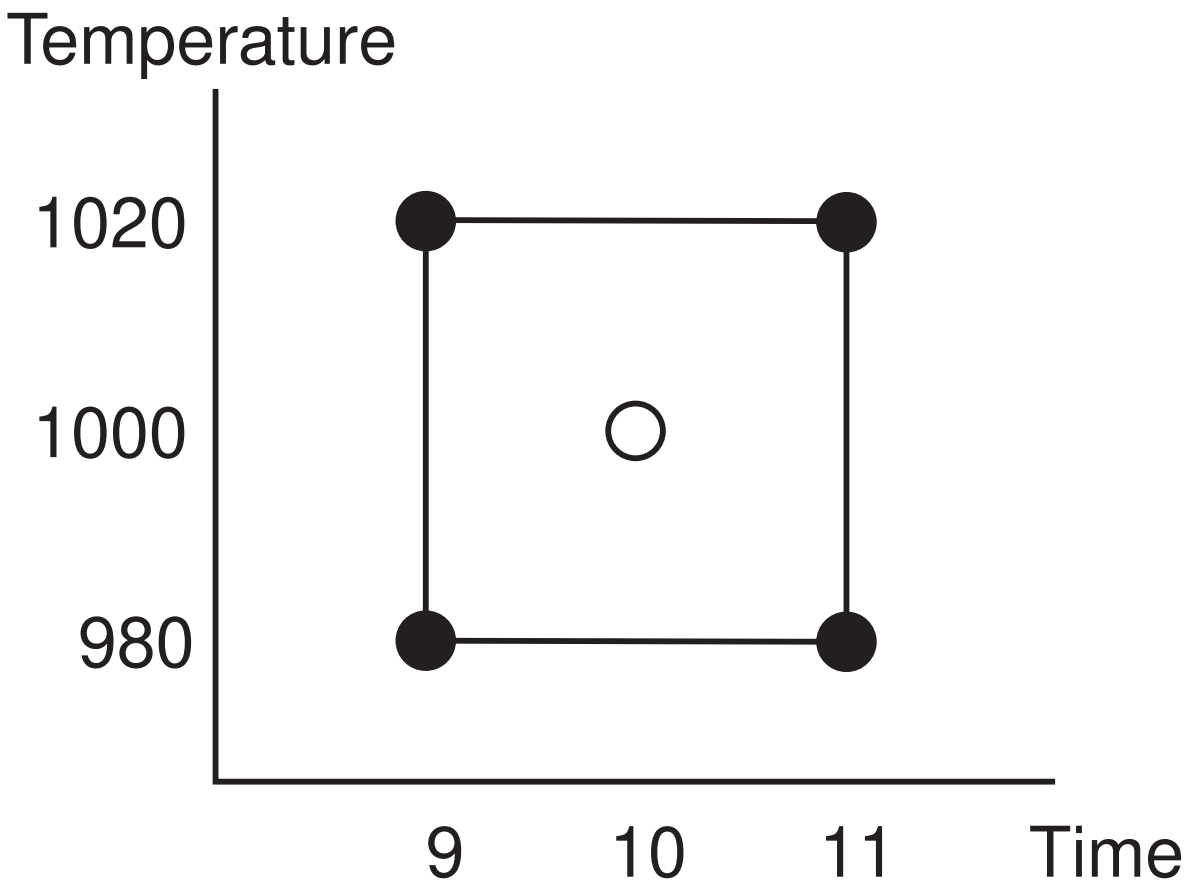
1. Estimates of the factor effects are more precise (all 48 wafers are used, while only 32 wafers in the OFAT design)
   (a) Variance of each effect is $\sigma^2/12$ for the 2^2 design
   (b) Variance of each factor effect for the OFAT is $\sigma^2/8$ (50% larger)
2. Interaction effect cannot be estimated by OFAT design
3. 2^2 design provides data over a broader factor space

Two factors in 4 or 5 runs unreplicated

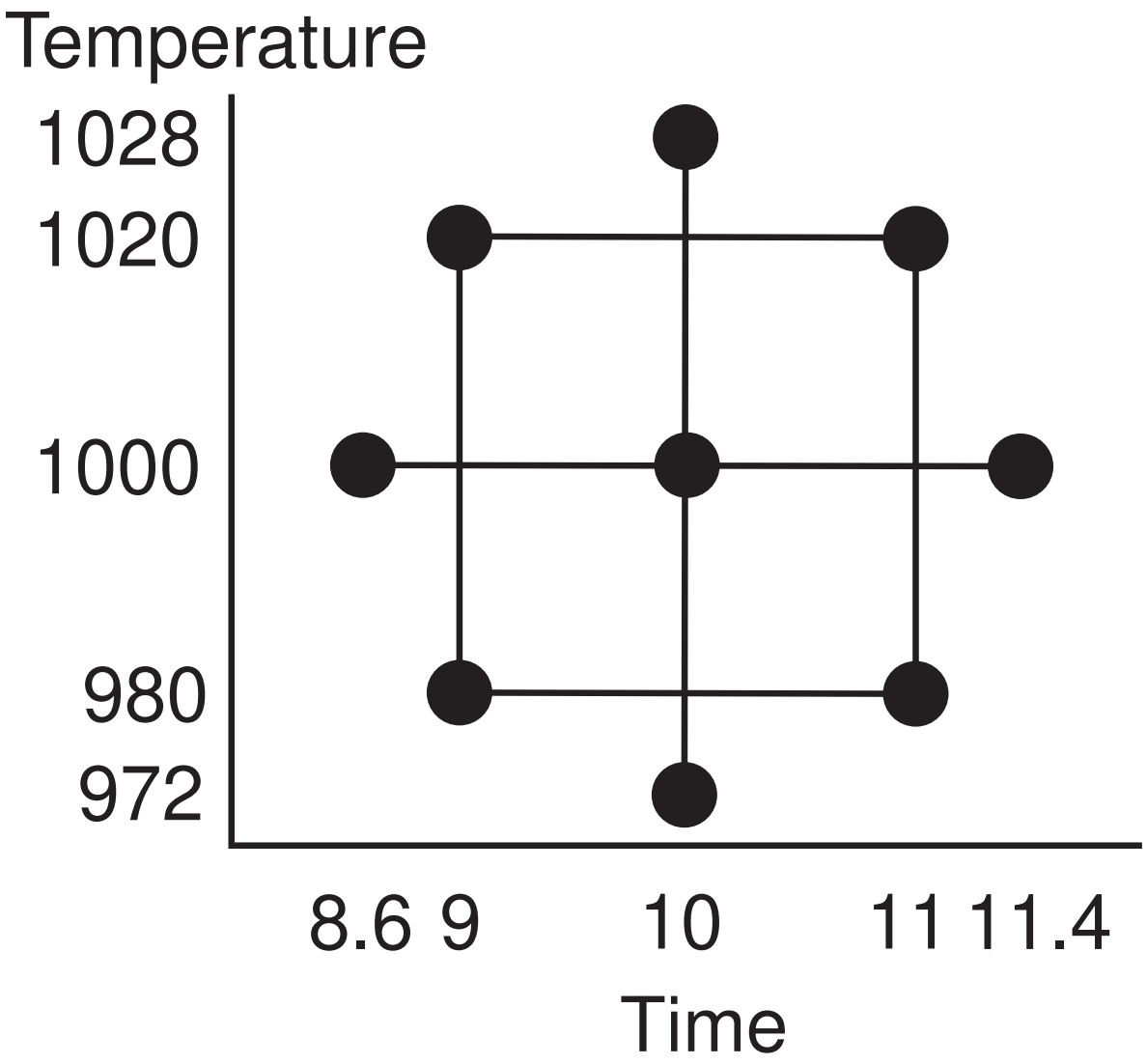
Two engineers planned an experiment for a rapid thermal anneal process. They wanted to study the sensitivity of the response sheet resistance to two factors—time and temperature. They used an OFAT with five runs.



2^2 design with or without center point



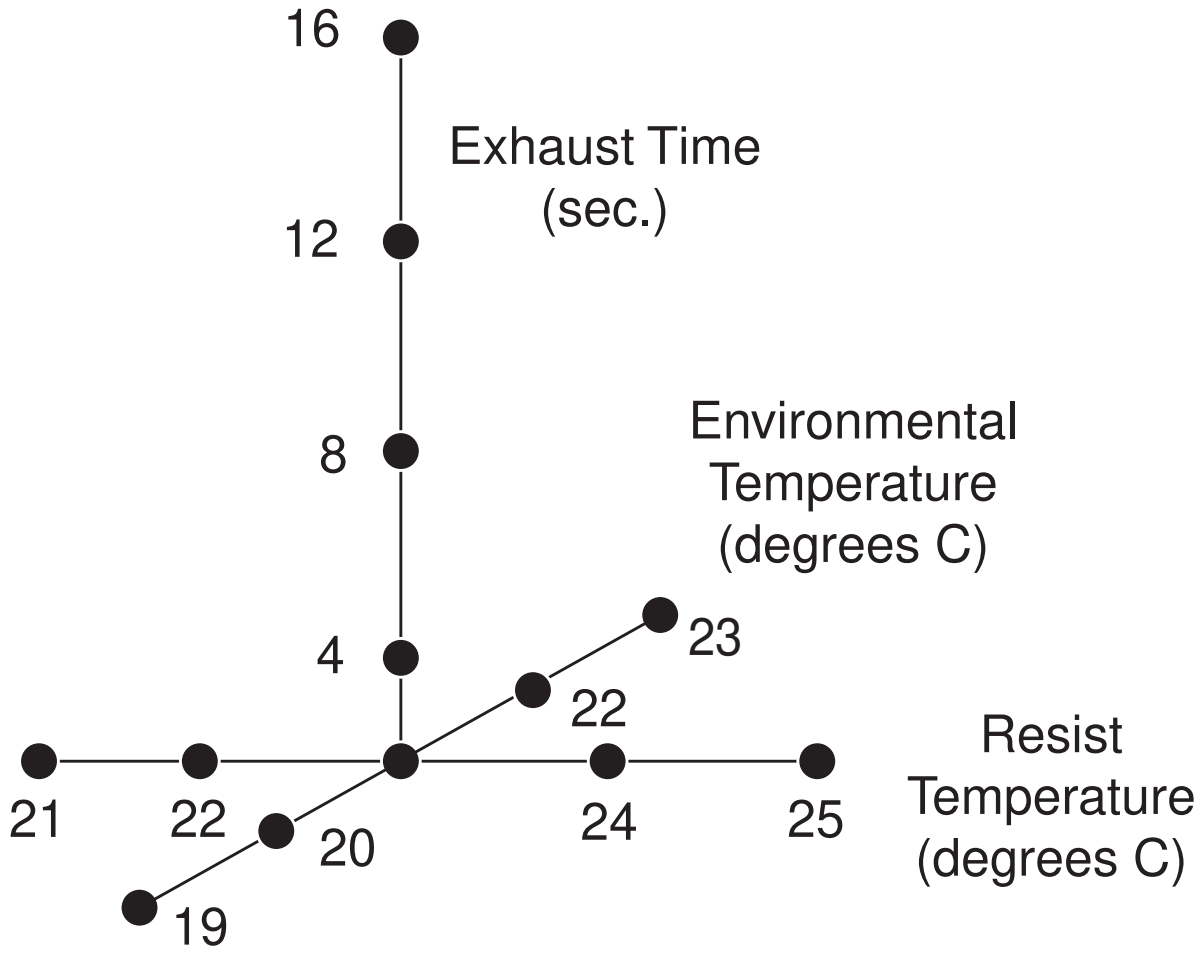
Central composite design



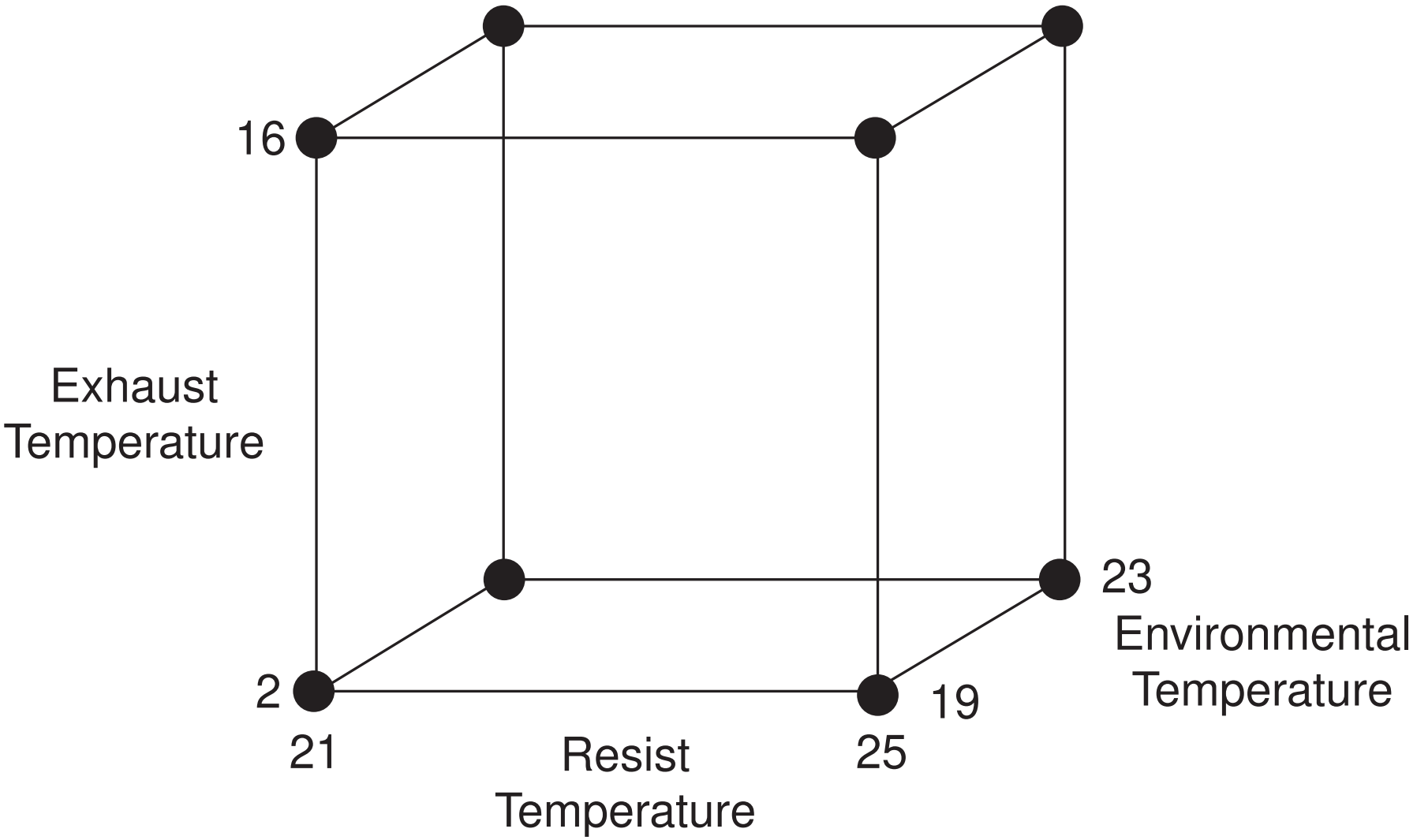
Advantages of central composite design

1. Estimates of factor effects are more precise
2. Interactions can be estimated
3. Curvature can be estimated in entire space
4. Optimization is possible over entire space
5. If center point is replicated 4 times, the design is rotatable, i.e., equal precision in estimation at all points equidistant from the center

Three-factor OFAT in 15 runs



2^3 experiment in 8 runs

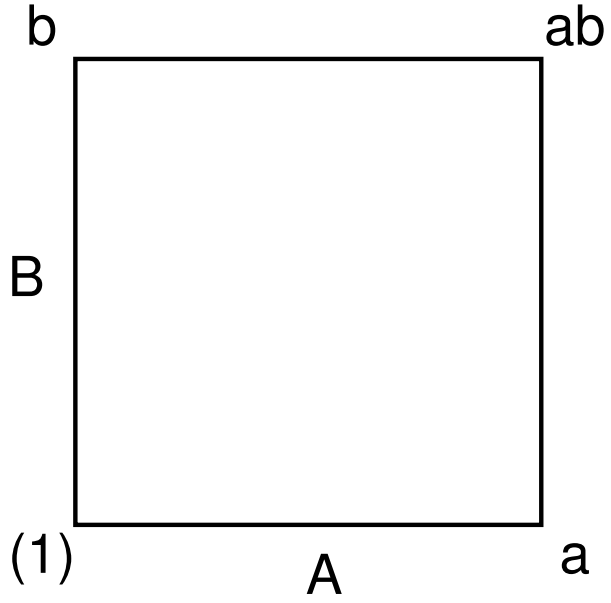


Advantages of the 2^3 design

1. Requires less resources—8 runs instead of 15
2. Estimates of factor effects are more accurate—all 8 runs used in each estimate versus 5 runs for OFAT experiment
3. Interaction effects cannot be estimated in OFAT experiment
4. 2^3 design yields information over a larger region of the factor space

Equivalent notations for 2^2 design

Run	A	B	Symbol
1	–	–	(1)
2	+	–	a
3	–	+	b
4	+	+	ab



- Two factors A and B, each at two levels
- 4 sets of experimental conditions: "low" level is indicated "–" and "high" level by "+"
- Use the notation $(1), a, b, ab$ to denote the mean responses ($\bar{y}$) at those treatment combinations

Model: $y_{ijk} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$

- Use sum-to-zero constraints $\alpha_1 + \alpha_2 = \beta_1 + \beta_2 = 0$ and

$$(\alpha\beta)_{11} + (\alpha\beta)_{12} = (\alpha\beta)_{21} + (\alpha\beta)_{22} = (\alpha\beta)_{11} + (\alpha\beta)_{21} = 0$$

- Define $\alpha_2 = \alpha$, $\beta_2 = \beta$ and $(\alpha\beta)_{22} = (\alpha\beta)$
- Then $\alpha_1 = -\alpha$, $\beta_1 = -\beta$, $(\alpha\beta)_{21} = -(\alpha\beta)$, $(\alpha\beta)_{11} = (\alpha\beta)$, $(\alpha\beta)_{12} = -(\alpha\beta)$

i	j	A	B		With subscripts	Without subscripts
1	1	–	–	(1)	$\eta + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\eta - \alpha - \beta + (\alpha\beta)$
2	1	+	–	a	$\eta + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$	$\eta + \alpha - \beta - (\alpha\beta)$
1	2	–	+	b	$\eta + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$	$\eta - \alpha + \beta - (\alpha\beta)$
2	2	+	+	ab	$\eta + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$	$\eta + \alpha + \beta + (\alpha\beta)$

Alternative "regression-type" model

- Let $x_1 = -1$ denote to level $-$ of A and $x_1 = 1$ to level $+$ of A
- Let $x_2 = -1$ denote to level $-$ of B and $x_2 = 1$ to level $+$ of B
- The model can be written as

$$y = \eta + \alpha x_1 + \beta x_2 + (\alpha\beta) x_1 x_2 + \epsilon$$

Estimating equations

$$\begin{aligned}(1) &= \eta - \alpha - \beta + (\alpha\beta) \\ a &= \eta + \alpha - \beta - (\alpha\beta) \\ b &= \eta - \alpha + \beta - (\alpha\beta) \\ ab &= \eta + \alpha + \beta + (\alpha\beta)\end{aligned}$$

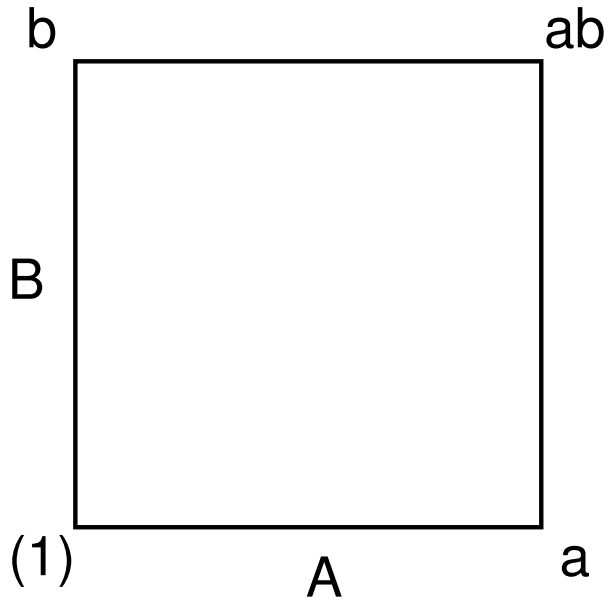
Estimates

$$\begin{aligned}\hat{\eta} &= [ab + a + b + (1)]/4 = \text{grand mean} \\ \hat{\alpha} &= [ab + a - b - (1)]/4 \\ \hat{\beta} &= [ab - a + b - (1)]/4 \\ \widehat{(\alpha\beta)} &= [ab - a - b + (1)]/4\end{aligned}$$

Definition of A main effect in 2^2 design

Main effect of A is defined as

$$\begin{aligned} A &= (1/2)\{[ab-b]+[a-(1)]\} \\ &= [ab+a]/2-[b+(1)]/2 \\ &= [ab+a-b-(1)]/2 \\ &= 2\hat{\alpha} \end{aligned}$$

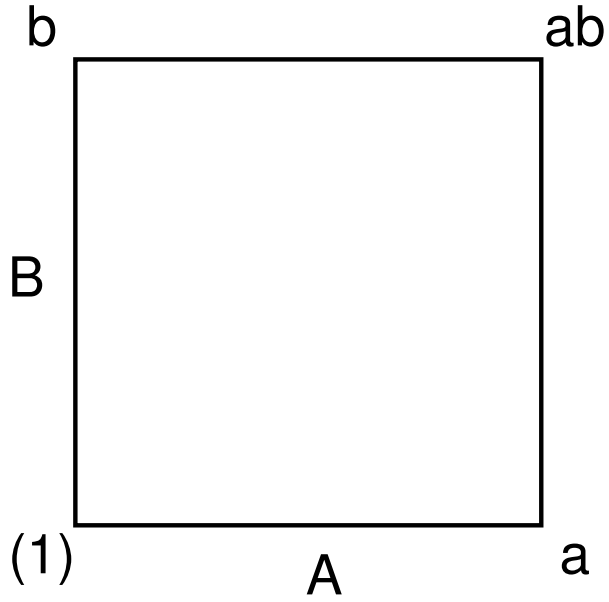


A = (1/2) $\times$ {(A effect at high B) + (A effect at low B)}

Definition of AB interaction in 2^2 design

Interaction effect is defined as

$$\begin{aligned} AB &= (1/2)\{[ab-b]-[a-(1)]\} \\ &= (1/2)\{[ab-a]-[b-(1)]\} \\ &= [ab-a-b+(1)]/2 \\ &= 2\widehat{(\alpha\beta)} \end{aligned}$$



$$\begin{aligned} AB &= (1/2) \times \{\text{(A effect at high B)} - \text{(A effect at low B)}\} \\ &= (1/2) \times \{\text{(B effect at high A)} - \text{(B effect at low A)}\} \end{aligned}$$

Important: $AB \neq A \times B$

Model expressed in terms of main and interaction effects

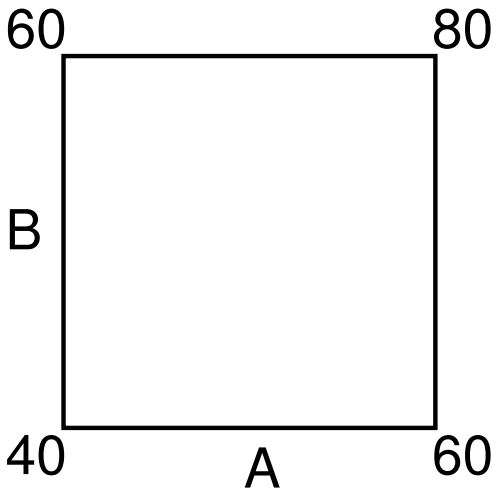
$$y = \eta + \alpha x_1 + \beta x_2 + (\alpha\beta) x_1 x_2 + \epsilon$$

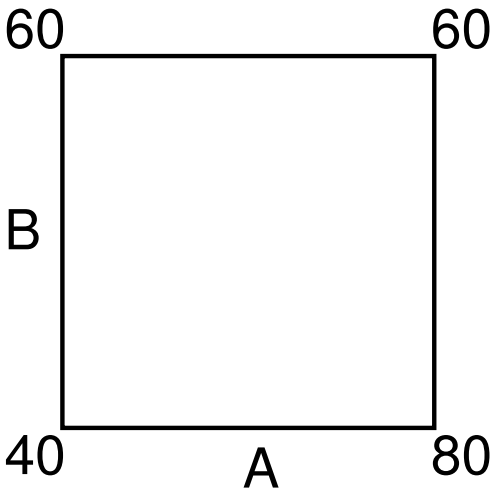
is equivalent to

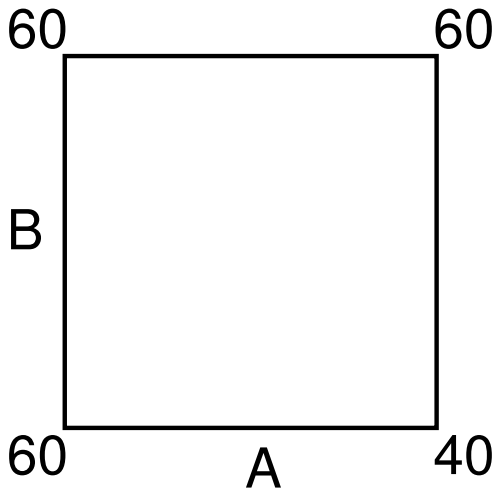
$$y = \eta + (A/2) x_1 + (B/2) x_2 + (AB/2) x_1 x_2 + \epsilon$$

Example

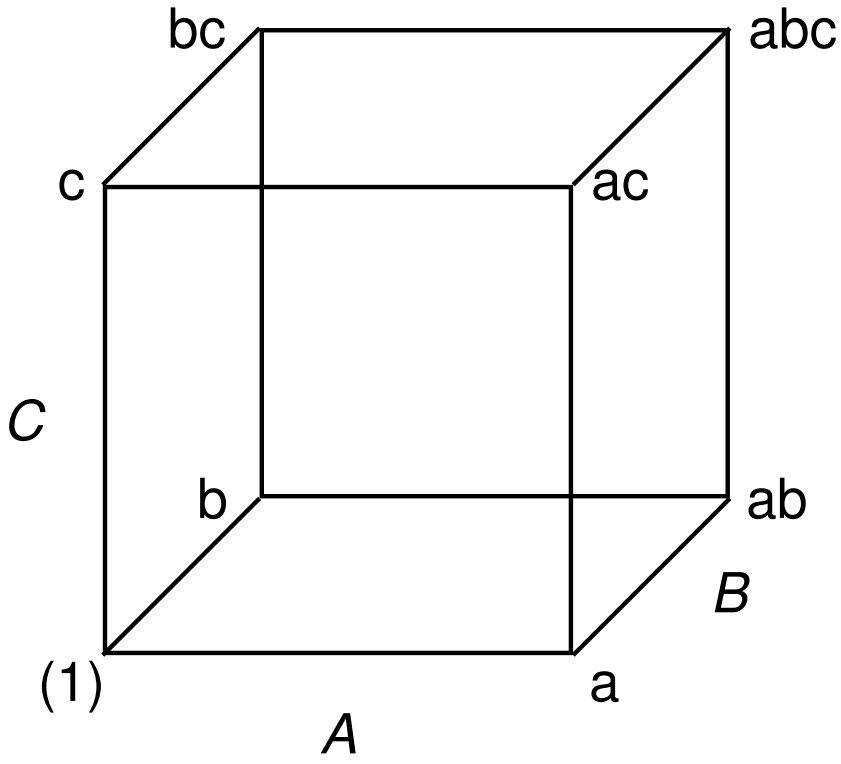
(i)				(ii)				(iii)			
b	60	ab	80	b	60	ab	60	b	60	ab	60
(1)	40	a	60	(1)	40	a	80	(1)	60	a	40



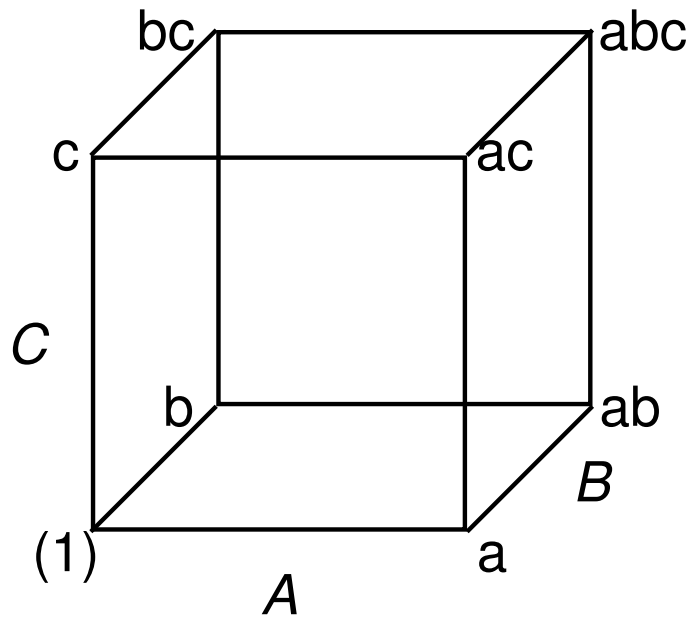




How are main and interaction effects in 2^3 defined?

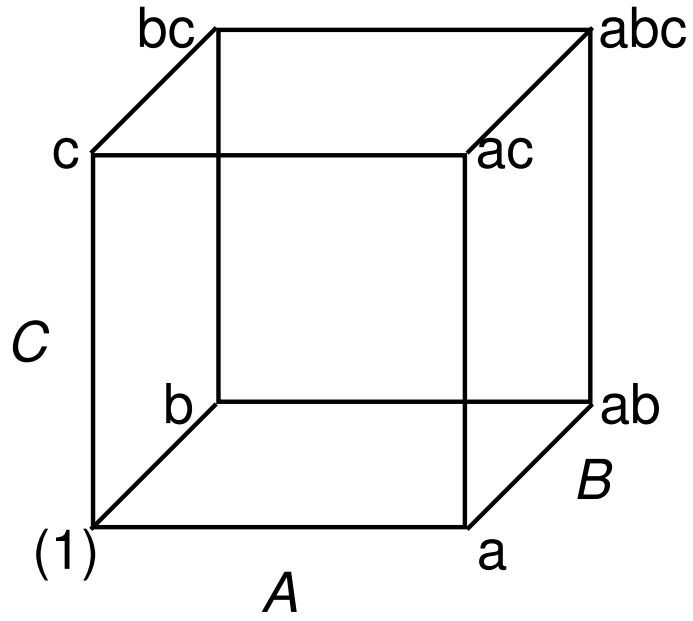


Main effect of A in 2^3 design



$$
\begin{aligned}
A &= (a + ab + ac + abc)/4 - ((1) + b + c + bc)/4 \\
&= \text{mean of high A} - \text{mean of low A} \\
&= (1/4)\{(a - (1)) + (ab - b) + (ac - c) + (abc - bc)\} \\
&= \text{mean of 4 univariate A effects} \\
&= (1/2)[\{abc + ac - bc - c\}/2 + \{ab + a - b - (1)\}/2] \\
&= (1/2)\{\text{A effect in A-B design at high C} + \text{A effect in A-B design at low C}\}
\end{aligned}
$$

AB and ABC interactions in 2^3 design



- The 2^3 consists of two 2^2 designs: one at high C and one at low C
- AB interaction at low C is $AB_1 = (ab + (1) - a - b)/2$
- AB interaction at high C is $AB_2 = (abc + c - ac - bc)/2$
- AB and ABC effects for 2^3 are defined as

$$\begin{aligned} AB &= (1/2)(AB_2 + AB_1) = (1/4)(abc + c - ac - bc + ab + (1) - a - b) \\ ABC &= (1/2)(AB_2 - AB_1) = (1/4)(abc + c - ac - bc - ab - (1) + a + b) \end{aligned}$$

Alternative "regression-type" model

Two ways to write the model:

$$\begin{aligned} E(y_{ijk}) &= \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \\ E(y) &= \eta + \alpha x_1 + \beta x_2 + \gamma x_3 + (\alpha\beta)x_1x_2 + (\alpha\gamma)x_1x_3 + (\beta\gamma)x_2x_3 \\ &\quad + (\alpha\beta\gamma)x_1x_2x_3 \end{aligned}$$

LSEs are:

$$\begin{aligned} \hat{\eta} &= = [abc + ab + ac + bc + a + b + c + (1)]/8 \\ A &= 2\hat{\alpha} = [abc + ab + ac + a - bc - b - c - (1)]/4 \\ AB &= 2\widehat{(\alpha\beta)} = [abc + ab - ac - bc - a - b + c + (1)]/4 \\ ABC &= 2\widehat{(\alpha\beta\gamma)} = [abc - bc - ac - ab + a + b + c - (1)]/4, \text{ etc.} \end{aligned}$$

Therefore model can be written as:

$$\begin{aligned} E(y) &= \eta + (A/2)x_1 + (B/2)x_2 + (C/2)x_3 + (AB/2)x_1x_2 + (AC/2)x_1x_3 \\ &\quad + (BC/2)x_2x_3 + (ABC/2)x_1x_2x_3 \end{aligned}$$

Symbolic expressions

The LSEs

$$\begin{aligned}
\hat{\eta} &= [abc + ab + ac + bc + a + b + c + (1)]/8 \\
A &= [abc + ab + ac + a - bc - b - c - (1)]/4 \\
AB &= [abc + ab - ac - bc - a - b + c + (1)]/4 \\
ABC &= [abc - bc - ac - ab + a + b + c - (1)]/4
\end{aligned}$$

can be obtained **symbolically** by expanding the expressions

$$\begin{aligned}
\hat{\eta} &= (a+1)(b+1)(c+1)/8 \\
A &= (a-1)(b+1)(c+1)/4 \\
AB &= (a-1)(b-1)(c+1)/4 \\
ABC &= (a-1)(b-1)(c-1)/4
\end{aligned}$$

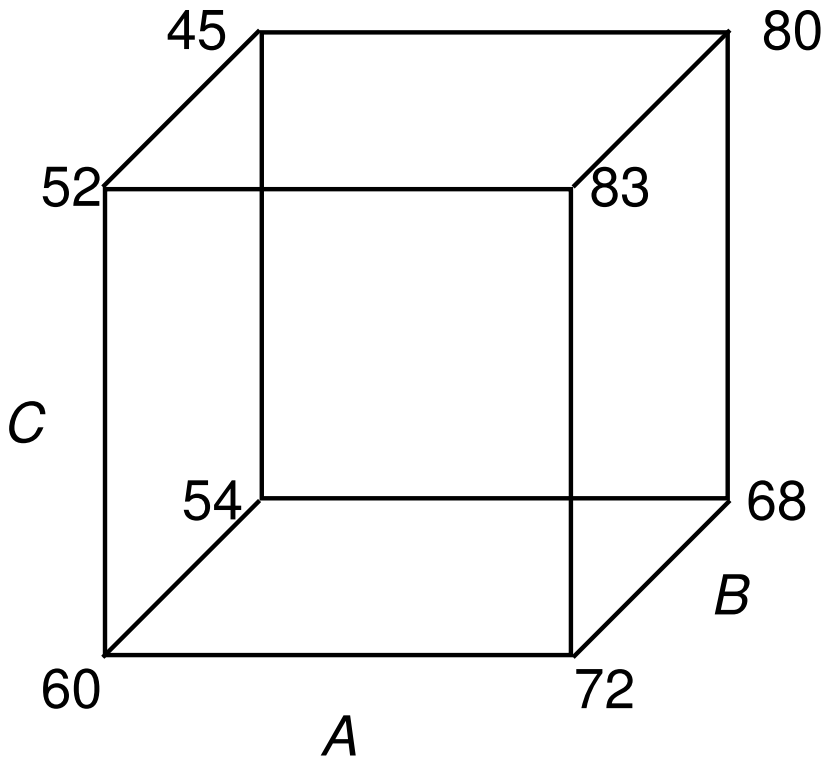
Properties

- **Variance** of any effect is var(effect) $= 4N^{-1}\sigma^2$, where N is total number of observations and $\sigma^2 = \text{var}(\epsilon)$
- Let s_i^2 be the estimate of σ^2 at the ith treatment combination ($i = 1, 2, \ldots, g$). Let ν_i be the degrees of freedom for s_i^2. Then the pooled estimate of σ^2 is:

$$s^2 = \frac{\nu_1 s_1^2 + \cdots + \nu_g s_g^2}{\nu_1 + \cdots + \nu_g}$$

- Confidence intervals for estimated effects may be obtained using the Bonferroni method.
- **Hidden replication property:** As far as the estimation of A is concerned, we have lost nothing in precision by adding B and C to the experiment because the variance of A in the experiment is the same as it would have been if we had not had B and C and had made instead only four runs at low A and four runs at high A

Example of a 2^3 experiment



Fast computation of effects: Table of contrast coefficients

1. Construct a table of signs as shown in following table
2. After the columns for A, B and C are written down as dictated by the actual runs, the interaction columns of signs are obtained by multiplying the corresponding columns
3. The effects are calculated by taking the dot products of the y column with each column and dividing by $N/2$

Analysis of unreplicated 2^k experiments

	A	B	C	AB	AC	BC	ABC	y
	-1	-1	-1	1	1	1	-1	60
	1	-1	-1	-1	-1	1	1	72
	-1	1	-1	-1	1	-1	1	54
	1	1	-1	1	-1	-1	-1	68
	-1	-1	1	1	-1	-1	1	52
	1	-1	1	-1	1	-1	-1	83
	-1	1	1	-1	-1	1	-1	45
	1	1	1	1	1	1	1	80
Dot product	92	-20	6.0	6.0	40	0	2.0	
Division by 4	23	-5	1.5	1.5	10	0	0.5	

Identification of significant effects from unreplicated 2^k experiments
Method I: Use interactions to estimate σ^2

- If no estimate of σ^2 is available, but some high-order interactions are assumed to be negligible, an estimate of the variance of an effect is given by the average of the squares of these high-order interactions
- The degrees of freedom associated with this estimate is equal to the number of high-order interactions used in the average

Illustration of Method I

Effects are A = 23, B = -5, C = 1.5, AB = 1.5, AC = 10, BC = 0, ABC = 0.5

1. If only 3fi (*ABC*) is assumed negligible,

$$\text{SD(effect)} = \sqrt{\text{Var(effect)}} = \sqrt{ABC^2} = 0.5$$

   with 1 df. Hence $t_{1,0.10/12} \times \text{SD(effect)} = 38.19 \times 0.5 = 19.1$ and simultaneous 90% Bonferroni intervals for A, B, C, AB, AC, and BC have the form "Effect $\pm$ 19.1" and only A significant

2. If 2fi's and 3fi are assumed negligible,

$$\begin{aligned} \text{Var(effect)} &= (AB^2 + AC^2 + BC^2 + ABC^2)/4 = 25.625 \\ \text{SD(effect)} &= \sqrt{25.625} = 5.06 \end{aligned}$$

   with 4 df. Hence 90% Bonferroni intervals for A, B and C have the form

$$\text{Effect} \pm t_{4,0.10/6} \times 5.06 = \text{Effect} \pm 3.19 \times 5.06 = 16.14$$

   and again only A is significant

Method II: Daniel's method

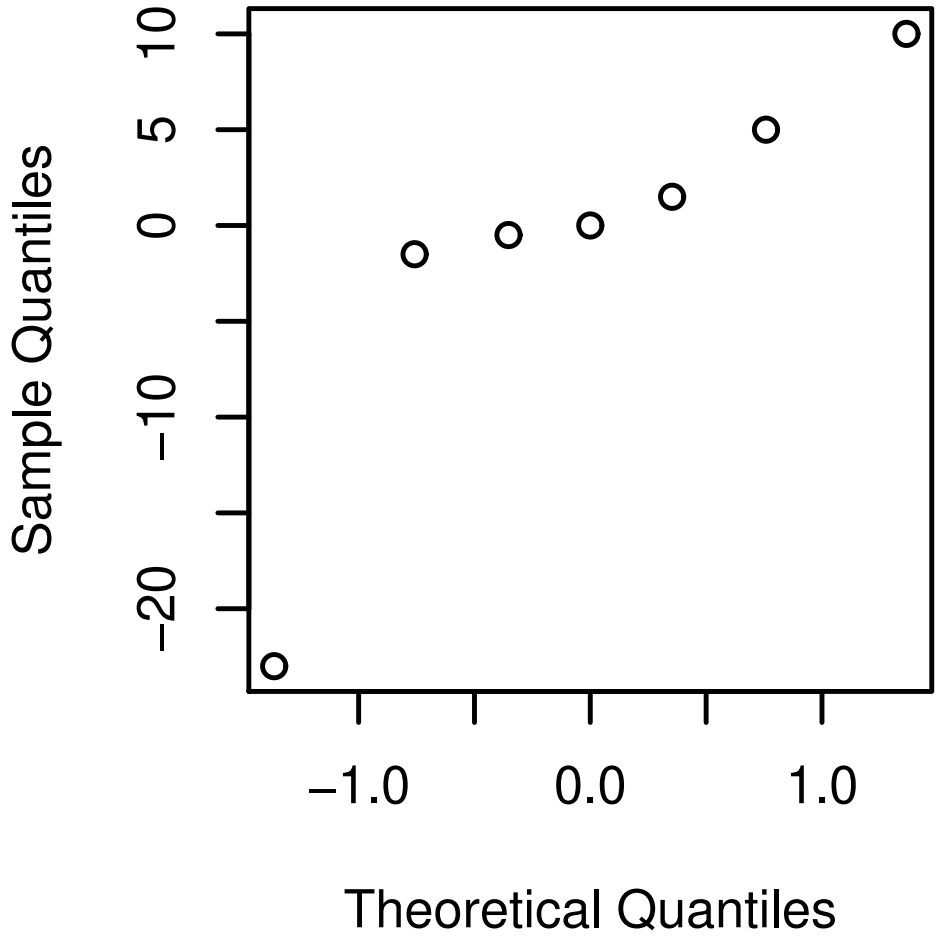
1. Make a normal qq-plot of the estimated effects
2. Fit a straight line to the middle portion of the data
3. Identify the outliers as significant effects

Motivation

1. Each effect $\hat{\theta}_i$ is distributed as $N(\theta_i, \tau^2)$
2. Because they are uncorrelated, they are also independent
3. If all the effects are not significant (i.e., all $\theta_i = 0$), then $\hat{\theta}_i$ are IID $N(0, \tau^2)$
4. Therefore a normal plot of the effects should be approximately straight

Illustration of Daniel's method

Effects are A = 23, B = -5, C = 1.5, AB = 1.5, AC = 10, BC = 0, ABC = 0.5

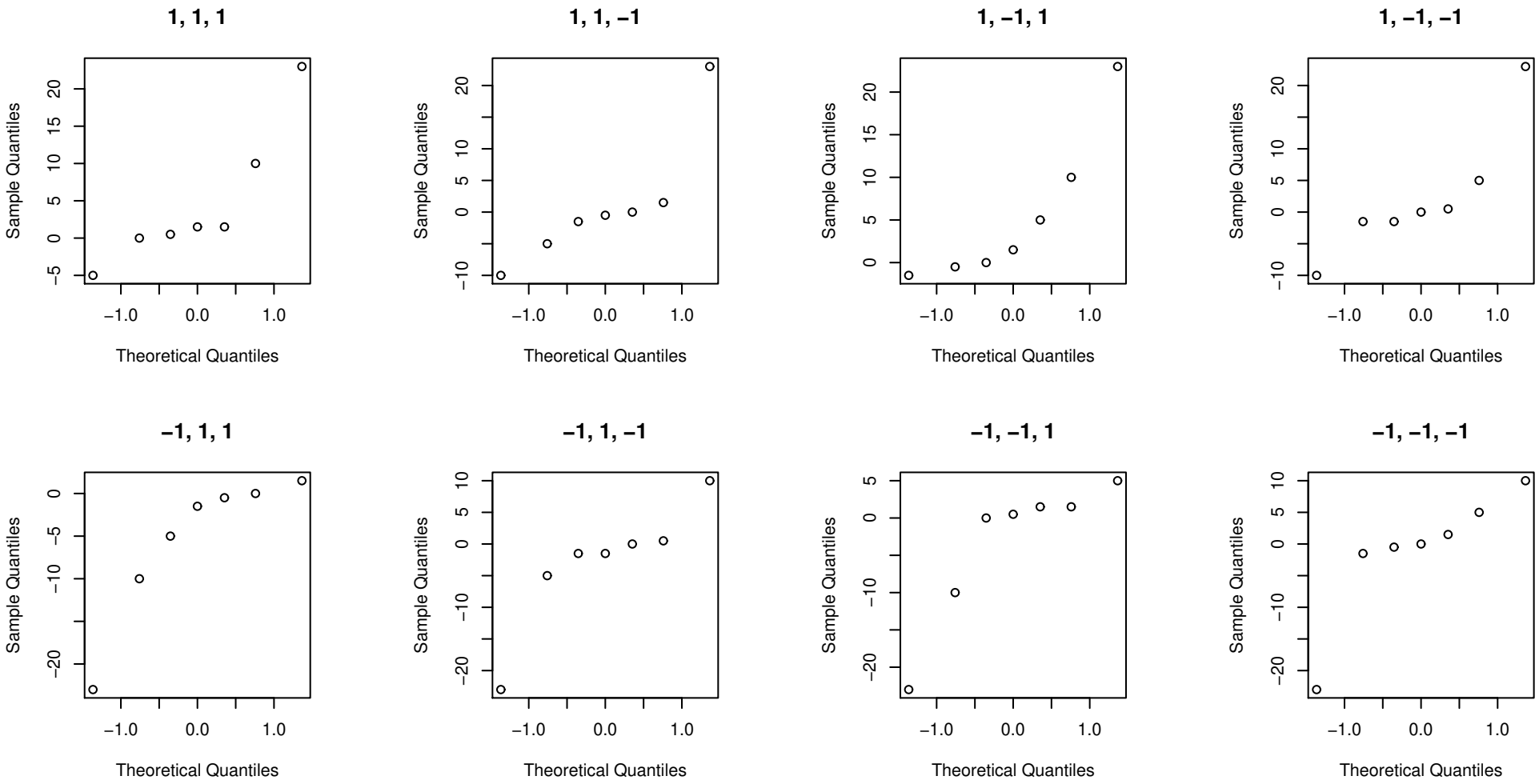


Problem with Daniel's method

Normal qq-plot depends on labeling of factor levels

	A	B	C	AB	AC	BC	ABC	y
	-1	-1	-1	1	1	1	-1	60
	1	-1	-1	-1	-1	1	1	72
	-1	1	-1	-1	1	-1	1	54
	1	1	-1	1	-1	-1	-1	68
	-1	-1	1	1	-1	-1	1	52
	1	-1	1	-1	1	-1	-1	83
	-1	1	1	-1	-1	1	-1	45
	1	1	1	1	1	1	1	80
Dot product	92	-20	6.0	6.0	40	0	2.0	
Division by 4	23	-5	1.5	1.5	10	0	0.5	

Daniel plots from switching factor labels



Method III: Lenth's method

Lenth (1989) proposed a 2-step estimate of the SD of the estimated effects

1. Given a set of effects $\hat{\theta}_1, \ldots, \hat{\theta}_g$, define initial estimate of standard error

$$s_0 = 1.5 \times \texttt{median}|\hat{\theta}_i|$$

   and the *pseudo standard error*

$$\texttt{PSE} = 1.5 \times \texttt{median}\{|\hat{\theta}_i| : |\hat{\theta}_i| < 2.5 s_0\}$$

2. Let $\nu = g/3$ and $\gamma = 0.5(1 - (1 - \alpha)^{1/g})$
3. Declare $\hat{\theta}_i$ significant at level α if $|\hat{\theta}_i| > t_{\nu,\gamma} \times \texttt{PSE}$

- Lenth's original method tends to be conservative
- More accurate critical values are given in, e.g., Wu and Hamada (2000)

Illustration of Lenth's method

1. A = 23, B = -5, C = 1.5, AB = 1.5, AC = 10, BC = 0, ABC = 0.5; i.e., $g = 7$
2. $|A|$ = 23, $|B|$ = 5, $|C|$ = 1.5, $|AB|$ = 1.5, $|AC|$ = 10, $|BC|$ = 0, $|ABC|$ = 0.5
3. $\texttt{median}(|\texttt{effects}|) = 1.5$ and $s_0 = 1.5 \times 1.5 = 2.25$
4. Only $|B|, |C|, |AB|, |BC|, |ABC|$ are less than $2.5 s_0 = 5.625$
5. Median $\{|B|, |C|, |AB|, |BC|, |ABC|\}$ = median $\{5, 1.5, 1.5, 0, 0.5\}$ = 1.5
6. $\texttt{PSE} = 1.5 \times 1.5 = 2.25$, $\nu = g/3 = 2.333$
7. Let $\alpha = 0.10$ so that $\gamma = 0.5(1 - (1 - \alpha)^{1/g}) = 0.00747$
8. Then $t_{\nu,\gamma} \times \texttt{PSE} = 6.566 \times 2.25 = 14.77$ and only A is significant

Method IV: Dong's method

Dong (1993) improved upon Lenth's method as follows:

1. Define (instead of PSE)

$$\begin{aligned} m_1 &= \#\{i \ : \ |\hat{\theta}_i| \leq 2.5 s_0\} \\ s_1^2 &= m_1^{-1} \sum_{|\hat{\theta}_i| \leq 2.5 s_0} \hat{\theta}_i^2 \end{aligned}$$

2. Repeat the previous step with s_1 in place of s_0:

$$\begin{aligned} m_2 &= \#\{i \ : \ |\hat{\theta}_i| \leq 2.5 s_1\} \\ s_2^2 &= m_2^{-1} \sum_{|\hat{\theta}_i| \leq 2.5 s_1} \hat{\theta}_i^2 \end{aligned}$$

3. Declare $\hat{\theta}_i$ as significant if $|\hat{\theta}_i| > t_{m_2,\gamma} \times s_2$

Illustration of Dong's method

1. A = 23, B = -5, C = 1.5, AB = 1.5, AC = 10, BC = 0, ABC = 0.5
2. $2.5s_0 = 5.625$, $m_1 = 5$ and $s_1^2 = (5^2 + 1.5^2 + 1.5^2 + 0 + 0.5^2)/5 = 5.95$
3. $s_1 = 2.439262$ and $|B|, |C|, |AB|, |BC|, |ABC|$ are less than $2.5s_1 = 6.098$
4. Hence $m_2 = 5$ (no change), $s_2 = s_1$,

$$t_{m_2,\gamma} \times s_2 = t_{5,0.00747} \times s_2 = 3.637741 \times 2.439262 = 8.873405$$

and A and AC are significant

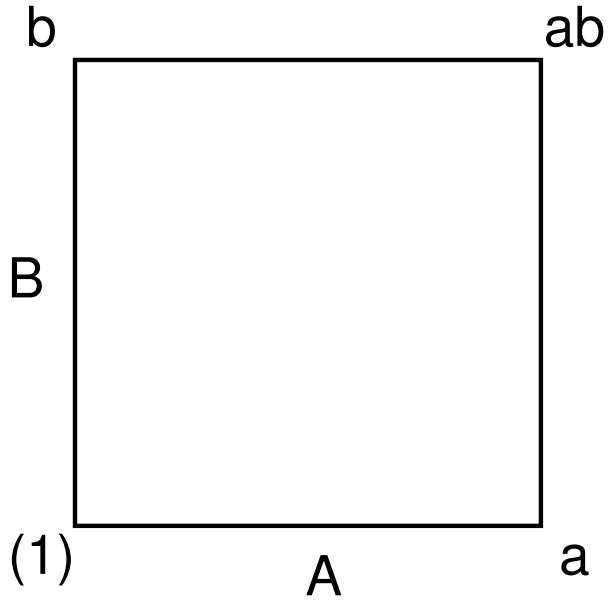
Blocking 2^2 in 2 blocks of size 2

Three possible designs:

D1. (1), a in block 1; b, ab in block 2

D2. (1), b in block 1; a, ab in block 2

D3. (1), ab in block 1; a, b in block 2



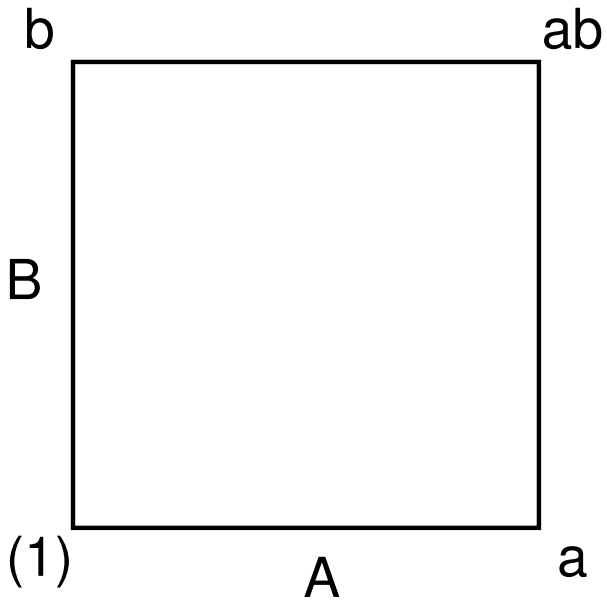
Blocking complete factorials: 2^2 in 2 blocks of size 2

Three possible designs:

D1. (1), a in block 1; b, ab in block 2

D2. (1), b in block 1; a, ab in block 2

D3. (1), ab in block 1; a, b in block 2



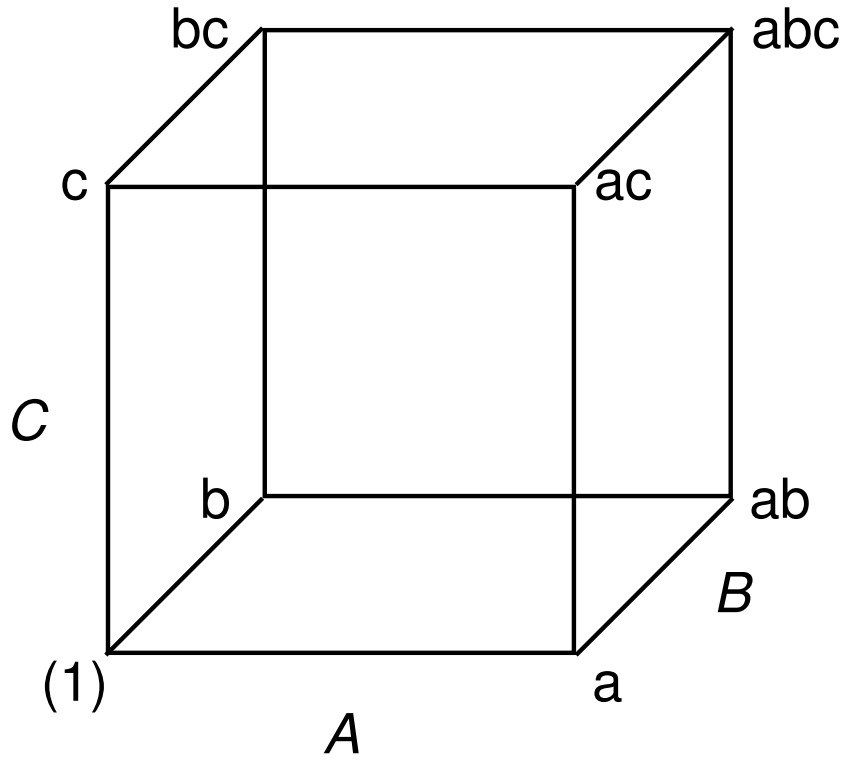
- Let $x_1, x_2 = \pm 1$ denote A and B levels and $z = \pm 1$ the block variable
- Model is $y = \eta + (A/2)x_1 + (B/2)x_2 + (C/2)x_1x_2 + \delta z + \epsilon$
- **D1.** B is confounded with block effect because

$$E(\hat{B}) = E(ab + b - a - (1))/2 = B + 2\delta \neq B$$

A and AB are not confounded

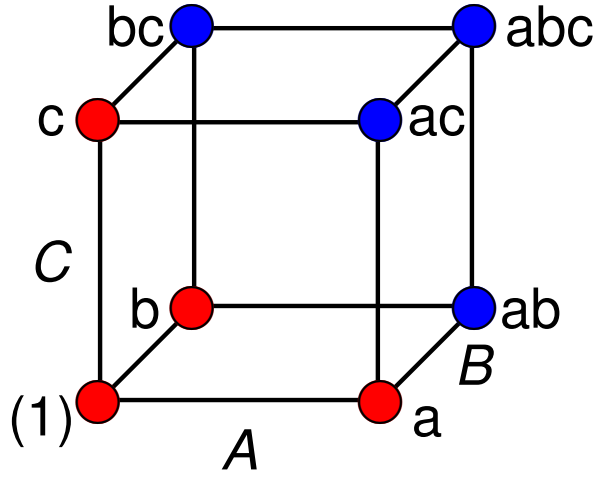
2^3 in 2 blocks of size 4

Question: How many different designs?



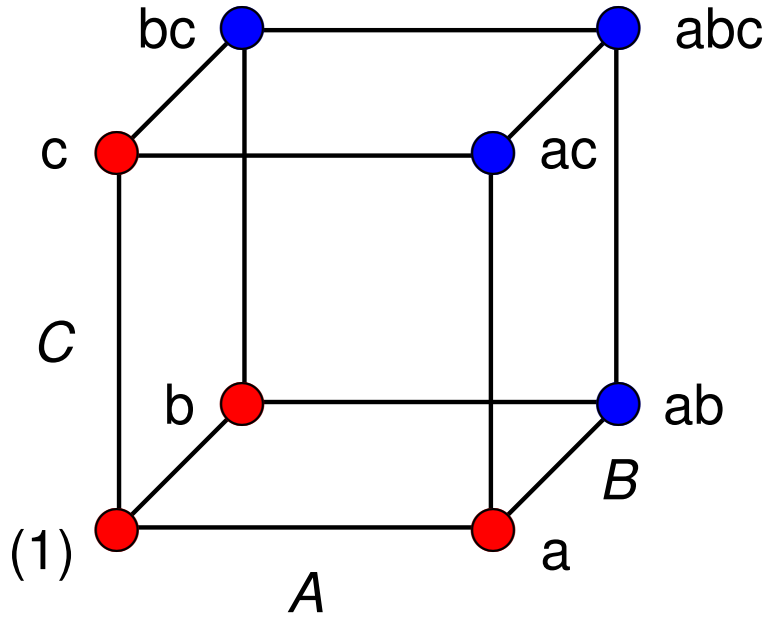
$$y = \eta + (A/2)x_1 + (B/2)x_2 + (C/2)x_3 + (AB/2)x_1x_2 + (AC/2)x_1x_3 + (BC/2)x_2x_3 + (ABC/2)x_1x_2x_3 + \delta z$$

OFAT design



$$
\begin{aligned}
(1) &= \eta - A/2 - B/2 - C/2 + AB/2 + AC/2 + BC/2 - ABC/2 - \delta + \epsilon_1 \\
a &= \eta + A/2 - B/2 - C/2 - AB/2 - AC/2 + BC/2 + ABC/2 - \delta + \epsilon_2 \\
b &= \eta - A/2 + B/2 - C/2 - AB/2 + AC/2 - BC/2 + ABC/2 - \delta + \epsilon_3 \\
c &= \eta - A/2 - B/2 + C/2 + AB/2 - AC/2 - BC/2 + ABC/2 - \delta + \epsilon_4 \\
ab &= \eta + A/2 + B/2 - C/2 + AB/2 - AC/2 - BC/2 - ABC/2 + \delta + \epsilon_5 \\
ac &= \eta + A/2 - B/2 + C/2 - AB/2 + AC/2 - BC/2 - ABC/2 + \delta + \epsilon_6 \\
bc &= \eta - A/2 + B/2 + C/2 - AB/2 - AC/2 + BC/2 - ABC/2 + \delta + \epsilon_7 \\
abc &= \eta + A/2 + B/2 + C/2 + AB/2 + AC/2 + BC/2 + ABC/2 + \delta + \epsilon_8
\end{aligned}
$$

OFAT design (cont'd.)



Only 2fi's unconfounded with block effect

$$
\begin{aligned}
E(\widehat{AB}) &= (1/4)(abc + ab - ac - bc - a - b + c + (1)) = AB \\
E(\widehat{AC}) &= (1/4)(abc - ab + ac - bc - a + b - c + (1)) = AC \\
E(\widehat{BC}) &= (1/4)(abc - ab - ac + bc + a - b - c + (1)) = BC \\
E(\widehat{ABC}) &= (1/4)(abc - ab - ac - bc + a + b + c - (1)) = ABC - \delta
\end{aligned}
$$

OFAT blocking

	A	B	C	AB	AC	BC	ABC	Block
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$	$-$
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$	$-$
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$-$
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$	$+$
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$-$
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$	$+$
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$	$+$
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$

Confounding only ABC

	A	B	C	AB	AC	BC	ABC	Block
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$	$-$
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$	$+$
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$+$
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$	$-$
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$	$-$
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$	$-$
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$

2^k in 2 blocks

Use the largest-order interaction column to define blocks

2^3 in 4 blocks of size 2

Need two block variables X and $Y = \pm 1$

	A	B	C	AB	AC	BC	ABC	X	Y
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$		
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$		
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$		
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$		
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$		
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$		
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$		
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$		

How not to do it: $X = ABC$ and $Y = BC$

	A	B	C	AB	AC	BC	ABC	X	Y
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$		
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$		
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$		
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$		
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$		
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$		
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$		
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$		

Confounds A with block effects

How to do it?

	A	B	C	AB	AC	BC	ABC	X	Y
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$		
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$		
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$		
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$		
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$		
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$		
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$		
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$		

Why $X = AB$ and $Y = BC$ works

	A	B	C	AB	AC	BC	ABC	X	Y
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$	$+$	$+$
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$	$-$	$+$
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$-$	$-$
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$	$+$	$-$
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$	$-$
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$	$-$	$-$
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$	$-$	$+$
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$

Main effects not confounded but all 2fi's are confounded with blocks

Partial confounding

- Suppose that a 2^3 design can only be run in blocks of size two
- Suppose further that, to achieve sufficient accuracy, replication is necessary and four replicates were to be run in 16 blocks of size two
- The design below would estimate main effects with greatest precision, providing less precision in the estimates of 2fi's and still less for the 3fi

	A	B	C	AB	AC	BC	ABC
Replicate 1	u	u	u	c	c	c	u
Replicate 2	c	u	u	u	u	c	c
Replicate 3	u	c	u	u	c	u	c
Replicate 4	u	u	c	c	u	u	c
# unconfounded reps.	3	3	3	2	2	2	1
# confounded reps.	1	1	1	2	2	2	3

2^4 in 4 blocks

Use ABC and ABD to define blocks; confounds CD as well

A	B	C	D	ABC	ABD	CD	Blk
−	−	−	−	−	−	+	1
+	−	−	−	+	+	+	4
−	+	−	−	+	+	+	4
+	+	−	−	−	−	+	1
−	−	+	−	+	−	−	3
+	−	+	−	−	+	−	2
−	+	+	−	−	+	−	2
+	+	+	−	+	−	−	3
−	−	−	+	−	+	−	2
+	−	−	+	+	−	−	3
−	+	−	+	+	−	−	3
+	+	−	+	−	+	−	2
−	−	+	+	+	+	+	4
+	−	+	+	−	−	+	1
−	+	+	+	−	−	+	1
+	+	+	+	+	+	+	4

2^5 in 4, 8 and 16 blocks

4 blocks. Use ABC and CDE to define blocks; confounds ABDE too

8 blocks. Use AB, CD and ACE to define blocks; confounds ABCD, ADE, BCE and BDE too

16 blocks. Use mirror-image pairs to form blocks; all 2fi's and 4fi's confounded

See Box et al. (2005, p. 221) for blocking 2^6 designs

Fractional two-level designs

Redundancy

- A 2^7 design permits 127 effects to be estimated
- Not all of these may be of appreciable size—there tends to be a hierarchy as in a Taylor expansion
- There tends to be redundancy in a 2^k design if k is not small—redundancy in terms of an excess number of interactions that can be estimated and sometimes an excess number of variables that are studied
- Fractional factorial designs exploit this redundancy

Use of fractions

Fractional factorials are used in:

1. industrial experimentation where runs are expensive and where an acceptable prior estimate of variance is available
2. screening experiments, which are performed in the early stages of an investigation in which many factors are considered, most of which may turn out to have little or no effect
   - The aim is to screen several factors with the purpose of spotting those factors, if any, that have appreciable effects
   - The emphasis is not so much upon estimating the effects with small standard errors, but upon finding out which factors merit further study

Assumptions

We shall assume here that:

1. The experimenter is interested in estimating the main effects, and perhaps the 2fi's
2. He is prepared to assume before carrying out the experiment that high-order interactions are negligible
3. He is not particularly concerned with using the actual experiment to provide himself with an estimate of the variance
4. The emphasis is on finding small experiments in which a high percentage of the degrees of freedom are used for estimation

Construction of 2^{5-1} design

- Write a full 2^4 design in variables A, B, C, and D
- Add a column of signs for factor E defined by

$$E = ABCD$$

This is called the **generator** of the design

2^{5-1} design using E = ABCD

A	B	C	D	E	A	B	C	D	E
−	−	−	−	+	−	−	−	+	−
+	−	−	−	−	+	−	−	+	+
−	+	−	−	−	−	+	−	+	+
+	+	−	−	+	+	+	−	+	−
−	−	+	−	−	−	−	+	+	+
+	−	+	−	+	+	−	+	+	−
−	+	+	−	+	−	+	+	+	−
+	+	+	−	−	+	+	+	+	+

Properties of E=ABCD design

- Multiplying both sides of the generator by E gives:

$$I = ABCDE \tag{1}$$

  where I is a column of plus signs

- Equation (1) is called the **defining relation** of the design
- Multiplying (1) on both sides by any main or interaction effect yields its confounding pattern, e.g., $A = BCDE$, $B = ACDE$, $AB = CDE$, etc.
- Let $\hat{A}$, $\widehat{AB}$, etc., be the usual LSEs of A, AB, etc.
- Then $E(\hat{A}) = A + BCDE$, $E(\widehat{AB}) = AB + CDE$, etc.

Concept of resolution

A design of **resolution** R is one in which no p-factor effect is confounded with any other effect containing less than $R - p$ factors

Examples:

1. A resolution III design does not confound main effects with one another but does confound main effects with 2fi's
2. A resolution IV design does not confound main effects and 2fi's but does confound 2fi's with other 2fi's
3. A resolution V design does not confound main effects and 2fi's with each other, but does confound 2fi's with 3fi's, and so on

In general, the **resolution** of a two-level fractional design is the **length of the shortest word in the defining relation**

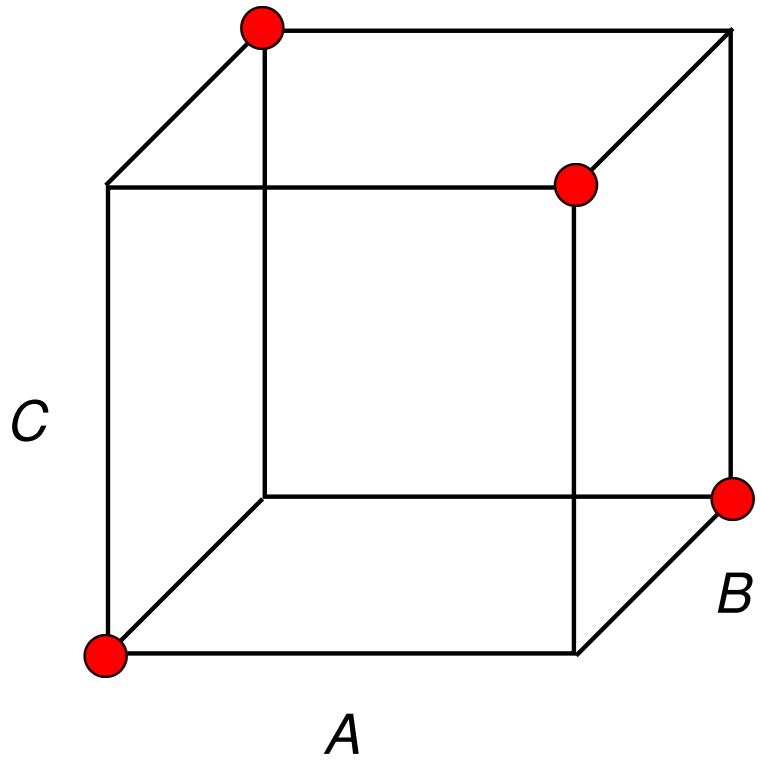
Half-fractions of the highest resolution

To construct a 2^{k-1} fractional factorial design of the highest possible resolution:

1. Write a full factorial design for the first $k-1$ variables
2. Associate the kth variable with plus or minus the interaction column $123\ldots(k-1)$

2^{3-1}_{III} design

This design gives a **complete** factorial in any two variables



Alternative rationale for the half-fraction design: factorials embedded in fractions

- In the 2_V^{5-1} design, a **complete** factorial in the remaining variables is obtained **whichever column is omitted**
- A fractional factorial design of resolution R contains complete factorials (possibly replicated) in every set of $R-1$ variables

Sequential use of fractional designs

It is almost always better for an experimenter planning a 2^5 design in 32 runs to run a half-fraction of 16 runs first, analyze the results, and think about them

1. The experimenter should randomize within each fraction
2. If eventually it is decided to run both fractions, they will be randomized orthogonal blocks of the complete factorial design
3. No information will be "lost" except that concerning the interaction which is confounded with the block effect
4. The design run as two randomized fractions can give greater precision than the whole design run in random order because the block effect is eliminated