Statistics 502 Lecture Notes

Peter D. Hoff

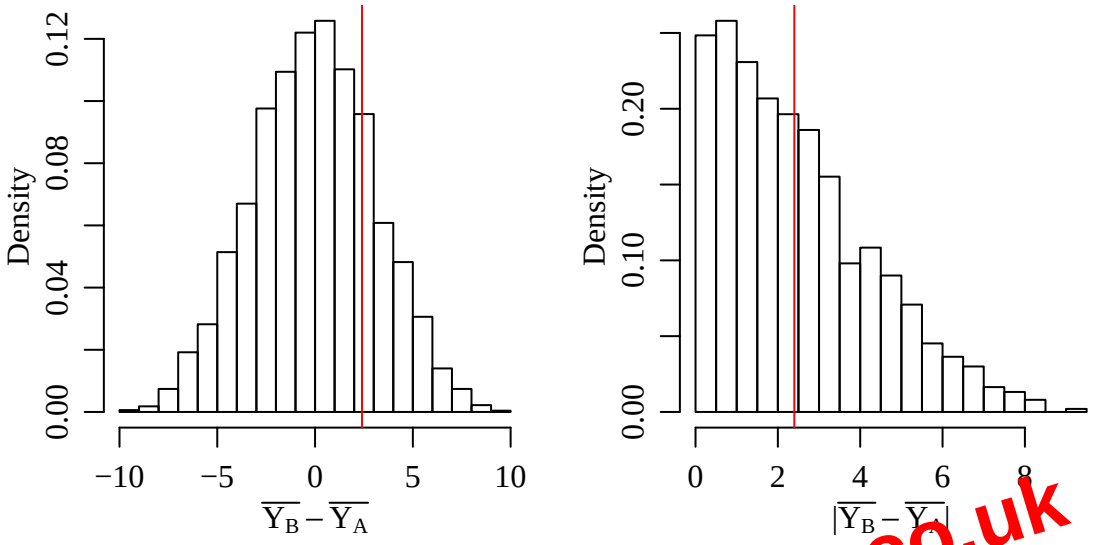

Figure 2.2: Randomization distribution for the wheat example

Approximating a randomization distribution We don't want to have to enumerate all $\binom{n}{n/2}$ possible treatment assignments. Instead, repeat the following `Nsim` times:

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

(b) compute the value of the test statistic, given the simulated treatment assignment and under H_0.

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

$$\frac{\#(|g_k|) \geq 2.4)}{\texttt{Nsim}} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq 2.4 | H_0)$$

The approximation improves if `Nsim` increased.

Here is some R-code:

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

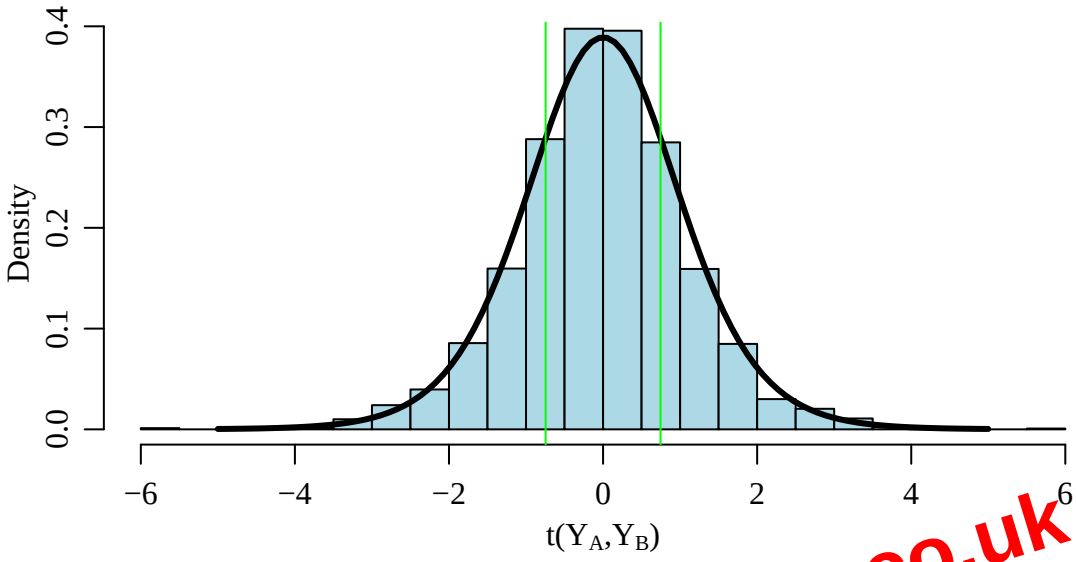

Figure 2.8: Randomization distribution for the t-statistic


```
xsim<-sample(x)
t.stat.sim[nsim]<- t.test(y[xsim=="B"],y[xsim=="A"],var.equal=T)$stat
                  }
mean( abs(t.stat.sim) >= abs(t.stat.obs) )
```

When I ran this, I got

$$\frac{\#(|t^{(j)}| \geq 0.75)}{\texttt{nsim}} = 0.48 \approx 0.47 = \Pr(|T_{n_A+n_B-2}| \geq 0.75)$$

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

Comparison:

- **Assumptions:**
  - Randomization Test: (1) Treatments are randomly assigned
  - t-test:

Questions:

- What does the fact that 0 is in the interval say about $H_0 : \mu_A = \mu_B$?
- What is the interpretation of this interval?
- Could we have constructed an interval via randomization tests?

2.9 Power and Sample Size Determination

Suppose that we plan to gather data, and then perform a hypothesis test.

Two sample t-test:

- H_0: $\mu_A = \mu_B$ $\quad$ H_1: $\mu_A \neq \mu_B$
- Gather data
- Perform a level α hypothesis test: reject H_0 if

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

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

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

$$\Pr(\text{type I error}|H_0 \text{ true}) = \Pr(\text{reject } H_0|H_0 \text{ true}) = 0.05$$

What about

$$\begin{aligned} \Pr(\text{type II error}|H_0 \text{ false}) &= \Pr(\text{accept } H_0|H_0 \text{ false}) \\ &= 1 - \Pr(\text{reject } H_0|H_0 \text{ false}) \end{aligned}$$

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

$$\Pr(\text{reject } H_0|\mu_B - \mu_A = .0001) < \Pr(\text{reject } H_0|\mu_B - \mu_A = 10,000)$$

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

$$1 - \Pr(\text{type II error} | \mu_B - \mu_A = \delta) = \Pr(\text{reject } H_0 | \mu_B - \mu_A = \delta)?$$

We define the **power** of a two-sample hypothesis test **against a specific alternative** to be:

$$\begin{aligned} \text{Power}(\delta, \sigma, n_A, n_B) &= \Pr(\text{reject } H_0 \mid \mu_B - \mu_A = \delta) \\ &= \Pr(|t(\mathbf{Y}_A, \mathbf{Y}_B)| \geq t_{1-\alpha/2, n_A+n_B-2} \big| \, \mu_B - \mu_A = \delta). \end{aligned}$$

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

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

If we suppose $Y_{A1}, \ldots, Y_{An_A} \sim$ i.i.d. normal(μ_A, σ) and $Y_{B1}, \ldots, Y_{Bn_B} \sim$ i.i.d. normal(μ_B, σ), where $\mu_B - \mu_A = \delta$ then we need to know the distribution of

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

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

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

but unfortunately

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

So, even though the **pairwise error rate** is 0.05 the **experiment-wise error rate** is 0.26.

This issue is called the problem of **multiple comparisons** and will be discussed further in Chapter 3. For now, we will discuss a method of testing the **global hypothesis** of no **variation due to treatment**:

$$H_0 : \mu_i = \mu_j \text{ for all } i \neq j \quad \text{versus} \quad H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j$$

To do this, we will compare **treatment variability** to **experimental variability**. First we need to have a way of quantifying these things.

3.1.1 A model for treatment variation

Data: $y_{i,j}$ = measurement from the jth replicate under th ith treatment.

$i = 1, \ldots, t$ indexes treatments

$j = 1, \ldots, r$ indexes observations or replicates

Treatment means model:

$$\begin{aligned} y_{i,j} &= \mu_i + \epsilon_{i,j} \\ E[\epsilon_{i,j}] &= 0 \\ V[\epsilon_{i,j}] &= \sigma^2 \end{aligned}$$

- μ_i is the ith treatment mean,
- $\epsilon_{i,j}$ represents **within treatment variation**/error/noise.

Treatment effects model:

$$\begin{aligned} y_{i,j} &= \mu + \tau_i + \epsilon_{i,j} \\ E[\epsilon_{i,j}] &= 0 \\ V[\epsilon_{i,j}] &= \sigma^2 \end{aligned}$$

- μ is the **grand mean**;
- $\tau_1, \ldots, \tau_t$ are the **treatment effects**, representing **between treatment variation**

We can "decompose" each observation as follows:

$$y_{i,j} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{i,j} - \bar{y}_i)$$

This leads to

$(y_{i,j} - \bar{y})$	$=$	$(\bar{y}_i - \bar{y})$	$+$	$(y_{i,j} - \bar{y}_i)$
total variation	$=$	between group variation	$+$	within group variation

All data can be decomposed this way, leading to the following vectors of length tr :

Total		Treatment		Error
$y_{1,1} - \bar{y}_{..}$	$=$	$(\bar{y}_{1.} - \bar{y}_{..})$	$+$	$(y_{1,1} - \bar{y}_{1.})$
$y_{1,2} - \bar{y}_{..}$	$=$	$(\bar{y}_{1.} - \bar{y}_{..})$	$+$	$(y_{1,2} - \bar{y}_{1.})$
.	$=$	.	$+$	.
.	$=$	.	$+$	.
.	$=$	.	$+$	.
$y_{1,r} - \bar{y}_{..}$	$=$	$(\bar{y}_{1.} - \bar{y}_{..})$	$+$	$(y_{1,r} - \bar{y}_{1.})$
$y_{2,1} - \bar{y}_{..}$	$=$	$(\bar{y}_{2.} - \bar{y}_{..})$	$+$	$(y_{2,1} - \bar{y}_{2.})$
.	$=$	.	$+$	.
.	$=$	.	$+$	.
.	$=$	.	$+$	.
$y_{2,r} - \bar{y}_{..}$	$=$	$(\bar{y}_{2.} - \bar{y}_{..})$	$+$	$(y_{2,r} - \bar{y}_{2.})$
⋮		⋮		⋮
$y_{t,1} - \bar{y}_{..}$	$=$	$(\bar{y}_{t.} - \bar{y}_{..})$	$+$	$(y_{t,1} - \bar{y}_{t.})$
.	$=$	.	$+$	.
.	$=$	.	$+$	.
.	$=$	.	$+$	.
$y_{t,r} - \bar{y}_{..}$	$=$	$(\bar{y}_{t.} - \bar{y}_{..})$	$+$	$(y_{t,r} - \bar{y}_{t.})$
SSTotal	$=$	SSTrt	$+$	SSE
$tr - 1$	$=$	$t - 1$	$+$	$t(r - 1)$

We've seen degrees of freedom before, in the definition of a χ^2 random variable:

- dof = the number of statistically independent elements in a vector

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

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

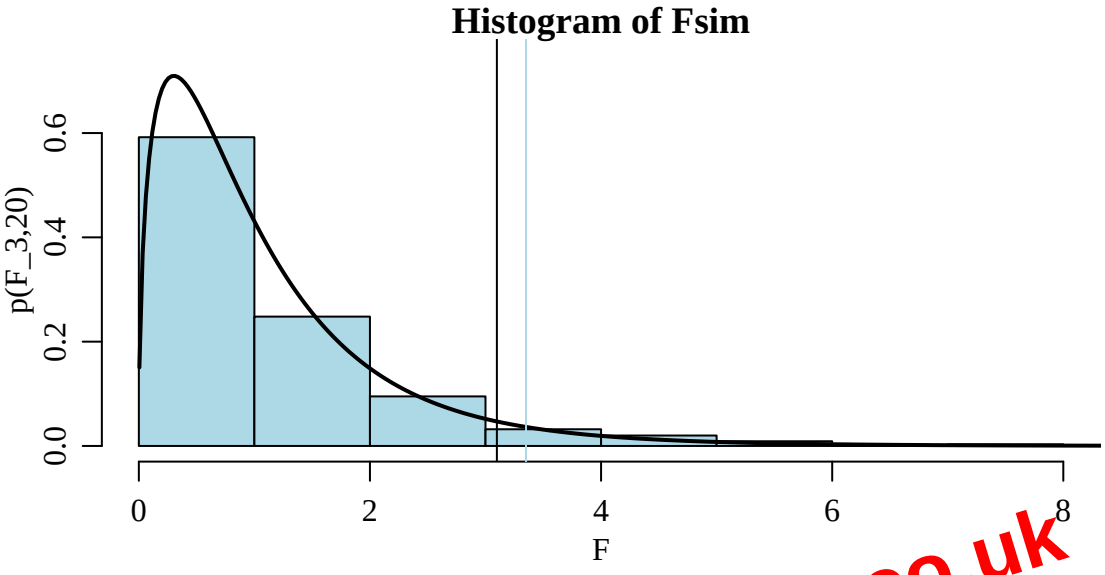

Figure 3.5: Normal-theory and randomization distributions of the F-statistic


```
Residuals 20  112.0          5.6

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

> mean(Fsim>=Fobs)
[1] 2e-04

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

3.1.10 Comparing group means

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

If H_0 is rejected we

- estimate μ_i with $\bar{y}_i$;
- estimate σ_i^2 with
  - s_i^2 : if variances are very unequal, this might be a better estimate.
  - MSE : if variances are close and r is small, this might be a better estimate.

Standard practice: Unless strong evidence to the contrary, we typically assume $V(Y_{i,j}) = V(Y_{k,l}) = \sigma^2$, and use $s^2 \equiv MSE$ to estimate σ^2. In this case,

$$\begin{aligned} V(\hat{\mu}_i) = V(\bar{Y}_{i\cdot}) &= \sigma^2/r_i \\ &\approx s^2/r_i \end{aligned}$$

The "standard error of the mean" $= SE(\hat{\mu}_i) = \sqrt{s^2/r_i}$ is an **estimate** of $V(\hat{\mu}_i) = V(\bar{Y}_i\cdot) = \sigma^2/r_i$

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

$$\begin{aligned} \hat{\theta} &= \hat{\theta}(\mathbf{Y}) \\ V(\hat{\theta}) &= \gamma^2 \\ SE(\hat{\theta}) &= \hat{\gamma} \end{aligned}$$

where $\hat{\gamma}^2$ is an estimate of γ^2.

Confidence intervals for treatment means: Similar to the one sample case.

$$\begin{aligned} \frac{\bar{Y}_{i\cdot} - \mu_i}{SE(\bar{Y}_{i\cdot})} = \sqrt{r_i}\frac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{MSE}} &= \frac{\bar{Y}_{i,\cdot} - \mu_i}{s/\sqrt{r_i}} \\ &\sim t_{N-t} \end{aligned}$$

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

$$\bar{Y}_i \pm SE(\bar{Y}_{i\cdot}) \times t_{1-\alpha/2,N-t}$$

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

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

However, if effects are **multiplicative** so that in fact:

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

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

$$Var(Y_{ij}) = \mu_i^2 Var(X_{ij1} \times X_{ij2} \times \cdots)$$

Log transform:

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

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

Crab data: Let $Y_{i,j} = \log(Y^{\text{raw}}_{i,j} + 1/6)$

Site	Mean	SD
6	0.82	2.21
4	0.91	1.87
5	1.01	1.74
3	1.75	2.41
1	2.16	2.27
2	2.30	2.44

Here are some common transformations:

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

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

More about the log transform

How did $\alpha = 1$ give us $y^*_{ij} = \log y_{ij}$, shouldn't it be $y^* = y^\alpha_{ij} = y^\lambda_{ij} = y^0_{ij} = 1$ in there?

Everything will make sense if we define for any $\lambda \neq 0$:

$$y^{*(\lambda)} = \frac{y^\lambda - 1}{\lambda} \propto y^\lambda + c.$$

For $\lambda = 0$, it's natural to define the transformation as:

$$\begin{aligned} y^{*(0)} = \lim_{\lambda \to 0} y^{*(\lambda)} &= \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} \\ &= \left. \frac{y^\lambda \ln y}{1} \right|_{\lambda=0} = \ln y \end{aligned}$$

Note that for a given $\lambda \neq 0$ it will not change the results of the ANOVA on the transformed data if we transform using:

$$y^* = y^\lambda \quad \text{or} \quad y^{*(\lambda)} = \frac{y^\lambda - 1}{\lambda} = ay^\lambda + b.$$

- Don't assume that the transformation is a magical fix: remember to look at residuals and diagnostics **after** you do the transform. If things haven't improved much, don't transform.
- Remember that the mean-variance relationship might not be cured by a transform in the Box-Cox class.
  - (e.g. if the response is a binomial proportion (= proportion of successes out of n), we have mean = p, s.d. = $\sqrt{p(1-p)}$; the variance stabilizing transformation in this case is $y^* = \arcsin\sqrt{y}$.)
- Keep in mind that statisticians disagree on the usefulness of transformations: some regard them as a 'hack' more than a 'cure':
  - It can be argued that if the scientist who collected the data had a good reason for using certain units, then one should not just transform the data in order to bang it into an ANOVA-shaped hole. (Given enough time and thought we could instead build a non-linear model for the original data.)
- **The sad truth**: as always you will need to exercise **judgment** while performing your analysis.

These warnings apply whenever you might reach for a transform, whether in an ANOVA context, or a linear regression context.

Example (Crab data): Looking at the plot of means vs. sd.s suggests $\alpha \approx 1$, implying a log-transformation. However, the zeros in our data lead to problems, since $\log(0) = -\infty$.

Instead we can use $y^*_{ij} = \log(y_{ij} + 1/6)$. (See plots.) For the transformed data this gives us a ratio of the largest to smallest standard deviation of approximately 2 which is acceptable based on the rule of 3.

site	sample sd	sample mean	log(sample sd)	log(sample mean)
4	17.39	9.24	2.86	2.22
5	19.84	10.00	2.99	2.30
6	23.01	12.64	3.14	2.54
1	50.39	33.80	3.92	3.52
3	107.44	50.64	4.68	3.92
2	125.35	68.72	4.83	4.23

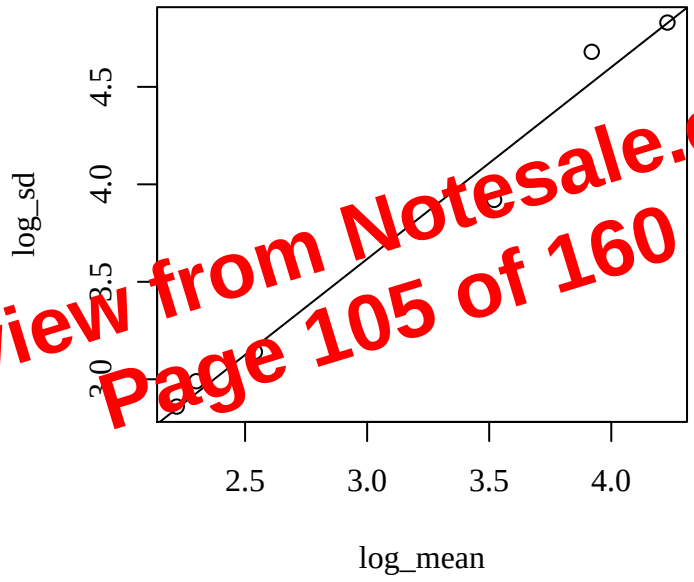

Figure 3.15: Mean variance relationship of the transformed data

4 assigned to (I, B),

$\vdots$

It might be helpful to visualize the design as follows:

	Delivery			
Type	A	B	C	D
1	$\mathbf{y}_{I,A}$	$\mathbf{y}_{I,B}$	$\mathbf{y}_{I,C}$	$\mathbf{y}_{I,D}$
2	$\mathbf{y}_{II,A}$	$\mathbf{y}_{II,B}$	$\mathbf{y}_{II,C}$	$\mathbf{y}_{II,D}$
3	$\mathbf{y}_{III,A}$	$\mathbf{y}_{III,B}$	$\mathbf{y}_{III,C}$	$\mathbf{y}_{III,D}$

This type of design is called a **factorial design**. Specifically, this design is a 3×4 **two-factor** design with 4 **replications**.

Factors: Categories of treatments

Levels of a factor: the different treatments in a category

So in this case, Type and Delivery are both factors. There are 3 levels of Type and 4 levels of Delivery.

4.1.1 Data analysis:

Lets first look at a series of plots:

Marginal Plots: Based on these marginal plots, it looks like (III, A) would be the most effective combination. But are the effects of Type consistent across levels of Delivery?

Conditional Plots: Type III looks best across delivery types. But the difference between types I and II seems to depend on delivery.

Cell Plots: Another way of looking at the data is to just view it as a CRD with $3 \times 4 = 12$ different groups. Sometimes each group is called a **cell**.

Notice that there seems to be a mean-variance relationship. Lets take care of this before we go any further: Plotting means versus standard deviations on both the raw and log scale gives the relationship in Figure 4.4. Computing the least squares line gives

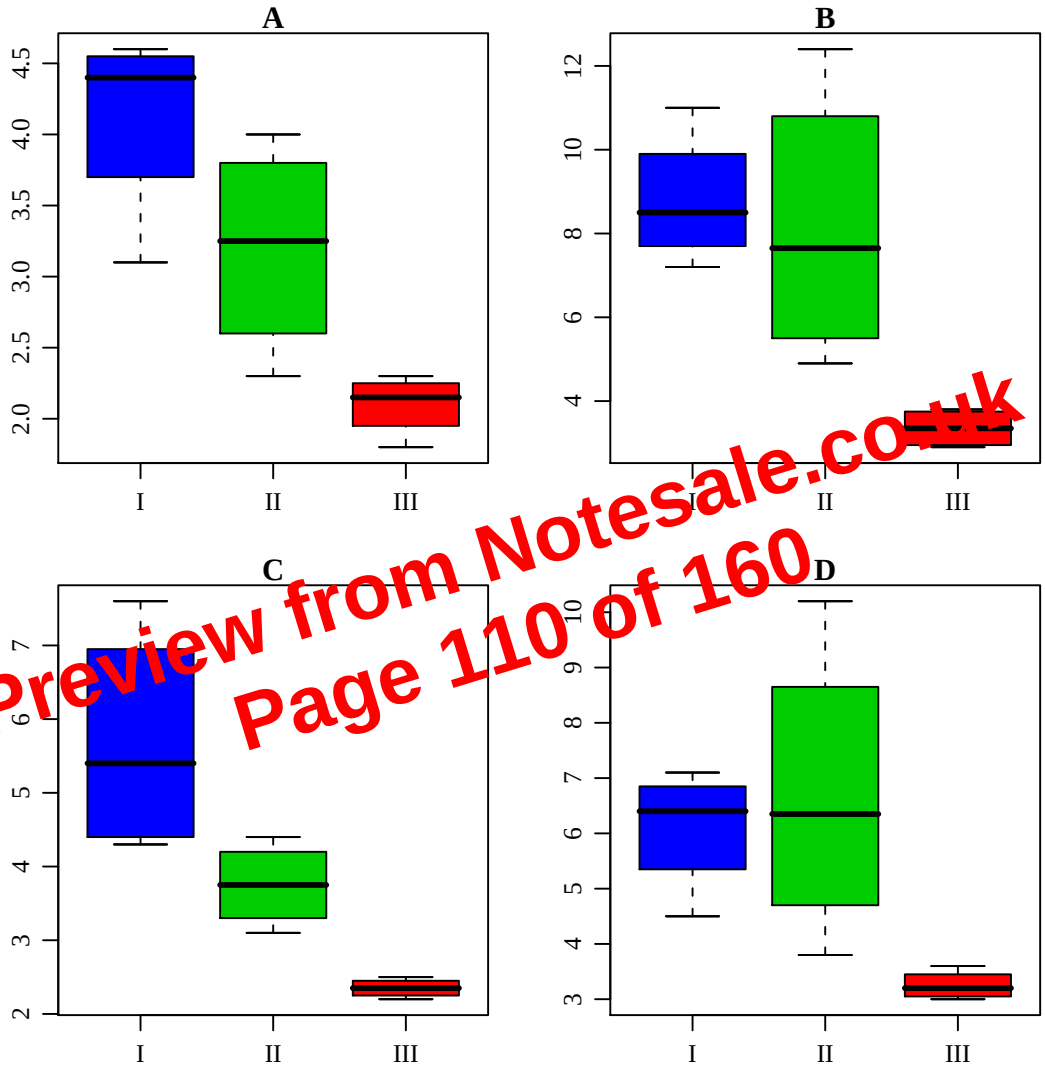

Figure 4.2: Conditional Plots.

1	parameter for μ
$t_1 - 1$	parameters for a_i's
$t_2 - 1$	parameters for b_j's
$t_1 + t_2 - 1$	parameters total.

Parameter estimation and ANOVA decomposition:

$$\begin{array}{rcccccccc} y_{ijk} & = & \bar{y}_{\cdots} & + & (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) & + & (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) & + & (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}) \\ & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j & + & \hat{\epsilon}_{ijk} \end{array}$$

These are the least-squares parameter estimates, under the **sum-to-zero** side conditions:

$$\sum \hat{a}_i = \sum (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) = n\bar{y}_{\cdots} - n\bar{y}_{\cdots} = 0$$

To obtain the **set-to-zero** side conditions, add $\hat{a}_1$ and $\hat{b}_1$ to $\hat{\mu}$, subtract $\hat{a}_1$ from the $\hat{a}_i$'s, and subtract $\hat{b}_1$ from the $\hat{b}_j$'s. Note that this does not change the fitted value in each group:

$$\begin{array}{rcccccc} \text{fitted}(y_{ijk}) & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j \\ & = & (\hat{\mu} + \hat{a}_1 + \hat{b}_1) & + & (\hat{a}_i - \hat{a}_1) & + & (\hat{b}_j - \hat{b}_1) \\ & = & \hat{\mu}^* & + & \hat{a}_i^* & + & \hat{b}_j^* \end{array}$$

As you might have guessed, we can write this decomposition out as vectors of length $t_1 \times t_2 \times r$:

$$\begin{array}{rcccccc} \boldsymbol{y} - \bar{y}_{\cdots} & = & \hat{\boldsymbol{a}} & + & \hat{\boldsymbol{b}} & + & \hat{\boldsymbol{\epsilon}} \\ v_T & = & v_1 & + & v_2 & + & v_e \end{array}$$

The columns represent

v_T variation of the data around the grand mean;

v_1 variation of factor 1 means around the grand mean;

v_2 variation of factor 2 means around the grand mean;

v_e variation of the data around fitted the values.

You should be able to show that these vectors are orthogonal, and so

$$\begin{array}{rcccccc} \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{\cdots})^2 & = & \sum_i \sum_j \sum_k \hat{a}_i^2 & + & \sum_i \sum_j \sum_k \hat{b}_i^2 & + & \sum_i \sum_j \sum_k \hat{\epsilon}_i^2 \\ \text{SSTotal} & = & \text{SSA} & + & \text{SSB} & + & \text{SSE} \end{array}$$

- The full model allows differences between Types to vary across levels of Delivery
- The reduced/additive model says differences are constant across levels of Delivery.

Therefore, the reduced model is appropriate if

$$(\mu_{IA} - \mu_{IIA}) = (\mu_{IB} - \mu_{IIB}) = (\mu_{IC} - \mu_{IIC}) = (\mu_{ID} - \mu_{IID})$$

How can we test for this? Consider the following **parameterization** of the full model:

Interaction model:

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}$$

μ = overall mean;

$a_1, \ldots, a_{t_1}$ = **additive** effects of factor 1;

$b_1, \ldots, b_{t_2}$ = **additive** effects of factor 2;

$(ab)_{ij}$ = **interaction terms** = **deviations from additivity**.

The **interaction term** is a correction for non-additivity of the factor effects. This is a full model: It fits a separate mean for each treatment combination:

$$E(Y_{ijk}) = \mu_{ij} = \mu + a_i + b_j + (ab)_{ij}$$

Parameter estimation and ANOVA decomposition:

$$\begin{array}{rcccccccccc} y_{ijk} & = & \bar{y}_{\cdots} & + & (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) & + & (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) & + & (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}) & + & (y_{ijk} - \bar{y}_{ij\cdot}) \\ & = & \hat{\mu} & + & \hat{a}_i & + & \hat{b}_j & + & (\hat{ab})_{ij} & + & \hat{\epsilon}_{ijk} \end{array}$$

Deciding between the additive/reduced model and the interaction/full model is tantamount to deciding if the variance explained by the $(\hat{ab})_{ij}$'s is large or not.

– $E(MSAB) = \sigma^2 + r\tau^2_{AB} > \sigma^2$.

This suggests

- An evaluation of the adequacy of the additive model can be assessed by comparing $MSAB$ to MSE. Under $H_0 : (ab)_{ij} = 0$,

$$F_{AB} = MSAB/MSE \sim F_{(t_1-1)\times(t_2-1),t_1t_2(r-1)}$$

  Evidence against H_0 can be evaluated by computing the p-value.

- If the additive model is adequate then MSE_{int} and $MSAB$ are two independent estimates of roughly the same thing (why independent?). We may then want to combine them to improve our estimate of σ^2.

```
                      Df  Sum Sq Mean Sq F value    Pr(>F)
pois$deliv             3 0.20414 0.06805 28.3431 1.376e-09 ***
pois$type              2 0.34877 0.17439 72.6347 2.240e-13 ***
pois$deliv:pois$type   6 0.01571 0.00262  1.0900    0.3867
Residuals             36 0.08643 0.00240
```

For these data, there is strong evidence of both treatment effects, and little evidence of non-additivity. We may want to use the additive model.

4.1.4 Inference for additive treatment effects

Consider a two-factor experiment in which it is determined that the effects of factor F_1 and F_2 are large. Now we want to compare means across levels of one of the factors.

Recall in the pesticide example we had 4 reps for each of 3 levels of Type and 4 levels of Delivery. So we have $4 \times 4 = 16$ observations for each level of Type.

The wrong approach: The two-sample t-test is

$$\frac{\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}}{s_{12}\sqrt{2/(4 \times 4)}}$$

For the above example,

- $\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} = 0.047$

(population) means:

	$F_2=1$	$F_2=2$	$F_2=3$	$F_2=4$	
$F_1=1$	$\bar{\mu}_{11\cdot}$	$\bar{\mu}_{12\cdot}$	$\bar{\mu}_{13\cdot}$	$\bar{\mu}_{14\cdot}$	$4\bar{\mu}_{1\cdot\cdot}$
$F_1=2$	$\bar{\mu}_{21\cdot}$	$\bar{\mu}_{22\cdot}$	$\bar{\mu}_{23\cdot}$	$\bar{\mu}_{24\cdot}$	$4\bar{\mu}_{2\cdot\cdot}$
$F_1=3$	$\bar{\mu}_{31\cdot}$	$\bar{\mu}_{32\cdot}$	$\bar{\mu}_{33\cdot}$	$\bar{\mu}_{34\cdot}$	$4\bar{\mu}_{3\cdot\cdot}$
	$3\bar{\mu}_{\cdot1\cdot}$	$3\bar{\mu}_{\cdot2\cdot}$	$3\bar{\mu}_{\cdot3\cdot}$	$3\bar{\mu}_{\cdot4\cdot}$	$12\bar{\mu}_{\cdots}$

So

$$\begin{aligned} a_1 - a_2 &= \bar{\mu}_{1\cdot\cdot} - \bar{\mu}_{2\cdot\cdot} \\ &= (\bar{\mu}_{11\cdot} + \bar{\mu}_{12\cdot} + \bar{\mu}_{13\cdot} + \bar{\mu}_{14\cdot})/4 - (\bar{\mu}_{21\cdot} + \bar{\mu}_{22\cdot} + \bar{\mu}_{23\cdot} + \bar{\mu}_{24\cdot})/4 \end{aligned}$$

Like any contrast, we can estimate/make inference for it using contrasts of sample means:

$a_1 - a_2 = \hat{a}_1 - \hat{a}_2 = \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$ is an unbiased estimate of $a_1 - a_2$

Note that this estimate is the corresponding **contrast** among the $t_1 \times t_2$ sample means:

	$F_2=1$	$F_2=2$	$F_2=3$	$F_2=4$	
$F_1=1$	$\bar{y}_{11\cdot}$	$\bar{y}_{12\cdot}$	$\bar{y}_{13\cdot}$	$\bar{y}_{14\cdot}$	$4\bar{y}_{1\cdot\cdot}$
$F_1=2$	$\bar{y}_{21\cdot}$	$\bar{y}_{22\cdot}$	$\bar{y}_{23\cdot}$	$\bar{y}_{24\cdot}$	$4\bar{y}_{2\cdot\cdot}$
$F_1=3$	$\bar{y}_{31\cdot}$	$\bar{y}_{32\cdot}$	$\bar{y}_{33\cdot}$	$\bar{y}_{34\cdot}$	$4\bar{y}_{3\cdot\cdot}$
	$3\bar{y}_{\cdot1\cdot}$	$3\bar{y}_{\cdot2\cdot}$	$3\bar{y}_{\cdot3\cdot}$	$3\bar{y}_{\cdot4\cdot}$	$12\bar{y}_{\cdots}$

So

$$\begin{aligned} \hat{a}_1 - \hat{a}_2 &= \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} \\ &= (\bar{y}_{11\cdot} + \bar{y}_{12\cdot} + \bar{y}_{13\cdot} + \bar{y}_{14\cdot})/4 - (\bar{y}_{21\cdot} + \bar{y}_{22\cdot} + \bar{y}_{23\cdot} + \bar{y}_{24\cdot})/4 \end{aligned}$$

Hypothesis tests and confidence intervals can be made using the standard assumptions:

- $E(\hat{a}_1 - \hat{a}_2) = a_1 - a_2$
- Under the assumption of constant variance:

$$\begin{aligned} V(\hat{a}_1 - \hat{a}_2) &= V(\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}) \\ &= V(\bar{y}_{1\cdot\cdot}) + V(\bar{y}_{2\cdot\cdot}) \\ &= \sigma^2/(r \times t_2) + \sigma^2/(r \times t_2) \\ &= 2\sigma^2/(r \times t_2) \end{aligned}$$

then it will increase the variance in response and also the experimental error variance/MSE if unaccounted for. If F_2 is a known, potentially large source of variation, we can control for it pre-experimentally with a **block design**.

Blocking: The stratification of experimental units into groups that are more homogeneous than the whole.

Objective: To have less variation among units within blocks than between blocks.

Typical blocking criteria:

- location
- physical characteristics
- time

Example(Nitrogen fertilizer timing): How does the timing of nitrogen additive affect nitrogen uptake?

- Treatment: Six different timing schedules $1, \ldots, 6$: Level 4 is "standard"
- Response: Nitrogen uptake (ppm$\times 10^{-2}$)
- Experimental material: One irrigated field

Soil moisture is thought to be a source of variation in response.

Design:

1. Field is divided into a 4×6 grid.
2. Within each row or **block**, each of the 6 treatments are randomly allocated.

1. The experimental units are **blocked** into presumably more homogeneous groups.
2. The blocks are **complete**, in that each treatment appears in each block.

Figure 4.13: Marginal plots and residuals

Consider comparing the F-stat from a CRD with that from an RCB: According to Cochran and Cox (1957)

$$\begin{aligned} MSE_{crd} &= \frac{SSB + r(t-1)MSE_{rcbd}}{rt-1} \\ &= MSB\left(\frac{r-1}{rt-1}\right) + MSE_{rcbd}\left(\frac{r(t-1)}{rt-1}\right) \end{aligned}$$

In general, the effectiveness of blocking is a function of MSE_{crd}/MSE_{rcb}. If this is large, it is worthwhile to block. For the nitrogen example, this ratio is about 2.

4.3 Unbalanced designs

Example: Observational study of 20 fatal accidents.

- Response: y = speed in excess of speed limit
- Recorded sources of variation:
  1. R = rainy (rainy/not rainy),
  2. T = interstate (interstate/two-lane highway),

	cell means		sum	marginal means
	interstate	two-lane		
rainy	15	5	130	13
	$r_{11} = 8$	$r_{12} = 2$	$r_{1\cdot} = 10$	
not rainy	20	10	120	12
	$r_{21} = 2$	$r_{22} = 8$	$r_{1\cdot} = 10$	
sum	160	90	250	
	$r_{\cdot 1} = 10$	$r_{\cdot 2} = 10$	$r_{\cdot\cdot} = 20$	
marginal mean	16	9		$\bar{y}_{\cdots} = 12.5$

Lets naively compute sums of squares based on the decomposition:

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

```
         50          60          70          80
-0.90902778 -0.02569444  0.19097222  0.74375000
```

What linear modeling commands in R will get you the same thing?

```
> options(contrasts=c("contr.sum","contr.poly"))
> fit_full<-lm( y~as.factor(ageg)*as.factor(trt))

> fit_full$coef[2:4]
as.factor(ageg)1 as.factor(ageg)2 as.factor(ageg)3
     -0.90902778      -0.02569444       0.19097222
> fit_full$coef[5:6]
as.factor(trt)1 as.factor(trt)2
      0.5347222       0.5826389
```

Note that the coefficients in the reduced/additive model are **not** the same:

```
> fit_add<-lm( y~as.factor(ageg)+as.factor(trt))
>
> fit_add$coef[2:4]
as.factor(ageg)1 as.factor(ageg)2 as.factor(ageg)3
      -0.7920935       -0.3607554        0.3075743
> fit_add$coef[5:6]
as.factor(trt)1 as.factor(trt)2
     1.207447717      -0.009899274
```


4.3.1 Non-orthogonal sums of squares:

Consider the following ANOVA table obtained from R:

```
                Df  Sum Sq Mean Sq F value  Pr(>F)
as.factor(ageg)  3  13.355   4.452  0.9606 0.42737
as.factor(trt)   2  28.254  14.127  3.0482 0.06613 .
Residuals       24 111.230   4.635
```

It might be somewhat unsettling that R also produces the following table:

```
> anova(  lm( y~as.factor(trt)+as.factor(ageg))  )
                Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(trt)   2  31.588  15.794  3.4079 0.0498 *
as.factor(ageg)  3  10.021   3.340  0.7207 0.5494
Residuals       24 111.230   4.635
```

This is actually what R presents in an ANOVA table:

```
> ss0<-sum(   lm( y~1 )$res^2 )
> ss1<-sum(   lm( y~as.factor(ageg) )$res^2 )
> ss2<-sum(   lm( y~as.factor(ageg)+as.factor(trt) )$res^2 )
> ss3<

> s0-ss1
[1] 13.3554
>
> ss1-ss2
[1] 28.25390
>
> ss2-ss3
[1] 53.75015

> ss3
[1] 57.47955

> anova(  lm( y~as.factor(ageg)*as.factor(trt) ) )
                          Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(ageg)            2 13.355   4.452  1.3941 0.27688
as.factor(trt)             2 28.254  14.127  4.4239 0.02737 *
as.factor(ageg):as.factor(trt)  6 53.750   8.958  2.8054 0.04167 *
Residuals                 18 57.480   3.193
```


Why does order of the variables matter?

- In a balanced design, the parameters are orthogonal, and $SSA = SSA|B$, $SSB = SSB|A$ and so on, so the order doesn't matter.
- In an unbalanced design, the estimates of one set of parameters depends on whether or not you are estimating the others, i.e. they are not orthogonal, and in general $SSA \neq SSA|B$, $SSB \neq SSB|A$.

I will try to draw a picture of this on the board.

The bottom line: For unbalanced designs, there is no "variance due to factor 1" or "variance due to factor 2". There is only "extra variance due to factor 1, beyond that explained by factor 2", and vice versa. This is essentially because of the non-orthogonality, and so the part of the variance

L	A	B
A	B

H	B	A
A	B

H	B	A
A	B

L	B	A
A	B

Randomization:

Sulfur type was randomized to whole plots;

Potato type was randomized to subplots.

Initial data analysis: Sixteen responses, 4 treatment combinations.

- 8 responses for each potato type
- 8 responses for each sulfur type
- 4 responses for each potato×type combination


```
> fit.full<-lm(y~type*sulfur) ;  fit.add<-lm(y~type+sulfur)
> anova(fit.full)
             Df  Sum Sq Mean Sq F value   Pr(>F)
type          1 1.48840 1.48840 13.4459 0.003225 **
sulfur        1 0.54022 0.54022  4.8803 0.047354 *
type:sulfur   1 0.00360 0.00360  0.0325 0.859897
Residuals    12 1.32835 0.11070

> anova(fit.add)
          Df  Sum Sq Mean Sq F value   Pr(>F)
type       1 1.48840 1.48840 14.5270 0.00216 **
sulfur     1 0.54022 0.54022  5.2727 0.03893 *
Residuals 13 1.33195 0.10246
```