STOR 664:
APPLIED STATISTICS I
Instructor: Richard L. Smith

Class Notes:
November 29, 2022

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Class Announcements

- Take-home exam: will be set 6:00 am – 9:00 pm Saturday, December 3, but with a 6-hour time limit
- Make-up exam: 12:00 pm – 6:00 pm Sunday, December 4 (by prior arrangement)
- Past exams with solutions are on course webpage
- **Usual Honor Code rules apply:** no consulting with other class members or any outside person but me
- Review session 5:00 pm – 6:00 pm Thursday, December 1 (room TBA)
- Final assignment due today (gradescope)
- Project also due today (gradescope or email)
- Office hour today: 2:00-3:00 pm (note change of usual time)
- Grades will be announced a.s.a.p. but won't be immediate (please check HW scores on gradescope)
- If I agreed to write a letter of recommendation for you and have not done so, please let me know
- Please complete CAS survey!

Chapter 8: Analysis of Designed Experiments

Basic definitions:

- Units, e.g. people, plots of land, industrial experiments
- Treatments, e.g. medical, fertilizer, temperature of an industrial process
- Blocks: other variables that affect the outcome but are not of direct interest (e.g. in medical studies, sex, age, race, prior medical condition
- *Interactions* arise when treatments perform better in some blocks than others

All involve *factor* (i.e. non-numeric) variables

Typically represent factors as 0–1 variables, e.g.

$$x_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is at level } j \\ 0 & \text{otherwise} \end{cases}$$

Use `model.matrix` to see representation in R

Completely Randomized Experiments (One-Way ANOVA)

Let y_{ij} be jth observation on treatment i, $1 \leq j \leq n_i$, $1 \leq i \leq r$
($n = \sum_{i=1}^{r} n_i$ is total sample size)

Model $y_{ij} = \mu_i + \epsilon_{ij}$ or $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ (independent)

LSE $\widehat{\mu}_i = \bar{y}_{i\cdot} = \frac{\sum_j y_{ij}}{n_i} = \widehat{\mu} + \widehat{\alpha}_i$.

Overdetermined, need a constraint:

- $\sum_i n_i \alpha_i = 0$, leads to $\widehat{\mu} = \frac{\sum_i \sum_j y_{ij}}{n} = \bar{y}_{\cdot\cdot}$, $\widehat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$.
- Set $\widehat{\mu} = 0$, $\widehat{\alpha}_i = \bar{y}_{i\cdot}$
- Fix $\alpha_1 = 0$, $\widehat{\mu} = \bar{y}_{1\cdot}$, $\widehat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{1\cdot}$.
- Last one is default in R but can change this with statements like
  `op = options(contrasts = c("contr.helmert", "contr.poly"))`

ANOVA Table

$$\begin{aligned} SSTO &= \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 \\ &= \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_i n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\ &= SSE + SSTR \\ \text{(DFs :)} \qquad n-1 &= (n-r) + (r-1) \end{aligned}$$

Estimate $s^2 = \frac{SSE}{n-r}$, test the null hypothesis H_0 that all means are equal by

$$F = \frac{SSTR/(r-1)}{SSE/(n-r)} \sim F_{r-1,n-r} \text{ if } H_0 \text{ true.}$$

Reject H_0 at level α if $F > F_{r-1,n-r,1-\alpha}$
(in R: `qf(1-alpha,r-1,n-r)`)

Testing Equality of Variances

Model $y_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2),\ i = 1, \ldots, r,\ j = 1, \ldots, n_i$,
test $H_0:\ \sigma_1^2 = \ldots = \sigma_r^2$

1. **Likelihood Ratio Test**

   Estimate $\hat{\sigma}_i^2 = \frac{\sum_j (y_{ij} - \bar{y}_{i\cdot})^2}{n_i}$, $\hat{\sigma}^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}{n}$, define

   $$T \;=\; 2 \log \frac{L_1}{L_0} \;=\; \sum_{i=1}^{r} n_i \log \frac{\hat{\sigma}^2}{\hat{\sigma}_i^2} \ \sim \chi^2_{r-1} \text{ asymptotically}$$

2. **Bartlett's Modification (1937)**

   (a) Replace n_i by $n_i - 1$, n by $n - r$ in definitions of $\hat{\sigma}_i^2$, $\hat{\sigma}^2$ and T.

   (b) Define $T' = \left\{1 + \frac{1}{3(r-1)} \sum_{i=1}^{r} \left(\frac{1}{n_i - 1} - \frac{1}{n - r}\right)\right\}^{-1} T$

   (c) If H_0 true, $T' \sim \chi^2_{r-1}$ approximately.

Round-robin test data

Laboratory i	n_i	Mean	S.D.	S_i	$\hat{\alpha}_i$	S.E.
1	5	102.1	48.1	9254.44		
2	9	92.8	8.3	551.12		
3	4	97.2	8.6	221.88		
4	5	79.9	9.2	338.56		
5	5	87.0	4.8	92.16		
6	5	93.1	5.5	121.00		
7	5	82.2	4.4	77.44		
8	6	54.9	1.9	18.05		
9	5	94.0	8.3	275.56		
10	5	90.4	2.2	19.36		
11	5	84.7	5.7	129.96		

p-value for equality of variances: 1.3×10^{-12}

p-value for equality of means: 0.0007

Round-robin test data

Laboratory i	n_i	Mean	S.D.	S_i	$\widehat{\alpha}_i$	S.E.
2	9	92.8	8.3	551.12		
3	4	97.2	8.6	221.88		
4	5	79.9	9.2	338.56		
5	5	87.0	4.8	92.16		
6	5	93.1	5.5	121.00		
7	5	82.2	4.4	77.44		
8	6	54.9	1.9	18.05		
9	5	94.0	8.3	275.56		
10	5	90.4	2.2	19.36		
11	5	84.7	5.7	129.96		

p-value for equality of variances: 0.05

p-value for equality of means: 7×10^{-13}

Round-robin test data

Laboratory i	n_i	Mean	S.D.	S_i	$\widehat{\alpha}_i$	S.E.
2	9	92.8	8.3	551.12	3.62	2.06
3	4	97.2	8.6	221.88	8.02	3.28
4	5	79.9	9.2	338.56	−9.28	2.90
5	5	87.0	4.8	92.16	−2.18	2.90
6	5	93.1	5.5	121.00	3.92	2.90
7	5	82.2	4.4	77.44	−6.98	2.90
9	5	94.0	8.3	275.56	4.82	2.90
10	5	90.4	2.2	19.36	1.22	2.90
11	5	84.7	5.7	129.96	−4.48	2.90

p-value for equality of variances: 0.18

p-value for equality of means: 0.003

Conclusions

- We threw out Lab 1 because the SD seemed obviously wrong — either Bartlett or Likelihood Ratio test decisively rejects hypothesis of equal variances

- We then threw out Lab 8 because the mean was discrepant — F-test decisively rejects hypothesis of equal means

- Among the rest, estimated treatment effect is significantly positive for Lab 3, negative for Labs 4 and 7

- However we could develop the last point in more detail with more formal multiple comparisons procedures — Least Significant Differences, Tukey test for pairwise differences, Scheffé test for contrasts (assuming equal variances)

Two-way ANOVA Without Interactions

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \ 1 \le i \le r, \ 1 \le j \le c.$$

Assume $\sum_i \alpha_i = \sum_j \beta_j = 0$ (but default in R is $\alpha_1 = \beta_1 = 0$)

Equality of treatments $H_0 : \ \alpha_1 = \ldots = \alpha_r = 0$

Equality of blocks $H_0' : \ \beta_1 = \ldots = \beta_c = 0$

Typically, H_0 is of interest but H_0' is not

ANOVA decomposition:

$$\begin{aligned}
\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 &= \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 + c \sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + r \sum_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 \\
&= \sum_i \sum_j (y_{ij} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j)^2 + c \sum_i \widehat{\alpha}_i^2 + r \sum_j \widehat{\beta}_j^2, \\
SSTO &= SSE + SSTR + SSB \\
rc - 1 &= (r-1)(c-1) + (c-1) + (r-1)
\end{aligned}$$

F test for H_0:

$$\frac{SSTR/(c-1)}{SSE/((r-1)(c-1))} \sim F_{c-1,(r-1)(c-1)} \text{ if } H_0 \text{ true.}$$

Two-way ANOVA With Interactions

Assume $t > 1$ observations for each treatment-block pair

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \ 1 \le i \le r, \ 1 \le j \le c, \ 1 \le k \le t.$$

Assume $\sum_i \alpha_i = \sum_j \beta_j = 0$, $\sum_j \gamma_{ij} = 0$ for each i, $\sum_i \gamma_{ij} = 0$ for each j.

$\widehat{\mu} = \bar{y}_{\cdots}$, $\widehat{\alpha}_i = \bar{y}_{i\cdot\cdot}$, $\widehat{\beta}_j = \bar{y}_{\cdot j\cdot}$, $\widehat{\gamma}_{ij} = \bar{y}_{ij\cdot}$.

ANOVA decomposition becomes

$$\begin{aligned} SSTO &= SSE + SSI + SSTR + SSB \\ rct - 1 &= rc(t-1) + (r-1)(c-1) + (c-1) + (r-1) \end{aligned}$$

F test for no treatment effect:

$$\begin{aligned} \frac{SSTR/(c-1)}{SSE/(rc(t-1)} &\sim F_{c-1,rc(t-1)} \text{ if no treatment effect} \\ \frac{SSI/((r-1)(c-1))}{SSE/(rc(t-1)} &\sim F_{(r-1)(c-1),rc(t-1)} \text{ if no interaction.} \end{aligned}$$

What if $t = 1$?

Tukey's 1DF Test for Additivity

Consider model

$$y_{ij} \;=\; \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j + \epsilon_{ij},\ 1 \le i \le r,\ 1 \le j \le c.$$

Assume $\sum_i \alpha_i = \sum_j \beta_j = 0$, test $H_0:\ \theta = 0$ against $H_1:\ \theta \neq 0$.
Define $z_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}$, then model

$$z_{ij} = \theta a_i b_j + e_{ij},\ e_{ij} \text{ random error}, a_i, b_j \text{ known s.t. } \sum_i a_i = \sum_j b_j = 0.$$

Estimate $\widehat{\theta} = \frac{\sum_i \sum_j z_{ij} a_i b_j}{\sum_i a_i^2 \cdot \sum_j b_j^2} = \frac{\sum_i \sum_j y_{ij} a_i b_j}{\sum_i a_i^2 \cdot \sum_j b_j^2}$, $\mathrm{Var}(\widehat{\theta}) = \frac{\sigma^2}{\sum_i a_i^2 \cdot \sum_j b_j^2}$.

Under H_0, $\frac{\widehat{\theta}^2 \sum_i a_i^2 \sum_j b_j^2}{\sigma^2} = \frac{(\sum_i \sum_j y_{ij} a_i b_j)^2}{\sigma^2 \sum_i a_i^2 \sum_j b_j^2} \sim \chi_1^2$.

Tukey's 1DF Test for Additivity, Page 2

ANOVA decomposition

$$\begin{aligned}
\sum_i \sum_j z_{ij}^2 &= \sum_i \sum_j (z_{ij} - \widehat{\theta} a_i b_j)^2 + \widehat{\theta}^2 \sum_i a_i^2 \sum_j b_j^2, \\
SSI &= SSIE + SSG, \\
(r-1)(c-1) &= (rc - r - c) + 1.
\end{aligned}$$

Calculations show SSG, $SSIE$ are statistically independent (not trivial). Hence, if H_0 true,

$$\frac{SSG}{SSIE/(rc-r-c)} \sim F_{1,rc-r-c} \qquad (*)$$

Now comes the key step: *All this is true for any choices of a_i, b_j, therefore, in particular, it's true if we take $a_i = \widehat{\alpha}_i$, $b_j = \widehat{\beta}_j$.*

With this substitution, (*) gives an exact test.

Fisher's data on barley varieties

Place	Year	Manchuria	Svansota	Velvet	Trebi	Peatland	Row Mean
1	1931	81.0	105.4	119.7	109.7	98.3	102.82
1	1932	80.7	82.3	80.4	87.2	84.2	82.96
2	1931	146.6	142.0	150.7	191.5	145.7	155.30
2	1932	100.4	115.5	112.2	147.7	108.1	116.78
3	1931	82.3	77.3	78.4	131.3	89.6	91.78
3	1932	103.1	105.1	116.5	139.9	129.6	118.84
4	1931	119.8	121.4	124.0	140.8	124.8	126.16
4	1932	98.9	61.9	96.2	125.5	75.7	91.64
5	1931	98.9	89.0	69.1	89.3	104.1	90.08
5	1932	66.4	49.9	96.7	61.9	80.3	71.04
6	1931	86.9	77.1	78.9	101.8	96.0	88.14
6	1932	67.7	66.7	67.4	91.8	94.1	77.54
Col Mean		94.392	91.133	99.183	118.200	102.542	101.09

Two models considered here:

1. Two-way ANOVA with interactions, $t = 2$ observations for each treatment-place combination (but ignoring possible year to year variation)

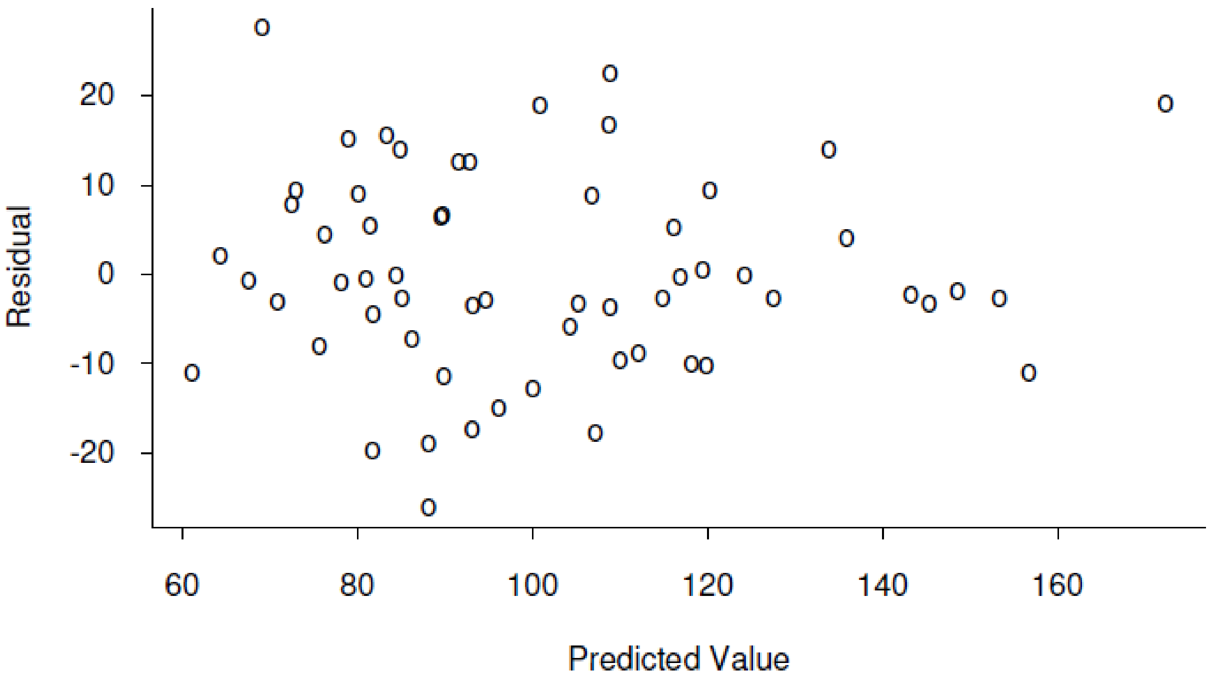
2. Treat each place $\times$ year combination as a block, so we have 5 treatments, 12 blocks, 1 observation for each treatment-place combination, but apply Tukey test for interaction

ANOVA Table for 2-way model with interactions (F-ratio for SSI is 0.48, not significant)

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE
SST	5309.97	4	**1327.5**
SSB	21220.90	5	4244.2
SSI	4433.02	20	221.7
SSE	13768.46	30	**458.9**
Total	44732.35	59	**F-ratio 2.89**

ANOVA Table for Tukey's 1-DF test (F-ratio for SSG is 3.27, p=0.077)

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE
SST	5309.97	4	**1327.5**
SSB	31913.32	11	2901.2
SSG	531.09	1	531.1
$SSIE$	6977.97	43	**162.3**
Total	44732.35	59	**F-ratio 8.18**



Plot of residuals vs. fitted values for barley data, 2-way model without interactions

Conclusions

- First model inadequate — ignores year to year variation, which masks the treatment effect.
- Second model seems OK — Tukey test accepts hypothesis of no interaction but the treatment effect *is* significant.
- However there are other possible models, e.g. model year effect explicitly as a 3-way ANOVA; make either the block effect or the interaction (or both) a random effect.
- Could also use Tukey multiple comparisons procedure to determine which pairwise treatment differences are significant.