Lecture 10

- Tests for homogenity of variance
- ANOVA remedial measures
- Two-way ANOVA

Homogeneity tests

- Homogeneity of variance (homoscedasticity)
- H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2$
- H_1: not all σ_i^2 are equal
- Several significance tests are available

Homogeneity tests (2)

- Text discusses Hartley and Levene

Homogeneity tests (3)

- There is a problem with assumptions
  - Anova is robust with respect to moderate deviations from normality
  - Anova results can be sensitive to the homogeneity of variance assumption
- Some homogeneity tests are sensitive to the normality assumption

Levene's Test

- Do anova on the absolute values of the residuals

Example

- NKNW p 765
- Compare the strengths of 5 types of solder flux (A has I=5 levels)
- Response variable is the pull strength, force in pounds required to break the joint
- There are 8 solder joints per flux (J=8)

Levene’s Test

```
flux<-
read.table('ch18ta02.txt',
col.names=c("strength",
"flux", "ind"));
flux$flux<-factor(flux$flux);
library(car);
leveneTest(flux$strength,
flux$flux, center=median);
```

Output

```
Levene's Test
      Df F value  Pr(>F)
group  4  2.9358 0.03414 *
      35
```

SDs

```
sd1<-ave(flux$strength,
flux$flux, FUN=sd)

[1]  1.2371396
[9]  1.2529708
[17] 2.4866440
[25] 0.8166034
[33] 0.7694154
```

Remedies

- **Delete outliers**
- **Use weights**
- **Transformations**
- **Nonparametric procedures**

Weighted least squares

- **We used this with regression**
  - **Obtained a model for how the sd depended on the explanatory variable (plotted absolute value of residual vs x)**
  - **Then used weights inversely proportional to the estimated variance**

Weighted LS (2)

- **Here we can compute the variance for each level**
- **Use these as weights in aov or lm**
- **We will illustrate with the soldering example**

Weighted ANOVA

```
wt<-1/sd1^2;
obj<-aov(strength~flux,
weight=wt, flux)
summary.aov(obj)
```

Output

```
    Df  SS    MS  F value Pr(>F)
flux 4 324.2 81.05 81.05< 2.2e-16
Res 35  35.0  1.00
```

Transformation Guides

- When σ_i^2 is proportional to μ_i, use
- $\sqrt{Y} + \sqrt{Y+1}$
- When σ_i is proportional to μ_i, use log(y)
- When σ_i is proportional to μ_i^2, use 1/y
- For proportions, use 2arcsin $\sqrt{Y}$

Nonparametric approach

- Based on ranks
- kruskal.test

- Two-way ANOVA
  - Data, model, parameter estimates
- Factor effects model
- Anova table with tests for main effects and interaction

Data

- For Y_{ijk} we use
  - i to denote the level of the factor A
  - j to denote the level of the factor B
  - k to denote the k^{th} observation in cell (i,j)
- i = 1, . . . , I levels of factor A
- j = 1, . . . , J levels of factor B
- k = 1, . . . , K observations in cell (i,j)

Cell means model

- $Y_{ijk} = \mu_{ij} + \xi_{ijk}$
  - where μ_{ij} is the theoretical mean or expected value of all observations in cell (i,j)
  - the ξ_{ijk} are iid $N(0, \sigma^2)$
  - $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$, independent

Parameters

- The parameters of the model are
  - μ_{ij}, for i = 1 to I and j = 1 to J
  - σ^2

Estimates

- Estimate μ_{ij} by the mean of the observations in cell (i,j), $\bar{Y}_{ij}$
- $\bar{Y}_{ij} = (\Sigma_k Y_{ijk})/K$
- For each (i,j) combination, we can get an estimate of the variance
- $s_{ij}^2 = (\Sigma(Y_{ijk} - \bar{Y}_{ij})^2)/(K-1)$
- We need to combine these to get an estimate of σ^2

Pooled estimate of σ^2

- In general we pool the s_{ij}^2, giving weights proportional to the df, K_{ij} -1
- The pooled estimate is
- $s^2 = (\Sigma\ (K_{ij}-1)s_{ij}^2) / (\Sigma\ (K_{ij}-1))$
- Here, $K_{ij} = K$, so
- $s^2 = (\Sigma s_{ij}^2) / (IJ)$

Factor effects model

- For the one-way anova model, we wrote $\mu_i = \mu + \alpha_i$
- Here we use $\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$

Constraints

- $\Sigma_i\ \alpha_i = 0$

  $\Sigma_j \beta_j = 0$
- $\Sigma_i\ \alpha\beta_{ij} = 0$ for all j
- $\Sigma_j\ \alpha\beta_{ij} = 0$ for all i

Factor effects model (2)

- $\mu = (\Sigma_{ij} \mu_{ij})/(IJ)$
- $\mu_{i.} = (\Sigma_j \mu_{ij})/J$
- $\mu_{.j} = (\Sigma_i \mu_{ij})/I$
- $\alpha_i = \mu_{i.} - \mu$
- $\beta_j = \mu_{.j} - \mu$
- $\alpha\beta_{ij}$ is difference between μ_{ij} and $\mu + \alpha_i + \beta_j$
- $\alpha\beta_{ij} = \mu_{ij} - (\mu + (\mu_{i.} - \mu) + (\mu_{.j} - \mu))$
- $= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$

Interpretation

- $\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$
- μ is average of means
- α_i is an adjustment for level i of A
- β_j is an adjustment for level j of B
- $\alpha\beta_{ij}$ is an additional adjustment that takes into account both i and j

Estimates for Factor effects model

$\hat{\mu} = Y_{...} = (\Sigma_{ijk} Y_{ijk})/(IJK)$

$\hat{\mu}_{i.} = Y_{i..} = (\Sigma_{jk} Y_{ijk})/(JK)$

$\hat{\mu}_{.j} = Y_{.j.} = (\Sigma_{ik} Y_{ijk})/(IK)$

- $\hat{\alpha}_i = \hat{\mu}_{i.} - \hat{\mu} = Y_{i..} - Y_{...}$
- $\hat{\beta}_j = \hat{\mu}_{.j} - \hat{\mu} = Y_{.j.} - Y_{...}$
- $\widehat{\alpha\beta}_{ij} = \hat{\mu}_{ij} - \hat{\mu}_{i.} - \hat{\mu}_{.j} + \hat{\mu} =$
- $Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$

SS for ANOVA Table

- $SSA = \Sigma_{ijk}(\hat{\alpha}_i)^2 = \Sigma_{ijk}(Y_{i..} - Y_{...})^2$
- $SSB = \Sigma_{ijk}(\hat{\beta}_j)^2 = \Sigma_{ijk}(Y_{.j.} - Y_{...})^2$
- $SSAB = \Sigma_{ijk}(\widehat{\alpha\beta}_{ij})^2 = \Sigma_{ijk}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$
- $SSE = \Sigma_{ijk}(Y_{ijk} - Y_{ij.})^2$
- $SST = \Sigma_{ijk}(Y_{ijk} - Y_{...})^2$

df for ANOVA Table

- dfA = I-1
- dfB = J-1
- dfAB = (I-1)(J-1)
- dfE = IJ(K-1)
- dfT = IJK-1 = n-1

MS for ANOVA Table

- MSA = SSA/dfA
- MSB = SSB/dfB
- MSAB = SSAB/dfAB
- MSE = SSE/dfE
- MST = SST/dfT

Hypotheses for two-way ANOVA

- H_{0A}: $\alpha_i = 0$ for all I
- H_{1A}: $\alpha_i \neq 0$ for at least one i
- H_{0B}: $\beta_j = 0$ for all j
- H_{1B}: $\beta_j \neq 0$ for at least one j
- H_{0AB}: $\alpha\beta_{ij} = 0$ for all (i,j)
- H_{1AB}: $\alpha\beta_{ij} \neq 0$ for at least one (i,j)

F statistics

- H_{0A} is tested by F_A = MSA/MSE; df=dfA, dfE
- H_{0B} is tested by F_B = MSB/MSE; df=dfB, dfE
- H_{0AB} is tested by F_{AB} = MSAB/MSE; df=dfAB, dfE

ANOVA Table

Source	df	SS	MS	F
A	I-1	SSA	MSA	MSA/MSE
B	J-1	SSB	MSB	MSB/MSE
AB	(I-1)(J-1)	SSAB	MSAB	MSAB/MSE
Error	IJ(K-1)	SSE	MSE	_
Total	IJK-1	SST	MST	

P-values

- P-values are calculated using the F(dfNumerator, dfDenominator) distributions
- If P $\leq$ 0.05 we conclude that the effect being tested is statistically significant

Example

- Y is the number of cases of bread sold
- A is the height of the shelf display, I=3 levels: bottom, middle, top
- B is the width of the shelf display, J=2: regular, wide
- K=2 stores for each of the 3x2 treatment combinations

ANOVA

```
bread<-read.table('ch19ta07.txt',
col.names=c("cases", "height",
"width", "store"));
bread$height<-factor(bread$height);
bread$width<-factor(bread$width);
obj<-aov(cases~height*width, bread);
summary(obj)
```

Output

```
Df Sum Sq Mean Sq F value    Pr(>F)
height 2 1544 772.0 74.71 5.754e-05
width  1   12  12.00 1.16 0.3226
h:wid  2   24  12.00 1.16 0.3747
Res    6   62  10.33
```

```
Note that there are 6 cells in
This design.
```

Output from lm

```
Residual standard error: 3.215
       on 6 degrees of freedom
Multiple R-squared: 0.9622
F-statistic: 30.58 on 5 and 6 DF,
p-value: 0.0003384
```

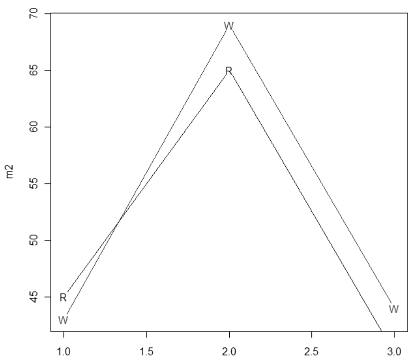
Results

- The main effect of height is statistically significant (F=74.71; df=2,6; P<0.0001)
- The main effect of width is not statistically significant (F=1.16; df=1,6; P=0.32)
- The interaction between height and width is not statistically significant (F=1.16; df=2,6; P=0.37)

Interpretation

- The height of the display affects sales of bread
- The width of the display has no apparent effect
- The effect of the height of the display is similar for both the regular and the wide widths

Plot of the means



Additional analyses

- We will need to do additional analyses to explain the height effect (factor A)
- There were three levels: bottom, middle and top
- We could rerun the data with a one-way anova and use the methods we learned in the previous chapters

R LM Constraints

- $\alpha_1 = 0$ (1 constraint)
- $\beta_1 = 0$ (1 constraint)
- $\alpha\beta_{1j} = 0$ for all j (J constraints)
- $\alpha\beta_{i1} = 0$ for all i (I constraints)
- The total is 1+1+I+J-1=I+J+1 (the constraint $\alpha\beta_{11}$ is counted twice above

Parameters and constraints

- The cell means model has IJ parameters for the means
- The factor effects model has (1+I+J+IJ) parameters
  - An intercept (1)
  - Main effect of A (I)
  - Main effect of B (J)
  - Interaction of A and B (IJ)

Factor effects model

- There are 1+I+J+IJ parameters
- There are 1+I+J constraints
- There are IJ unconstrained parameters (or sets of parameters), the same number of parameters for the means in the cell means model

Solution output

```
obj2<
lm(cases~height*width,
bread); summary(obj2);
    Est Sd    t   Pr(>|t|)
Int 45.0 2.3 19.8 1.08e-06
ht2 20.0 3.2  6.2 0.000797
ht3 -5.0 3.2 -1.5 0.170844
wd2 -2.0 3.2 -0.6 0.556718
h2w2 6.0 4.5  1.3 0.235013
h3w2 6.0 4.5  1.3 0.235013
```

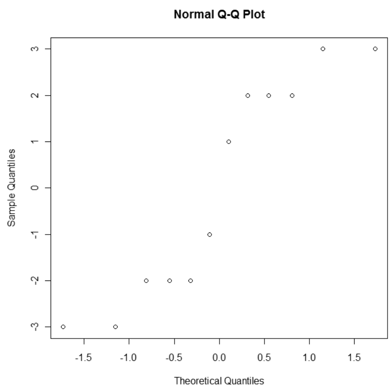
Means

height	width	Mean
1	1	45=45
1	2	43=45-2
2	1	65=45+20
2	2	69=45+20-2+6
3	1	40=45-5
3	2	44=45-5-2+6

Check the normal assumption

```
r<-residuals(obj2);
qqnorm(r);
```

The plot



ANOVA Table

Source	df	SS	MS	F
A	I-1	SSA	MSA	MSA/MSE
B	J-1	SSB	MSB	MSB/MSE
AB	(I-1)(J-1)	SSAB	MSAB	MSAB/MSE
Error	IJ(K-1)	SSE	MSE	_
Total	IJK-1	SST	MST	

Expected Mean Squares

- $E(MSE) = \sigma^2$
- $E(MSA) = \sigma^2 + KJ(\Sigma_i \alpha_i^2)/(I-1)$
- $E(MSB) = \sigma^2 + KI(\Sigma_j \beta_j^2)/(J-1)$
- $E(MSAB) = \sigma^2 + K(\Sigma_{ij} \alpha\beta_{ij}^2)/((I-1)(J-1))$
- Here, α_i, β_j, and $\alpha\beta_{ij}$ are defined with the usual factor effects constraints

An analytical strategy

- Run the model with main effects and the two-way interaction
- Plot the data, the means and look at the normal quantile plot
- Check the significance test for the interaction

AB interaction ns

- If the AB interaction is not statistically significant
  - Rerun the analysis without the interaction
  - For a main effect with more than two levels that is significant, use the means statement with the tukey multiple comparison procedure

Rerun without interaction

```
obj3<-aov(cases~height+width,
bread);
summary(obj3)
TukeyHSD(obj3)$height
```

Anova output

	Df	SS	MS	F	Pr(>F)
ht	2	1544	772.0	71.8	7.749e-06
wd	1	12	12.0	1.1	0.3216
Res	8	86	10.75		

MSh and MSw have not changed, MSE, F's, and P-values have

Comparison of MSEs

Model with interaction

Error 6 62 10.33

Model without interaction

Error 8 86 10.75

Tukey Output

	diff	lwr	upr	p adj
2-1	23	16.4	29.6	2.36e-05
3-1	-2	-8.6	4.6	6.77e-01
3-2	-25	-31.6	-18.4	1.26e-05

Regression Approach

- **Similar to what we did for one-way**
- **Use I-1 variables for A**
- **Use J-1 variables for B**
- **Multiply each of the I-1 for A times each of the J-1 for B to get (I-1)(j-1) for AB**

Pooling SS

- **Data = Model + Residual**
- **When we remove a term from the \`model', we put this variation and the associated df into \`residual'**
- **This is called pooling**
- **A benefit is that we have more df for error and a simpler model**

Pooling SSE and SSAB

- **For model with interaction**
  - **SSAB=24, dfAB=2**
  - **SSE=62, dfE=6**
  - **MSE=10.33**
- **For the model with main effects only**
  - **SSE=62+24=86, dfE=6+2=8**
  - **MSE=10.75**