Problem 8.1 of GO

Data: A completely randomized 2×5 factorial trial on barley germination. Lots of 100 seeds were randomly allocated to 10 conditions, determined from the amount of water used in the germination (values $4, 8$ ml) and the harvest time after germination (values $1, 3, 6, 9, 12$ weeks), and there were 3 replicates for each condition. The outcome was the number of seeds that germinated (out of 100); we could equivalently use the germination rate obtained by dividing by 100.

Model: The statistical model of interest is two-way ANOVA model:

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad \text{or} \\ \text{Seed count} &= \text{Water} + \text{Age} + \text{Water} * \text{Age} + \text{Error}, \end{aligned}$$

where

- μ = overall mean,
- α_i = effect of water group i,
- β_j = effect of age group j,
- $(\alpha\beta)_{ij}$ = interaction effect of (water,age) group (i, j),
- ε_{ijk} = error for obs. (i, j, k), assumed to be i.i.d. and $\mathrm{N}(0, \sigma^2)$.

Our first step is to run the full two-way ANOVA model with the relevant model checks.

General Linear Model: seeds versus age, water

Factor Information

Factor	Type	Levels	Values
age	Fixed	5	1, 3, 6, 9, 12
water	Fixed	2	4, 8

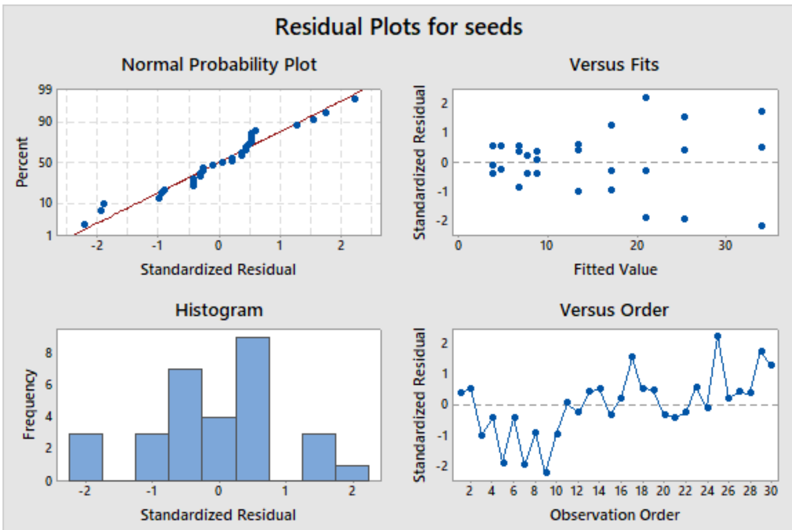
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
age	4	1321.1	330.28	5.53	0.004
water	1	1178.1	1178.13	19.72	0.000
age*water	4	208.9	52.22	0.87	0.497
Error	20	1194.7	59.73		
Total	29	3902.8			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.72873	69.39%	55.61%	31.13%

Residual Plots for seeds



The residual plot shows a clear fanning shape with larger variances at higher fitted values. This is also quite visible in the data table. Nor should it be unexpected as the counts could be assumed to follow a binomial distribution, in which the variance decreases towards the ends of the probability interval. Here the proportions do not exceed 0.5, so the only issue is with low counts/proportions.

Our usual approach to resolve such a problem is to apply a suitable transformation to the outcome. We can either try a power transformation selected by the Box-Cox method or the "standard" approach for proportion data ($y \mapsto \text{arsin}(\sqrt{\hat{p}})$), see Lecture 1a or Table 6.3 of GO. The Box-Cox analysis gives $\hat{\lambda} = 0.357$ with a 95% CI of $(0.083, 0.65)$ (from Minitab, the Stata CI equals $(0.077, 0.64)$). From this it seems reasonable to try a square-root transformation (i.e., $\lambda = 0.5$). We show the results below; note that because the proportions are all reasonably small in magnitude the arcsine-squareroot transform gives very similar results (not shown, but one needs to try it to make such a statement!).

General Linear Model: rootseed versus age, water

Factor Information

Factor	Type	Levels	Values
age	Fixed	5	1, 3, 6, 9, 12
water	Fixed	2	4, 8

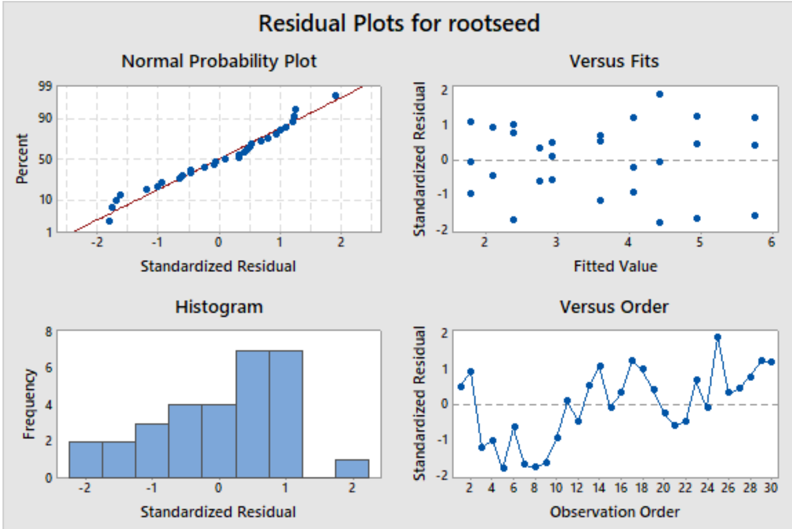
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
age	4	21.895	5.4737	5.94	0.003
water	1	21.893	21.8930	23.76	0.000
age*water	4	2.248	0.5621	0.61	0.660
Error	20	18.428	0.9214		
Total	29	64.464			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.959896	71.41%	58.55%	35.68%

Residual Plots for rootseed

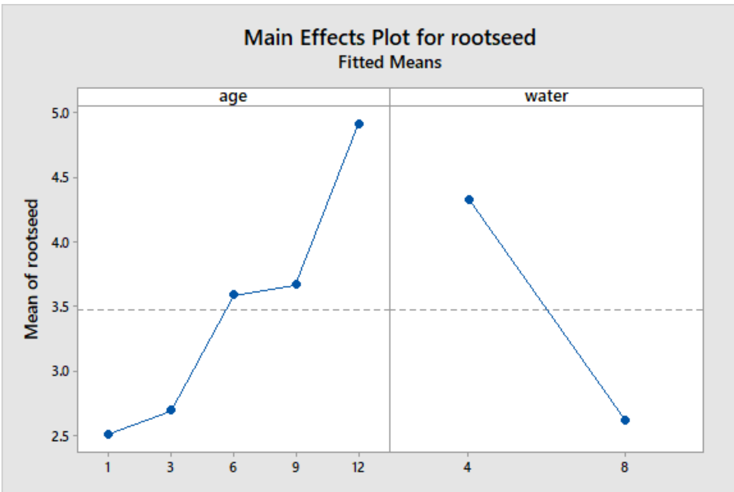


The residual plot is much improved, and we may consider the compliance with model assumptions as satisfactory. The ANOVA table shows no evidence of an interaction ($F = 0.61, P = 0.66$), so the two factors can be assessed separately (on square-root scale). Both are strongly significant, giving us evidence of effects of both water and age on germination.

Presentation of effects on transformed (square-root) scale is relatively straightforward in the balanced design where the estimates are simple means and standard errors are the same across factor levels. We already knew from the ANOVA table about the significance difference between the two water levels, so no further comparisons are necessary.

Means

Term	Fitted Mean	SE Mean
age		
1	2.510	0.392
3	2.691	0.392
6	3.585	0.392
9	3.663	0.392
12	4.909	0.392
water		
4	4.326	0.248
8	2.617	0.248



The germination increases with age, so it is perhaps of more interest to fit a functional relation than carry out pairwise comparisons. For the latter, the (unadjusted) LSD-value is computed as $\mathrm{LSD}_{.95} = t^* \sqrt{\mathrm{MSE}/6} = 0.82$. Due to the age values not being exactly equidistant, polynomial contrasts are not useful. Instead we can fit polynomial models, here shown for with linear and quadratic terms (for simplicity without the interaction term).

General Linear Model: rootseed versus age, water

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
water	1	21.893	33.96%	21.893	21.8930	26.60	0.000
age	1	20.346	31.56%	20.346	20.3464	24.72	0.000
Error	27	22.225	34.48%	22.225	0.8231		
Lack-of-Fit	7	3.797	5.89%	3.797	0.5424	0.59	0.757
Pure Error	20	18.428	28.59%	18.428	0.9214		
Total	29	64.464	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.907274	65.52%	62.97%	27.2508	57.73%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.185	0.307	(1.555, 2.816)	7.11	0.000	
water						
4	0.854	0.166	(0.514, 1.194)	5.16	0.000	1.00
age	0.2074	0.0417	(0.1218, 0.2931)	4.97	0.000	1.00

General Linear Model: rootseed versus age, water

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
water	1	21.8930	33.96%	21.8930	21.8930	26.04	0.000
age	1	20.3464	31.56%	0.2186	0.2186	0.26	0.614
age*age	1	0.3632	0.56%	0.3632	0.3632	0.43	0.517
Error	26	21.8617	33.91%	21.8617	0.8408		
Lack-of-Fit	6	3.4337	5.33%	3.4337	0.5723	0.62	0.711
Pure Error	20	18.4280	28.59%	18.4280	0.9214		
Total	29	64.4643	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.916970	66.09%	62.17%	28.9032	55.16%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.418	0.471	(1.450, 3.386)	5.14	0.000	
water						
4	0.854	0.167	(0.510, 1.198)	5.10	0.000	1.00
age	0.092	0.181	(-0.279, 0.463)	0.51	0.614	18.33
age*age	0.0089	0.0136	(-0.0189, 0.0368)	0.66	0.517	18.33

It is seen that the quadratic term does not contribute significantly (nor substantially) to the fit, so a linear effect for age seems most appropriate as it furthermore shows no indication of lack of fit. The estimated regression slope is $\hat{\beta}(\text{age}) = 0.207$, corresponding to the weekly increase in (square-root) germination. For prediction on original scale, we need to output the predicted values for each combination of water and age, and backtransform those values by squaring them. Confidence intervals can be backtransformed as well, if desired (but omitted in the graph below).