Analysis of Keener/Bashor data on
Bacterial counts for washer-treated chickens

Jason A. Osborne, Dept. of Statistics, NCSU, May 2002

Bacterial counts y were taken on a number of chickens from two plants (Goldkist, Perdue) over three seasons: Fall (Perdue), Winter (Goldkist) and Spring (Perdue) for periods of three days within each season. Ten chickens were measured at each of 3 or 4 (Goldkist only) locations within the plants for periods of three days. The total sample sizes by season were 90,120 and 90 chickens in Fall, Winter and Spring.

The response variable is bacterial count (y). As is often the case with count data, the variability of responses appears to increase with the mean, as seen by inspection of plots of residuals against predicted values or examination of means and standard deviations (shown rounded to nearest 1000 on page 4). The square root transformation can be taken to reduce this heterogeneity of variances, but it doesn't seem to cure it. So, the observations were transformed by a more extreme exponent, the 5^{th} root for the purpose of tests involving location effects: $\tilde{y} = y^{0.2}$. One other issue which will have to be addressed when the final analysis is carried out is the modality at 0. Some of the measurements came out negative, suggesting the absence of any bacteria. This results in a discreteness of the response variable which causes problems for general linear modelling. One option will be to consider *zero-inflated* models, which have been developed considerably in recent statistical literature.

For a given plant/season in the experiment, 10 chickens were sampled at each of three locations on each of three days. So, multiple observations were made on the same day. Because day-to-day variability in chickens that the plants process is possible, day was treated as a random effect in the following repeated measures model:

$$Y_{ijk} = \mu + \alpha_i + D_j + (\alpha D)_{ij} + E_{ijk}$$

where μ denotes an overall average bacterial count, α_i denote fixed location effects, D_j denote possible random day effects, D_{ij} denote possible random day-by-location random interaction effects and E_{ijk} denotes departure from the mean of chicken k at location-day combination ij. The random effects have error components denotes as follows:

$$\begin{aligned} \mathrm{Var}[D_j] &= \sigma^2_D \\ \mathrm{Var}[(\alpha D)_{ij}] &= \sigma^2_{\alpha D} \\ \mathrm{Var}[E_{ijk}] &= \sigma^2 \end{aligned}$$

The ANOVA table for each season breaks down the sources of variation in bacterial counts on page 2. The table on page 3 shows that all estimated variance components for the random day effects are negative. This suggests that allowing for intraday correlation via a repeated measures model is unnecessary. So, the classical fixed effects analysis was used for subsequent investigation of effects. The p-values from such an analysis can be found in the ANOVA tables on page 2. SAS code appears on the last page.

The GLM Procedure

Season=Fall

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	200350.2701	25043.7838	3.64	0.0011
Error	81	556830.5317	6874.4510		
Corrected Total	89	757180.8019			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Location	2	177035.3831	88517.6915	12.88	<.0001
day	2	3984.2999	1992.1500	0.29	0.7492
day*Location	4	19330.5872	4832.6468	0.70	0.5922

Season=Spring

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	332885.8924	41610.7365	7.80	<.0001
Error	81	432332.6172	5337.4397		
Corrected Total	89	765218.5096			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Location	2	329522.3097	164761.1549	30.87	<.0001
day	2	327.9218	163.9609	0.03	0.9698
day*Location	4	3035.6608	758.9152	0.14	0.9659

Season=Winter

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	873309.326	79391.757	9.52	<.0001
Error	108	900954.612	8342.172		
Corrected Total	119	1774263.938			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Location	3	846461.8031	282153.9344	33.82	<.0001
day	2	553.5269	276.7634	0.03	0.9674
day*Location	6	26293.9964	4382.3327	0.53	0.7880

```
------------------------------ Season=Fall ------------------------------

                Variance Components Estimation Procedure

                         MIVQUE(0) Estimates

                    Variance Component           sy

                    Var(day)                -94.68323
                    Var(day*Location)      -204.18042
                    Var(Error)                 6874.5

----------------------------- Season=Spring -----------------------------

                    Variance Component           sy

                    Var(day)                -19.83181
                    Var(day*Location)      -457.85245
                    Var(Error)                 5337.4

----------------------------- Season=Winter -----------------------------

                    Variance Component           sy

                    Var(day)               -102.63923
                    Var(day*Location)      -395.98396
                    Var(Error)                 8342.2
```

Multiple comparisons among location means

For each season, the hypothesis of equal mean bacterial counts for all 3 or 4 locations is implausible (p¡0.01). Tukey's procedure was used to carry out pairwise comparisons among the location means within each season. Results are summarized on page 4. The raw bacterial counts are reported here, though the power transformed counts were used for the pairwise comparisons. The p-values for all pairwise comparisons are given in matrix form. Inspection of this matrix indicates that for Fall and Spring, the mean bacterial count for location 1 is significantly greater than that of locations 2 and 3. For the winter, the mean bacterial count for locations 1 and 2 are significantly higher than that of locations 3 and 4.

Adjustment for Multiple Comparisons: Tukey

------------------------------ Season=Fall ------------------------------

Location	Results LSMEAN	LSMEAN Number	
1	50470.0000	1	(sd=38000)
2	17683.3333	2	(sd=14000)
3	13633.3333	3	(sd=10000)

Least Squares Means for effect Location
Pr > |t| for H0: LSMean(i)=LSMean(j)

i/j	1	2	3
1		0.0004	<.0001
2	0.0004		0.7678
3	<.0001	0.7678	

------------------------------ Season=Spring ------------------------------

Location	Results LSMEAN	LSMEAN Number	
1	53833.3333	1	(sd=27000)
2	12900.0000	2	(sd=8000)
3	7133.3333	3	(sd=5000)

Least Squares Means for effect Location
Pr > |t| for H0: LSMean(i)=LSMean(j)

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		0.2635
3	<.0001	0.2635	

------------------------------ Season=Winter ------------------------------

Location	Results LSMEAN	LSMEAN Number	
1	77083.3333	1	(sd=86000)
2	41623.3333	2	(sd=28000)
3	3833.3333	3	(sd=2500)
4	2633.3333	4	(sd=2600)

Least Squares Means for effect Location
Pr > |t| for H0: LSMean(i)=LSMean(j)

i/j	1	2	3	4
1		0.1366	<.0001	<.0001
2	0.1366		<.0001	<.0001
3	<.0001	<.0001		0.9165
4	<.0001	<.0001	0.9165	

```
options ls=75 nodate;

data one;
   infile "recent2.txt" firstobs=2;
   /* note changes to dataset:  identification of outliers as
      misrecordings and mislabelled treatment combinations,
      after phone conversation with Bashor, May 8, 2002 */
   input Season $ Plant $ Rep Location Results;
   day=rep;
   sy=sqrt(results);
run;

proc sort;  by season; run;

proc glm;
   by season;
   class day location;
   *model results sy=location|day;
   model sy=location|day;
   lsmeans location/pdiff adj=tukey;
   output out=two p=p r=r;
run;

proc varcomp;
   by season;
   class day location;
   model sy=location|day/fixed=1;
run;

proc gplot;
   by season;
   plot r*p;
run;

proc univariate normal plot;
   by season;
   var r;
   histogram r;
run;
```