START Mon 2/21/22 (Week 6, Lecture 15)

HANDOUT # 5

Evaluation of Model Conditions

1. Model Conditions
2. Evaluation of Normality: Graphical and Shapiro-Wilk Test
3. Evaluation of Equal Variances: Brown-Forsythe Version of Levene
4. Example: Crab Data
5. Transformations: Power and Box-Cox
6. Rank-Based Procedures: Kruskal-Wallis
7. Generalized Linear Model
8. Evaluation of Independence: Autocorrelation
9. Runs Test and Durbin-Watson Tests
10. Impact of Correlation on C.I. and Test Statistics
11. Supplemental Reading: Design & ANOVA Book - Chapter 5

Model Conditions

In the linear models that we will be using throughout this course, the basic conditions imposed on the model

$$y_{ij} = \mu_i + e_{ij} \quad \text{for } i = 1, 2, \ldots, t; \quad j = 1, 2, \ldots, n_i :$$

- e_{ij}'s are iid $N(0, \sigma_e^2)$ which implies

C1. The t treatment populations have a normal distribution

C2. The t treatment populations have the same standard deviation σ_e

C3. The experiments are conducted so that the observed data values are independent

If the above conditions are violated, then the inferences that we conduct assuming that the model conditions are valid would yield invalid p-values, tests with incorrect power values (hence incorrect Type I and II error rates), confidence intervals that are either too wide or too narrow and hence have incorrect coverage probabilities. Thus, it is imperative that a careful evaluation of the model conditions be conducted prior to making inferences.

Evaluation of Normality Condition

Case 1: Each of the t treatment sample sizes n_i's is large

In this situation, construct a normal probability plot and run the Shapiro-Wilk test separately for each of the t random samples of observations from the t treatment populations. Thus, an assessment of whether or not each of the t treatment populations had a normal distribution would be determined. This case rarely occurs in most experiments.

Measure for the Normal Distribution: Shapiro-Wilk Measure

Shapiro and Wilk's W statistic is one of the most powerful procedures for assessing the fit of the normal distribution. The W statistic is a measure of the straightness of the normal reference plot, and small values of W indicate a departure from normality. The values of μ and σ do not need to be specified for the computation of the W statistic:

$$W = \frac{\left(\sum_{i=1}^{k} a_{n-i+1}[X_{(n-i+1)} - X_{(i)}]\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad \text{where}$$

1. $k = \frac{n}{2}$ if n is even and $k = \frac{(n-1)}{2}$ if n is odd
2. $X_{(i)}$'s are the order statistics of the X_i 's
3. a_i's are given in Table A28 on the following page
4. $a_i = -a_{n-i+1}$ for $i = 1, 2, \ldots, k$ and $a_{(n+1)/2} = 0$ for n odd

The upper percentiles of W are given in Table A29 on the following page to assess the p-value associated with a computed value of W.

The computation of W can also be obtained from SAS. The following SAS commands will provide both the Shapiro-Wilk test and a normal probability plot:

Suppose our response variable is labelled as **y** and the treatment variable is **TRT**. After inputting the data, use the following commands to obtain a SW test of normality and a normal probability plot for each treatment:

```
proc sort;

by TRT;
proc univariate plot normal;
var y:
by TRT;
run;
```

The following R code will produce the p-value for the Shapiro-Wilk test of normality for each level of the treatment:

$y_1 = c(y_{11}, y_{12}, ..., y_{1n_1})$
$y_2 = c(y_{21}, y_{22}, ..., y_{2n_2})$
.
$y_t = c(y_{t1}, y_{t2}, ..., y_{tn_t})$
$\mathbf{SW}_1 = \mathbf{shapiro.test(y}_1\mathbf{)}$
$\mathbf{SW}_2 = \mathbf{shapiro.test(y}_2\mathbf{)}$
.
$\mathbf{SW}_t = \mathbf{shapiro.test(y}_t\mathbf{)}$

Table A28 Coefficients Used in the Shapiro–Wilk Test for Normality*

	a_{n-i+1}											
i	$n=3$	4	5	6	7	8	9	10	11	12	13	14
1	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
2		0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
3				0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
4						0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
5								0.0399	0.0695	0.0922	0.1099	0.1240
6										0.0303	0.0539	0.0727
7												0.0240
	$n=15$	16	17	18	19	20	21	22	23	24	25	26
1	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407
2	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043
3	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533
4	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151
5	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836
6	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563
7	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316
8		0.0196	0.0359	0.0496	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089
9				0.0163	0.0303	0.0422	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876
10						0.0140	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672
11								0.0122	0.0228	0.0321	0.0403	0.0476
12										0.0107	0.0200	0.0284
13												0.0094
	$n=27$	28	29	30	31	32	33	34	35	36	37	38
1	0.4366	0.4328	0.4291	0.4254	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015
2	0.3018	0.2992	0.2968	0.2944	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774
3	0.2522	0.2510	0.2499	0.2487	0.2475	0.2463	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391
4	0.2152	0.2151	0.2150	0.2148	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110
5	0.1848	0.1857	0.1864	0.1870	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881
6	0.1584	0.1601	0.1616	0.1630	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686
7	0.1346	0.1372	0.1395	0.1415	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513
8	0.1128	0.1162	0.1192	0.1219	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356
9	0.0923	0.0965	0.1002	0.1036	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211
10	0.0728	0.0778	0.0822	0.0862	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075
11	0.0540	0.0598	0.0650	0.0697	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947
12	0.0358	0.0424	0.0483	0.0537	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824
13	0.0178	0.0253	0.0320	0.0381	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706
14		0.0084	0.0159	0.0227	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592
15				0.0076	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481
16						0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372
17								0.0062	0.0119	0.0172	0.0220	0.0264
18										0.0057	0.0110	0.0158
19												0.0053
	$n=39$	40	41	42	43	44	45	46	47	48	49	50
1	0.3989	0.3964	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2755	0.2737	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2380	0.2368	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2104	0.2098	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1880	0.1878	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1689	0.1691	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1520	0.1526	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1366	0.1376	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1225	0.1237	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1092	0.1108	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.0967	0.0986	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0848	0.0870	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0733	0.0759	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0622	0.0651	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0515	0.0546	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0409	0.0444	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0305	0.0343	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0203	0.0244	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0101	0.0146	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20		0.0049	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21				0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22						0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23								0.0039	0.0076	0.0111	0.0143	0.0174
24										0.0037	0.0071	0.0104
25												0.0035

*$a_i = -a_{n-i+1}$ for $i = 1, 2, \ldots, k$ where $k = n/2$ if n is even and $k = (n-1)/2$ if n is odd.

Source: Shapiro, S. S. and Wilk, M. B. (1965). "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, **52**, 591–611. Copyright Biometrika Trustees. Reprinted with permission.

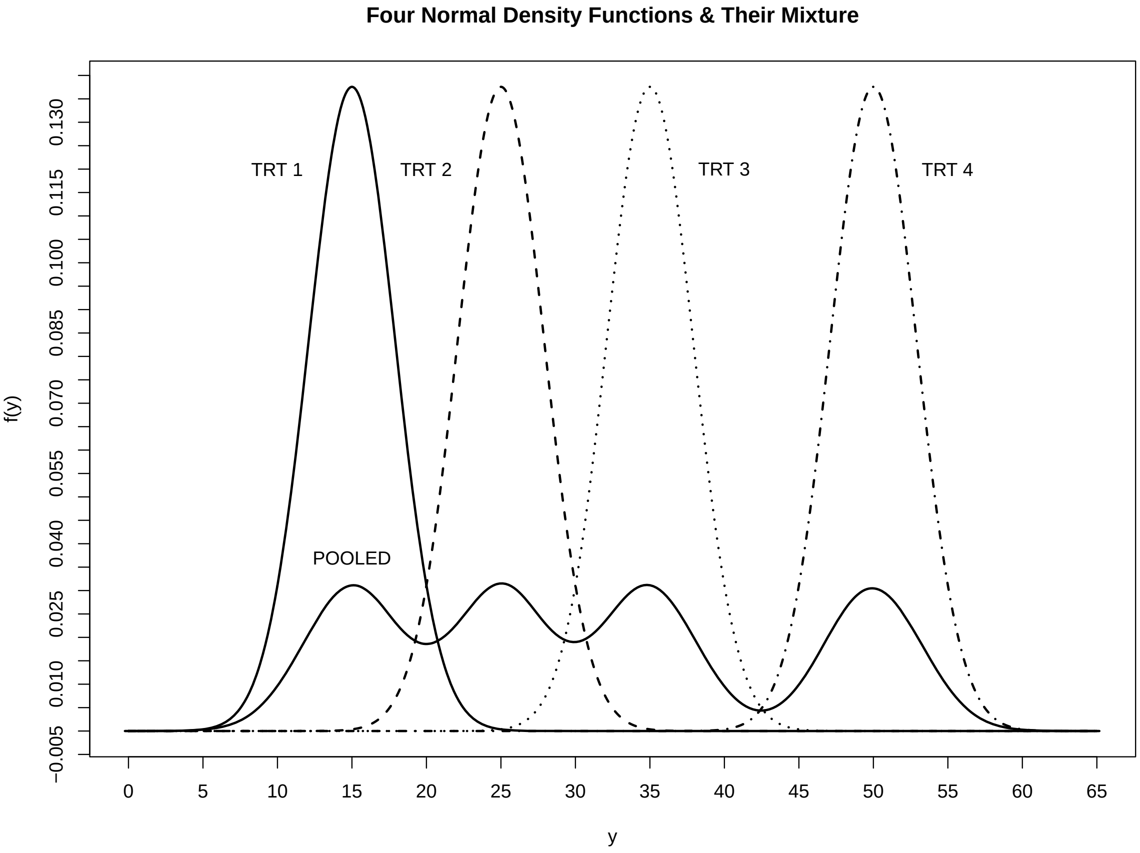
Table A29 Critical Values for the Shapiro–Wilk Test for Normality

			Critical Value		
n	$\alpha = 1\%$	2%	5%	10%	50%
3	0.753	0.756	0.767	0.789	0.959
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
6	0.713	0.743	0.788	0.826	0.927
7	0.730	0.760	0.803	0.838	0.928
8	0.749	0.778	0.818	0.851	0.932
9	0.764	0.791	0.829	0.859	0.935
10	0.781	0.806	0.842	0.869	0.938
11	0.792	0.817	0.850	0.876	0.940
12	0.805	0.828	0.859	0.883	0.943
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
16	0.844	0.863	0.887	0.906	0.952
17	0.851	0.869	0.892	0.910	0.954
18	0.858	0.874	0.897	0.914	0.956
19	0.863	0.879	0.901	0.917	0.957
20	0.868	0.884	0.905	0.920	0.959
21	0.873	0.888	0.908	0.923	0.960
22	0.878	0.892	0.911	0.926	0.961
23	0.881	0.895	0.914	0.928	0.962
24	0.884	0.898	0.916	0.930	0.963
25	0.888	0.901	0.918	0.931	0.964
26	0.891	0.904	0.920	0.933	0.965
27	0.894	0.906	0.923	0.935	0.965
28	0.896	0.908	0.924	0.936	0.966
29	0.898	0.910	0.926	0.937	0.966
30	0.900	0.912	0.927	0.939	0.967
31	0.902	0.914	0.929	0.940	0.967
32	0.904	0.915	0.930	0.941	0.968
33	0.906	0.917	0.931	0.942	0.968
34	0.908	0.919	0.933	0.943	0.969
35	0.910	0.920	0.934	0.944	0.969
36	0.912	0.922	0.935	0.945	0.970
37	0.914	0.924	0.936	0.946	0.970
38	0.916	0.925	0.938	0.947	0.971
39	0.917	0.927	0.939	0.948	0.971
40	0.919	0.928	0.940	0.949	0.972
41	0.920	0.929	0.941	0.950	0.972
42	0.922	0.930	0.942	0.951	0.972
43	0.923	0.932	0.943	0.951	0.973
44	0.924	0.933	0.944	0.952	0.973
45	0.926	0.934	0.945	0.953	0.973
46	0.927	0.935	0.945	0.953	0.974
47	0.928	0.928	0.946	0.954	0.974
48	0.929	0.937	0.947	0.954	0.974
49	0.929	0.937	0.947	0.955	0.974
50	0.930	0.938	0.947	0.955	0.974

Source: Adapted from Shapiro, S. S. and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611. Copyright Biometrika Trustees. Reprinted with permission.

Case 2: One of more of the sample sizes n_i is small

When the n_i's are relatively small it is not possible to obtain meaningful normal probability plots or S-W tests of normality. Ideally, we would apply the evaluation of the normality condition to all n data values. However, if there is a difference in the treatment means, we would be pooling observations from populations having different means. Thus, even if all t population distributions were normal distributions, the pooled data would often indicate a multimodal distribution.



Therefore, for the case when n_i's are small, it is necessary to examine the sample residuals from the fitted model: $y_{ij} = \mu_i + e_{ij}$

$$e_{ij} = y_{ij} - \mu_i \quad \Rightarrow \quad \hat{e}_{ij} = y_{ij} - \hat{\mu}_i = y_{ij} - \bar{y}_{i.},$$

where $\hat{\mu}_i = \bar{y}_{i.}$ is the LSE of μ_i.

There are $n = n_1 + n_2 + \cdots + n_t$ residuals so that in most experiments, there would be a large enough sample to yield valid plots and a S-W test with reasonable power would be possible.

Properties of the residuals

1. $\hat{e}_{ij}$ have a normal distribution

$$\hat{e}_{ih} = y_{ih} - \bar{y}_{i.} = y_{ih} - \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \left(1 - \frac{1}{n_i}\right) y_{ih} - \frac{1}{n_i}\sum_{j\neq h} y_{ij}$$

   Thus, $\hat{e}_{ih}$ is a linear combination of independent normal r.v.'s and hence has a normal distribution.

2. $Var\left(\hat{e}_{ij}\right) = \left(1 - \frac{1}{n_i}\right)\sigma_e^2$

   Thus, if n_i are unequal, then the residuals have unequal variances.

$$Var\left(\hat{e}_{ih}\right) = Var\left(y_{ih} - \bar{y}_{i.}\right) = Var\left(y_{ih}\right) + Var\left(\bar{y}_{i.}\right) - 2cov\left(y_{ih}, \bar{y}_{i.}\right) = \text{ messy calculations}$$

$$\begin{aligned}
Var\left(\hat{e}_{ih}\right) &= Var\left(y_{ih} - \bar{y}_{i.}\right) \\
&= Var\left(\left(1 - \frac{1}{n_i}\right) y_{ih} - \frac{1}{n_i}\sum_{j\neq h} y_{ij}\right) \\
&= \left(1 - \frac{1}{n_i}\right)^2 Var(y_{ih}) + \frac{1}{n_i^2}\sum_{j\neq h}^{n_i} Var(y_{ij}) \\
&= \left(1 - \frac{1}{n_i}\right)^2 \sigma_e^2 + \frac{1}{n_i^2}(n_i - 1)\sigma_e^2 \\
&= \left(1 - \frac{1}{n_i}\right)\sigma_e^2 \\
&= \left(1 - \frac{1}{r}\right)\sigma_e^2 \quad (\text{if} \;\; n_1 = n_2 = \cdots = n_t = r)
\end{aligned}$$

3. $\hat{e}_{ij}$'s are correlated

$$\begin{aligned} Cov(\hat{e}_{ij}, \hat{e}_{i^{'}h}) &= Cov(y_{ij} - \bar{y}_{i.}, y_{i^{'}h} - \bar{y}_{i^{'}.}) \\ &= Cov(y_{ij}, y_{i^{'}h}) - Cov(y_{ij}, \bar{y}_{i^{'}.}) - Cov(y_{i^{'}h}, \bar{y}_{i.}) + Cov(\bar{y}_{i.}, \bar{y}_{i^{'}.}) \end{aligned}$$

a. If $i \neq i^{'}$, then $Cov(\hat{e}_{ij}, \hat{e}_{i^{'}h}) = 0$, because y_{ij} is independent of $y_{i^{'}h}$ for all values of j and h.

b. If $i = i^{'}$ and $j = h$, then $Cov(\hat{e}_{ij}, \hat{e}_{i^{'}h}) = Cov(\hat{e}_{ij}, \hat{e}_{ij}) = Var(\hat{e}_{ij}) = (1 - \frac{1}{n_i})\sigma_e^2$

c. If $i = i^{'}$ and $j \neq h$, then $Cov(\hat{e}_{ij}, \hat{e}_{i^{'}h}) = Cov(\hat{e}_{ij}, \hat{e}_{ih}) = -\frac{\sigma_e^2}{n_i}$

$$\begin{aligned} Cov(\hat{e}_{ij}, \hat{e}_{ih}) &= 0 - 2Cov(y_{ih}, \bar{y}_{i.}) + Cov(\bar{y}_{i.}, \bar{y}_{i.}) \\ &= 0 - 2Cov\left(y_{ih}, \frac{1}{n_i}\sum_{k=1}^{n_i} y_{ik}\right) + Var(\bar{y}_{i.}) \\ &= -2\frac{1}{n_i}\sum_{k=1}^{n_i} Cov(y_{ih}, y_{ik}) + \frac{\sigma_e^2}{n_i} \\ &= -2\frac{1}{n_i}Cov(y_{ih}, y_{ih}) + \frac{\sigma_e^2}{n_i} \\ &= -2\frac{1}{n_i}Var(y_{ih}) + \frac{\sigma_e^2}{n_i} = -\frac{\sigma_e^2}{n_i} \end{aligned}$$

d. $Corr(\hat{e}_{ij}, \hat{e}_{i^{'}h}) = \frac{Cov(\hat{e}_{ij}, \hat{e}_{i^{'}h})}{Var(\hat{e}_{ij})} = \frac{Cov(\hat{e}_{ij}, \hat{e}_{i^{'}h})}{\left(1 - \frac{1}{n_i}\right)\sigma_e^2}$

Therefore,

$$Corr(\hat{e}_{ij}, \hat{e}_{i^{'}h}) = \begin{cases} 0 & if \quad i \neq i^{'} \\ 1 & if \quad i = i^{'}, j = h \\ -\frac{1}{n_i - 1} & if \quad i = i^{'}, j \neq h \end{cases}$$

e. From the above formula for the correlation, it is obvious that in most experiments the correlation between the residuals is minimal. For example,

if $n_i > 6$, then $-.2 <$ Corr < 0 or

$n_i > 11$, then $-.1 <$ Corr < 0 or

$n_i > 21$, then $-.05 <$ Corr < 0.

Evaluation of Normality Using the Sample Residuals:

To evaluate whether or not the e_{ij}'s have a normal distribution, we would thus apply

1. the Shapiro-Wilk's (S-W) test to the n sample residuals: $\hat{e}_{ij}$

2. Construct normal probability plots using the sample residual: $\hat{e}_{ij}$.

3. Construct box plots using the sample residual: $\hat{e}_{ij}$.

4. When the sample sizes are unequal, the results are somewhat affected by the unequal variances but only minimally.

5. The slight correlation in the residuals likewise has a slight affect on the sensitive of the S-W test.

6. An approach to overcome the above problems is to standardized the residuals by dividing by the estimated standard errors, yielding the Studentized Residuals:

   $$\hat{e}^*_{ij} = \frac{\hat{e}_{ij}}{\hat{\sigma}_e\sqrt{1 - \frac{1}{n_i}}} \qquad \text{with } \hat{\sigma}_e = \sqrt{MSE}$$

   $\hat{e}^*_{ij}$ has a t-distribution with $df = df_E$

   The studentized residuals are still correlated however.

7. In order to detect if an observation y_{ij} is an **outlier** relative to observations from a normally distributed population, declare y_{ij} to be an outlier if $|\hat{e}^*_{ij}| \geq t_{.0005,df_E} \approx 3.3$ for large n.

8. Alternatively, a box plot of the studentized residuals could be used to detect outliers by declaring an observation an outlier if the corresponding studentized residual is beyond three IQR's of the quartiles (extreme outliers in box plot).

9. The Grubbs test provides a formal statistical test of whether or not an observation is an outlier. See the article by Beckman and Cook(1983), *Technometrics*, **25**, pp. 119-149.

Evaluation of Equal Variance Condition: $\sigma_1 = \sigma_2 = \cdots = \sigma_t$

The deviation of the actual size and power of tests, and actual coverage probability of confidence intervals relative to their nominal values is greater for a violation of constant variance than for moderate violations of the normality condition. Although the AOV F-test is robust to moderate deviations from constant variances provided $n_1 = n_2 = \cdots = n_t$.

Brown-Forsythe-Levene Test of Homogeneity of Variance $(n_i \geq 3)$

The Brown-Forsythe version of the Levene test allows an assessment of the equality of the t population variances without the necessity of the normality of the distributions provided all $n_i \geq 3$.

The Levene test involves replacing the data value, y_{ij}, with the random variable
$z^*_{ij} = |y_{ij} - \bar{y}_{i.}|$, and then computing the test statistic using the z^*_{ij}'s.

The Brown-Forsythe modification to the Levene test involves replacing the observation, y_{ij}, with the random variable $z_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}_i$ is the sample median of the ith sample. This produces a procedure which is less affected by outliers and skewed distributions.

$H_o : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_t^2$ homogeneity of variances

$H_a :$ Population variances are not all equal

T.S.: $L = \dfrac{\sum_{i=1}^t n_i(\bar{z}_{i.} - \bar{z}_{..})^2/(t-1)}{\sum_{i=1}^t \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2/(N-t)}$

R.R. For a specified value of α, reject H_o if L $\geq F_{\alpha,df_1,df_2}$, where $df_1 = t-1$, $df_2 = n-t$, $n = \sum_{i=1}^t n_i$, and F_{α,df_1,df_2} is the upper α percentile from the F-distribution.

Note: The B-F-L test statistic is simply the AOV F-Test applied to the data values $z_{ij} = |y_{ij} - \tilde{y}_i|$. Thus, any software package that computes the AOV F-test can be used by first transforming the data y_{ij} to z_{ij} and then applying the AOV procedure to the z_{ij}'s. Furthermore, unless $n_i > 3$, the B-F-L test is not very meaningful.

When the sample sizes within the levels of the groups are odd, at least one value of $z_{ij} = |y_{ij} - \tilde{y}_i|$ will always be zero. This artificially dampens the variance estimate for that group. Thus, Hajek-Sidak recommend replacing the zero with the minimum non-zero value if there is only one z_{ij} which is zero. If more than one z_{ij} within a given group is zero, the zeros are kept.

The following SAS code will obtain the B-F-L test directly without having to transform the data.

```
proc glm data=old;
class trt;
model y=trt;
means trt/hovtest=bf;
run;
```

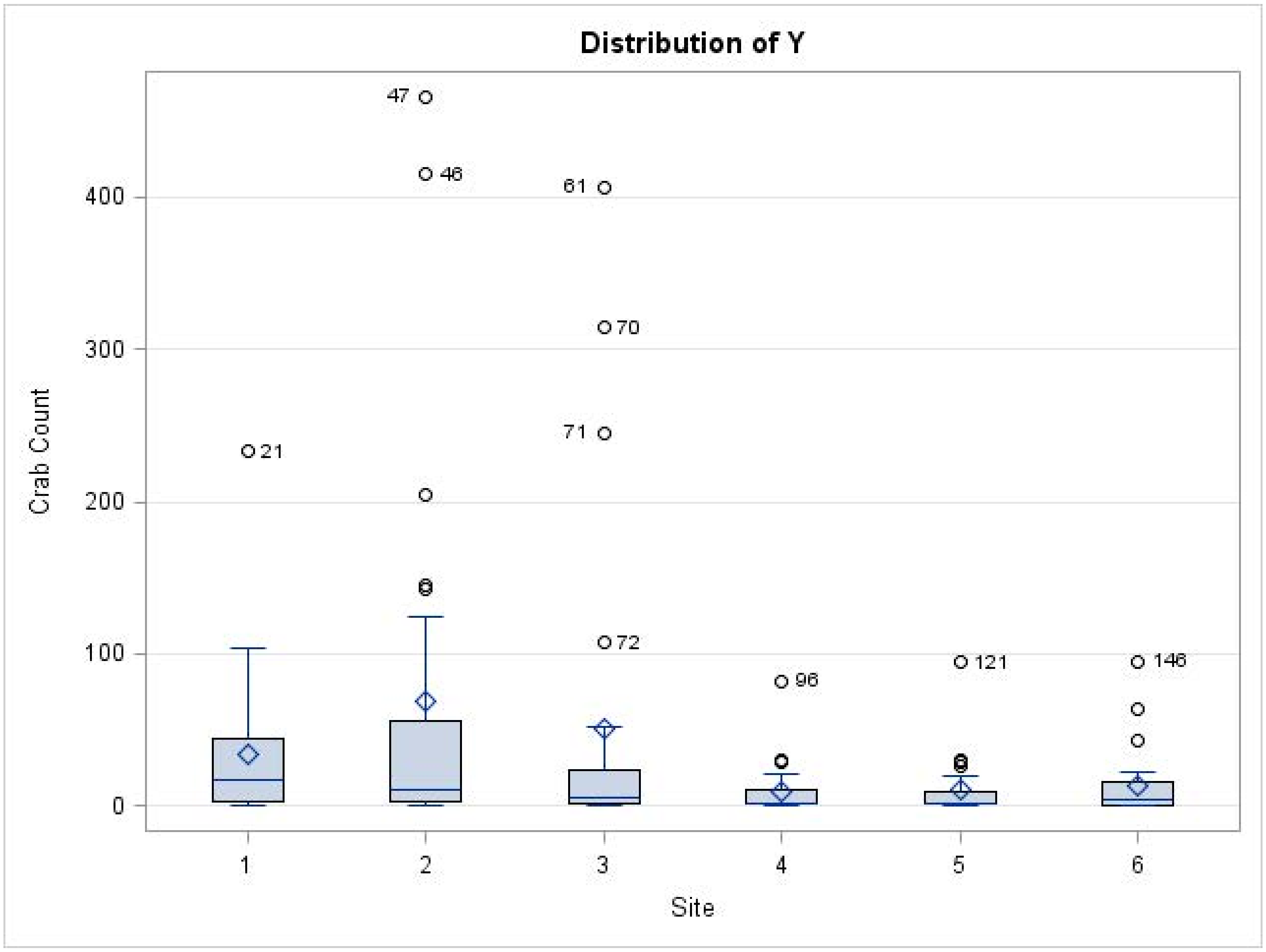
Graphical Detections of Unequal Variances

The following graphs could also be used to detect unequal variances across the t treatment populations:

1. **Box Plots**

   If $n_1, n_2, \cdots, n_t$ are all reasonably large then side-by-side box plots of the residuals would display the variability in the t samples and hence provide evidence of a difference in the t treatment population variances.

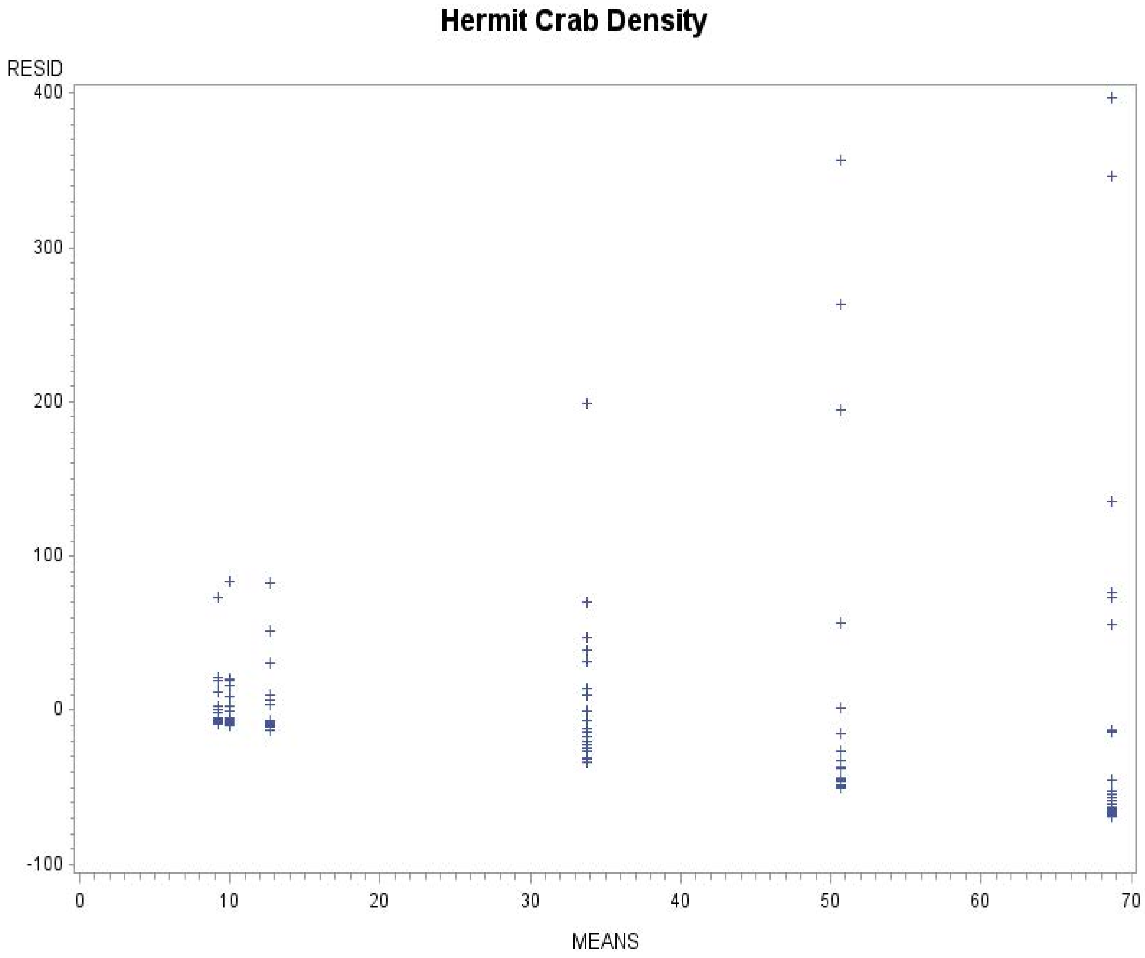
   Box Plots of Crab Counts:



2. **Residual Plots**

If some of $n_1, n_2, \cdots, n_t$ are somewhat small, the individual box plots are not very informative. A plot of the residuals versus the estimated treatment means provides an indication of a relationship between the population variances and population means:

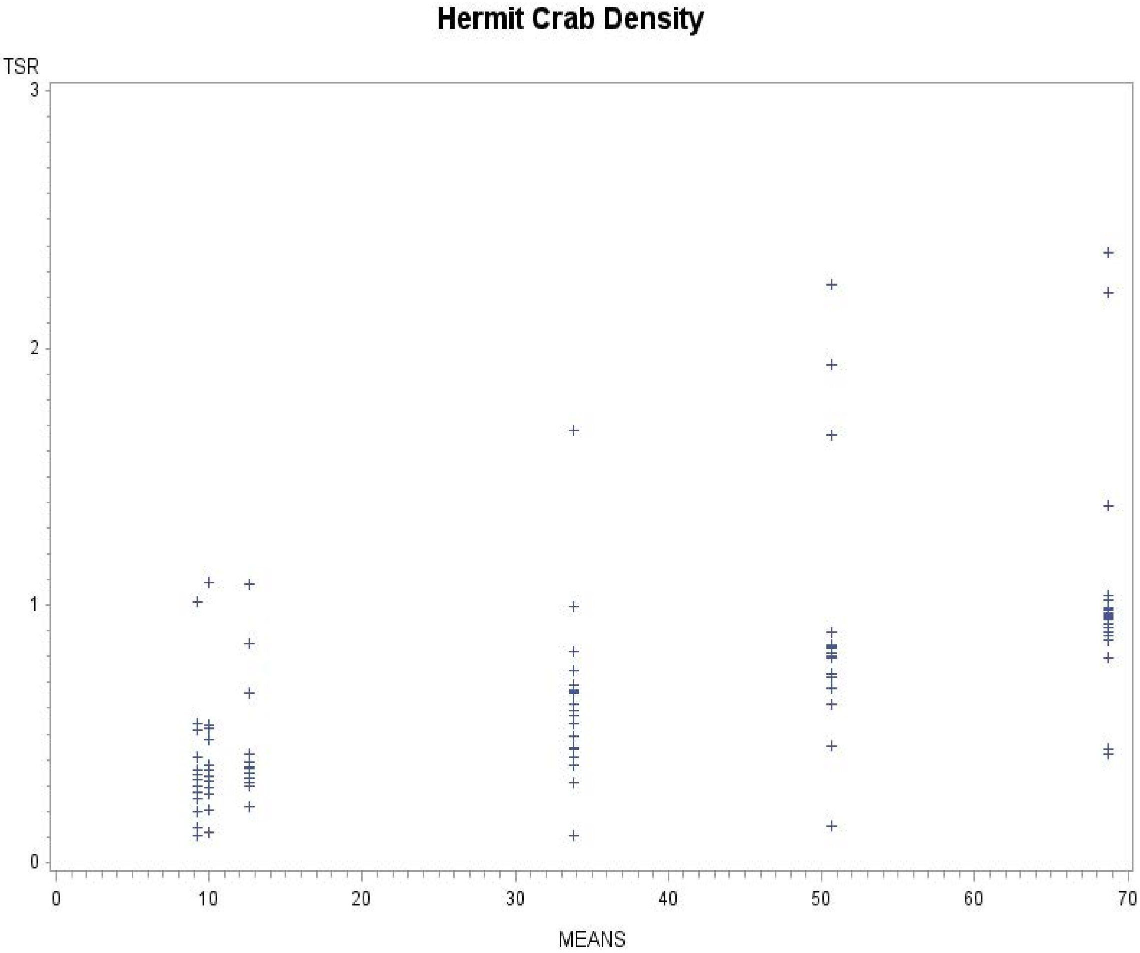
Residual Plots of Crab Counts:



3. **Standardized Residual Plots**

   In place of plotting the residuals versus the treatment means, a plot which is somewhat more informative is a plot of the square root of the absolute value of the studentized residual $\sqrt{|\hat{e}^*_{ij}|}$ versus the treatment means. The absolute values of the residuals reflects the degree of variability with a treatment group. The absolute studentized residuals have an asymmetric distribution. Taking the square root, somewhat removes this asymmetry.

Square Root of Absolute Residual Plots of Crab Counts:



The AOV F-test is relatively robust to departures from normality and unequal variances, especially when the sample sizes are equal. Unless the group variances are extremely different or the number of groups, t is large, the usual ANOVA test is relatively robust when the sample sizes n_i's are all about the same size. As Box is quoted, "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!"

EXAMPLE: Assessment of Model Conditions

The following data from Kuehl, *Design of Experiments* will be used to illustrate the above methods for assessment of deviations from normality and homogeneity of variances.

Hermit Crab Counts in Coastline Habitats

A marine biologist was interested in the relationship between different coastline habitats and the populations of Hermit crabs inhibiting the site. The biologist counted Hermit crabs on 25 transects randomly located in each of six different sites of a coastline habitat. The data and summary statistics are given in the following tables.

Number of Crabs per Transect at 6 Habitats (H)

	North to South Orientation of Transects																								
H	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0	0	22	3	17	0	0	7	11	11	73	33	0	65	13	44	20	27	48	104	233	81	22	9	2
2	0	0	56	0	8	0	3	1	16	55	142	10	2	145	6	4	5	124	24	204	415	466	6	14	12
3	0	0	4	13	5	1	1	4	4	36	407	0	0	18	4	14	0	24	52	314	245	107	5	6	2
4	0	0	0	4	2	2	5	4	2	1	0	12	1	30	0	3	28	2	21	8	82	12	10	2	0
5	0	1	1	2	2	1	2	29	2	2	0	13	0	19	1	3	26	30	5	4	94	1	9	3	0
6	0	0	0	2	3	0	0	4	0	5	4	22	0	64	4	4	43	3	16	19	95	6	22	0	0

Summary Statistics for Count Data

Habitat	Mean ($\hat{\mu}_i$)	Median ($\tilde{\mu}_i$)	StdDev ($\hat{\sigma}_i$)	Minimum ($Y_{(1)}$)	Maximum ($Y_{(25)}$)
1	33.80	17	50.39	0	233
2	68.72	10	125.35	0	466
3	50.64	5	107.44	0	407
4	9.24	2	17.39	0	82
5	10.00	2	19.84	0	94
6	12.64	4	23.01	0	95

```
* crab.sas;
* The relationship between different habitats and the population densities of
Hermit crabs. There are 6 sites. At each site 25 transects are run and the
number of crabs are counted.;

ods html; ods graphics on;

option ls=70 ps=50 nocenter nodate;
title 'Hermit Crab Density';

*Input Data;
data count;
infile 'u:\meth2\kuehl\expl4-1.dat';
input Y Site;
label Y = 'Crab Count';

*Generate BoxPlots;
proc boxplot;
plot y*site/boxstyle=schematic;
run;

*Analysis of Variance;
proc glm data=count;
class Site;
model  Y = Site;

*Brown-Forsythe-Levene Test;
means Site/hovtest=bf;

means Site/ LSD tukey snk;

*Residual analysis;
output out=ASSUMP r=RESID p=MEANS STUDENT=SR;
DATA TRANSRESID; SET ASSUMP; TSR=SQRT(ABS(SR));
proc univariate def=5 plot normal; var RESID;
proc gplot data=assump; plot resid*means;
PROC gplot DATA=TRANSRESID; PLOT TSR*MEANS;
RUN;
ods graphics off;
ods html close;
```

```
Class         Levels    Values
Site               6    1 2 3 4 5 6
Number of observations    150

Dependent Variable: Y   Crab Count

                                    Sum of
Source                     DF      Squares    Mean Square   F Value  Pr > F
Model                       5   76695.0400     15339.0080      2.97  0.0140
Error                     144  744493.1200      5170.0911
Corrected Total           149  821188.1600

Source                     DF  Type III SS    Mean Square   F Value  Pr > F
Site                        5  76695.04000    15339.00800      2.97  0.0140
```

```
   Brown and Forsythe's Test for Homogeneity of Y Variance
       ANOVA of Absolute Deviations from Group Medians

                      Sum of        Mean
Source        DF     Squares      Square    F Value    Pr > F

Site           5     71145.7     14229.1       2.93    0.0151
Error        144      699845      4860.0
```

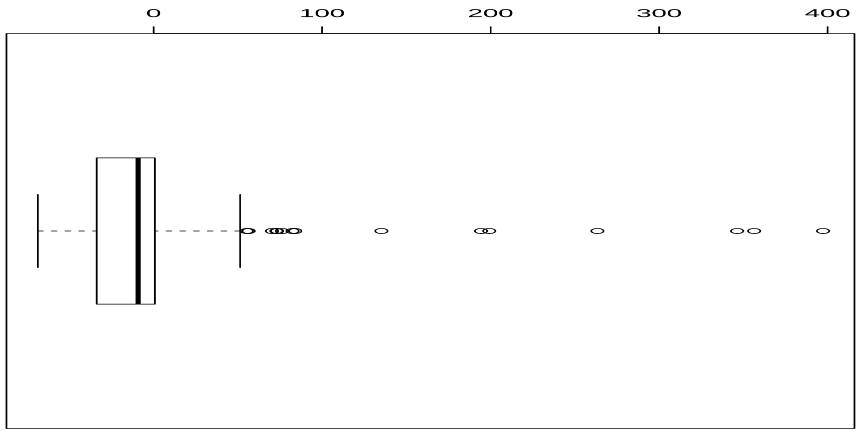
```
Level of            --------------Y--------------
Site          N             Mean          Std Dev

1            25       33.8000000        50.385183
2            25       68.7200000       125.353673
3            25       50.6400000       107.437920
4            25        9.2400000        17.386010
5            25       10.0000000        19.841035
6            25       12.6400000        23.010650
```

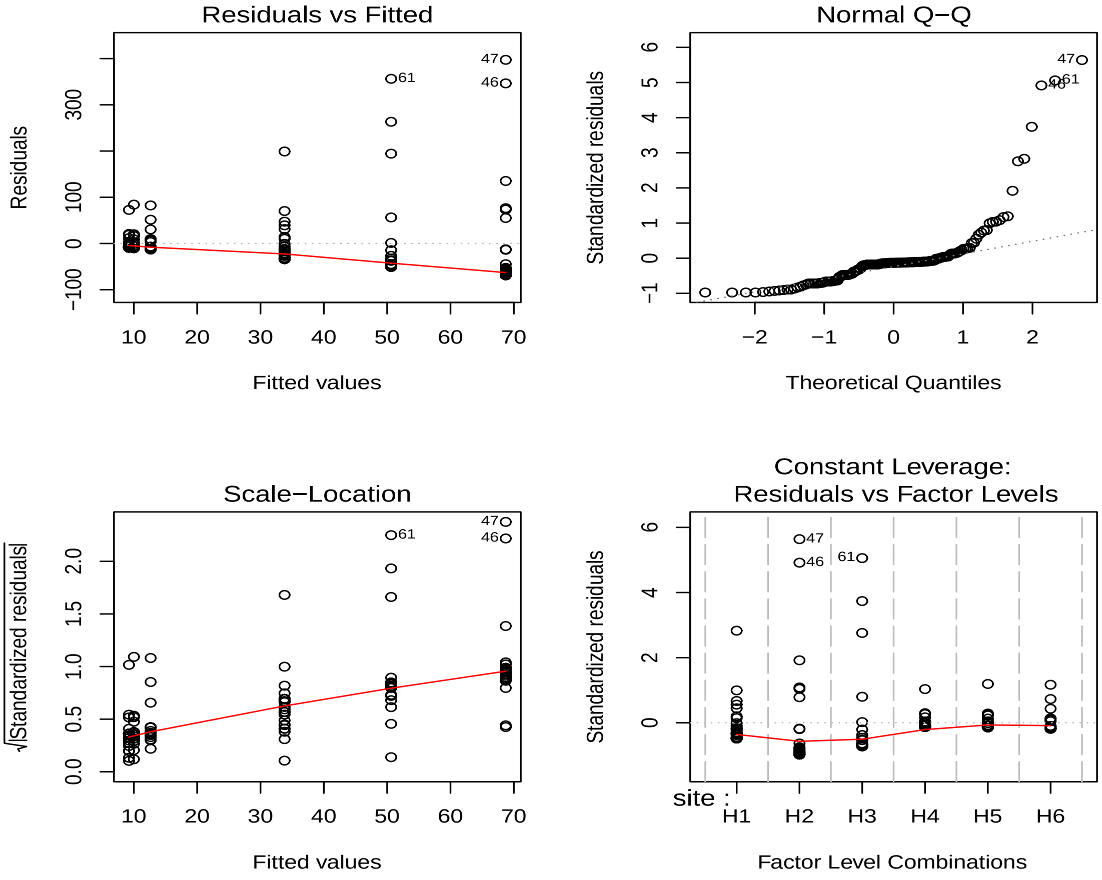
```
Tukey's Studentized Range (HSD) Test for Y

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.


Alpha                                   0.05
Error Degrees of Freedom                 144
Error Mean Square                   5170.091
Critical Value of Studentized Range  4.08495
Minimum Significant Difference        58.744


Means with the same letter are not significantly different.


Tukey
Grouping       Mean      N    Site

     A        68.72     25    2
     A
B    A        50.64     25    3
B    A
B    A        33.80     25    1
B    A
B    A        12.64     25    6
B    A
B    A        10.00     25    5
B
B              9.24     25    4
```

```
Variable:  RESID

                   Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.615573    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.267711    Pr > D     <0.0100
Cramer-von Mises      W-Sq  3.029193    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  16.18001    Pr > A-Sq  <0.0050

Variable:  RESID

                  Histogram                  #  Boxplot
    390+*                                    1     *
       .
       .*                                    2     *
       .
       .
       .
       .*                                    1     *
       .
       .
       .
       .*                                    2     *
       .
       .
       .*                                    1     *
       .
       .*                                    2     0
       .**                                   4     0
       .**                                   4     0
       .***                                  5     |
       .********                            16  +--+--+
       .********************************    64  *-----*
       .********                            16  +-----+
       .**********                          20     |
    -70+******                              12     |
        ----+----+----+----+----+----+--
        * may represent up to 2 counts

Variable:  RESID

                       Normal Probability Plot
     390+                                                  *
        |
        |                                              * *
        |
        |
        |
        |                                             *
        |
        |
        |
        |                                           **
        |                                                +++
        |                                             +++
        |                                          *+++
        |                                        ++
        |                                     +++ *
        |                                  +++  **
        |                               +++   **
        |                            +++     **
        |                         +++    ****
        |                    *************
        |                 ****++
        |           *******++
     -70+* * ********  +++
         +----+----+----+----+----+----+----+----+----+----+
             -2        -1         0        +1        +2
```

R code to obtain Residual Plots - Crab-AOVResAnal.R in CANVAS

```
site_1  = c(0,0,22,3,17,0,0,7,11,11,73, 33, 0, 65, 13, 44, 20, 27, 48, 104, 233, 81, 22, 9, 2)
site_2  = c(0,0,56,0,8,0,3,1,16,55,142, 10, 2, 145, 6, 4, 5, 124, 24, 204, 415, 466, 6, 14, 12)
site_3  = c(0,0,4,13,5,1,1, 4, 4, 36, 407, 0, 0, 18, 4, 14, 0, 24, 52, 314, 245, 107, 5, 6, 2)
site_4  = c(0, 0, 0, 4, 2, 2, 5, 4, 2, 1, 0, 12, 1, 30, 0, 3, 28, 2, 21, 8, 82, 12, 10, 2, 0)
site_5  = c(0, 1, 1, 2, 2, 1, 2, 29, 2, 2, 0, 13, 0, 19, 1, 3, 26, 30, 5, 4, 94, 1, 9, 3, 0)
site_6  = c(0, 0, 0, 2, 3, 0, 0, 4, 0, 5, 4, 22, 0, 64, 4, 4, 43, 3, 16, 19, 95, 6, 22, 0, 0)
   y = c(site_1, site_2, site_3, site_4, site_5, site_6)
   S1  = rep("H1", 25)
   S2  = rep("H2", 25)
   S3  = rep("H3", 25)
   S4  = rep("H4", 25)
   S5  = rep("H5", 25)
   S6  = rep("H6", 25)
   hab  = c(S1, S2, S3, S4, S5, S6)
   site  = as.factor(hab)
   Habit_aov  = aov(y~site)
   summary(Habit_aov)
   Habit_model = lm(y~site)
   par(mfrow=c(2,2))
   plot(Habit_model)
   Habit_res = Habit_model$residuals

#Shapiro-Wilks test of normality of residuals
   shapiro.test(Habit_res)
          Shapiro-Wilk normality test
data:  Habit_res
W = 0.61557, p-value < 2.2e-16

# B-F-L test of homogeneity of Variances
 m1 = median(site_1)
 z1 = abs(site_1-m1)
 m2 = median(site_2)
 z2 = abs(site_2-m2)
 m3 = median(site_3)
 z3 = abs(site_3-m3)
 m4 = median(site_4)
 z4 = abs(site_4-m4)
 m5 = median(site_5)
 z5 = abs(site_5-m5)
 m6 = median(site_6)
 z6 = abs(site_6-m6)

 z = c(z1,z2,z3,z4,z5,z6)
 lev1  =  aov(z~site)

 summary(lev1)
             Df Sum Sq Mean Sq F value Pr(>F)
site          5  71146   14229   2.928 0.0151 *
Residuals   144 699845    4860
par(ask=TRUE)
boxplot(Habit_res)
```





STOP Monday 2/21/22 (Week 6, Lecture 15)

START Wednesday 2/23/22 (week 6, lecture 16)

Non-normal Residuals - Non-constant Variance

From the tests and plots of the crab count data, there is strong evidence that both the normality condition and the equal variance condition are not valid. What are alternatives to using the AOV F-test and multiple comparison procedures when the conditions are not met?

Transformations of the data to obtain constant variance

I. Power Transformations

Suppose that we have determined that the unequal variances are related to the treatment means by

$\sigma_i = \sqrt{Var(y_{ij})} = C(\mu_i)^\beta$, that is,

$\sigma_i = h(\mu_i) = C(\mu_i)^\beta$.

Let $x_{ij} = (y_{ij})^{1-\beta} = g(y_{ij}) \quad \Rightarrow \quad g^{(1)}(\mu_i) = (1-\beta)(\mu_i)^{-\beta}$

From the Taylor series approximation,

$Var(x_{ij}) \approx (g^{(1)}(\mu_i))^2(\sigma_i)^2 = (g^{(1)}(\mu_i)h(\mu_i))^2 = ((1-\beta)(\mu_i)^{-\beta})^2(C)^2(\mu_i)^{2\beta} = C^2(1-\beta)^2$

The transformed data x_{ij} have approximately equal variances.

How do we determine if the relationship, $\sigma_i = C(\mu_i)^\beta$, and what is the value of β?

$\sigma_i = C(\mu_i)^\beta \;\Rightarrow log(\sigma_i) = log(C) + \beta log(\mu_i)$

Thus, one approach is to plot $\log(S_i)$ versus $\log(\bar{y}_{i.})$ and determine if there is a straight-line relationship between $\log(S_i)$ and $\log(\bar{y}_{i.})$.

If the regression is valid, then $\log(S_i) = \beta_o + \beta_1 \log(\bar{y}_{i.})$.

Thus, use the LSE of β_1 from the regression as an estimate for the value of β in the transformation:

$x_{ij} = (y_{ij})^{1-\hat{\beta}_1}$

If $\hat{\beta}_1 \approx 1$, then use the transformation

$x_{ij} = log(y_{ij})$.

The justification for using $x_{ij} = log(y_{ij})$ is as follows:

Suppose $Var(y_{ij}) = \sigma_i^2 = [C\mu_i]^2$ and $x_{ij} = log(y_{ij}) = g(y_{ij})$. Then

$Var(x_{ij}) = \left[\frac{dg(y)}{dy})|_{y=\mu_i}\right]^2 [\sigma_i]^2 = \left[\frac{1}{\mu_i}\right]^2 [C\mu_i]^2 = C^2$

II. Box-Cox Transformation:

Suppose the data $Y_1, Y_2, \ldots, Y_n$ consist of iid r.v.'s with positive values and a pdf f_Y which is skewed.

A power transformation defined by

$$x_{ij} = \begin{cases} (y_{ij}^{\theta} - 1)/(\theta * (gmy)^{\theta-1}) & : \quad \theta \neq 0 \\ (gmy)log(y_{ij}) & : \quad \theta = 0 \end{cases}$$

where gmy is the geometric mean of the y_{ij}'s:

$$\text{gmy} = \prod_{i=1}^{t} \prod_{j=1}^{n_i} y_{ij}^{1/n} = e^{\frac{1}{n} \sum_{i=1}^{t} \sum_{i=1}^{n_i} log(y_{ij})}$$

The Box-Cox transformation can often produce 'nearly' a normal distribution for $y^{(\theta)}$. That is, the pdf of $y^{(\theta)}$ is a $N(\mu, \sigma^2)$ pdf.

Note: $\lim_{\theta \to 0} \frac{(y^{\theta}-1)}{\theta} = log(y)$.

If in fact the power transformation is successful, and $y^{(\theta)}$ has a normal distribution, $N(\mu, \sigma^2)$ then the pdf of y, f_Y, is given by

$$f_Y(y) = y^{\theta-1} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(y^{(\theta)}-\mu)^2} = y^{\theta-1} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}\left[\frac{(y^{\theta}-1)}{\theta}-\mu\right]^2}$$

Determination of θ:

1. Select grid of possible values for θ
2. For each θ, let $x_{ij} = (y_{ij}^{\theta} - 1)/(\theta * (gmy)^{\theta-1})$
3. Run AOV on x_{ij} and obtain MSE(θ)
4. Select θ having maximum value of $L(\theta) = -\frac{1}{2}log(MSE(\theta))$

In Case I and Case II, x_{ij}'s have approximately equal variances and we can then conduct the AOV F-test and multiple comparison procedures. However, it is important to note that the tests and confidence intervals are being constructed using the transformed data, x_{ij}. Thus, we are making inferences about μ_i^x and not about μ_i. For example, we are testing

$H_o^x : \mu_1^x = \mu_2^x = \cdots = \mu_t^x$ versus H_1^x : not all μ_i^x's are equal

when in fact we want to test

$H_o^y : \mu_1^y = \mu_2^y = \cdots = \mu_t^y$ versus H_1^y : not all μ_i's are equal

Are the sets of hypotheses, (H_o^x, H_1^x) and (H_o^y, H_1^y) equivalent?

If the 1st order Taylor Series expansion approximation is accurate then:

$\mu_i^x \approx g(\mu_i)$. However, the two sets of hypotheses may not be equivalent.

Consider the following example.

EXAMPLE:

Suppose y_{ij}'s have a log normal distribution, with mean and standard deviation: μ_i and σ_i.

Then, $x_{ij} = \log(y_{ij})$ has a $N(\mu_i^x, (\sigma_i^x)^2)$ distribution. Therefore,

$E[x_{ij}] = \mu_i^x$ and $Var[x_{ij}] = (\sigma_i^x)^2$

$E[y_{ij}] = \mu_i = e^{\mu_i^x + \frac{1}{2}(\sigma_i^x)^2} \;\Rightarrow\; log(\mu_i) = \mu_i^x + \frac{1}{2}(\sigma_i^x)^2$

Thus, we conclude that $\mu_i^x \approx \log(\mu_i)$ only if $\sigma_i^x \approx 0$.

However, if the t population variances for the x_{ij}'s are equal, $\sigma_1^x = \sigma_2^x = \cdots = \sigma_t^x = \sigma^x$, then we have the following equivalences between testing hypotheses about the means of the y_{ij}'s and the means of the x_{ij}'s:

$H_o : \mu_1 = \mu_2 = \cdots = \mu_t$ holds if and only if $H_o : \mu_1^x = \mu_2^x = \cdots = \mu_t^x$

The above follows from $\mu_i = \mu_h$ if and only if $e^{\mu_i^x + \frac{1}{2}(\sigma^x)^2} = e^{\mu_h^x + \frac{1}{2}(\sigma^x)^2}$ if and only if $e^{\mu_i^x} = e^{\mu_h^x}$ if and only if $\mu_i^x = \mu_h^x$.

Warning: Great care must be taken in making inferences about the treatment populations using tests and confidence intervals obtained from transformed data.

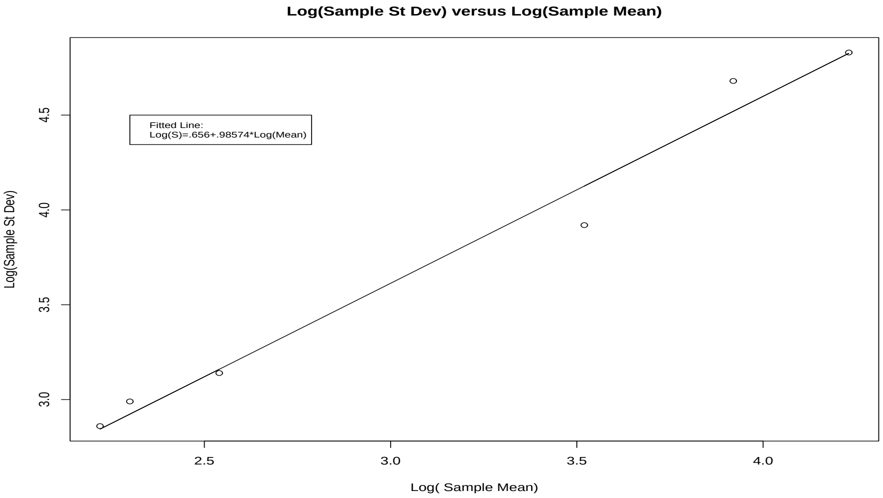
EXAMPLE

Determining The Appropriate Transformation - Using Power Relationship

For the crab count data we have the following values for S_i and $\bar{y}_{i.}$:

Site	$\bar{y}_{i.}$	S_i	$log(\bar{y}_{i.})$	$log(S_i)$
1	33.80	50.39	3.5205	3.9198
2	68.72	125.35	4.2300	4.8311
3	50.64	107.44	3.9247	4.6769
4	9.24	17.39	2.2235	2.8559
5	10.00	19.84	2.3026	2.9877
6	12.64	23.01	2.5369	3.1359

A plot of $log(S_i)$ versus $log(\bar{y}_{i.})$ yields



From the plot there appears to be a strong linear relationship between $log(S_i)$ and $log(\bar{y}_{i.})$. The regression of $log(S_i)$ on $log(\bar{y}_{i.})$ yields (see R code)

```
mean = c(33.8,68.72,50.64,9.24,10.00,12.64)
stdv = c(50.39,125.35,107.44,17.39,19.84,23.01)
x = log(mean)
y = log(stdv)
reg.fit = lm(y˜x)
summary(reg.fit)
z = .656+.98574*x
plot(x,y,type="p",main="Log(Sample St Dev) versus Log(Sample Mean)",
xlab="Log( Sample Mean)",ylab="Log(Sample St Dev)")
matlines(x,z,type="l",main="Log(Sample St Dev) versus Log(Sample Mean)",
xlab="Log( Sample Mean)",ylab="Log(Sample St Dev)")
legend(2.3,4.5,c("Fitted Line:","Log(S)=.656+.98574*Log(Mean)"),cex=.75)
```

$log(S_i) = 0.656 + 0.98574 log(\bar{y}_{i.})$ with $\mathrm{R}^2 = .981$.

That is, $\hat{\beta} = .9857$. Therefore, $\hat{\beta} \approx 1$ and we use the transformation $X = log(Y)$.

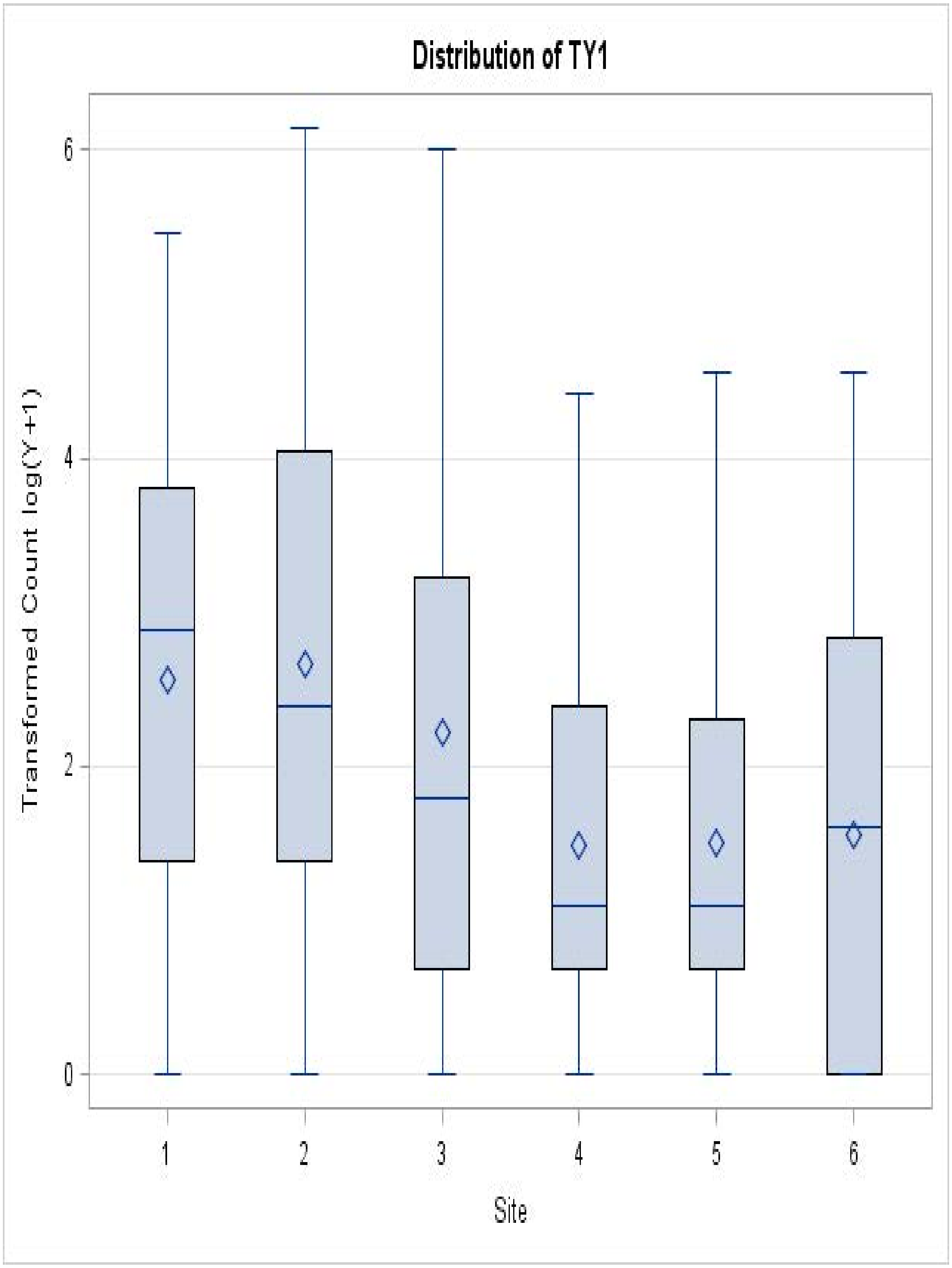
The impact of this transformation is given in the SAS output on the next pages.

```
* transcrab.sas;
option ls=80 ps=50 nocenter nodate;
title 'Transformation of Crab Data';
data old;
input Y S;
LY=log(Y);
LS=log(S);
cards;
33.80  50.39
68.72 125.35
50.64 107.44
 9.24  17.39
10.00  19.84
12.64  23.01
run;
proc reg data=old;
model  LS=LY;
run;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: LS

                             Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1        3.74287        3.74287     213.44    0.0001
Error                     4        0.07014        0.01754
Corrected Total           5        3.81301

                        Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1        0.65606        0.21754       3.02      0.0393
LY            1        0.98574        0.06747      14.61      0.0001
```

```
*SAS Code for analysis of transformed data;
* crab_logtrans.sas;
* A natural logarithm transformation is applied to the crab counts;

option ls=120 ps=55 nocenter nodate;
title 'Hermit Crab Density - LogTransformation';
data count;
infile 'u:\meth2\kuehl\expl4-1.dat';
input Y Site;
label Y = 'Crab Count';
*analysis of transformed data;
data trans;
set count;
label TY1 = 'Transformed Count #2';
TY1 = log(Y+1);
proc boxplot;
plot ty1*site/boxstyle=schematic;
run;
proc glm data=trans;
class Site;
model  TY1 = Site;
*BFL Test;
means Site/hovtest=bf;
means Site/ tukey;
output out=ASSUMP2 r=RESID2 p=MEANS2 STUDENT=SR2;
data TRANSRESID2; set ASSUMP2; TRS2 = sqrt(abs(SR2));
proc univariate def=5 plot normal; var RESID2;
proc gplot data=ASSUMP2; plot RESID2*MEANS2;
proc gplot data=TRANSRESID2; plot TRS2*MEANS2;
run;
```

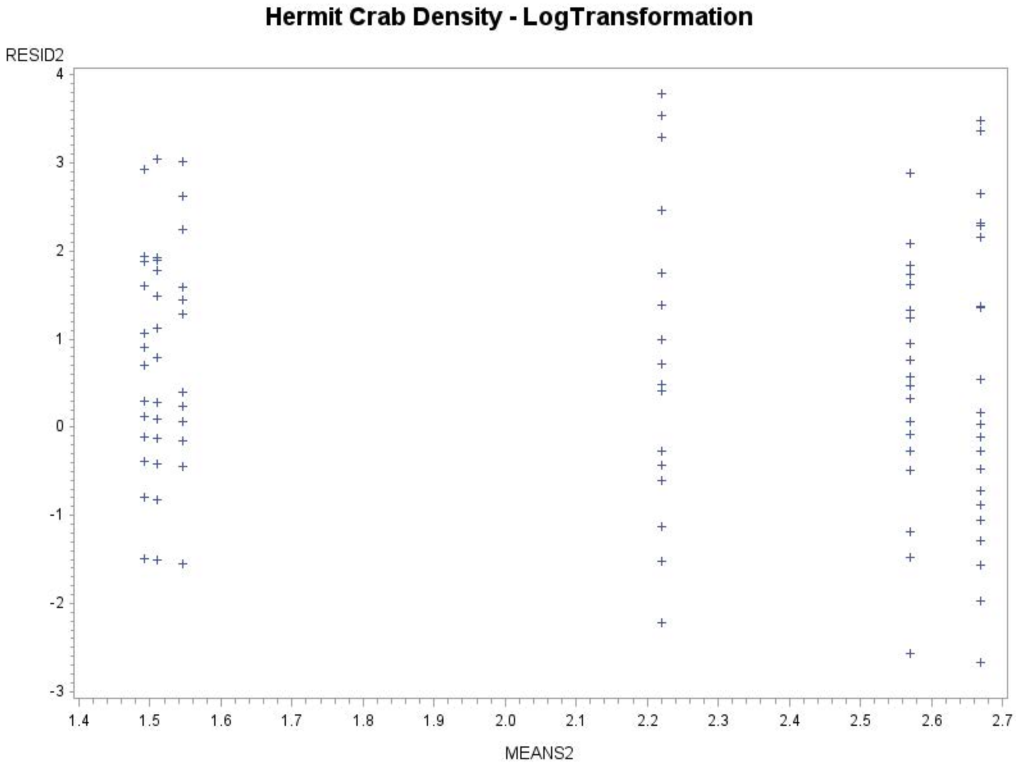
Box Plots of Crab Counts: TY1 = $\log(Y_{ij} + 1)$



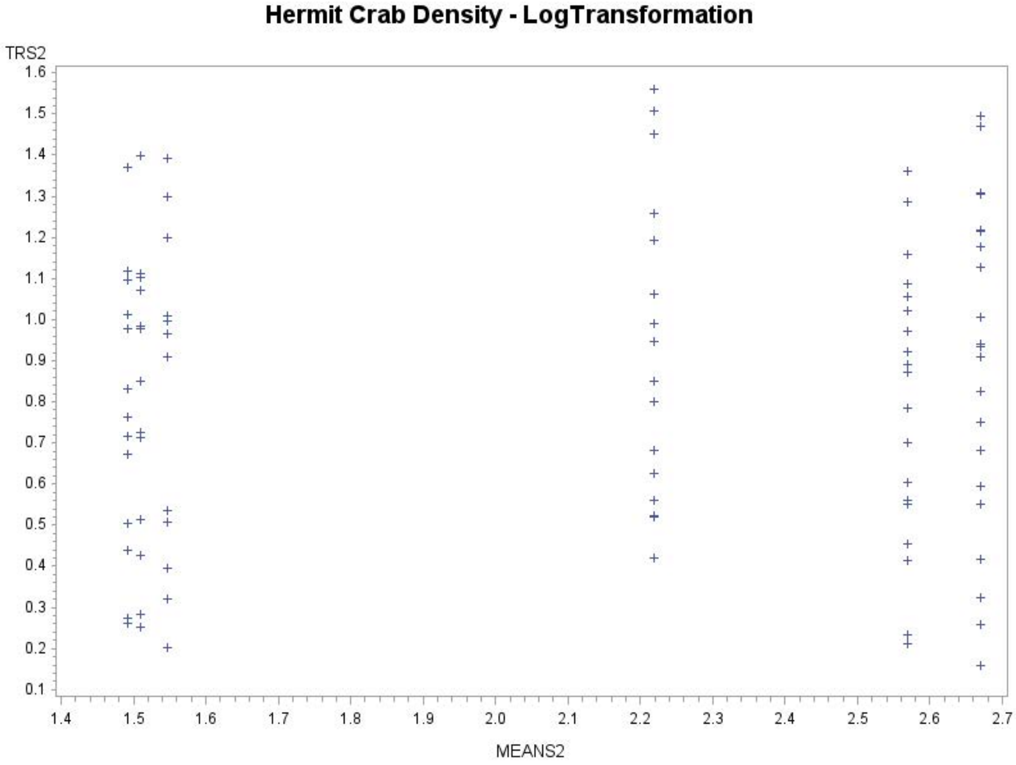
```
Hermit Crab Density - LogTransformation

      Class Level Information

Class         Levels    Values
Site               6    1 2 3 4 5 6
Number of Observations Read         150

Dependent Variable: TY1   Transformed Count log(Y+1)

                                        Sum of
Source                     DF          Squares     Mean Square    F Value    Pr > F
Model                       5       38.1018490       7.6203698       3.02    0.0128
Error                     144      363.6739366       2.5255134
Corrected Total           149      401.7757856

Source                     DF        Type I SS     Mean Square    F Value    Pr > F
Site                        5      38.10184900      7.62036980       3.02    0.0128


  Brown and Forsythe's Test for Homogeneity of TY1 Variance
       ANOVA of Absolute Deviations from Group Medians

                      Sum of        Mean
Source        DF     Squares      Square    F Value    Pr > F
Site           5      7.7658      1.5532       1.53    0.1851
Error        144       146.5      1.0174


Level of            -------------TY1-------------
Site          N             Mean          Std Dev

1            25       2.56993820       1.63724006
2            25       2.66937890       1.93144336
3            25       2.21980088       1.86120401
4            25       1.49249880       1.25111453
5            25       1.51019426       1.24609121
6            25       1.54677926       1.46967377

Tukey's Studentized Range (HSD) Test for TY1


Means with the same letter are not significantly different.



Groups     Mean      N    Site

A        2.6694     25    2
A
A        2.5699     25    1
A
A        2.2198     25    3
A
A        1.5468     25    6
A
A        1.5102     25    5
A
A        1.4925     25    4
```

```
                   Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.971165    Pr < W      0.0030
Kolmogorov-Smirnov    D     0.080179    Pr > D      0.0190
Cramer-von Mises      W-Sq  0.181786    Pr > W-Sq   0.0090
Anderson-Darling      A-Sq  1.140381    Pr > A-Sq   0.0055



   Stem Leaf                         #  Boxplot                        Normal Probability Plot
      3 558                          3     |        3.75+                                                *+*
      3 0034                         4     |            |                                          ******+
      2 56799                        5     |            |                                        *****+
      2 12233                        5     |            |                                     ****++
      1 5666678889999               13     |            |                                  ****
      1 001112334444                12  +-----+         |                               ****
      0 5556677889                  10  |     |         |                             ***
      0 0111111112233344            16  |  +  |         |                         +****
     -0 444444444444433322111111    24  *-----*         |                      *****
     -0 9888888877666655            16  |     |         |                   ****
     -1 3211                         4  |     |         |                +***
     -1 65555555555555555555555     23  +-----+         |            ******
     -2 222220                       6     |            |         +***
     -2 777766666                    9     |       -2.75+* * ******
        ----+----+----+----+----                         +----+----+----+----+----+----+----+----+----+----+
                                                             -2        -1         0        +1        +2
```

Residual Plots for Transformed Crab Counts:



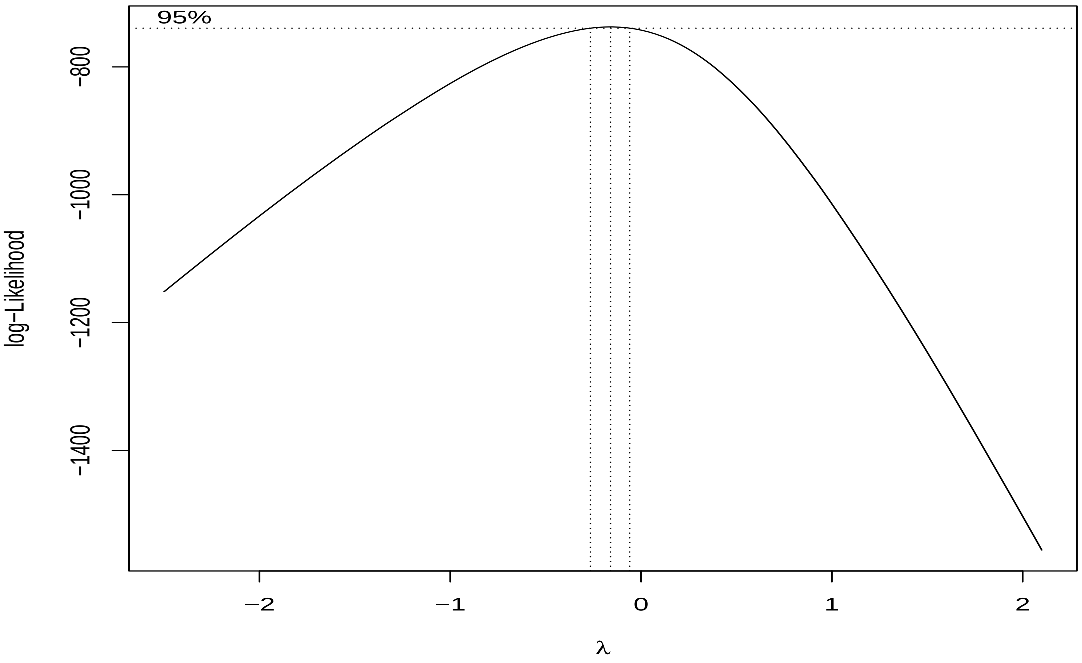
Square Root of Absolute Residual Plots for Transformed Crab Counts:

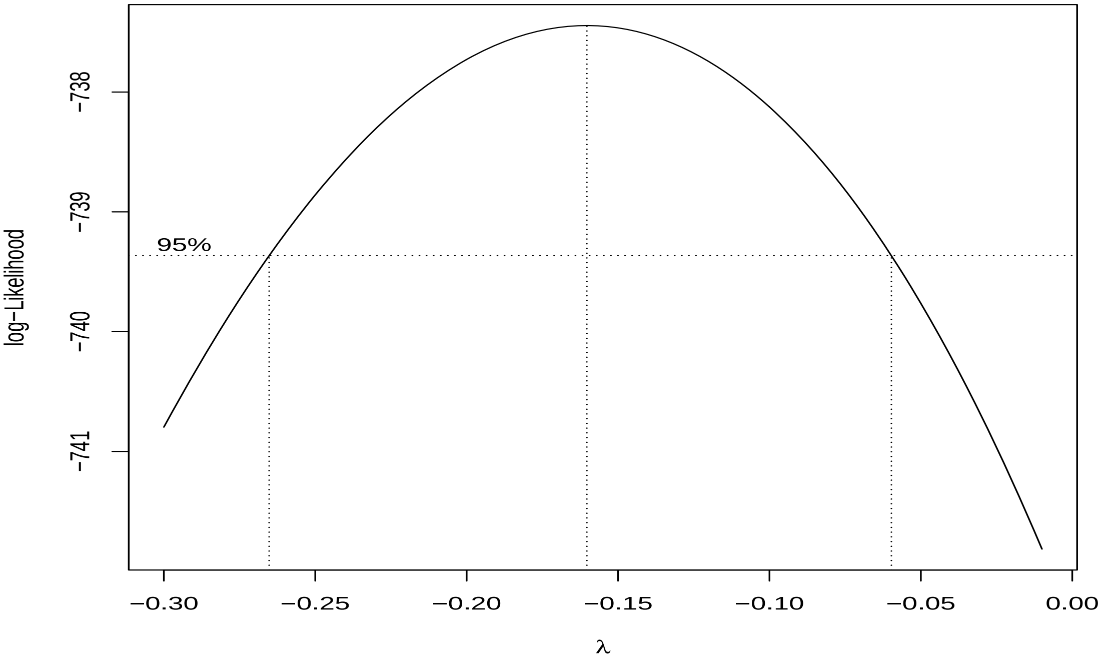


EXAMPLE: Box-Cox transformation

```
# boxcox_Crabs_V2.R
# R CODE FOR THE COMPUTATION FOR BOX-COX TRANSFORMATION
# OF CRAB COUNT DATA:  EXAMPLE 4.1 IN TEXTBOOK
 data =  matrix(0,150,2)
 y =  matrix(0,150,1)

yhab1 = c(0,0,22,3,17,0,0,7,11,11,73,33,0,65,13,44,20,27,48,104,233,81,22,9,2)
yhab2 = c(0,0,56,0,8,0,3,1,16,55,142,10,2,145,6,4,5,124,24,204,415,466,6,14,12)
yhab3 = c(0,0,4,13,5,1,1,4,4,36,407,0,0,18,4,14,0,24,52,314,245,107,5,6,2)
yhab4 = c(0,0,0,4,2,2,5,4,2,1,0,12,1,30,0,3,28,2,21,8,82,12,10,2,0)
yhab5 = c(0,1,1,2,2,1,2,29,2,2,0,13,0,19,1,3,26,30,5,4,94,1,9,3,0)
yhab6 = c(0,0,0,2,3,0,0,4,0,5,4,22,0,64,4,4,43,3,16,19,95,6,22,0,0)
y = c(yhab1,yhab2,yhab3,yhab4,yhab5,yhab6)

s1 = rep("h1",25)
s2 = rep("h2",25)
s3 = rep("h3",25)
s4 = rep("h4",25)
s5 = rep("h5",25)
s6 = rep("h6",25)
hab = c(s1,s2,s3,s4,s5,s6)

site = as.factor(hab)

library(MASS)

like=boxcox(y+1˜site,lambda=seq(-2.5,2.1,.01))

like_max=max(like$y)
imax = which(like$y==like_max)
thmax=like$x[imax]

thmax
[1] -0.16

like2=boxcox(y+1˜site,lambda=seq(-0.3,-0.01,.0001))
```

Note that the maximum value of the likelihood function and hence the minimum MSE occurs at $\lambda = -0.16$ which is very close to 0. Hence the log transformation would be a logical choice for the transformation.


95%
-800
-1000
-1200
-1400
log-Likelihood
−2
−1
0
1
2
λ


95%
-738
-739
-740
-741
log-Likelihood
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
λ

A Distribution-Free AOV Procedure: Kruskal-Wallis Test

The Kruskal-Wallis test is a generalization of the Wilcoxon Rank Sum procedure. It is used for testing the research hypothesis that there is a shift difference in the t treatment populations. It retains all the conditions required of the AOV F-test except the normality condition. In particular, the required conditions are as follows:

C_1. The n random variables $[y_{i1}, y_{i2}, \cdots, y_{in_i}], i = 1, 2, \cdots, t$ are mutually independent.

C_2. For each fixed $i = 1, 2, \cdots, t$, the n_i r.v.'s $[y_{i1}, y_{i2}, \cdots, y_{in_i}]$ are a random sample from a continuous distribution with cdf G_i.

C_2. The distribution functions $G_1, \cdots, G_t$ are related through the relationship

$$G_i(y) = G(y - \tau_i), \quad \text{for} \quad -\infty < y < \infty,$$

for $i = 1, \cdots, t$, where G is a distribution function from a continuous distribution with unknown location parameter θ and τ_i is the unknown treatment effect of the ith population.

Conditions, C_1, C_2, C_3, are equivalent to the following model:

$$y_{ij} = \theta + \tau_i + e_{ij}, \quad i = 1, \cdots, t; \quad j = 1, \cdots, n_i,$$

where θ is the overall median, τ_i is the effect of Treatment i, and the n e_{ij}'s are iid r.v.'s from a continuous distribution G with location parameter 0.

Note, if we specified that G was a normal cdf, then we would have **exactly** the same conditions as in the normal based AOV tests. Since the median would be the mean under a symmetric distribution and the common cdf G would require that all the variances are equal (provided that they exist). Because the K-W test is appropriate for any cdf G, whereas the AOV F is just for normally distributed responses, the K-W test tends to be less powerful than the F test when the data is normally distributed.

The research hypothesis is that there is a Treatment difference, that is, at least one τ_i is different from the rest:

$$H_o : \tau_1 = \cdots = \tau_t \quad \text{vs} \quad H_1 : \tau_1, \cdots, \tau_t \text{ not all equal}$$

The null hypothesis is equivalent to having $G_1 \equiv G_2 \equiv G_t \equiv G$, that is, the treatment populations have **identical** distributions. This would be identical to our normal-based AOV procedures since under the null hypothesis the treatment means are identical and hence the distributions are identical since they have a normal distribution with a common variance under both the null and alternative hypotheses.

Kruskal-Wallis Procedure

1. Combine the $n = \sum_{i=1}^{t} n_i$ observations from the t samples and rank them from smallest to largest. Let R_{ij} be the rank of y_{ij} in the combined sample. Note: R_{ij}'s are an arrangement of the numbers $1, \cdots, n$.

2. Compute the total and mean rank for each treatment:

   $$R_{i.} = \sum_{j=1}^{n_i} R_{ij}; \qquad \bar{R}_{i.} = \frac{R_{i.}}{n_i}; \qquad \bar{R}_{..} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i} R_{ij}}{n} = \frac{1+2+\cdots+n}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2}$$

   and note that $\sum_{i=1}^{t}\sum_{j=1}^{n_i} R_{ij}^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$

   We now have $\bar{R}_{i.}$ is the mean rank of the observations from the ith treatment and $\bar{R}_{..}$ is the mean rank of all observations.

3. The Kruskal-Wallis statistic KW is then computed as follows:

   $$\begin{aligned} KW = \frac{SS_{TRT}}{SS_{TOT}} &= \frac{\sum_{i=1}^{t} n_i \left(\bar{R}_{i.} - \bar{R}_{..}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(R_{ij} - \bar{R}_{..}\right)^2} \\ &= \frac{\sum_{i=1}^{t} \frac{R_{i.}^2}{n_i} - n\left(\frac{n+1}{2}\right)^2}{\frac{1}{n-1}\left(\sum_{i=1}^{t}\sum_{j=1}^{n_i} R_{ij}^2 - n\left(\frac{n+1}{2}\right)^2\right)} \\ &= \frac{\sum_{i=1}^{t} \frac{R_{i.}^2}{n_i} - \frac{n(n+1)^2}{4}}{\frac{1}{n-1}\left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}\right)} \\ &= \frac{\sum_{i=1}^{t} \frac{R_{i.}^2}{n_i} - \frac{n(n+1)^2}{4}}{\frac{n(n+1)}{12}} \\ &= \frac{12}{n(n+1)}\sum_{i=1}^{t}\frac{R_{i.}^2}{n_i} - 3(n+1) \end{aligned}$$

always use this.

4. The distribution of KW, under H_o, does not depend on the distribution of data values: y_{ij}

5. We reject H_o at level α if $KW \geq h_\alpha$, where h_α are given in Table A.12 of the book, *Nonparametric Statistical Methods, 2nd Ed.*, by Hollander and Wolfe.

6. ✱ A large sample approximation is obtained by using:

   Reject H_o at level α if $KW \geq \chi^2_{t-1,\alpha}$, where $\chi^2_{t-1,\alpha} = qchisq(1-\alpha, t-1)$ is the upper α-percentile from the Chi-square Distribution with $df = t - 1$.

7. If there are ties among the n y_{ij}'s, assign each of the observations in a tied group the **average** of the integer ranks that are associated with the tied group and compute H with these average ranks. The following modification must be made to H due to the occurrence of the ties:

   $$KW^{'} = \frac{KW}{1 - \left(\sum_{k=1}^{m}(w_k^3 - w_k)/(n^3 - n)\right)}$$

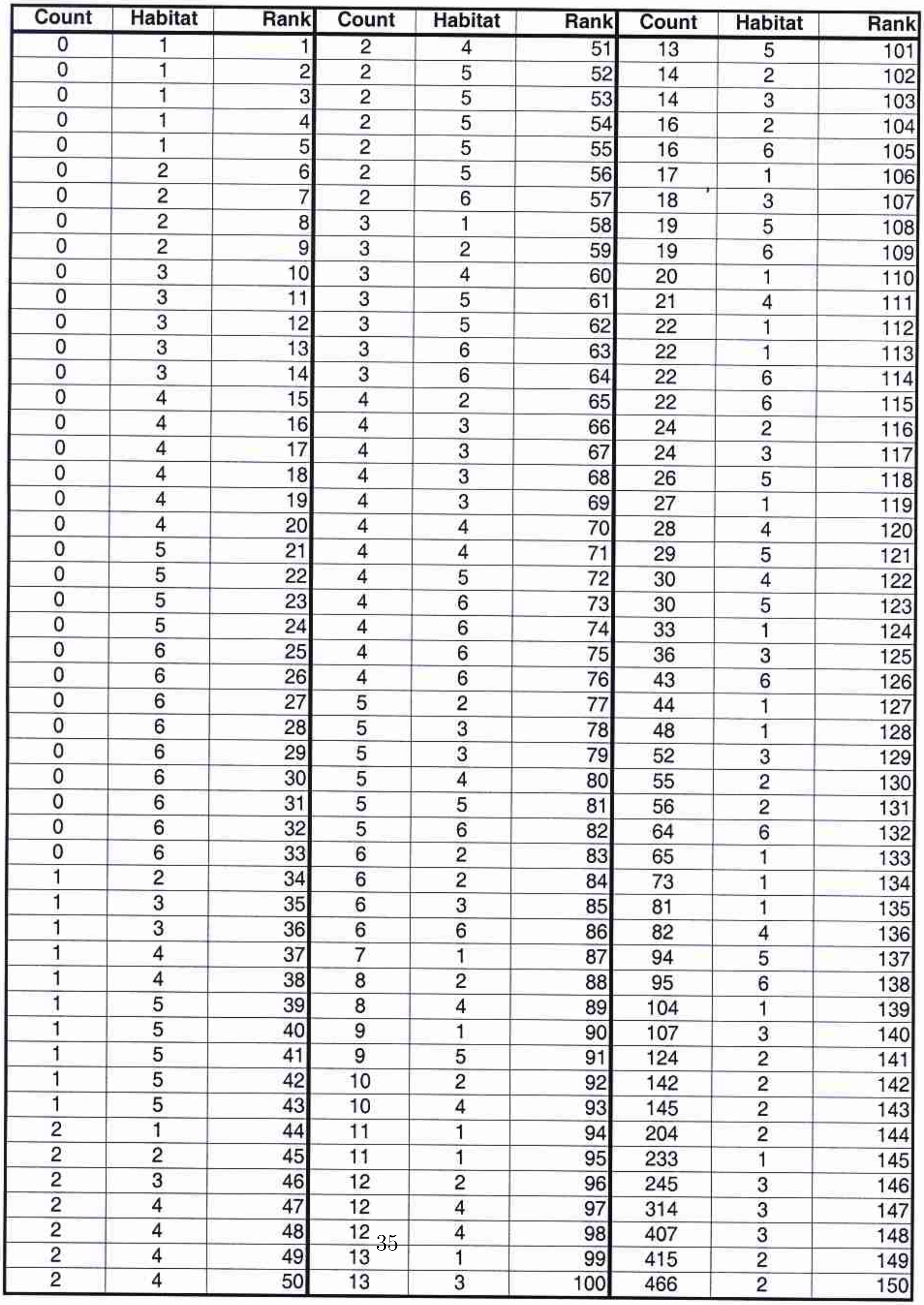
   where KW is computed using average ranks, m is the number of groups of $y_{ij} - ties$, w_k is the number of y_{ij}'s in the kth group of ties. Using $KW^{'}$ results in only an approximate test but in most cases the differences between KW and $KW^{'}$ is minimal unless there is large number of groups of ties.

SAS code to compute the Kruskal-Wallis Test with a Multiple Comparison Procedure - DSCF;

```
ods html; ods graphics on;

option ls=70 ps=50 nocenter nodate;
title 'Hermit Crab Density';

*Input Data;
data count;
infile 'u:\meth2\kuehl\expl4-1.dat';
input Y Site;
label Y = 'Crab Count';
proc npar1way anova wilcoxon DSCF;
var Y;
class Site;
run;
```

The SAS output for the Kruskal-Wallis test is given below.

```
Hermit Crab Density Analysis Using Nonparametric Procedures

The NPAR1WAY Procedure


             Wilcoxon Scores (Rank Sums) for Variable Y
                    Classified by Variable Site

                    Sum of      Expected       Std Dev          Mean
Site       N        Scores      Under H0      Under H0         Score
--------------------------------------------------------------------
1         25       2290.00       1887.50    197.059089        91.600
2         25       2268.50       1887.50    197.059089        90.740
3         25       1999.00       1887.50    197.059089        79.960
4         25       1577.50       1887.50    197.059089        63.100
5         25       1582.50       1887.50    197.059089        63.300
6         25       1607.50       1887.50    197.059089        64.300

                Average scores were used for ties.

   Kruskal-Wallis Test

Chi-Square         12.5996
DF                       5
Pr > Chi-Square     0.0274
```



Count	Habitat	Rank	Count	Habitat	Rank	Count	Habitat	Rank
0	1	1	2	4	51	13	5	101
0	1	2	2	5	52	14	2	102
0	1	3	2	5	53	14	3	103
0	1	4	2	5	54	16	2	104
0	1	5	2	5	55	16	6	105
0	2	6	2	5	56	17	1	106
0	2	7	2	6	57	18	3	107
0	2	8	3	1	58	19	5	108
0	2	9	3	2	59	19	6	109
0	3	10	3	4	60	20	1	110
0	3	11	3	5	61	21	4	111
0	3	12	3	5	62	22	1	112
0	3	13	3	6	63	22	1	113
0	3	14	3	6	64	22	6	114
0	4	15	4	2	65	22	6	115
0	4	16	4	3	66	24	2	116
0	4	17	4	3	67	24	3	117
0	4	18	4	3	68	26	5	118
0	4	19	4	3	69	27	1	119
0	4	20	4	4	70	28	4	120
0	5	21	4	4	71	29	5	121
0	5	22	4	5	72	30	4	122
0	5	23	4	6	73	30	5	123
0	5	24	4	6	74	33	1	124
0	6	25	4	6	75	36	3	125
0	6	26	4	6	76	43	6	126
0	6	27	5	2	77	44	1	127
0	6	28	5	3	78	48	1	128
0	6	29	5	3	79	52	3	129
0	6	30	5	4	80	55	2	130
0	6	31	5	5	81	56	2	131
0	6	32	5	6	82	64	6	132
0	6	33	6	2	83	65	1	133
1	2	34	6	2	84	73	1	134
1	3	35	6	3	85	81	1	135
1	3	36	6	6	86	82	4	136
1	4	37	7	1	87	94	5	137
1	4	38	8	2	88	95	6	138
1	5	39	8	4	89	104	1	139
1	5	40	9	1	90	107	3	140
1	5	41	9	5	91	124	2	141
1	5	42	10	2	92	142	2	142
1	5	43	10	4	93	145	2	143
2	1	44	11	1	94	204	2	144
2	2	45	11	1	95	233	1	145
2	3	46	12	2	96	245	3	146
2	4	47	12	4	97	314	3	147
2	4	48	12	4	98	407	3	148
2	4	49	13	1	99	415	2	149
2	4	50	13	3	100	466	2	150

Kruskal-Wallis Tests Applied to Crab Count Data

$$\begin{aligned}
KW &= \frac{12}{n(n+1)}\sum_{i=1}^{t}\frac{R_{i.}^2}{n_i} - 3(n+1) \\
&= \frac{12}{150(150+1)}\sum_{i=1}^{6}\frac{R_{i.}^2}{25} - 3(151) \\
&= \left(\frac{12}{150(150+1)(25)}\right)\left[(2290)^2+(2268.5)^2+(1999)^2+(1577.5)^2+(1582.5)^2+(1607.5)^2\right] - 3(151) \\
&= 12.4424
\end{aligned}$$

$p - value = P[KW \geq 12.4424] \approx 1 - G(12.4424) = 1 - pchisq(12.4424, 5) = .0292,$

Adjustment for ties: w_k is the number of data values in the kth Group of Ties

Tied Value	0	1	2	3	4	5	6	8	9	10	11	12	13	14	16	19	22	24	30
w_k	33	10	14	7	12	6	4	2	2	2	2	3	3	2	2	2	4	2	2

Correction to H for Ties:

$$\begin{aligned}
C &= 1 - \frac{1}{(n^3-n)}\sum_{k=1}^{g}(w_k^3 - w_k) \\
&= 1 - \frac{1}{(150^3-150)}\left[(33^3-33)+(10^3-10)+(14^3-14)+\cdots+(2^3-2)\right] \\
&= 1 - \frac{1}{(150^3-150)}(42108) = .987523
\end{aligned}$$

$$KW^{'} = KW/C = (12.4424)/(.987523) = 12.5996$$

$p - value = P[KW \geq 12.5996] \approx 1 - G(12.5996) = 1 - pchisq(12.5996, 5) = .0274,$

where G is the cdf of chi-square distribution with $df = t - 1 = 5$

Did the crab data satisfy the conditions for applying the Kruskal-Wallis Test?

Probably not because there is a large difference in the variances for the 6 sites. However, if we apply the log-transformation to the data which yields populations with approximately the same level of variability, we will obtain exactly the same results that we obtained for the untransformed data. Why?

Using the R code on the next page for the Kruskal-Wallis Test for Crab Data, we can determine that the R function **kruskal.test()** is correcting for ties.

```
RCode: KruskalWallis_Crabs.S

 data =  matrix(0,150,2)
 y =  matrix(0,150,1)

yhab1 = c(0,0,22,3,17,0,0,7,11,11,73,33,0,65,13,44,20,27,48,104,233,81,22,9,2)
yhab2 = c(0,0,56,0,8,0,3,1,16,55,142,10,2,145,6,4,5,124,24,204,415,466,6,14,12)
yhab3 = c(0,0,4,13,5,1,1,4,4,36,407,0,0,18,4,14,0,24,52,314,245,107,5,6,2)
yhab4 = c(0,0,0,4,2,2,5,4,2,1,0,12,1,30,0,3,28,2,21,8,82,12,10,2,0)
yhab5 = c(0,1,1,2,2,1,2,29,2,2,0,13,0,19,1,3,26,30,5,4,94,1,9,3,0)
yhab6 = c(0,0,0,2,3,0,0,4,0,5,4,22,0,64,4,4,43,3,16,19,95,6,22,0,0)
y = c(yhab1,yhab2,yhab3,yhab4,yhab5,yhab6)
s1 = rep("h1",25)
s2 = rep("h2",25)
s3 = rep("h3",25)
s4 = rep("h4",25)
s5 = rep("h5",25)
s6 = rep("h6",25)
hab = c(s1,s2,s3,s4,s5,s6)
site = as.factor(hab)
d=data.frame(y,site)

kruskal.test(y,site,y˜site)

#Output from R:
#data:  y and site
#Kruskal-Wallis chi-squared = 12.5996, df = 5, p-value = 0.02743
```

```
Kruskal-Wallis Test applied to ytrans = log(y+1)

kruskal.test(ytrans,site,ytrans˜site)

        Kruskal-Wallis rank sum test

data:  ytrans and site
Kruskal-Wallis chi-squared = 12.5996, df = 5, p-value = 0.02743
```

Multiple Comparison Procedure Using Ranks

Procedure I (Hollander-Wolfe)

1. Calculate the t(t-1)/2 absolute differences $|\bar{R}_{i.} - \bar{R}_{h.}|$ for $i < h$, where the $\bar{R}_{i.}$'s are the mean ranks for the *ith* treatment.
2. At an familywise error rate $\alpha_F < \alpha$,

$$\text{Declare} \;\; \tau_i \neq \tau_h \;\; \text{if} \;\; |\bar{R}_{i.} - \bar{R}_{h.}| \geq \sqrt{h_\alpha \left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i} + \frac{1}{n_h}\right)}$$

$$\text{If} \;\; n_i = r, \;\; \text{Declare} \;\; \tau_i \neq \tau_h \;\; \text{if} \;\; |\bar{R}_{i.} - \bar{R}_{h.}| \geq \sqrt{h_\alpha \left(\frac{2t(n+1)}{12}\right)},$$

where h_α is the critical value for the Kruskal-Wallis test (Table A.12 in Hollander-Wolfe Book).

Procedure II (Miller,1966)- Large Sample Approximation When $n_1 = \cdots = n_t = r$, with r large:

1. Calculate the t(t-1)/2 absolute differences $|\bar{R}_{i.} - \bar{R}_{h.}|$ for $i < h$, where the $\bar{R}_{i.}$'s are the mean ranks for the *ith* treatment.
2. At an familywise error rate $\alpha_F < \alpha$,

$$\text{Declare} \;\; \tau_i \neq \tau_h \;\; \text{if} \;\; |\bar{R}_{i.} - \bar{R}_{h.}| \geq q(\alpha, t, \infty)\sqrt{\left(\frac{t(n+1)}{12}\right)},$$

where $q(\alpha, t, \infty) = qtukey(1-\alpha, t, 10000)$ is the critical value for the Studentized Range (Table VII in Kuehl's Book).

Procedure III (Dunn,1964)- Large Sample Approximation When n_i's are unequal, with n_i's large:

1. Calculate the t(t-1)/2 absolute differences $|\bar{R}_{i.} - \bar{R}_{h.}|$ for $i < h$, where the $\bar{R}_{i.}$'s are the mean ranks for the *ith* treatment.
2. At an familywise error rate $\alpha_F < \alpha$,

$$\text{Declare} \;\; \tau_i \neq \tau_h \;\; \text{if} \;\; |\bar{R}_{i.} - \bar{R}_{h.}| \geq Z_{\frac{\alpha/2}{M}}\sqrt{\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i} + \frac{1}{n_h}\right)},$$

where $Z_{\frac{\alpha/2}{M}}$ is the upper $\frac{\alpha/2}{M}$-percentile of $N(0,1)$ distribution and $M = \frac{t(t-1)}{2} = 15$.

Procedure IV Dwass, Steel, Critchlow-Fligner Method

Procedure used in SAS for multiple comparison. See Critchlow, D. E., Fligner, M. A. (1991). On distribution-free multiple comparisons in the one-way analysis of variance. Commun. Statist. Theor. Meth. 20:127–139.

EXAMPLE OF MULTIPLE COMPARISON USING RANKS

Determine the pairs of Sites for which there is significant evidence ($\alpha_E = .05$) of a difference in the distribution of crab counts.

Using the crab count data, $n_i = 25$ for $i = 1, \ldots, 6$; $n = 6(25) = 150$

- Using Miller's Method: Let

$$D_M(\alpha, t) = q(\alpha, t, \infty)\sqrt{\tfrac{t(n+1)}{12}} = (4.03)\sqrt{\tfrac{6(150+1)}{12}} = 35.02$$

where $q(.05, 6, \infty) = qtukey(.95, 6, 10000)$

- Using Dunn's Method: Let

$$D_D(\alpha, t) = Z_{\frac{\alpha/2}{M}}\sqrt{\left(\tfrac{n(n+1)}{12}\right)\left(\tfrac{2}{r}\right)} = Z_{.025/15}\sqrt{\left(\tfrac{150(150+1)}{12}\right)\left(\tfrac{2}{25}\right)} = 36.07$$

- For Miller: Declare $\tau_i \neq \tau_h$ if $|\bar{R}_{i.} - \bar{R}_{h.}| \geq 35.02$
- For Dunn: Declare $\tau_i \neq \tau_h$ if $|\bar{R}_{i.} - \bar{R}_{h.}| \geq 36.07$

Not much difference in the two procedures for large n.

The following R function will perform the Dunn procedure on the Crab data:

```
# Multiple Comparison using the Dunn Procedure

Add following to R code on page 42


install.packages("pgirmess")
library("pgirmess")
kruskalmc(y~site)

Output from R:

Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
      obs.dif critical.dif difference
h1-h2    0.86     36.06833      FALSE
h1-h3   11.64     36.06833      FALSE
h1-h4   28.50     36.06833      FALSE
h1-h5   28.30     36.06833      FALSE
h1-h6   27.30     36.06833      FALSE
h2-h3   10.78     36.06833      FALSE
h2-h4   27.64     36.06833      FALSE
h2-h5   27.44     36.06833      FALSE
h2-h6   26.44     36.06833      FALSE
h3-h4   16.86     36.06833      FALSE
h3-h5   16.66     36.06833      FALSE
h3-h6   15.66     36.06833      FALSE
h4-h5    0.20     36.06833      FALSE
h4-h6    1.20     36.06833      FALSE
h5-h6    1.00     36.06833      FALSE
```

Site	1	2	3	4	5	6
$\bar{R}_{i.}$	91.60	90.74	79.96	63.10	63.30	64.30
Groupings	A	A	A	A	A	A

Thus, there is not significant evidence that any of the pairs of Sites differ with respect to their distributions of Hermit Crab Counts. Using the Kruskal-Wallis tests, there was significant evidence of a difference in the distribution of Hermit Crabs across the six sites. The multiple comparison were unable to determine which sites differ because these types of procedures have relatively low power. Hence, the differences in the distributions would need to be quite large in order for the Miller or Dunn procedure to declare pairs of sites to be different.

Hermit Crab Density

The NPAR1WAY Procedure

Pairwise Two-Sided Multiple Comparison Analysis

Dwass, Steel, Critchlow-Fligner Method

Variable: Y

Site	Wilcoxon Z	DSCF Value	Pr > DSCF
1 vs. 2	0.0681	0.0963	1.0000
1 vs. 3	0.8964	1.2678	0.9475
1 vs. 4	2.3919	3.3826	0.1589
1 vs. 5	2.2802	3.2247	0.2021
1 vs. 6	2.2281	3.1510	0.2247
2 vs. 3	0.9348	1.3220	0.9376
2 vs. 4	2.2235	3.1446	0.2268
2 vs. 5	2.2108	3.1265	0.2326
2 vs. 6	2.1353	3.0198	0.2690
3 vs. 4	1.3683	1.9350	0.7461
3 vs. 5	1.3562	1.9180	0.7532
3 vs. 6	1.2978	1.8353	0.7864
4 vs. 5	0.0196	0.0277	1.0000
4 vs. 6	-0.0887	0.1254	1.0000
5 vs. 6	-0.0785	0.1110	1.0000

STOP Wednesday 2/23/22 (Week 6, lecture 16)

START Friday 2/25/22 (week 6, lecture 17)

GLIM: Generalized Linear Models

The GENMOD Procedure in SAS provides a methodology for fitting non-normal error terms to a data set.

This is a brief introduction to the theory of generalized linear models (most of the material is from SAS documentation). The modelling of discrete data is a central topic in STAT 659, STAT 645, and STAT 646.

Response Probability Distributions

In generalized linear models, the response is assumed to possess a probability distribution of the exponential form. That is, the probability density of the response Y for continuous response variables, or the probability mass function for discrete responses, can be expressed as

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

for specified functions $a(\cdot), b(\cdot)$, and $c(\cdot)$ that determine the specific distribution. For fixed ϕ, this is a one parameter exponential family of distributions with canonical parameter θ. The function $a(\cdot)$ is such that $a(\phi) = \phi/w_i$, where w_i is a known weight for each observation, y_i. A variable representing w_i in the input data set may be specified in the WEIGHT statement. If no WEIGHT statement is specified, $w_i = 1$ for all observations. For example, suppose in our normal based AOV model, suppose we take $y_i = \bar{y}_{i\cdot}$. Then, $a(\phi) = \frac{\sigma_e^2}{n_i}$ hence, $w_i = n_i, \phi = \sigma_e^2$.

Standard theory for this type of distribution gives expressions for the mean and variance of Y:

$$E[Y] = b^{'}(\theta) \qquad \text{and} \qquad Var(Y) = \frac{b''(\theta)\phi}{w}$$

where the primes denote derivatives with respect to θ. If μ_i represents the mean of Y_i, then the variance expressed as a function of the mean is given by

$$Var(Y_i) = \frac{V(\mu_i)\phi}{w},$$

where V is the variance function. Probability distributions for the response Y_i in generalized linear models are usually parameterized in terms of the mean and dispersion parameter instead of the natural parameter .

The probability distributions that are available in the GENMOD procedure are shown in the following list. The PROC GENMOD scale parameter and the variance of Y are also shown.

Normal: $f(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$; for $-\infty < y < \infty$ with

$\phi = \sigma^2$; Scale $= \sigma$ $b(\theta) = \theta^2/2$ $Var(Y) = \sigma^2$

Inverse Gaussian: $f(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi y^3}} exp\left[\frac{-1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right]$; for $0 < y < \infty$ with

$\phi = \sigma^2$; Scale $= \sigma$; $b(\theta) = -(-2\theta)^{1/2}$; $Var(Y) = \mu^3\sigma^2$

Gamma: $f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)y} exp\left(\frac{y\nu}{\mu}\right)^\nu exp\left(\frac{-y\mu}{\mu}\right)$; for $0 < y < \infty$ with

$\phi = \frac{1}{\nu}$; Scale $= \nu$; $b(\theta) = -log(-\theta)$; $Var(Y) = \frac{\mu^2}{\nu}$

Negative Binomial: $f(y;\alpha) = \frac{(y+k-1)!}{y!(k-1)!}\frac{\alpha^y}{(1+\alpha)^{y+k}}$ for $y = 0, 1, 2\ldots$;

Scale $= k$; $b(\theta) = -klog(1-e^\theta)$; $Var(Y) = \mu + k\mu^2$

Poisson: $f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$ for $y = 0, 1, 2, \ldots$

$\phi = 1$; $b(\theta) = exp(\theta)$; $Var(Y) = \mu$

Binomial: $f(y;\mu) = \binom{n}{r}\mu^r(1-\mu)^{n-r}$ for $y = \frac{r}{n}; r = 0, \ldots, n$

$\phi = 1$; $b(\theta) = log(1+e^\theta)$; $Var(Y) = \frac{\mu(1-\mu)}{n}$

where θ is related to μ through $\mu = E[Y] = b^{'}(\theta)$.

Link Function:

In our linear model, we have the mean vector of the response variable related to a linear combination of explanatory variables through

$$\mu_i = \mathbf{X}_i\boldsymbol{\beta}$$

where $\mathbf{X}_i$ is a fixed known vector of explanatory variables, and $\boldsymbol{\beta}$ is a vector of unknown parameters.

In the generalized linear model, the mean μ_i of the response in the ith observation,y_i is related to a linear predictor through a monotonic differentiable link function $g(\cdot)$.

$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$

In the classical linear model, $g(\cdot)$ is just the identity function, $g(\mu_i) = \mu_i$

Log-Likelihood Functions:

Log-likelihood functions for the distributions that are available in GENMOD are parameterized in terms of the means μ_i and the dispersion parameter ϕ. The term y_i represents the response of the ith observation and w_i represents the known dispersion weight. The log-likelihood functions are of the form:

$$L(\mathbf{y}, \boldsymbol{\mu}, \phi) = \sum_{i=1}^{n} log(f(y_i, \mu_i, \phi)).$$

The forms of the individual contributions: $l_i = log(f(y_i, \mu_i, \phi))$ are shown below. The parameterizations are expressed in terms of the mean and dispersion parameters.

1. Normal: $l_i = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{\phi} + log\left(\frac{\phi}{w_i}\right) + log(2\pi)\right]$

2. Inverse Gaussian: $l_i = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{y_i\mu_i^2\phi} + log\left(\frac{\phi y_i^3}{w_i}\right) + log(2\pi)\right]$

3. Gamma: $l_i = \frac{w_i}{\phi} log\left(\frac{w_i y_i}{\phi\mu_i}\right) - \frac{w_i y_i}{\phi\mu_i} - log(y_i) - log\left(\Gamma\left(\frac{w_i}{\phi}\right)\right)$

4. Negative Binomial: $l_i = y_i log\left(\frac{k\mu_i}{w_i}\right) - (y_i + \frac{w_i}{k}) log(1 + \frac{k\mu_i}{w_i}) + log\left(\frac{\Gamma\left(y_i + \frac{w_i}{k}\right)}{\Gamma(y_i+1)\Gamma(w_i/k)}\right)$

5. Poisson: $l_i = \frac{w_i}{\phi}\left[y_i log(\mu_i) - \mu_i)\right]$

6. Binomial: $l_i = \frac{w_i}{\phi}\left[r_i log(p_i) + (n_i - r_i) log(1 - p_i)\right]$

Maximum Likelihood Fitting:

The GENMOD procedure uses a ridge-stabilized Newton-Raphson algorithm to maximize the log-likelihood function with respect to the regression parameters. By default, the procedure also produces maximum likelihood estimates of the scale parameter as defined in the ”Response Probability Distributions” section for the normal, inverse Gaussian, negative binomial, and gamma distributions.

Over-dispersion:

Over-dispersion is a phenomenon that sometimes occurs in data that are modelled with the binomial or Poisson distributions. If the estimate of dispersion after fitting, as measured by the deviance or Pearson's chi-square, divided by the degrees of freedom, is not near 1, then the data may be over-dispersed if the dispersion estimate is greater than 1 or under-dispersed if the dispersion estimate is less than 1. A simple way to model this situation is to allow the variance functions of these distributions to have a multiplicative over-dispersion factor ϕ.

binomial: $V(\mu) = \phi\mu(1-\mu)$

Poisson: $V(\mu) = \phi\mu$

The models are fit in the usual way, and the parameter estimates are not affected by the value of ϕ.

Goodness of Fit:

For a set of observations $\mathbf{y} = (y_1, \ldots, y_n)$, let $\mathrm{L}(\boldsymbol{\mu}; \mathbf{y})$ denote the log-likelihood function, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, and $\mu_i = E(y_i)$.

Let $\mathrm{L}(\hat{\boldsymbol{\mu}}; \mathbf{y})$ denote the maximum of the log likelihood for the model. Considering all possible models, the maximum achievable value of the log likelihood is $\mathrm{L}(\mathbf{y}; \mathbf{y})$. That is the model in which there is a separate parameter for every observation and the perfect fit $\hat{\boldsymbol{\mu}} = \mathbf{y}$

Such a model is called a saturated model and it is not useful because it does not provide any reduction in the data. However, this model serves as the baseline for comparison to any model which is fit to the data.

The *deviance* of a Poisson regression model is defined to be

$$-2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})]$$

This is the likelihood-ratio statistics for testing the null hypothesis that there is not difference in the 6 sites against the hypothesis that the sites differ. Under the null hypothesis, the deviance has a chi-squared asymptotic distribution with df = n-p, where p is the number of model parameters. In the crab data, p=6, the number of sites. This statistics provides a test of model fit.

A reference to this topic is McCullogh and Nelder(1989), *Generalized Linear Models*, Chapman-Hall.

STAT 645, STAT 646, and STAT 659 cover this topic.

EXAMPLE: Using Proc Genmod in SAS to Fit Over-dispersed Poisson Model

```
*ods html;ods graphics on;
* crab,genmod.sas
 The relationship between different habitats and the population densities of
Hermit crabs. There are 6 sites. At each site 25 transects are run and the
number of crabs are counted. Analyze using an overdispersed Poisson Model ;

option ls=120 ps=50 nocenter nodate;
title 'Hermit Crab Density';
data count;
infile 'u:\meth2\kuehl\expl4-1.dat';
input Y Site;
label Y = 'Crab Count';

title "Poisson Regression on Hermit Crab Data";
proc genmod data=count;
class Site;
model  Y = Site/Dist=P link=log;
output out=outdata stdresdev = stdresdev p = predicted;
proc PLOT data = outdata;
plot stdresdev*Site;
run;
Title "Overdispersed Poisson Regression on Hermit Crab Data";
proc genmod data=count;
class Site;
model  Y = Site/dist=P link=log  scale = pearson;
contrast 'S2 vs S1' Site -1  1  0  0  0  0;
contrast 'S2 vs S3' Site  0  1 -1  0  0  0;
contrast 'S2 vs S4' Site  0  1  0 -1  0  0;
contrast 'S2 vs S5' Site  0  1  0  0 -1  0;
contrast 'S2 vs S6' Site  0  1  0  0  0 -1;
run;
output out=outdata2 stdresdev = stdresdev2 p = predicted;
proc PLOT data = outdata2;
plot stdresdev2*Site;
run;
*ods graphics off;
*ods html close;
```

```
                         Poisson Regression on Hermit Crab Data

                                  The GENMOD Procedure

                                   Model Information

                   Data Set                    WORK.COUNT
                   Distribution                   Poisson
                   Link Function                      Log
                   Dependent Variable                   Y    Crab Count


                            Number of Observations Read          150
                            Number of Observations Used          150


                               Class Level Information

                            Class       Levels    Values

                            Site             6    1 2 3 4 5 6


                         Criteria For Assessing Goodness Of Fit

                    Criterion                 DF           Value        Value/DF

                    Deviance                 144      10254.6846         71.2131
                    Scaled Deviance          144      10254.6846         71.2131
                    Pearson Chi-Square       144      15496.3115        107.6133
                    Scaled Pearson X2        144      15496.3115        107.6133
                    Log Likelihood                    12475.6559


            Algorithm converged.


                               Analysis Of Parameter Estimates

                                        Standard     Wald 95% Confidence       Chi-
   Parameter          DF    Estimate       Error           Limits            Square    Pr > ChiSq

   Intercept           1      2.5369      0.0563      2.4266      2.6471    2033.68        <.0001
   Site          1     1      0.9836      0.0659      0.8544      1.1128     222.51        <.0001
   Site          2     1      1.6932      0.0612      1.5732      1.8131     765.18        <.0001
   Site          3     1      1.3879      0.0629      1.2646      1.5111     487.10        <.0001
   Site          4     1     -0.3133      0.0866     -0.4830     -0.1437      13.10        0.0003
   Site          5     1     -0.2343      0.0846     -0.4002     -0.0684       7.66        0.0056
   Site          6     0      0.0000      0.0000      0.0000      0.0000        .            .
   Scale               0      1.0000      0.0000      1.0000      1.0000

 NOTE: The scale parameter was held fixed.
```

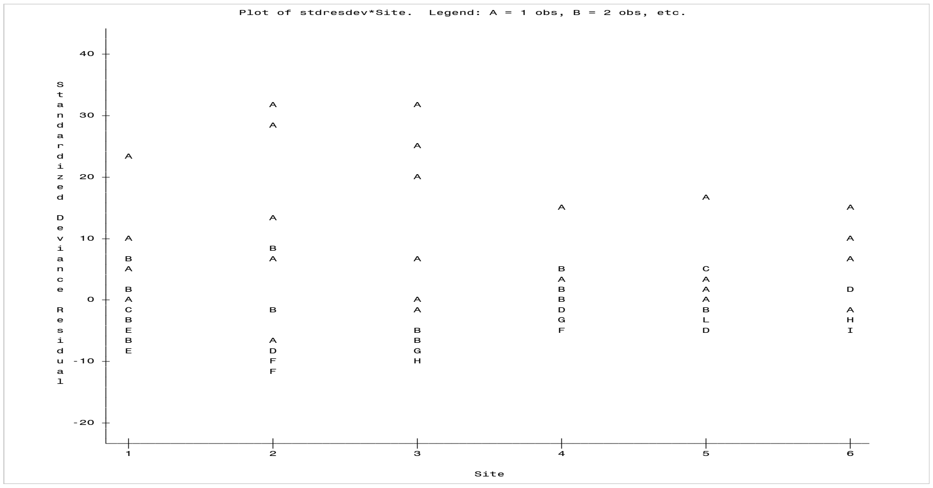
good fit if this is ≈ 1.

- rerunning model w/ dispersion parameter

```
  Overdispersed Poisson Regression on Hermit Crab Data

                      The GenMod Procedure

Distribution                Poisson
Link Function                   Log
Dependent Variable                Y    Crab Count
Number of Observations Read            150
Number of Observations Used            150

Class       Levels     Values
Site           6     1 2 3 4 5 6

             Criteria For Assessing Goodness Of Fit

Criterion                     DF            Value        Value/DF
Deviance                     144       10254.6846         71.2131
Scaled Deviance              144          95.2920          0.6618     ≈ 1
Pearson Chi-Square           144       15496.3115        107.6133
Scaled Pearson X2            144         144.0000          1.0000
Log Likelihood                           115.9305
Full Log Likelihood                      -49.9595
AIC (smaller is better)                  111.9189
AICC (smaller is better)                 112.5063
BIC (smaller is better)                  129.9827


                      Analysis Of Maximum Likelihood Parameter Estimates

                                    Standard     Wald 95% Confidence          Wald
Parameter          DF    Estimate       Error          Limits            Chi-Square    Pr > ChiSq
Intercept           1      2.5369      0.5836      1.3931      3.6806         18.90        <.0001
Site          1     1      0.9836      0.6840     -0.3571      2.3243          2.07        0.1505
Site          2     1      1.6932      0.6350      0.4487      2.9377          7.11        0.0077
Site          3     1      1.3879      0.6523      0.1093      2.6664          4.53        0.0334
Site          4     1     -0.3133      0.8980     -2.0734      1.4467          0.12        0.7272
Site          5     1     -0.2343      0.8781     -1.9553      1.4867          0.07        0.7896
Site          6     0      0.0000      0.0000      0.0000      0.0000           .             .
Scale               0     10.3737      0.0000     10.3737     10.3737

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

                               Contrast Results

                                                           Chi-
Contrast     Num DF     Den DF     F Value     Pr > F     Square     Pr > ChiSq     Type
S2 vs S1          1        144        2.82     0.0953       2.82         0.0931     LR
S2 vs S3          1        144        0.64     0.4255       0.64         0.4242     LR
S2 vs S4          1        144       11.92     0.0007      11.92         0.0006     LR
S2 vs S5          1        144       11.43     0.0009      11.43         0.0007     LR
S2 vs S6          1        144        9.88     0.0020       9.88         0.0017     LR
```

SAS Output Page 2 of 5

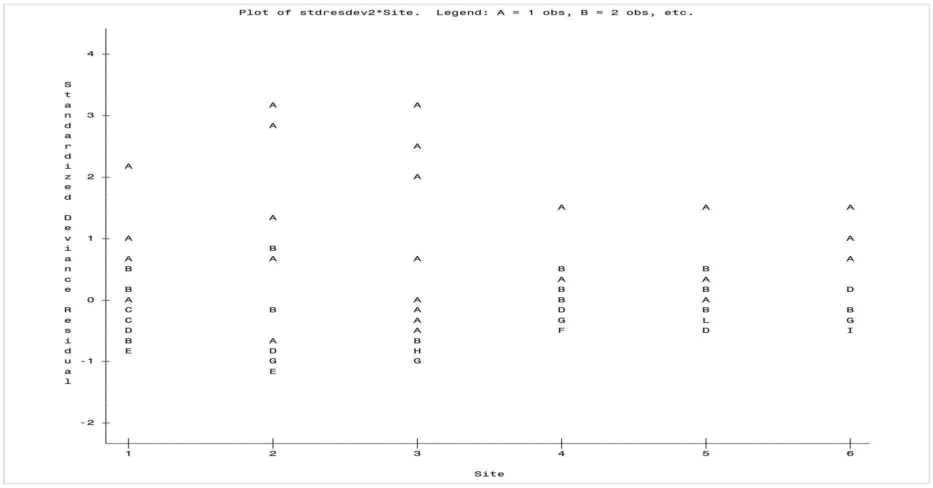
Poisson Regression on Hermit Crab Data



file:///C:/Users/longneck/AppData/Local/Temp/SAS%20Temporary%20Files/_TD8500_B... 2/18/2019

SAS Output Page 5 of 5

Overdispersed Poisson Regression on Hermit Crab Data



file:///C:/Users/longneck/AppData/Local/Temp/SAS%20Temporary%20Files/_TD8500_B... 2/18/2019

Tests for Correlation in Residuals

When we have **positive correlation** in the data, the inferences in the AOV F-test and multiple comparisons procedures can have a dramatic increase in the size of the test, that is, the type I error rate may be much larger than the nominal value. This also produces an increase in the power but in a practical sense, the increase in size of the test negates the positive aspect of an increase in power. Confidence intervals for treatment means and model parameters will have a coverage probability which is less than the stated value. One approach for detecting a first-order autocorrelation in the data is the Durbin-Watson test.

This test requires that the residuals have a **normal** distribution. A first-order autocorrelation in the residuals, e_t, where t represents a time sequencing in the data or a spatial ordering, can be represented by

$$e_t = \rho e_{t-1} + w_t \;\Rightarrow\; Corr(e_{t-1}, e_t) = E[e_t(\rho e_{t-1} + w_t)]/\sigma^2 = \rho\sigma^2/\sigma^2 = \rho$$

where e_t and w_t are independent and the w_t are iid $N(0, \sigma^2)$.

Similarly, we have $Corr(e_{t-k}, e_t) = \rho^k$

A test of $\rho = 0$ is equivalent to testing that the e_t's are independent.

An estimate of ρ is given by $\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$.

The Durbin-Watson test statistic is given by

$$\text{DW} = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^n e_t^2 + \sum_{t=1}^{n-1} e_t^2 - 2\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \approx \frac{2\sum_{t=1}^n e_t^2 - 2\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$$

which yields DW $\approx 2(1 - \hat{\rho})$

One-sided Test of $H_o : \rho = 0$ vs $H_1 : \rho > 0$ or

Decision rule: If $DW < d_L$, reject $H_o : \rho = 0$ at level α

If $DW > d_U$, do not reject $H_o : \rho = 0$ at level α

If $d_L \le DW \le d_U$, the test is said to be inconclusive

The values of d_L and d_U are given in the table on the following pages where k is the number of treatments.

Significance Points of d_L and d_U:1%

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65
150	1.61	1.64	1.60	1.65	1.58	1.67	1.57	1.68	1.56	1.69
200	1.66	1.68	1.65	1.69	1.64	1.70	1.63	1.72	1.62	1.72

Significance Points of d_L and d_U:1% (*Continued*)

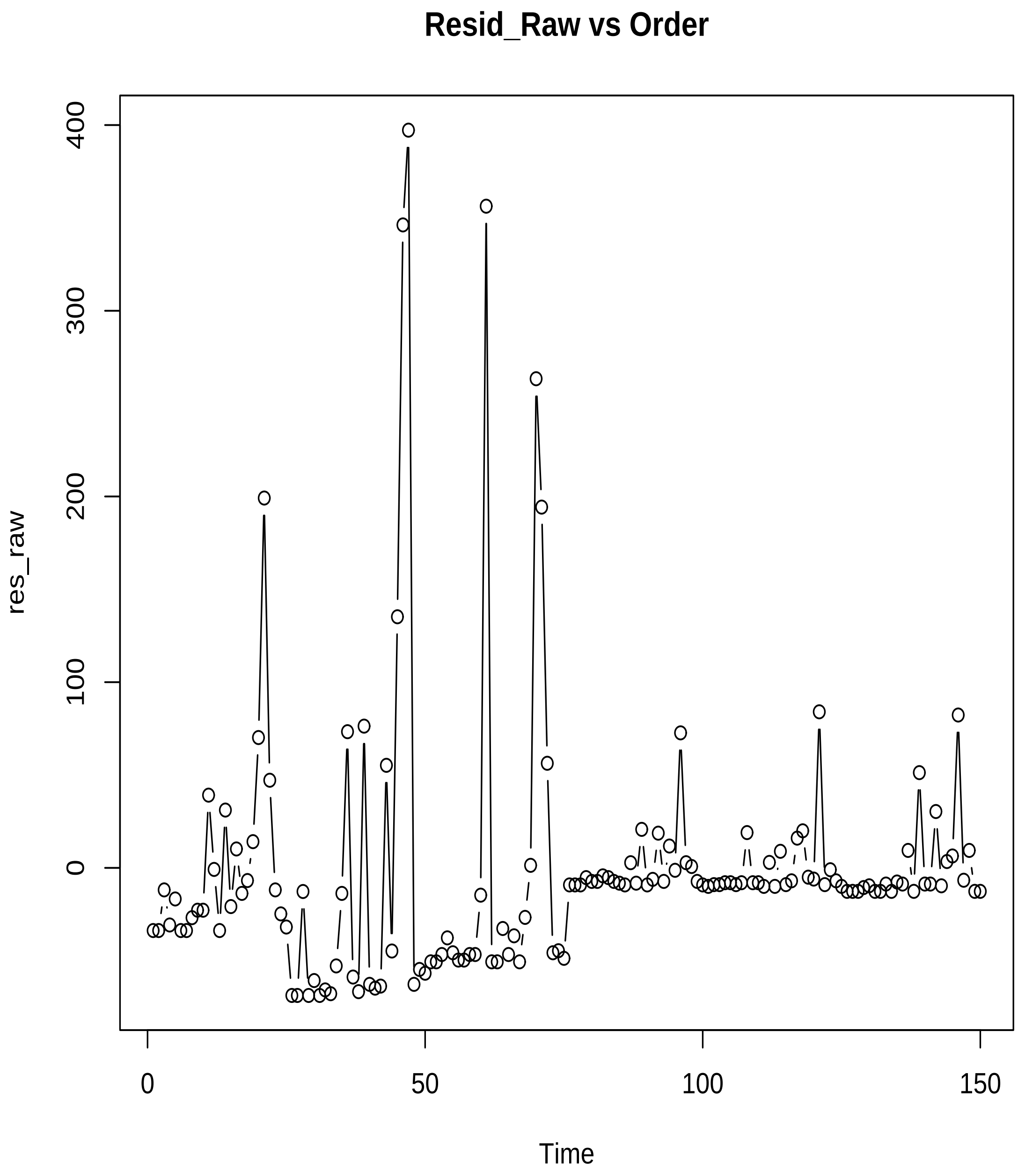
	$k = 6$		$k = 7$		$k = 8$		$k = 9$		$k = 10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
20	0.52	1.92	0.44		0.36		0.29		0.23	
21	0.55	1.88	0.47		0.40		0.33		0.27	
22	0.59	1.85	0.51		0.44		0.37		0.30	
23	0.62	1.82	0.55	1.98	0.47		0.40		0.34	
24	0.65	1.80	0.58	1.94	0.51		0.44		0.38	
25	0.68	1.78	0.61	1.92	0.54		0.47		0.41	
26	0.71	1.76	0.64	1.89	0.57		0.51		0.44	
27	0.74	1.74	0.67	1.88	0.60		0.54		0.47	
28	0.76	1.73	0.70	1.85	0.63	1.97	0.57		0.50	
29	0.79	1.72	0.72	1.83	0.66	1.95	0.60		0.53	
30	0.81	1.71	0.75	1.81	0.68	1.93	0.62		0.56	
31	0.83	1.70	0.77	1.80	0.71	1.91	0.65		0.59	
32	0.86	1.69	0.79	1.79	0.73	1.89	0.67		0.61	
33	0.88	1.68	0.82	1.78	0.76	1.87	0.70	1.98	0.64	
34	0.90	1.68	0.84	1.77	0.78	1.86	0.72	1.96	0.67	
35	0.91	1.67	0.86	1.76	0.80	1.85	0.74	1.94	0.69	
36	0.93	1.67	0.88	1.75	0.82	1.84	0.77	1.93	0.71	
37	0.95	1.66	0.90	1.74	0.84	1.83	0.79	1.91	0.73	
38	0.97	1.66	0.91	1.74	0.86	1.82	0.81	1.90	0.75	1.99
39	0.98	1.66	0.93	1.73	0.88	1.81	0.83	1.89	0.77	1.97
40	1.00	1.65	0.95	1.72	0.90	1.80	0.84	1.88	0.79	1.96
45	1.07	1.64	1.02	1.70	0.97	1.77	0.97	1.83	0.88	1.90
50	1.12	1.64	1.08	1.69	1.04	1.75	1.00	1.81	0.96	1.86
55	1.17	1.64	1.13	1.69	1.10	1.73	1.06	1.79	1.02	1.84
60	1.21	1.64	1.18	1.68	1.14	1.73	1.11	1.77	1.07	1.82
65	1.25	1.64	1.22	1.68	1.19	1.72	1.15	1.76	1.12	1.80
70	1.28	1.65	1.25	1.68	1.22	1.72	1.19	1.75	1.16	1.79
75	1.31	1.65	1.28	1.68	1.26	1.72	1.23	1.75	1.20	1.79
80	1.34	1.65	1.31	1.68	1.29	1.71	1.26	1.75	1.23	1.78
85	1.36	1.66	1.34	1.69	1.31	1.71	1.29	1.74	1.26	1.77
90	1.38	1.66	1.36	1.69	1.34	1.71	1.31	1.74	1.29	1.77
95	1.40	1.67	1.38	1.69	1.36	1.72	1.34	1.74	1.31	1.77
100	1.42	1.67	1.40	1.69	1.38	1.72	1.36	1.74	1.34	1.77
150	1.54	1.71	1.53	1.72	1.52	1.74	1.50	1.75	1.49	1.77
200	1.61	1.74	1.60	1.75	1.59	1.76	1.58	1.77	1.57	1.80

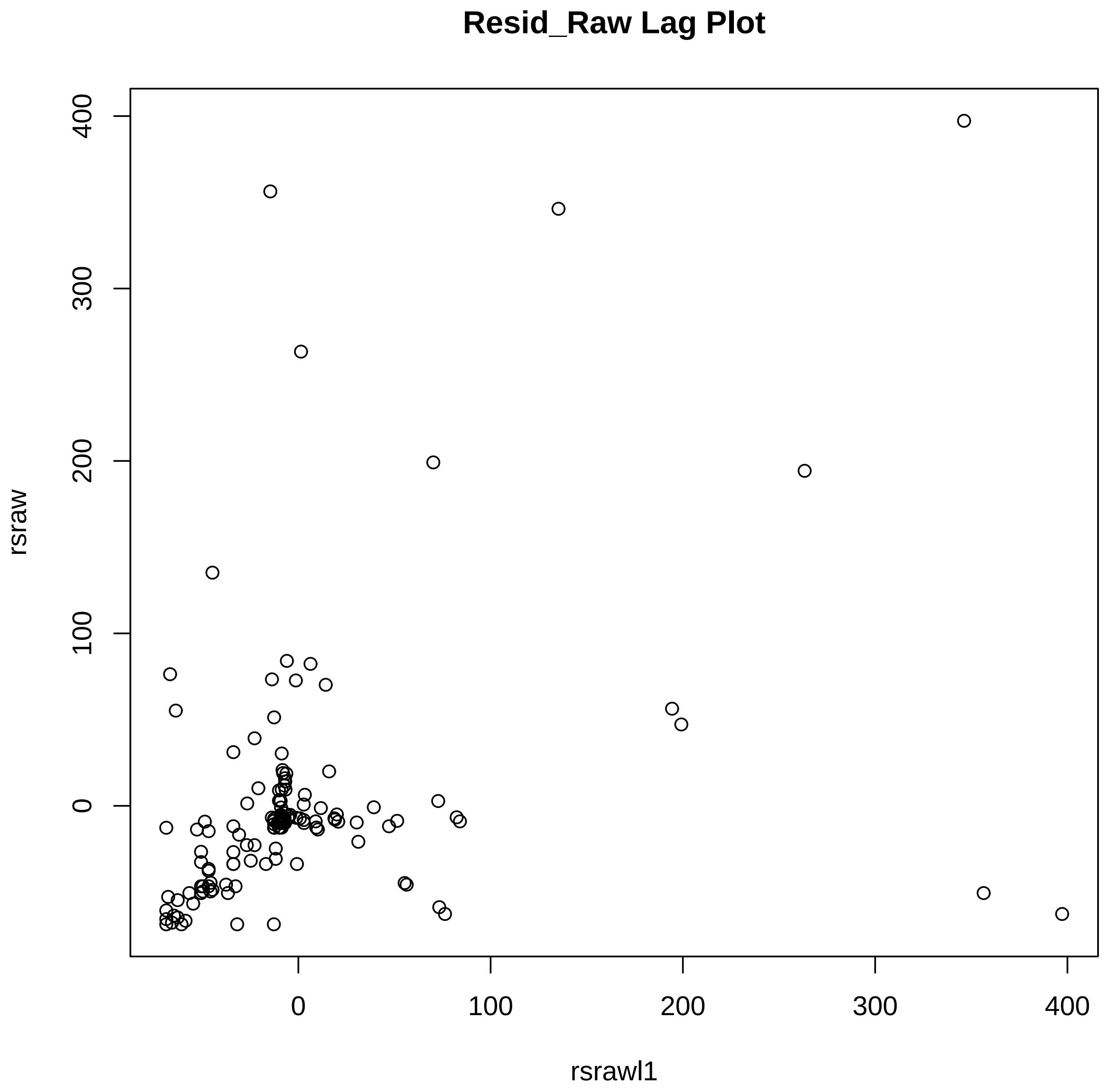
The R code on the following page yields diagnostic plots for correlation in the residuals.

```
 R CODE FOR DISPLAYING CORRELATION PLOTS FOR HERMIT CRAB DATA EXAMPLE

Crab_corrplots.R in CANVAS

  library(ts)
 data = matrix(0,150,2)
 y  = matrix(0,150,1)
 data  = scan("u:\meth2/s,files/expl4-1.dat",list(a = 0,b = ""))
 y  = data$a
 site  = data$b
 d = data.frame(y,site)
 anal1  = aov(y ~ site,data = d)
 rs1  = resid(anal1,type = "response")
 rstime1  = ts(rs1,start = 1,frequency = 1)
 abline(h = 0,lty = 2)
 rsraw  = rs1[2:150]
 rsrawl1  = rs1[1:149]
 plot(rstime1,type = "b",ylab = "res_raw",main = "Resid_Raw vs Order")
 plot(rsrawl1,rsraw,main = "Resid_Raw Lag Plot")

#Calculation of Durbin-Watson Statistics

dif1  = (rsraw-rsrawl1)^2
num1  = sum(dif1)
rs12  = rs1^2
den1  = sum(rs12)
DW1  = num1/den1
prd1  = rsraw*rsrawl1
prdsum1  = sum(prd1)
rho1  = prdsum1/den1


-------------------------------------------
Output from Corrplot.s:
> rho1
[1] 0.3775777
> DW1
[1] 1.243095
```

$d_L = 1.54$

$d_U = 1.71$


Resid_Raw vs Order
res_raw
0
100
200
300
400
0
50
100
150
Time


Resid_Raw Lag Plot
rsraw
rsrawl1
0
100
200
300
400
0
100
200
300
400

Runs Test for Correlation

When the data is nonnormal, the Durbin-Watson test is invalid. An alternative distribution-free procedure, the Runs Test, will be presented.

Let $X_1, X_2, \ldots, X_T$ be T equally spaced observations on a random process.

To test if the t observations are correlated:

1. Center the observations: $Y_t = X_t - \bar{X}$, where $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$
2. Count the number of runs (R), where a run is defined as a sequence of observations of all positive values or all negative values
3. Count the number of positive Y_ts (n_1) and the number of negative Y_ts (n_2)
4. When $n_1 \leq 20$ and $n_2 \leq 20$, we can use the following decision rule where R_L and R_U are values given in the table from *Annals of Mathematical Statistics,* **14**, pp. 66-87.:

   a. the data indicates that X_t is positively correlated if $R \leq R_L$

   b. the data indicates that X_t is negatively correlated if $R \geq R_U$,

   c. the data is indeterminate if $R_L \leq R \leq R_U$,

5. Large sample size critical values are obtained by declaring that the data indicates that X_t is correlated if $Z > Z_{\alpha/2}$, where

$$Z = \frac{|R - \mu| - 0.5}{\sigma}, \quad \mu = 1 + \frac{2n_1 n_2}{n_1 + n_2}, \quad \sigma^2 = \frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

   where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the $N(0,1)$ distribution.

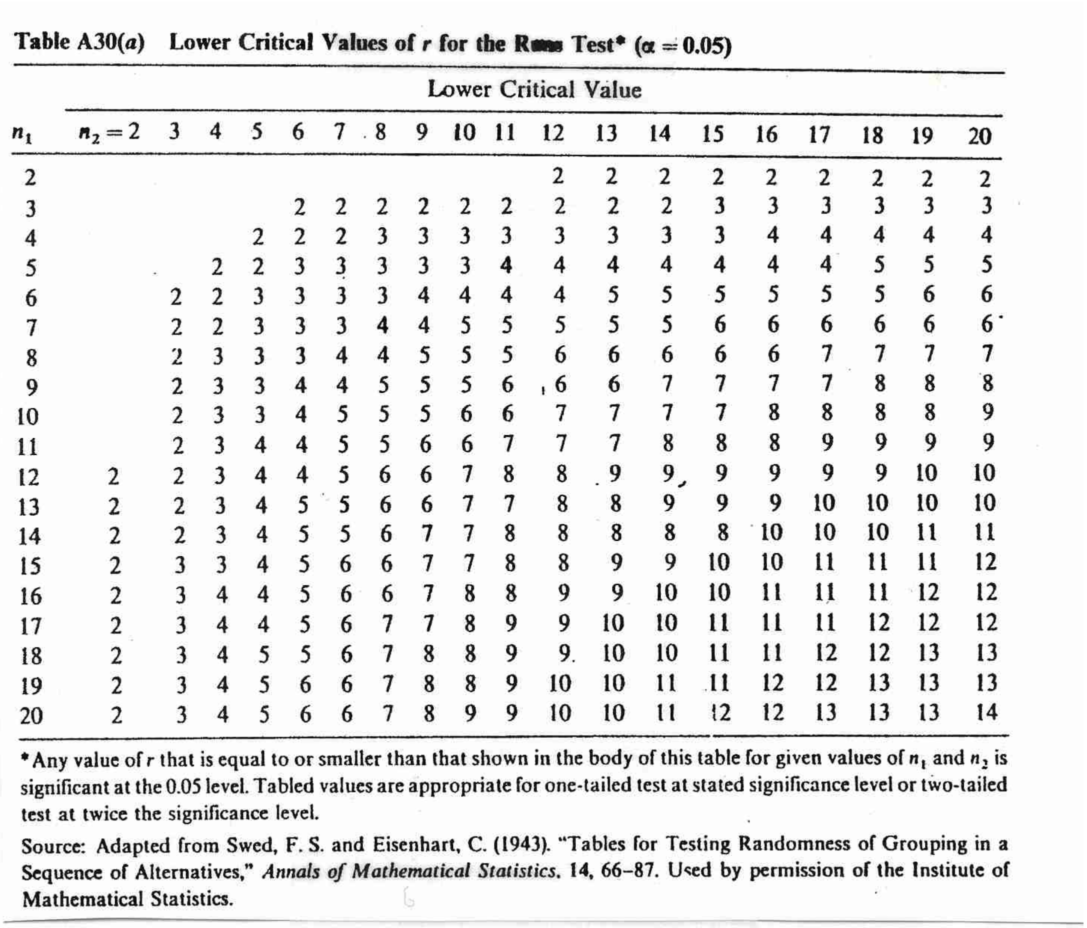


Table A30(a) **Lower Critical Values of r for the Runs Test* ($\alpha = 0.05$)**

	Lower Critical Value																		
n_1	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	8	8	9	9	9	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	8	8	8	10	10	10	11	11
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	12
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

*Any value of r that is equal to or smaller than that shown in the body of this table for given values of n_1 and n_2 is significant at the 0.05 level. Tabled values are appropriate for one-tailed test at stated significance level or two-tailed test at twice the significance level.

Source: Adapted from Swed, F. S. and Eisenhart, C. (1943). "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," *Annals of Mathematical Statistics*, 14, 66–87. Used by permission of the Institute of Mathematical Statistics.

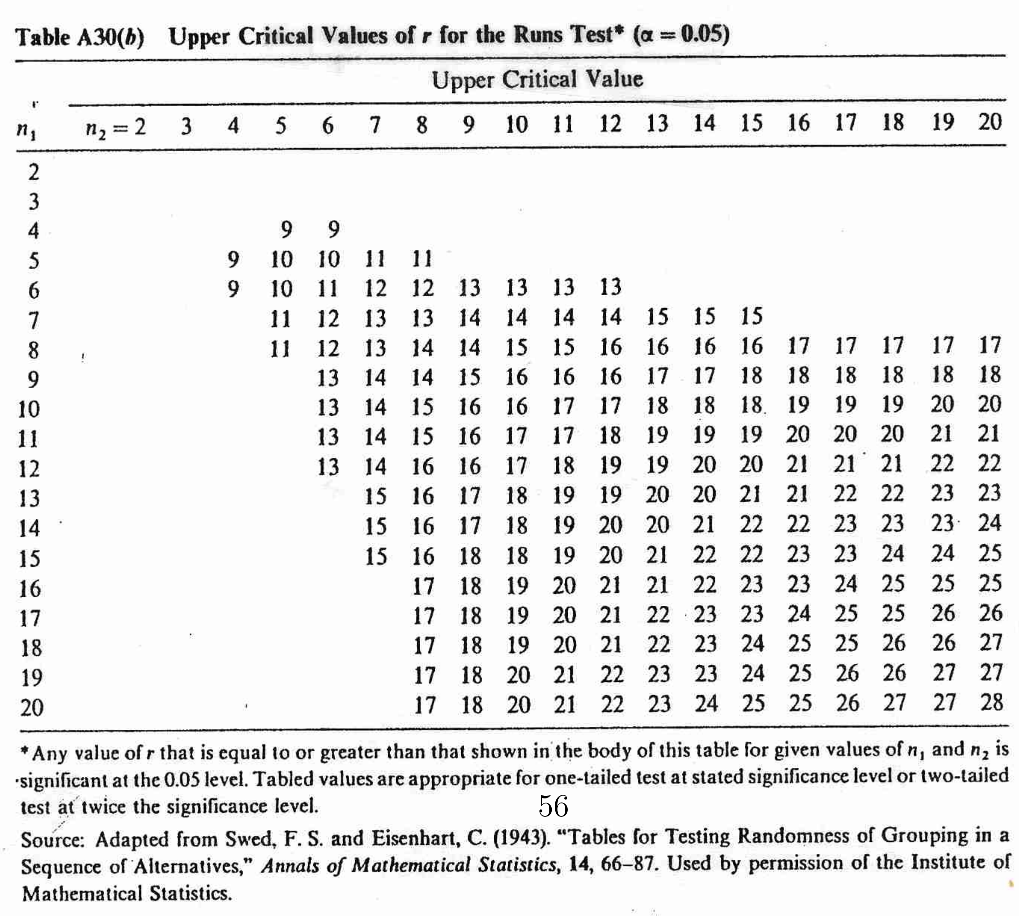


Table A30(b) **Upper Critical Values of r for the Runs Test* ($\alpha = 0.05$)**

	Upper Critical Value																		
n_1	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4				9	9														
5			9	10	10	11	11												
6			9	10	11	12	12	13	13	13	13								
7				11	12	13	13	14	14	14	14	15	15	15					
8				11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9					13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10					13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11					13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12					13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13						15	16	17	18	19	19	20	20	21	21	22	22	23	23
14						15	16	17	18	19	20	20	21	22	22	23	23	23	24
15						15	16	18	18	19	20	21	22	22	23	23	24	24	25
16							17	18	19	20	21	21	22	23	23	24	25	25	25
17							17	18	19	20	21	22	23	23	24	25	25	26	26
18							17	18	19	20	21	22	23	24	25	25	26	26	27
19							17	18	20	21	22	23	23	24	25	26	26	27	27
20							17	18	20	21	22	23	24	25	25	26	27	27	28

*Any value of r that is equal to or greater than that shown in the body of this table for given values of n_1 and n_2 is significant at the 0.05 level. Tabled values are appropriate for one-tailed test at stated significance level or two-tailed test at twice the significance level.

Source: Adapted from Swed, F. S. and Eisenhart, C. (1943). "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," *Annals of Mathematical Statistics*, 14, 66–87. Used by permission of the Institute of Mathematical Statistics.

STOP Friday 2/25/22 (Week 6, Lecture 17)

START Monday 2/28/22 (week 7, Lecture 14)

$n_1=12$ (+)

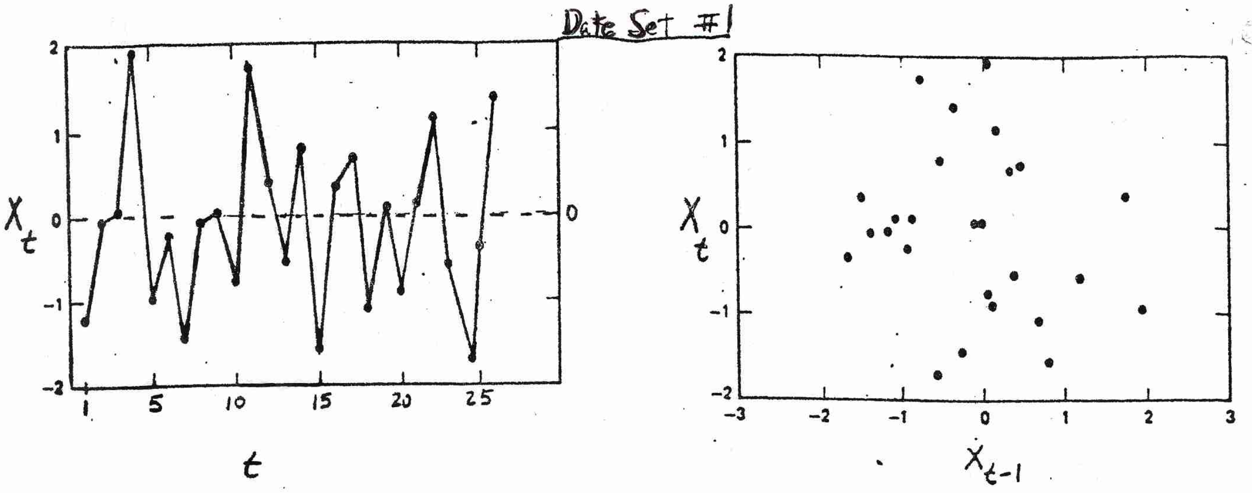
$n_2=14$ (-)

$r=16$

$r_L=9$

$r_U=20$

Data Set #1



$n_1=12$

$n_2=14$

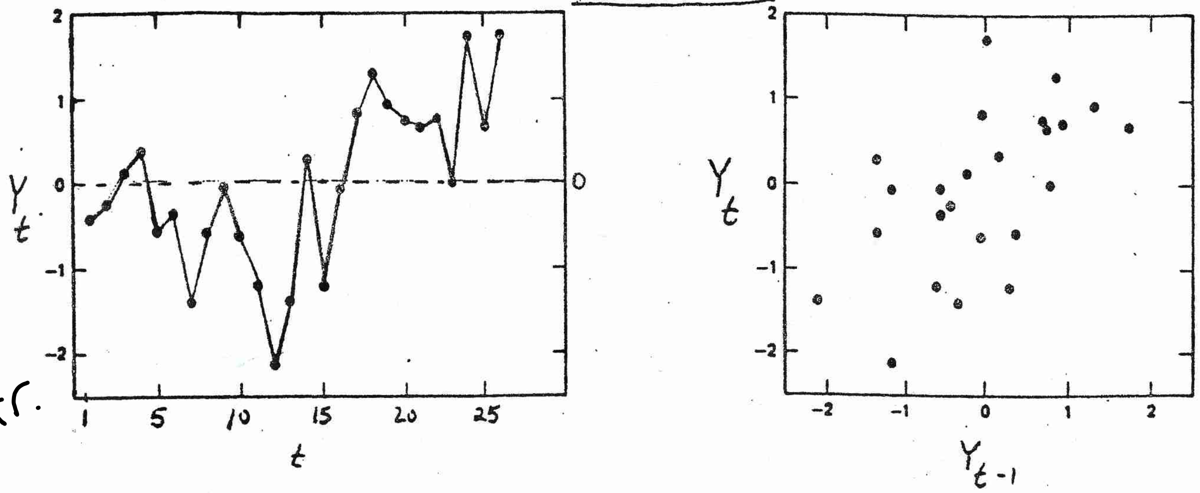
$r=8$

$r_L=9$

$r_U=20$

⇒ positive corr.

Data Set # 2



$n_1=12$

$n_2=14$

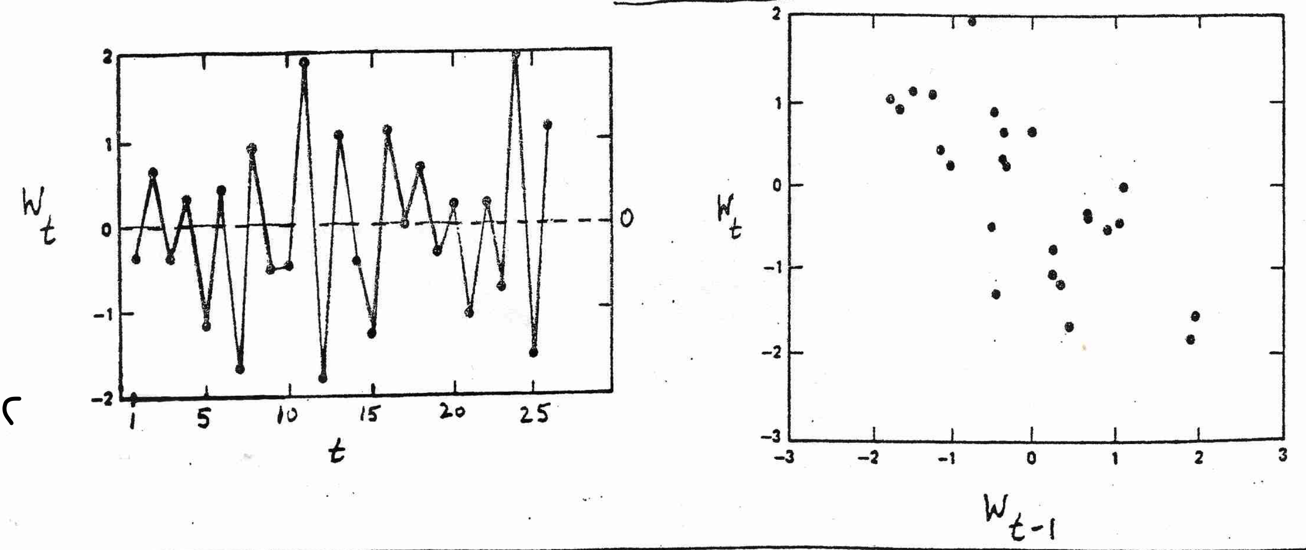
$r_L=9$

$r_U=20$

$r=24$

⇒ neg. corr

Data Set #3



Three Time Series from Applied Regression Analysis, 3rd Ed, Draper - Smith

EXAMPLE

The residuals from the 150 Crab Count data values are given next to illustrate the Runs Test. We will assume that there is a spatial ordering in the data as given below by site. We want to determine if the 25 residuals for each site have a correlation in them. We will evaluate each site individually because it may not be too realistic to have correlation between sites but correlation within sites is a possibility.

```
Site 1:

    1     2     3     4     5     6     7     8     9    10   11   12
 -33.8 -33.8 -11.8 -30.8 -16.8 -33.8 -33.8 -26.8 -22.8 -22.8 39.2 -0.8

   13   14    15   16    17   18   19   20    21   22    23    24    25
 -33.8 31.2 -20.8 10.2 -13.8 -6.8 14.2 70.2 199.2 47.2 -11.8 -24.8 -31.8

Site 2:

   26     27     28     29     30     31     32     33     34     35    36     37
 -68.72 -68.72 -12.72 -68.72 -60.72 -68.72 -65.72 -67.72 -52.72 -13.72 73.28 -58.72

     38    39     40     41     42    43     44     45     46     47     48     49    50
 -66.72 76.28 -62.72 -64.72 -63.72 55.28 -44.72 135.28 346.28 397.28 -62.72 -54.72 -56.72

Site 3:

     51     52     53     54     55     56     57     58     59    60     61     62
  -50.64 -50.64 -46.64 -37.64 -45.64 -49.64 -49.64 -46.64 -46.64 -14.64 356.36 -50.64

     63     64     65     66     67     68   69   70     71      72    73     74     75
 -50.64 -32.64 -46.64 -36.64 -50.64 -26.64 1.36 263.36 194.36 56.36 -45.64 -44.64 -48.64

Site 4:

  76    77    78    79    80    81    82     83    84    85    86   87
 9.24 -9.24 -9.24 -5.24 -7.24 -7.24 -4.24  -5.24 -7.24 -8.24 -9.24 2.76

   88    89    90    91    92    93    94    95   96    97   98    99   100
 -8.24 20.76 -9.24 -6.24 18.76 -7.24 11.76 -1.24 72.76 2.76 0.76 -7.24 -9.24

Site 5:

 101   102   103   104   105   106   107   108   109   110   111   112
 -10    -9    -9    -8    -8    -9    -8    19    -8    -8   -10     3

 113   114   115   116   117   118   119   120   121   122   123   124   125
 -10     9    -9    -7    16    20    -5    -6    84    -9    -1    -7   -10

Site 6:

    126    127    128   129   130    131    132   133    134   135   136  137
  -12.64 -12.64 -12.64-10.64 -9.64 -12.64 -12.64 -8.64 -12.64 -7.64 -8.64 9.36

   138   139   140   141   142   143  144  145   146   147  148    149    150
 -12.64 51.36 -8.64 -8.64 30.36 -9.64 3.36 6.36 82.36 -6.64 9.36 -12.64 -12.64
```

From the above data, we can use the following R program to determine the number of Runs for each site and the values of n_1 and n_2:

The following R program is labeled as **runstestCrabdata.R** in the CANVAS R Files folder:

```
site1=c(0,0,22,3,17,0,0,7,11,11,73,33,0,65,13,44,20,27,48,104,233,81,22,9,2)
site2=c(0,0,56,0,8,0,3,1,16,55,142,10,2,145,6,4,5,124,24,204,415,466,6,14,12)
site3=c(0,0,4,13,5,1,1,4,4,36,407,0,0,18,4,14,0,24,52,314,245,107,5,6,2)
site4=c(0,0,0,4,2,2,5,4,2,1,0,12,1,30,0,3,28,2,21,8,82,12,10,2,0)
site5=c(0,1,1,2,2,1,2,29,2,2,0,13,0,19,1,3,26,30,5,4,94,1,9,3,0)
site6=c(0,0,0,2,3,0,0,4,0,5,4,22,0,64,4,4,43,3,16,19,95,6,22,0,0)
site = c(site1,site2,site3,site4,site5,site6)
data6 = matrix(site,nrow=6,byrow=T)
resid = matrix(0,6,25)
for (i in 1:6) {
  means6[i] = mean(data6[i,])
   resid[i,]   = data6[i,]-means6[i]
   resid1[i,]  = resid[i,2:25]
   residl1[i,] = resid[i,1:24]
  for (j in 1:24){
    dif1[i,j] = (resid1[i,j]-residl1[i,j])^2
    prd1[i,j] = resid1[i,j]*residl1[i,j]
  }
   rho[i] = sum(prd1[i,])/sum((resid[i,])^2)
   DW[i] = sum(dif1[i,])/sum((resid[i,])^2)
  }
n.neg =rep(0,6)
n.pos =rep(0,6)
for (i in 1:6) {
  n.neg[i] =length(resid[i,][resid[i,]<0])
  n.pos[i] =length(resid[i,][resid[i,]>0])
  }
numb.runs =rep(1,6)
for (i in 1:6) {
  for (j in 2:25) {
    if (sign(resid[i,j]) != sign(resid[i,j-1])) {numb.runs[i] =numb.runs[i] + 1}
    }
  }
residruns.result =as.data.frame(cbind(numb.runs, n.pos, n.neg))
names(residruns.result) =c("No. runs", "N+", "N-")
```

From the output of the R program we obtain the following values along with the critical values from Tables 30(a) and 30(b) on Page 54.

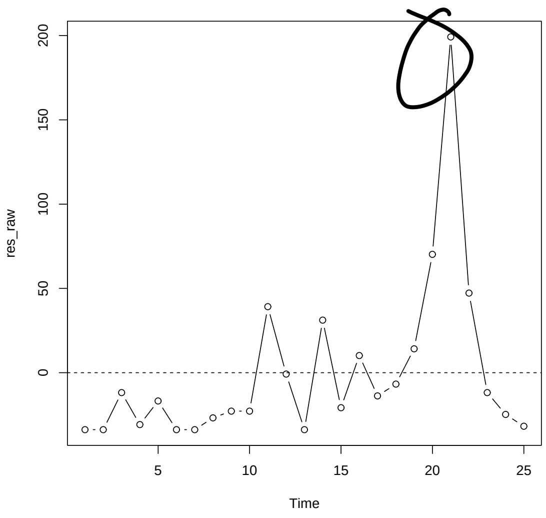
Site	n_1	n_2	R	R_L	R_U	Decision	$\hat{\rho}$
1	7	18	9	6	15	Indeterminant	.462
2	6	19	9	6	13	Indeterminant	.475
3	5	20	5	5	11	Positive Correlation	.252
4	8	17	11	7	17	Indeterminant	.0006
5	6	19	11	6	13	Indeterminant	-.097
6	7	18	11	6	15	Indeterminant	-.048

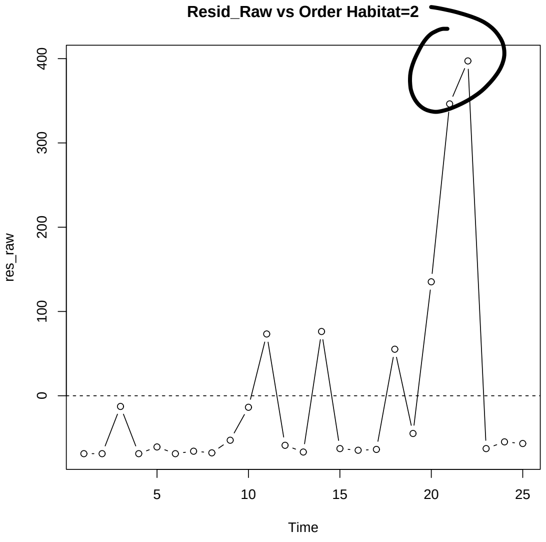
From the above Runs' tests, only Site 3 appears to have positive correlation.

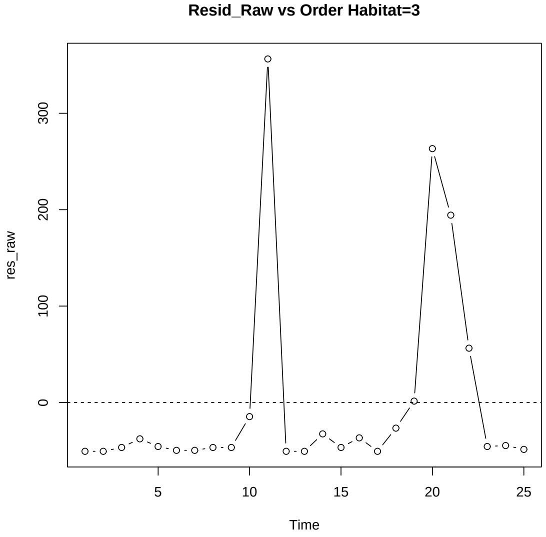
Why do the estimated coefficients for ρ appear to contradict our conclusions about the presence/absence of correlation in the crab counts at the six sites?

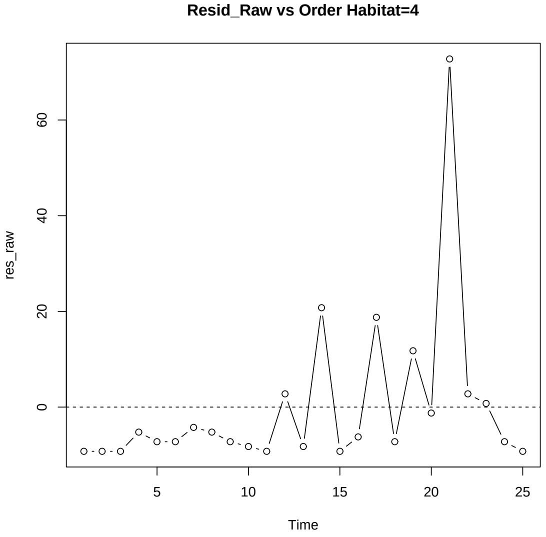
Plots of the residuals versus the spatial ordering of the counts by Habitat are given on the next page.

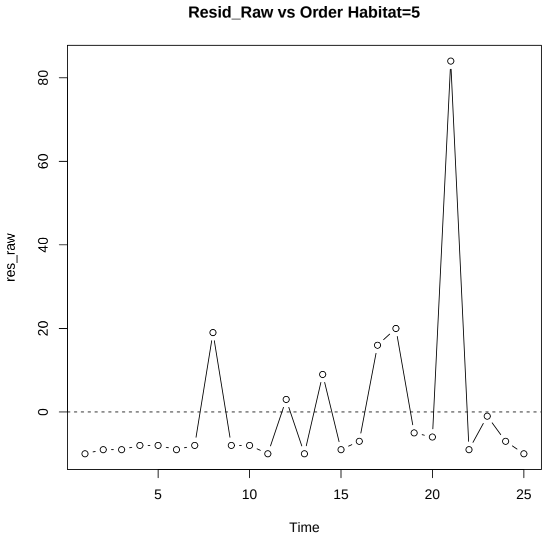
corelation is affected by outliers.

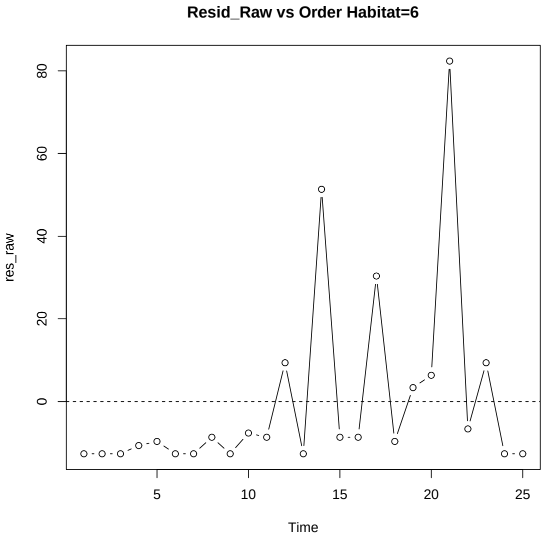

Resid_Raw vs Order Habitat=1


Resid_Raw vs Order Habitat=2


Resid_Raw vs Order Habitat=3


Resid_Raw vs Order Habitat=4


Resid_Raw vs Order Habitat=5


Resid_Raw vs Order Habitat=6

Impact of Correlated Data on Inference Procedures

What happens to the F-test or Confidence Intervals if the data is not independent but is correlated:

1. Equicorrelated: $Cov(y_{ij}, y_{ih}) = \rho\sigma^2$ for all $j \neq h$ $\frac{-1}{n_i - 1} < \rho < 1$

   $E[\bar{y}_{i.}] = \mu_i$ and $Var(\bar{y}_{i.}) = \frac{\sigma^2}{n_i}[1 + (n_i - 1)\rho]$

2. 1st Order Autoregressive: $Cov(y_{ij}, y_{ih}) = \rho^{|j-h|}\sigma^2$ for $h \neq j$ $1 < \rho < 1$

   $E[\bar{y}_{i.}] = \mu_i$ and

$$\begin{aligned} Var(\bar{y}_{i.}) &= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(Y_i) + \sum\sum_{i \neq j} Cov(Y_i, Y_j)\right] \\ &= \frac{n\sigma^2}{n^2} + \frac{2\sigma^2\rho}{n^2(1-\rho}\left[n + \frac{1-\rho^n}{1-\rho}\right] \\ &\approx \frac{\sigma^2}{n}\left[\frac{1+\rho}{1-\rho}\right] \end{aligned}$$

In both cases, if $\rho > 0$, then $Var(\bar{y}_{i.}) > \frac{\sigma^2}{n}$.
Thus, $\frac{\hat{\sigma}^2}{\sqrt{n}}$ underestimates $Var(\bar{y}_{i.})$.
This results in a C.I. for μ_i: $\bar{y}_{i.} \pm t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n_i}}$

which is too narrow and hence the coverage probability is less than $100(1-\alpha)\%$.

Also, the F-test statistic $\frac{MS_{TRT}}{MSE}$ is too large in comparison to the ratio with the correct value for the estimated variance was used in the denominator. Thus, the probability of Type I error is inflated above α, resulting in an inflated proportion of Type I errors. However, the power of the test is also inflated but this gain in power is paid for by an inflated Type I error rate.

We need to detect when correlation is present and adjust C.I.'s and tests of hypotheses for the correlation. If possible we need to determine the type of correlation present and then estimate the standard error of $\hat{\mu}_i$ to take into account the correlation in the data. The critical value of the test statistic would need to be adjusted also.
Time series STAT 626 and STAT 673 deal with Temporally Correlated Data.
STAT 647 deals with Spatially Correlated Data.

Linear models- STAT 612, applied multivariate analysis- STAT 636, and theoretical multivariate analysis- STAT 616, deal with modelling situations in which the correlated in the residuals is generally specified as

$$\sigma_e^2\mathbf{V} \neq \sigma_e^2\mathbf{I}$$

We will examine choices for $\mathbf{V}$ when discussing experiments involving repeated measures.

Finished Monday 2/28/22 (Week 7, lecture 18) @ 25min mark.