Jack Roden
STAT 642
Assignment 05
3/10/22

Read H.O.5 | Design's Anova book Chp 5:

1.)

(H.O.5 pg. 2)
(a) Model Conditions: ① t treatment populations $\sim N$

② t treatment populations have equal variance σ_e^2

③ The observed data values are independent

(H.O.5 pg 6) → [• NOTE: Our $n_i = 6$ are relatively small, thus we shouldn't do S.W. tests on the data for each of the treatment means. ⇒ examine the sample residuals from the fitted model $y_{ij} = \mu_i + e_{ij} \Rightarrow \hat{e}_{ij} = y_{ij} - \bar{y}_{i\cdot}$]

① Normality: • Using the Shapiro Wilks test of normality on the residuals we get $W = 0.92835$, p-value $= 0.08956$. Thus, we dont have significant evidence at the $\alpha = 0.05$ level to conclude that the residuals arent normally distributed.

• looking at a qq plot for the residuals, they don't seem to fit the line very well, but this could be due in part to our relatively small sample size.

(H.O.5 pg 10-18) → ② Constant variance: • Using the BFL test of homogeneity of variance

$H_0: \sigma_{40°C} = \sigma_{45°C} = \sigma_{55°C} = \sigma_{70°C}$

H_1: Not all σ_i are equal

• running this test, we get

$L = 23.43 \geq F_{0.05, 4-1, 24-4} = 3.098391$

$P[F_{3,20} \geq 23.43] = 9.4 \times 10^{-7} \approx 0.$

• Thus we would reject our hypothesis of equal variances and conclude that we have significant evidence that the treatment populations have different variances.

• This can also be seen in the boxplot of the responses for the different treatments.

(H.O.5 pg 50) → ③ Independence: NOTE: our residuals are normally distributed ⇒ use Durbin-Watson test

• conducting the DW test we get

$\boxed{DW = 0.8471899}$, $dL_{\alpha=0.05} = 1.01$ $du_{\alpha=0.05} = 1.78$ ← ask about this, (conflicting w/ r-output.)

; $dL_{\alpha=0.01} = 0.80$, $du_{\alpha=0.01} = 1.53$

• Thus we would reject the hypothesis that the residuals are independent at the $\alpha = 0.05$ level, but we would fail to reject the hypothesis that the residuals are independent at the $\alpha = 0.01$ level.

Ask about approach taken in R. (is my s_i correct)

1.) (contd)

(H.O.5 pg 20) → (b) Determine a reasonable transformation of the data using the slope of the regression line based on $\log(s_i)$ vs $\log(\bar{y}_i)$.

Answer to be graded.

- Fitting a the regression model: $\log(s_i) = \beta_0 + \beta_1 \log(\bar{y}_{i\cdot})$ ⇒ use $x_{ij} = y_{ij}^{1-1.7295} = y_{ij}^{-0.7295}$

We get an estimate $\hat{\beta}_1 = 1.7295$.

$SE(\hat{\beta})$ given in r. output

Talked to Prof, said not to do this. Do NOT Grade

- A 95% CI on $\hat{\beta}_1$ is given by: $(1.7295 \pm qt(0.975, 2,) * 0.2768)$
$= (0.538, 2.920)$

Thus, we don't have significant evidence that β_1 is different from 1. Thus, the appropriate transformation is given by $x_{ij} = \log(y_{ij})$

Ask if transforms are correct or should I round?

ask about example in notes. 95% CI for θ doesn't include 0 but says log is appropriate

(H.O.5 pg 30) → (c) Use the Box-cox technique for selecting a transformation of the data. Is the transformation from Box-cox procedure consistent w/ your transformation from part (b)?

- No, in part (b) I found the appropriate trans form to be $\theta = -0.7295$ ~~log transformation to be appropriate~~. However using the box cox technique I get $\theta = -0.64$. But the CI for θ on the box-cox transformation does include -0.7295, so the two methods give very similar results.

?

- NOTE: If we wanted to round each to the nearest quarter (i.e. says that 0.25, 0.5, ...) then in (b) I would have $\hat{\beta}_1 = 1.75$ ⇒ $x_{ij} = y_{ij}^{-0.75}$; in part (c) I would get $\theta_{max} = -0.75$

(d) Using the transformation from part c, is the transformed data appropriate for conducting AOV?

Normality: SW p-value $= 0.5852$ ✓

Equal Var: BFL: $L = 2.261$, p-value $= 0.113$ ✓

Independence: $DW = 1.4939417$; $d_{L\, \alpha=0.05} = 0.80$, $d_{U\, \alpha=0.01} = 1.53$ ✓

- Yes, the transformed data is appropriate for conducting AOV.

1.) (contd)

(e) Perform an AOV on both the original data and the transformed data. Compare the results of the two analyses.

Original Data:

	DF	Sum Sq	Mean Sq	Fval	P(>F)
Temps	3	39183995	13061332	11.15	0.000162
Residuals	20	23427429	1171371		

Transformed Data:

	DF	Sum Sq	Mean Sq	Fvalue	P(>F)
Temps	3	0.0003676	0.00012253	31.95	0.0000000798
Residuals	20	0.00007767	0.000003840		

- For both data sets we would reject:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

but the conclusion is somewhat stronger for the transformed data than it is for the original data. However, the p-value from the original data is likely not valid because of the violation of the constant error variance assumption.

Should we use the transformed data →

(f) Use Tukey's HSD to group the 4 temps relative to mean time to failure.

- Using the original Data: $G_1 = \{40°C\}$, $G_2 = \{45°C, 55°C, 70°C\}$
- Using transformed Data: $G_1 = \{40°C, 45°C\}$ $G_2 = \{55°C\}$, $G_3 = \{70°C\}$
- In grouping the treatment means I would use the transformed data b/c the original data violated the assumption of constant variance.

(g) Test for a trend in the time to failures as a function of Temperature. B/c the temperatures are unequally spaced, the following contrast coefficients we obtained from R. The contrasts for the three contrasts, linear, quadratic & cubic are given below.

- Using $\alpha_{PC} = 1-(1-0.95)^{1/3} = 0.01695$
  - Transformed Data: There is significant evidence of a linear trend (p-value $= 5.5973 \times 10^{-9}$) The quadratic & cubic trends aren't significant.
  - Untransformed Data: There is significant evidence of a linear trend (p-value: 7.991×10^{-5}) and a quadratic trend (p-value $= 0.01424$) The cubic trend is not significant.
- Note we get different conclusions for the transformed & untransformed data. I would ignore the conclusions of the untransformed data as it violates the assumption of constant error variance.

(See H.O.5 pg 32) →
(H.O.5 pg 37 for code)

2.) For the time to failure data in Problem 1:

(a) Use a rank based test to compare the average time to failure for the four temperatures

- Kruskal-Wallis Test: KW chi-squared = 18.278 w/ df = 3
  P-value = 0.0003853.

We can thus conclude at the $\alpha = 0.01$ level that we have significant evidence that the treatment populations have different location parameters.

* [Q: Why doesn't nonequal variances matter (see H.O.5 pg 36)?]
  ↳ consistent under monotonic transformations

(H.O.5 pg 38) → (b) Use a rank based multiple comparisons procedure to group the 4 temperatures relative to average time to failure.

$\bar{R}_{1.} = 20.\overline{333}$ $\bar{R}_{2.} = 16.00$, $\bar{R}_{3.} = 9.\overline{333}$, $\bar{R}_{4.} = 4.\overline{333}$

- Using Hollander-Wolfe Procedure two pairs are said to be different if:

$$|\bar{R}_{i.} - \bar{R}_{n.}| \geq \sqrt{h_\alpha \left(\frac{2t(n_r+1)}{12}\right)}$$

* found table online $h_\alpha = 7.453$ ($\alpha = 0.05$, groups = 4, $n_i = r = 6$)

$$|\bar{R}_{i.} - \bar{R}_{n.}| \geq \sqrt{7.453\left(\frac{2(4)(24+1)}{12}\right)}$$

Q: Which procedure should we use here?
HW fits what we have better but results from miller rank are more sane.

- $G_1 = \{40°C, 45°C, 55°C\}$ $G_2 = \{55°C, 70°C\}$

use results from Miller Rank procedure.

- Using Miller Rank procedure:
  - $G_1 = \{40°C, 45°C\}$, $G_2 = \{45°C, 55°C\}$, $G_3 = \{55°C, 70°C\}$

(c) Compare your results to your analysis of the untransformed data

- For the untransformed data: $G_1 = \{40°C\}$, $G_2 = \{45°C, 55°C, 70°C\}$
- For the transformed data: $G_1 = \{40°C, 45°C\}$ $G_2 = \{55°C\}$, $G_3 = \{70°C\}$

- The rank based procedures give us very different results than the untransformed data does using the Tukey HSD procedure. However the rank based procedures (specifically the miller rank procedure) is very similar to the results we obtained from the transformed data using the Tukey HSD procedure.

1x6

3.) (HO.5 pg 41-49) An entomologist counted the number of eggs laid by female moths on successive days in three strains of tobacco budworm (USDA, Field, Resistant) from each of 15 matings. The entomologist is interested in evaluating whether the average number of eggs was different from the three strains. The number of eggs laid on the 3rd day after mating for each female is given in the following table (see HW for table).

(a) The entomologist suspects that the data is from poisson distributions. Based on the data do Poisson distributions appear to be reasonable distributions for the egg data.

- We know if a variable $X \sim Poisson(\lambda)$

$$E[X] = \lambda = var(X) \quad \text{i.e.} \quad \mu_x = \sigma_x^2$$

- For the moth data:

Strain	Mean	Variance
USDA	368.00	70,554.71
Field	181.27	44,517.28
Resistant	90.80	13,949.17

Thus the regular poisson distribution doesn't seem like a reasonable fit, but an overdispersed poisson may be a reasonable fit.

(b) Using PROC GENMOD in SAS (glm in R), perform an analysis using a model having a poisson distribution for the three egg count distributions. Make sure to check for variance inflation.

- From the output in R we get the scaled deviance / df = 200.32 which is not very close to 1 therefore the results of the poisson analysis isn't valid.
- Using the overdispersed model we get scaled deviance = 42.71254 w/ df = 42

  Thus scaled deviance / df = 1.0170 which is ≈ 1. Thus, the results from the overdispersed poisson analysis would appear to be valid.
- From SAS output:

Contrast	NumDF	DenDF	Fvalue	Pr > F	Chi-Sq	Pr > Chi-Sq	Type
Field Vs USDA	1	42	4.93	0.0318	4.93	0.0264	LR
Field Vs Resist	1	42	2.33	0.1340	2.33	0.1265	LR
Resist Vs Field	1	42	13.67	0.0006	13.67	0.0002	LR

- Using $\alpha_{pc} = 0.05/3 = 0.0167$ there is significant evidence of a difference in the mean egg count of the USDA & Resistant strains only.

4.) Answer the following using at most 20 words.

(H.O.5 pg 50) → (a) Yes, she is correct, But the benefit of increased power is offset by the increase in P[Type I error] when correlation is present

(H.O.5 pg 20-25) → (b) Constant variance seems to be violated. I would attempt the transformation:

$$z_{ij} = (y_{ij})^{-1/2}$$

(c) The largest y_{ij} will be the smallest z_{ij}. Use the best is smallest definition on the transformed data.

(H.O.5 pg 32) → (d) Kruskal-Wallis Test Assumptions:

All the same assumptions other than normality of data

- i.e. w/in treatments data are iid,
- treatment populations are a part of the same location scale family and only potentially differ in their location parameters.

.5 pg 9) → (e) Declare an obs y_{ij} an outlier if $|e_{ij}^*| \geq \hat{\sigma}_e \sqrt{1 - \frac{1}{n_i}}\ t_{0.005, df_E}$

$\Rightarrow |\hat{e}_{ij}| \geq \sqrt{9}\sqrt{1-\frac{1}{10}}\,(3.52) = 10.02$

- no, none of the residuals are outliers.

3.) (c) Does there appear to be correlation in the egg counts?

: B/c eggs data not normally distributed, use runs tests.

N	N+	N-	n_{Lower}	n_{upper}
9	7	8	4	13
6	6	9	4	13
8	6	9	4	10

There doesn't appear to be correlation in the data for any of the three strains.