Key to Q1 Day 2- Jan 2021

a. For estimating p_1, both E_1 *and* E_2 offer comparable performance. However, for estimating p_2, E_2 is a better choice because the standard error part associated with E_1 can collapse to 0 when the p is at the boundary, but the standard error part associated with E_2 is bounded away from 0 even when the p is at the boundary. Consequently, E_2 is a better choice when the point estimate suggests that the population proportion is considerably far away from 0.5 (suggested weight 40%)

b. Large sample size improves the coverage of both E_1 *and* E_2, in terms of achieving the nominal level when estimating p_1. However when p is at the boundary E_1 is still less reliable as compared to E_2 because sample size does not play any role in the collapse of standard error parts in E_1 (suggested weight 20%)

c. Convert the 500 standard normal samples obtained in part (b) into exponential samples with mean =2, i.e

xx=rnorm(500)

Fxx=pnorm(xx)

yy=qexp(Fxx, rate=2)

Use these samples to compute 95% CI of rate parameter.

f(y)=$\lambda e^{-\lambda y}, y > 0, \lambda > 0$

Take the transform $W = 2\lambda Y, then\ W \sim \chi^2_2, \sum W_i \sim \chi^2_{2n}$.

Then the 100(1-α)% CI for λ is given by

$$\left(\frac{\chi^2_{2n;1-\frac{\alpha}{2}}}{2\sum W_i}, \frac{\chi^2_{2n;\frac{\alpha}{2}}}{2\sum W_i}\right)$$

(suggested weight 40%)

Key Problem 2

2. Iron Deficiency Chlorosis (IDC) is a common condition in Northern US in soybeans. Symptoms of IDC include yellowing of leaf tissue and necrosis of meristem and leaf tissue. Earlier studies have shown that increased seeding rate improves IDC response. Researchers have developed two soybean lines: one IDC resistant (RS) and one IDC susceptible (SS). They designed a study to see if the occurrence of IDC between the two lines differed at differing seeding rates.

The Experimental design was an RCBD using 10 fields. Each field was divided into six plots and the six treatment combinations were randomly assigned. They evaluated 100 plants in each plot for occurrence of IDC and recorded the numbers with and without IDC. A graduate student working on the analysis came to the SC3L with the data collected from the study and her preliminary analysis. Following is a screenshot of part of the Excel file. IDC=1 are plants that have IDC, IDC=0 are plants that are IDC free.

	A	B	C	D	E
1	field_id	line	seeding	idc	y
2	1	R	25	0	49
3	1	R	25	1	51
4	1	R	50	0	60
5	1	R	50	1	40
6	1	R	150	0	78
7	1	R	150	1	22
8	1	S	25	0	22
9	1	S	25	1	78
10	1	S	50	0	34
11	1	S	50	1	66
12	1	S	150	0	44
13	1	S	150	1	56
14	2	R	25	0	60
15	2	R	25	1	40

IDC

She had taken STAT 801 and STAT 802. She also applied what she learned in 802 about using estimate statements to look at simple effects of line at fixed levels of seeding and seeding at fixed levels of line.

```
PROC IMPORT OUT= WORK.IDCALL
            DATAFILE= "C:my research\ prob2data 010622.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="IDC";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES; RUN;
proc glimmix data=idcall;
class line seeding field_id idc;
model y =line|seeding|idc/ s;
random field_id;
lsmeans line*seeding/slicediff=(line seeding);  run;
```

Following are partial results. She was concerned about the very large line*seeding*idc interactions that she was finding. She also was curious why she was not seeing any differences in line, seeding or line*seeding, and why all of the lsmeans for the line*seeding were 50.

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
line	1	99	0.00	1.0000
seeding	2	99	0.00	1.0000
line*seeding	2	99	0.00	1.0000
idc	1	99	0.41	0.5252
line*idc	1	99	724.92	<.0001
seeding*idc	2	99	172.98	<.0001
line*seeding*idc	2	99	7.06	0.0014

line*seeding Least Squares Means						
line	seeding	Estimate	Standard Error	DF	t Value	Pr > |t|
R	25	50.0000	1.1524	99	43.39	<.0001
R	50	50.0000	1.1524	99	43.39	<.0001
R	150	50.0000	1.1524	99	43.39	<.0001
S	25	50.0000	1.1524	99	43.39	<.0001
S	50	50.0000	1.1524	99	43.39	<.0001
S	150	50.0000	1.1524	99	43.39	<.0001

Upon seeing the design of the experiment and how the data were collected, you knew that the data were not Normally distributed and should be reanalyzed using the appropriate distribution. Note that these can be included in the Appendix.

a) What is it about her data and analysis that would explain the results that concerned her?

b) What is an appropriate model for analyzing this experiment?

You explain this to the graduate student and agreed to rerun her analysis under the correct distribution and provide her with a short summary report, which would include an explanation of how the data are distributed and how to correctly interpret the results including any significant interaction effect, and any conclusions based on the original research question. The summary report should include

- 2-3 pages of written report not including any relevant supporting summary tables — 25 pts
- Appendix including answers to questions a) and b) — 5 pts each
- Appendix with supporting summary tables — 5
- Appendix with SAS program that you used. — 5

Note that the attached draft report is fairly bare-bones. Hopefully, the students will expand a bit. It would be nice to include figures, but it takes more time programming those. I've also included in the appendix further explanation of odds and odds ratios. This is not necessary, but I always encourage the SC3L staff to try and include explanations of non-standard results. The important things are that the student recognizes that the data are Binomial, they investigate the interaction, use and interpret the odds ratios and write a conclusion based on the original research question.

Report:

Research Question:

Are there differences in the occurrence of IDC between the two lines (Resistant and Susceptible) and do the differences different at differing seeding rates (25, 50 and 150).

Treatment and Experimental Design:

The Treatment design was a factorial with two lines (Resistant and Susceptible) and three seeding rates (25, 50 and 150). The Experimental design was an RCBD using 10 fields. Each field was divided into six plots and the six treatment combinations were randomly assigned. One hundred (100) plants were randomly selected in each plot and evaluated for occurrence of IDC.

Statistical Model:

The PROC GLIMMIX procedure in SAS 9.4 was used to analyze the occurrence of IDC. A generalized linear mixed model was used to account for the underlying Binomial distribution of the data. Plants with IDC were considered a "success" for the proportions (see the appendix for more detailed description). The linear model included the fixed effects of line, seeding rate and their interaction, as well as the random blocking effect of field. An α= 0.05 level was used for determining significance.

Results:

Table 1 in the Appendix presents the ANOVA table of the results from the statistical analysis. There was a significant line by seeding rate effect on the proportion of plants with IDC, which means that the differences in the response among the seeding rates were different for the two different lines. Therefore, we need to look at the simple effects (or the differences among the seeding rate within each line).

Table 2 presents the line by seeding rate LSMeans. The "estimate" is in the model scale (see the appendix for more information), while the "mean" is in the data scale and is an estimate of the proportion of plants with IDC. For example the estimated proportion of Resistant plants with IDC at the 150 seeding rate is .257, while the estimated proportion of Susceptible plants with IDC at the 150 seeding rate is .542. The Susceptible plants had estimated proportions over .50 for all three seeding rates, while the Resistant plants had estimated proportions under .50 for all three seeding rates.

Table 3 presents the simple effects between the seeding rates within each line. The odds ratios are the odds of having IDC under the first level of the treatment over the odds of having IDC under the second level of the treatment (see the appendix for further description of odds ratios). For example, in the Resistant line, the odds of having IDC at the 25 seeding rate is around 1.4 times greater than the odds of having IDC at the 50 seeding rate. In other words, the higher seeding rate is significantly better for controlling IDC in the Resistant line. In the Susceptible line, the odds of having IDC at the 25 seeding rate is around 2 times greater than the odds of having IDC at the 50 seeding rate. Which means that the difference between the seeding rates is greater for the Susceptible line than the Resistant line. When

comparing the odds ratios for 50 vs 150, the difference between the seeding rates is smaller for the Susceptible line than the Resistant line. This is what is causing the significant interaction. Overall, the odds of having IDC were greater at lower seeding rates within each of the lines.

Table 4 presents the simple effects between lines within each seeding rate. For all three seeding rates, the odds of having IDC were smaller for the Resistant line than for the Susceptible line.

Based on the original research question, IDC occurrence did differ between seeding rates for both the Resistant and the Susceptible lines. Both lines had decreases in the proportion of plants with IDC as the seeding rate increased. However, the Resistant line was better at controlling for IDC at all seeding rates with less than half of the plants having IDC compared to over half having IDC for the Susceptible line.

a. The original analysis used treated the data as Normally distributed, with the response being the number of plants (out of 100) with idc (idc=1) or without idc (idc=0). So plants can only fall into one of two idc categories, which means that the data are Binomially distributed. The original analysis treated the count within each idc category as the response variable and included idc as a factor. The LSMeans for Line*Seeding rate are all 50, because they are calculated using the model average of any other factors in the model. Because y response always totals 100 over the two IDC levels, all of the IDC (and interactions involving IDC) average 50.

b. The following describe the distribution and statistical model.

$y_{ijk} \mid B_i \sim Binomial(n_{ijk}, \pi_{ijk})$

$B_i \sim N(0, \sigma_b^2)$

n_{ijk} is the number of trials within the ijkth group

π_{ijk} is the population proportion of "success" or proportion of plants under tihe ijkth combination with IDC

Linear predictor:

$\eta_{ijk} = \eta + B_i + L_j + S_k + LS_{jk}$

η is the intercept

B_i denotes the ith block (field) effect

L_j denotes the jth line effect

S_k denotes the kth seeding rate effect

LS_{jk} denotes the kth seeding rate by lth line interaction effect

Because the data are Binomially distributed, a logit link function is used

$$\eta_{ijk} = \log it(\pi_{ijk}) = \log\left[\frac{\pi_{ijk}}{(1-\pi_{ijk})}\right]$$

Note that the estimates that will be determined by SAS will be in this model scale. In order to interpret the results in the original scale, differences between levels of effects are described using odds ratios, which are the odds of success under one level of the treatment over the odds of success under another level of the treatment.

For example, in Table 3, the odds ratio for occurrence of IDC at seeding rate 25 vs 50 in the Resistant line is calculated as follows, using the estimated proportions from Table 2.

Odds for IDC at seeding rate 25 and 50, respectively are

$$\left[\frac{p_{r25}}{(1-p_{r25})}\right] = \frac{0.4780}{(1-0.4780)} = 0.9157 \qquad \left[\frac{p_{r50}}{(1-p_{r50})}\right] = \frac{0.3940}{(1-0.3940)} = 0.6502$$

So the odds ratio is $\left[\dfrac{\dfrac{p_{r25}}{(1-p_{r25})}}{\dfrac{p_{r50}}{(1-p_{r50})}}\right] = \dfrac{0.9157}{0.6502} = 1.408$

Which agrees with Table 3.

Table 1:

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
line	1	45	389.43	<.0001
seeding	2	45	97.21	<.0001
line*seeding	2	45	5.56	0.0069

Table 2:

line*seeding Least Squares Means													
line	seeding	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Mean	Standard Error Mean	Lower Mean	Upper Mean
R	25	-0.08806	0.06331	45	-1.39	0.1711	0.05	-0.2156	0.03945	0.4780	0.01580	0.4463	0.5099
R	50	-0.4305	0.06472	45	-6.65	<.0001	0.05	-0.5609	-0.3002	0.3940	0.01545	0.3633	0.4255
R	150	-1.0616	0.07237	45	-14.67	<.0001	0.05	-1.2074	-0.9159	0.2570	0.01382	0.2302	0.2858
S	25	1.0933	0.07293	45	14.99	<.0001	0.05	0.9464	1.2402	0.7490	0.01371	0.7204	0.7756
S	50	0.3971	0.06450	45	6.16	<.0001	0.05	0.2672	0.5270	0.5980	0.01550	0.5664	0.6288
S	150	0.1684	0.06347	45	2.65	0.0110	0.05	0.04056	0.2962	0.5420	0.01576	0.5101	0.5735

Table 3:

Simple Effect Comparisons of line*seeding Least Squares Means By line													
Simple Effect Level	seeding	seeding	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Odds Ratio	Lower Odds Ratio	Upper Odds Ratio
line R	25	50	0.3425	0.09053	45	3.78	0.0005	0.05	0.1601	0.5248	1.408	1.174	1.690
line R	25	150	0.9736	0.09615	45	10.13	<.0001	0.05	0.7799	1.1672	2.647	2.181	3.213
line R	50	150	0.6311	0.09708	45	6.50	<.0001	0.05	0.4356	0.8266	1.880	1.546	2.286
line S	25	50	0.6961	0.09736	45	7.15	<.0001	0.05	0.5001	0.8922	2.006	1.649	2.441
line S	25	150	0.9249	0.09668	45	9.57	<.0001	0.05	0.7302	1.1196	2.522	2.075	3.064
line S	50	150	0.2287	0.09049	45	2.53	0.0151	0.05	0.04649	0.4110	1.257	1.048	1.508

Table 4:

Simple Effect Comparisons of line*seeding Least Squares Means By seeding													
Simple Effect Level	line	line	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Odds Ratio	Lower Odds Ratio	Upper Odds Ratio
seeding 25	R	S	-1.1813	0.09658	45	-12.23	<.0001	0.05	-1.3759	-0.9868	0.307	0.253	0.373
seeding 50	R	S	-0.8277	0.09137	45	-9.06	<.0001	0.05	-1.0117	-0.6436	0.437	0.364	0.525
seeding 150	R	S	-1.2300	0.09626	45	-12.78	<.0001	0.05	-1.4239	-1.0361	0.292	0.241	0.355

```
PROC IMPORT OUT= WORK.IDCALL
            DATAFILE= "C:\Users\Kathy\Statistics Department\comprehensive
exams\prob2data 010622.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="IDC";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
*need to first change the data to binomial;

proc transpose data=idcall out=idcc;
by field_id line seeding;
var y;
run;
*col2 has the number of IDC plants, so that is the number to put into the idc
variable;
data idc2;
keep field_id line seeding idc total;
set idcc;
idc=col2;
total=col1+col2;
run;

proc glimmix data=idc2;
class line seeding field_id;
model idc/total  =line|seeding/link=logit htype=3 dist=bin s;
random intercept/subject=field_id;
lsmeans line*seeding/slicediff=(line seeding) ilink or cl;
run;
```