1. Let $X \sim N(0,1)$. Define the following two Bernoulli random variables $Y_1 = I_{X \leq 0}$ and $Y_2 = I_{X \leq -2}$. Let $p_1 = P(Y_1 = 1)$ and $p_2 = P(Y_2 = 1)$

a. Draw n =50 iid copies of X and obtain the corresponding samples of Y_1 and Y_2. We wish to obtain 95% CI for p_1 and p_2 from the foregoing 50 samples of Y_1 and Y_2

Two interval estimators of $p_i, i = 1,2$ are given by

$$E_1 = \widehat{p_\iota} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{p_\iota}(1 - \widehat{p_\iota})/n} \, , i = 1,2$$

$$E_2 = \widetilde{p_\iota} \pm z_{\frac{\alpha}{2}} \widetilde{s_\iota}, \text{ where } \widetilde{p_\iota} = \frac{\hat{p} + z^2_{\alpha/2}/2n}{1 + z^2_{\alpha/2}/n} \text{ and } \widetilde{s_\iota} = \frac{\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}, i = 1,2$$

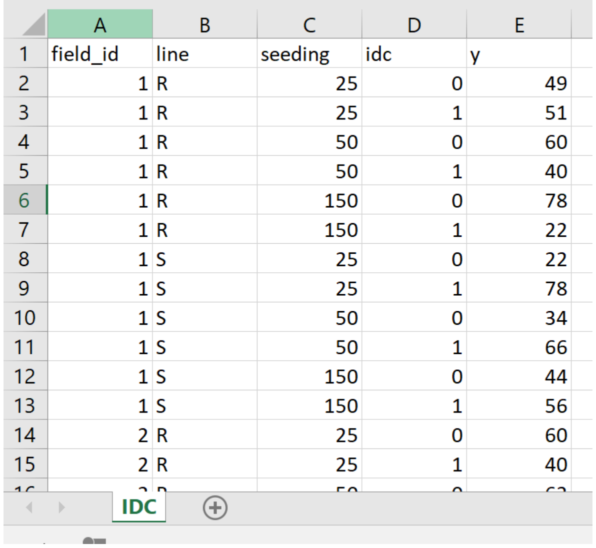
Obtain the coverage probabilities of E_1 and E_2 based on 1000 replicates. Which estimator is better at achieving the nominal coverage level?

b. Now draw n=500 iid copies of X. Obtain the corresponding samples of Y_1 and Y_2 and compute the coverage probabilities of E_1 and E_2 based on 1000 replicates. Which estimator is better at achieving the nominal coverage level? Does your answer match with part (a)

c. Again let X~$N(0,1)$. Find a transformation $Y_3 = g(X)$ such that Y_3 has an exponential distribution with mean parameter 0.5. Draw n =500 iid copies of X and obtain the corresponding samples of Y_3. Construct a 95% CI for the rate parameter of the exponential distribution and assess the coverage probabilities based on 1000 replicates.

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.

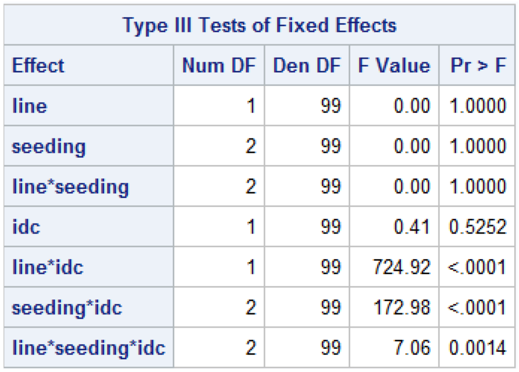


	A	B	C	D	E
1	field_id	line	seeding	idc	y
2	1	R	25	0	49
3	1	R	25	1	51
4	1	R	50	0	60
5	1	R	50	1	40
6	1	R	150	0	78
7	1	R	150	1	22
8	1	S	25	0	22
9	1	S	25	1	78
10	1	S	50	0	34
11	1	S	50	1	66
12	1	S	150	0	44
13	1	S	150	1	56
14	2	R	25	0	60
15	2	R	25	1	40
16	2	R	50	0	[illegible]

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.

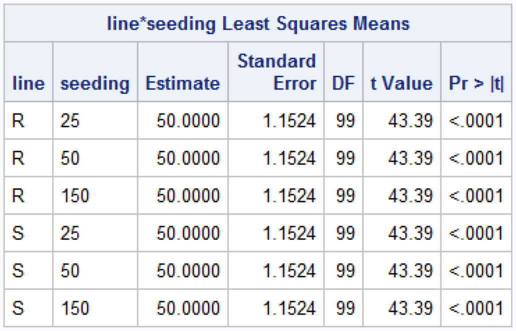
```
PROC IMPORT OUT= WORK.IDCALL
            DATAFILE= "C:my research\ prob2data 010622.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="IDC";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
proc glimmix data=idcall;
class line seeding field_id idc;
model 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
- Appendix with answers to questions a) and b)
- Appendix with supporting summary tables
- Appendix with SAS program that you used.