Sommertagung der AG Landwirtschaftliches Versuchswesen
Monheim 29. Juni, 2023

FINDING NEMAS – TRANSFORMING COUNT DATA IS THE WORST OF BOTH WORLDS

Edzard van Santen
University of Florida
Institute for Food and Agricultural Sciences
Statistical Consulting Unit

Two Questions

1. What century do we live in?

2. Shouldn't our approach to data analysis reflect the times we live in?

20^{th} century approach

- Make the data conform to a model that we are familiar with
  - Assuming everything is normal and
  - If it isn't, make it so. Fitting a square peg into a round hole.
- Why?
  - Researchers recognized the limitations of the classic LM approach
  - Developed empirical transformations that 'fixed the problem'
  - A classic is the arcsine(sqrt) transformation for weed proportion, which 'fixed' the heterogeneous variance issue
- Please Notice:
  - I am talking about the dependent (= response) variables, i.e., stuff we measure or count

Deficiencies of the linear model may be addressed

- By ignoring the deficiencies
- By transforming the response variable
- By creating of variance groups
- Choosing an appropriate distribution

Approaches to Fixing Problems

- Transformations to
  - stabilize variances
  - obtain a linear relationship
  - normalize the distribution
- Problems with transformations
  - Log transformation to stabilize variance implies that the model on the original scale is multiplicative
-

Main point from these approaches

- These solutions are appealing because they are
  - Simple and cheap
  - Produce reasonable results because of robustness
- but do not address the problem that the model is incorrect
  - As what Stroup puts it: It does not answer the question of the likely underlying probabilistic process leading to a given body of data
- Generalized linear mixed models to the rescue
  - Utilize models that are appropriate for the data

21st century approach

- Fit a model to the data, not data to a model, i.e., the characteristics of the data should drive the model

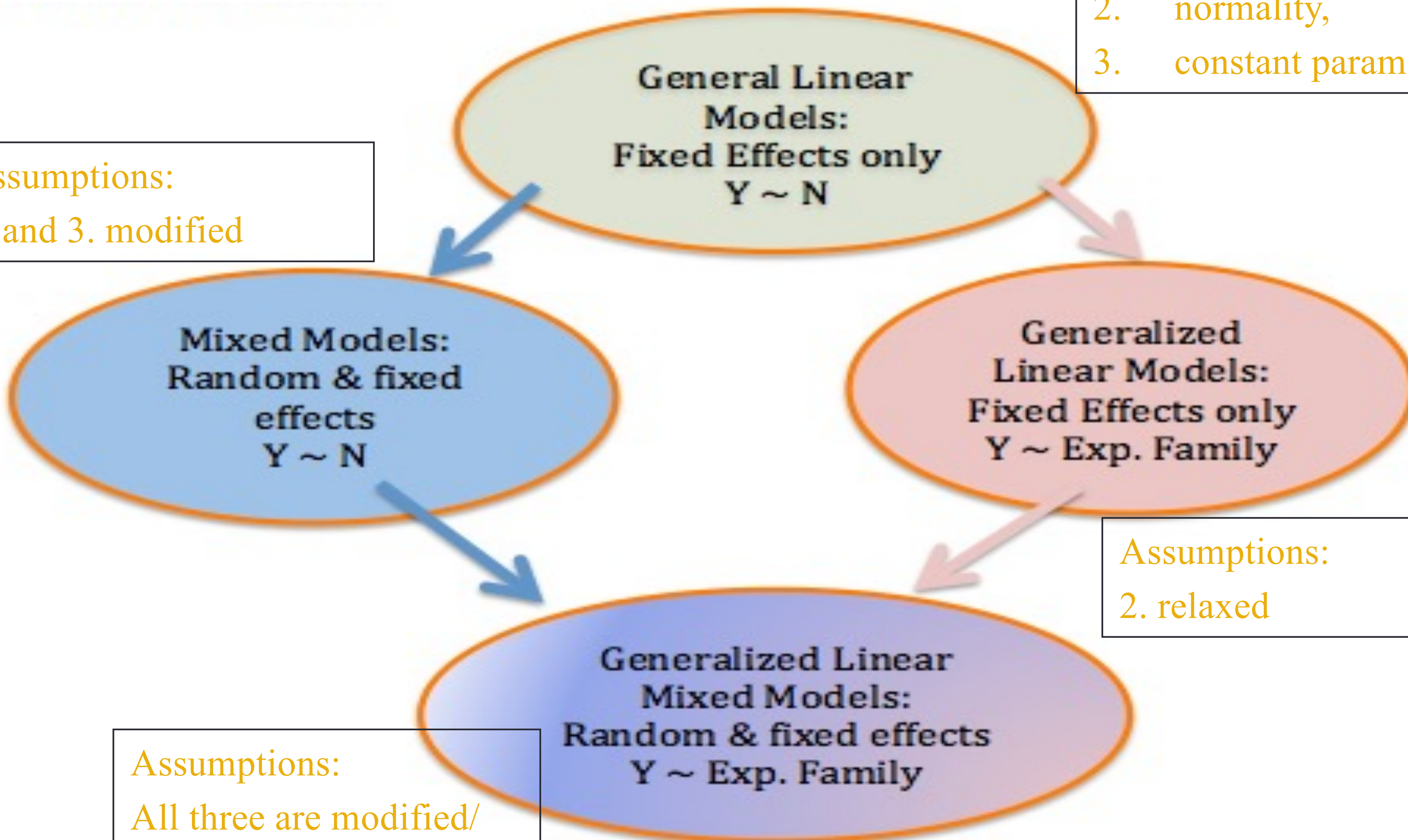
- Walt Stroup encouraged researchers to ask the question
  - What is a likely probabilistic process that led to the generation of the body of data at hand?
  - Determining the mass of something.
  - Determining the length of something.
  - Determining the proportion of a constituent.
  - Determining the ratio or product of something.
  - Applying a certain number of insects and counting how many can be found on several objects.
  - Counting the number of nematodes in a soil sample.

Probabilistic processes

- Determining the mass of something - Gaussian
- Determining the length of something - Gaussian
- Determining the proportion of a constituent - Beta
- Determining the ratio or product of something - lognormal
- Applying a certain number of insects and counting how many can be found on several objects - Binomial
- Counting the number of nematodes in a soil sample – Poisson? Probably not

General Modeling Framework

Assumptions:
1. independence,
2. normality,
3. constant parameters



Assumptions:
1. and 3. modified

Assumptions:
2. relaxed

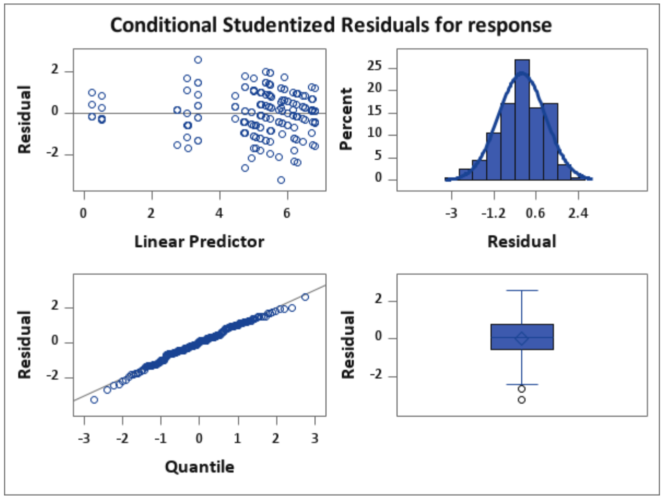
Assumptions:
All three are modified/ relaxed

What happens when you transform count data?

- Case study: ***Meloidogyne enterolobii***
  - GH-study
  - Comparing effect of antagonists and nematicides in two soils
  - 3 experimental repeats
  - 5 replicate pots per experimental repeat
  - Response variables: galling index, eggs/plant, eggs/g of root

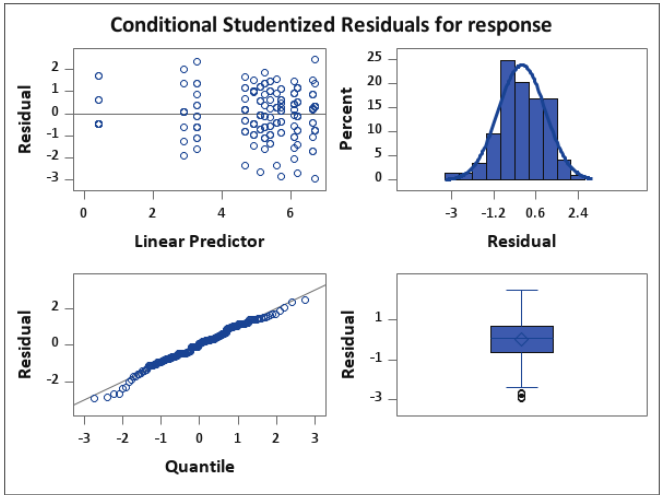
Gall index

Normal



AICc = 787.5

Normal with 4 Vgroups



AICc = 768.82

Index is a Likert type scale ➔ multinomial

With 15 obs per treatment we can capitalize on the CLT

Gall index

Normal

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Soil	1	181.1	63.35	<.0001
Trt	6	181.9	9.27	<.0001
Soil*Trt	6	181.1	5.93	<.0001

AICc = 787.5

Normal with 4 Vgroups

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Soil	1	163	65.53	<.0001
Trt	6	118.6	10.28	<.0001
Soil*Trt	6	118.6	7.24	<.0001

AICc = 768.82

Gall index: LSmeans

Normal

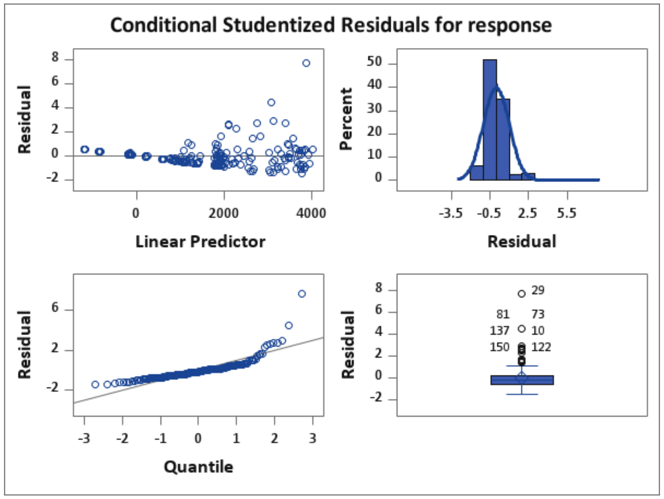
Soil	Trt	Mu	Lmu	Umu
NP	T0	6.10	5.11	7.09
NP	T1	6.20	5.21	7.19
NP	T2	5.74	4.75	6.73
P	T0	0.44	-0.59	1.46
P	T1	4.67	3.68	5.66
P	T2	5.13	4.14	6.12

Normal with 4 Vgroups

Soil	Trt	Mu	Lmu	Umu
NP	T0	6.10	5.03	7.17
NP	T1	6.20	5.37	7.03
NP	T2	5.74	4.67	6.81
P	T0	0.43	-0.10	0.96
P	T1	4.67	3.60	5.74
P	T2	5.13	4.30	5.96

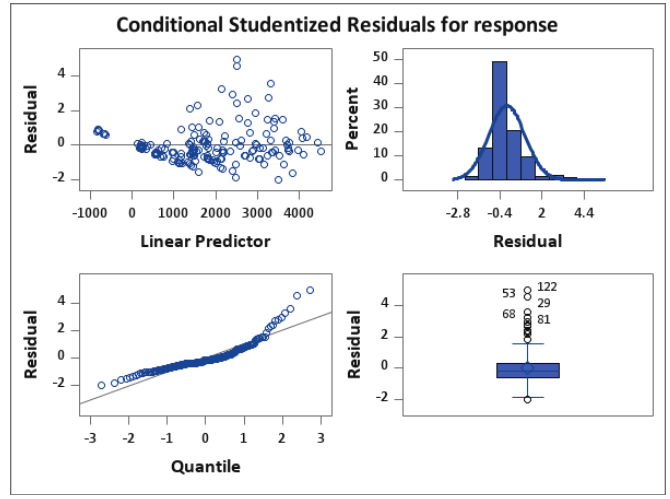
Eggs per g of root

Normal



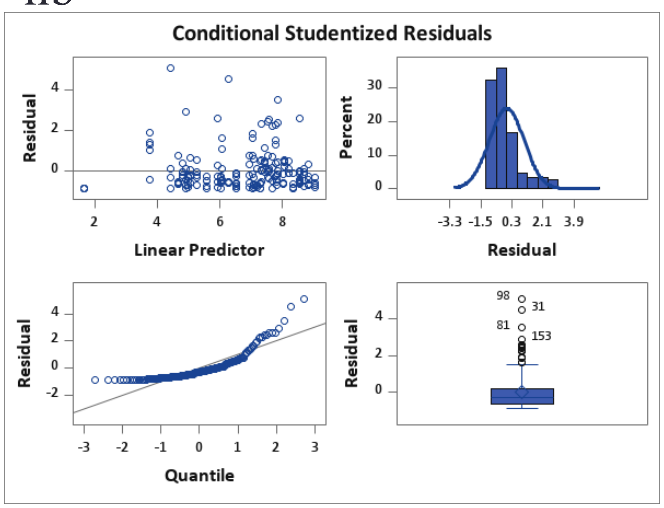
AICc = 3214.6

Normal with 4 Vgroups



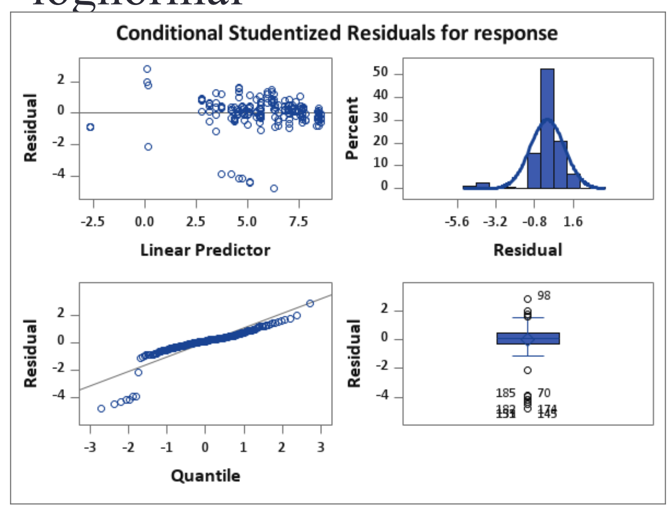
AICc = 3142.9

nb



Pearson Chi-Square / DF = 0.68

lognormal



Eggs per g of root

Normal

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Soil	1	159	18.40	<.0001
Trt	6	159	0.69	0.6597
Soil*Trt	6	159	1.75	0.1121

Normal with 4 Vgroups

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Soil	1	159	20.85	<.0001
Trt	6	159	0.88	0.5089
Soil*Trt	6	159	3.30	0.0043

nb

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Soil	1	159	13.36	0.0003
Trt	1	159	11.47	0.0009
Soil*Trt	6	159	27.57	<.0001

lognormal

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Soil	1	159	62.23	<.0001
Trt	6	159	5.64	<.0001
Soil*Trt	6	159	6.56	<.0001

Eggs per g of root: LSmeans

Normal

Soil	Trt	Mu	Lmu	Umu
NP	T0	2619	612	4625
NP	T1	2796	980	4611
NP	T2	3162	1347	4977
P	T0	523	-1539	2586
P	T1	1196	-620	3011
P	T2	1234	-581	3049

Normal with 4 Vgroups

Soil	Trt	Mu	Lmu	Umu
NP	T0	2456	484	4429
NP	T1	2796	1009	4582
NP	T2	3162	271	6053
P	T0	332	-1221	1886
P	T1	1196	-277	2668
P	T2	1234	-239	2706

nb

Soil	Trt	Mu	Lmu	Umu
NP	T0	2319	1381	3893
NP	T1	2040	882	4718
NP	T2	2873	1242	6645
P	T0	28	14	56
P	T1	581	137	1089
P	T2	657	344	1215

lognormal

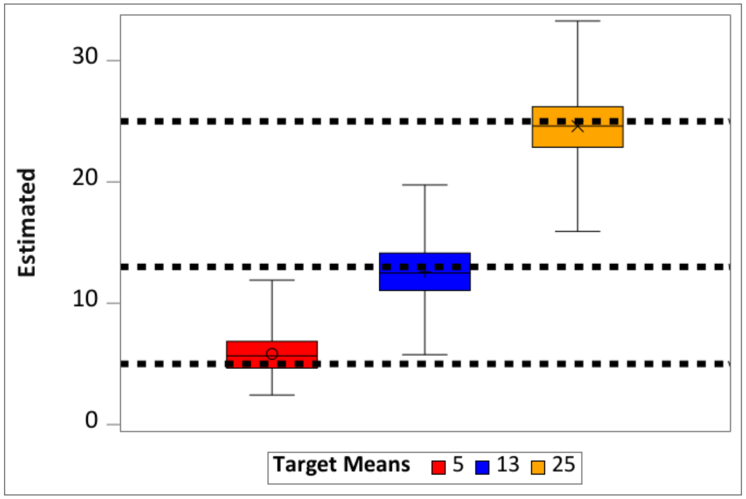
Soil	Trt	Mu	Lmu	Umu
NP	T0	1417	185	10822
NP	T1	1407	226	8779
NP	T2	1317	211	8214
P	T0	0	0	3
P	T1	396	63	2467
P	T2	456	73	2847

Simulation

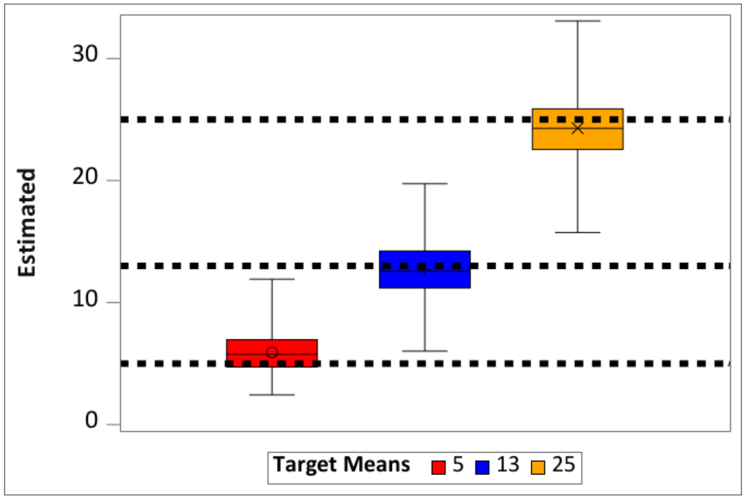
- Repeats: 1000
- Blocks: 5 or 15
- Block effect ~ N(0, 20);
- Trt means: 5, 20, 35 or 5, 50 500
- Data generating distribution: nb
- Data analysis distributions: NB, Poisson, Normal, lognormal
  - Calculated estimated means and variances

Estimated means (r = 5)

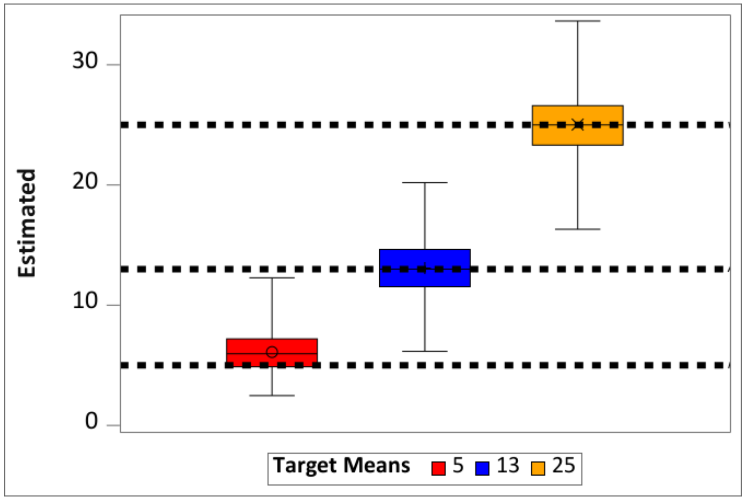
nb



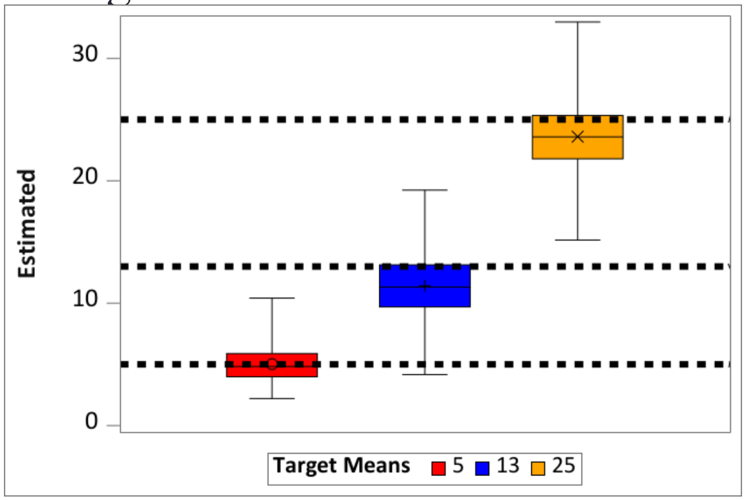
Poisson



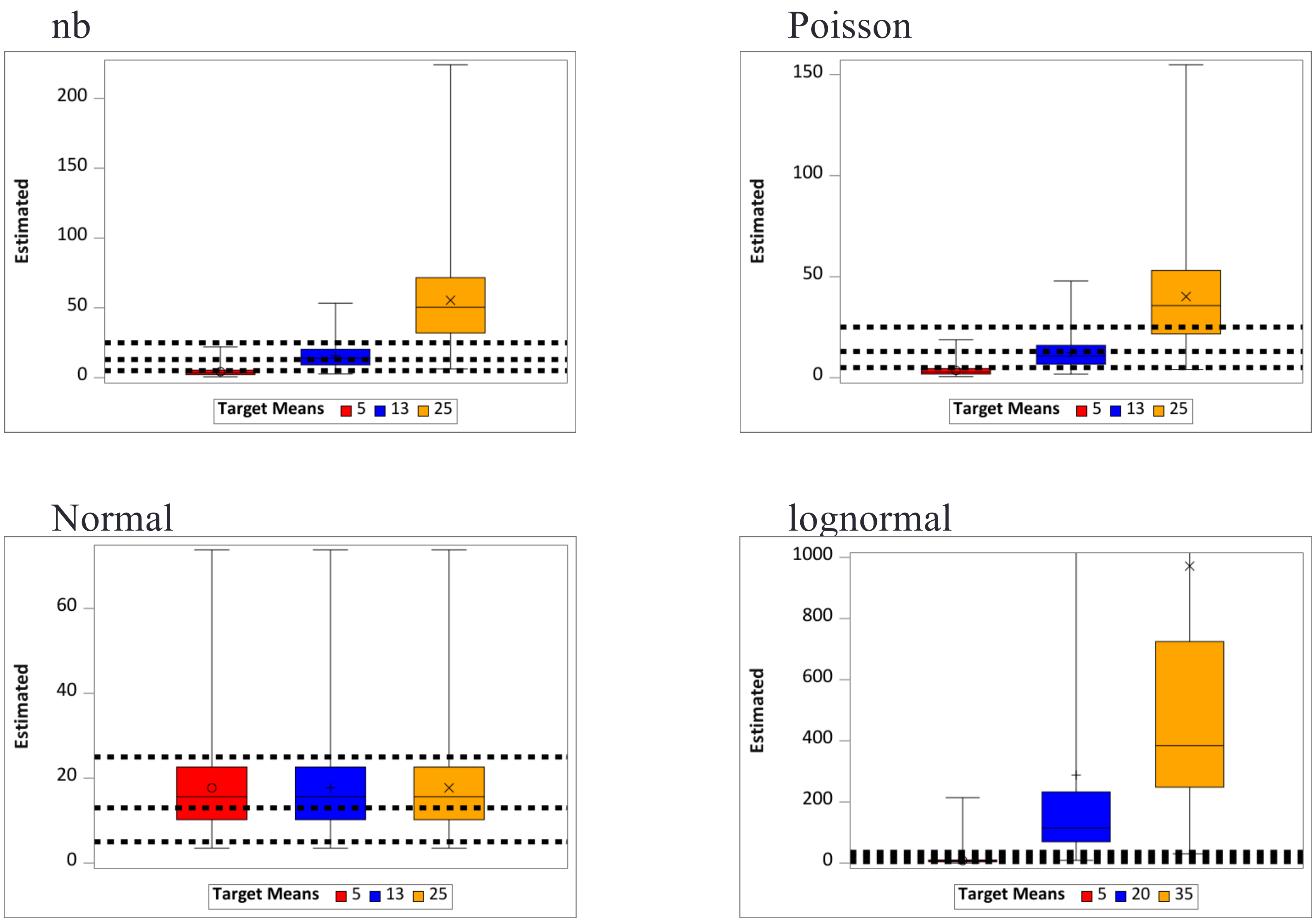
Normal



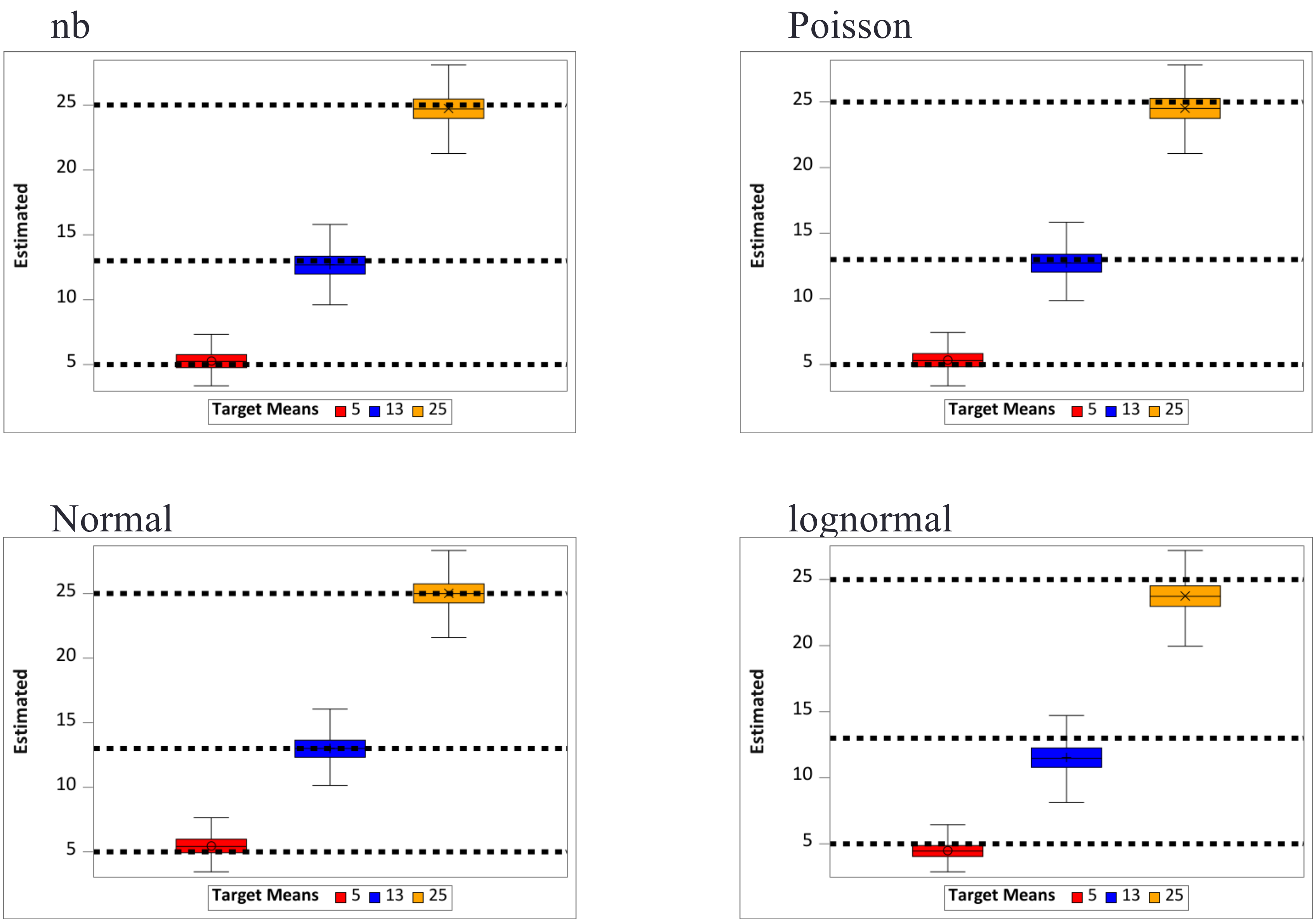
lognormal



Estimated variances (r =5)

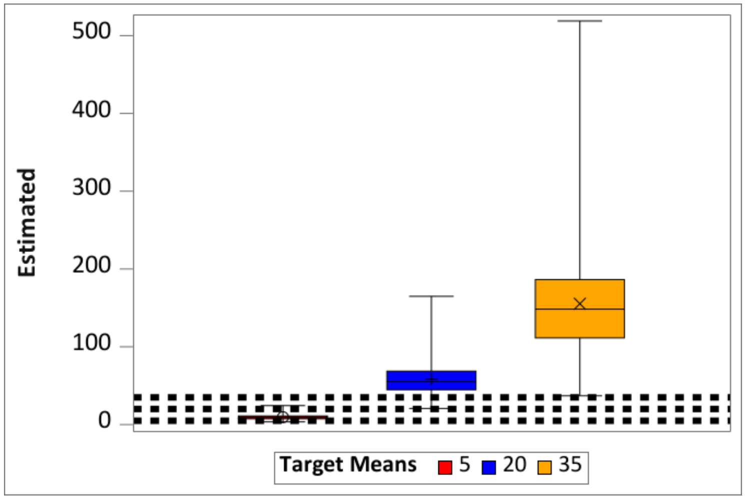


Estimated means (r = 15)

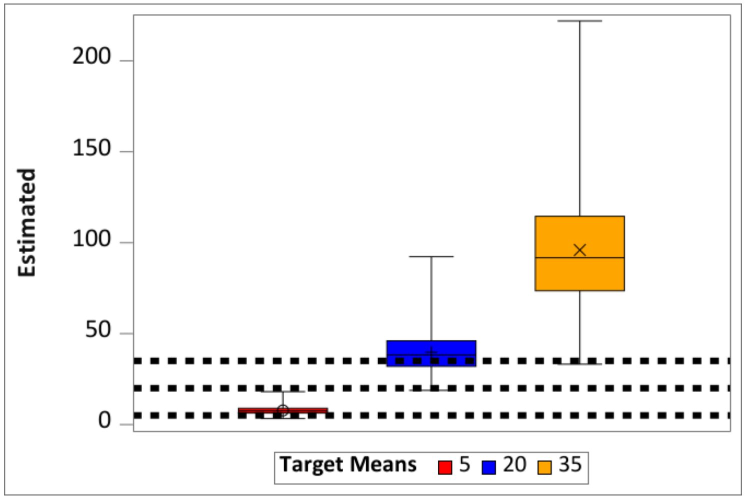


Estimated variances (r = 15)

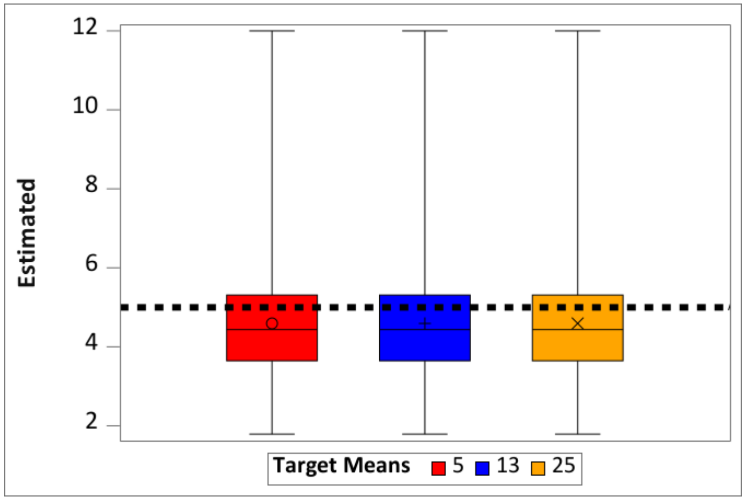
nb



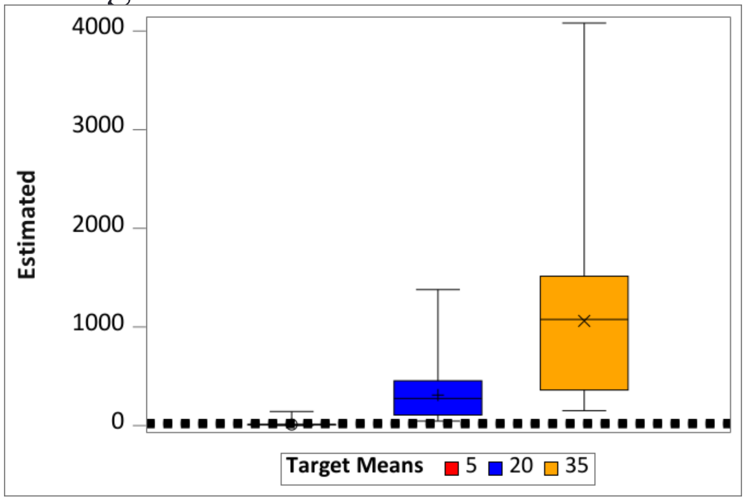
Poisson



Normal



lognormal



My conclusion

- Use GLMM with nb
- Sometimes, a Poisson will work and may lead to faster convergence

- Alternate approach: Use the bacteriologist's approach
  - Express counts as decalogs and stick with it, i.e.,
  - argue on the decalog scale

Power calculations using PROC GLIMMIX

- Egg laying capacity of Monarch butterflies on 4 species of milkweed
- Blocks of 4 pots are placed at isolated locations
- How many blocks are needed?

Species_N	Species	Common	Native	Native to	Egg_N	Expct prop	H_0
1	*A. incarnata*	pink swamp	Yes	N America	102	0.34	0.25
2	*Asclepias curassavica*	tropical	No	tropical Americas	97	0.32	0.25
3	*Gomphocarpus physocarpu:*	balloon	No	tropical Africa	61	0.20	0.25
4	*Calotropis gigantea*	giant	No	SE Asia	41	0.14	0.25
					300		

Power calculation: Step 1

```
%let N = 300;
proc datasets library=work kill memtype=data nolist;quit;
data power;
     input trt p;
     N=&N;
     mu=N*p;
     do block = 1 to 10;
     output;
     end;
     datalines;
     1 0.34
     2 0.32
     3 0.20
     4 0.14
;run;
/*proc print data=power;* (obs=5);run;*/
```

Power calculation: Step 2

```
ods select none;
proc GLIMMIX data=power plots=studentpanel;
    parms (10 .1)/hold=1,2;
    class Block trt;
    model Mu=trt/dist=nb;* ddfm=kr;
    random Intercept/subject=Block;
    Contrast 'trt 1 vs 2' trt -1 1 0 0;
    Contrast 'trt 1 vs 3' trt -1 0 1 0;
    Contrast 'trt 1 vs 4' trt -1 0 0 1;
    Contrast 'trt 2 vs 3' trt 0 -1 1 0;
    Contrast 'trt 2 vs 4' trt 0 -1 0 1;
    Contrast 'trt 3 vs 4' trt 0 0 -1 1;
/*  lsmeans loc*trt;*/
    ods output tests3=F_overall contrasts=F_contrasts;
run;
ods select all;
```

Power calculation: Step 3

```
Data power_calc;
    set F_contrasts;
    nc_parm=numdf*fvalue ;
    alpha=0.05;
    F_crit=Finv(1-alpha,numdf,dendf,0);
    Power_F=1-probF(F_crit,numdf,dendf,nc_parm);
run;
proc print;run;
```

Based on a total of 300 eggs and 10 blocks

Obs	Label	NumDF	DenDF	FValue	ProbF	nc_parm	alpha	F_crit	Power_F
1	trt 1 vs 2	1	27	0.17	0.6861	0.1669	0.05	4.21001	0.06798
2	trt 1 vs 3	1	27	12.43	0.0015	12.4328	0.05	4.21001	0.92480
3	trt 1 vs 4	1	27	33.70	<.0001	33.7013	0.05	4.21001	0.99986
4	trt 2 vs 3	1	27	9.73	0.0043	9.7279	0.05	4.21001	0.85230
5	trt 2 vs 4	1	27	29.18	<.0001	29.1768	0.05	4.21001	0.99941
6	trt 3 vs 4	1	27	5.29	0.0294	5.2902	0.05	4.21001	0.60173

Fewer or more total eggs

Based on a total of 100 eggs and 10 blocks

Obs	Label	NumDF	DenDF	FValue	ProbF	nc_parm	alpha	F_crit	Power_F
1	trt 1 vs 2	1	27	0.14	0.7102	0.1410	0.05	4.21001	0.06517
2	trt 1 vs 3	1	27	10.08	0.0037	10.0771	0.05	4.21001	0.86428
3	trt 1 vs 4	1	27	26.17	<.0001	26.1703	0.05	4.21001	0.99850
4	trt 2 vs 3	1	27	7.85	0.0093	7.8543	0.05	4.21001	0.77088
5	trt 2 vs 4	1	27	22.58	<.0001	22.5783	0.05	4.21001	0.99556
6	trt 3 vs 4	1	27	3.96	0.0569	3.9579	0.05	4.21001	0.48352

Based on a total of 900 eggs and 10 blocks

Obs	Label	NumDF	DenDF	FValue	ProbF	nc_parm	alpha	F_crit	Power_F
1	trt 1 vs 2	1	27	0.18	0.6765	0.1779	0.05	4.21001	0.06918
2	trt 1 vs 3	1	27	13.50	0.0010	13.4971	0.05	4.21001	0.94298
3	trt 1 vs 4	1	27	37.32	<.0001	37.3201	0.05	4.21001	0.99996
4	trt 2 vs 3	1	27	10.58	0.0031	10.5791	0.05	4.21001	0.87999
5	trt 2 vs 4	1	27	32.36	<.0001	32.3645	0.05	4.21001	0.99978
6	trt 3 vs 4	1	27	5.97	0.0214	5.9679	0.05	4.21001	0.65374

Conclusion

- Using a 21st century approach to count data is straight forward
  - for data analysis as well as
  - for power calculations.
- Standard software implementation in SAS and in R

- What's your pain threshold as far as the asymptotic properties are concerned?

Further avenues

- ZINB models
- Bayesian approaches

Further Reading

- Walter F. Stroup (2013). Generalized Linear Mixed Models. CRC Press, Boca Raton, FL.
  - The writing is dense, but it has good examples in SAS