(330 pts total):

Section 1 – Short answer

1. (10 pts) In equations or words, what is the definition of a probability density function (PDF) for a continuous variable, and how it is related to probability?

2.(10 pts) Define pseudoreplication (in words).

3. (18 pts)

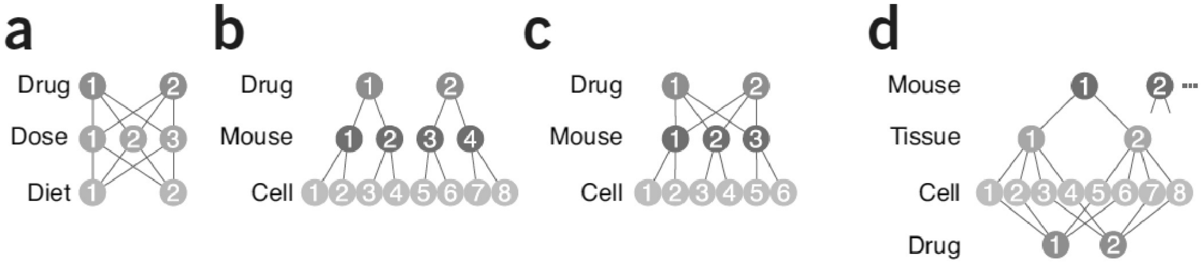


Figure 2 from Krzywinski et al. 2014

a) (3 pts) Which factors are crossed and which are nested in the experimental design illustrated in panel a?

b) (3 pts) Which factors are crossed and which are nested in the experimental design illustrated in panel b?

c) (3 pts) Which factors are crossed and which are nested in the experimental design illustrated in panel c?

d) (3 pts) Which factors are crossed and which are nested in the experimental design illustrated in panel d?

e) (6 pts) Which factors are best considered fixed effects and which best considered random effects (Circle one for each Factor):

Drug: Fixed / Random

Dose: Fixed / Random

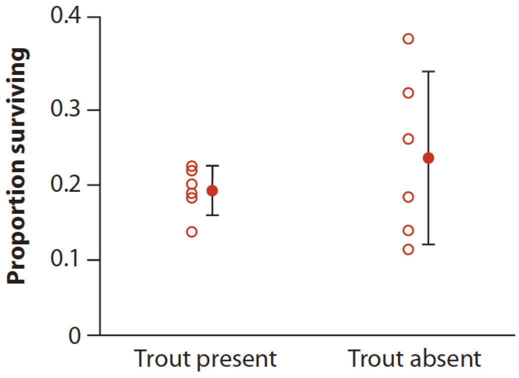
Diet: Fixed / Random

Mouse: Fixed / Random

Cell: Fixed / Random

Tissue: Fixed / Random

4. (20 pts) Below is a scatterplot of data on the survivorship of chinook salmon in 12 streams, 6 of which had brook trout (an introduced species) and 6 of which did not have brook trout. The open circles represent the original data and the solid circle represents the mean across the 6 streams in each category with their 95th percentile confidence intervals.



a) (10 pts) The authors want to test whether survival of chinook salmon is higher when brook trout are absent, and they consider using an unpaired t-test. These data violate 2 assumptions of the t-test; what are these violations?

b) (10 pts) Using words and/or equations, describe a more appropriate method of testing this hypothesis.

5. (30 pts)
a) (15 pts) What is the complete model description (i.e., equation(s)) for ordinary least squares (OLS) regression?

b) (15 pts) Illustrate heteroskedacity. In what way does heteroskedacity violate the model as described in part a?

Section 2 – Long answer

10. (142 pts)

In 1985, Doug Futuyma and Lawrence Harshman wrote a paper in *American* Naturalist entitled "Survivorship and Growth of Sexually and Asexually Derived Larvae of *Alsophila pometaria* (Lepidoptera: Geometridae)". For this paper, geometrid moths were collected and reared in bags on Oak and Red Maple trees on Stony Brook's campus.

	# moth survived on...		
Genetic Group	Oak	Maple	Both Oak and Maple
A	24/30=80%	6/30=20%	30/60=50%
B	33/60=55%	2/60=3%	35/120=29%
C	29/45=64%	6/45=13%	35/90=39%
D	28/45=62%	6/45=13%	34/90=38%
All groups	114/180=63%	20/180=11%	134/360=37%

Note that the 37% in the lower right hand cell is the overall survival rate of moths in the experiment.

Part I (65 pts)

a) (5 pts) What is the expected number of moths to survive (out of 30) in Genetic Group A on Oak trees? (In other words, based on this experiment, what is the number of moths you would expect to have survive if you were to repeat this experiment with 30 moths sampled in Genetic Group A on Oak trees?)

b) (10 pts) What is the (approximate) 95th percentile confidence interval on the expected number of moths to survive (out of 30) in Genetic Group A on Oak trees?

c) (10 pts) Under the null hypothesis that genetic group and host tree species have independent effects on moth survival, what is the expected number of moths to survive in Genetic Group A on Oak trees? (Note that this question is different from Part a) because in Part a) we did not assume independence, and here I am asking you for the expected number under the null hypothesis of independence between genetic group and host tree species.)

d) (15 pts) What is the (approximate) 95th percentile confidence interval on the expected number of moths to survive (out of 30) in Genetic Group A on Oak trees, under the null hypothesis of independence between genetic group and host tree species?

e) (10 pts) Is this (i.e., using your answers to parts b and d) evidence in favor of rejecting the null hypothesis? Why or why not?

f) (15 pts) Our test in part e has low power. What other information could we add to create a more powerful test of this null hypothesis?

Part II (37 pts)

A reasonable model for this kind of data would be the following

$$Y_{ijk} \sim Binom(n_{ij}, p_{ij})$$

$$logit(p_{ij}) = \beta_0 + \beta_1 I[Group = B] + \beta_2 I[Group = C] + \beta_3 I[Group = D] + \beta_4 I[Tree = Maple]$$

g) (5 pts) What is the biological interpretation of $\widehat{\beta_0}$?

h) (5 pts) What is the biological interpretation of $\widehat{\beta_2}$?

i) (12 pts) Briefly describe how you could use bootstrapping to estimate the standard error of $\widehat{\beta_2}$.

j) (15 pts) Name 3 possible methods for testing whether tree species has a significant effect on survival.

Part III (40 pts)

Futuyma and Harshman report on the dry weight of female pupae of 4 different genotypes reared on Oak Trees. Their data is summarized as follows:

Genetic Group	$\bar{Y}$	N	s^2
A	12.68 mg	10	34.388
B	11.44 mg	10	9.860
C	8.37 mg	19	5.138
D	7.63 mg	11	13.943

k) (5 pts) What is the implied null hypothesis H_0 for a one-way ANOVA based on these data, assuming Genetic Group is treated as a random effect.

l) (5 pts) What is the implied null hypothesis H_0 for a one-way ANOVA based on these data, assuming Genetic Group is treated as a fixed effect.

m) (15 pts) What assumption of ANOVA is violated in the data summarized above, and what entry in the one-way ANOVA table is impacted by the violation?

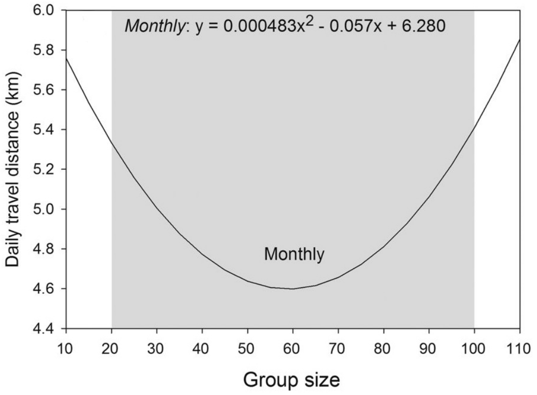
n) (15 pts) Using the data in the table, calculate MS_{within}. (Since you don't have calculators, you can just write out the expression with the appropriate numbers plugged in without simplifying it further.)

11. (100 pts)

Part I (20 pts)

Markham et al. (2015) studied optimal group size and foraging dynamics in wild baboons (*Papio cynocephalus*).

The figure below shows the best fitting line for data on Daily travel distance (km) and Group size.



a) (20 pts) Name two methods to test whether a quadratic model is to be preferred over a linear model. For each method, be sure to describe the test statistic and its distribution under the null hypothesis (be sure to state the null hypothesis being tested).

Part II (40 pts)

Below are the parameter estimates for a multiple regression model (covariates: group size, group size squared, cumulative rainfall, average maximum temperature) for the average daily distance travelled by a baboon troop in each month.

	Estimate	SE	t-statistic	p-value
Intercept	0.125		5.209	0.022
Group size	0.002		3.045	0.081
Group size^2	-1.61×10^{-5}		3.890	0.049
Rainfall	6.33×10^{-5}		2.398	0.122
Ave. Max temp.	0.001		1.572	0.210

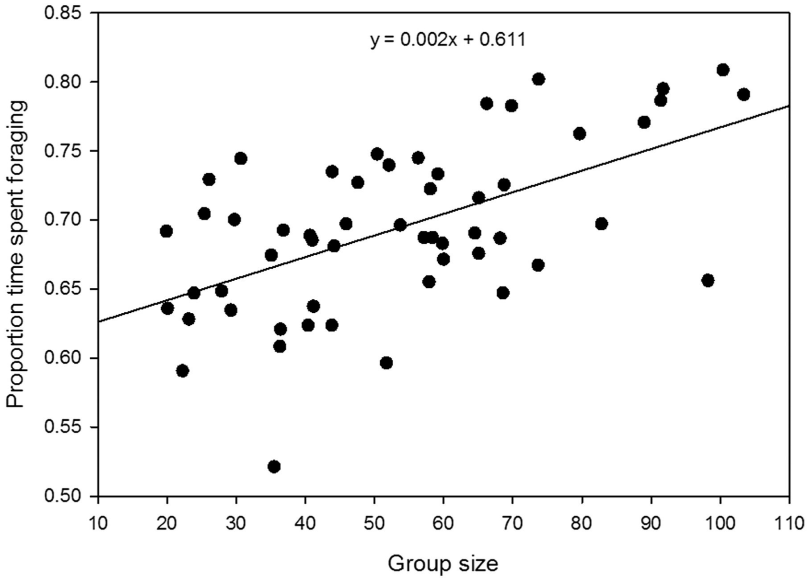
a) (15 pts) The original paper does not provide standard error estimates for these parameters. How could you calculate them based on the information provided?

b) (15 pts) It is reasonable to expect that Rainfall and Average Maximum Temperature are correlated with one another. What effect might this have on the parameter estimates? (Note: If Group size is standardized so that the mean is equal to zero, Group size and Group size^2 are not correlated.)

c) (10 pts) Describe one solution to the problems that might be introduced by correlations between Rainfall and Average Maximum Temperature.

Part III (40 pts)

The questions in Part III relate to the plot below, which represents the proportion of time spent foraging for groups of varying sizes.



a) (5 pts) In the figure above, draw a circle around the point that has the largest 'leverage'.

b) (5 pts) In the figure above, draw a square around the point that has the largest 'influence'.

c) (15 pts) The authors note that "To approach normality, we used an arcsin-square root transform for the proportion of time spent foraging". What is therefore the implied statistical model indicated by this figure. Write the model equation in full.

d) (15 pts) What might be a better statistical model? Write the preferred model equation in full. Why would this be preferred over the model used in the original analysis?