Worksheet 9 - Analysis of frequencies

Frequency Analysis

- Quinn & Keough (2002) - Chpt 13-14

Question 1 - Goodness of fit test

A fictitious plant ecologist sampled 90 shrubs of a dioecious plant in a forest, and each plant was classified as being either male or female. The ecologist was interested in the sex ratio and whether it differed from 50:50. The observed counts and the predicted (expected) counts based on a theoretical 50:50 sex ratio follow.

Format of fictitious plant sex ratios - note, not a file

Expected and Observed data (50:50 sex ratio).

	Female	Male	Total
Observed	40	50	90
Expected	45	45	90

Note, it is not necessary to open or create a data file for this question.

Q1-1. First, what is the **appropriate test** to examine the sex ratio of these plants?

Q1-2. What null hypothesis is being tested by this test?

Q1-3. What are the degrees of freedom are associated with this data for this test?

Q1-4. **Perform a Goodness-of-fit test** to test the null hypothesis that these data came from a population with a 50:50 sex ratio (hint). Identify the following:

a. X^2 statistic

b. df

c. P value

Q1-5. What are your conclusions (statistical and biological)?

Lets now extend this fictitious endeavor. Recent studies on a related species of shrub have suggested a 30:70 female:male sex ratio. Knowing that our plant ecologist had similar research interests, the authors contacted her to inquire whether her data contradicted their findings.

Q1-6. Using the same observed data, **test the null hypothesis** that these data came from a population with a 30:70 sex ratio (hint). From a 30:70 female:male sex ratio, what are the expected frequency counts of females and males from 90 individuals and what is the X^2 statistic?.

a. Expected number of females

b. Expected number of males

c. X^2 statistic

Q1-7. Do the plant ecologist's data dispute the findings of the other studies? (y or n)

Question 2 - Contingency tables

Here is a modified example from Quinn and Keough (2002). Following fire, French and Westoby (1996) cross-classified plant species by two variables: whether they regenerated by seed only or vegetatively and whether they were dispersed by ant or vertebrate vector. The two variables could not be distinguished as response or predictor since regeneration mechanisms could just as conceivably affect dispersal mode as vice versa.

Format of french.csv data files

REGEN	DISP	COUNT
seed	ant	25
seed	vert	6
veg	ant	36
veg	vert	21

REGEN Categorical listing of the plants regeneration mode.

DISP Categorical listing of the plants dispersal mode.

COUNT The observed number of individuals in each

category.

REGENERATION MODE	DISPERSAL MODE	
	ANT	Vertebrate
SeedOnly	25	6
Vegetative	61	21

Open the french data file. HINT.

Q2-1. What null hypothesis is being tested by this test?

Q2-2. Generate a **cross table** out of the dataset in preparation for frequency analysis (HINT).

Q2-3. **Fit** a **2 x 2 (two way) contingency table** (HINT), and explore the main assumption of the test by examining the expected frequencies (HINT).

Q2-4. If the assumption is OK, test this null hypothesis and identify the following.

a. X^2 statistic

b. df

c. P value

Q2-5. What are your conclusions (statistical and biological)?

Question 3 - Contingency tables

Arrington et al. (2002) examined the frequency with which African, Neotropical and North American fishes have empty stomachs and found that the mean percentage of empty stomachs was around 16.2%. As part of the investigation they were interested in whether the frequency of empty stomachs was related to dietary items. The data were separated into four major trophic classifications (detritivores, omnivores, invertivores, and piscivores) and whether the fish species had greater or less than 16.2% of individuals with empty stomachs. The number of fish species in each category combination was calculated and a subset of that (just the diurnal fish) is provided.

Format of arrington.csv data file

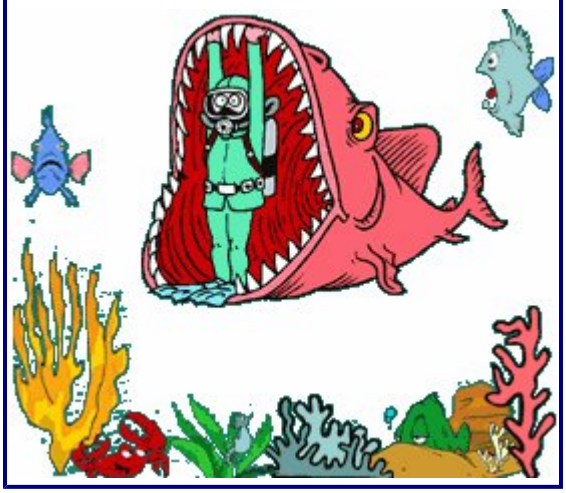
	% Stomachs empty

STOMACH	TROPHIC
< 16.2	DET
..	..
< 16.2	OMN
..	..
< 16.2	PISC
..	..
< 16.2	INV
..	..

STOMACH Categorical listing of the proportion of individuals in the species with empty stomachs (< 16.2% or > 16.2%).

TROPHIC Categorical listing of the trophic classification (DET = detritovore, OMN = omnivore, INV = invertivore, PISC = piscivore).

Trophic classification	< 16.2	> 16.2
DET	18	4
OMN	45	8
INV	58	15
PISC	16	34



Open the arrington data file (HINT).

Note the format of the data file. Rather than including a compilation of the observed counts, this data file lists the categories for each individual. This example will demonstrate how to analyse two-way contingency tables from such data files. Each row of the data set represents a separate species of fish that is then cross categorised according to whether the proportion of individuals of that species with empty stomachs was higher or lower than the overal average (16.2%) and to what trophic group they belonged.

Q3-1. Generate a cross table out of the raw data file in preparation for the contingency table (HINT).

Q3-2. Fit the model (HINT), test the assumptions (HINT) and, using a **two-way contingency table**, test the null hypothesis that the percentage of empty stomachs was independent of trophic classification (HINT). What would you conclude form the analysis?
Write the results out as though you were writing a research paper/thesis. For example (select the phrase that applies and fill in gaps with your results):
The percentage of empty stomachs was (choose the correct option)
(choose correct optio ☐ trophic classification. (X^2 = ☐, df = ☐, P = ☐).

Q3-3. Generate the residuals (HINT) associated with the above contingency test and complete the following table of standardized residuals.

	< 16.2%	> 16.2%
DET	☐	☐
OMN	☐	☐
INV	☐	☐
PISC	☐	☐

Q3-4. What further conclusions would you draw from the standardized residuals?

Question 4 - Contingency tables

Here is an example (13.5) from Fowler, Cohen and Parvis (1998). A field biologist collected leaf litter from a 1 m^2 quadrats randomly located on the ground at night in two locations - one was on clay soil the other on chalk soil. The number of woodlice of two different species (*Oniscus* and *Armadilidium*) were collected and it is assumed that all woodlice undertake their nocturnal activities independently. The number of woodlice are in the following contingency table.

Format of Woodlice data set

SOIL TYPE	WOODLICE SPECIES	
	Oniscus	Armadilidium
Clay	14	6
Chalk	22	46

Open the woodlice data file. HINT.

Q4-1. What null hypothesis is being tested by this test?

Q4-2. Generate a **cross table** out of the dataset in preparation for frequency analysis (HINT).

Q4-3. **Fit** a **2 x 2 (two way) contingency table** (HINT), and explore the main assumption of the test by examining the expected frequencies (HINT).

Q4-4. If the assumption is OK, test this null hypothesis (HINT) and identify the following.

a. X^2 statistic ☐

b. df ☐

c. P value ☐

Q4-5. Generate the residuals (HINT) associated with the above contingency test and complete the following table of standardized residuals.

	oniscus	armadilidium
CLAY		
CHALK		

Q4-6. What are your conclusions (statistical and biological)?

Question 5 - Logistic regression

Polis *et al.* (1998) were intested in modelling the presence/absence of lizards (*Uta sp.*) against the perimeter to area ratio of 19 islands in the Gulf of California.

Format of polis.csv data file

ISLAND	RATIO	PA
Bota	15.41	1
Cabeza	5.63	1
Cerraja	25.92	1
Coronadito	15.17	0
..	..	..

ISLAND Categorical listing of the name of the 19 islands used - variable not used in analysis.

RATIO Ratio of perimeter to area of the island.

PA Presence (1) or absence (0) of *Uta* lizards on island.

Open the polis data file (HINT).

Q5-1. What is the null hypothesis of interest?

Q5-2. Test this null hypothesis by **fitting the general linear model with a binomial error distribution (logit linkage)** (HINT). Identify and interpret the following (HINT);

a. sample constant (β_0)

b. sample slope (β_1)

c. Wald statistic (z value) for main H_0 ☐

d. P-value for main H_0 ☐

e. r^2 value (HINT) ☐

Q5-3. Another way ot test the fit of the model, and thus test the H_0 that $\&beta_1 = 0$, is to compare the fit of the full model to the reduce model via ANOVA. Perform this ANOVA (HINT) and identify the following

a. G^2 statistic ☐

b. df ☐

c. P value ☐

Q5-4. Construct a scatterplot of the presence/absence of *Uta* lizards against perimeter to area ratio for the 19 islands (HINT). Add to this plot, the predicted probability of occurances from the logistic regression. (HINT)

Q5-5. Calculate the LD50 (in this case, the perimeter to area ratio with a predicted probability of 0.5) from the fitted model (HINT). Islands above this ratio are not predicted to contain lizards and islands above this ratio are expected to have lizards.

Q5-6. What are your conclusions (statistical and biological)?

Question 6 - Log-linear modelling

Roberts (1993) was intested in examining the interaction between the presence of dead coolibah trees and the position of quadrats along a transect.

Format of roberts.csv data file

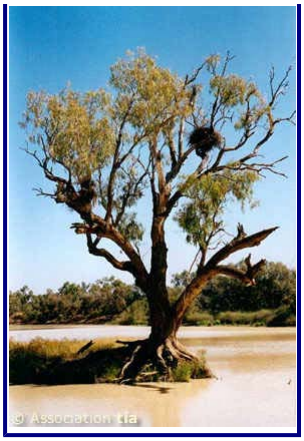
COOLIBAH	POSITION	COUNT
WITH	BOTTOM	15
WITH	MIDDLE	4
WITH	TOP	0
WITHOUT	BOTTOM	13
WITHOUT	MIDDLE	8

WITHOUT	TOP	17

COOLIBAH Categorical listing whether or not dead coolibahs are present (WITH) or absent (WITHOUT) from the quadrat

POSITION Position of the quadrat along a transect

PA Number of quadrats in each classification



Open the roberts data file (HINT).

Q6-1. What is the null hypothesis of interest?

Q6-2. Fit a log-linear model to examine the interaction between presence of dead coolibah trees and position along transect by first fitting a reduced model (one without the interaction, HINT), then fitting the full model (one with the interaction, HINT) and finally comparing the reduced model to the full model (HINT). Alternatively, we can just generate an ANOVA (deviance) table from the full model and ignore the non-interaction terms (HINT). Identify the following:

a. G^2

b. df

c. P value

Question 7 - Log-linear modelling

Roberts (1993) was intested in examining the interaction between the presence of dead coolibah trees and the position of quadrats along a transect.

Format of sinclair.csv data file

SEX	MARROW	DEATH	COUNT
FEMALE	SWF	PRED	26
MALE	SWF	PRED	14
FEMALE	OG	PRED	32
MALE	OG	PRED	43
FEMALE	TG	PRED	8
MALE	TG	PRED	10
FEMALE	SWF	NPRED	6

MALE	SWF	NPRED	7
FEMALE	OG	NPRED	26
MALE	OG	NPRED	12
FEMALE	TG	NPRED	16
MALE	TG	NPRED	26

SEX	Categorical listing sex of the wildebeast carcasses
MARROW	Categorical listing of the bone marrow type (SWF: solid white fatty, OG: opaque gelatinous, TG: translucent gelatinous).
DEATH	Categorical listing of the cause of death (predation or non-predation)
COUNT	Number of carcasses encountered in each cross-classification.

Open the sinclair data file (HINT).

Q7-1. What is the null hypothesis of interest?

Q7-2. Log-linear models for three way tables have a greater number of interactions and therefore a greater number of combinations of terms that should be tested. As with ANOVA, it is the higher level interactions (three way) interaction that is of initial interest, followed by the various two way interactions.
Test the null hypothesis that there is no three way interaction (the cause of death is independent of sex and bone marrow type). First fit a reduced model (one that contains all two way interactions as well as individual effects but without the three way interaction, HINT), then fitting the full model (one with the interaction and all other terms, HINT) and finally comparing the reduced model to the full model (HINT).

	G^2	df	P
SEX:MARROW:DEATH			

Q7-3. We would clearly reject the null hypothesis of no three way interaction. As with ANOVA, following a significant interaction it is necessary to slip the data up acording to the levels of one of the factors and explore the patterns further within the multiple subsets.

a. Sinclair and Arcese (1995) might have been interesting in investigating the associations between cause of death and marrow type separately for each sex. Split the sinclair data set up by sex. (HINT)

b. **Perform the log-linear modelling for the associations between cause of death and marrow type separately for each sex.**

	G^2	df	P
Females - MARROW:DEATH			
Males - MARROW:DEATH			

Q7-4. It appears that there is a significant interactions between cause of death and bone marrow type for both females and males. Given that there was a significant interaction between cause of death and sex and bone marrow type, it is likely that the nature of the two way interactions in females is different to the two way interactions in males. To explore this further, we will examine the odds ratios for each pairwise comparison of bone marrow type with respect to predation level (for each sex)!!!!!!.
Calculate the odds ratios for wildebeast killed by prediation for each pair of marrow types separately for males and females.

	Odds ratio	95% CI min	95% CI max
Females			
OG vs TG			
SWF vs TG			
SWF vs OG			
Males			
OG vs TG			
SWF vs TG			
SWF vs OG			

Q7-5. What would your conclusions be?

Question 8 - Power and continguency tables

A marine ecologist was interested in investigating whether hermit crabs on North Stradbroke Island (what a wet ecologist does on holidays I guess!). He intended to score shells according to whether or not they were occupied and whether they what type of gastropod they were from (*Austrocochlea* or *Bembicium*). Shells with living gastropods were to be ignored. Essentially, the NERD wanted to know whether or not hermit crabs occupy shells in the proportions that they are available. A quick count of shells on the rocky shore revealed that approximately 40% of available gastropod shells were occupied and that *Austrocochlea* or *Bembicium* shells were approximately equally available.
The ecologist scratched his sparsely haired scalp, raised one eyebrow and contemplated performing a quick power analysis to determine how many observation would be required to have an 80% chance of detecting a 20% preference for *Austrocochlea* shells.

This task is best broken down into parts.

Q8-1.
First we compile what is know about the availability of shells:

a. Create a variable that contains the proportion of occupied shells and another that contains the proportion of unoccupied shells (HINT, HINT).

b. Now create two variables that contains the proportion of *Austrocochlea* and *Bembicium* shells respectively (HINT, HINT).

c. Next create a table of proportions that reflect the null hypothesis situation - that is, the proportions expected when hermit crabs show no preferences at all. (HINT, HINT). Provide

row (HINT) and column (HINT) titles for this table. Examine this table (HINT)

d. Now create a variable that contains the anticipated deviance from the null hypothesis - the proportion (20%=0.2) by which the preferences of the hermit crabs deviate from the null hypothesis (HINT).

e. Use this deviance to calculate the expected proportions when this alternative hypothesis (H1) is true (HINT, HINT, HINT). Examine this table (HINT)

f. Calculate the effect size (HINT).

g. Calculate appropriate degrees of freedom (HINT).

h. Finally, we can calculate the approximate number of total observations needed to have an 80% chance of detecting the 20% preference (HINT). How many gastropod shells need to be sampled?

Welcome to the end of Worksheet 9!