

# 13

## Categorical data

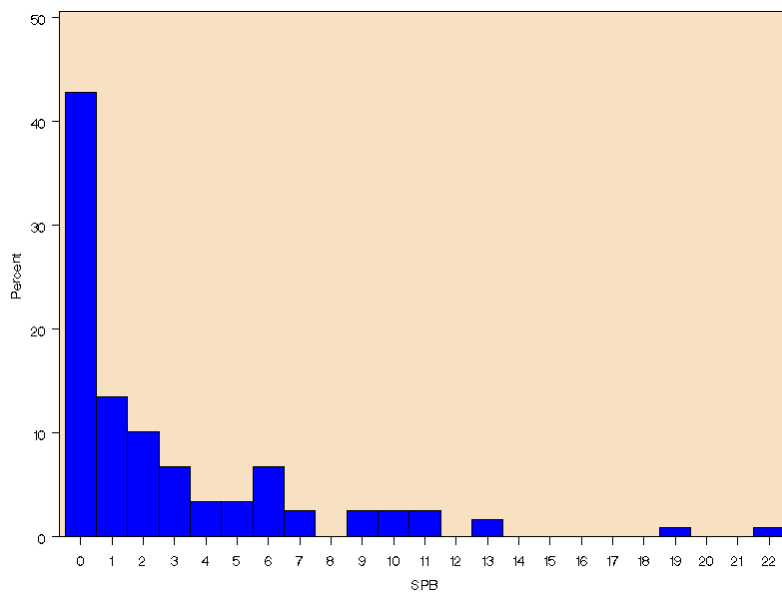
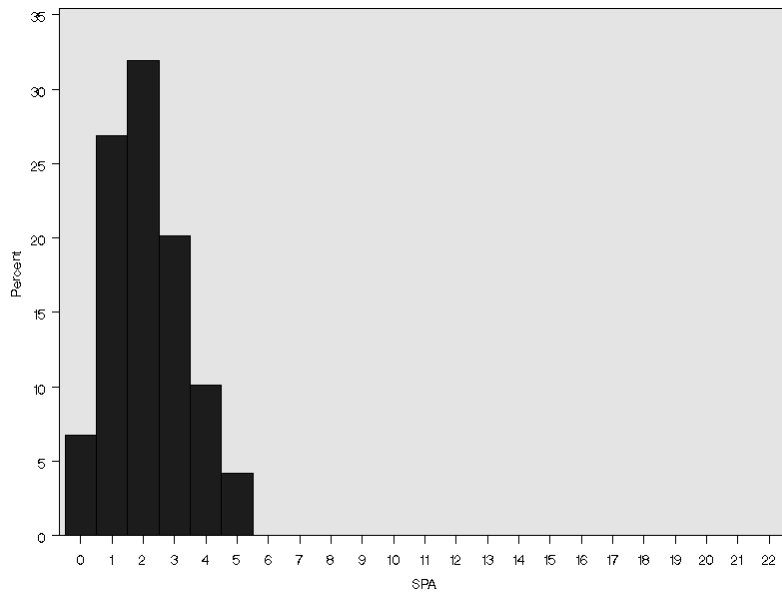
---

### 13.2 The Poisson distribution

The number of parasites per host are plotted here on the same scale. Combining the appearance of these histograms with descriptive data (means and variances), it is thought that Species A may follow a Poisson distribution, and Species B an aggregated distribution. This hypothesis is tested in the main text.

SAS COMMANDS FOR FIGURE 13.2 <b>Two plots of the parasite data</b>	
Commands	<pre>proc capability noprint;   var SPA SPB;   histogram / midpoints = 0 to 22 by 1; run;</pre>
Menu route	Graphs > Histogram SPA SPB → Analysis <input type="button" value="Display"/> Type 0, 22, 1 into :Midpoints of histogram intervals:

SAS OUTPUT FOR FIGURE 13.2 **Two histograms of the parasite data**



SAS produces both of the histograms, but shows only one of them. In the command line route, both histograms are in the same window, and you need only scroll down from one to see the other (or navigate through the results pane on the left-hand side). The menu route is more difficult. It shows you a window with one histogram. To see the other, you find the left hand pane of the Analyst window, which contains a list of all the results from the session. You need to double-click on 'Histogram for SPB' to see the second histogram. This awkwardness for first sight of the histogram must be set against the advantage of being able to go back and see it at any later time in the session by the same route.

We can also obtain descriptive statistics for these distributions by the normal route:

#### SAS COMMANDS FOR BOX 13.1 Descriptive statistics of the parasite data

```
Commands  proc means data=gandh.Chapter13 mean std median q1 q3 min max;
           var SPA SPB;
           run;
```

```
Menu route  Statistics > Descriptive > Summary Statistics...
           SPA SPB → Analysis
```

#### SAS OUTPUT FOR BOX 13.1 Descriptive statistics of the parasite data

The MEANS Procedure

Variable	Label	Mean	Std Dev	Median	Lower Quartile	Upper Quartile
SPA	SPA	2.1260504	1.2320207	2.0000000	1.0000000	3.0000000
SPB	SPB	2.7394958	4.0452152	1.0000000	0	4.0000000

Variable	Label	Minimum	Maximum
SPA	SPA	0	5.0000000
SPB	SPB	0	22.0000000

This information can then be used to perform a dispersion test, as described in the main text. The first step is to calculate the chi-squared value, and the second to obtain a p-value. This is done for species A in the next box.

In the menu route, the result is repeated in all the rows of the last column, which is far from neat, but does the job. The command route shows how to use SAS as a calculator. If you have the log window and the editor window visible at the same time, then you see the result instantly. One way to find out how to perform a function in the command line route is to do it in the menu route, and then look at the log window.

SAS COMMANDS FOR BOX 13.2 <b>A dispersion test for the helminth parasites in bird species A</b>	
Commands	<pre>data;       pA=probchi(118*1.2320207**2/2.1260504,118);       put 'pA=' pA;       run;</pre>
Menu route	<pre>Edit &gt; Mode &gt; Edit Data &gt; Transform &gt; Compute...       Type "pA" into top left box       Type "probchi(118*1.2320207**2/2.1260504,118)" into the main box</pre>

### *Step 1*

We first use  $118 \times 1.2320207^2 / 2.1260504$  to calculate the chi-square statistic, and then use this in the probchi function to find out the area to the left of this value in a chi-square distribution with 118 degrees of freedom. The numbers come from the degrees of freedom (118), the mean (1.232) and the variance (2.126). SAS tells us the area to the left is 0.00806.

### *Step 2*

Is this the final probability for our test? No—the dispersion test is applied here as a two-tailed test, because we are interested in whether the variance is greater or less than the mean. To convert this p-value into the appropriate p-value for our test, it should be multiplied by 2. So the final answer for the dispersion test is a p-value of 0.0162. In this case the variance is significantly less than the mean, and we conclude that the distribution is underdispersed (more uniform than expected).

This may then be repeated for Species B using the same commands, as discussed in more detail in the main text.

### 13.3 The chi-squared test in contingency tables

The data need to be entered in the usual subscripts form (i.e. mating strategy coded as 1, 2 or 3 and trees size as 1, 2 or 3), and the FREQ procedure is used. Unfortunately, SAS will not produce the standardised residuals discussed in the main text, and we will make do with the deviations (=observed–expected) instead. To obtain the standardised residuals, divide this by the square root of the expected in each case.

#### SAS COMMANDS FOR BOX 13.4 Chi-squared analysis of Darwin's hypothesis

```
Commands  proc freq data=gandh.Chapter13;
           tables PLSIZE*STRATEGY / chisq expected deviation;
           weight NOSP;
           run;
```

Menu route Statistics > Table Analysis...

PLSIZE → Row

STRATEGY → Column

NOSP → Cell Counts

Statistics

Chi-Square statistics

SAS OUTPUT FOR BOX 13.4 **Chi-squared analysis of Darwin's hypothesis**

The FREQ Procedure

Table of PLSIZE by STRATEGY

PLSIZE (PLSIZE)	STRATEGY (STRATEGY)			Total
	1,	2,	3,	
Frequency,				
Expected,				
Deviation,				
Percent,				
Row Pct,				
Col Pct,				
	1,	2,	3,	Total
-----	-----	-----	-----	-----
1,	379,	102,	19,	500
,	345.44,	113.61,	40.951,	
,	33.557,	-11.61,	-21.95,	
,	50.07,	13.47,	2.51,	66.05
,	75.80,	20.40,	3.80,	
,	72.47,	59.30,	30.65,	
-----	-----	-----	-----	-----
2,	88,	30,	31,	149
,	102.94,	33.855,	12.203,	
,	-14.94,	-3.855,	18.797,	
,	11.62,	3.96,	4.10,	19.68
,	59.06,	20.13,	20.81,	
,	16.83,	17.44,	50.00,	
-----	-----	-----	-----	-----
3,	56,	40,	12,	108
,	74.616,	24.539,	8.8454,	
,	-18.62,	15.461,	3.1546,	
,	7.40,	5.28,	1.59,	14.27
,	51.85,	37.04,	11.11,	
,	10.71,	23.26,	19.35,	
-----	-----	-----	-----	-----
Total	523	172	62	757
	69.09	22.72	8.19	100.00

## Statistics for Table of PLSIZE by STRATEGY

Statistic	DF	Value	Prob
Chi-Square	4	63.2823	<.0001
Likelihood Ratio Chi-Square	4	56.3200	<.0001
Mantel-Haenszel Chi-Square	1	36.6381	<.0001
Phi Coefficient		0.2891	
Contingency Coefficient		0.2778	
Cramer's V		0.2044	

Sample Size = 757

No new commands are introduced for the rest of this chapter, so we move onto the exercises.

---

## 13.6 Exercises

### Soya beans revisited

---

#### SAS COMMANDS FOR BOX 13.14 **Soya beans revisited**

---

Commands

```
data;
    set gandh.Chapter13;
    SQRTDAM =sqrt(DAMAGE);

proc glm;
    class WDKLR;
    model SQRTDAM = WDKLR / solution;
run;
```

Menu route

Edit > Mode > Edit  
Data > Transform > Compute...  
Type "SQRTDAM" into top left box  
sqrt(DAMAGE) → Main pane

Statistics > Anova > Linear Models...  
SQRTDAM → Dependent  
WDKLR → Class  
 Statistics  
 Parameter estimates

SAS OUTPUT FOR BOX 13.14 **Soya beans revisited**

The GLM Procedure

Dependent Variable: SQRTDAM

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	46.18031892	23.09015946	83.18	<.0001
Error	21	5.82921058	0.27758146		
Corrected Total	23	52.00952949			

R-Square	Coeff Var	Root MSE	SQRTDAM Mean
0.887920	9.890775	0.526860	5.326782

Source	DF	Type I SS	Mean Square	F Value	Pr > F
WDKLR	2	46.18031892	23.09015946	83.18	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WDKLR	2	46.18031892	23.09015946	83.18	<.0001

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	7.038319138 B	0.18627314	37.78	<.0001
WDKLR 1	-3.397516750 B	0.26343000	-12.90	<.0001
WDKLR 2	-1.737095661 B	0.26343000	-6.59	<.0001
WDKLR 3	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

**Fig trees in Costa Rica**

The SAS output for this exercise may be found in the answers to exercises.