

13

Categorical data

13.2 The Poisson distribution

The number of parasites per host are plotted here. 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.

Plots and descriptive data are obtained using the EXAMINE command, or the EXPLORE menu option, as follows:

SPSS COMMANDS FOR BOX 13.1 AND FIGURE 13.2

Two plots of the parasite data, with descriptive statistics

Syntax `examine SPA SPB`
 `/plot boxplot histogram`
 `/compare variables.`

Menu route Analyze > Descriptive Statistics > Explore
 SPA SPB → Dependent List
 Plots
 Boxplots:
 Dependents together
 Descriptive:
 Histogram
 Stem-and-Leaf

SPSS OUTPUT FOR BOX 13.1 **Descriptive statistics of the parasite data****Explore****Total sample****Case Processing Summary**

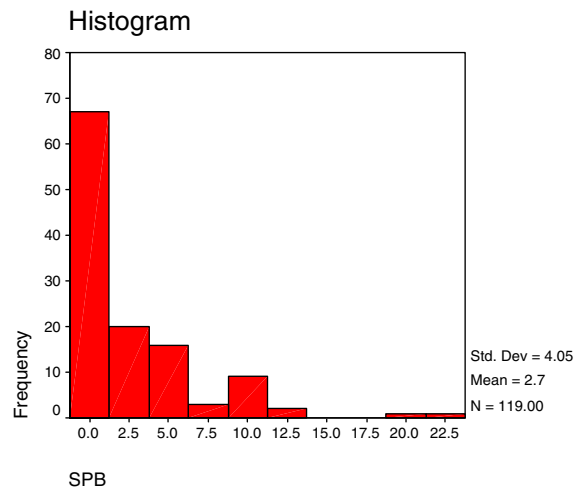
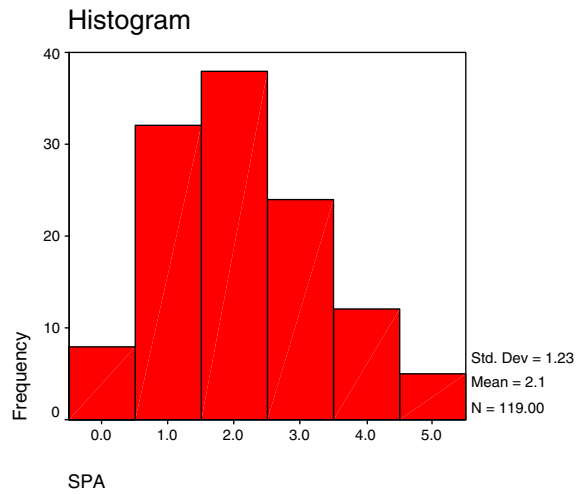
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SPA	119	39.7%	181	60.3%	300	100.0%
SPB	119	39.7%	181	60.3%	300	100.0%

Descriptives

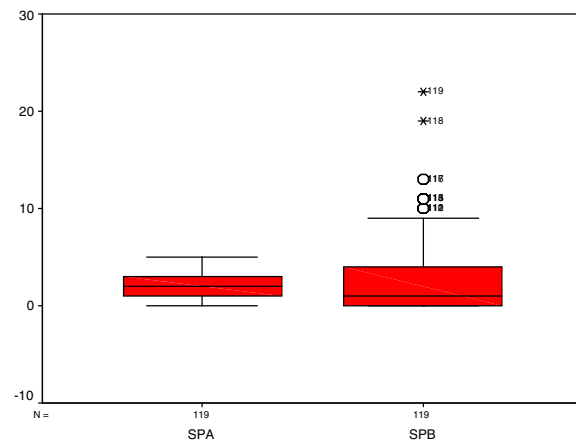
			Statistic	Std. Error
SPA	Mean		2.13	.113
	95% Confidence Interval for Mean	Lower Bound	1.90	
		Upper Bound	2.35	
	5% Trimmed Mean		2.09	
	Median		2.00	
	Variance		1.518	
	Std. Deviation		1.232	
	Minimum		0	
	Maximum		5	
	Range		5	
	Interquartile Range		2.00	
	Skewness		.420	.222
	Kurtosis		-.299	.440
SPB	Mean		2.74	.371
	95% Confidence Interval for Mean	Lower Bound	2.01	
		Upper Bound	3.47	
	5% Trimmed Mean		2.22	
	Median		1.00	
	Variance		16.364	
	Std. Deviation		4.045	
	Minimum		0	
	Maximum		22	
	Range		22	
	Interquartile Range		4.00	
	Skewness		2.162	.222
	Kurtosis		5.635	.440

SPSS OUTPUT FOR FIGURE 13.2 **Boxplots and histograms of the parasite data**

Histograms



Box plots



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 can be done for species A using the following commands. The result goes into each row of a new column called PDISPA (p-value for dispersion test in Species A), which is far from neat but does the job. SPSS automatically reports the answer to two decimal places in the new column. To see the result more accurately, click on one of the cells in the PDISPA column, and the full decimal places will be shown in the box above the variable names. (Remember to select both lines before running if using the syntax route).

SPSS COMMANDS FOR BOX 13.2 A dispersion test for the helminth parasites in bird species A	
Syntax	<code>compute PDISPA=cdf.chisq(118*1.518/2.13,118) .</code> <code>execute.</code>
Menu route	Transform > Compute PDISPA → Target Variable <code>cdf.chisq(118*1.518/2.13,118)</code> → Numeric Expression

Step 1

The chi-squared value is calculated from the variance:mean ratio and degrees of freedom for species A, (118*1.518/2.13). Under the null hypothesis these data are from a Poisson distribution, so K1 should come from the chi-squared distribution with 118 degrees of freedom. The chi-squared value is part of an expression which asks for the cumulative probability of that value, with 118 degrees of freedom (cdf.chisq).

In other words, it asks for the area of the probability distribution of χ_{118}^2 to the left of the chi-squared value. In reply, SPSS gives the probability of 0.007796. This differs slightly from the value given in the main text due to rounding errors. SPSS reported the mean of SPA to only two decimal places.

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 0.015592. 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.

13.3 The chi-squared test in contingency tables

SPSS will not calculate chi-square tests for data in the 'table form'. Instead, it assumes that the data are in subscripts form, with one variable coding for mating strategy and one for plant size. If each species had its own row, then we could use CROSSTABS straightforwardly. However, we have the data in a summary form in which each of the nine combinations of mating strategy (STRATEGY) and plant size (PLSIZE) occupies only one row, and a third variable (NOSP) indicates how many species possess that combination. To use this weighted form of the data, we need to employ SPSS's facility for weighting cases. *It is very important that you remember to turn it off again*, as it affects all subsequent analyses. Turn weighting on as follows:

Syntax	<code>weight by NOSP .</code>
Menu route	Window > SPSS Data Editor Data > Weight Cases... NOSP → Frequency Variable

Now perform the chi-square test with

SPSS COMMANDS FOR BOX 13.4 Chi-squared analysis of Darwin's hypothesis	
Syntax	<code>crosstabs</code> <code> /tables = STRATEGY by PLSIZE</code> <code> /statistic = chisq</code> <code> /cells = count expected sresid .</code>
Menu route	Analyze > Descriptive Statistics > Crosstabs STRATEGY → Row(s) PLSIZE → Column(s) Statistics <input checked="" type="checkbox"/> Chi-square Cells Counts: <input checked="" type="checkbox"/> Expected Residuals: <input checked="" type="checkbox"/> Standardized

192 Categorical data

Now turn the weights off with

Syntax	weight off .
Menu route	Window > SPSS Data Editor Data > Weight Cases... <input checked="" type="radio"/> Do not weight cases

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

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
STRATEGY * PLSIZE	757	100.0%	0	.0%	757	100.0%

STRATEGY * PLSIZE Crosstabulation

			PLSIZE			Total
			1	2	3	
STRATEGY 1	Count	379	88	56	523	
	Expected Count	345.4	102.9	74.6	523.0	
	Std. Residual	1.8	-1.5	-2.2		
2	Count	102	30	40	172	
	Expected Count	113.6	33.9	24.5	172.0	
	Std. Residual	-1.1	-.7	3.1		
3	Count	19	31	12	62	
	Expected Count	41.0	12.2	8.8	62.0	
	Std. Residual	-3.4	5.4	1.1		
Total	Count	500	149	108	757	
	Expected Count	500.0	149.0	108.0	757.0	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	63.282 ^a	4	.000
Likelihood Ratio	56.320	4	.000
Linear-by-Linear Association	36.638	1	.000
N of Valid Cases	757		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.85.

The use of the weights generates a 'Log' report in the output, which gives the following warning:

```
>Warning # 3211
>On at least one case, the value of the weight variable was zero,
negative,
>or missing. Such cases are invisible to statistical procedures
and graphs
>which need positively weighted cases, but remain on the file and
are
>processed by non-statistical facilities such as LIST and SAVE.
```

This is because NOSP is missing for all but the first nine rows. SPSS is concerned, but we need not be, about the fact that data in other variables in the tenth row and beyond is being ignored (because we are using a data file which contains other datasets).

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

13.6 Exercises

Soya beans revisited

This exercise uses familiar commands to produce Box 13.14.

SPSS COMMANDS FOR BOX 13.14 **Soya bean data revisited**

Syntax	<pre>compute SRDAM = sqrt(DAMAGE) . glm SQRTDAM by WDKLR /print parameters /design WDKLR.</pre>
--------	---

Menu route	<pre>Transform > Compute SQRTDAM → Target Variable sqrt(DAMAGE) → Numeric Expression Analyze > General Linear Model > Univariate SQRTDAM → Dependent Variable WDKLR → Fixed Factor(s) Options <input checked="" type="checkbox"/> Parameter estimates</pre>
------------	---

SPSS OUTPUT FOR BOX 13.14 **Soya bean data revisited****General linear model****Between-Subjects Factors**

	N
WDKLR 1	8
2	8
3	8

Tests of Between-Subjects Effects

Dependent Variable: SQRTDAM

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	46.180 ^a	2	23.090	83.183	.000
Intercept	680.990	1	680.990	2453.300	.000
WDKLR	46.180	2	23.090	83.183	.000
Error	5.829	21	.278		
Total	733.000	24			
Corrected Total	52.010	23			

a. R Squared = .888 (Adjusted R Squared = .877)

Parameter Estimates

Dependent Variable: SQRTDAM

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.038	.186	37.785	.000	6.651	7.426
[WDKLR=1]	-3.398	.263	-12.897	.000	-3.945	-2.850
[WDKLR=2]	-1.737	.263	-6.594	.000	-2.285	-1.189
[WDKLR=3]	0 ^a

a. This parameter is set to zero because it is redundant.

Fig trees in Costa Rica

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