

SPSS Statistics Portfolio
Creativity and Value Creation
ACBA5014
Rowan Moses 2203230

Table of Contents

Introduction	2
Exercise 1: Chi-square Test of Independence	2
Exercise 2: Pearson Product-Moment Correlation	5
Exercise 3: Model 1 Linear Regression	7
Exercise 4: Two-Way ANOVA	9
Conclusion	11
References	12

Introduction

The use of statistics is important in ecological and environmental fields, as the natural world is complex and highly variable. Statistical inference is able to show patterns, or lack thereof, in large amounts of data, which would be much harder to spot otherwise. In a field where there is much uncertainty and potential for observation error, this is invaluable (Fowler *et al*, 2003). It could be impossible to count every individual present in a community, for example, but population numbers are still needed for conservation action to take place. Statistical inference allows this information to still be collected (McCarthy, 2015). Statistics can also demonstrate trends in complex systems, such as management of waste for island nations, and show potential correlations between factors such as waste production and economic status, allowing informed management to take place (Eckelman *et al*, 2014). However, with many statistical tests possible to perform, and the diversity of data types collected in ecological fields, errors can be made. Correctly selecting tests and using them to fully explore data is vital (Steel *et al*, 2013).

This portfolio demonstrates the reasoning, method and completion of four statistical tests, as completed using the IBM SPSS Statistics software. Images of inputs and outputs are included, and interpretation of results follows for each exercise.

Exercise 1: Chi-square Test of Independence

The Chi-square test of independence is used to show whether two variables have a significant relationship between them. Two variables that are either categorical or nominal are required. This exercise is examining the species counts of two species of fish over a ten year span, the seabass and the common dab. The Chi-square test of independence is the correct test to determine if there is a statistically significant relationship between these two sets of data as the data is categorical. This means it is based on a “group”, such as hair colour, or for this exercise, species (Hazra and Gogtay, 2016).

See figure one for data input.



Year	Seabass	Dab
1987	14	94
1988	11	39
1998	193	85
1990	146	14
1991	37	75
1992	48	72
1993	23	40
1994	25	58
1995	86	128
1996	46	11

Figure 1: Data input into SPSS.

See figure two for questions answered by this test.

Perform the relevant Chi-square to answer the following:

- 1) Is there any difference in the distribution of seabass across the 10 years?
- 2) Is there any difference in the distribution of common dab across the 10 years?
- 3) Is the distribution of the fish species similar across the 10 years?

Figure 2: Questions to be answered by this test.

For question one of this exercise, the null hypothesis is that there is no significant relationship between year and number of sea bass counted. The alternative hypothesis here is that there is a significant relationship between the year, and the number of sea bass. As shown in figure three, a crosstab has been produced, and the degrees of freedom are eighty one based on the size of the crosstabulation.

Year * Seabass

Crosstab												
Count		Seabass										
		11	14	146	193	23	25	37	46	48	86	Total
Year	1987	0	1	0	0	0	0	0	0	0	0	1
	1988	1	0	0	0	0	0	0	0	0	0	1
	1990	0	0	1	0	0	0	0	0	0	0	1
	1991	0	0	0	0	0	0	1	0	0	0	1
	1992	0	0	0	0	0	0	0	0	1	0	1
	1993	0	0	0	0	1	0	0	0	0	0	1
	1994	0	0	0	0	0	1	0	0	0	0	1
	1995	0	0	0	0	0	0	0	0	0	1	1
	1996	0	0	0	0	0	0	0	1	0	0	1
	1998	0	0	0	1	0	0	0	0	0	0	1
Total		1	1	1	1	1	1	1	1	1	1	10

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	90.000 ^a	81	.231
Likelihood Ratio	46.052	81	.999
N of Valid Cases	10		

a. 100 cells (100.0%) have expected count less than 5. The minimum expected count is .10.

Figure 3: Seabass and year results.

The p-value is a measure of statistical significance, measuring the likelihood of getting the observed result, if the null hypothesis is correct. The lower the p-value, the more likely statistical

significance is present (Newcastle University, n.d). A p-value typically has to be under 0.05 to be considered significant, and here that is not the case. This indicates there is not an association between the year and the number of seabass.

For question two of exercise one, see figure four of the crosstabulation of dab species count per year.

Year * Dab

		Crosstab										
Count		Dab										
		11	128	14	39	40	58	72	75	85	94	Total
Year	1987	0	0	0	0	0	0	0	0	0	1	1
	1988	0	0	0	1	0	0	0	0	0	0	1
	1990	0	0	1	0	0	0	0	0	0	0	1
	1991	0	0	0	0	0	0	0	1	0	0	1
	1992	0	0	0	0	0	0	1	0	0	0	1
	1993	0	0	0	0	1	0	0	0	0	0	1
	1994	0	0	0	0	0	1	0	0	0	0	1
	1995	0	1	0	0	0	0	0	0	0	0	1
	1996	1	0	0	0	0	0	0	0	0	0	1
1998	0	0	0	0	0	0	0	0	0	1	0	1
Total		1	1	1	1	1	1	1	1	1	1	10

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	90.000 ^a	81	.231
Likelihood Ratio	46.052	81	.999
N of Valid Cases	10		

a. 100 cells (100.0%) have expected count less than 5. The minimum expected count is .10.

Figure 4: Dab crosstab and chi-square.

As with the seabass, there is no statistical significance here. The p-value is higher than 0.05. This means the null hypothesis must be accepted, there is no statistical significance between dab species number and the year. The degrees of freedom are also eighty one in this test.

Addressing question three of exercise one, see figure five, the two species chi-square test. The null hypothesis is that the number of seabass per year and the number of dab per year are not associated, and the alternative hypothesis is that they are associated. The degrees of freedom are also eighty one here.

Seabass * Dab Crosstabulation												
Count		Dab										
		11	14	39	40	58	72	75	85	94	128	Total
Seabass	11	0	0	39	0	0	0	0	0	0	0	39
	14	0	0	0	0	0	0	0	0	94	0	94
	23	0	0	0	40	0	0	0	0	0	0	40
	25	0	0	0	0	58	0	0	0	0	0	58
	37	0	0	0	0	0	0	75	0	0	0	75
	46	11	0	0	0	0	0	0	0	0	0	11
	48	0	0	0	0	0	72	0	0	0	0	72
	86	0	0	0	0	0	0	0	0	0	128	128
	146	0	14	0	0	0	0	0	0	0	0	14
	193	0	0	0	0	0	0	0	85	0	0	85
Total		11	14	39	40	58	72	75	85	94	128	616

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	5544.000 ^a	81	<.001
Likelihood Ratio	2619.940	81	<.001
Linear-by-Linear Association	38.071	1	<.001
N of Valid Cases	616		

a. 52 cells (52.0%) have expected count less than 5. The minimum expected count is .20.

Figure 5: Seabass and dab chi-square.

Here it is shown that there is statistical significance, as the p-value is below 0.05. This means the alternative hypothesis is accepted. The Chi-square does not explain causation, but indicates that this area should be investigated further (University of Southampton, 2024). This is useful information for the river management team. The population of seabass and the population of dab are correlated across the ten year period.

Exercise 2: Pearson Product-Moment Correlation

The Pearson Correlation is used to measure the linear relationship between two continuous variables. A continuous variable has values that can be split into infinite smaller values, such as into fractions or decimals, such as height and volume in this exercise. Variables need to be normally distributed, and have a linear relationship (Turney, 2024). The null hypothesis is that there is not a correlation between tree height and volume, and the alternative hypothesis is that there could be a correlation between tree height and volume.

See figure six for input into SPSS.

	 Height	 Volume
1	70	10
2	65	10
3	63	10
4	72	16
5	81	19
6	83	20
7	66	16
8	75	18
9	80	23
10	75	20
11	79	24
12	76	21
13	76	21
14	69	21
15	75	19
16	74	22
17	85	34
18	86	27
19	71	26
20	64	25
21	78	35
22	80	32
23	74	36
24	72	38
25	77	43
26	81	55
27	82	56
28	80	58
29	80	52
30	80	51
31	87	77

Figure 6: Input.

Please see figure seven for results.

Correlations			
		Height	Volume
Height	Pearson Correlation	1	.598**
	Sig. (2-tailed)		<.001
	N	31	31
Volume	Pearson Correlation	.598**	1
	Sig. (2-tailed)	<.001	
	N	31	31

** . Correlation is significant at the 0.01 level (2-tailed).

Figure 7: Results.

The Pearson Correlation here is .598, which is significant, as indicated by the asterisks present (Kent State University, 2021). The P-value displayed is <.001, which as it is under 0.05 also indicates significance. The null hypothesis is therefore rejected, the evidence shows there is positive association between tree height and tree volume. As tree height increases, tree volume increases as well. This allows growth patterns to be more accurately predicted, as well as having implications for carbon storage by the trees in this study (Swenson and Enquist, 2007).

Exercise 3: Model 1 Linear Regression

Linear regression is often used in environmental fields to show correlations and trends in large amounts of data (Boldina and Beninger, 2016). This exercise uses a simple linear regression to analyse the level of correlation between spider mass, the independent variable, and spider web size, the dependent variable. As there is just one independent variable in this data, a simple linear regression is suitable. This test requires there to not be outliers in either variable data set, and they also need to be normally distributed. The null hypothesis is that spider body mass and spider web size are not correlated. The alternative hypothesis is that they are correlated.

See figure eight for SPSS input.

Sample	BodyMass	WebArea
1	23	314
2	28	182
3	31	290
4	29	301
5	44	324
6	33	254
7	26	351
8	41	146
9	7	148
10	20	615
11	7	71
12	7	105
13	5	87
14	1	31
15	1	39
16	38	359
17	53	658
18	65	765
19	27	429
20	9	271
21	9	160
22	46	493
23	36	355
24	6	136

Figure 8: Input.

See figure nine for the regression coefficients calculated from the input.

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.201	63.177		.114	.910
	Mass	8.715	1.426	.791	6.112	<.001
	Sample	5.239	3.598	.188	1.456	.160

a. Dependent Variable: Area

Figure 9: Correlation coefficients.

There is significance in spider body mass ($p = <.001$). See figure ten for the scatterplot with a regression line.

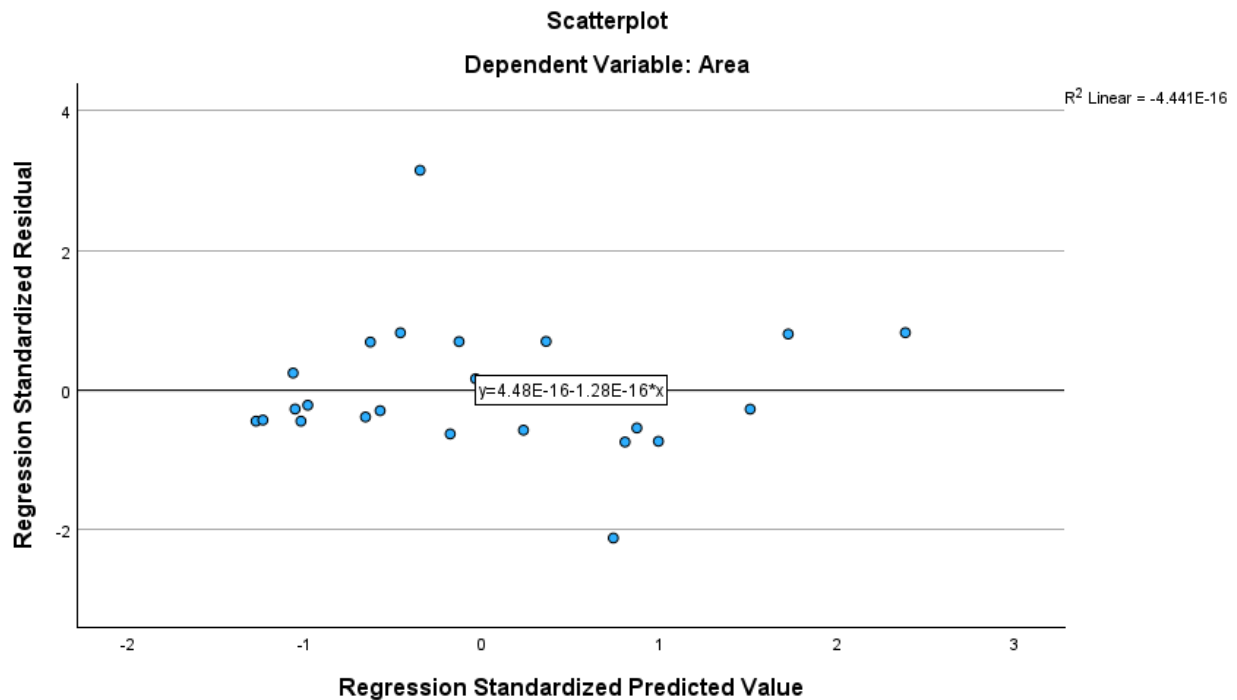


Figure 10: Scatterplot chart.

The degrees of freedom for this regression are two. The null hypothesis is rejected, there is positive correlation between spider body mass and web size. As spider body mass increases, so does web size. This allows surveyors of the *Eustala illicita* species to more effectively count population numbers, as survey methods can be adjusted to identify smaller or larger individuals when webs are seen.

Exercise 4: Two-Way ANOVA

A two-way ANOVA is a test used to compare the differences between groups that have two independent variables, analysing the interactions between the independent variables on the dependent variable (University of Manchester, 2020). In this exercise, researchers wanted to determine factors affecting the crawling speed of ragworms, and the independent variables are substrate and hair length. The dependent variable is crawling speed. If there are three independent variables, a three-way ANOVA is to be used instead. The dependent variable is required to be continuous, and the independent variables must be categorical, independent groups. The dependent variable data must be normally distributed, and there should not be significant outliers (National University Library, 2024). The null hypothesis would be that there is no interaction between substrate and hair length, and crawling speed. The alternative hypothesis is that substrate and hair length does affect crawling speed.

See figure eleven for input into SPSS.

	 Mud_Intact	 Mud_Haircut	 Sand_Intact	 Sand_Haircut
1	4.2	3.9	6.2	6.8
2	3.8	5.2	10.3	2.1
3	7.7	4.8	7.3	3.6
4	5.1	2.3	5.1	2.7
5	4.5	6.4	10.3	2.6
6	4.9	6.4	5.6	7.1
7	8.7	4.1	7.2	3.8
8	4.0	4.3	5.3	6.8
9	8.0	2.4	11.4	4.5

Figure 11: SPSS Data Input.

See figures twelve and thirteen for results tables.

Tests of Within-Subjects Effects						
Measure: distance						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
substrate	Sphericity Assumed	9.000	1	9.000	5.126	.053
	Greenhouse-Geisser	9.000	1.000	9.000	5.126	.053
	Huynh-Feldt	9.000	1.000	9.000	5.126	.053
	Lower-bound	9.000	1.000	9.000	5.126	.053
Error(substrate)	Sphericity Assumed	14.045	8	1.756		
	Greenhouse-Geisser	14.045	8.000	1.756		
	Huynh-Feldt	14.045	8.000	1.756		
	Lower-bound	14.045	8.000	1.756		
hairiness	Sphericity Assumed	44.001	1	44.001	6.800	.031
	Greenhouse-Geisser	44.001	1.000	44.001	6.800	.031
	Huynh-Feldt	44.001	1.000	44.001	6.800	.031
	Lower-bound	44.001	1.000	44.001	6.800	.031
Error(hairiness)	Sphericity Assumed	51.764	8	6.470		
	Greenhouse-Geisser	51.764	8.000	6.470		
	Huynh-Feldt	51.764	8.000	6.470		
	Lower-bound	51.764	8.000	6.470		
substrate * hairiness	Sphericity Assumed	8.604	1	8.604	1.764	.221
	Greenhouse-Geisser	8.604	1.000	8.604	1.764	.221
	Huynh-Feldt	8.604	1.000	8.604	1.764	.221
	Lower-bound	8.604	1.000	8.604	1.764	.221
Error(substrate*hairiness)	Sphericity Assumed	39.021	8	4.878		
	Greenhouse-Geisser	39.021	8.000	4.878		
	Huynh-Feldt	39.021	8.000	4.878		
	Lower-bound	39.021	8.000	4.878		

Figure 12: Results output.

Tests of Within-Subjects Contrasts

Measure: distance

Source	substrate	hairiness	Type III Sum of Squares	df	Mean Square	F	Sig.
substrate	Linear		9.000	1	9.000	5.126	.053
Error(substrate)	Linear		14.045	8	1.756		
hairiness		Linear	44.001	1	44.001	6.800	.031
Error(hairiness)		Linear	51.764	8	6.470		
substrate * hairiness	Linear	Linear	8.604	1	8.604	1.764	.221
Error(substrate*hairiness)	Linear	Linear	39.021	8	4.878		

Tests of Between-Subjects Effects

Measure: distance

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1104.454	1	1104.454	438.046	<.001
Error	20.171	8	2.521		

Figure 13: Results output.

The effects of substrate on crawling speed are not statistically significant, but it is shown that hair length does have an association with crawling speed ($p = .031$). Ragworm speed does not appear to be affected by sand vs mud, but they are less able to move with the hairs on their appendages removed. This implies that their hairs play a role in their mobility. This may mean that ragworms with longer hairs are more able to feed and successfully reproduce. The researchers can carry out further tests based on the results of this test. Two-way ANOVAs are useful in environmental fields where it is common to have many influences at play in a test area (Cunningham, 2000).

Conclusion

Statistical inference can be used in a wide variety of studies in environmental fields, from comparing tree height and volume, to assessing the mobility of ragworms. Using the appropriate test for the data is vital. Using these tests shows significance that would not be detectable to the eye. In this portfolio, the Chi-square test of Independence shows correlation between the population of seabass and dab over a ten year period. The Pearson product-moment correlation shows positive association between tree height and tree volume, and model one linear regression shows positive association between spider body mass and spider web size. Finally, a two-way ANOVA is used to determine that the presence of hairy appendages is significant for ragworm mobility.

References

- Boldina, I. and Beninger, P.G. (2016) 'Strengthening statistical usage in marine ecology: Linear regression', *Journal of experimental marine biology and ecology*, 474, pp. 81-91 Available at: 10.1016/j.jembe.2015.09.010.
- Cunningham, S.A. (2000) 'Effects of Habitat Fragmentation on the Reproductive Ecology of Four Plant Species in Mallee Woodland', *Conservation Biology*, 14(3), pp. 758-768 Available at: 10.1046/j.1523-1739.2000.98609.x.
- Eckelman, M.J., Ashton, W., Arakaki, Y., Hanaki, K., Nagashima, S. and Malone-Lee, L. (2014) 'Island Waste Management Systems', *Journal of Industrial Ecology*, 18(2), pp. 306-317 Available at: 10.1111/jiec.12113.
- Fowler, J., Cohen, L. and Jarvis, P. (2003) *Practical statistics for field biology*. 2. ed., reprint. edn. Chichester [u.a.]: Wiley.
- Hazra, A. and Gogtay, N. (2016) 'Biostatistics Series Module 4: Comparing Groups - Categorical Variables', *Indian journal of dermatology*, 61(4), pp. 385-392 Available at: 10.4103/0019-5154.185700.
- Kent State University (2021) *Pearson Correlation*. Available at: <https://libguides.library.kent.edu/SPSS/PearsonCorr> (Accessed: 13th May 2024).
- McCarthy, M. (2015) 'Approaches to Statistical Inference', in Fox, G., Negrete-Yankelevich, S. and Sosa, V. (eds.) *Ecological Statistics: Contemporary Theory and Application* Oxford: Oxford Academic.
- National University Library (2024) *Statistics Resources Two-Way ANOVA*. Available at: <https://resources.nu.edu/statsresources/Two-WayANOVA> (Accessed: 16th May 2024).
- Newcastle University (n.d) *P-value*. Available at: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/hypothesis-testing/p-value.html> (Accessed: 12th May 2024).
- Steel, E.A., Kennedy, M.C., Cunningham, P.G. and Stanovick, J.S. (2013) 'Applied statistics in ecology: common pitfalls and simple solutions', *Ecosphere*, 4(9), pp. art115 Available at: 10.1890/ES13-00160.1.
- Swenson, N.G. and Enquist, B.J. (2007) 'Ecological and evolutionary determinants of a key plant functional trait: wood density and its community-wide variation across latitude and elevation', *American Journal of Botany*, 94(3), pp. 451-459 Available at: 10.3732/ajb.94.3.451.
- Turney, S. (2024) *Pearson Correlation Coefficient (r) | Guide & Examples*. Available at: <https://www.scribbr.com/statistics/pearson-correlation-coefficient/> (Accessed: 16th May 2024).
- University of Manchester (2020) *Two-Way Analysis of Variance*. Available at: https://www.mub.eps.manchester.ac.uk/helm/wp-content/uploads/sites/81/2020/09/44_2.pdf (Accessed: 16th May 2024).

University of Southampton (2024) *Chi Square*. Available at:
https://www.southampton.ac.uk/passs/full_time_education/bivariate_analysis/chi_square.page
(Accessed: 12th May 2024).