**Chapter 7 Exercises**

**Concepts**

1. If the correlation coefficient is low, is it necessarily correct to assume that the variables are not related? What other factors might be in play?
2. What are the assumptions of Pearson’s Correlation? How might a failure to meet each assumption affect an analysis?
3. What are the benefits of having a large sample size? What are the drawbacks, i.e. what should a researcher keep in mind upon finding that two variables are significantly correlated in a study with a large sample size?
4. Spatial scales, levels of aggregation, and lack of independence can all influence the likelihood of accepting the null hypothesis. Explain each of these.

**Exercises**

1. Use the following data to answer the questions:

|  |  |  |
| --- | --- | --- |
| **Record** | **X** | **Y** |
| ***1*** | 164 | 782 |
| ***2*** | 89 | 401 |
| ***3*** | 98 | 456 |
| ***4*** | 140 | 759 |
| ***5*** | 150 | 895 |

* 1. Find the correlation coefficient for the sample data.
	2. Test the null hypothesis that ρ = 0.
	3. Find Spearman’s rank correlation coefficient for the data, and test whether it is significantly different from zero.
1. Plot the following points and calculate the correlation coefficient *r* for them. State whether you believe a correlation exists between the variables, and why the correlation coefficient reflects or fails to reflect the relationship.

|  |  |  |
| --- | --- | --- |
| **Record** | **X** | **Y** |
| ***1*** | 58 | 20378 |
| ***2*** | 19 | 2 |
| ***3*** | 53 | 3752 |
| ***4*** | 11 | 0.1 |
| ***5*** | 27 | 14 |
| ***6*** | 70 | 99575 |
| ***7*** | 89 | 1248286 |
| ***8*** | 42 | 135 |

1. Repeat the analysis described in Exercise (2) with the following data.

|  |  |  |
| --- | --- | --- |
| **Record** | **X** | **Y** |
| ***1*** | 62 | 329 |
| ***2*** | 63 | 314 |
| ***3*** | 51 | 297 |
| ***4*** | 83 | 443 |
| ***5*** | 63 | 323 |
| ***6*** | 65 | 204 |
| ***7*** | 72 | 388 |
| ***8*** | 56 | 314 |
| ***9*** | 45 | 276 |
| ***10*** | 30 | 498 |

1. Given the following sample sizes, find the minimum value the correlation coefficient can assume and still be significant:
	1. 75
	2. 500
	3. 1000
2. Give a Pearson Correlation of *r* = 0.50, conduct a two-tailed significance test at the α = 0.05 level for the statistic when the:
	1. Original sample size was *n* = 15
	2. Original sample size was *n* = 20
	3. Would your interpretation of the correlation change between a and b?
3. A researcher wants to study the factors that influence economic growth in different counties. A sample size of 850 counties yields a correlation coefficient of 0.112 between the variables for the number of comic book stores and the average income in counties.
	1. Should the researcher accept or reject the null hypothesis that the true correlation is zero?
	2. If the null hypothesis is rejected, should the researcher conclude that a causal relationship exists between the two variables?
4. Given the following data, find Spearman’s rank correlation coefficient, and test whether the true correlation is equal to zero.

|  |  |  |
| --- | --- | --- |
| **Record** | **X** | **Y** |
| ***1*** | 10.4 | 16.7 |
| ***2*** | 3.4 | 5.9 |
| ***3*** | 7.5 | 11.4 |
| ***4*** | 12.6 | 21.4 |
| ***5*** | 8.6 | 14.2 |
| ***6*** | 10.7 | 18.7 |
| ***7*** | 11.5 | 16.4 |
| ***8*** | 11.1 | 11.8 |

1. Use the following data to answer the questions:

|  |  |  |
| --- | --- | --- |
| **Record** | **X** | **Y** |
| ***1*** | 57 | 730 |
| ***2*** | 54 | 568 |
| ***3*** | 48 | 476 |
| ***4*** | 45 | 198 |
| ***5*** | 55 | 445 |
| ***6*** | 51 | 285 |
| ***7*** | 52 | 310 |
| ***8*** | 49 | 494 |
| ***9*** | 50 | 261 |
| ***10*** | 56 | 546 |
| ***11*** | 53 | 420 |

* 1. Find the (Pearson’s) correlation coefficient for the variables. Give the significance.
	2. Find Spearman’s rank correlation coefficient for the variables. Give the significance.
1. Use the following data to answer the questions:

|  |  |  |
| --- | --- | --- |
| **Record** | **X** | **Y** |
| ***1*** | 66 | 29.8 |
| ***2*** | 73 | 30.5 |
| ***3*** | 72 | 29.7 |
| ***4*** | 63 | 31.8 |
| ***5*** | 52 | 28.9 |
| ***6*** | 39 | 31.4 |
| ***7*** | 33 | 33.6 |
| ***8*** | 30 | 32.8 |
| ***9*** | 33 | 33.8 |
| ***10*** | 37 | 33.4 |

* 1. Find the (Pearson’s) correlation coefficient for the variables. Give the significance.
	2. Find Spearman’s rank correlation coefficient for the variables. Give the significance.
	3. Plot the data and examine if the value appear to meet the assumptions.
1. Using Excel or SPSS and the SPSS Housing dataset, analyse the correlation between the date a property was constructed, its price, its floor area, and the unemployment rate. Which of these relationships are significant? What does this suggest about how the age of the property and the price are related? The age of the property and its floor area?
2. Using Excel or SPSS and the Milwaukee Housing Dataset, analyze the correlation between the age of a property, the lot size, the number of bedrooms, the number of bathrooms, and the sale price. Which of these relationships are significant? What does this dataset suggest about how the age of a property and the price are related? The age and the lot size?
3. Using the Singapore Census Dataset, analyze the correlation between English Speakers, Unemployment Rate, Percent Long Commute, Percent Renter. Which of these relationships are significant?
4. Repeat the analysis in question 12, however use a Spearman Rank correlation. Has the significance of any of the relationships changed? Why might this be the case?

**Chapter 7 Solutions**

* 1. r = .942
	2. t = 4.873, with significance = .017. Thus, we reject the null hypothesis.

The correlation is significantly different from zero.

* 1. ri = .900, which has significance .037. Again, we reject the null hypothesis. The correlation is significantly different from zero.
1. r = .691, with p = .057. The plot is given below, and suggests an exponential relationship rather than a linear one. Thus, a correlation exists, but the assumption of linearity is invalid.



1. r = -.107, with p = .769. The plot below shows the impact of the outliers on the overall graph, so that the correlation coefficient does not reflect the otherwise fairly linear spread.



* 1. 0.231
	2. 0.089
	3. 0.063
	4. t = 2.08, tcrit13 = 2.160. We therefore fail to reject the null hypothesis that the correlation is significantly different from zero.
	5. t = 2.45, tcrit18 = 2.101. We therefore reject the null hypothesis that the correlation is significantly different from zero.
	6. The change in the significance of the statistic suggests that our interpretation may change. However, both samples are relatively small and the Person Correlation value of 0.5 is very close to the minimum value needed to attain significance with sample sizes this small. It would therefore be prudent to exercise caution in interpreting these results, and a good idea to carefully inspect the scatterplots or collection additional data.
	7. Given that the minimum value of *r* required for signifiance in a sample of size 850 is 0.069, and that the *t*-score is 3.282 which exceeds the critical value of 1.960, we can reject the null hypothesis.
	8. The researcher should not assume a causal relationship. While the evidence suggests that the two are correlated, when there are many variables being measured there can be spurious relationships. Assuming that the relationship is not spurious, they may both derive from the same phenomenon, rather than one being dependent on the other.
1. ri = .738, with p = .037
	1. r = .698, with p = .017
	2. ri = .636, with p = .035
	3. r = -0.769, with p = 0.009. There is enough evidence to suggest we reject the null hypothesis. The correlation is significantly different from zero.
	4. ri = -0.742. with p = 0.014. There is enough evidence to suggest we reject the null hypothesis. The correlation is significantly different from zero.
	5. The scatter plot does not show any obvious signs of assumption violations. However, this is a small sample, so some caution should be exercised.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Date Built | Price | Floor Area | Unemp. % |
| Date Built | r | 1 | .307\*\* | -0.088 | -.119\*\* |
|  | Sig. |  | 0 | 0.05 | 0.008 |
| Price | r | .307\*\* | 1 | .684\*\* | -.266\*\* |
|  | Sig. | 0 |  | 0 | 0 |
| Floor Area | r | -0.088 | .684\*\* | 1 | -.199\*\* |
|  | Sig. | 0.05 | 0 |  | 0 |
| Unemployment % | r | -.119\*\* | -.266\*\* | -.199\*\* | 1 |
|  | Sig. | 0.008 | 0 | 0 |  |

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

Significant correlations: Date with Price and Unemployment. Price with all. Floor Area with Price and Unemployment. Unemployment with all.

This suggests that properties built at later dates tend to retail at a higher cost, and bigger properties also retail at higher costs.

1. Correlations:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Age | LotSize | Bedrms | Baths | SalePrice |
| Age | r | 1 | -.400\*\* | .186\*\* | .066\* | .052\* |
|  | Sig. |  | 0 | 0 | 0.012 | 0.049 |
| LotSize | r | -.400\*\* | 1 | -0.005 | .095\*\* | .211\*\* |
|  | Sig. | 0 |  | 0.857 | 0 | 0 |
| Bedrms | r | .186\*\* | -0.005 | 1 | .544\*\* | .280\*\* |
|  | Sig. | 0 | 0.857 |  | 0 | 0 |
| Baths | r | .066\* | .095\*\* | .544\*\* | 1 | .526\*\* |
|  | Sig. | 0.012 | 0 | 0 |  | 0 |
| SalePrice | r | .052\* | .211\*\* | .280\*\* | .526\*\* | 1 |
|  | Sig. | 0.049 | 0 | 0 | 0 |  |

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

\* Correlation is significant at the 0.05 level (2-tailed).

Significant correlations: Age with all other variables. Lot Size with all except Bedrooms. Bedrooms with all except Lot Size. Baths with all other variables. Sale Price with all other variables.

This suggests that older properties tend to retail at a higher cost, and that the older the lot is, the smaller it tends to be.

1. Correlations:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | EngSpk | Unemp | PctLngComm | PctRent |
| EngSpk | r | 1 | -.721\*\* | -.526\*\* | .368\* |
|  | Sig. |  | 0 | .001 | 0.027 |
| Unemp | r | -.721\*\* | 1 | .549\*\* | -.403\* |
|  | Sig. | 0 |  | 0.001 | .015 |
| PctLngComm | r | -.526\*\* | .549\*\* | 1 | -.728\*\* |
|  | Sig. | .001 | 0.001 |  | 0 |
| PctRent | r | .3.68\* | -.403\* | -.728\*\* | 1 |
|  | Sig. | 0.027 | .015 | 0 |  |

All these correlations are significant at the 0.05 level, or lower. The negative correlations between Unemployment Rate and English Speakers and Unemployment Rate and Percent Renters indicate that districts with lower unemployment tend to have more English Speakers and Renters. The positive correlation between English Speakers and Percent Renters suggests that districts with English Speakers have more renters.

1. Spearman Rank Correlations:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | EngSpk | Unemp | PctLngComm | PctRent |
| EngSpk | r | 1 | -.652\*\* | -.424\*\* | .277 |
|  | Sig. |  | 0 | .010 | 0.102 |
| Unemp | r | -.652\*\* | 1 | .467\*\* | -.268 |
|  | Sig. | 0 |  | 0.004 | .114 |
| PctLngComm | r | -.424\*\* | .467\*\* | 1 | -.818\*\* |
|  | Sig. | .010 | 0.004 |  | 0 |
| PctRent | r | .277 | -.268 | -.818\*\* | 1 |
|  |  | 0.102 | .114 | 0 | 0 |

Many of the results and significance levels remained largely the same. However, the significance of the relationship between Percent Renter and English Speaker, and Percent Renter and Unemployment changed with the Spearman statistics. Both relationship are no longer significant at the 0.05 level.

Examining the scatter plots of these variable relationships, we observe wider variation at the high end of the percent renter values. These is a possibility that some of these values may be outliers altering the Pearson statistic.



