Pearson Correlation with Excel

1. Description of r

The Pearson product moment correlation coefficient, r, is a measure of linear relation between two quantitative variables. The values of r range from -1.00 (a perfect negative relationship) to 1.00 (a perfect positive relationship). A value of 0.00 indicates no linear relationship, but a non-linear relation could exist when r = 0.00. One should always inspect a scatterplot of variables to visually determine whether a non-linear, or curvilinear, relation exists, or whether outliers or other odd results exist when calculating Pearson r.

The Pearson r is a descriptive statistic, but one can also perform hypothesis tests with Pearson r. Why conduct hypothesis tests? We know samples have variability and don't always match the population from which they were drawn. It is possible, for example, for a sample to indicate a relation exists between two variables, or for means to differ between two groups, when neither a relation nor a mean difference exists in the population. Hypothesis tests help us decide whether our sample results appear to be due to random sample variation or whether the results appear to be reflective of real population relations or differences.

With Pearson r the null hypothesis of no relation is tested. For example, a correlation of r = .23 may be observed between test grades and hours studied in a sample of 5th grade students. The null states that test grades and hours studied are unrelated. Performing a hypothesis test will help us decide whether the correlation of .23 is due to random variation caused by this sample or whether it represents a relation between test grades and hours studied in the population of 5th grade students. The hypothesis test helps us decide whether the observed correlation is likely random or real.

2. Free Excel for GSU Students

Calculating Pearson r is not difficult, but formulas for Pearson r won't be covered in this course. Instead, we will rely on Excel to calculate Pearson r. If you don't have Excel, note that current Georgia Southern students may freely download Microsoft Office 365. The link is provided in Folio in the Technology Resources section (see Figure 1).

Figure 1: Link to Free MS Office 365

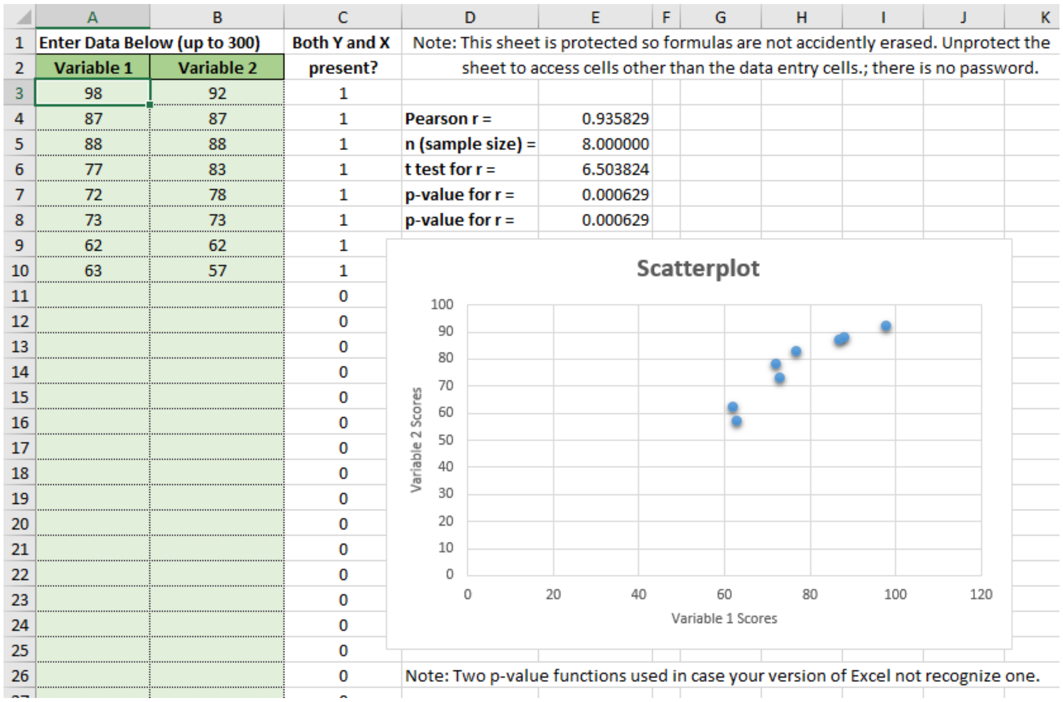
Technology Resources

Download Software
Antivirus, Office365, GS App Store

3. Calculating Pearson r with Excel

The Correlation spreadsheet appears as shown in Figure 2. The spreadsheet has been protected to prevent users from accidentally erasing formulas, but if you wish to edit the sheet, it can be unprotected by clicking on the Unprotect icon under Review. There is no password.

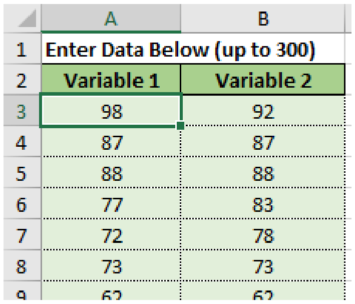
Figure 2: Excel Sheet to Calculate Pearson r



	A	B	C	D	E
1	Enter Data Below (up to 300)		Both Y and X	Note: This sheet is protected so formulas are not accidently erased. Unprotect the	
2	Variable 1	Variable 2	present?	sheet to access cells other than the data entry cells.; there is no password.	
3	98	92	1		
4	87	87	1	Pearson r =	0.935829
5	88	88	1	n (sample size) =	8.000000
6	77	83	1	t test for r =	6.503824
7	72	78	1	p-value for r =	0.000629
8	73	73	1	p-value for r =	0.000629
9	62	62	1		
10	63	57	1		
11			0		
12			0		
13			0		
14			0		
15			0		
16			0		
17			0		
18			0		
19			0		
20			0		
21			0		
22			0		
23			0		
24			0		
25			0		
26			0	Note: Two p-value functions used in case your version of Excel not recognize one.	

One can enter data in the two green columns labeled Variable 1 and Variable 2 (see Figure 3).

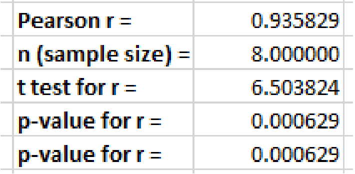
Figure 3: Data Entry in Green Columns



	A	B
1	Enter Data Below (up to 300)	
2	Variable 1	Variable 2
3	98	92
4	87	87
5	88	88
6	77	83
7	72	78
8	73	73
9	62	62

Results are presented to the right of the green columns (see Figure 4).

Figure 4: Pearson r, sample size, t-test, and p-value Results



Pearson r =	0.935829
n (sample size) =	8.000000
t test for r =	6.503824
p-value for r =	0.000629
p-value for r =	0.000629

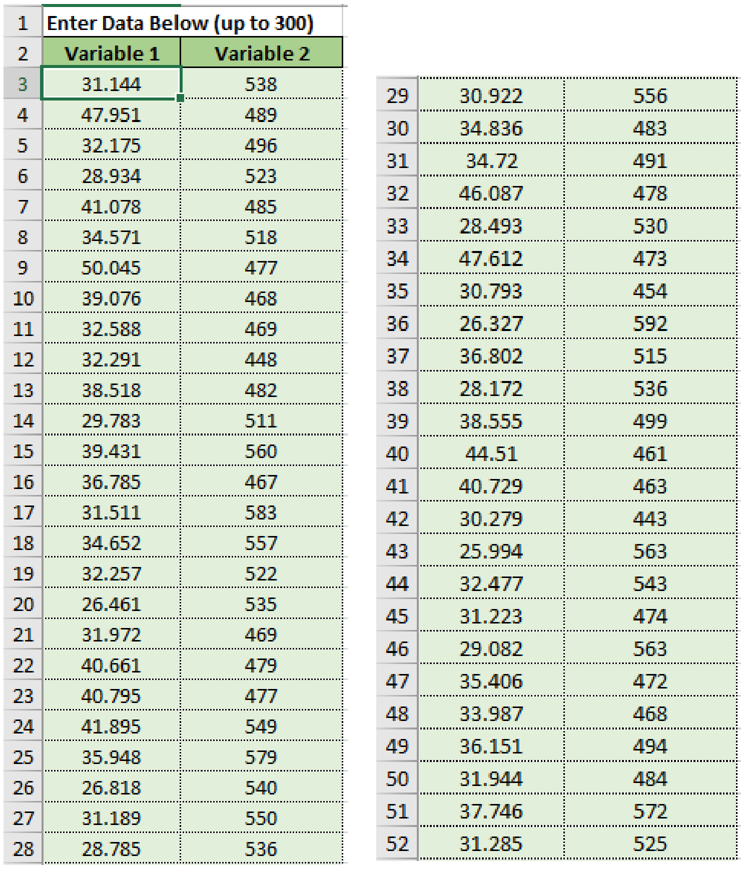
A scatterplot is also provided to allow visualization of the relation between the two variables.

4. Example 1: State Mean SAT Mathematics and Mean Teacher Salary

Is there a relation between the mean SAT mathematics score by state and mean teacher salary by state? Some may hypothesize states that pay their teachers more will have higher SAT scores, so a positive relation would be expected.

Figure 5 shows the data as entered in the Excel spreadsheet. The data are shown in two columns to save space.

Figure 5: Teacher Salary and Math SAT Scores for Each State

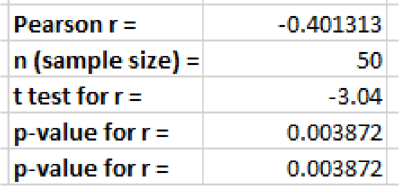


1	Enter Data Below (up to 300)	
2	Variable 1	Variable 2
3	31.144	538
4	47.951	489
5	32.175	496
6	28.934	523
7	41.078	485
8	34.571	518
9	50.045	477
10	39.076	468
11	32.588	469
12	32.291	448
13	38.518	482
14	29.783	511
15	39.431	560
16	36.785	467
17	31.511	583
18	34.652	557
19	32.257	522
20	26.461	535
21	31.972	469
22	40.661	479
23	40.795	477
24	41.895	549
25	35.948	579
26	26.818	540
27	31.189	550
28	28.785	536
29	30.922	556
30	34.836	483
31	34.72	491
32	46.087	478
33	28.493	530
34	47.612	473
35	30.793	454
36	26.327	592
37	36.802	515
38	28.172	536
39	38.555	499
40	44.51	461
41	40.729	463
42	30.279	443
43	25.994	563
44	32.477	543
45	31.223	474
46	29.082	563
47	35.406	472
48	33.987	468
49	36.151	494
50	31.944	484
51	37.746	572
52	31.285	525

Variable 1 is mean teacher salary in thousands of dollars and variable 2 is mean SAT math scores. For the first state listed, teacher salary is 31.144 which means $31,144, and Math SAT is 538.

Results are reported in Figure 6. The correlation is r = -.40. This negative correlation indicates that as teacher salary increases, Math SAT scores tend to decline.

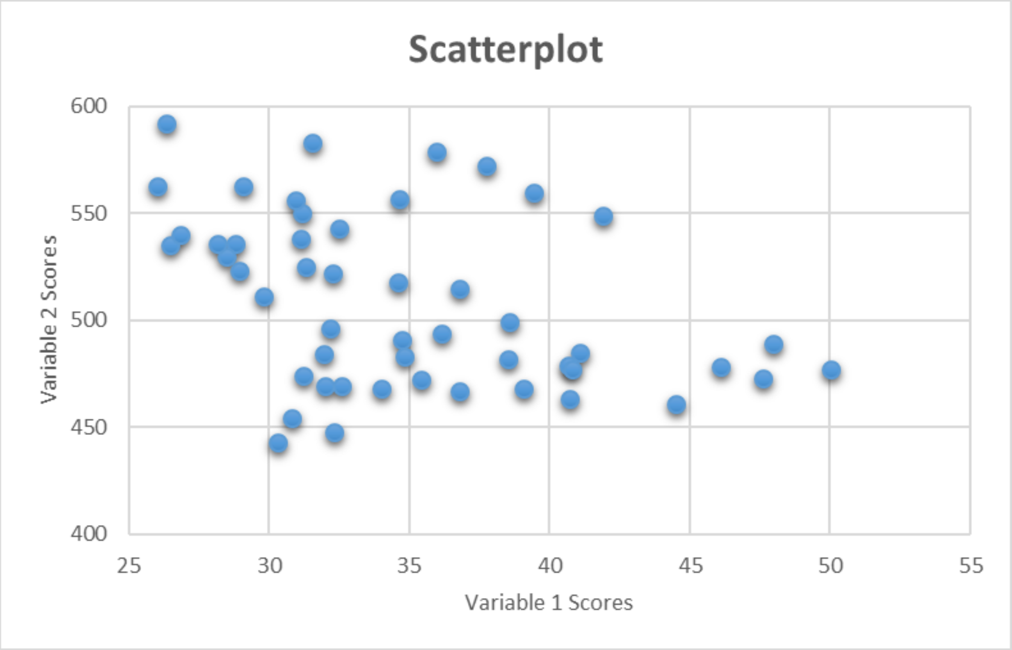
Figure 6: Correlation Results for Teacher Salary and Math SAT



Pearson r =	-0.401313
n (sample size) =	50
t test for r =	-3.04
p-value for r =	0.003872
p-value for r =	0.003872

Figure 7 shows the scatterplot for these data, and a negative trend is discernable.

Figure 7: Scatterplot of Mean Teacher Salary and Mean Math SAT



The null hypothesis for these variables follows:

> Null: There is no relation between state level mean teacher salary and mean math SAT scores.

With hypothesis testing we wish to know whether the sample data provide enough evidence to determine whether a relation is possible in the population. The test for Pearson r = -.40 is performed by a t-test and the value of the calculated t, shown in Figure 6, is t = -3.04.

To help decide whether the data are consistent, or inconsistent, with the stated null hypothesis, we use the p-value which is derived from the t-test. The reported p-value, shown in Figure 6, is 0.003. What does this mean?

> p-value = .003: The probability, through random chance, of obtaining a correlation r this large or larger (in absolute value) if the null hypothesis of no relationship is true in the population.

What does this mean? To calculate this p-value we must assume that the null is true for the population, so we assume there is no relation between math SAT scores and teacher salary. If the null is true, then what are the chances of taking a random sample of 50 scores from a given year and finding a t-ratio of -3.04 or larger in absolute value (i.e., t ≥ 3.04 or t ≤ -3.04); the p-value tells us this chance, or probability, and it is .003. If we were to repeatedly sample from this population, the chance of getting a sample Pearson correlation like we observed, or a more extreme Pearson correlation, is about .003 x 100 = 0.03%. The key here is the assumption that the null hypothesis – no correlation between math SAT scores and teacher salary – is true in the population. We don't know if it is true, but we do know that if it were true, we could expect to see a correlation this large or larger only about 0.03% of the time. So, if the null is true, then the correlation observed (r = -.40) is a rare event.

When collecting sample data, like the data in this example, random variability is expected from sample to sample. When random variation occurs, it may create random relationships among variables. The point of conducting hypothesis tests and examining p-values is to help us decide whether the relationship we observed is likely due to random chance, or due to something real. The fact that we observed a correlation of r = -.40 in a sample does not mean it exists in the population; the hypothesis test helps us decide whether that correlation could be due to random variation or due to a real relationship in the population.

In hypothesis testing small p-values lead researchers to reject the null and conclude that a relationship exists; if the p-value is large, researchers won't reject the null hypothesis and will conclude that no relationship exists. Since this p-value is 0.003, which is very small, it appears that the sample data are inconsistent with the stated null, so our sample data seem to indicate there is a relationship between salary and SAT scores.

Recall the discussion of hypothesis testing errors. Two error probabilities were presented, alpha (α) and beta (β). Alpha is the probability of a Type 1 error (rejecting the null and claiming there is a relation when in fact there is no relation in the population), and beta is the probability of a Type 2 error (failing to reject a false null and claiming there is no relation when in fact there is a relation in the population).

Researchers use a decision rule when deciding to reject or not reject the null hypothesis:

Decision Rule: If p-value ≤ α reject the null; if p-value > α do not reject the null

A common alpha level used for hypothesis testing decisions is .05 which means there is a 5% chance of making a Type 1 error in hypothesis testing.

If alpha = .05, would the null hypothesis of no relation between mean teacher salary and mean math SAT scores be rejected? The p-value reported above for these data was .003, so complete the decision rule:

Decision Rule: If .003 ≤ .05 reject the null; if .003 > .05 do not reject the null

Since .003 is less than .05, the null is rejected and we claim there is a "statistically significant" correlation between mean teacher salary and mean math SAT scores across the states. Interpretation: A correlation of r = -.40 indicates a negative association was found. This suggest that the higher teacher salaries, the lower math SAT scores.

If you are interested in replicating this analysis, the raw data can be downloaded from this link:

http://www.bwgriffin.com/gsu/courses/edur7130/statistics/State-Salary-and-SAT-Scores-1994-1995.xls

Source of data: http://www.stat.ucla.edu/labs/pdflabs/sat.pdf

A key variable was omitted from this analysis – the percent of students in each state who took the SAT. If we consider the percentage of students in each state who took the SAT, the relation between teacher salary and mean SAT scores changes – it is no longer negative. Why might the variable percent of students who take the SAT change the nature of the relation between salary and SAT?

5. Example 2: Doctoral Student Efficacy and Anxiety toward the Dissertation Process

Doctoral students were asked to respond to a questionnaire designed to assess two constructs: anxiety toward the dissertation process and efficacy toward the dissertation process. The items represent an attempt to measure students' anxiety (concern, fear, nervousness) and efficacy (confidence, certainty) about undertaking and completing the process of developing and defending a dissertation. The questionnaire is presented in Display 1. The even-numbered items measure anxiety and the odd-numbered items measure efficacy.

Display 1: Dissertation Process Questionnaire

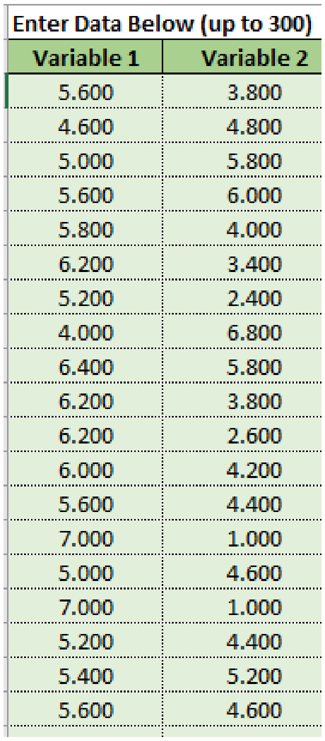
The purpose of the questionnaire is to ascertain doctoral students' thoughts about the dissertation process. Your honest responses will help provide a better understanding of doctoral students' experience with this process. In the context of this questionnaire, dissertation process means the entire process students experience to construct and defend the dissertation. This includes, for example, developing the research idea, developing and defending the prospectus, collecting and analyzing data, writing the dissertation, and defending the dissertation before of a committee.

The following 10 statements refer to the dissertation process that you will soon experience. There are no right or wrong answers, so please answer as accurately as possible. Use the scale below to respond to each statement. If you think the statement is very true of you, circle 7; if the statement is not at all true of you, circle 1. If the statement is more or less true of you, find the number between 1 and 7 that best describes you.

		not at all true of me						very true of me
1.	I believe I will do well on the dissertation.	1	2	3	4	5	6	7
2.	I feel uneasy or uncomfortable with the dissertation process as a whole.	1	2	3	4	5	6	7
3.	I am confident that I can address even the hardest aspects of the dissertation process.	1	2	3	4	5	6	7
4.	Thinking about the upcoming dissertation process makes me feel anxious.	1	2	3	4	5	6	7
5.	The process of writing and defending the dissertation may be difficult or hard, but I think I will be successful anyway.	1	2	3	4	5	6	7
6.	I am worried about how well I will do during the dissertation defense.	1	2	3	4	5	6	7
7.	I know that I have learned the literature and theories that will be necessary to report in the dissertation.	1	2	3	4	5	6	7
8.	I feel my heart beating faster as I start to think about the dissertation.	1	2	3	4	5	6	7
9.	I am sure that I will be able to answer some of the more challenging or difficult questions posed by the dissertation committee.	1	2	3	4	5	6	7
10.	Thinking about the consequences of failing some component of the dissertation process makes me uptight.	1	2	3	4	5	6	7

For each respondent the mean of responses to the five anxiety items was used to form a composite score for anxiety, and the mean for the efficacy items was used to form a composite score for efficacy. The mean responses are reported in Figure 8; Variable 1 represents the efficacy scores and Variable 2 the anxiety scores.

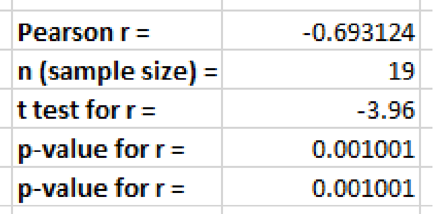
Figure 8: Variable 1 = Efficacy, Variable 2 = Anxiety



Enter Data Below (up to 300)	
Variable 1	Variable 2
5.600	3.800
4.600	4.800
5.000	5.800
5.600	6.000
5.800	4.000
6.200	3.400
5.200	2.400
4.000	6.800
6.400	5.800
6.200	3.800
6.200	2.600
6.000	4.200
5.600	4.400
7.000	1.000
5.000	4.600
7.000	1.000
5.200	4.400
5.400	5.200
5.600	4.600

Prior research has shown that anxiety and efficacy tend to correlate negatively, so a similar correlation was expected for these data. Results are reported in Figure 9.

Figure 9: Results for Dissertation Process Anxiety and Efficacy



Pearson r =	-0.693124
n (sample size) =	19
t test for r =	-3.96
p-value for r =	0.001001
p-value for r =	0.001001

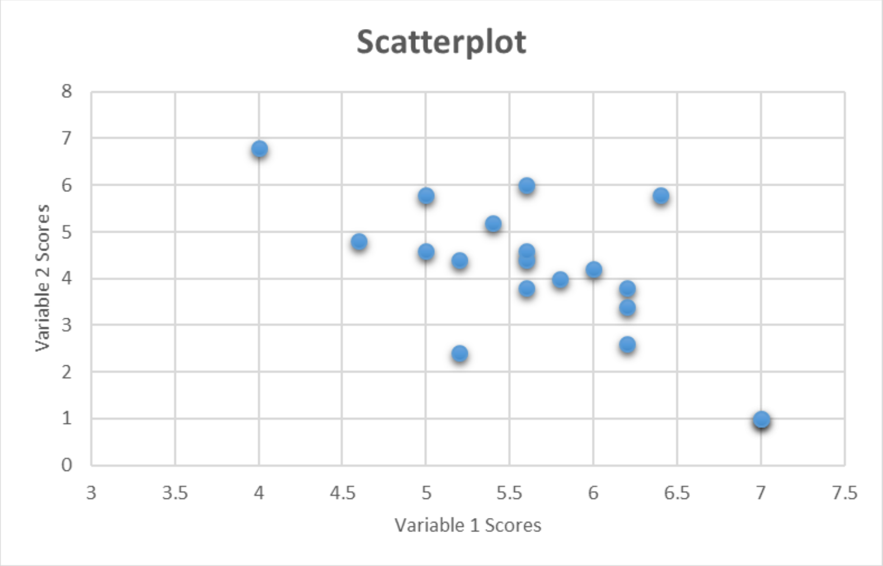
The correlation is r = -.69 with a sample size of 19 respondents. The p-value for this correlation is .001. Since this value is less than alpha of .05, i.e.,

Decision Rule: If .001 ≤ .05 reject the null; if .001 > .05 do not reject the null

we can reject the null hypothesis (null: there is no relation between dissertation process anxiety and efficacy) and conclude, based upon this sample of 19, that there appears to be a negative relation between anxiety and efficacy for the dissertation process.

A scatterplot of dissertation process efficacy and anxiety is presented in Figure 10. The plot shows a negative trend which is consistent with the negative Pearson r value.

Figure 10: Scatterplot of Dissertation Process Efficacy (Variable 1) and Anxiety (Variable 2)



6. Review of Hypothesis Testing with Correlations with Excel

(a) Enter data in green columns
(b) Examine Pearson r value and scatterplot to assess nature of relationship
(c) If p-value is less than alpha (e.g., .05), reject the null and claim relationship was identified; if the p-value is greater than .05 do not reject the null and state that the sample data indicate no relationship was identified
(d) Interpret the results – explain what the r and scatterplot indicate are occurring with the relationship between variable 1 and variable 2