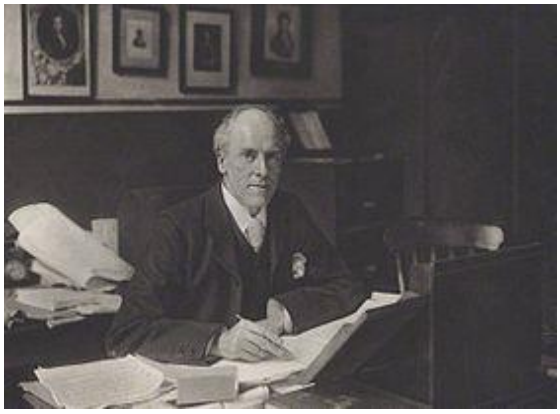
Critical Reasoning 19 - Correlation, Contingency Tables and the χ^2 Test



Karl Pearson (1857-1936) English Biostatistician and Founder of Mathematical Statistics

In this Study Unit we consider hypotheses that suggest an association between two (or more) variables from the same sample. Our task is to describe and quantify such associations, if any, that involve or hint at a statistical relation of dependence between one variable and another. We do this for both ordinal and nominal scale variables. Once again, we have relied on Professors Kruger and Janeke's UNISA study guide to undergraduate Psychology Statistics both as a curriculum guide and model of clear explanation (Kruger and Janeke, 2012)

Measuring the association between variables

Correlation, as we have seen, is a statistical measure that indicates the extent to which two or more variables vary together or the interdependence of variable quantities. Recall that variables are observed attributes of events or objects that are measurable or quantifiable in some way. Usually this involves assigning a numerical value according to some agreed upon scale; other times (for categorical variables) we might simply wish to encode a phenomenon such as "Vaccinated? Yes = 1; No = 0".

One way to visualise the relation between two variables is to create a **scatter plot**, a graphical representation of data points for typically two variables of interest using Cartesian coordinates. By convention the independent variable is plotted on the x-axis and the dependent variable on the y-axis; however as, in the following example, we may sometimes not even know for certain which is which - no matter.

The following example is adapted from a post at the resourceful website *statstutor.ac.uk* entitled "Statistical Analysis 2: Pearson Correlation", available here. Suppose a student dietician is interested in the relation between calcium intake and knowledge about calcium in the diet and the human body. As a ready group of subjects she convinces 20 sports science students to enrol in her study. At the beginning of the study she administers a quick multiple choice questionnaire out of 50 that has been shown to reliably measure a person's knowledge about calcium in the diet and in the human body. She then asks the participants keep a diary of weighed and measured portions of all that they have to eat and drink over a period of a week. Using specialised dietetic software she is able to calculate each person's mean daily intake of calcium in mg per day.

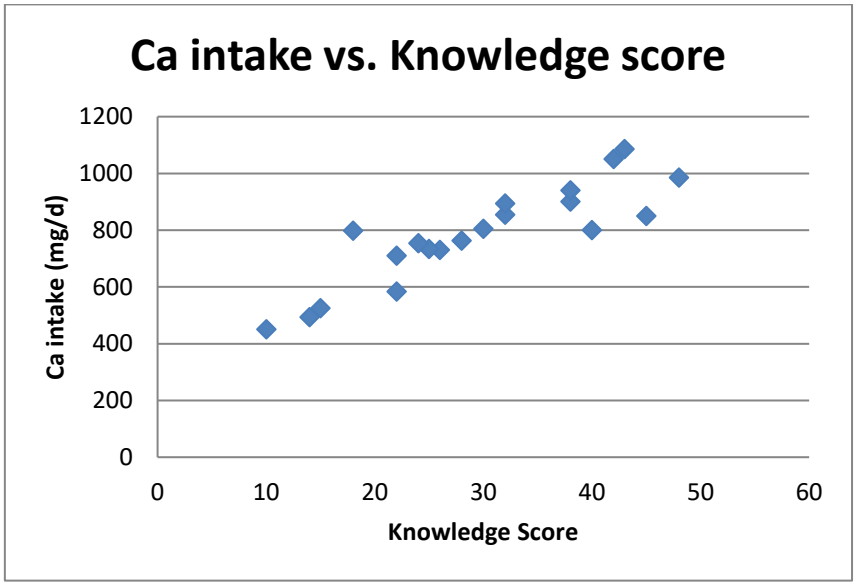
Before the results are even in she has a clear, operationalised **research question** in mind, in other words, a question that her research project sets out to answer: Is there a relation between "knowledge about calcium" as measured by the questionnaire and "mean daily dietary intake of calcium" as calculated from the food diaries? Two hypotheses suggest themselves:

- The **null hypothesis** (H_0): There is no relation between "knowledge about calcium" and "mean daily dietary intake of calcium".
- The **alternative hypothesis** (H_1): There is a relation between "knowledge about calcium" and "mean daily dietary intake of calcium".

At this stage she has no idea as to the direction, if any, of the alternative hypothesis; however as soon as the data are in she decides to tabulate the results and construct a scatter plot. The table below shows the data she collected.

Case No.	Knowledge score (/50)	Ca (mg/d)	Case No.	Knowledge score (/50)	Ca (mg/d)
1	10	450	**11**	38	940
2	42	1050	**12**	25	733
3	38	900	**13**	48	985
4	15	525	**14**	28	763
5	22	710	**15**	22	583
6	32	854	**16**	45	850
7	40	800	**17**	18	798
8	14	493	**18**	24	754
9	26	730	**19**	30	805
10	32	894	**20**	43	1085

From this table of results we can draw up a scatter plot of the data points, either by hand using graph paper and a pencil or with the aid of a spreadsheet program, which is less laborious. We used Microsoft Excel™ for which there are numerous online tutorials for creating charts and graphs for those unfamiliar with the task.



Our student researcher perceives a pattern in the way the data points seem to cluster around a diagonal (not shown) from middle left to top right on the scatter plot. She interprets this as evidence for a liner relation between the two variables: "knowledge about calcium" and "mean daily dietary intake of calcium". To make this manifest she could have actually drawn in a **line of best fit** *i.e.* a

straight line that lies closest to most of the data points her scatter plot. Alternatively, she could have selected the option to overlay one from among the graphic options from the spreadsheet program she was using, However she is keen to quantify the relation, that she perceived – was it a strong or weak relation and if so how probable is her result? What were the chances that they might have been the result of a mere fluke? Such questions can only be guessed at from a line of best fit.

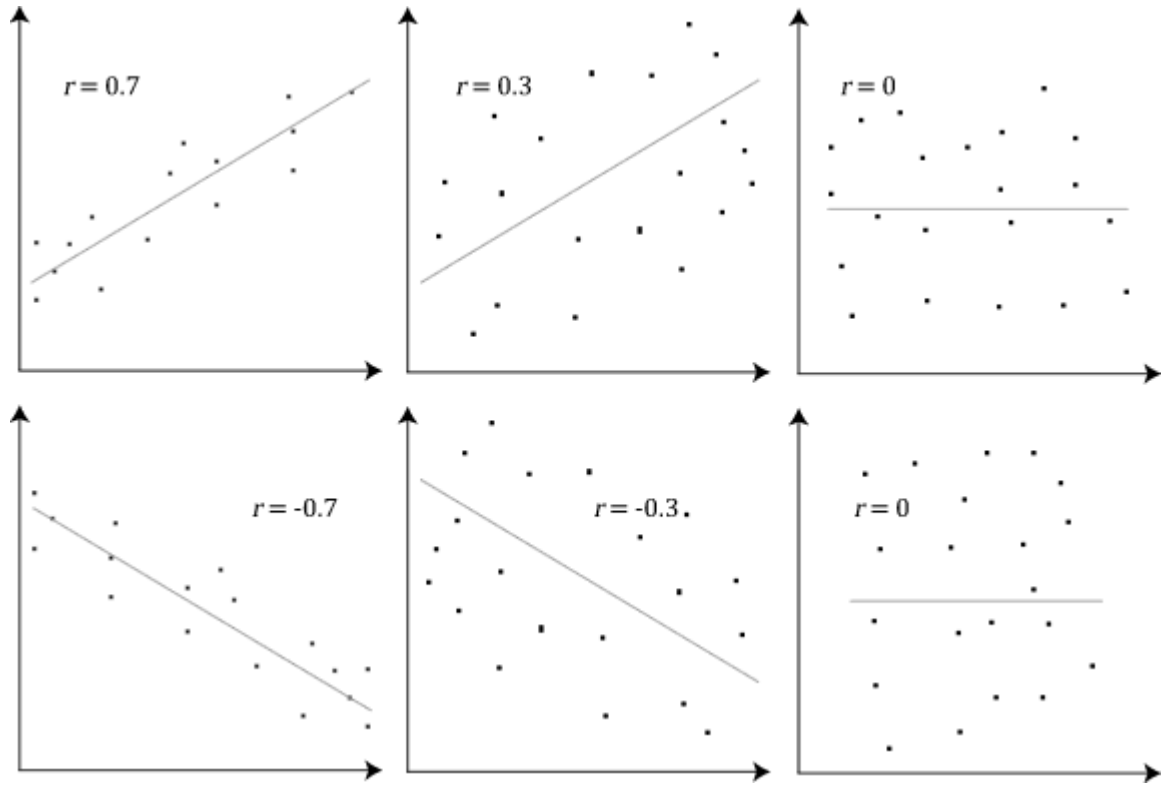
Fortunately **correlation coefficients** which measure the strength of a linear relationship between two variables can be calculated. The one our researcher is interested in is known as the **Pearson product-moment correlation coefficient** or just **Pearson's r**, symbolised simply as 'r'. This measure can take on any continuous value from +1 to -1 inclusively, such that:

r = 1 implies a perfect positive linear relation between variables,

r = 0 implies no liner relation at all, and

r = -1 implies a perfect negative linear relationship.

Informally, a positive relationship indicates that two variables get bigger together or smaller together, whereas a negative relationship indicates that when one variable gets bigger the other gets smaller or *vice versa*. The absence of such a relationship indicates that the variables vary quite independently of each other. The figure below shows six scatter plots including a line of best fit for each as well as the corresponding value of r for each.

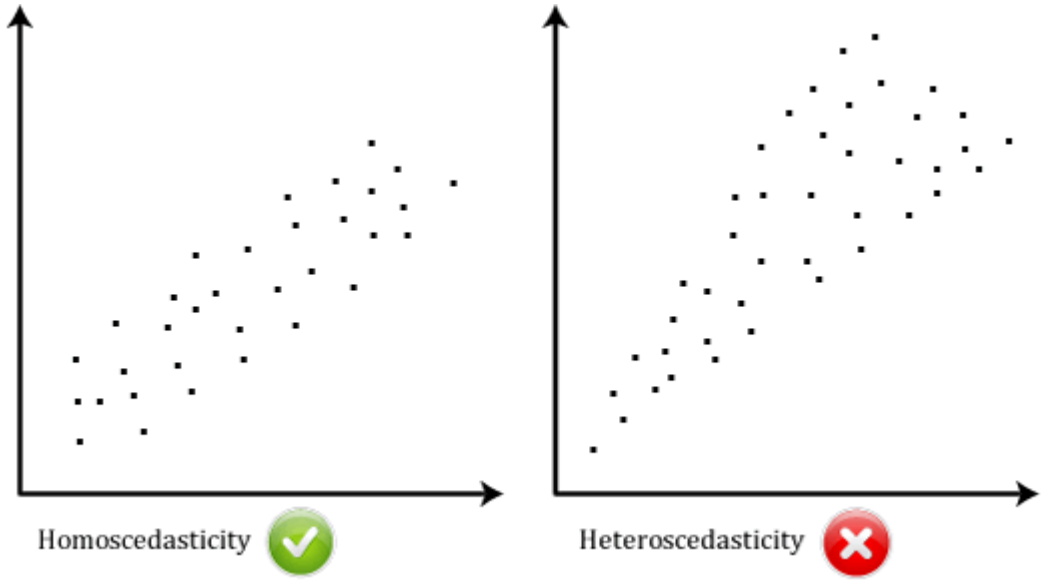


A perfect correlation of r = 1 (not shown) would have all the data points perfectly aligned on a diagonal from bottom left to top right, whereas a perfect correlation of r = -1 (also not shown) would have all the data points perfectly aligned on a diagonal from top left to bottom right. Perfect correlations in the Life Sciences and Humanities are almost never seen. On the other hand, data that

appear to fall on a U shaped or S shaped line are quite common, although The Pearson correlation coefficient tells us (almost) nothing about such nonlinear relationships.

Besides the requirement of linearity there are several other assumptions that should be met before deciding utilise Pearson's correlation:

- Both variables should be continuous interval or ratio measurements. If one is ordinal Spearman's correlation should be considered. (See below)
- Both variables should be approximately normally distributed.
- Each participant or observation should have a pair of variables associated with it. Participants or observations with a missing value may not be included.
- **Outliers** (data points typically more than ± 3.29 standard deviations from the mean) are usually not included as these tend to skew the result too far one way or the other.
- The data should be **homoscedastic**, *i.e.* they should have equal or similar finite variance. Graphically[1], homoscedastic and heteroscedastic distributions can be seen in the following scatter plots: Note how the latter tends to splays out along the line of best fit.



Our student researcher's dataset meets all the above requirements therefore we can see how she would have legitimately calculated Pearson's *r* for the two variable of interest. If we let x stand for one variable and y for the other variable in her sample, then, in general, Pearson's r can be expressed as follows:

$$r = \frac{\mathrm{cov}(x; y)}{\sqrt{\mathrm{var}(x)\mathrm{var}(y)}}$$

Here $\mathrm{cov}(x; y)$ refers to the **covariance** of x and y, *i.e.* the extent to which they vary together; while $\mathrm{var}(x)$ and $\mathrm{var}(y)$ refer, respectively, to the variances of x and y on their own. Recall from Critical

[1] Image courtesy of https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Reasoning 13 that variance is just the square of the standard deviation. Expressed this way, we can see that r is in fact a *ratio* between how two variables vary together and how they vary apart. And because two variables taken together cannot vary more than they can taken separately, their ratio, r can never exceed 1 (or be less than -1). (p. 134)

To manually calculate Pearson's r, we have to flesh out the above formula using some of the sigma notation with which we are already familiar, thus

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

Rather than substituting each pair of values of x and y in our dataset directly into the formula we can use a pocket calculator or spreadsheet to first calculate their respective squares, x^2 and y^2 and their product xy and then sum these so that we can substitutes *these* values into the formula. Importantly, Kruger and Janeke, are at pains to point out to the reader that the product of the sums, $(\sum x)(\sum y)$ should not be confused with the sum of the products, $(\sum xy)$. (p. 135)

We have tabulated the required calculations from our data set for substitution into the formula.

Case No.	*x*	x^2	*y*	y^2	*xy*
1	10	100	450	202500	4500
2	42	1764	1050	1102500	44100
3	38	1444	900	810000	34200
4	15	225	525	275625	7875
5	22	484	710	504100	15620
6	32	1024	854	729316	27328
7	40	1600	800	640000	32000
8	14	196	493	243049	6902
9	26	676	730	532900	18980
10	32	1024	894	799236	28608
11	38	1444	940	883600	35720
12	25	625	733	537289	18325
13	48	2304	985	970225	47280
14	28	784	763	582169	21364
15	22	484	583	339889	12826
16	45	2025	850	722500	38250
17	18	324	798	636804	14364
18	24	576	754	568516	18096
19	30	900	805	648025	24150
20	43	1849	1085	1177225	46655
Sum (Σ)	592	19852	15702	12905468	497143

We can mow substitute the required information and simplify, as follows

$$r = \frac{20(497143) - (592)(15702)}{\sqrt{[20(19852) - (592)^2][20(12905468) - (15702)^2]}}$$

$$= \frac{9942860 - 9295584}{\sqrt{[397040 - 350464][258109360 - 246552804]}}$$

$$= \frac{647276}{\sqrt{46576 \times 11556556}}$$

$$= \frac{647276}{\sqrt{538258152256}} = \frac{647276}{733660{,}79}$$

$$r \approx 0.882$$

The way we have calculated our r is admittedly tedious and there are plenty of steps in which human error could have crept in; however unless we first understand the "nuts and bolts" of the manual calculation we will be unlikely to appreciate just what is being calculated for us when we hand over the task to a spreadsheet or statistical program. Such programs will nonetheless calculate a value for r without complaint even if the initial assumptions are wildly off, which may result in a value without meaning, or without an appropriate meaning. The temptation to calculate "because we can" and cast about for an *ad hoc* explanation afterwards is ever present when the power to crunch numbers exceeds the power to reason about them. The better programs do however ask about some of the assumptions such as the equality of variance, either homoscedastic, heteroscedastic or unknown.

Note that the answer that we have arrived at is positive, which is relatively large and consistent with the scatter plot. As a rule of thumb the *effect size* of Person's *r* (whether positive or negative) is considered:

'small' at around r = 0.1

'medium' at around r = 0.3 and

'large' for any value $r \geq 0.5$.

So clearly there is a strong positive correlation in our student's sample between "knowledge about calcium" and "mean daily dietary intake of calcium"; although we cannot say for certain whether their knowledge about calcium *caused* people in the sample to consume more calcium daily or whether those who had more calcium in their daily diet were more likely to be more knowledgeable on the subject or to seek out such knowledge. Of course, we have our suspicions as to which way the causal arrow points, although the value of r on its own cannot answer that question.

Although we have rather large effect size, we have still to test our hypothesis formally. Suppose we set our confidence level at 95%, that is α = 0.05. Because we technically do not know whether our alternative hypothesis is directional we calculate the value for the two-tailed probability using a reliable online calculator and discover that p = 0.00000027. This is a vanishingly small probability - much smaller than our α of 0.05. Therefore we reject the null hypothesis (H_0) in favour of the alternative hypothesis (H_1), to wit that there is a relation between "knowledge about calcium" and "mean daily dietary intake of calcium" in our student's sample.

Since the p-value for the one tailed alternative hypothesis is smaller yet, (precisely by half, p = 0.00000014) we would be similarly justified in rejecting the null hypostasis in favour of a directional alternative hypothesis. Informed by judgement rather than any further calculation, this is just what our student does. She reasons that, given the very many benefits of incorporating sufficient calcium in one's diet, anyone who has learned of such benefits will be inclined to do so as a matter of practice. On the other hand, while calcium is essential for many vital functions in the body, including the nervous system, it would be a fantastical sort of element to actually induce behaviour as complex as acquiring knowledge about itself. Without hesitation therefore, she concludes that the causal arrow points *from* knowledge about calcium in the diet and in the body *to* the respondents' mean daily dietary intake of calcium. In fact, the size of the effect detected in her sample leads her to hypothesise further that the same will be generalizable to the population of all students and possibly all people.

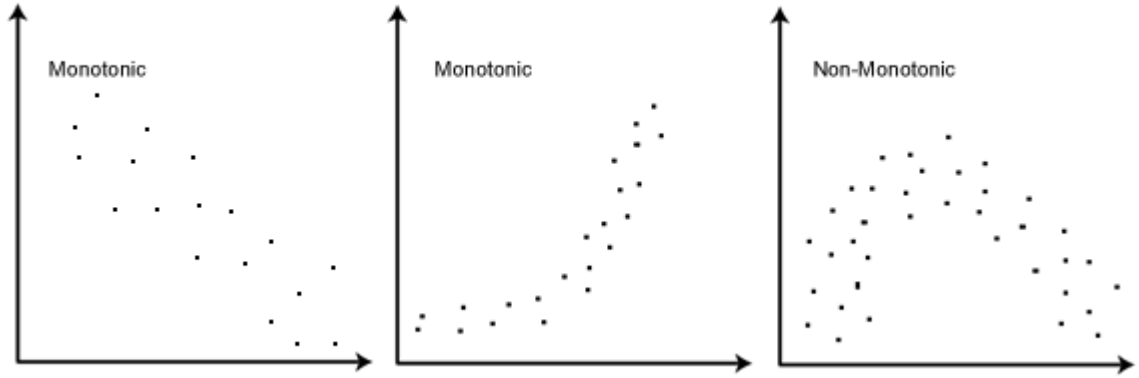
Kruger and Janeke (2012) devote a further section to "Using Pearson's r to test an hypothesis". The principles involved are the same as those discussed in Critical Reasoning 15, however they do not discuss what to do if one (or more) of variables is ordinal.

Spearman's rank correlation coefficient

Named after the English psychologist Charles Spearman, **Spearman's ρ** (pronounced 'rho') is a non-parametric version of Pearson's r that measures **rank correlation** *i.e.* the strength and direction of statistical association between two ranked variables. We can rank several observations of a particular variable by assigning order 'first', 'second', 'third' *etc.* to them. A client satisfaction survey is a common example: After being served a client is asked to rate the service on, say, a five point **Likert scale** from 1 for 'very bad' to 5 for 'excellent'. Clearly, we can rank such responses because 5 is always better than 4, which is better than 3, and so on. Because such responses typically tend to cluster around the middle value and because we do not have an objective measure as to just how far, say, 5's differ from 4's *vs.* 3's from 2's, Pearson's requirement of linearity is not met. On the other hand, so long as the variables involved are measured on an ordinal, interval or ratio scale and the relation between the variables is monotonic, *i.e.* always approximating an increasing function or always approximating a decreasing function, the criteria for using Spearman's ρ are met. There are formal tests of monotonicity; however the easiest method is to simply look at a scatter plot of the data and compare them to the following three graphics[2] by way of example.

A Likert scale (pronounced LIK-ərt) is a ranked psychometric scale used in survey questionnaires that offers a quick, standardised, ordered or ranked response method to assess level of agreement (or disagreement) towards attitudes or experiences. Their chief advantage is that respondents are not forced to express an either-or opinion, when some may wish to be accounted more or less neutrally. Survey forms that use multiple Likert scales are useful because different items of interest can be assessed for inter-correlation between scores and hence also for consistency.

[2] Courtesy of https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php



The one on the right is clearly not monotonic because it is not *only* increasing or *only* decreasing over the entire interval plotted.

Before Spearman's ρ can be calculated we have to rank the scores from our raw data. This can be done by assigning a 1 to the lowest score, a 2 to the second lowest, and so on, and an n to the highest ranking score out of a sample size of n. If two scores are identical, they are known a 'tie' and should be ranked according to average of the two positions that they otherwise would have occupied. Suppose two scores are tied for 3rd spot - we have no way of telling which should be put in 3rd and which should be 4th place - therefore we take the average of 3 and 4 = (3 + 4) ÷ 2 = 3,5th place. Of course this means that no two other scores may occupy 3rd or 4th place. The same general procedure is followed for three way ties or more.

Once our raw scores X_i and Y_i for sample size n have been converted into ranks rgX_i and rgY_i, Spearman's ρ is calculated as Pearson's r for ranked variables as follows:

$$\rho = r_{rg_X\ rg_Y} = \frac{\mathrm{cov}(rg_X; rg_Y)}{\sqrt{\mathrm{var}(rg_X)\mathrm{var}(rg_Y)}}$$

where

- $r_{rg_X\ rg_Y}$ is Person's correlation coefficient applied to ranked variables rgX_i and rgY_i
- $\mathrm{cov}(rg_X; rg_Y)$ is the covariance of the ranked variables, and
- $\mathrm{var}(rg_X)$ and $\mathrm{var}(rg_Y)$ refer, respectively, to the variances of rg_X and rg_Y on their own.

If, as happens when there are no ties, *i.e.* when all n ranks are distinct integers, then Spearman's ρ may be calculated as follows

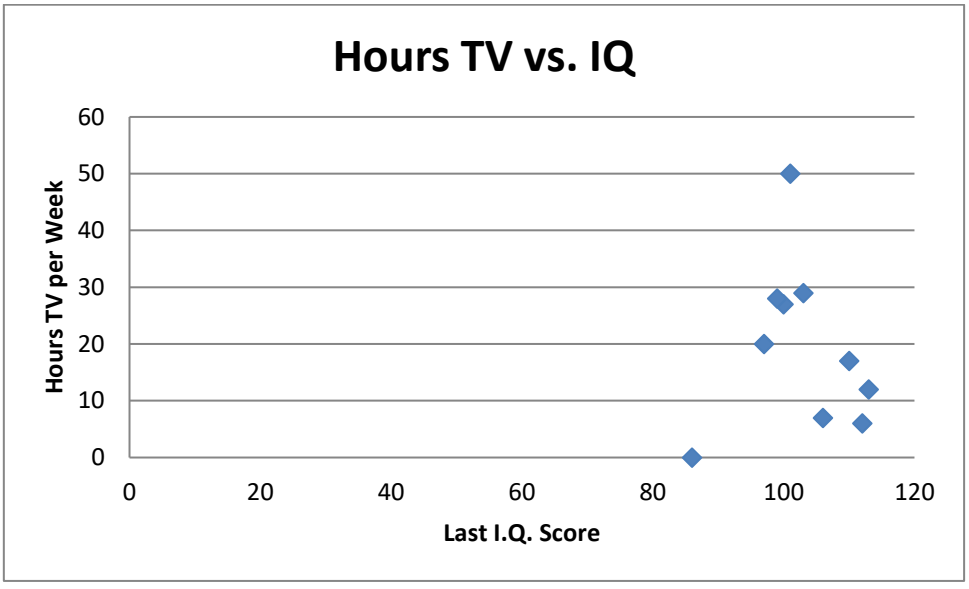
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where,

- $d_i = rg(X_i) - rg(Y_i)$ is the difference between the two ranks of each observation, and
- n is the number of observations. (Wikipedia: Spearman's rank correlation coefficient)

The following example is adapted from the one at the same Wikipedia entry: Suppose a grade 5 educator believes that watching a lot of TV is correlated with low intelligence. He operationalises his variables as X = IQ and Y= Hours spent watching TV per week. He picks ten learners at random and asks each one's parents to keep a diary of the time that their child spends watching TV for one week. Being an educator, he has access to each of his learner's permanent record on which his or her most recent IQ test result is recorded. His hypothesis is that 'Hours spent watching TV' is negatively correlated with 'IQ'. He wishes to test his hypothesis at the 90% confidence level, *i.e.* α = 0.1. His results are tabulated as follows, followed by a scatter plot of the same:

IQ, X_i	TV, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17



The scatterplot does not reveal much, given the paucity of the data. The data points clearly do not approximate a linear function, although they do appear to sweep from the top left of the graph, slightly down towards the right and in that way appears to approximate a monotonic function. Can we test the educator's hypothesis using Spearman's ρ? Yes, given that the basic assumptions are met. First we have to sort the first column above according to rank, from lowest to highest, assigning a 1 to the lowest value and a 10 to the highest. This we recorded in an adjacent column. Next, we sort the second column above in the same way and record the rankings alongside, so that we now have four columns - two for the raw data and two for the rankings, thus

IQ	TV	rank x_i	rank y_i
106	7	7	3
86	0	1	1
100	27	4	7
101	50	5	10
99	28	3	8
103	29	6	9
97	20	2	6
113	12	10	4
112	6	9	2
110	17	8	5

We observe that there are no ties, therefore we can use the second of the two formulae above for calculating Spearman's ρ. We know that n = 10 however, we will have to calculate d_i for each case and thence d_i^2 because these are the variables that need to be summed in the formula. To do this we calculate the difference in rank d_i between each entry in column x_i and its corresponding entry in column y_i and record them in yet another two columns alongside. In order to facilitate summation, we create a final column in which we square the quantity d_i to provide d_i^2 for each entry, thus

IQ	TV	rank x_i	rank y_i	d_i	d_i^2
106	7	7	3	4	16
86	0	1	1	0	0
100	27	4	7	-3	9
101	50	5	10	-5	25
99	28	3	8	-5	25
103	29	6	9	-3	9
97	20	2	6	-4	16
113	12	10	4	6	36
112	6	9	2	7	49
110	17	8	5	3	9

When we add up the last column we find that $\sum d_i^2$ = 194, which together with our n = 10, we can substitute into the formula above, as follows

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 194}{10(10^2-1)}$$

$$\rho = -0.1\dot{7}\dot{5}$$

The sign of ρ is at least in the right direction of the hypothesis being tested, however when we use a reliable online calculator, we find that this value is associated with a p-value of 0.627188. Obviously, this is much larger than the α of 0.1 at which the educator was prepared to test his hypothesis,

therefore we must reject his hypothesis in favour of the null hypothesis, to wit that 'Hours spent watching TV' is not correlated with 'IQ'.

Note that this example was no doubt tweaked so as to be easily computable using the second formula. In most cases however, and especially if the data set is large, a researcher will use a statistical package to calculate the value of ρ directly from the raw data. When using such a package it is advisable to find out just how the programme calculates the value so as to check that key assumptions are not being overlooked.

The χ^2 test for association between nominal variables

So far we have been considering variables that can be measured on a quantitative scale, whether ratio, interval or ordinal. Nominal or categorical variables, such as sex, marital status, employment sector, hair or eye colour, HIV status *etc.* cannot be measured in this way but they can be counted and so do have a frequency distribution. We might, for example, assign the codes 1 to brown eyes, 2 to green eyes and, 3 to blue eyes. Clearly, a 1 for brown eyes and a 2 for green eyes do not add up to a 3 for blue eyes in anything like an arithmetical sense. If wanted to test whether eye colour is associated with some other factor, such as marital status, say, we could look at the frequency of distribution of eye colours and marital status in a random sample and extrapolate what we would have *expected* to find among members of the sample and use that information to test an hypothesis. Our first step would be to formally state the statistical hypotheses, thus

- The **null hypothesis** (H_0): There is no relation between "eye colour" and "marital status"
- The **alternative hypothesis** (H_1): There is a relation between "eye colour" and "marital status"

Next we would have to decide on a level of confidence at which to test our hypothesis, say at the 90% level (*i.e.* a level of significance of 10% or α = 0.1). Once we have collected and codified our data we need to arrange it in a standardised format known as a contingency table.

A **contingency table** (or cross tabulation) is a two dimensional table in the form of a matrix that displays the multivariate distribution of frequencies of the variables involved. (Wikipedia: Contingency table) Each frequency is recorded in a unique cell designated by its row number i and its column number j, just as cells in a spreadsheet are referred to by their unique combination of row number and column letter. The following contingency table represents the type that would be required to record the raw data from our example. Note the "O" in each case stands for the *observed* frequency.

		Eye Colour			
		1 Brown	2 Green	3 Blue	Row total
Marital Status	1 Single 2 Married	O11 O21	O12 O22	O13 O23	**O1.** **O2.**
	Column total	**O.1**	**O.2**	**O.3**	**O..**

Once we have our data in hand, the contingency table should be filled out as follows: In

- cell O11 (row 1, column 1) record the number of subjects who are single and brown eyed
- cell O12 (row 1, column 2) record the number of subjects who are single and green eyed
- cell O13 (row 1, column 3) record the number of subjects who are single and blue eyed
- cell O21 (row 2, column 1) record the number of subjects who are married and brown eyed
- cell O22 (row 2, column 2) record the number of subjects who are married and green eyed
- cell O23 (row 2, column 3) record the number of subjects who are married and blue eyed

The frequencies for each row are summed and recorded in cells **O1.** and **O2.** respectively; meanwhile the frequencies for each column are summed and recorded in cells **O.1**, **O.2** and **O.3** respectively. Note that the sum of the row totals **O1.** and **O2.** should be the same as the sum off the column totals **O.1**, **O.2** and **O.3**, which should be the same number of subjects in the sample. This number is recorded in cell **O..** Verifying that these totals matchup is a useful (but not fool proof) way of checking that the data have been correctly entered.

If the null hypothesis is true and there really is no significant relation between the variables "marital status" and "eye colour" we would expect the observed frequencies to be distributed in a predictable way. We designate these **expected cell frequencies** as $\mathrm{E_{ij}}$ and calculate them as follows: For each cell, take its observed row total and multiply it by its observed column total and divide the result by the overall total. Thus

$$\mathrm{E_{ij}} = \frac{(\mathrm{Oi.} \times \mathrm{O.j})}{\mathrm{O..}}$$

Having calculated the expected frequencies for each cell it is a good idea to record them alongside the observed frequencies in our contingency table because we want to know if, and to what extent, our observed frequencies differ from the expected ones. If there are such differences and we have ruled out the **interaction effect** (where one or more of the variables differs depending on the level of another) we can proceed calculating Pearson's chi-squared test statistic according to the formula

$$\chi_p^2 = \sum \mathrm{ij} \frac{(\mathrm{O_{ij}} - \mathrm{E_{ij}})^2}{\mathrm{E_{ij}}}$$

This says that we take the observed minus expected frequency for each cell $\mathrm{E_{ij}}$ in turn, square it and divide it by the expected frequency for that cell, then we sum up all the terms to yield a value for χ_p^2. Kruger and Janeke (p. 145) recommend drawing up a table to accommodate the various steps involved in the calculation. For a 3×2 contingency table, in general, such a table of calculations might look as follows,

Row	**Column**	**O**	**E**	$(\mathbf{O}-\mathbf{E})$	$(\mathbf{O}-\mathbf{E})^{\mathbf{2}}$	$\frac{(\mathbf{O}-\mathbf{E})^{\mathbf{2}}}{\mathbf{E}}$
1	1	O11	E11	$(O11-E11)$	$(O11-E11)^2$	$\frac{(O11-E11)^2}{E11}$
1	2	O12	E12	$(O12-E12)$	$(O12-E12)^2$	$\frac{(O12-E12)^2}{E12}$
1	3	O13	E13	$(O13-E13)$	$(O13-E13)^2$	$\frac{(O13-E13)^2}{E13}$
2	1	O21	E21	$(O21-E21)$	$(O21-E21)^2$	$\frac{(O21-E21)^2}{E21}$
2	2	O22	E22	$(O22-E22)$	$(O22-E22)^2$	$\frac{(O22-E22)^2}{E22}$
2	3	O23	E23	$(O23-E23)$	$(O23-E23)^2$	$\frac{(O23-E23)^2}{E23}$
						SUM = χ_p^2

This table contains only the mathematical steps required to calculate χ_p^2 not the raw scores. If you chose to use a spreadsheet to compute χ_p^2 in this way you will have to both encode the steps and enter the raw data into the program. Using a statistical package to generate χ_p^2 directly from the observed frequencies will be much quicker but will not reveal any of the operations "under the hood" as it were; therefore it is worth studying the table in its present form.

Notice that we are essentially interested in *differences*, in particular differences between the observed and expected frequencies. If the null hypothesis is correct and the observed frequencies coincide exactly with those of the expected frequencies, all terms to the right go to zero, and hence also their sum, χ_p^2. Of course this almost never happens due to sampling error but if such differences are sufficiently small χ_p^2 will have a low enough value that we might decide not to reject the null hypothesis anyway. Note also that all differences are squared, therefore χ_p^2 can never be less than zero. Furthermore, because the differences are squared the χ_p^2 test is not sensitive to deviations one way or the other, *i.e.* it is not a directional test statistic, although a careful study of the contingency table, with the expected frequencies pencilled in, will reveal in which direction certain differences are leaning.

Like all statistical tests, Pearson's χ^2 test works best with larger samples, however in addition there should be adequate representation in the expected frequencies of each cell of the contingency table. As a rule of thumb, the expected frequencies per cell should be 5 or more in a 2×2 table or 5 or more in 80% of cells of larger tables. Sometimes it may be practical to combine certain categories in order to increase the expected frequencies. Finally, it goes, almost without saying, that none of

the expected frequencies may be zero because then one of the terms on the right hand side of the table above will be undefined and so too will the sum. (Wikipedia: Pearson's chi-squared test)

Once we obtain a value for χ_p^2 when testing an hypothesis, that statistic still needs to converted into a corresponding p-value, which can be done using a statistical package or reliable online calculator, in which case we will be prompted for the degrees of freedom. For a contingency table with r rows and k columns, this is given by

$$df = (r-1)(k-1)$$

Having set out the theoretical and mathematical considerations for the χ^2 test for association between nominal variables, we can proceed to test our hypothesis, to wit that there is a relation between "eye colour" and "marital status" according to the following data collected by an immigration official by looking at the passports of 100 people as they crossed a border post into South Africa. Since the eye colour and marital status of the passport holder are recorded on the first page of all passports, this could be done without a fuss. We have summarised the observed data in the following contingency table, including the row and column totals:

Observed

		Eye Colour			
		1 Brown	2 Green	3 Blue	Row total
Marital Status	1 Single	40	6	2	**48**
	2 Married	45	4	3	**52**
	Column total	**85**	**10**	**5**	**100**

As can be seen, subjects with brown eyes vastly outnumbered those with either green or blue eyes, as expected, since the genotype for brown eyes is dominant over that for green or blue, which are recessive. We also note that about half of each eye colour grouping were single and about half of each were married. So just by looking at the contingency table we suspect our alternative hypothesis may be in trouble, never the less we press on with testing it formally. Next we draw up a table of expected frequencies based on the formula for $\mathrm{E_{ij}}$ above, thus

Expected

		Eye Colour		
		1 Brown	2 Green	3 Blue
Marital Status	1 Single	40.8	4.8	2.4
	2 Married	44.2	5.2	2.6

Ideally, we would have liked to have an expected frequency of 5 or more in at least 80% of the cells above according to our rule of thumb. In a genuine research project we could have achieved this by either combining the categories 'Green' and 'Blue' (both recessive) or by increasing our sample size until we had adequate representation. However our purpose here is simply explanatory, therefore we proceed to calculate the test statistic χ_p^2. We did so both by the table method and by using the

program Microsoft Excel™ and verified that both values are the same, (except for some small difference due to rounding). Since we have already demonstrated the table method, you may wish to explore the much quicker spreadsheet option, as follows.

Open a new Excel™ document. Copy both the observed and expected frequency contingency tables either side-by-side or one under the other. Left click an open cell and click the formula bar where the 'fx' symbol appears. Select the category "Statistical" and from the pull-down menu select 'CHISQ.TEST'. You will be prompted to input "Actual range" and "Expected range". This can be done by left clicking in the top left cell of the observed frequencies, holding down the mouse key and sweeping the mouse pointer across to the right and down to the bottom right of the observed frequencies. Repeat this for the expected frequencies. Click OK. If everything has been entered correctly and the correct fields have been selected, the spreadsheet program will calculate the χ_p^2 test statistic, the degrees of freedom and will display the associated two-tailed p-value in the open cell selected. The p-value retuned for this example was ≈ 0.69. This is much larger than our α = 0.1, therefore we decide not to reject the null hypothesis, to wit that there is no relation between "eye colour" and "marital status".

The value of spreadsheet programs and statistical packages is obviously enormous, especially when it comes to processing large amounts of data that could potentially take a human weeks to calculate by hand. However to reiterate our earlier warning: Such programs are dumb (for now at least)! They will blindly calculate almost any statistic that we demand that is not an illegal function, such as dividing by zero, or a syntax error. So unless we already have an understanding what sort of statistic we require and why, and have some idea as to how it is arrived at, the output of our program may be meaningless.

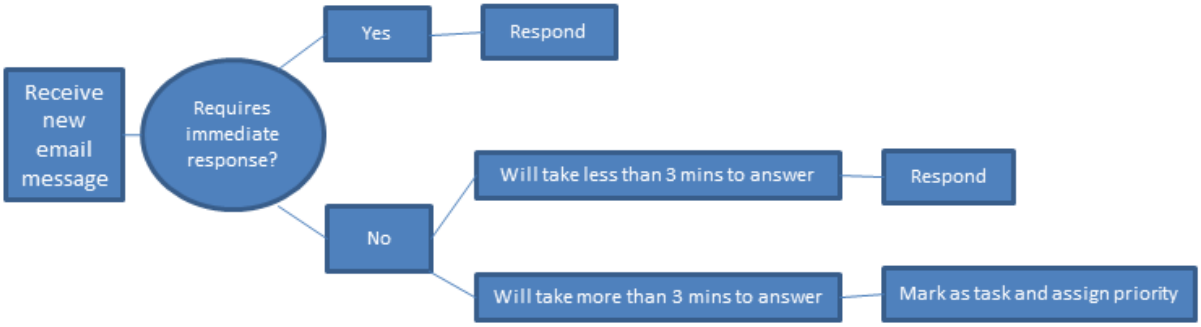
Endnote on curriculum

As of 2012 this topic was the last one covered in the final undergraduate Psychology Statistics module at UNISA. If you have worked through all of the Critical Reasoning study units, including the relevant tasks and feedback, that deal with probability and statistics posted on this website thus far you will have covered most, if not all, of the undergraduate material on the subject. Of course, psychology students are not required to take modules on logic or set theory but a good practical knowledge of informal logic (Critical Reasoning 01 - 04) as well as the Scientific Method (Critical Reasoning 12) will be an advantage. In some universities these have been incorporated into the general Humanities curriculum. Finally the topics on Heuristics (Critical Reasoning 06) and Groupthink (Critical Reasoning 08) form part of the undergraduate Cognitive Science and Economics syllabi at all universities in South Africa.

If you are preparing for the final undergraduate examination in Psychology Statistics or simply wish to consolidate your knowledge of the statistical topics posted so far, please attempt the task below. There will be further posts on statistical topics a philosophy.org.za, the next beginning with Linear Regression and Analysis of Variance (ANOVA).

Task

On the last page of their study guide Kruger and Janeke (2012) summarise the various statistical tests dealt with the syllabus in the form of a **decision tree**, which is a flowchart-like structure where each internal node represents a decision to be made and each branch represents the outcome of a decision. The end nodes meanwhile represent various final outcomes of the decision process. Decision trees are especially useful in illustrating algorithms, machine diagrams and formal classifications. Even very simple decision processes can be represented as decision trees as in the following example of what to do when checking e-mail, with not much time to spare.[3]



Of the statistical tests covered so far, the first decision to be made is what we wish to do with our data: Do we wish to test

1. the difference between one group and a constant
2. the difference between two groups with σ unknown, or
3. the relation between two variables?

These can be represented as the "root" nodes of a decision tree. The next decision to be made is about the nature of our data: If we selected the first node above we would want to know whether the population standard deviation is known or unknown. This would lead to two further nodes

1.1. σ known

1.2. σ unknown

These in turn lead to two terminal nodes for this part of the decision tree

1.1.1 $z_{\bar{x}}$ test

1.1.2 $t_{\bar{x}}$ test.

In a similar fashion, if we selected the second root node above, we would want to know whether our groups were independent or dependent. These would lead to the terminal nodes: t_c test and t_d test respectively. Finally if we selected the third root node above, we would want to know whether the measurement scales were both continuous, interval or ratio, one or both ordinal, or one or both

[3] This one can no longer be found at its original location.

nominal or categorical. These would lead to the terminal nodes: Pearson's r, Spearman's ρ and Pearson's χ_p^2 respectively. The task is to represent these options and statistical test outcomes that they lead to as a decision tree.

Feedback

Kruger and Janeke present theirs using rather austere grey boxes joined by branching lines from left to right. Although there is no single right way of doing so, there are plenty of wrong ways, such as a decision node pointing to the wrong statistical test. Our idea was to turn the page landscape and draw three trees representing the thee initial tasks, branching upwards with the various statistical tests represented as oversize leaves. This allowed us enough space to note other assumptions such as the population distributions and the sample sizes. If you have devised another way of representing the decisions that go into selecting the right statistical test for the task and data at hand that is clearer or easier to remember, so much the better.

The Next Critical Reasoning study unit concerns "truthiness and worse".

References:

KRUGER, P. & JANEKE, H. C. (2012) *Psychological Research - Study Guide for PYC3704*. UNISA Department of Psychology