Week 10 - Correlations

Reading: 147-159

Correlations overview

Correlations qunatitatively describe how two variables can co-vary, where a change in one variable is associated with a change in the other variable. For example, in a study, we may find a positive correlation between time spent studying and exam performance, meaning that as time spent studying increases, exam performance also tends to increase.

However, correlation *does not equal* causation, meaning that just because two variables are related, it doesn't necessarily say anything as to whether one *causes* the other, which requires experimental manipulations. For example, we may find a correlation between ice cream sales and crime rates, but this does not mean that eating ice cream causes crime. Instead, there could be a third variable, such as temperature, that affects both ice cream sales and crime rates. On balance, correlations can help to identify potential alternative explanations which may be subsequently investigated using experimental methods (to establish causality).

Is there a linear relationship between two continuous variables?

Strength of relationship (correlation coefficient = *r)*

Direction of relationship (negative/positive *r)*

Quality of relationship (significant?)

Pearson's *r* for continuous variables (e.g., height, weight)

Spearson'srfor ranked variables (e.g., class performance, height)

- The goal of running a correlation test is to quantify the strength between two continuous (Pearson) or ranked (Spearman) quantities.
- The two variables we associate can be represented as *x* and *y*
- To represent each quantity within our variable, we can add the subscript j. For example, Zi,z_2,z_3. ..*Xi* and *yi,y-2,*y_s*. ..yi*
- To represent the mean of each variable, we can add a " estimate. For example, the means of *x* and *y* are respectively represented by *x* ˆand *y.* ˆ
- *\We* can then estimate the correlation coefficient *r,* which can range from — 1 ≤ 0 ≤ 1. The *closer* to 0, the *weaker* the correlation.

A working example

Suppose we distributed a Physical Activity Scale to 6 female and 4 male participants to measure how much time they spend each week on vigorous, moderate, walking and sedentary physical activities. For consistency, all data is collected in minutes. The data for the hypothetical 10 participants are illustrated in the table below.

ID	Sex	Age	Walking	Vigorous	Moderate	Sedentary
1	Male	21	237	120	183	399
2	Male	25	262	116	153	276
3	Male	29	201	104	136	327
4	Male	24	297	73	108	336
5	Female	22	228	67	128	387
6	Female	44	393	57	297	332
7	Female	48	361	59	224	309
8	Female	40	344	48	263	212
9	Female	49	340	56	268	204
IO	Female	46	366	45	201	337

Descriptive statistics

The sample is aged between **21** and **48** years (Al = **34.8,** *SD* — **~~11~~.61** years), with **4** males and **6** females.

Research hypothesis

You suspect younger persons are more likely to engage in vigorous physical activities. You want to know whether there might be an association between participants' **age** and the amount of **vigorous** activity engaged in.

Statistical hypothesis

The **Pearson correlation coefficient,** summarily presented as *r,* will not significantly vary from a null estimate, so H_Q **: r = 0.**

Pearson correlation coefficients

The Pearson correlation coefficient is a measure of the *linear* association between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

The Pearson coefficient quantifies the strength and direction of the relationship between two variables, providing a single value summary of the relationship. It is commonly used in psychological research to assess the relationship between variables, but it should be used with caution as it assumes a linear relationship and may be influenced by outliers and confounding variables.

> While there are multiple ways to estaimte coefficients, we focus on the Pearson correlation coefficient as this is the most commonly used method for quantifying linear relationships between two continuous variables.

Computing a Pearson correlation coefficient involves the following steps:

1. *Prepare the data:* Ensure that the data is in the form of two continuous variables. If necessary, clean and pre-process the data to remove any missing values or outliers.
2. *Compute the means of both variables:* Calculate the mean of each of the two variables.
3. *Subtract the mean from each data point:* For each data point, subtract the mean of that variable to get the deviation from the mean.
4. *Compute the product of deviations:* Multiply the deviation of each data point for both variables.
5. *Compute the numerator:* Sum all the products from step 4 to get the **numerator** of the Pearson correlation coefficient.
6. *Compute the denominator:* Square the standard deviations for each variable, take the sum of their product then square root the value to get the **denominator** of the Pearson correlation coefficient.
7. *Compute the Pearson correlation coefficient:* Divide the **numerator** by the **denominator** to get the Pearson correlation coefficient. The resulting value will range from -1 to 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation.

Recall our hypothesis was to explore whether age *(age)* was statistically associated with vigorous physical activity *(vigor).*

Suppose the average age across participants is represented by *$\hat{age}$,* and each participant's individual age is represented by *age*^ Similarly, suppose the average anount of vigorous activity engaged across participants is *$\hat{vigor}$,* and each participant's individual vigorous activity is *$vigor_i$*.

The Pearson correlation coefficient *(r)* can be computed using the following formula:

$$r = — \frac{\text{^}(agei - age)(\hat{v}igor- vigor)_\lambda \quad \hat{}}{yj\text{^^}agei — age)^{2}\,\hat{}\,\text{^}\{vigori — vigor)^{2} \quad \hat{}} = — \frac{\sum(x_i - x)\hat{}\text{^} - y) \quad \hat{}}{\sqrt{\sum(x_i \sim \hat{x})^2(y_i \sim \hat{y})^2}}$$

where *x* and y represent the two continuous variables being associated.

Let's compute the values necessary to run the above formula.

agei	*$\hat{age}$*	*agei — $\hat{age}$*	*(agei - $\hat{age}$)^2*	*vigors*	*$\hat{vigor}$*	*vigori*	*— $\hat{vigor}$ (vigori — $\hat{vigor}$)^2*	*(agei — $\hat{age}$) (vigori — $\hat{vigor}$)*
21	34.8	-13.8	190.44	120	74.5	45.5	2070.25	-627.9
25	34.8	-9.8	96.04	116	74.5	41.5	1722.25	-406.7
29	34.8	-5.8	33.64	104	74.5	29.5	870.25	-171.1
24	34.8	-10.8	116.64	73	74.5	-1.5	2.25	16.2
22	34.8	-12.8	163.84	67	74.5	-7.5	56.25	96
44	34.8	9.2	84.64	57	74.5	-17.5	306.25	-161
48	34.8	13.2	174.24	59	74.5	-15.5	240.25	-204.6
40	34.8	5.2	27.04	48	74.5	-26.5	702.25	-137.8
49	34.8	14.2	201.64	56	74.5	-18.5	342.25	-262.7
46	34.8	11.2	125.44	45	74.5	-29.5	870.25	-330.4

We can first estimate the **numerator,** which requires summing the product of individual deviations (Steps 4 & 5).

Sum the product of the two differences, represented in the last column in the above table

- *U(agei - age)(v̂igori - vigor)*
- £[-627.9, -406.7, -171.1,16.2,96, -161, -204.6, -137.8, -262.7, -330.4] = -2190

We can then estimate the **denominator,** which is the squared root of the summed squared deviations (Step 6).

- $\sqrt{\sum(age_i$ *~ agê)^2 ^(vigori - vigor)^2*
  # Multiply the sum of the two squared differences and take the square root

£[190.44,96.04,33.64,116.64,163.84,84.64,174.24, 27.04, 201.64,125.44] and £[2070.25,1722.25,870.25, 2.25,56.25,306.25,240.25,702.25,342.25,870.25] gives us 1213.6 and 7182.5 respectively.

The product of the two estimates are 1213.6 x 7182.5 = 8716682.

The square root is V8716682 = 2952.4

Finally, divide the **numerator** from the **denominator** to acquire the **Pearson correlation coefficient**

£(a^ej - *age)(v̂igori - vigor)* -2190

r = ——===== =-- = —.742

^YXᵃgᵉi - ᵃgᵉY m™g°ʳi - vigor)$^{2\ 2952,4}$

We can report that **age** and **vigorous activity** are *highly negatively correlated, r* = —.742.

In relation to our research hypothesis, being *younger* appears highly associated with morevigrous physical activity.

The next step involves determining whether this relationship is significantly different from a chance estimate *(p* = .05).

Determining the p-value of the correlation coefficient

Frequentist statistical tests require assuming that the variables being investigated have a fictional null relationship (that which woul be expected by chance).

The goal of the test is to determine whether the hypothesis of this null relationship can be statistically **retained** *(p* > .05) or **rejected** *(p* < .05).

In relation to our target variables, the null hypothesis (Ho) would be that there is **no** significant relationship between *age* and *vigorous physical activity*

Manually determining a p-value

The p-value for a Pearson correlation coefficient (r) requires, first, converting the coefficient into a test statistic (t). This is done by:

$$t = \frac{r_{Age,Vigor}\sqrt{n_{=10} - 2}}{\sqrt{1 - r^2_{Age,Vigor}}}$$

where *r* is the Pearson correlation coefficient between variables *Age* and *Vigor* and *n* is the sample size (which in our example is *n* = 10).

The p-value can then be estimated using the *t-distribution* with n — 2 degrees of freedom:

P = 2(1 - i„_2,|t|)

where i_n-2,|t| ^{is} the cumulative distribution function of the t-distribution evaluated at \t\.

Years ago, the t-distribution had to be manually looked up using tables. This is no longer the case thanks to modern statistical software like R, which includes these tables 'built-in'.

> This p-value can be used to test the null hypothesis that there is no correlation between the two variables, against the alternative hypothesis that there is a non-zero correlation. A small p-value (< 0.05) indicates that it is unlikely to observe a correlation as strong as the one calculated under the assumption of no correlation, and provides evidence to **reject** the null hypothesis.

Running a correlation in R

The steps in computing correlation coefficients (r) and their probability of rejecting the null hypothesis (p) are a two step process in /?:

Step 1: Input the values to two variables to be correlated

Assign data to two variables called age and vigor within the R console.

```
age   <- c(21,25,29,24,22,44,48,40,49,46)     # Age in Years
vigor <- c(120,116,104,73,67,57,59,48,56,45)  # Vigorous Activity in minutes
```

We can check whether the values have been entered correctly by typing in the variable name and pressing ‘Enter’ on the keyboard

```
age    # years
```

```
##  [1] 21 25 29 24 22 44 48 40 49 46
```

```
vigor # minutes
```

```
##    [1] 120 116 104      73    67    57    59    48    56    45
```

Step 2: Place the values into the built-in cor.test() function

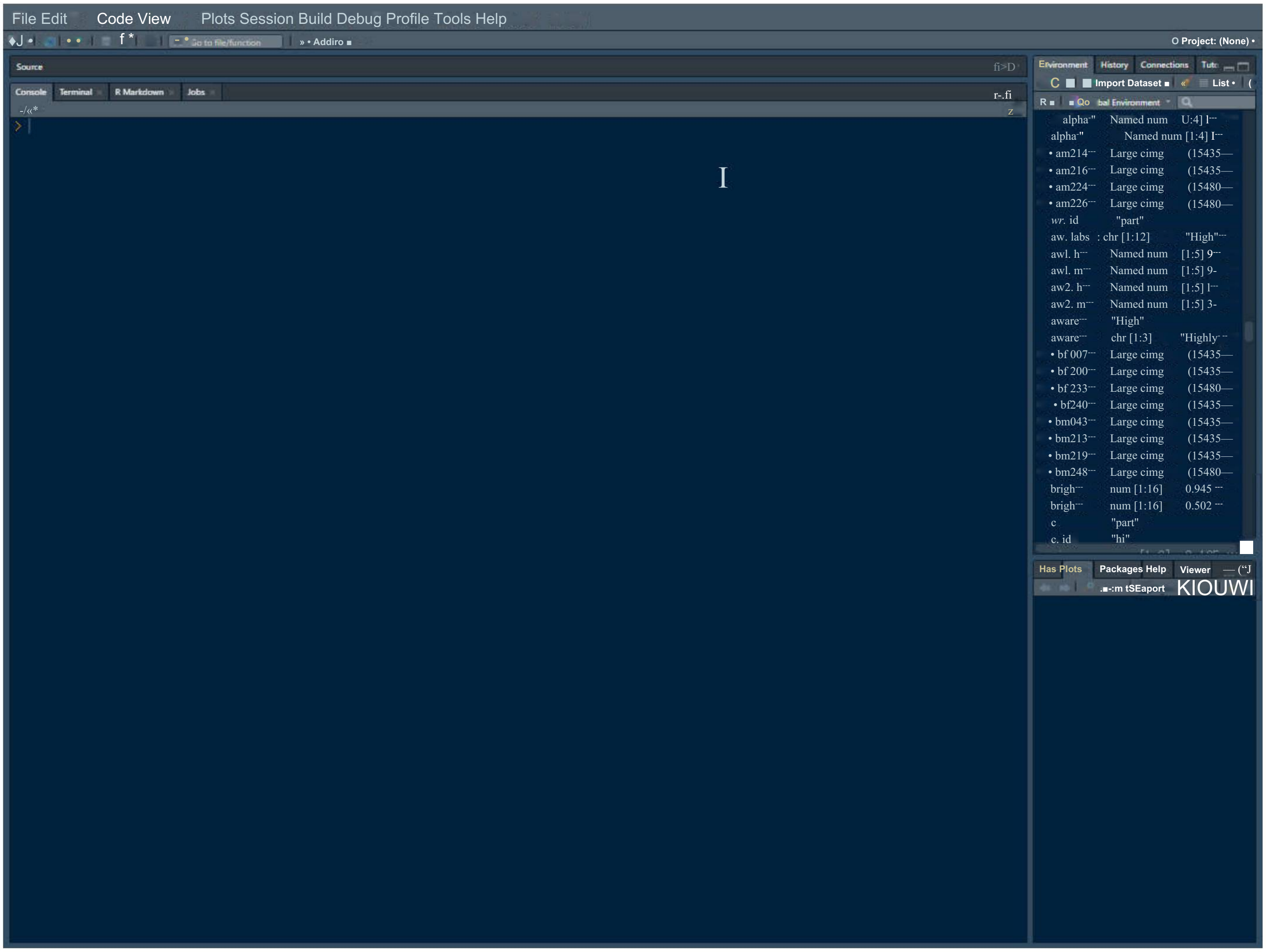
```
cor.test(age,vigor)          # For correLations, the ordering of the 'x' and 'y' variabLes are not
important
```

```
it#
## Pearson's product-moment correlation
##
## data: age and vigor
## t = -3.1283, df = 8, p-value = 0.01405
##    alternative    hypothesis:    true    correlation    is    not    equal    to    0
## 95 percent confidence interval:
##    -0.9348060 -0.2104132
## sample estimates:
##              cor
## -0.7417687
```

The cor value is our correlation coefficient (r). The p-value gives us the probability of assuming the present data assuming the null hypothesis was true. Ap = .014 implies there is a 1.4% likelihood of observing the present correlation if the null hypothesis was true.

Because the likelihood of such a result is unacceptably low to assume the null hypothesis is true (represented as any *p* value **less** than .05), we can *reject* the former and claim that correlation is statistically significant.

These steps are shown in action below.



alpha·"	Named num	U:4] I···
alpha·"	Named num [1:4] I···	
• am214···	Large cimg	(15435—
• am216···	Large cimg	(15435—
• am224···	Large cimg	(15480—
• am226···	Large cimg	(15480—
wr. id	"part"	
aw. labs	: chr [1:12]	"High"···
awl. h···	Named num	[1:5] 9···
awl. m···	Named num	[1:5] 9-
aw2. h···	Named num	[1:5] 1···
aw2. m···	Named num	[1:5] 3-
aware···	"High"	
aware···	chr [1:3]	"Highly···
• bf 007···	Large cimg	(15435—
• bf 200···	Large cimg	(15435—
• bf 233···	Large cimg	(15480—
• bf240···	Large cimg	(15435—
• bm043···	Large cimg	(15435—
• bm213···	Large cimg	(15435—
• bm219···	Large cimg	(15435—
• bm248···	Large cimg	(15480—
brigh···	num [1:16]	0.945 ···
brigh···	num [1:16]	0.502 ···
c	"part"	
c. id	"hi"	

(The above image is a GIF file that only works for the HTML version)

> Results of the Pearson correlation indicated a significant negative correlation between participants' age and vigorous physical activity, r(8) = —.74,p = .014. The null hypothesis that age and vigorous activity are not related can be rejected.

The value inside brackets (8) refers to the degrees of freedom *(df)* in your data. For the present correlation, this is calculated as the total number of participants *(N* = 10) minus the total number of variables *(K* — 2).

The degrees of freedom for the present data is therefore *N* — *K* = 10 — 2 = 8. The *df* informs us how many parameters within our model are 'free to vary'.

Teaching Evaluation Instructions

To complete this week's evaluation, you will have to extract data from the table near the beginning of the page (the first table under the *A working example* header).

Your task is to first **manually** correlate the values for walking and moderate activity times using the steps discussed.

Show how you estimated the **numerator** and **denominator** estimates, then calculate the correlation coefficient.

Next, run the correlation test in R. To do this, first assign the walking and moderate activity times to two variables. For example:

```
walk <- c(l,2,3)
moder <- c(l,2,3)
```

Apply the cor.test() function to the two variable and examine the p-value and report the results of your correlation in the format illustrated above in blue highlights.

Remember that*p* < .05 implies a significant relationship, whereas/? > .05 implies a relationship is *not* significant.

Show the series of steps you used to calculate the correlation coefficient on a document file (.doc, .pdf), and interpret the value *(Hint:* Positive or negative values indicate positive or negative relationships; the further the coefficient is from 0, the *stronger* the relationship).

Take a screenshot of your code (or copy & paste it onto a text file) within your submission.

Good luck.