Spatial Analysis and Data Exploration in History and Archaeology, Spring 2021
LDA-H313

WEEK 4 EXERCISE: Correspondence Analysis

This week we will look at non-spatial analysis, with the main themes being the Chi-squared test and correspondence analysis, or *CA*. First, please place the **week4** folder, which contains the data we will examine, in the **sade2021** folder and set your workspace. Make sure that the path is exactly as indicated, for example without a space in the folder name:

```
Windows:
      setwd("C:\\sade2021\\week4")
Mac:
      setwd("~/Documents/sade2021/week4")
```

1. Chi-squared Test

1.1 Introduction

Correspondence analysis is related to a fundamental and much used statistical test called the Chi-squared test, also written as X^2. It is widely used for testing a statistical hypothesis in data, such as relationships between **categorical** (nominal and ordinal, see week 2 presentation powerpoint) variables in data. You can see that the below examples have in practice a spatial component, but the method itself is not limited to spatial statistics. In biology and medical science, for instance, we might use for testing whether there is a statistically significant relationship between having certain genes and suffering from certain diseases.

There are two types of Chi-squared test, and we will look at both of them. In archaeology, a classic use of the Chi-squared test is checking whether there is a significant relationship between archaeological sites and landscape properties such as soils. We will carry out a simple one-sample example looking at a hypothetical distribution of Finnish Iron Age monument sites.

1.2 Toy dataset

First, let us create some data. We will do this by creating vectors that contain data, and then combining them into a data frame. Create vectors:

```
soils <- c("clay", "morainic", "peat")
```

```
monuments <- c(21, 7, 24)
pct_land <- c(0.45, 0.25, 0.3)
```

Now combines these into a data frame with a command that turns each of the vectors into a column in the new table.

```
mydata <- cbind.data.frame(soils, monuments, pct_land)
```

Look at the data by typing the name of the new object. You should see a data frame of three columns and three rows (plus one "row" at the top giving the names of the columns). The first gives a soil type, the second tells you how many archaeological monuments have been observed on areas with that soil type in the study area, and the third gives the percentage of the study area covered by that soil type. So "clay" soils cover 45% percent of the study area, and contain 21 out of the 52 observed monuments.

This is a useful command so make a note of it, as is its sibling `rbind.data.frame` which creates data frames by binding them by row. Create some vectors and try `rbind.data.frame` on them to see what it does.

1.3 One-sample Chi-squared test

We will start with a one-sample Chi-squared test, also known as a **goodness to fit test**. Our **null hypothesis** is that there is *no significant relationship between the distribution of the observed monuments and the soil type areas they are found in*.

A one-sample Chi-squared test compares a sample of events (here: recorded monuments) to a theoretical expected distribution of events; in other words how well the actual and the theoretic populations correspond or "fit". To illustrate the point, lets create a fourth column to our data frame. This column gives the expected numbers of monuments per soil area, if there distributions were completely even.

```
mydata$expected <- mydata$pct_land * sum(mydata$monuments)
```

Do you understand how the above command is constructed? Now type **`mydata`** to check the new column has been included. You'll see there is clearly a variation between the observed and the "expected" (average) distribution of monuments. Is this significant? Carry out a Chi-squared test to find out.

```
chisq.test(mydata$monuments, p=mydata$pct_land)
```

In the above function the values read by the argument **p** must add up to 1, which is the reason we denoted the percentages as decimals. If the numbers in the original vector **pct_land** has been denoted as 45, 25 and 30, you could have divided them by 100 within the function (note the extra pair of brackets): **p=(mydata$pct_land/100)**.

Completing the test gives you some information:

```
> chisq.test(mydata$monuments, p=mydata$pct_land)

     Chi-squared test for given probabilities

data:  mydata$monuments
X-squared = 7.5385, df = 2, p-value = 0.02307
```

df means *degree of freedom,* a concept that relates to variation of values within the test. X-squared is the Chi-squared result. If we were doing this on pen and paper, we would now use these numbers to check where this results falls on the *critical values table for* X^2 to obtain the significance level (or p-value) for the results. Since we are doing this on a statistical programme, this step is not necessary for us and R calculates a p-value (probability value) to us directly: p=0.02307, or 2.307%.

This means that we can **reject the null hypothesis** with ~97.7% confidence and conclude that the distribution is statistically significant, as opposed to being a product of random chance. Remember that the accepted threshold for rejecting the null hypothesis is p=0.05.

Of course, correlation does not equal causation, and it is a separate task to interpret the results as to why the distribution might be an indication of historical processes. We might dig deeper into the data with the following test, however.

1.4 Two-sample Chi-squared test

A two-sample Chi-squared test is a test for independence of classification between two types of category. In other words, we can check whether the probability distribution of one variable is affect by the presence of another. A simple but illustrative example that Stephen Shennan uses in *Quantifying Archaeology* (2nd ed 1997, see chapter 7 for an excellent discussion of this test in archaeology) is whether there is a relationship between being buried in a Bronze Age cemetery on the left-hand side or the right-hand side, and the body being male or female. The 87 bodies are divided as follows:

	MALE	FEMALE
RIGHT-HAND	29	14
LEFT-HAND	11	33

A two-sample Chi-squared test shows that the pattern of burials is statistically highly significant ($p<0.001$), and as a simply glance at the table would indicate there is a probable relationship between the observed sex of the bodies and the manner they are buried. The categorical variables *observed sex* and *burial position* are not likely to be independent of each other.

To carry out a similar test in R, lets divide our monuments data by monument type (house, manufacturing site, ritual site) and create a new data frame to look at. Our null hypothesis that *there is no relationship between the type of monument and the soil type they are found on*. The two categorical variables we are looking at here is the type of monument and the type of soil area it is found on.

```
soils <- c("clay", "morainic", "peat")
houses <- c(6, 3, 19)
manu <- c(10, 3, 2)
ritual <- c(5, 1, 3)
pct_land <- c(0.45, 0.25, 0.3)
newdata <- cbind.data.frame(soils, houses, manu, ritual,
pct_land)
```

Type **`newdata`** to look at the table. In this type of Chi-squared test, we need to simplify the cross table, so lets pick the columns with the relevant numbers. We won't need the information on soil area percentages, since we are not looking at expected populations.

```
newdata2 <- data.frame(newdata$houses, newdata$manu,
newdata$ritual)
```

Now you can carry out the two-sample Chi-squared test:

```
chisq.test(newdata2)
```

You should get the below result.

```
> chisq.test(newdata2)

     Pearson's Chi-squared test

data:  newdata2
X-squared = 12.919, df = 4, p-value = 0.01168

Warning message:
In chisq.test(newdata2) : Chi-squared approximation may be incorrect
```

The p-value of 0.01168 indicates that we can reject the null hypothesis. You can see that this further refines our investigation into the data. If we had not been able to reject the null hypothesis, then it would have been unlikely (from a statistical perspective) that the distribution of monuments was related to the interpretation of what these sites were used for. We would have to look for other explanations of historical processes underlying the pattern.

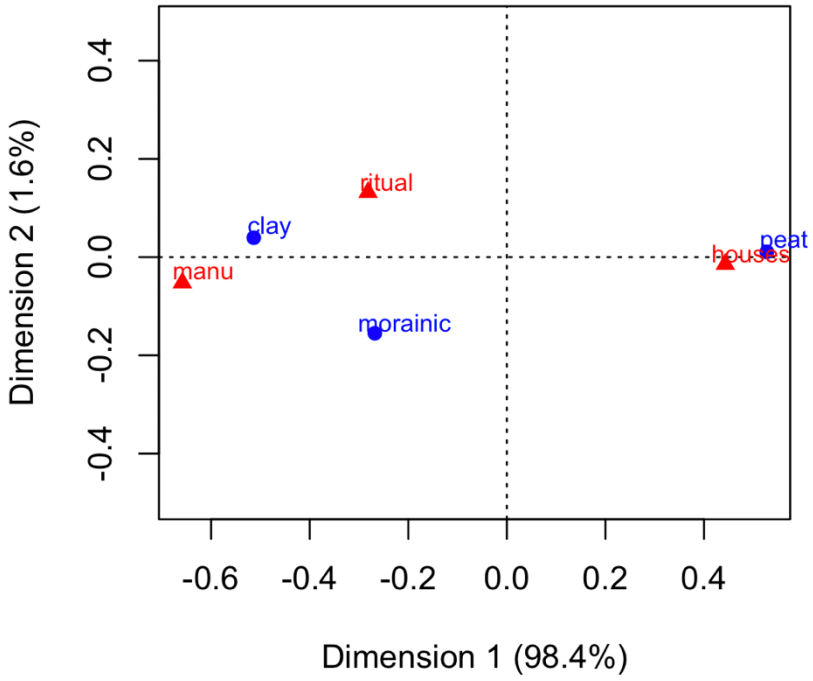
Note, however, the warning message. It is given because our dataset is quite small, and therefore the statistical validity of the pattern is in question. More monuments and more data would allow to make assertions with greater confidence.

2. Correspondence Analysis with Roman Coins

2.1 Introduction

Correspondence analysis is related to the Chi-squared test. Both can be used to study categorical (nominal and ordinal) data presented in a cross table format, with CA modifying and building upon the principles used in the Chi-squared test. You do not need to understand the detailed mathematics underlying CA in order to used basic CA approaches and to carry out these exercises, but for a brief read Carlson 2017 chapter 13 (see *Key bibliography* folder on Moodle, unfortunately not available as an ebook).

In essence, CA reduced complex datasets into a simple graphical display, a biplot, that can be interpreted to study relationships present within the data. Consider the data monument data in section 1.4 above: this has been presented as a cross table with soils as rows and monument types of columns. The soil and monument types are two different classes of categorical variable. If we run this very simple data through CA we can build the following biplot to represent relations in the dataset. The nearness of the “houses” and “peat” dots indicates that these variables are closely associated in the data.



The below exercises will take you through the analysis and Roman hoard data discussed Lockyear 2017 'Simplifying Complexity', so make sure to have read the article first.

2.2 Data

Running CA on R is quite straightforward, and the analytical work comes from interpreting and digging into the results. We will use the package ***ca***, but there are several other R packages with somewhat different functionalities, such as ***FactoMineR*** and ***MASS***. Install ***ca*** using **`install.packages()`** and then load it.

```
library(ca)
```

Then load the Roman hoards data used by Lockyear. The **`read.csv`** function is slightly different from what you have used before. Note that we must create a cross table, which means that the first row in the CSV sheet ('hoards') is not to be assigned as a regular column in the data frame, but rather must be indicated by the **`row.names`** argument to contain the hoard names as the names of the rows. After loading the data, type the name of the object to inspect the contents and to get a sense of what the data looks like. The rows give the names of the coin hoards (by the place where they were recovered) whereas the columns give the names of the mints where coins were minted

```
mydata<-read.csv(file="romanhoards.csv", header=TRUE, sep=",",
row.names="hoards")
```

Use the **`summary()`** function to study the breakdowns by mint. If you want to study he breakdowns by mint, a quick way to do is to use the **`t()`** function to transpose (or "flip") the columns and rows in a data frame. Note that we'll want to delete the 19th row "Total", as it would now render the summary statistics meaningless! Try:

```
t.mydata <- t(mydata)
t.mydata
t.mydata <- t.mydata[-c(19),]
summary(t.mydata)
```

Try some basic data visualisation methods like creating boxplot, bar charts and histograms. Not all many be equally useful or appropriate: think about which of those work and give meaningful results?

2.2 Performing CA

Now let us run CA for the hoards data. We don't want the $Total column and the $uncertain column to be included. They can be excluded by subsetting with the [] operator so that we only use columns numbered 1 to 17.

```
mydata.ca <- ca(mydata[,1:17])
```

You could have also deleted a column entirely from the original data frame with the the command **`mydata$Total <- NULL`** but thanks to subsetting you don't need to do it and we can keep all the original data for reference.

Having run CA, examine the diagnostic statistics with the **`summary()`** function. Study the resulting tables.

```
summary(mydata.ca)
```

Eigenvalues refer to the amount of variation in the data that can be explained by each axis in a plot. Note that under the heading principal inertias, dimensions (*dim*) 1 & 2 add up to 96.6% (see column *cum%*) of the variation in the data, meaning that almost everything about the dataset can be explained just by taking these dimensions into account. This means we can plot the data with confidence into just a 2-dimensional graph (i.e. one with x- and y-axis). This is a deliberately constructed dataset, however, and most archaeological and historical data will not conform to

such level of analytical cleanliness. The * symbols under *scree plot* are just a short graphical shorthand for the values.

Reference Lockyear 2017, pp. 18-21, for a fuller explanation of the data table you are presented with. Make sure you understand what the key columns *mass*, *cor* and *qlt* represent.

2.3 Plotting CA

Next, create a CA plot.

```
plot(mydata.ca)
```

This is called a biplot. The first two dimensions are represented by default: the x-axis gives 61.8% of the variation, and the y-axis 35.8%.

The resulting plot is unfortunately in places a bit hard to read, as the labels fall on top of each other. The are many graphics packages in R that can offer solutions to this and produce nice, publishable quality maps. Some of the CA packages in R also offer nicer graphics. We won't be going into these in this course, but be aware that there is a number of guides to improving visual graphics online. Mike Baxter and Hilary Cool's *Basic Statistical Graphics for Archaeology with R: Life Beyond Excel* is one good guide to statistical visualisation in general that is available for free online as a pdf. I have linked to it in the *Key bibliography* folder.

You can interrogate the plot function in the package ***ca*** with **`?plot.ca()`** and find ways to change the visualisation. Try out some of the arguments presented, for example arrows or changing the colour of the labels or symbols.

Importantly, you can plot only the rows or the columns with the **`what`** argument. Lets plot them side by side:

```
dev.new(device=pdf, height=5, width=10)
par(mfrow=c(1,2), mai=c(1, 1, 1, 1))
plot(mydata.ca, what=c("none","all"), main="Columns")
plot(mydata.ca, what=c("all","none"), main="Rows")
```

In this straightforward example, it is easy to see from the right-hand *Rows* plot that the hoards divide into three groups of two by their composition. The coins in the Jezzine and Nebek hoards, for example, tend to some from the same mints. This is the impression you probably would have received simply looking at the original data frame, and the CA confirms it. Similarly (looking at the left-and *Columns* plot), coins

from Antioch and Alexandra tend to come from the same hoards. And the combined plot showing symbols for both columns and rows indicates which hoards and which mints have a relationship in the data.

2.4 Dimensions

As noted above, the first two dimensions nearly represent almost all of the data. However, in actual research it is a good idea to at least look into the lower dimensions, so as to check whether there important information that would be otherwise missed. The argument for setting the dimensions in the graph is **`dim`** and the below command allows you to construct the plot with the first and the third dimension.

```
dev.new(device=pdf)
plot(mydata.ca, dim=c(1,3))
```

Examine how changing the "angle" changes the plot. Of course, with these two dimensions combined you are now accounting for a smaller proportion of the variation in the data than before.

3. Correspondence Analysis with Burial Data

In the final portion of this week's exercise, we will look at using CA with a somewhat more complex dataset. This comes from an instructional data compiled by Mike Baxter for *Notes on Quantitative Archaeology and R* (2015), which is a good academic manual for a variety of statistical and data exploration methods. It, alongside the data, has been made available for free on mikemetrics.com and on Mike Baxter's academia.edu page.

3.1. Data

First, load the data and look at the data frame.

```
baxter<-read.csv(file="baxter_burials.csv", header=TRUE,
sep=",", row.names="pottery")
baxter
```

This represents 52 Iron Age tombs (designated with the letters **a** to **p** in columns), from which various amounts of different pottery types have been recovered (pottery

types **1** to **52**). You can create some basic statistical graphs to examine the data, as before.

Then run CA, naming the CA object as **baxter.ca**, and obtain the diagnostic information. Look above for the relevant commands and functions. Note that as this dataset is a more complex than the previous one, the first two dimensions account only for 44.9% of the variation, but this is still quite good for CA.

3.2 Dealing with outliers

Now plot the CA.

```
plot(baxter.ca)
```

The burials should appear as labelled red triangles and the pottery types as blue dots. The overall patterns is rather lop-sided, with burial **c**, associated in particular with pottery types **13**, **25** and **28-29**, appearing towards one corner of the plot. This could be important information, and you would note it down if you were carrying out a formal research analysis of the data. You can examine this relationship by looking at the third dimensions as well, as per the example above.

But we are also interested in patterns in the rest of the data, and the presence of this outlier cramps the plot and makes it harder to interpret. A common approach to CA is to remove such outliers, allowing patterns and relationships in the mass of the data to be examined with greater facility. Having identified the variable you want to remove, you can rerun the CA command.

```
baxter.ca <- ca(baxter[,-c(3)])
```

Can you work out how the row **c** was removed by subsetting?

3.3 Seriation

Now recreate the edited CA plot.

```
plot(baxter.ca)
```

It display a distinctive horse-shoe shape on the graph, a result that is often hoped for in archaeological CA. What you see here is an example of *seriation*. In archaeological applications of CA it is frequently hoped for that this ordering makes a chronological interpretation. Let us assume, for example, that these tombs across a span of time

and contain ceramics vessels manufactured in overlapping but roughly sequentially progressing styles. The plot would capture this information, also suggesting a relative chronology for the tombs.

Replot the data with the Columns and Rows as separate plots, as in section 2.3 above, and examine the distributions.

The first (horizontal) axis is the best approximation of the seriation order. This should be confirmed by other, independent evidence, however, such as evidence backed by radiocardon dates or other archaeological criteria such as, in this case, has been done for the relative chronological ordering of the tombs **g**, **h**, **i** and **j** (though you'll note that when read left to right they are not in alphabetical order).

Finally, bear in mind that while chronological seriation may work if the data is chronologically sensitive, seriation could easily represent other information as well. Composition of object assemblages could be associated with gender or some other non-temporal cultural expression. In your study diary, consider an example of what kind of data seriation could potentially be applied to (not necessarily only in historical and archaeological evidence), as well as other possible applications of the techniques you have learned this week.

Save the data in your workspace. You are now finished with this week's exercise.