

Using R to Detect Changes Within Individuals

Michael A. Covington
Institute for Artificial Intelligence
The University of Georgia

2011 December 9

Introduction

In what follows I will demonstrate statistical analysis of an experiment that looks for a correlation between two measurements on each of a set of texts, using Excel to edit and prepare the data and R to analyze it.

My “Using R” tutorials partly repeat each other’s content so that each one will be complete by itself.

About R

R is freeware, downloadable from www.r-project.org. It is good for calculation, matrix arithmetic, statistics, and various kinds of machine learning.

Suggested reading: Peter Dalgaard (2002) *Introductory Statistics with R*. Berlin: Springer.

R is derived from an earlier language called S. To learn more about its advanced statistical functions see Venables and Ripley, *Modern Applied Statistics with S* (which includes R). To learn more about R as a programming language, see Braun and Murdoch, *A First Course in Statistical Programming with R*.

R will mystify you until you learn something about its data types and syntax. The data types are:

Numbers

Strings

Vectors (1-dimensional arrays of numbers or strings)

Factors (vectors of discrete named values, like an array of enums in C)

Lists (sequences whose components are named, not numbered)

Matrices (2-dimensional arrays)

Arrays (with any number of dimensions)

Data frames (2-dimensional arrays with named columns, for statistical data)

The syntax of R is unusual. Note that:

There are no semicolons between statements.

Vectors and arrays are indexed from 1, not 0.

All names are case-sensitive.

`<-` is assignment

`.` within a name is just an ordinary character

`$` picks out parts of lists or frames (such as `a$b`, which means the part of `a` named `b`)

`#` makes the rest of the line a comment

`%` marks special operators, such as `/%` (integer division), `%%` (matrix multiplication)

Vectors (sequences) have to be created, usually with the function `c()`.

Compare:	In Prolog:	<code>X = [1,2,3]</code>
	In R:	<code>X <- c(1,2,3)</code>

The help system is accessed with commands such as `help(t.test)` (for finding out about the function named `t.test`).

About data files

Excel's own file formats, `.xls` and `.xlsx`, are generally not understood by other software. Instead, we exchange spreadsheets with other software by using two other formats, **tab-delimited text** and **comma-separated values (CSV)**. These are text files that use, respectively, tab characters (Unicode 0009) and commas to separate the columns.

R reads tab-delimited text with the function `read.delim()` and CSV with `read.csv()`. In each case what you get is a data frame with the first line used as column labels.

The experiment

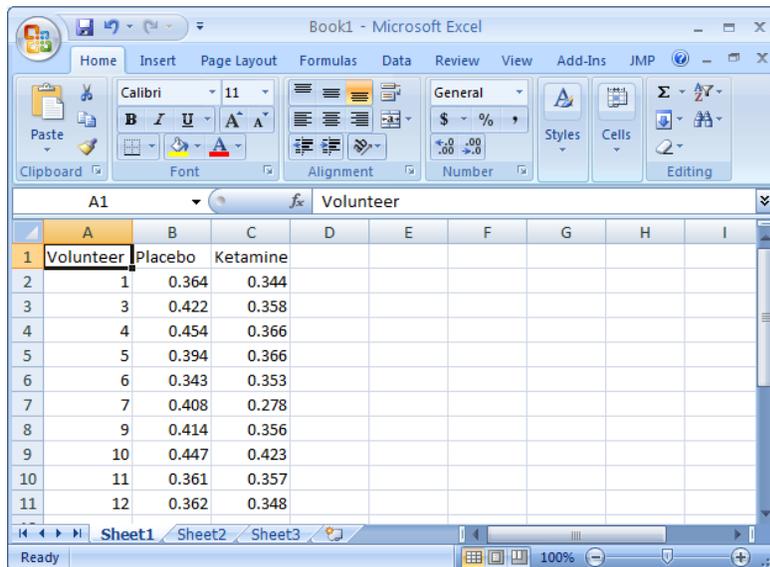
This example uses data from Cati Brown et al., *Reduced Idea Density in Speech as an Indicator of Schizophrenia and Ketamine Intoxication* (poster, ICOSR 2005). The idea density in the

speech of eleven volunteers was measured with and without the drug ketamine. The experiment was done at Cambridge University.

The question is whether ketamine affects idea density.

Preparing the data file

I entered the data in Excel as shown here and then saved it as a **tab-delimited text file**. Note that the first row consists of column labels.



Volunteer	Placebo	Ketamine
1	0.364	0.344
3	0.422	0.358
4	0.454	0.366
5	0.394	0.366
6	0.343	0.353
7	0.408	0.278
9	0.414	0.356
10	0.447	0.423
11	0.361	0.357
12	0.362	0.348

The research question

We are interested in the columns called **Placebo** and **Ketamine**. Both are measures of idea density. The question is, on the whole, is there the same kind of change, from one column to the other, for each individual?

Loading the data into R

This is how to read the whole data set into R and store it in a variable called `d`:

```
> d <- read.delim("c:\\Users\\mc\\Desktop\\ketadata.txt")
```

Note the doubled backslashes. If you want R to put up an open file dialog box and let you pick a file, you can do it in this somewhat clumsy way:

```
> require(tcltk) # do this just once, to bring in the tcltk library
> d <- read.delim(tclvalue(tkgetOpenFile()))
```

To confirm that it was read in correctly, we can immediately display it:

```
> d
  Volunteer Placebo Ketamine
1          1  0.364  0.344
2          3  0.422  0.358
3          4  0.454  0.366
4          5  0.394  0.366
5          6  0.343  0.353
6          7  0.408  0.278
7          9  0.414  0.356
8         10  0.447  0.423
9         11  0.361  0.357
10        12  0.362  0.348
```

The columns we are interested in are **Placebo** and **Ketamine**.

We can look at the individual columns as vectors (sequences of values):

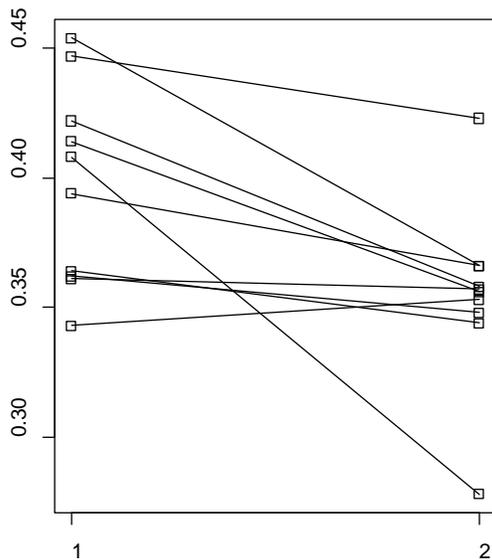
```
> d$Placebo
[1] 0.364 0.422 0.454 0.394 0.343 0.408 0.414 0.447 0.361 0.362
> d$Ketamine
[1] 0.344 0.358 0.366 0.366 0.353 0.278 0.356 0.423 0.357 0.348
```

Visualizing the changes

We can visualize the changes by using the following two commands:

```
> stripchart(list(d$Placebo,d$Ketamine),vertical=T)
> for (i in 1:length(d$Placebo)) { segments(1,d$Placebo[i],2,d$Ketamine[i]) }
```

The first command creates a chart consisting of a pair of one-dimensional scatter plots. The second command adds line segments joining the corresponding points:



Sure enough, almost all of the volunteers showed a dramatic drop in idea density.

Significance testing

Is there enough evidence to confirm our impression that ketamine makes idea density go down?

What we need here is a **paired *t*-test**. We aren't just comparing the two groups (as in a plain *t*-test) – we're comparing the two values *for each individual*. Here's how it's done:

```
> t.test(d$Placebo, d$Ketamine, paired=T)
```

Paired t-test

```
data: d$Placebo and d$Ketamine
t = 3.0985, df = 9, p-value = 0.01275
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01133685 0.07266315
sample estimates:
mean of the differences
      0.042
```

The important number here is the p -value, which is the probability of getting at least this much difference in the samples even if there is no real difference in the populations from which the samples are drawn. (That is, even if ketamine had no real effect.)

In the biosciences, we generally conclude that an effect is real when $p < 0.05$. Here p is much lower than that, so we conclude that ketamine does indeed reduce idea density.

Was that the right test?

If we had known in advance that idea density could only decrease, not increase, we could have done a one-tailed t -test and gotten an even lower p -value. But in this case that is not appropriate.

The t -test assumes that the two sets of numbers have bell-shaped (normal) distributions. We may not want to assume that. In that situation, we do a Wilcoxon test:

```
> wilcox.test(d$Placebo, d$Ketamine, paired=T)
```

```
Wilcoxon signed rank test
```

```
data: d$Placebo and d$Ketamine
```

```
V = 53, p-value = 0.005859
```

```
alternative hypothesis: true location shift is not equal to 0
```

Again we get a highly significant p -value.