Review: Confidence Intervals and Hypothesis Tests

Vineyard Soil Data: Potassium in 2004 compared to 2007

Background
Soil Potassium was measured at 10 randomly sampled locations in a vineyard in 2004 and again in 2007.

- Has soil potassium (ppm) changed over time?
- If so, how much has it changed?

We’ll begin by assuming we essentially have two independently chosen random samples of soil. One sample from 2004 and another from 2007. (As a matter of fact, the data was collected differently and we shall see the impact of this on the analysis later).

Import and Graph the Data
First, we read the data in R using the function `read.csv`. The data is a csv file (comma separated value). We will regularly be reading in csv files. These files can be created in Excel or in a basic text editor like notepad. The first row of these files contains the names of the columns with each name separated by a comma. The subsequent rows contain the data, again with each value separated by a comma. Missing values can be left blank. This imported data object (a data frame) is saved under the local name soil.

```r
> soil = read.csv(file.choose())
```

Now view the first few rows of the data using the `head` command.

```r
> head(soil)
   Year Slope     ID  pH  EC  CEC  OM K.perc Mg.perc Ca.perc Na.perc Ca.Mg N03 PO4  K  Mg  Ca  S  Zn  Mn  Fe  Cu  WHC
1 2004  bottom blue 8.2 0.2 33.1 3.9 0.7  3.5 95.4  0.4  27.3  8 15 88 142 6334 62 0.9 3 4 3.2 17.1
2 2004  bottom orange 8.1 0.3 38.7 3.4 0.5  9.1 90.1  0.9  9.9  8 12 82 428 6991 67 0.3 2 3 1.6 17.1
3 2004  bottom red 7.8 0.3 38.5 3.4 0.7  7.8 91.1  0.4 11.7  8 15 100 365 7025 68 0.6 2 3 1.6 17.1
4 2004  bottom white 8.0 0.3 34.6 3.8 0.7  7.7 91.3  0.3 11.9  8 16 93 326 6334 61 0.4 3 5 2.6 17.1
5 2004  bottom yellow 7.9 0.4 33.0 3.2 0.8  7.3 91.6  0.3 12.5 13 14 105 293 6049 60 0.6 3 4 0.4 17.1
6 2004  top blue 7.9 0.3 41.1 4.4 1.3 12.4 86.0  0.3 6.9 13 19 216 619 7087 66 1.9 2 4 4.8  8.5
```

Here’s some information about the dimensions of the data frame (`dim`) as well as a command to get the names of all of the columns (`names`).

```r
> dim(soil)
[1] 6 35
> names(soil)
[1] "Year"  "Slope"  "ID"   "pH"  "EC"  "CEC"  "OM"   "K.perc"  "Mg.perc"  "Ca.perc"  "Na.perc"  "Ca.Mg"  "N03"  "PO4"  "K"   "Mg"   "Ca"   "S"   "Zn"   "Mn"   "Fe"   "Cu"   "WHC"
```
> dim(soil)
[1] 20 23

> names(soil)
[1] "Year" "Slope" "ID" "pH" "EC" "CEC" "OM"
[8] "K.perc" "Mg.perc" "Ca.perc" "Na.perc" "Ca.Mg" "N03" "PO4"
[15] "K" "Mg" "Ca" "S" "Zn" "Mn" "Fe"
[22] "Cu" "WHC"

A `stripchart` produces side-by-side dotplots. Here’s the basic syntax:

> stripchart(K ~ Year, data = soil)

or

> with(soil, stripchart(K ~ Year))

The plot below is made with the following options.

> stripchart(K ~ Year, data = soil, vertical = TRUE, pch = 16,
+     xlim = c(.5, 2.5), xlab = "Year", ylab = "Potassium (ppm)",
+     method = 'jitter')

The 2004 data appears rather skewed. Let’s look at a histogram and a normal probability plot of that data. We can examine a subset of the data by using the `subset` command. Here we only look at data for which the variable Year is 2004.

> with(subset(soil, subset = Year == 2004), hist(K, main = "2004 Data Only"))
> with(subset(soil, subset = Year == 2004), qqnorm(K))
Produce Basic Summary Statistics

Here are some ways to compute the mean (\texttt{mean}), sd (\texttt{sd}), and sample size (\texttt{length}) of all of the K data.

\begin{verbatim}
> with(soil, mean(K))
[1] 216.2
> mean(soil$K)
[1] 216.2
> sd(soil$K)
[1] 147.0759
> length(soil$K)
[1] 20
\end{verbatim}

Now we get the summaries for each year separately, which is what we really want to do. The command \texttt{tapply} allows us to repeatedly apply a function to data separately for each group. The syntax is \texttt{tapply(response\_variable, grouping\_variable, function)}

\begin{verbatim}
> group.mean = with(soil, tapply(K, Year, mean))
> group.sd = with(soil, tapply(K, Year, sd))
> group.n = with(soil, tapply(K, Year, length))
> group.se = group.sd/sqrt(group.n)
> rbind(group.mean, group.sd, group.se, group.n)
\end{verbatim}

\begin{verbatim}
   2004      2007
group.mean 118.90000 313.50000
group.sd    47.47034 149.58034
group.se    15.01144  47.30146
group.n     10.00000  10.00000
\end{verbatim}

Barplot

A barplot of the means with SE intervals is a common plot. Here’s how you can make one in R.

\begin{verbatim}
> tmp = barplot(group.mean, ylim = c(0,400), ylab = "Potassium (ppm")
\end{verbatim}
Recall the one sample t-interval

Here I’m not considering the entire dataset, but rather each year separately. This is not the appropriate analysis, but useful as a starting point for discussion. A t-test or t-interval is computed using the function t.test.

> with(subset(soil, subset = Year == 2004), t.test(K))

One Sample t-test

data:  K
t = 7.9206, df = 9, p-value = 2.397e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 84.94176 152.85824
sample estimates:
mean of x
118.9

> with(subset(soil, subset = Year == 2007), t.test(K))

One Sample t-test

data:  K
t = 6.6277, df = 9, p-value = 9.616e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 206.4967 420.5033
sample estimates:
mean of x
313.5

Here’s a shortcut for this using tapply.
> with(soil, tapply(K, Year, t.test))

**Conduct two-sample t-test**

This test is appropriate for two-independently chosen random samples. The data should come from a normal population if the sample sizes are small.

> with(soil, t.test(K ~ Year))

    Welch Two Sample t-test

    data:  K by Year
    t = -3.9213, df = 10.795, p-value = 0.002474
    alternative hypothesis: true difference in means is not equal to 0
    95 percent confidence interval:
    
    -304.0808  -85.1192
    sample estimates:
    mean in group 2004 mean in group 2007
    118.9      313.5

**Conduct Wilcoxon-Mann Whitney test**

If we are uncomfortably with the small sample sizes and possible non-normality of the data, we can use a non-parametric test such as the Wilcoxon-Mann-Whitney.

> with(soil, wilcox.test(K ~ Year))

    Wilcoxon rank sum test

    data:  K by Year
    W = 9, p-value = 0.00105
    alternative hypothesis: true location shift is not equal to 0

**Conduct paired t-test**

In actuality, the data was collected at the same sites for each year (noted by the variables ID and slope). Thus there aren’t really two independently chosen random samples of size 10 each, but rather one sample of 10 locations measured twice: once in 2004 and again in 2007. *(Note: an easy way to grab a variable from a data frame is to use the $ notation. The following are equivalent: dataframe$variable and subset(dataframe, select = variable).)*

> soil.2004 = subset(soil, subset = Year == 2004)
> soil.2007 = subset(soil, subset = Year == 2007)
> t.test(soil.2004$K, soil.2007$K, paired = TRUE)

    Paired t-test

    data:  soil.2004$K and soil.2007$K
    t = -5.1558, df = 9, p-value = 0.0005984
    alternative hypothesis: true difference in means is not equal to 0
    95 percent confidence interval:
    
    -279.9820  -109.2180
    sample estimates:
    mean of the differences
    -194.6