The bootstrap method

Data

For this lab, we will use the dataset sphagnum_cover.csv, from the paper:

> Maanavilja, L., Kangas, L., Mehtätalo, L. and Tuittila, E.-S. (2015), Rewetting of drained boreal spruce swamp forests results in rapid recovery of Sphagnum production. J Appl Ecol, 52: 1355-1363. doi:10.1111/1365-2664.12474)

This dataset contains measurements of the percentage cover of sphagnum moss (*sphcover*) for 36 boreal swamps divided into three types (*habitat*): *Dr* = drained, *Re* = rewetted et *Un* = undrained.

```
cover <- read.csv("sphagnum_cover.csv")
str(cover)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ site    : chr  "AmLuxx" "EvLuPa" "EvLuVK" "HeLuxx" ...
##  $ habitat : chr  "Un" "Un" "Un" "Un" ...
##  $ sphcover: num  35.3 56.2 46.6 56 54.3 ...
```

1. Estimation of mean cover for drained swamps

a) From the dataset, extract the *sphcover* values for drained swamps. Calculate the mean percentage cover and its standard error from the usual formula (based on standard deviation and sample size). Finally, calculate the 95% confidence interval based on the t distribution:

$$(\bar{x} + t_{(n-1)0.025} s_{\bar{x}}, \bar{x} + t_{(n-1)0.975} s_{\bar{x}})$$

Reminder: The function `qt(p, df)` gives the quantile corresponding to a given cumulative probability `p`, for a t distribution with `df` degrees of freedom.

b) Simulate 10,000 bootstrap samples for the mean calculated in a). What is its standard error according to the bootstrap? Does this statistic appear biased?

c) How does the bootstrap distribution differ from a normal distribution? To answer this question, it may be useful to draw a quantile-quantile plot (in the code below, `res` is the result of the bootstrap):

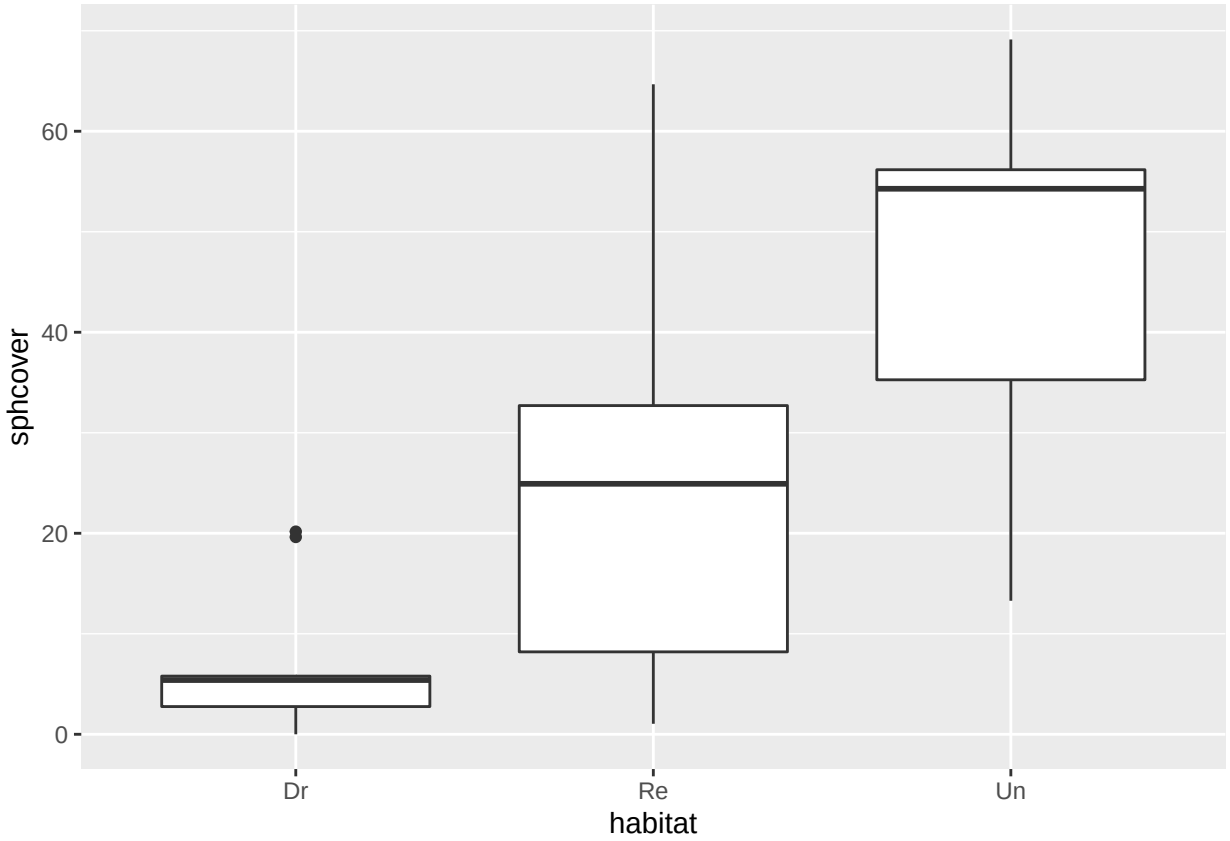
```
qqnorm(res$t)
qqline(res$t)
```

d) Calculate the 95% confidence interval of the mean using the BCa method. How does it differ from the one calculated in a) using the classical formula? Can you explain this difference based on the result in c)?

2. Estimation of differences between habitats

a) Here is the distribution of *sphcover* values in each habitat type.

```
library(ggplot2)
ggplot(cover, aes(x = habitat, y = sphcover)) +
    geom_boxplot()
```



What are the assumptions of a classical ANOVA model that would describe sphagnum moss cover by habitat type? Do these assumptions appear to be met here?

b) Fit the `sphcover ~ habitat` linear model to the `cover` dataset. See the summary of the model results with the `summary` function and the confidence intervals of the coefficients with the `confint` function. What is the interpretation of each coefficient? Are the confidence intervals plausible?

c) Create a function with arguments `x` and `i`, which fits the linear model in b) by replacing the original data set (`data = cover`) with `data = x[i, ]`, and then returns the coefficients of the model with the function `coef`. Then, apply `boot` to the `cover` dataset with the function created and perform 10,000 replicates.

Notes

- When the first argument of 'boot' is a dataset, it is the rows of that dataset that are resampled.
- Since the statistic calculated by the function has several values (each of the coefficients), the `t` element of the `boot` result is a matrix rather than a vector. The columns of this matrix correspond to each of the coefficients in order. You can calculate a statistic for each column with the function `apply`, e.g. `apply(res$t, 2, mean)`. Here, `2` indicates to calculate the `mean` function per column (`1` would mean per row).

d) The application of the bootstrap in c) resamples across all rows, so that the number of observations in each habitat type varies from sample to sample. If it is preferable to consider these numbers as fixed quantities,

habitat types can be defined as strata by adding the argument `strata = as.factor(cover$habitat)` to the `boot` function. (The conversion of the variable `habitat` to a factor is necessary here.)

Repeat the analysis in c) with stratified resampling and compare the standard errors obtained for each coefficient.

e) Calculate the confidence interval for the `habitatUn` coefficient according to the result of the bootstrap in d). Note that you need to add the argument `index = 3` to the function `boot.ci` to tell R to calculate the interval for the 3rd coefficient.

f) Finally, we will resample the model residuals.

- Fit a linear model as in b), then add to the dataset `cover` a column for the model's `fitted` values.
- Write a function that creates a new dataset by adding a resampled vector `x[i]` to the expected values to produce a new response variable, then fits a model with this new response variable to the habitat.
- Simulate 10,000 samples with the `boot` function, using as arguments (1) the residuals vector (`residuals`) of the model and (2) the function created above. Do not specify strata. Recalculate the mean, standard error and 95% confidence interval of the coefficients.

Is resampling the residuals a good choice for these data?