1

The Central Limit Theorem states that the sum $Z = X_1 + \ldots + X_n$ of n independent identically distributed (iid) random variables is approximately normally distributed for large n. In fact, if $X_1, \ldots X_n$ have mean μ and variance σ^2, then the standardised sum $Z_s = ((X_1 - \mu) + \ldots + (X_n - \mu))/(\sqrt{n\sigma^2})$ has a standard normal distribution $N(0, 1)$ for large n. Let us illustrate this result in the case where X_i follows the uniform distribution on the interval [-1,1].

Generate a random sample x of $n = 500$ independent draws from the uniform distribution on the interval [-1,1]. Plot the result in a histogram.

Solution (./04_R-ws-soln.html#-1%0A)

Write a function `func` that takes n as input, generates a random sample x of n independent draws from the uniform distribution on the interval [-1,1] and returns the sum of these draws.

Use this function to generate $N = 1000$ such sums for n = 500, that is, generate a sample of size $N = 1000$ for the variable Z. Plot a histogram of the result. Use `qqnorm` to check that the sums follow a normal distribution.

Now adjust the function `func` so that it returns a list containing the sample variance as well as the sum for each sample x.

Use the adjusted function to generate a sample of size $N = 1000$ for the standardised sum Z_s. Note that the mean μ for the distribution of X_i is 0 and the variance σ^2 can be approximated by the sample variance for each x. Use a histogram of the result and `qqnorm` as before to check that the standardised sum Z_s follows a standard normal distribution.

2

Data in `R` are often held in data frames. A data frame is a table in which each column is a vector containing the values of one variable. The variables can be of either numeric, factor or character type.

Solution (./04_R-ws-soln.html#-2%0A)

One of the built in datasets in R is called `trees`. Use the command `dat <- trees` to create a new data frame with this data.

Have a look at the data using, for example, `summary(dat)` and `pairs(dat)`.

Fit a simple linear model with response variable `Volume` and predictor variable `Girth`, storing the results as an object `mod`.

Use `plot(mod)` and `summary(mod)` to assess the goodness of fit of the model. Does `Girth` predict `Volume` well?

From the formula for the volume of a cylinder, we may expect the variable `Girth`^{2} to be a better predictor for `Volume`. Add this variable as a new column `Girth2` to the data frame dat.

Fit a linear model as before but with `Girth` replaced by `Girth2`. Did this improve the model?