Introduction to R

Holger Sennhenn-Reulen[a], Nordwestdeutsche Forstliche Versuchsanstalt (NW-FVA)

July 07, 2021 (Version 0.1)

Contents

[a]Private webpage: uncertaintree.github.io

1 Introduction: What is R?

...

2 Data

We will use two data-sets within this course.

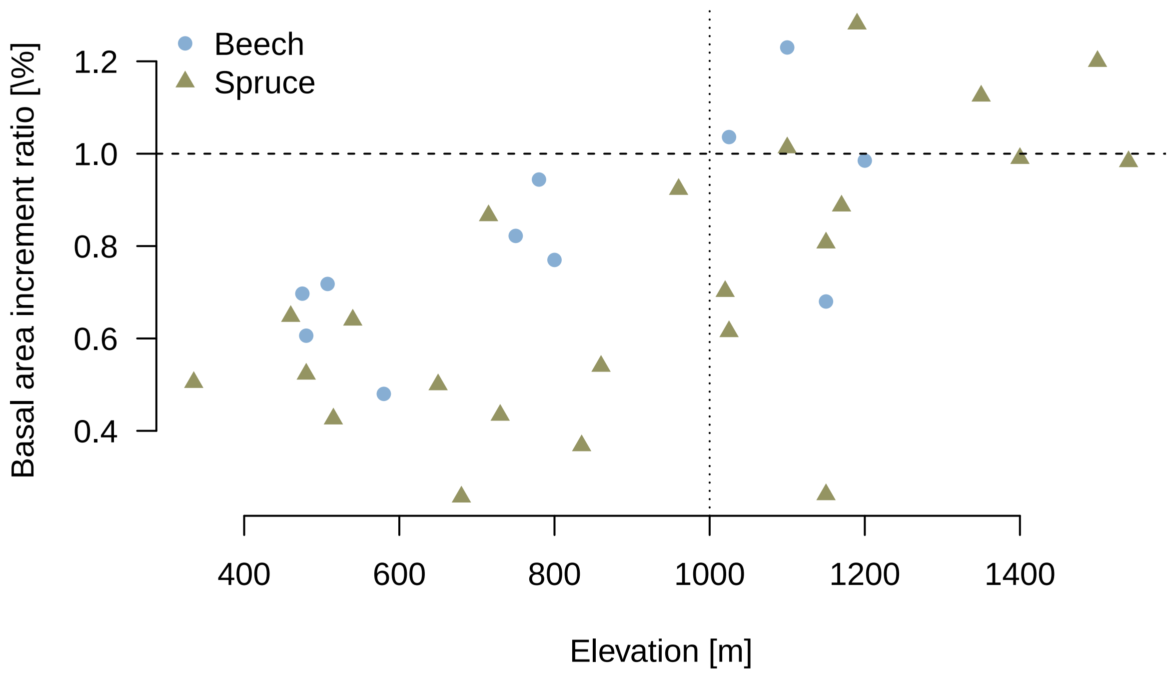
Drought

```r
bair <- c(.505, .648, .523, .426, .64, .5, .257, .866, .434, .368, .54, .923, .702,
          .615, 1.013, .807, .262, .887, 1.281, 1.125, .99, 1.2, .983, .697, .606,
          .718, .48, .822, .944, .77, 1.036, 1.23, .68, .985)
elev <- c(335, 460, 480, 515, 540, 650, 680, 715, 730, 835, 860, 960,
          1020, 1025, 1100, 1150, 1150, 1170, 1190, 1350, 1400, 1500, 1540,
          475, 480, 507.5, 580, 750, 780, 800, 1025, 1100, 1150, 1200)
species <- c("Spruce", "Spruce", "Spruce", "Spruce", "Spruce", "Spruce",
             "Spruce", "Spruce", "Spruce", "Spruce", "Spruce", "Spruce", "Spruce",
             "Spruce", "Spruce", "Spruce", "Spruce", "Spruce", "Spruce", "Spruce",
             "Spruce", "Spruce", "Spruce", "Beech", "Beech", "Beech", "Beech",
             "Beech", "Beech", "Beech", "Beech", "Beech", "Beech", "Beech")
drought <- data.frame(bair = bair,
                 elev = elev,
                 species = species)
summary(drought)
```

```
##       bair             elev          species
##  Min.   :0.2570   Min.   : 335.0   Beech :11
##  1st Qu.:0.5272   1st Qu.: 597.5   Spruce:23
##  Median :0.7100   Median : 847.5
##  Mean   :0.7489   Mean   : 888.3
##  3rd Qu.:0.9732   3rd Qu.:1150.0
##  Max.   :1.2810   Max.   :1540.0
```

Plot

```r
pch_here <- c(16, 17)
col_here <- colorspace::diverging_hcl(n = 20, pal = "Lisbon")[c(4, 16)]
plot(drought$elev, drought$bair, las = 1, bty = "n",
     xlab = "Elevation [m]", ylab = "Basal area increment ratio [\\%]",
     pch = pch_here[1 + (drought$species == "Spruce")],
     col = col_here[1 + (drought$species == "Spruce")])
abline(h = 1, lty = 2)
abline(v = 1000, lty = 3)
legend("topleft", pch = pch_here, col = col_here, legend = c("Beech", "Spruce"),
       bg = "white", bty = "n")
```



A Simple Model

```
m <- lm(bair ~ elev*species, data = drought)
summary(m)
```

```
##
## Call:
## lm(formula = bair ~ elev * species, data = drought)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58062 -0.07711  0.01943  0.14246  0.41573
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.969e-01  2.035e-01   1.951   0.0605 .
## elev                  5.201e-04  2.403e-04   2.164   0.0386 *
## speciesSpruce        -2.056e-01  2.397e-01  -0.858   0.3979
## elev:speciesSpruce    4.619e-05  2.723e-04   0.170   0.8664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2106 on 30 degrees of freedom
## Multiple R-squared:  0.463,  Adjusted R-squared:  0.4093
## F-statistic: 8.621 on 3 and 30 DF,  p-value: 0.0002811
```

3 Section 1

Outcomes

Definition (Outcome): The *outcome* of a random variable X is a value that the random variable might (in theory/model) take on when applying a measurement method.

3.1 Tree species

3.1.0.1 Example (Tree species)

Random selection of a single tree from the last German National Forest Inventory data-base builds a random variable X – for characteristic *tree species group*[i] – with sample space Ω:

$$\Omega = \{\text{oak}, \text{beech}, \text{odtwlle}, \text{odtwsle}, \text{spruce}, \text{fir}, \text{douglas fir}, \text{pine}, \text{larch}\}\,.$$

All possible ω are:

$$\omega = \{\text{oak}\}\,, \omega = \{\text{beech}\}\,, \ldots, \omega = \{\text{larch}\}\,.$$

Random variables are actually only truly helpful if we can define them such that mathematical manipulations can be applied on them as we do with real numbers. So instead of X, we better define eight indicators [Kadane, 2020] $X_1, \ldots, X_8$ with samples spaces $\Omega_1 = \{\text{not oak}, \text{oak}\}$, $\Omega_2 = \{\text{not beech}, \text{beech}\}$, …, $\Omega_8 = \{\text{not larch}, \text{larch}\}$:

$$X_1 = \begin{cases} 0, & \omega = \{\text{not oak}\} \\ 1, & \omega = \{\text{oak}\} \end{cases}, \; X_2 = \begin{cases} 0, & \omega = \{\text{not beech}\} \\ 1, & \omega = \{\text{beech}\} \end{cases}, \; \ldots, \; X_8 = \begin{cases} 0, & \omega = \{\text{not larch}\} \\ 1, & \omega = \{\text{larch}\} \end{cases},$$

and thus with outcomes $x_j = 0$, or $x_j = 1$, $j = 1, \ldots, 8$.

For repeated random experiments, such outcomes are stored in the form of a vector – columns of a data-set in R.

3.1.0.2 Example (Tree species) continued

Random selection of $n = 10$ trees results in a vector:

$$\mathbf{x}_1 = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)^\top$$

with $x_{1,1} = 0$, $x_{1,2} = 0$, $x_{1,3} = 1$, …, $x_{1,10} = 0$.

3.1.0.3 Example (Drought)

The *drought* data contains measures on the ratio of this year's basal area increment (BAI) in relation to last year's BAI: % So BAI ratios X:

$$X = \frac{\text{BAI this year}}{\text{BAI last year}}$$

BAI measures can be positive and negative[ii], but in order to avoid $-1 = \frac{-1}{1} = \frac{1}{-1}$, we take a 'growing process definition' [iii] by which basal area increments become strictly positive. Consequently, the sample space of X is the positive real line, $\Omega = \mathbb{R}_+$, with samples such as:

$$\omega = \{1.37\}\,, \omega = \{0.93\}\,, \ldots$$

and positive real valued outcomes such as:

$$x = (1.37, .93, \ldots)^\top$$

[i] odtwlle: other deciduous trees with long life expectancy, odtwsle: other deciduous trees with short life expectancy
[ii] https://www.sciencedirect.com/science/article/pii/S0378112720314948#t0005
[iii] http://www.personal.psu.edu/agl/Johnson&Abrams%20TRACE.pdf

References

J.B. Kadane. *Principles of Uncertainty*. Chapman and Hall/CRC, 2020. ISBN 9781138052734.