Table of Contents

Section: Diversity analysis

Rarefaction

Theory R functions **Examples** Exercise

Example 1: Comparing forest diversity along elevation standardized to sample area, number of individuals and sample completeness

We will use vegetation data from One-hectar plots in different forest types across Taiwan, containing the survey of woody species at seven localities sampled in different elevations in Taiwan. At each locality, a 1-ha plot has been established, and within the plot, 25 10×10-m subplots have been sampled on an even grid (since there are gaps between the subplots, in total 25×0.01-ha = 0.25-ha area was surveyed within each locality; see data description for details). Dataset is prepared in two forms: abundance data (`hp.abund`) contains numbers of individuals for each species at each locality; incidence-based data (`hp.incid`) contains incidences of each species at each locality (incidence is the presence of species in a subplot made within each locality; each species at the locality can incidence number up to 25, i.e. occurring in up to 25 subplots).

Our aim will be to compare diversities of forest vegetation in different elevation. For this, we need to standardize data to a common base. The original data are standardized to the area (at each locality, the total of 0.25 ha was surveyed), which is a common approach for vegetation ecologists. Another option is to standardize data to the same number of individuals (this is possible in case of abundance-based data) or the same coverage (this is possible for both abundance- and incidence-based data). Let's see the result of all three options.

First, upload the dataset:

```
hp.abund <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/hp.abund.txt')
hp.incid <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/hp.incid.txt')
```

Both `hp.abund` and `hp.incid` are data frames, with species in rows and localities (1-ha plots) in columns; the cells are filled either by numbers of individuals of given species (`hp.abund`) or the number of incidences (out of 25) of each species (`hp.incid`). In case of `hp.incid`, the first row additionally contains the information about the number of subplots for which incidences were recorded (25 in case of all localities); this is necessary for the calculation of incidence-based rarefaction (see further).

For all calculations in this exercise, we will use the library `iNEXT`, developed by the team of prof. Anne Chao (Tsing-Hua University, Hsinchu, Taiwan). You may need to install the package from CRAN first if you haven't used it before:

```
# install.packages ('iNEXT')
library (iNEXT)
```

Before the analysis, let's oversee the data in each of the data frames first. The package `iNEXT` offers function `DataInfo` for that:

```
DataInfo (hp.abund, datatype = 'abundance')
```

```
  site    n S.obs     SC f1 f2 f3 f4 f5 f6 f7 f8 f9 f10
1   FT  232    30 0.9658  8  8  1  1  0  1  3  1  2   0
2  YLN 1056    59 0.9924  8 11  5  2  3  1  1  3  1   1
3   LJ  838    40 0.9905  8  4  2  1  1  1  1  0  0   3
4   WJ 1731    60 0.9919 14  3  4  0  1  0  1  0  1   3
5  YYH 1551    34 0.9916 13  1  0  0  0  0  0  1  0   1
6   PL  394    29 0.9823  7  4  3  0  1  0  1  0  0   0
7   GY  362    20 0.9890  4  4  1  2  1  0  2  0  0   0
```

```
DataInfo (hp.incid, datatype = 'incidence')
```

```
  site  T   U S.obs     SC Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10
1   FT 25 140    30 0.9396  9  7  2  2  1  2  1  1  2   1
2  YLN 25 425    59 0.9761 11 11  5  4  1  4  2  1  3   2
3   LJ 25 264    40 0.9627 10  2  2  1  5  2  2  2  3   2
4   WJ 25 564    60 0.9738 15  3  6  0  1  2  3  1  3   2
5  YYH 25 313    34 0.9587 13  1  0  0  0  1  1  2  3   0
6   PL 25 186    29 0.9533  9  4  1  0  2  1  0  1  2   3
7   GY 25 134    20 0.9490  7  2  2  2  0  2  0  0  0   0
```

The meaning of the variables in the output of `DataInfo` is the following:

- for `datatype = "abundance"`: `site` is the abbreviation of the locality (e.g. FT is Feng-Tien, see the data description), n is the number of individuals, `S.obs` is the observed number of species in the locality, SC is sample coverage (estimated values of sampling completeness for each locality, between 0 and 1). The values in columns `f1`, `f2`, `f3` ... `f10` are numbers of species in the locality represented by only 1, 2, 3, ...10 individuals (singletons, doubletons, tripletons etc.)
- for `datatype = "incidence"`: `site` and `S.obs` the same as above; T is the number of plots within each locality (25 subplots in our case), U is the overall number of species incidences within locality (i.e. the sum of incidences of individual species, where incidence = presence of the species in one subplot). `Q1`, `Q2`, `Q3`, ... `Q10` are numbers of species occurring in only 1, 2, 3, ... 10 subplots within each locality (unique species, duplicate species, etc.).

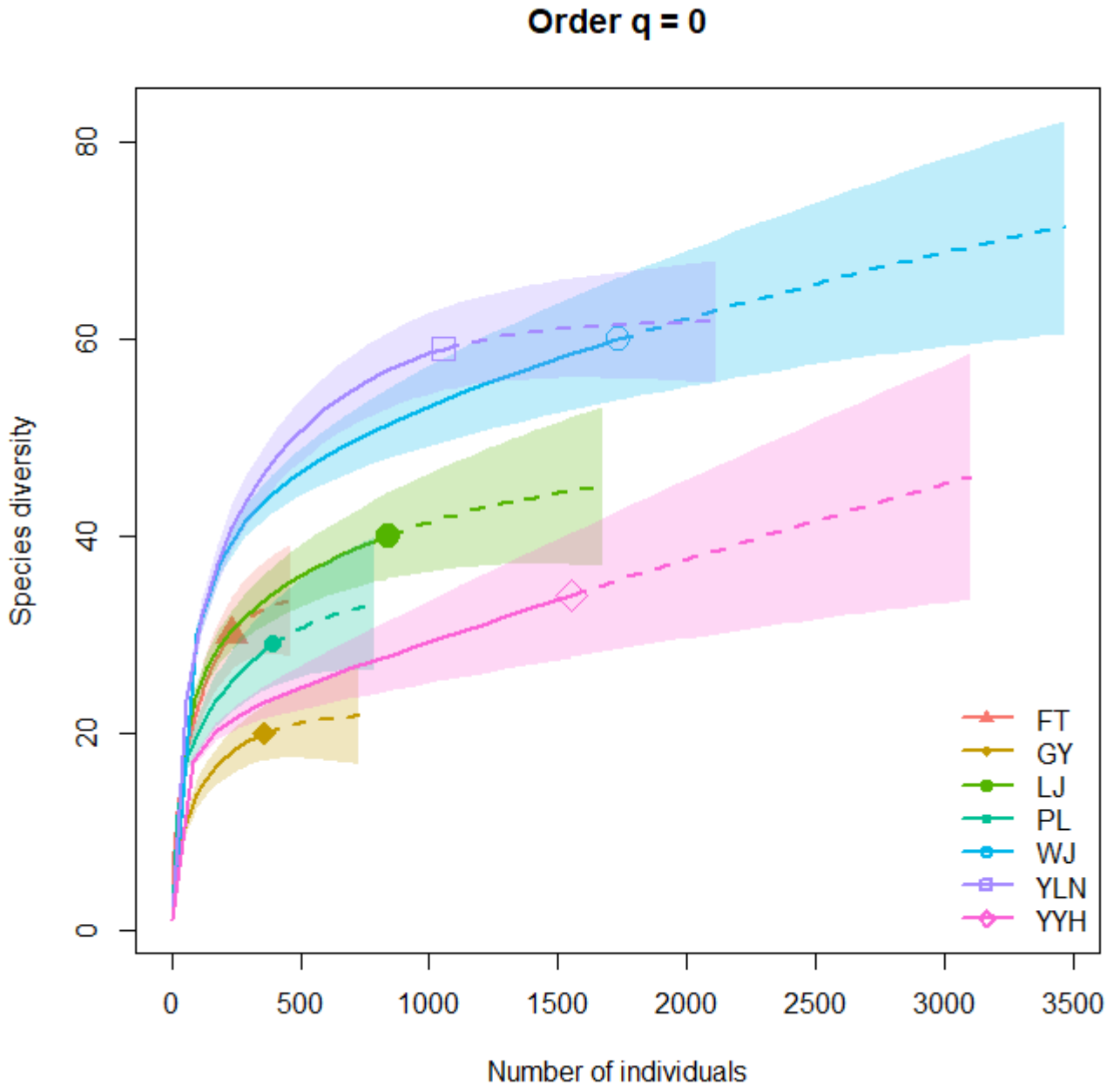
(note that both abundance and incidence data are based on the same original dataset, which contains number of individuals surveyed within each of 25 10×10-m subplots; for abundance data, individuals of each species have been summed across all 25 subplots within the locality, while for incidence data, only presences-absences of species within the subplots (i.e. not the number of their individuals) were considered and summed accross all 25 subplots within the locality).

Let's focus on abundance-based data first (`hp.abund`). You can see that localities quite remarkably differ in numbers of individuals (n), with the lowest number in FT (Feng-Tien, 232 individuals, low elevation) and highest in WJ (Wu-Jie, 1731 individuals, middle elevation). The numbers of species somehow copy the number of individuals (the correlation between n and `S.obs` in `DataInfo (hp.abund)` is 0.7: `cor (DataInfo (hp.abund)$S.obs, DataInfo (hp.abund)$n)`). This may be suspicious; what if the middle elevation localities are diverse simply because they have a

higher density of individuals per fixed sampled area?

To make sure that this is not the case, let's standardize the data to fixed number of individuals. We can first draw the rarefaction curves to see differences between individual localities:

```
D_abund <- iNEXT (hp.abund, datatype = 'abundance')
plot (D_abund)
```

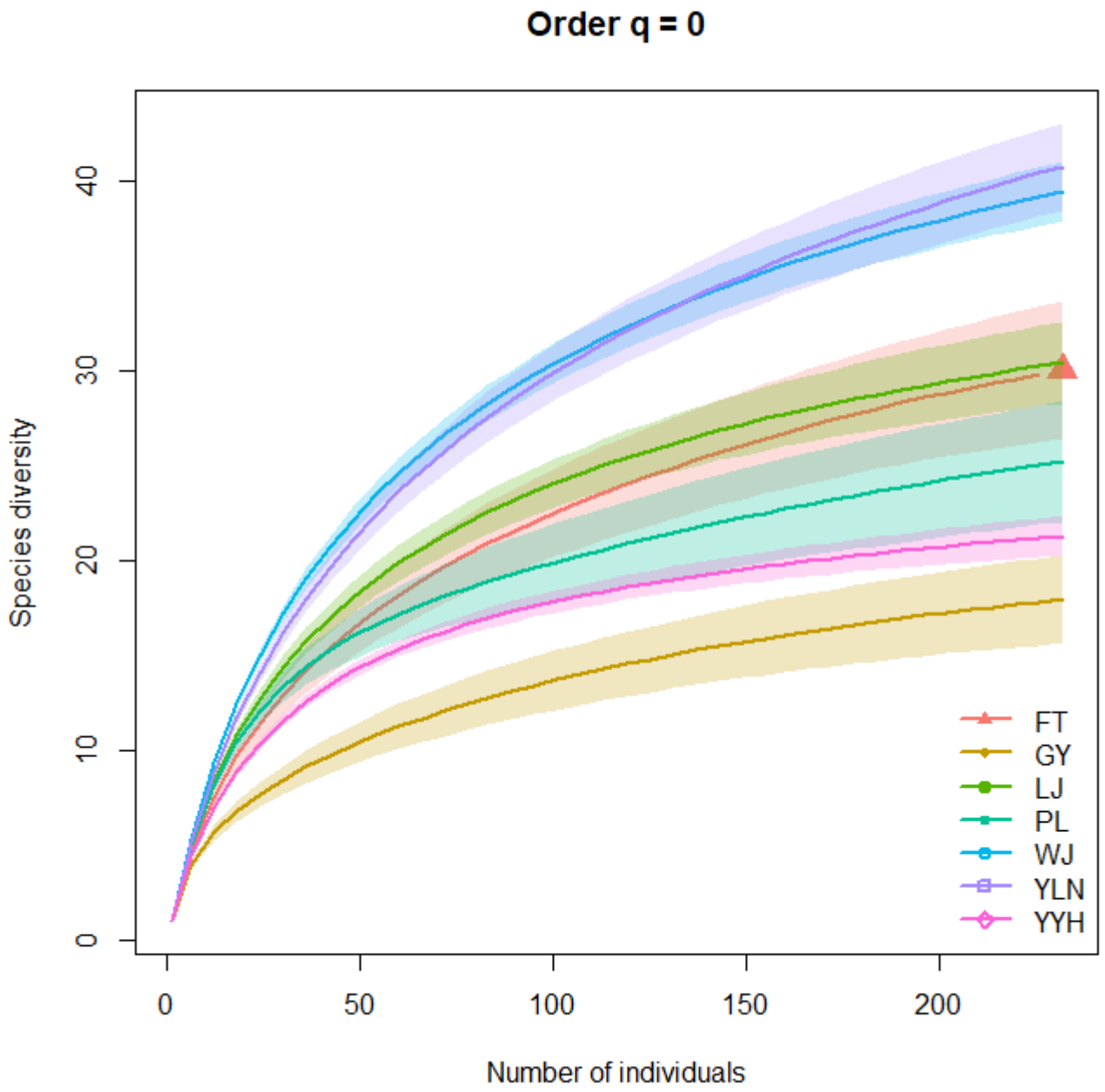


Localities very much differ by the number of individuals, which is why the rarefaction curves have rather different length; additionally, by default, the function calculates and plots also extrapolated part of rarefaction curve (up to double the number of individuals, dashed line). Important are also confidence intervals (C.I., envelopes around each curve): only diversity estimates with non-overlapping C.I. can be considered as significantly different. (Note that localities in figure legend are sorted by alphabet, while in the original dataset they are sorted by elevation).

To rarefy the diversities of all localities to the lowest number of observed individuals per locality (232

individuals in FT), we can use:

```
D_abund_232 <- iNEXT (hp.abund, datatype = 'abundance', endpoint = 232)
plot (D_abund_232)
```



The estimated numbers are, more efficiently, calculated by the function `estimateD`, with the same arguments as `iNEXT`:

```
est_D_abund <- estimateD (hp.abund, datatype = 'abundance', conf = NULL)
est_D_abund
```

```
  site   m       method    SC  q = 0  q = 1  q = 2
1   FT 232     observed 0.966 30.000 13.811  8.360
2  YLN 232 interpolated 0.944 40.742 21.907 14.028
3   LJ 232 interpolated 0.969 30.416 17.323 12.245
```

```
4   WJ 232 interpolated 0.957 39.403 24.666 18.115
5  YYH 232 interpolated 0.983 21.316 11.060  7.093
6   PL 232 interpolated 0.970 25.229 15.268 11.799
7   GY 232 interpolated 0.979 17.965  7.373  5.236
</code>
Note that, for simplicity, in the further comparisons I decided to ignore confidence intervals (''conf = NULL'' in ''estimateD''). If you do the comparison seriously, you should, however, consider them; in that case, the output of ''estimateD'' will be slighlty more complex.

In the output of ''estimateD'', the columns ''q = 0'', ''q = 1'' and ''q = 2'' contains estimated diversity by Hill number 1 (species richness), 2 (Shannon diversity) and 3 (Simpson diversity); check [[en:div-ind#hill_numbers|Hill numbers]] to refresh what it means. We care about species richness in this case, so the estimates of richness standardized to the same number of individuals (232) is in the column ''q = 0'':
<code rsplus>
D_individuals <- est_D_abund$`q = 0`
</code>

Alternatively, we may standardize the same data not to the same number of individuals, but to the same sample coverage (completeness of our survey when compared to the expected number of species occurring in the surveyed community). We may first check the coverage rarefaction curve for our localities, which can be plotted by the same ''plot'' function applied on the result of ''iNEXT'', just with modified ''type'' argument (check ''?plot.iNEXT'' for the meaning of individual arguments):
<code rsplus>
plot (D_abund, type = 3)
</code>
{{:obrazky:hp_abund_raref_curve_indiv_coverage.png?direct|}}
Since sample coverage values for all localities are quite high (lowest 0.9658 for FT, highest 0.9924 for YLN), we may zoom to the end of the x-axis, using the ''xlim'' argument of the ''plot'' function:
<code rsplus>
plot (D_abund, type = 3, xlim = c(.95, 1))
</code>
{{:obrazky:hp_abund_raref_curve_indiv_coverage_xlim.png?direct|}}

The locality with the lowest coverage is again Feng-Tien (it had also the lowest number of individuals, while being a potentially highly diverse lowland forest). We can use ''estimateD'' to standardize the diversities to this lowest coverage. Function ''estimateD'' has (additionally to ''datatype'' argument) also arguments ''base'' (either ''size'' if we want to do comparison based on sample size (number of individuals in abundance-based data or number of plots in incidence-based data), or ''coverage'' if we do comparison based on coverage) and ''level'' - the absolute value to which to standardize (either size or coverage). If the ''level'' argument is left default (''NULL''), the standardization is done to the lowest level (of size or coverage) among localities:
<code rsplus>
```

```
est_D_abund_coverage <- estimateD (hp.abund, datatype = 'abundance', base =
'coverage', conf = NULL)
est_D_abund_coverage
</code>
<file>
  site   m       method    SC  q = 0  q = 1  q = 2
1   FT 232     observed 0.966 30.000 13.811  8.360
2  YLN 378 interpolated 0.966 47.089 23.005 14.341
3   LJ 210 interpolated 0.966 29.703 17.173 12.183
4   WJ 280 interpolated 0.966 41.247 25.145 18.348
5  YYH 124 interpolated 0.966 18.766 10.566  6.933
6   PL 200 interpolated 0.966 24.196 15.101 11.711
7   GY 151 interpolated 0.966 15.777  7.188  5.185
```

The estimates of species richness standardized to sample coverage 0.966 are again in the column `q = 0`:

```
D_coverage <- est_D_abund_coverage$`q = 0`
```

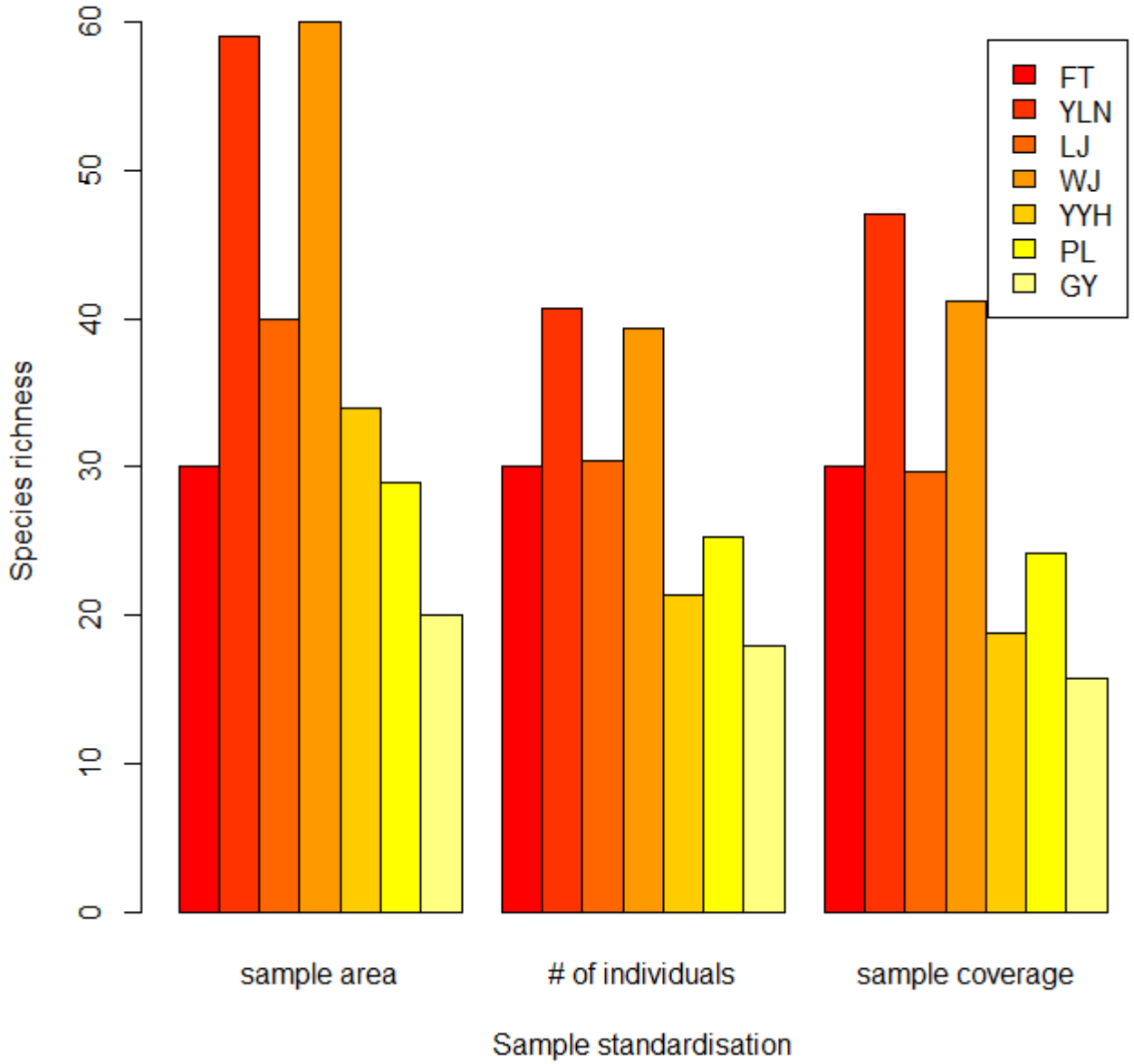
We need to add also the original numbers of species at each locality (i.e. diversity standardised to sample area, `D_area`; this can be taken e.g. from `DataInfo` output object, argument `S.obs`), and store all three variables into a single object, to allow for comparison (`D_est`):

```
D_area <- DataInfo (hp.abund, datatype = 'abundance')$S.obs
D_est <- cbind (D_area, D_individuals, D_coverage)
rownames (D_est) <- D_area <- DataInfo (hp.abund, datatype =
'abundance')$site
D_est
```

```
    D_area D_individuals D_coverage
FT      30        30.000     30.000
YLN     59        40.742     47.089
LJ      40        30.416     29.703
WJ      60        39.403     41.247
YYH     34        21.316     18.766
PL      29        25.229     24.196
GY      20        17.965     15.777
```

To visually compare diversity, we can use a simple `barplot` function, with argument `beside = TRUE` to make sure that bars for individual sites will be displayed beside, not stacked:

```
barplot (D_est, beside = T, legend.text = T, col = heat.colors (7), xlab =
'Sample standardisation',
         ylab = 'Species richness', names.arg = c('sample area', '# of
individuals', 'sample coverage'))
```



Note that it make no sense to compare diversities of individual plots across different standardisations (e.g. richness of FT standardised to area, individuals and coverage), but it make sense to compare them within the same standardisation (e.g. FT standardised to number of individuals with YYH standardised to number of individuals). From the barplot it is clear that the rank of localities according to their richness changes after standardisation; e.g., after standardisation to sample coverage, the YYH (Yuan-Yang-Hu, the plot close to famous Yuan-Yang lake, 鴛鴦湖, perhaps the foggiest locality in Taiwan) became species poorer than FT (Feng-Tien, lowland subtropical forest), although in the original data YYH is richer than FT, perhaps due to remarkably higher number of individuals surveyed in YYH (1551 ind.) than in FT (232 ind.).

Another option is to remove absolute difference in diversity among standardisations (use function `scale` to standardise diversity within the same standardisation to zero mean and unit variance) and plot the pattern along elevation (the script is already a bit more advanced, you may check the `?matplot` for details):

```
matplot (scale (D_est), type = 'b', axes = F, xlab = 'Locality', ylab =
'Standardized species richness')
```

```
axis (1, at = 1:7, labels = rownames (D_est))
axis (2)
box ()
legend ('topright', title = 'Diversity standardised by:', legend = c('sample
area', '# individuals', 'sample coverage'), pch = as.character (1:3), col =
1:3, lty = 1:3)
```