Summary stats - Lithic analysis from three sites: Balver Höhle, Buhlen & Ramioul

Lisa Schunk

2021-02-05

# Goal of the script

This script computes standard descriptive statistics for each group.  
The groups are based on:

* Tool type
* state of the tool (Complete, distal/proximal fragment, medial fragment)

It computes the following statistics:

* n (sample size = length): number of measurements
* smallest value (min)
* largest value (max)
* mean
* median
* standard deviation (sd)

dir\_in <- "analysis/all\_sites/derived\_data/"  
dir\_out <- "analysis/all\_sites/summary\_stats/"

Raw data must be located in ~/analysis/all\_sites/derived\_data/.  
Formatted data will be saved in ~/analysis/all\_sites/summary\_stats/. The knit directory for this script is the project directory.

# Load packages

library(openxlsx)

Warning: package 'openxlsx' was built under R version 4.0.3

library(R.utils)  
library(tools)  
library(doBy)

Warning: package 'doBy' was built under R version 4.0.3

# Get name, path and information of the file

data\_file <- list.files(dir\_in, pattern = "\\.xlsx$", full.names = TRUE)  
md5\_in <- md5sum(data\_file)  
info\_in <- data.frame(file = basename(names(md5\_in)), checksum = md5\_in,   
 row.names = NULL)

The checksum (MD5 hashes) of the imported file is

file checksum  
1 all\_sites\_analysis.xlsx 6dc873265ae237ede148dfdbfd89cac5

# Load data into R object

imp\_data <- loadObject(paste0(dir\_in, "all\_sites\_analysis.Rbin"))

The imported file is: “~/analysis/all\_sites/derived\_data/all\_sites\_analysis.xlsx”

# Define numeric variables

# changes the order of columns   
imp\_data <- imp\_data[c(1:22, 27,23:26, 28:34)]  
num.var <-19:23

The following variables will be used:

[19] length  
[20] width  
[21] thickness  
[22] weight  
[23] thickness.back

# Compute summary statistics

## Create function to compute the statistics at once

nminmaxmeanmedsd <- function(x){  
 y <- x[!is.na(x)]  
 n\_test <- length(y)  
 min\_test <- min(y)  
 max\_test <- max(y)  
 mean\_test <- mean(y)  
 med\_test <- median(y)  
 sd\_test <- sd(y)  
 out <- c(n\_test, min\_test, max\_test, mean\_test, med\_test, sd\_test)  
 names(out) <- c("n", "min", "max", "mean", "median", "sd")  
 return(out)  
}

## Compute the summary statistics

### Dimensions

# Dimensions Keilmesser, Keilmesser-points, Pradnik scraper &   
# Later sharpening spall   
dimensions <- summaryBy(length + width + thickness ~ technological.class +  
 artefact.state, data = imp\_data, FUN = nminmaxmeanmedsd)  
str(dimensions)

'data.frame': 13 obs. of 20 variables:  
 $ technological.class: chr "flake" "hammerstone" "Keilmesser" "Keilmesser" ...  
 $ artefact.state : chr "complete" "complete" "complete" "distal\_fragment" ...  
 $ length.n : num 2 1 278 1 36 2 13 146 7 3 ...  
 $ length.min : num 54 90 29.7 22 13 ...  
 $ length.max : num 99 90 135.6 22 91.9 ...  
 $ length.mean : num 76.5 90 55.5 22 38.1 ...  
 $ length.median : num 76.5 90 51.9 22 35.6 ...  
 $ length.sd : num 31.8 NA 16.9 NA 16.1 ...  
 $ width.n : num 2 1 278 1 36 2 13 146 7 3 ...  
 $ width.min : num 27 62 14 44 19 ...  
 $ width.max : num 42 62 81.4 44 53.2 ...  
 $ width.mean : num 34.5 62 33.7 44 33.3 ...  
 $ width.median : num 34.5 62 32.4 44 32 ...  
 $ width.sd : num 10.61 NA 8.84 NA 8.47 ...  
 $ thickness.n : num 2 1 278 1 36 2 13 146 7 3 ...  
 $ thickness.min : num 16 52 7 12 8 10 13.6 2 4 2.32 ...  
 $ thickness.max : num 25 52 31 12 23.6 ...  
 $ thickness.mean : num 20.5 52 16.3 12 14.3 ...  
 $ thickness.median : num 20.5 52 15.8 12 13.6 ...  
 $ thickness.sd : num 6.36 NA 4.33 NA 3.58 ...

# Save data

## Format name of output file

file\_out <- "all\_sites\_analysis\_stats"

The file will be saved as “~/analysis/all\_sites/summary\_stats/.[ext]”.

## Write to XLSX

write.xlsx(list(dimensions = dimensions),   
 file = paste0(dir\_out, file\_out, ".xlsx"))

## Save R object

saveObject(list(dimensions = dimensions),   
 file = paste0(dir\_out, file\_out, ".Rbin"))

## Show file information

file\_out <- c(paste0(dir\_out, file\_out, ".xlsx"), paste0(dir\_out,   
 file\_out, ".Rbin"))  
md5\_out <- md5sum(file\_out)  
info\_out <- data.frame(files = basename(names(md5\_out)), checksum = md5\_out,  
 row.names = NULL)

The checksum (MD5 hashes) of the exported files are:

files checksum  
1 all\_sites\_analysis\_stats.xlsx 7a5de567c202d14ab9c3912f2e287459  
2 all\_sites\_analysis\_stats.Rbin 8ff732498420a3d51de58f477d0c8b60

# sessionInfo() and RStudio version

sessionInfo()

R version 4.0.2 (2020-06-22)  
Platform: x86\_64-w64-mingw32/x64 (64-bit)  
Running under: Windows 10 x64 (build 19041)  
  
Matrix products: default  
  
locale:  
[1] LC\_COLLATE=German\_Germany.1252 LC\_CTYPE=German\_Germany.1252   
[3] LC\_MONETARY=German\_Germany.1252 LC\_NUMERIC=C   
[5] LC\_TIME=German\_Germany.1252   
  
attached base packages:  
[1] tools stats graphics grDevices utils datasets methods   
[8] base   
  
other attached packages:  
[1] doBy\_4.6.8 R.utils\_2.10.1 R.oo\_1.24.0 R.methodsS3\_1.8.1  
[5] openxlsx\_4.2.3   
  
loaded via a namespace (and not attached):  
 [1] zip\_2.1.1 Rcpp\_1.0.6 compiler\_4.0.2 pillar\_1.4.7   
 [5] digest\_0.6.27 lattice\_0.20-41 evaluate\_0.14 lifecycle\_0.2.0   
 [9] tibble\_3.0.5 gtable\_0.3.0 pkgconfig\_2.0.3 rlang\_0.4.10   
[13] Matrix\_1.2-18 DBI\_1.1.1 yaml\_2.2.1 xfun\_0.20   
[17] dplyr\_1.0.3 stringr\_1.4.0 knitr\_1.31 generics\_0.1.0   
[21] vctrs\_0.3.6 grid\_4.0.2 tidyselect\_1.1.0 glue\_1.4.2   
[25] R6\_2.5.0 rmarkdown\_2.6 tidyr\_1.1.2 purrr\_0.3.4   
[29] ggplot2\_3.3.3 magrittr\_2.0.1 backports\_1.2.0 scales\_1.1.1   
[33] ellipsis\_0.3.1 htmltools\_0.5.1.1 MASS\_7.3-53 assertthat\_0.2.1   
[37] colorspace\_2.0-0 Deriv\_4.1.2 stringi\_1.5.3 munsell\_0.5.0   
[41] broom\_0.7.4 crayon\_1.4.0

RStudio version 1.3.1073.

END OF SCRIPT