Two proportion tests

Brian Yurk

3/10/2021

```r
rm(list=ls())
gc()
```

```
##          used (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
## Ncells 423595 22.7     881232 47.1         NA   658077 35.2
## Vcells 805963  6.2    8388608 64.0      32768  1802945 13.8
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(ggplot2)
```

Some functions. The first creates 2x2 contingency tables. Events are grouped according to whether drift potential exceeds a threshold for a particular coast-station pair or not. The number of events with system (low or high pressure) centers in a particular direction or not is compared between the two groups. Ultimately a Fisher exact test is run on each contingency table (e.g., for each direction for a coast-station pair) and the proportions are plotted for the two groups for each system center direction.

```r
#build contingency table
con_tab <- function(data,cdir_stat,sysDir,dpthresh=1,lo=TRUE){
  if(lo){
    tib <- data %>% select(sTart = loDirStart,eNd = loDirEnd,dp=all_of(cdir_stat))
  } else {
    tib <- data %>% select(sTart = hiDirStart,eNd = hiDirEnd,dp=all_of(cdir_stat))
  }

  tib <- tib %>% mutate( boolCDir = (dp >= dpthresh),
                         boolSysDir =( (sTart == sysDir) |
                                         (eNd == sysDir) ) ) %>%
      select(-dp,-sTart,-eNd)

  tib <- tib %>% drop_na() %>% group_by(boolCDir) %>%
    summarize(SysDirT = sum(boolSysDir),SysDirF = sum(!boolSysDir))

  if(tib$boolCDir[1]){
    roLabs <- c(cdir_stat,paste0("not_",cdir_stat))
  } else {
    roLabs <- c(paste0("not_",cdir_stat),cdir_stat)
  }

  coLabs <- c(as.character(sysDir),paste0("not_",sysDir))

  ct <- as.matrix(tib)[,2:3]
  dimnames(ct) <- list(roLabs,coLabs)

  return(ct)
}

#run fisher exact test for direction and output proportions
fisher_test_coast <- function(data,cdir_stat,sysDirs,dpthresh=1,lo=TRUE){

  #build a list contingency tables for the different system directions
  ct_list <- sysDirs %>%
    map(~con_tab(data=data,cdir_stat=cdir_stat,sysDir=.x,dpthresh=dpthresh,lo=lo))

  #run fisher test on the contingency tables and output a vector of p values
  p_vals <- ct_list %>% map_dbl(~fisher.test(.x)$p.value)
  names(p_vals) <- as.character(sysDirs)

  #grepl is used to make sure that the proportion is the proportion
  #of successes rather than failures and to make sure that the rows
  #are orderd by coast, not coast
  prop_df <- ct_list %>% map_dfc(~.x[1+grepl("not",rownames(.x)),
                                     !grepl("not",colnames(.x))]/
                                   rowSums(.x[1+grepl("not",rownames(.x)),]))
  names(prop_df) <- as.character(sysDirs)
  prop_df$cdir_stat <- c(cdir_stat,paste0("not_",cdir_stat))


  return(list(p_vals=p_vals,prop_df=prop_df))
}

# given the output of fisher_test_coast, plot proportions of events in with
# system center in the given direction for the two groups
plot_props <- function(ftc){
  prop_df <- ftc$prop_df
  print(prop_df)
  p_vals <- ftc$p_vals
  print(p_vals)
  prop_df_long <- prop_df %>%
    pivot_longer(!cdir_stat,names_to="sysDir",values_to="proportion") #long format better for plotting
  (p <- ggplot(data=prop_df_long,aes(x=reorder(sysDir,as.numeric(sysDir)),
                                     y=proportion,fill=cdir_stat)) +
      geom_bar(stat="identity",color="black", position=position_dodge()) +
      xlab("system center") +
      theme_minimal() +
      scale_fill_brewer(palette="Blues"))

  return(p)
}
```

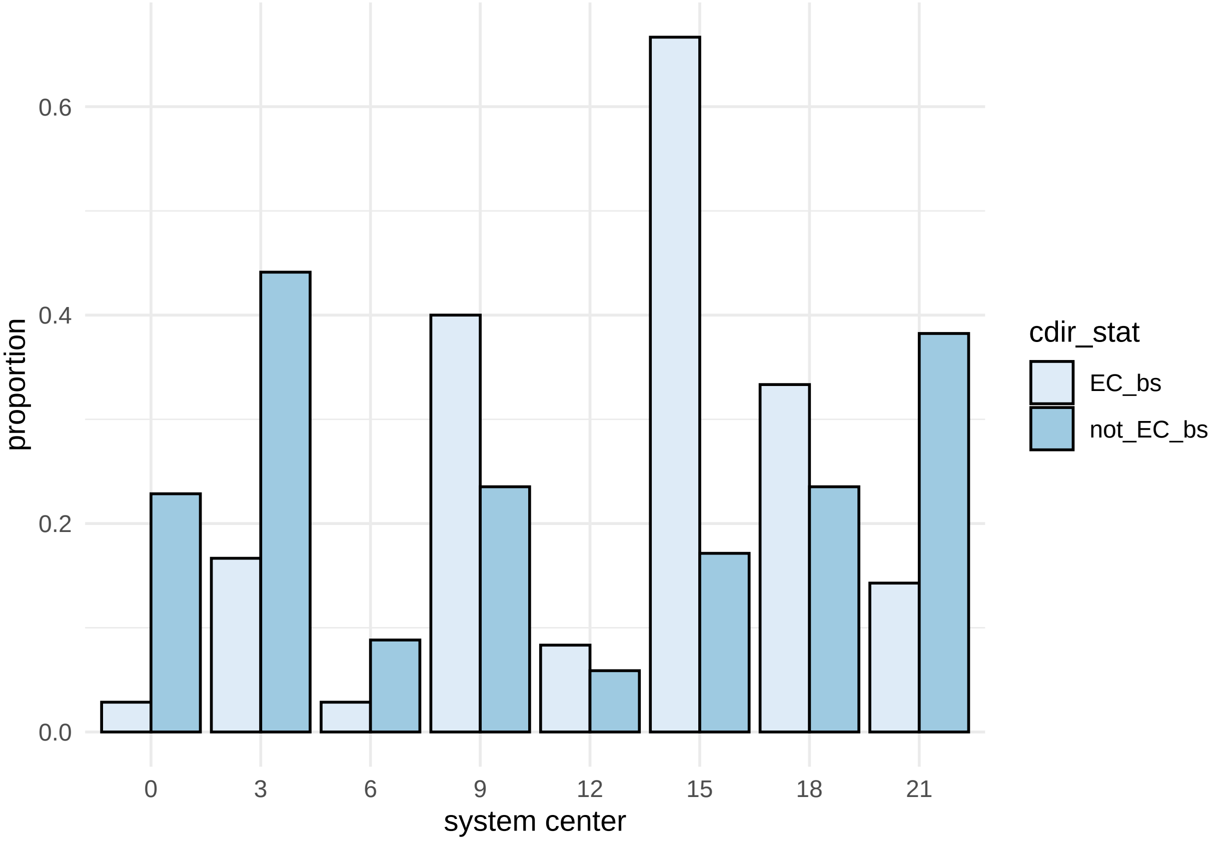
Import the data, run the test for a specified coast-station pair, and plot the proportions.

```
library(xlsx)
data <- read.xlsx("raw_data/Table R-1.xlsx",sheetIndex=1,colIndex=1:22,rowIndex=1:84,
                  header=TRUE,stringsAsFactors=FALSE)
names(data)<-c("Event","tStart","tEnd","loDirStart","loDirEnd","hiDirStart",
               "hiDirEnd","NC_na","NC_pi","NEC_na","EC_ho","EC_bs","SEC_mc",
               "SEC_gt","SC_mc","SC_bh","SWC_ch","WC_ke","WC_mw","NWC_fa")

cdir_stations <- c("NC_na","NC_pi","NEC_na","EC_ho","EC_bs","SEC_mc","SEC_gt",
                   "SC_mc","SC_bh","SWC_ch","WC_ke","WC_mw","NWC_fa")
sysDirs <- seq(0,21,by=3)

#example
#ct <- con_tab(data,"EC_bs",0,lo=TRUE)
#ft <- fisher.test(ct)
#ft$p.value
#props <- ct[,!grepl("not", colnames(ct))]/rowSums(ct)

#example
#ftc <- fisher_test_coast(data,"EC_bs",sysDirs,dpthresh=1,lo=TRUE)
#prop_df <- ftc$prop_df
#p_vals <- ftc$p_vals


ftc <- fisher_test_coast(data,"EC_bs",sysDirs,dpthresh=1,lo=FALSE)
```

```
## New names:
## * NA -> ...1
## * NA -> ...2
## * NA -> ...3
## * NA -> ...4
## * NA -> ...5
## * ...
```

```
ftc %>% plot_props()
```

```
## # A tibble: 2 x 9
##      `0`   `3`    `6`   `9`   `12`  `15`  `18`  `21` cdir_stat
##    <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <chr>
## 1 0.0286 0.167 0.0286 0.4   0.0833 0.667 0.333 0.143 EC_bs
## 2 0.229  0.441 0.0882 0.235 0.0588 0.171 0.235 0.382 not_EC_bs
##            0            3            6            9           12           15
## 2.750462e-02 1.839949e-02 3.564785e-01 1.974938e-01 1.000000e+00 3.262383e-05
##           18           21
## 4.330639e-01 3.001159e-02
```



```
prop_df <- ftc$prop_df
p_vals <- data.frame(ftc$p_vals)
```