spidercluster-permutations

Claire and Ethan

October 12, 2017

```
require(dplyr)
require(ggplot2)
```

```
# CREATE NULL DISTRIBUTIONS of clusters of (1) all spiders and (2) spiders who clustered. Uses 2014 data.
# (1) all spiders can cluster
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.2.4
```

```
library(backports)
```

```
## Warning: package 'backports' was built under R version 3.2.5
```

```
library(dplyr)
library(ggplot2)
spiders <- read_excel("spiders.xlsx")
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [211, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [397, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [441, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [442, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [461, 10]: expecting numeric: got 'NA'
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [514, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [516, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [537, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [541, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [555, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [556, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [557, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [562, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [566, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [569, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [570, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [924, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [936, 10]: expecting numeric: got 'NA'
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [948, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [962, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [963, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [965, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1021, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1022, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1029, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1149, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1174, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1176, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1192, 10]: expecting numeric: got 'NA'
```

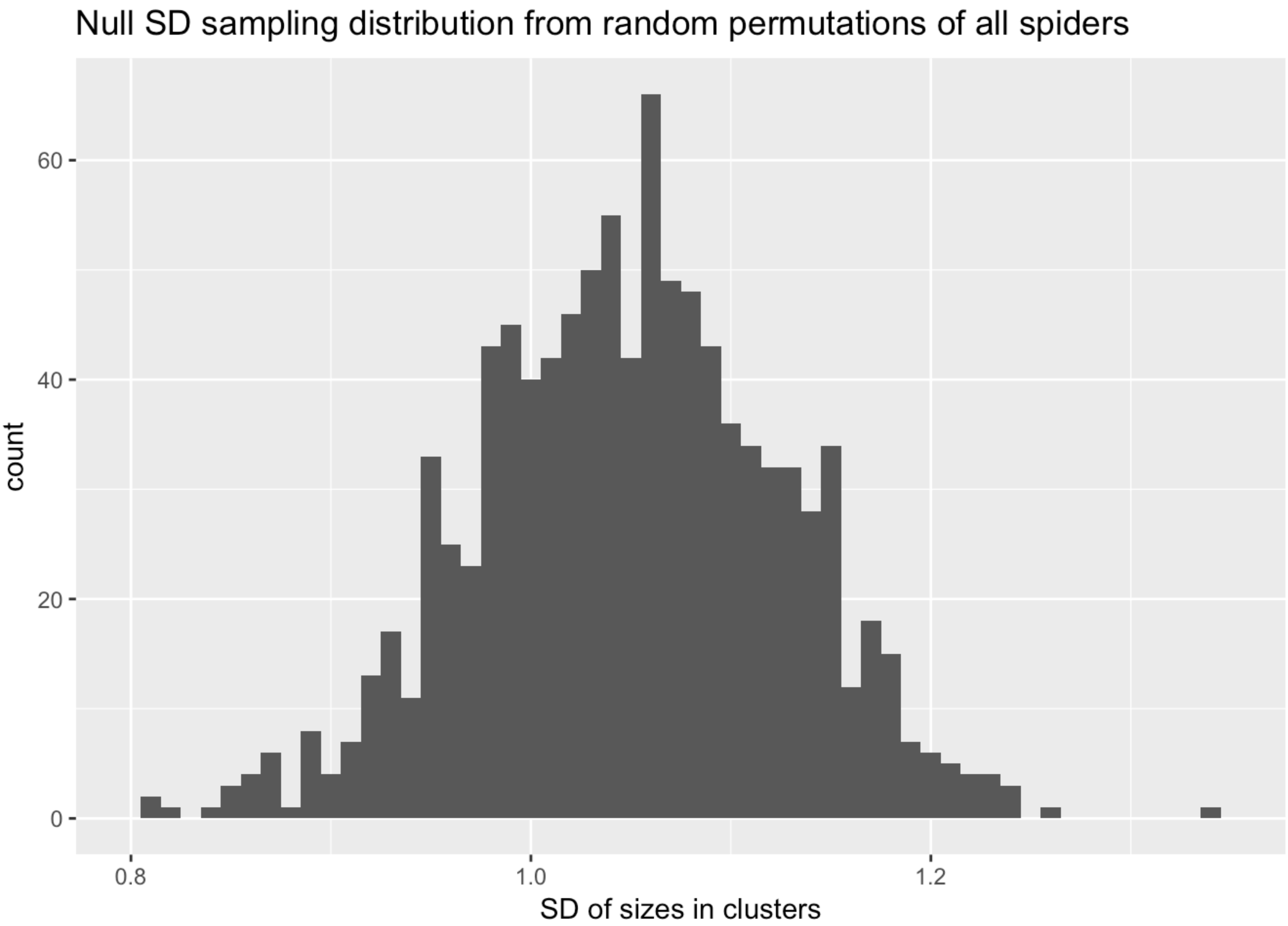
```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1204, 10]: expecting numeric: got 'NA'
```

```
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1540, 10]: expecting numeric: got 'NA'
## Warning in read_xlsx_(path, sheet, col_names = col_names, col_types =
## col_types, : [1541, 10]: expecting numeric: got 'NA'
```

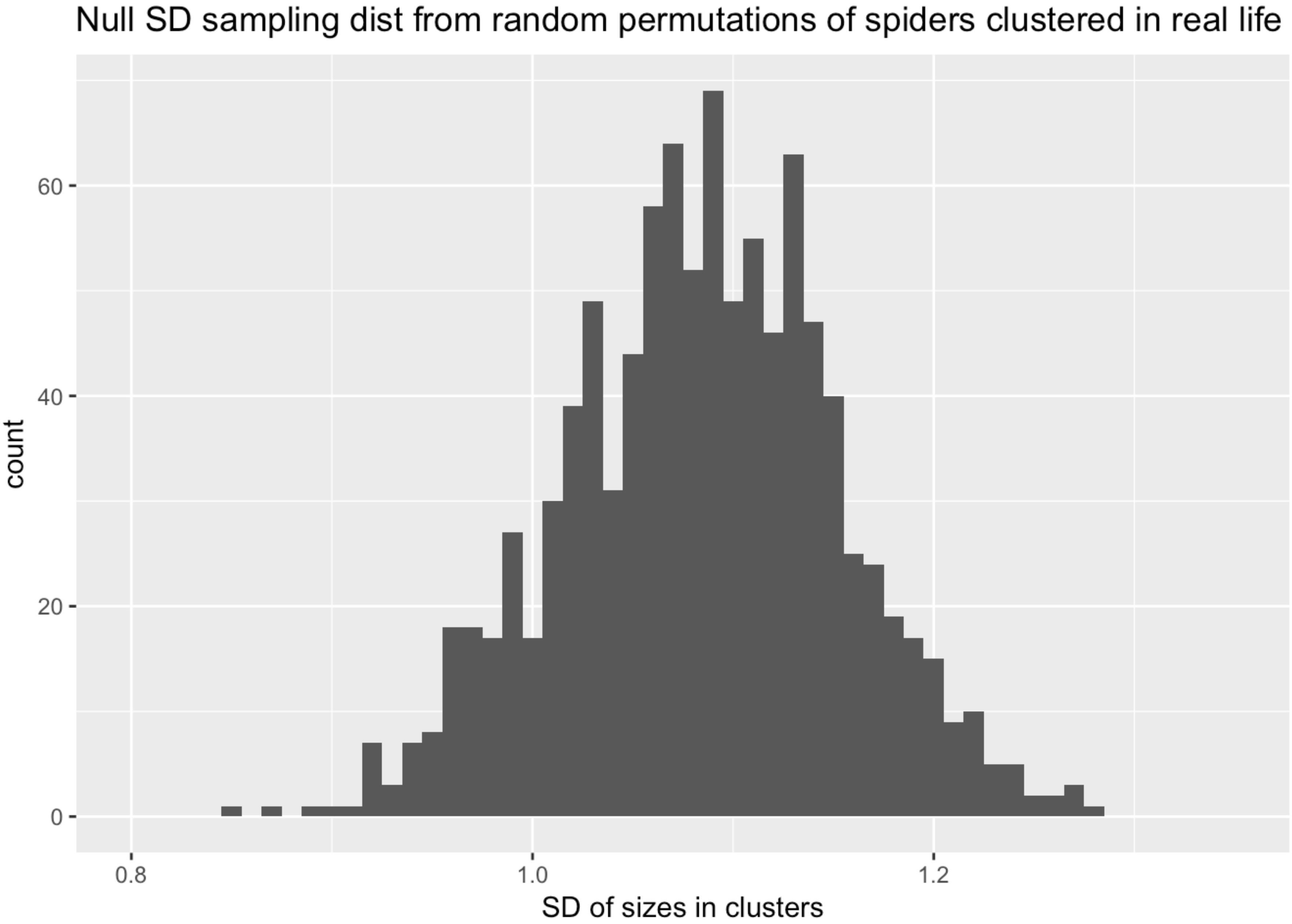
```
#ANALYSIS
# the actual clusters
clusters14 <- spiders %>%
  filter(Year == 2014)%>%
  filter(grepl('C', WebID)) %>%
  group_by(WebID) %>%
  filter(!is.na(SpiderSizemm)) %>%
  summarise(meanSize = mean(SpiderSizemm), sdSize = sd(SpiderSizemm), maxSize = max(S
piderSizemm), minSize = min(SpiderSizemm))
```

```
## Warning: package 'bindrcpp' was built under R version 3.2.5
```

```
meanActualsdSize = mean(clusters14$sdSize, na.rm=TRUE)  # 0.92 average SD of actual c
lusters



reps = 1000

nullDist <- c()

for(i in 1:reps){
  clusterStats <- spiders %>%
    filter(Year == 2014) %>%
    mutate(sizePerm = sample(SpiderSizemm, replace=FALSE)) %>%
    group_by(WebID) %>%
    filter(!is.na(SpiderSizemm)) %>%
    summarise(meanSize = mean(sizePerm), sdSize = sd(sizePerm), maxSize =
                max(sizePerm), minSize = min(sizePerm), sumSq = sum(sizePerm^2))
  meanNullsdSize = mean(clusterStats$sdSize, na.rm=TRUE)
  nullDist <- c(nullDist, meanNullsdSize)
#  if (meanNullsdSize<meanActualsdSize) {
 #   nullDist = c(nullDist, 1)
  #}
}

# (2) only use spiders that originally clustered
nullDist2 <- c()

for(i in 1:reps){
  clusterStats <- spiders %>%
    filter(Year == 2014) %>%
    filter(grepl('C', WebID)) %>%
    mutate(sizePerm = sample(SpiderSizemm, replace=FALSE)) %>%
    group_by(WebID) %>%
    filter(!is.na(SpiderSizemm)) %>%
    summarise(meanSize = mean(sizePerm), sdSize = sd(sizePerm), maxSize =
                max(sizePerm), minSize = min(sizePerm))
  meanNull1sdSize = mean(clusterStats$sdSize, na.rm=TRUE)
  nullDist2 <- c(nullDist2, meanNull1sdSize)
 # if (meanNullsdSize>meanActualsdSize) {
  #  nullDist2 = c(nullDist2, 1)
  #}
}

nullDistdf <- data.frame(matrix(nrow = reps,ncol=3))
nullDistdf[,1] <- c(1:reps)
nullDistdf[,2] <- nullDist
nullDistdf[,3] <- nullDist2

ggplot(data=nullDistdf) + geom_histogram(aes(X2), binwidth=0.01) + labs(x = 'SD of si
zes in clusters', title = 'Null SD sampling distribution from random permutations of
all spiders') + coord_cartesian(xlim = c(0.8, 1.35))
```



```
ggplot(data=nullDistdf) + geom_histogram(aes(X3), binwidth=0.01) + labs(x = 'SD of si
zes in clusters', title = 'Null SD sampling dist from random permutations of spiders
clustered in real life') + coord_cartesian(xlim = c(0.8, 1.35))
```



```
# analysis by aggregating the actual clusters

pSD = sum(nullDist<meanActualsdSize)/reps   # .039 proportion of random clusters of a
ll spiders whose SDs were more extreme
pSD2 = sum(nullDist2<meanActualsdSize)/reps
#pSD2 = sum(nullDist2)/reps # 0.012 proportion of random clusters of clustered spider
s whose SDs were more extreme

pSD
```

```
## [1] 0.044
```

```
pSD2
```

```
## [1] 0.009
```

```
# analysis on each actual cluster
pSDnull1 <- c()
pSDnull2 <- c()

# for(cluster in 1:61){
#   pSDnull1 <- c(pSDnull1, 2*sum(nullDist$sdSize<clusters14$sdSize[cluster])/1000)
#   pSDnull2 <- c(pSDsnull2, 2*sum(nullDist2$sdSize<clusters14$sdSize[cluster])/1000)
# }
#
# mean(pSDnull1) # 0.63 average p-value from the actual clusters based on null1
# mean(pSDnull2) # 0.67 average p-value from the actual clusters based on null2
```