Raw vs Summarized

Contents

0.0.1 Same plot from raw and summarized data

For the following data structure

```
dsN<-data.frame(
  id=rep(1:100, each=4),
  yearF=factor(rep(2001:2004, 100)),
  attendF=sample(1:8, 400, T, c(.2,.2,.15,.10,.10, .20, .15, .02))
)
dsN[sample(which(dsN$yearF==2001), 5), "attendF"]<-NA
dsN[sample(which(dsN$yearF==2002), 10), "attendF"]<-NA
dsN[sample(which(dsN$yearF==2003), 15), "attendF"]<-NA
dsN[sample(which(dsN$yearF==2004), 20), "attendF"]<-NA

attcol8<-c("Never"="#4575b4",
           "Once or Twice"="#74add1",
           "Less than once/month"="#abd9e9",
           "About once/month"="#e0f3f8",
           "About twice/month"="#fee090",
           "About once/week"="#fdae61",
           "Several times/week"="#f46d43",
           "Everyday"="#d73027")
dsN$attendF<-factor(dsN$attendF, levels=1:8, labels=names(attcol8))
head(dsN,13)
```

```
   id yearF              attendF
1   1  2001   Several times/week
2   1  2002   Several times/week
3   1  2003 Less than once/month
4   1  2004     About once/month
5   2  2001     About once/month
6   2  2002                Never
7   2  2003                Never
8   2  2004                 <NA>
9   3  2001                Never
10  3  2002   Several times/week
11  3  2003        Once or Twice
12  3  2004      About once/week
13  4  2001    About twice/month
```

we can obtain a series of a stacked bar charts

```
require(ggplot2)
# p<- ggplot( subset(dsN,!is.na(attendF)), aes(x=yearF, fill=attendF)) # leaving NA out of
```

```
p<- ggplot( dsN, aes(x=yearF, fill=attendF))  #  keeping NA in calculations
p<- p+ geom_bar(position="fill")
p<- p+ scale_fill_manual(values = attcol8,
                         name="Response category" )
p<- p+ scale_y_continuous("Prevalence: proportion of total",
                          limits=c(0, 1),
                          breaks=c(.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
p<- p+ scale_x_discrete("Waves of measurement",
                        limits=as.character(c(2000:2005)))
p<- p+ labs(title=paste0("In the past year, how often have you attended a worship service?"))
p
```

The graph above is produced from the raw data. However, it is sometimes convenient to produce graphs from summarized data, especially if one needs control over statistical functions. Below is transformation of dsN into ds that contains only the values that are actually mapped on the graph above:

```
require(dplyr)
ds<- dsN %.%
  dplyr::filter(!is.na(attendF)) %.%
  dplyr::group_by(yearF,attendF) %.%
  dplyr::summarize(count = sum(attendF)) %.%
  dplyr::mutate(total = sum(count),
              percent= count/total)
head(ds,10)
```

```
Source: local data frame [10 x 5]
Groups: yearF
```

```
   yearF              attendF count total percent
1   2001                Never    15   356 0.04213
2   2001        Once or Twice    42   356 0.11798
3   2001 Less than once/month    51   356 0.14326
4   2001     About once/month    24   356 0.06742
5   2001    About twice/month    35   356 0.09831
6   2001      About once/week    96   356 0.26966
7   2001   Several times/week    77   356 0.21629
8   2001              Everyday    16   356 0.04494
9   2002                Never    11   374 0.02941
10  2002        Once or Twice    28   374 0.07487
```

```
# verify
summarize(filter(ds, yearF==2001), should.be.one=sum(percent))
```

```
Source: local data frame [1 x 2]
```

```
  yearF should.be.one
1  2001             1
```

0.0.2 Question:

How would one re-create a graph from above using this summary dataset ds?