Simple Analysis and Data Vis

Bethany Growns

3/8/23

Welcome to Data Visualisation in `R`! We'll start with a quick introduction to R, and then plot some data. I always like to remember the `2 important rules` of R: 1) most people really do not know what they're doing, and 2) using R is spending more time on Google than in R Studio. Some more great resources are here: here and here

In order to run lines of code, highlight the ones you want and the press `Ctrl+Enter (Windows)`

Here we're going to load the packages that we need. Install these using `install.packages("PackageName")` in the Console below if you haven't already

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.0      v stringr 1.4.0
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(sjmisc)
```

```

Attaching package: 'sjmisc'

The following object is masked from 'package:purrr':
    is_empty

The following object is masked from 'package:tidyr':

    replace_na

The following object is masked from 'package:tibble':

    add_case
```

```r
library(ggplot2)
library(viridis)
```

```
Loading required package: viridisLite
```

```r
library(cowplot)
```

Read in the CSV file. This line of code basically says create a datafile (called "df") in the Global Environment by reading in this .csv file

```r
df <- read.csv("./Workshop_Data.csv")
```

Calculate the `frequency` of gender and ethnicity

```r
frq(df$gender)
```

```
x <character>
# total N=174 valid N=174 mean=1.95 sd=0.99

Value                             |  N | Raw % | Valid % | Cum. %
-----------------------------------------------------------------
Female                            | 89 | 51.15 |   51.15 |  51.15
Gender diverse (please specify):  |  5 |  2.87 |    2.87 |  54.02
Male                              | 80 | 45.98 |   45.98 | 100.00
<NA>                              |  0 |  0.00 |    <NA> |   <NA>
```

```r
frq(df$background)
```

```
x <character>
# total N=174 valid N=174 mean=5.30 sd=2.40

Value                        |   N | Raw % | Valid % | Cum. %
-------------------------------------------------------------
African American             |  14 |  8.05 |    8.05 |   8.05
Asian                        |  32 | 18.39 |   18.39 |  26.44
Hispanic                     |  11 |  6.32 |    6.32 |  32.76
Indian                       |   1 |  0.57 |    0.57 |  33.33
Indigenous Native American   |   1 |  0.57 |    0.57 |  33.91
Please specify:              |   3 |  1.72 |    1.72 |  35.63
White                        | 112 | 64.37 |   64.37 | 100.00
<NA>                         |   0 |  0.00 |    <NA> |   <NA>
```

Calculate the `mean` and `sd`

```
df %>%
  summarise(M = mean(age),stdev=sd(age))
```

```
        M    stdev
1 30.2931 11.50221
```

Conduct an `ANOVA` on Accuracy by BaseRate and Feedback

```
ANOVA <- aov(data = df, Accuracy ~ BaseRate * Feedback)
```

Print a `table` of the ANOVA

```
summary(ANOVA)
```

```
                   Df Sum Sq Mean Sq F value   Pr(>F)    
BaseRate            1 0.2510  0.2510   27.74 4.15e-07 ***
Feedback            1 0.2603  0.2603   28.77 2.64e-07 ***
BaseRate:Feedback   1 0.6088  0.6088   67.28 5.55e-14 ***
Residuals         170 1.5382  0.0090                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t.test(df$Accuracy, mu = 0.5)
```

```
	One Sample t-test

data:  df$Accuracy
t = 23.532, df = 173, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.7025901 0.7396864
sample estimates:
mean of x
0.7211382
```

Where are the means?

```r
df %>%
  group_by(Feedback,BaseRate) %>%
  summarise(M = mean(Accuracy),SD=sd(Accuracy))
```

```
`summarise()` has grouped output by 'Feedback'. You can override using the
`.groups` argument.
```

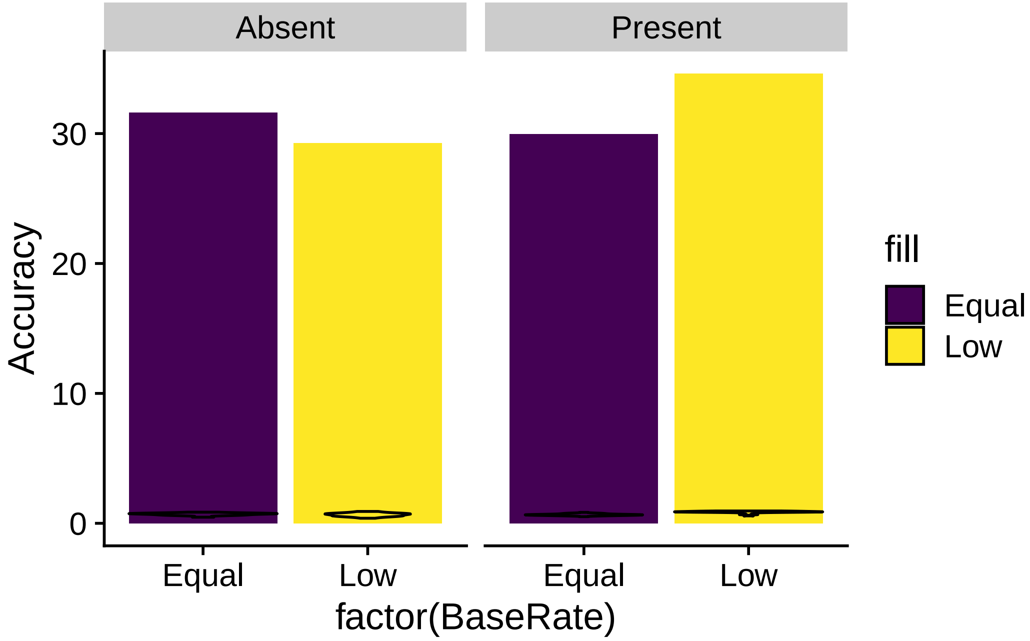
```
# A tibble: 4 x 4
# Groups:   Feedback [2]
  Feedback BaseRate     M     SD
  <chr>    <chr>    <dbl>  <dbl>
1 Absent   Equal    0.703 0.0916
2 Absent   Low      0.665 0.129
3 Present  Equal    0.666 0.0662
4 Present  Low      0.865 0.0810
```

Bar Charts

Ok but can I actually see the data now?

Here is a good resource

```r
ggplot(data = df, aes(x=factor(BaseRate), y = Accuracy, fill = "Feedback")) +
   geom_bar(stat = "identity", aes(fill = factor(BaseRate))) +
   facet_grid(. ~ Feedback) +
   theme_cowplot() +
   scale_fill_viridis(discrete = TRUE) +
   geom_violin(colour = "black", aes(fill = factor(BaseRate)))
```



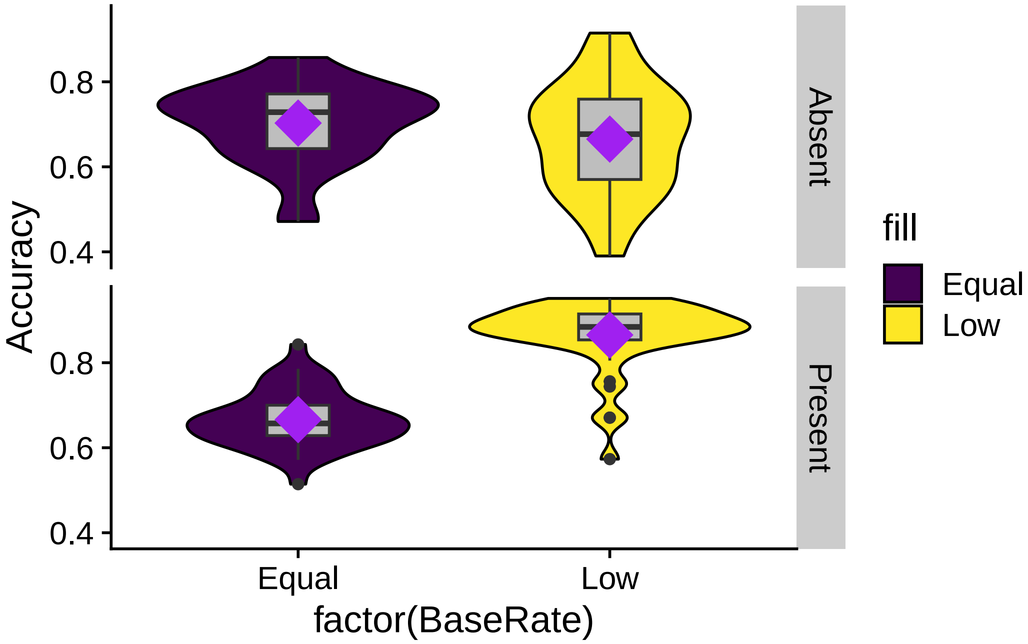
Seperate by Feedback: facet_grid(. ~ Feedback)

Make it less ugly?: theme_cowplot()

`Pretty colours!`: scale_fill_viridis(discrete = TRUE)

Violin Plots

```
ggplot(data = df, aes(x=factor(BaseRate), y = Accuracy, fill = "Feedback")) +
  facet_grid(Feedback ~ .) +
  theme_cowplot() +
  scale_fill_viridis(discrete = TRUE) +
  geom_violin(colour = "black", aes(fill = factor(BaseRate))) +
  geom_boxplot(width=0.2, fill = "grey") +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 8, color = "purple", fill =
```



What about the `range`?: geom_boxplot(width=0.2, fill = “grey”)

Where’s the mean??: stat_summary(fun = mean, geom = “point”, shape = 18, size = 8, color = “black”, fill = “black”)