IntroSlides

Basic steps in workflow

1. *Define environment*
2. *Import*
3. *Transform*
4. *Explore (general/outlier/distribution) (go back to 3?)*
5. *Classify scale level / distribution (based on 3/4)*
6. *Describe*
7. *Test / Model (may include step 6)*
8. *Report*

Define environment

- *Activate packages to use: library() / pacman::p_load()*
- *ggplot theme: theme_set() / theme_update()*
- *flextable settings: set_flextable_defaults()*
- *knitr::opts_chunk$set()*

```
if(!require(pacman)){
  install.packages("pacman")
}
pacman::p_load(wrappedtools,
               ggbeeswarm, ggsignif, ggridges,
               car, flextable)

theme_set(theme_light(base_size = 20))

set_flextable_defaults(
  font.size = 9,
  theme_fun = theme_zebra,
  padding.bottom = 1, padding.top = 3, padding.left = 2, padding.right = 4)

knitr::opts_chunk$set(message = FALSE,
                      warning = FALSE,
                      comment = NA)
```

Import / Transform

Import

- *read_xlsx() / read_csv() / read_csv2()*
- *options related to separators, number formats, ranges etc.*

Transform

- *rename() / rename_with()*
- *mutate() / mutate(across())*
- *e.g. for log-transformation, creation of factors, text recodings*

Transformations: colnames

```
data(faketrial) # from wrappedtools
colnames(faketrial)[1:10]
```

```
[1] "Sex"                          "Agegroup"
[3] "Treatment"                    "HR"
[5] "sysRR"                        "diaRR"
[7] "Responder"                    "Med Consectetur FakePharm"
[9] "Med Sollicitudin FakePharm"   "Med Suspendisse FakePharm"
```

```
faketrial <-
  rename(.data = faketrial,
         Heartrate = HR) #newname = oldname
faketrial <-
  rename_with(faketrial,
              .fn = ~str_replace(string = .,  #. is placeholder
                                 pattern = "Fa.+$",
                                 replacement = "generic"))
colnames(faketrial)[1:10]
```

```
[1] "Sex"                          "Agegroup"
[3] "Treatment"                    "Heartrate"
[5] "sysRR"                        "diaRR"
[7] "Responder"                    "Med Consectetur generic"
[9] "Med Sollicitudin generic"     "Med Suspendisse generic"
```

Transformations: content

```
ksnormal(faketrial$`Biomarker 1 [units]`)
```

```
[1] 0.01800259
```

```
faketrial <-
  mutate(faketrial,
         `Biomarker 1 ln`=log(`Biomarker 1 [units]`))
ksnormal(faketrial$`Biomarker 1 ln`)
```

```
[1] 0.2090613
```

```
# faketrial |> select(contains('Biomarker 1 ')) |> str()
faketrial <-
  mutate(faketrial,
         across(matches('Biom.+\\]'),
                .fns = ~.x*1000,
                .names = "{.col}x1000"),
         across(starts_with('Med'),
                .fns = factor))
faketrial |>
  select(contains('Biomarker 1 ')) |>
  str()
```

```
tibble [300 × 3] (S3: tbl_df/tbl/data.frame)
 $ Biomarker 1 [units]     : num [1:300] 155 151 140 130 152 ...
 $ Biomarker 1 ln          : num [1:300] 5.05 5.02 4.94 4.86 5.02 ...
 $ Biomarker 1 [units]x1000: num [1:300] 155306 151185 140376 129510 152014 ...
```

Explore / group variables

Explore (general/outlier/distribution)

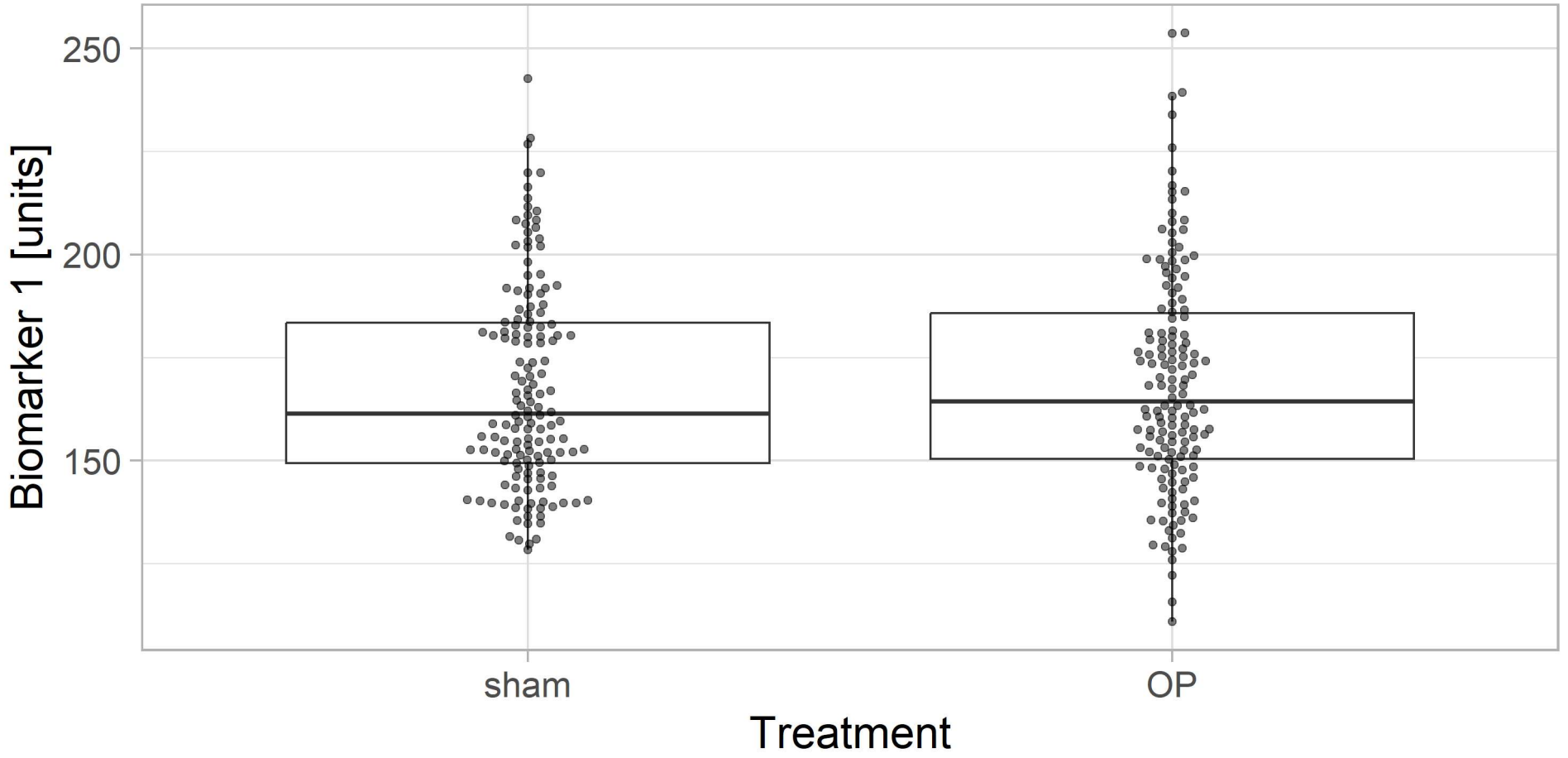
- *ggplot()+geom_boxplot() / geom_beeswarm() / geom_density()*
- *ks.test() / ksnormal() / shapiro.test()*

Classify scale level / distribution

- *gaussvars / ordvars / factvars, possibly more…*
- *Store variables accordingly, e.g. FindVars()*

Explore: Outlier

```
ggplot(faketrial,
       aes(x = Treatment,
           y = `Biomarker 1 [units]`))+
  geom_boxplot(outlier.alpha = 0) + #hide outliers, beeswarm will plot them
  geom_beeswarm(alpha=.5)
```

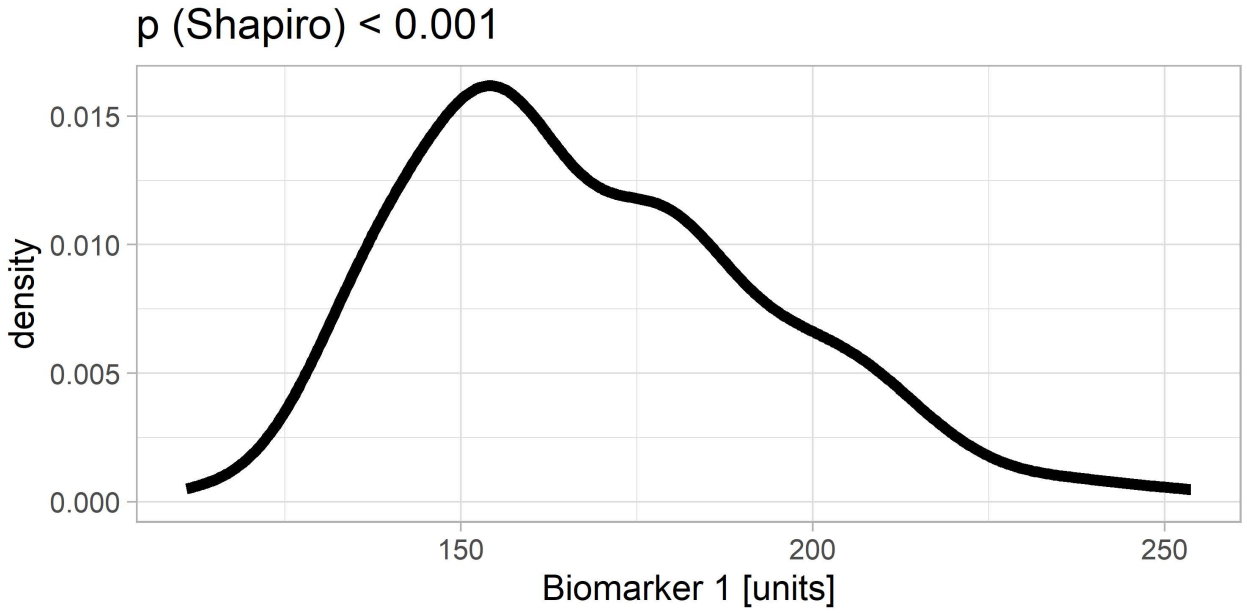


Explore: Normal distribution 1

```r
p_normal <-
  shapiro.test(faketrial$`Biomarker 1 [units]`)
p_normal
```

```
    Shapiro-Wilk normality test

data:  faketrial$`Biomarker 1 [units]`
W = 0.96812, p-value = 3.448e-06
```

```r
ggplot(faketrial,aes(x = `Biomarker 1 [units]`))+
  geom_density(linewidth=3)+
  ggtitle(paste('p (Shapiro)',
                formatP(p_normal$p.value, pretext = T)))
```



Explore: Normal distribution 2

```
faketrial |>
  summarize(across(.cols = starts_with('Biom'),
                   .fns = ksnormal)) |>
  pivot_longer(cols = everything(),
               names_to = 'Variable',values_to = 'pKS')
```

```
# A tibble: 21 × 2
   Variable                 pKS
   <chr>                  <dbl>
 1 Biomarker 1 [units]  0.0180
 2 Biomarker 2 [units]  0.479
 3 Biomarker 3 [units]  0.170
 4 Biomarker 4 [units]  0.976
 5 Biomarker 5 [units]  0.00741
 6 Biomarker 6 [units]  0.675
 7 Biomarker 7 [units]  0.0873
 8 Biomarker 8 [units]  0.944
 9 Biomarker 9 [units]  0.150
10 Biomarker 10 [units] 0.163
# … with 11 more rows
```

Explore: Group variables

```
gaussvars <- FindVars(varnames = c('He','RR'),
                      allnames = cn(faketrial))
gaussvars
```

```
$index
[1] 4 5 6

$names
[1] "Heartrate" "sysRR"     "diaRR"

$bticked
[1] "`Heartrate`" "`sysRR`"     "`diaRR`"

$count
[1] 3
```

```
ordvars <- FindVars(c('B'),
                    allnames = cn(faketrial),
                    exclude = c('x1','ln$'))
ordvars$names |> head(n = 6)
```

```
[1] "Biomarker 1 [units]" "Biomarker 2 [units]" "Biomarker 3 [units]"
[4] "Biomarker 4 [units]" "Biomarker 5 [units]" "Biomarker 6 [units]"
```

```
factvars <- FindVars(c('Sex','Res','generic'),
                     allnames = cn(faketrial))
factvars$bticked |> head(n=4)
```

```
[1] "`Sex`"                        "`Responder`"
[3] "`Med Consectetur generic`"  "`Med Sollicitudin generic`"
```

Model

Describe

- *mean() / sd() / meansd()*
- *median() / quantile() / median_quart()*
- *table() / prop.table() / cat_desc_stats()*

Test

- *t.test() / lm()+[Aa]nova() / compare2numvars()*
- *wilcox.test()*
- *fisher.test() / glm(family=binomial)*

Model: Describe

```r
desc_gauss <- faketrial |>
  summarize(across(.cols = gaussvars$names,
                   .fns = meansd))
desc_gauss
```

```
# A tibble: 1 × 3
  Heartrate sysRR    diaRR
  <chr>     <chr>    <chr>
1 200 ± 22  113 ± 13 83 ± 13
```

```r
desc_ord <- faketrial |>
  summarize(across(ordvars$names,
                   .fns=~median_quart(.x))) |>
  pivot_longer(everything(),
               names_to = 'Measure',
               values_to = 'Median[1Q/3Q]')
desc_ord
```

```
# A tibble: 10 × 2
   Measure               `Median[1Q/3Q]`
   <chr>                 <chr>
 1 Biomarker 1 [units]   163 (149/184)
 2 Biomarker 2 [units]   149 (140/160)
 3 Biomarker 3 [units]   164 (146/188)
 4 Biomarker 4 [units]   148 (139/158)
 5 Biomarker 5 [units]   163 (149/189)
 6 Biomarker 6 [units]   149 (139/159)
 7 Biomarker 7 [units]   167 (147/189)
 8 Biomarker 8 [units]   149 (139/160)
 9 Biomarker 9 [units]   165 (145/191)
10 Biomarker 10 [units] 148 (140/157)
```

Model: Test 1 / single variables

```r
#t-Test with test for equal variances
t.test(formula=sysRR~Treatment, data=faketrial,
       var.equal=var.test(formula=sysRR~Treatment,
                          data=faketrial)$p.value>.05)
```

```
    Welch Two Sample t-test

data:  sysRR by Treatment
t = -9.4166, df = 273.41, p-value < 2.2e-16
alternative hypothesis: true difference in means between group sham and group OP is not
equal to 0
95 percent confidence interval:
 -15.063528  -9.854138
sample estimates:
mean in group sham   mean in group OP
          107.0578           119.5166
```

```r
#Wilcoxon-Test
wilcox.test(`Biomarker 1 [units]`~Treatment,
            data = faketrial)
```

```
    Wilcoxon rank sum test with continuity correction

data:  Biomarker 1 [units] by Treatment
W = 10905, p-value = 0.6465
alternative hypothesis: true location shift is not equal to 0
```

Model: Test 2 / multiple variables

```r
test_gauss <- compare2numvars(data = faketrial,
                              dep_vars = gaussvars$names,
                              indep_var = 'Treatment',
                              gaussian = TRUE,
                              round_p = 5)
test_gauss
```

```
# A tibble: 3 × 5
  Variable  desc_all `Treatment sham` `Treatment OP` p
  <fct>     <chr>    <chr>           <chr>          <chr>
1 Heartrate 200 ± 22 201 ± 24        200 ± 20       0.65363
2 sysRR     113 ± 13 107 ± 13        120 ± 10       0.00001
3 diaRR     83 ± 13  76 ± 13         90 ± 9         0.00001
```

```r
test_ord <- compare2numvars(data = faketrial,
                            dep_vars = ordvars$names,
                            indep_var = 'Treatment',
                            gaussian = FALSE)
test_ord |> slice_head(n = 5)
```

```
# A tibble: 5 × 5
  Variable             desc_all      `Treatment sham` `Treatment OP` p
  <fct>                <chr>         <chr>            <chr>          <chr>
1 Biomarker 1 [units] 163 (149/184) 161 (149/184)    164 (150/186)  0.647
2 Biomarker 2 [units] 149 (140/160) 147 (139/159)    150 (141/160)  0.205
3 Biomarker 3 [units] 164 (146/188) 162 (147/186)    166 (146/189)  0.627
4 Biomarker 4 [units] 148 (139/158) 148 (137/158)    148 (141/158)  0.493
5 Biomarker 5 [units] 163 (149/189) 161 (147/187)    168 (150/194)  0.260
```

Model: linear models 1 / univariable

```r
lm1<- lm(sysRR~Agegroup, data=faketrial)
lm1
```

```
Call:
lm(formula = sysRR ~ Agegroup, data = faketrial)

Coefficients:
   (Intercept)  Agegroupmiddle     Agegroupold
       110.669          -1.202           9.056
```

```r
anova(lm1)
```

```
Analysis of Variance Table

Response: sysRR
           Df Sum Sq Mean Sq F value   Pr(>F)
Agegroup    2   6289 3144.70      21 2.95e-09 ***
Residuals 297  44476  149.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#post-hoc
pairwise.t.test(x = faketrial$sysRR, g = faketrial$Agegroup, p.adjust.method = 'fdr')
```

```
    Pairwise comparisons using t tests with pooled SD

data:  faketrial$sysRR and faketrial$Agegroup

       young   middle
middle 0.49    -
old    4.7e-07 2.6e-08
```

Model: linear models 2 / multivariable

```r
lm2<- lm(sysRR~(Sex+Agegroup)*Treatment,
         data=faketrial)
lm2
```

```
Call:
lm(formula = sysRR ~ (Sex + Agegroup) * Treatment, data = faketrial)

Coefficients:
               (Intercept)                     Sexmale
                   99.9852                      2.8812
           Agegroupmiddle                  Agegroupold
                   -0.8378                     18.2524
              TreatmentOP         Sexmale:TreatmentOP
                   21.3694                     -4.7366
Agegroupmiddle:TreatmentOP     Agegroupold:TreatmentOP
                   -1.2218                    -19.0720
```

```r
Anova(lm2,type = 3)
```

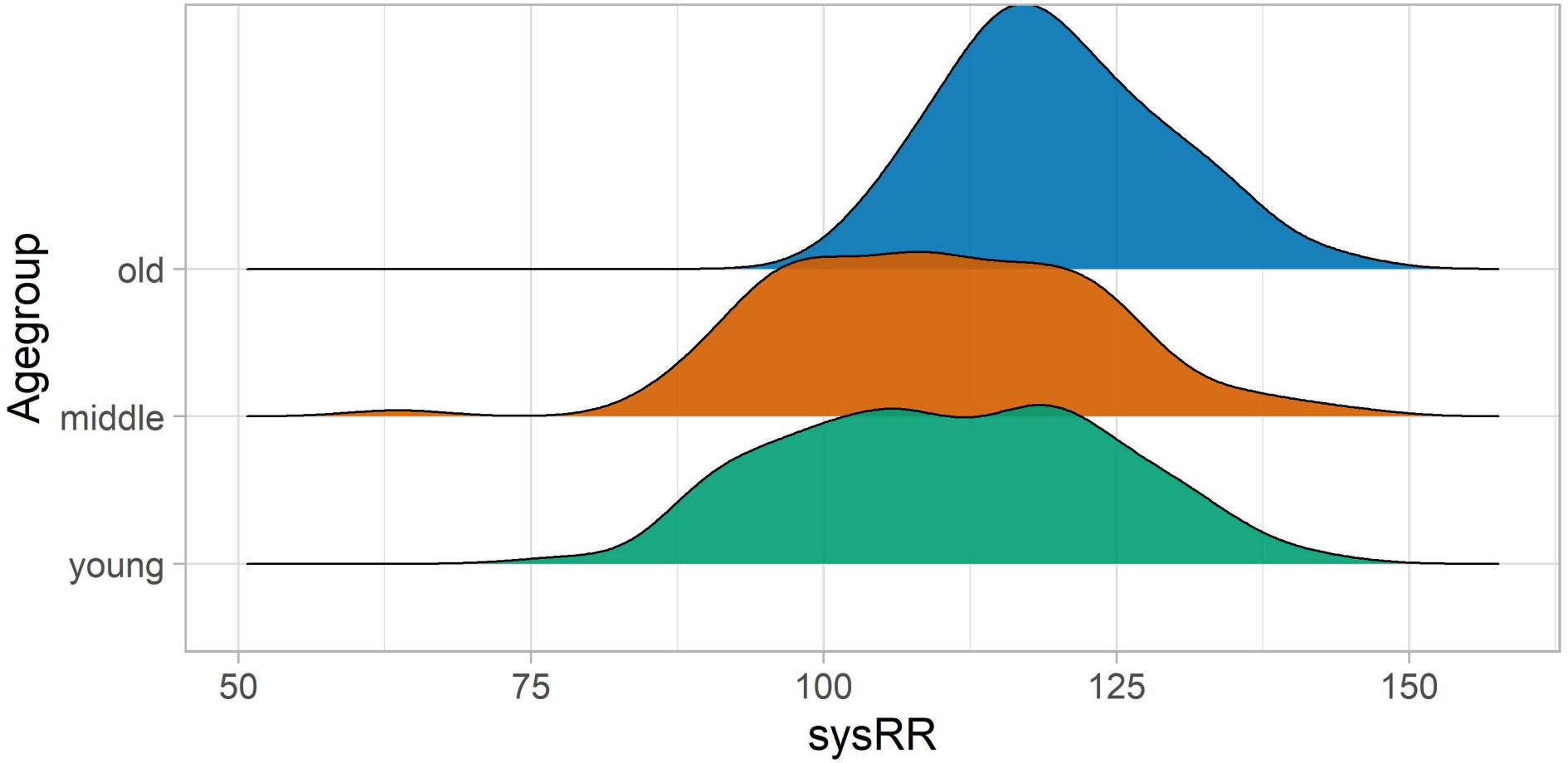
```
Anova Table (Type III tests)

Response: sysRR
                    Sum Sq  Df   F value    Pr(>F)
(Intercept)         424455   1 4632.6739 < 2.2e-16 ***
Sex                    303   1    3.3052   0.07009 .
Agegroup             11605   2   63.3323 < 2.2e-16 ***
Treatment             8761   1   95.6165 < 2.2e-16 ***
Sex:Treatment          410   1    4.4724   0.03529 *
Agegroup:Treatment    5635   2   30.7511 7.597e-13 ***
Residuals            26754 292
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualize 1

```
ggplot(faketrial,aes(x = Treatment, y = sysRR))+
  geom_violin(draw_quantiles = c(.25,.5,.75))+
  geom_signif(comparisons = list(c(1,2)),
              annotations =
                paste('p',
                      formatP(test_gauss$p[2],
                              pretext = T)))+
  scale_y_continuous(expand = expansion(mult = .1))
```

Visualize 2

```
1  agecolors <- c("#009E73", "#D55E00", "#0072B2")
2  ggplot(faketrial,aes(x = sysRR, y = Agegroup, fill=Agegroup))+
3    geom_density_ridges(alpha=.9)+
4    guides(fill='none')+
5    scale_fill_manual(values = agecolors)
```



Report

- *RMarkdown and quarto are powerful tools to create reports*
- *Package flextable provides nice features for table formatting*

```
test_gauss |>
  flextable() |>
  bg(~p<=0.05,j = 5,bg = 'yellow') |>
  set_caption('Treatment effects, measures following a normal distribution') |>
  add_footer_lines('Significance level is set at 0.05')
```

Treatment effects, measures following a normal distribution

Variable	desc_all	Treatment sham	Treatment OP	p
Heartrate	200 ± 22	201 ± 24	200 ± 20	0.65363
sysRR	113 ± 13	107 ± 13	120 ± 10	0.00001
diaRR	83 ± 13	76 ± 13	90 ± 9	0.00001

Significance level is set at 0.05