Lab: Complete ANOVA lab

Introduction

In this lab we will work on a complete problem based on ANOVA, from start to finish. We will load the data, clean it up, produce some graphs, and run appropriate analyses.

```
library(haven)
targeting <- read_sav("~/statsLabPractice/targetting/targeting.sav")
# View(targeting)
```

Factor variables:

- target race (White/Black): Coded in the variable name
- object (armed/unarmed): Coded in the variable name but implicitly
- shot action taken (correct/incorrect): Coded in the variable name but implicitly

Let's take a look at the variables:

```
names(targeting)
```

We will only need the first 12 variables, the remaining are computed quantities. We start by gathering the 8 columns that contain observations (don't worry about the warning):

```
targetingLong <- targeting %>%
    select(1:12) %>%
    gather(key="key", value="meanRT", 3:10)
```

Next we need break the key variable into two parts, one showing the race and another showing the outcome. We'll first split at the underscore, and basically discard the left part:

```
targetingLong <- targeting %>%
    select(1:12) %>%
    gather(key="key", value="meanRT", 3:10) %>%
    separate(key, c("_ignore", "key"), "_")
```

Now we split the new key variable in two parts, splitting after the first 5 characters:

```
targetingLong <- targeting %>%
    select(1:12) %>%
    gather(key="key", value="meanRT", 3:10) %>%
    separate(key, c("_ignore", "key"), "_") %>%
    separate(key, c("race", "outcome"), 5) %>%
    select(-starts_with("_ignore"))
```

Next we need to work on the outcome variable, which actually contains two different pieces of information:

- whether the object was armed (Hits/Misses) vs unarmed (CRs/FAs)
- whether the subject took the correct action (Hits/CRs) or incorrect action (Misses/FAs)

We will use mutate and recode_factor to create these:

```
targetingFinal <- targetingLong %>%
    mutate(object=recode_factor(outcome, Hits="Armed", Misses="Armed",
                                CRs="Unarmed", FAs="Unarmed"),
           action=recode_factor(outcome, Hits="Correct", Misses="Incorrect",
                                CRs="Correct", FAs="Incorrect"))
```

To double-check that we did this correctly, we'll create counts:

```
targetingFinal %>%
    group_by(race, outcome, object, action) %>%
    summarize(count=n())
```

We should see 49 cases for each, corresponding to our initial 49 data rows.

Finally, a couple more cleanup steps are in order before we move on:

- We should fix the names of some of the variables. We will use rename for that.
- We should drop the outcome column as it is no longer needed. We will use select for that.
- The gender, race and age variables need to be coded as factors. We will use mutate and factor for that (we would use recode_factor if we wanted to change the names of the labels, but we don't).

This can all be done in a series of pipelined steps.

```
targetingFinal <- targetingFinal %>%
    rename(subject="script.subjectid",
           iat="expressions.d",
           gender="gender_response",
           age="age_response") %>%
    select(-starts_with("outcome")) %>%
    mutate(gender=factor(gender), age=factor(age), race=factor(race))
```

Lastly, we will filter the action column to only include the correct answers, as that was the focus of the study.

```
targetingCorrect <- targetingFinal %>% filter(action=="Correct")
```

There are of course numerous graphs we can construct, and we can choose to log-transform the mean reaction time or not. We will leave it as is for now as the data did not show any signs of extreme skewness. Here is a starting plot that shows the density distribution for meanRT for armed and unarmed objects, and with different graphs for each race and gender combination:

```
ggplot(targetingCorrect) +
    aes(x=meanRT, color=object) +
    geom_density() +
    facet_grid(race~gender)
```

We can see that mean reaction times were slower for the unarmed objects.

Let us compute some numerical summaries:

```
targetingCorrect %>%
    group_by(race, object, gender) %>%
    summarize(mean=mean(meanRT),
              se=sd(meanRT)/sqrt(n()))
```

We can also plot these:

```
ggplot(targetingCorrect) +
    aes(x=object, y=meanRT, color=race) +
    stat_summary(fun.data=mean_se, position=position_dodge(0.2)) +
    facet_wrap(~gender)
```

We probably expected the marked difference in reaction times between armed and unarmed subjects. For female subjects, the race of the subject seems to play a small factor.

```
fit1 <- lm(meanRT~race*object, data=targetingCorrect)
summary(fit1)
anova(fit1)
```

We can see a significant overall effect, but we can also see that the interaction terms are not significant. We remove them from the model:

```
fit2 <- lm(meanRT~race+object, data=targetingCorrect)
summary(fit2)
anova(fit2)
```

We can compare the two models to see if there are differences, and there is no significant difference:

```
anova(fit1, fit2)
```

We can get some default diagnostics from plotting the fit:

```
par(mfrow=c(2,2))
plot(fit2)
```

The residuals appear to be normal and with constant variance. We can visualize their effect against the other predictors:

```
ggplot(targetingCorrect) +
    aes(x=object, y=resid(fit2), color=race) +
ggplot(targetingCorrect) +
    aes(x=race, y=resid(fit2), color=object) +
    geom_point(position=position_jitter(0.1))
ggplot(targetingCorrect) +
    aes(x=race, y=resid(fit2), color=object) +
    geom_point(position=position_dodge(0.1))
ggplot(targetingCorrect) +
    aes(x=gender, y=resid(fit2), color=interaction(race, object)) +
    geom_point(position=position_dodge(0.2))
```

We can look at how iat might be related to those residuals, there's clearly a relation there:

```
ggplot(targetingCorrect) +
    aes(x=iat, y=resid(fit2), color=object) +
    geom_point() +
    geom_smooth()
```

Now let's add the subject's gender into the model:

```
fit3 <- lm(meanRT~race+object+gender+race:gender+object:gender, data=targetingCorrect)
summary(fit3)
anova(fit3)
anova(fit2, fit3)
```

We see that the subject's gender does not appear to be significant.

Finally, we look at whether we should remove race from the model as well:

```
fit4 <- lm(meanRT~object, data=targetingCorrect)
summary(fit4)
anova(fit4)
anova(fit4, fit3)
```

```
fit5 <- lm(meanRT~poly(iat, 2)+object, data=targetingCorrect)
summary(fit5)
anova(fit5)
```

ggplot(targetingCorrect) + aes(x=race, y=meanRT, color=gender) + geom_point() + geom_line(aes(group=subject))