Plotting and NHST

Take-home quiz #1

The Solution

Getting started

Identify and prepare data

ToothGrowth is the correct dataset for this assignment.

```
data(ToothGrowth)
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Load required packages

```
pacman::p_load(multcomp, tidyverse, gvlma, xtable)
  # Load xtable instead of pander to support cross-referencing.
  # It is only available for knitting to .pdf.
  # pander is a good option when knitting to Word.
# Load custom function table_glht() from
# https://gist.github.com/ajpelu/194e721077ec045a2b864088908e7aca
source("https://gist.githubusercontent.com/ajpelu/194e721077ec045a2b864088908e7aca/raw/e8002861fd1b99d8
```

Distribution of response variable

Plot

```
ToothGrowth %>%
  ggplot(aes(x=len)) + theme_bw(16) +
      geom_histogram(aes(y=..density..),
                     binwidth=2,
                     colour="black",
                     fill="lightgreen") +
      geom_density(alpha=.2, fill="#FF6666") +
      xlim(c(-5, 45))
```

Data model

- A Probability Density Function describes the range of all possible values of a variable, and gives the likelihood of each value occuring in a randomly-drawn sample. The area below the PDF and above the X axis integrates

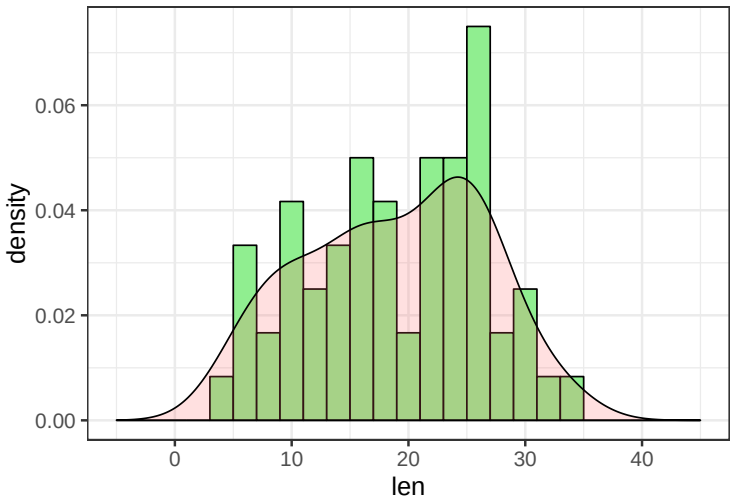

Figure 1: Histogram and density of `len` in the `ToothGrowth` dataset.

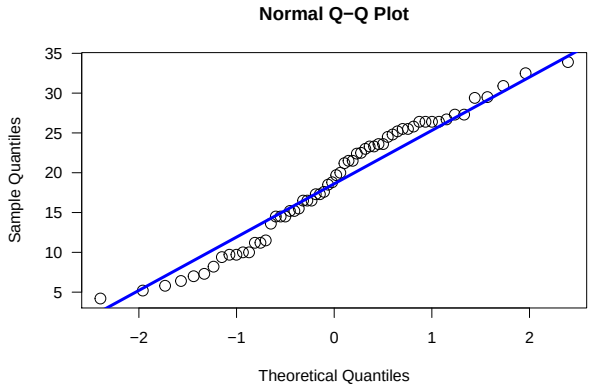

Figure 2: Q-Q plot for variable `len` in `ToothGrowth` dataset.

to 1, or the sum of all probabilities. When fit to a sample, it represents the theoretical distribution of data in the population from which the sample was drawn.

- Data in Fig. 1 are best modelled with a normal (Gaussian) distribution.

```
qqnorm(ToothGrowth$len, cex=1.5, pch=1, las=1,
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")
qqline(ToothGrowth$len, distribution = qnorm,
       probs = c(0.025, 0.975), col="blue", lwd=3)
```

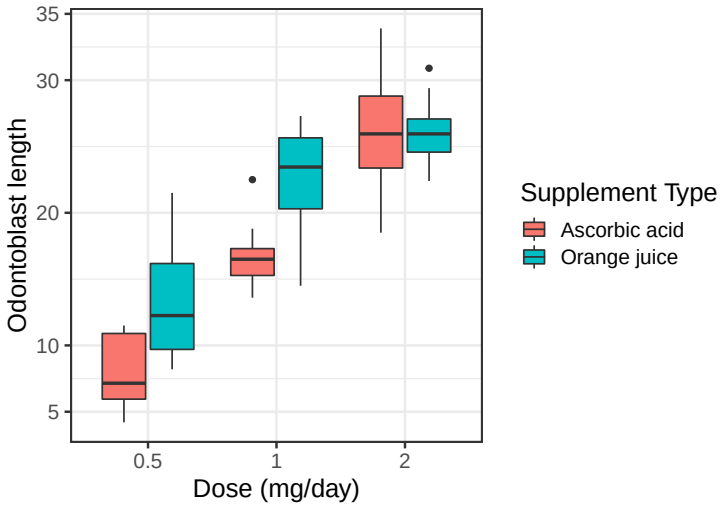

Figure 3: Boxplot of `len` against `dose` conditioned by `supp` from `ToothGrowth`.

Data visualization

```
ToothGrowth %>%
  mutate(supp = plyr::revalue(as.character(supp),
                              c("OJ"="Orange juice",
                                "VC"="Ascorbic acid"))) %>%
ggplot(aes(x=dose, y=len)) + theme_bw(16) +
 geom_boxplot(aes(fill=supp)) +
 labs(y="Odontoblast length",
      x="Dose (mg/day)") +
  scale_fill_discrete(name="Supplement Type") +
  scale_y_continuous(breaks=c(5,10, 20, 30, 35))
```

Model fitting

Which statistical model?

- As these data are a continuous variable fit against two categorical variables, an ANOVA is most appropriate.
- H_0: Odontoblast length is unaffected by either supplement at any level of dosage.
  H_1: Odontoblast length increases as supplement dosage increases.

Run a test

```
lm(len ~ dose + supp, ToothGrowth) %>%
  anova( )%>%
    xtable(caption="ANOVA results testing the fit of len against dose + supp.",
         label = "tab:anova" ) %>%
      print(comment=FALSE)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dose	2	2426.43	1213.22	82.81	0.0000
supp	1	205.35	205.35	14.02	0.0004
Residuals	56	820.43	14.65		

Table 1: ANOVA results testing the fit of len against dose + supp.

Post-hoc model validation

```
mod.sum <-summary(gvlma(len ~ dose + supp, ToothGrowth) )
```

```
xtable(mod.sum, caption="The ANOVA model meets the necessary assumptions.",
       label="tab:gvlma") %>%
  print(comment=FALSE)
```

	Value	p-value	Decision
Global Stat	8.05	0.09	Assumptions acceptable.
Skewness	3.31	0.07	Assumptions acceptable.
Kurtosis	0.05	0.82	Assumptions acceptable.
Link Function	4.40	0.04	Assumptions NOT satisfied!
Heteroscedasticity	0.29	0.59	Assumptions acceptable.

Table 2: The ANOVA model meets the necessary assumptions.

Post-hoc comparisons

```
lm(len ~ dose + supp, ToothGrowth) %>%
  glht(., linfct = mcp(dose = "Tukey")) %>%
    summary( ) %>%
      table_glht( ) %>%
    xtable(caption="Results of Tukey post-hoc pairwise comparison on dose.",
           label="tab:glht")   %>%
      print(comment=FALSE)
```

	Estimate	Std. Error	t value	Pr(>|t|)
1 - 0.5	9.13	1.21	7.54	0.00
2 - 0.5	15.49	1.21	12.80	0.00
2 - 1	6.37	1.21	5.26	0.00

Table 3: Results of Tukey post-hoc pairwise comparison on dose.

Conclusions

ANOVA confirms that both supplement type and daily dosage have significant effects on odontoblast length (Table 1). Pair-wise poc-hoc comparisons indicate that each of the three dosage levels are different from each other (Table 3). These results provide evidence to accept H_1 over H_0.

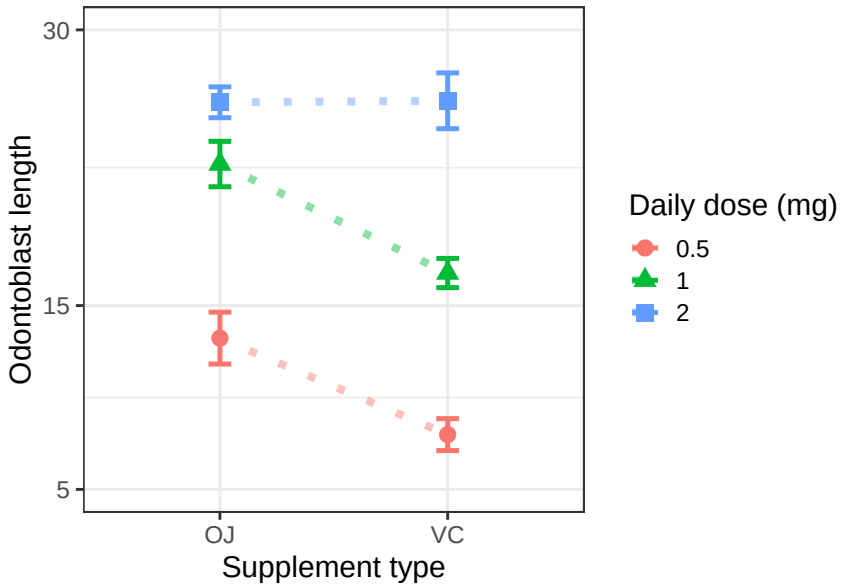

Figure 4:

Another statistical test?

Additional data visualization

```
ToothGrowth %>%
  group_by(supp, dose) %>%
     summarize(mean=mean(len),
               se=sd(len)/sqrt(length(len)) ) %>%
ggplot(aes(x=supp, y=mean,
           color=dose,
           shape=dose)) +
           theme_bw(16) +
              labs(x="Supplement type",
                   y="Odontoblast length") +
              scale_shape_discrete(name="Daily dose (mg)") +
              scale_color_discrete(name="Daily dose (mg)")+
  geom_line(aes(group=dose),
            linetype= "dotted",
               size=2, alpha=0.45) +
 geom_errorbar(aes(ymax=mean+se,
                   ymin=mean-se),
                width=0.1, size=1.25) +
 geom_point(size=4) +
  scale_y_continuous(breaks=c(5,15,30)) +
  coord_cartesian(ylim = c(5,30))
```

Fig. 4 connects the mean response of each dosage by supplement, and highlights disparities between the supplements at each dose. The slopes of the three lines are not consistent, but nor do they cross within the bound of the data plotted, which suggests the ANOVA model should be updated to test for a significant ordinal interaction.

```
lm(len ~ dose + supp + dose:supp, ToothGrowth) %>%
  anova() %>%
    xtable(caption = "Updated ANOVA model testing for interaction.",
           label = "intmod") %>%
      print(comment=FALSE)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dose	2	2426.43	1213.22	92.00	0.0000
supp	1	205.35	205.35	15.57	0.0002
dose:supp	2	108.32	54.16	4.11	0.0219
Residuals	54	712.11	13.19		

Table 4: Updated ANOVA model testing for interaction.

```
mod.sum2 <-summary(gvlma(len ~ dose + supp + dose:supp, ToothGrowth) )
```

```
xtable(mod.sum2, caption="The updated ANOVA model also meets the necessary assumptions.",
       label="tab:gvlma2") %>%
  print(comment=FALSE)
```

	Value	p-value	Decision
Global Stat	0.49	0.97	Assumptions acceptable.
Skewness	0.27	0.60	Assumptions acceptable.
Kurtosis	0.05	0.82	Assumptions acceptable.
Link Function	0.00	1.00	Assumptions acceptable.
Heteroscedasticity	0.16	0.69	Assumptions acceptable.

Table 5: The updated ANOVA model also meets the necessary assumptions.

```
lm(len ~ dose + supp + dose:supp, ToothGrowth) %>%
  glht(., linfct = mcp(dose = "Tukey")) %>%
    summary( ) %>%
      table_glht( ) %>%
    xtable(caption="Results of Tukey post-hoc pairwise comparison on dose in the ANOVA
           model testing for an interaction between dose and supp.",
           label="tab:glht2")   %>%
      print(comment=FALSE)
```

	Estimate	Std. Error	t value	Pr(>|t|)
1 - 0.5	9.47	1.62	5.83	0.00
2 - 0.5	12.83	1.62	7.90	0.00
2 - 1	3.36	1.62	2.07	0.11

Table 6: Results of Tukey post-hoc pairwise comparison on dose in the ANOVA model testing for an interaction between dose and supp.

There is a signficant interaction between dosage level and supplement type ($F = 4.12$, $P = 0.02$). However, because the interaction appears ordinal (Fig. 4), we can still make claims about the main effects in the model. Specifically, greater dosages of both supplement types increase odontoblast length, although orange juice is more effective at low and moderate dosages. At 2 mg/day, there is no difference between the supplement types. Furthermore, in the presence of the significant interaction, the two highest dosages are no longer significantly different from each other in the pairwise comparison ($t = 2.1$, $P = 0.11$; Table 6).