SHC 798 Assignment 2, 2025

Richard Lubega

2025-10-06

SHC 798 Assignment 2, 2025

Analysis of Variance (ANOVA)

Question 4: # Compressive strength results

```
pacman::p_load(tidymodels)

# Getting started with the dataset in timber.csv :
timber <- read.csv(file.choose(), header = TRUE, na.strings = c("NA"))
# timber
head(timber)
```

```
##   species stiffness
## 1    pine    7897.6
## 2    pine    8239.5
## 3    pine    7740.3
## 4    pine    7722.1
## 5    pine    8982.9
## 6    pine    8696.7
```

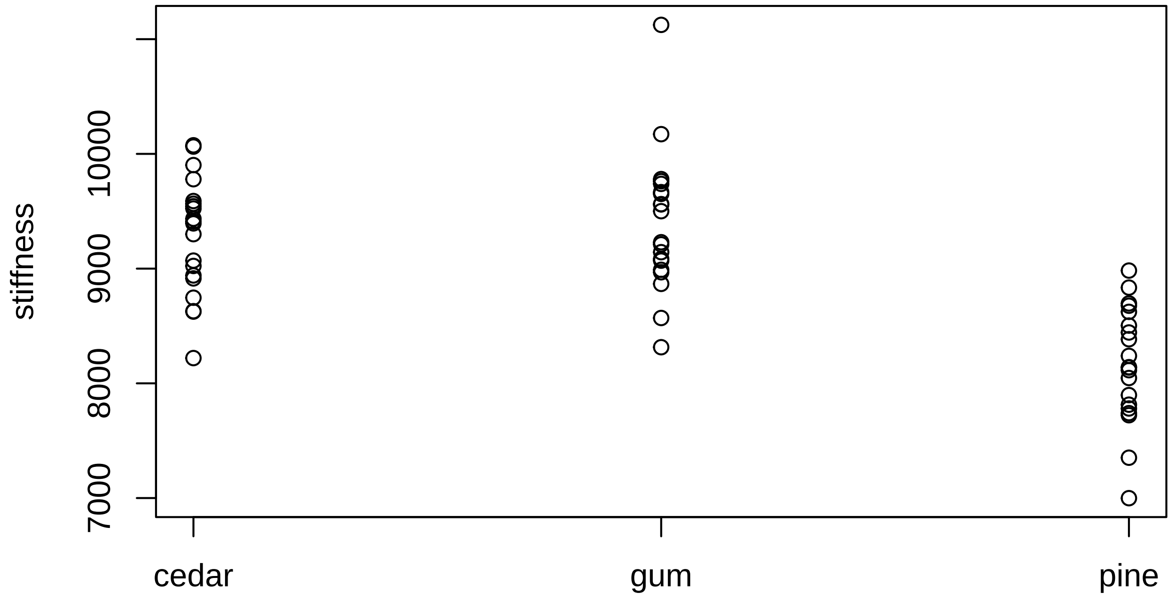
```
str(timber)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ species  : chr  "pine" "pine" "pine" "pine" ...
##  $ stiffness: num  7898 8240 7740 7722 8983 ...
```

```
## Convert species column to a factor
timber$species <- factor(timber$species)
## Check levels
levels(timber$species)
```

```
## [1] "cedar" "gum"   "pine"
```

```
## Visualize data
stripchart(stiffness ~ species, data = timber, pch = 1, vertical = TRUE)
```



```
# Summary statistics by species
summary_stats <- timber %>%
  group_by(species) %>%
  summarise(
    Mean = mean(stiffness),
    SD = sd(stiffness),
    Median = median(stiffness),
    IQR = IQR(stiffness),
    Q1 = quantile(stiffness, 0.25),
    Q3 = quantile(stiffness, 0.75),
    Min = min(stiffness),
    Max = max(stiffness)
  )

# Identify outliers using IQR method
outliers <- timber %>%
  group_by(species) %>%
  mutate(
    Q1 = quantile(stiffness, 0.25),
    Q3 = quantile(stiffness, 0.75),
    IQR = Q3 - Q1,
    Lower_Bound = Q1 - 1.5 * IQR,
    Upper_Bound = Q3 + 1.5 * IQR,
    Outlier = stiffness < Lower_Bound | stiffness > Upper_Bound
  ) %>%
  filter(Outlier) %>%
```

```r
  select(species, stiffness)

# Print summary statistics and outliers
print("Summary Statistics:")
```

```
## [1] "Summary Statistics:"
```

```r
print(summary_stats)
```

```
## # A tibble: 3 x 9
##   species  Mean    SD Median   IQR    Q1    Q3   Min    Max
##   <fct>   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 cedar   9288.  506.  9403.  638. 8934. 9571. 8220. 10075.
## 2 gum     9398.  607.  9365.  635. 9050. 9684. 8315. 11124.
## 3 pine    8156.  506.  8139.  728. 7806. 8534. 6999.  8983.
```

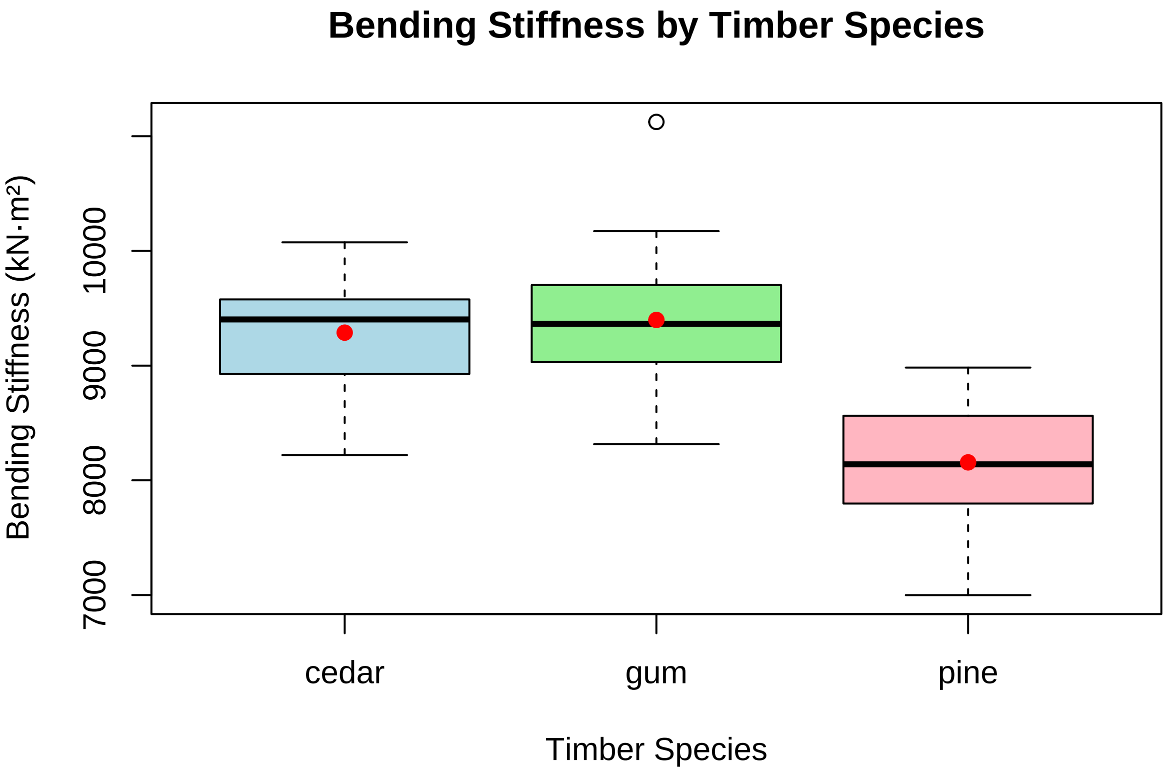
```r
print("Outliers:")
```

```
## [1] "Outliers:"
```

```r
print(outliers)
```

```
## # A tibble: 1 x 2
## # Groups:   species [1]
##   species stiffness
##   <fct>       <dbl>
## 1 gum        11124.
```

```r
# Part a): # Box Plots
boxplot(stiffness ~ species,
        data = timber,
        main = "Bending Stiffness by Timber Species",
        xlab = "Timber Species",
        ylab = "Bending Stiffness (kN·m²)",
        col = c("lightblue", "lightgreen", "lightpink"),
        border = "black")
#Adding means as points
means <- tapply(timber$stiffness, timber$species, mean)
points(1:3, means, pch = 19, col = "red")
```



```
# Annotate outliers on the plot
# text(x = bp$group, y = bp$out, labels = bp$out, pos = 3, cex = 0.7, col = "darkblue")
```

Commenting on Variability and Outliers *Variability*: Standard Deviation (SD): Gum has the highest variability (SD = 641.0 kN · m²), followed by cedar (SD = 572.8 kN · m²), and pine has the lowest (SD = 552.6 kN · m²). This suggests that gum's stiffness values are more spread out compared to pine and cedar.

Interquartile Range (IQR): Pine has the highest IQR (696.8 kN · m²), indicating a slightly wider spread of the middle 50% of stiffness values compared to gum (673.8 kN · m²) and cedar (659.1 kN · m²). However, the differences in IQR are small, suggesting comparable spread in the central data across species.

Range (Max - Min): Gum shows the largest range (11124.5 - 8314.9 = 2809.6 kN · m²), followed by cedar (10074.9 - 8220.3 = 1854.6 kN · m²), and pine (8982.9 - 6999.2 = 1983.7 kN · m²). This reinforces that gum has the most extreme values.

Central Tendency: Gum has the highest median stiffness (9425.3 kN · m²), followed by cedar (9387.7 kN · m²), and pine (8139.2 kN · m²). This indicates that gum and cedar generally have higher bending stiffness than pine.

In short;

Variability: Gum exhibits the highest variability in bending stiffness, as seen in its larger SD and range, suggesting less consistency in its mechanical properties compared to pine and cedar. Cedar and pine have similar variability, but pine's stiffness values are generally lower.

Outliers: Pine has one low outlier, indicating a single timber sample with unusually low stiffness, possibly due to defects or testing conditions. Gum has both a high and a low outlier, suggesting it can exhibit extreme stiffness values (both stronger and weaker), which may reflect natural variability or quality differences in the samples. Cedar's lack of outliers suggests greater consistency in its stiffness properties.

Practical Implications: If consistency is desired, cedar may be preferable due to its lack of outliers and moderate variability. Gum's higher median stiffness is appealing for strength, but its outliers and variability suggest a need for quality control. Pine's lower stiffness and single outlier may indicate it's less suitable for applications requiring high or consistent stiffness.

```r
# Part b):# Fit a one-way ANOVA test
timber$species <- relevel(timber$species, ref = "gum")
options(contrasts = c("contr.sum", "contr.poly"))
# options(contrasts = c("contr.treatment", "contr.poly")) # used as default anyway

stiff <- aov(stiffness ~ species, data = timber)
summary(stiff) ## ANOVA table including F-test
```

```
##             Df   Sum Sq Mean Sq F value   Pr(>F)    
## species      2 18889629 9444815   32.17 4.45e-10 ***
## Residuals   57 16734248  293583                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Extract coefficients
coef(stiff)       ## be careful with interpretation
```

```
## (Intercept)    species1    species2 
##   8947.4233    450.8267    340.1017
```

```r
dummy.coef(stiff) ## full coefficients, easier to interpret
```

```
## Full coefficients are 
##                                          
## (Intercept):     8947.423                
## species:              gum     cedar      pine
##                  450.8267  340.1017 -790.9283
```

Model Interpretation

Null hypothesis (H): $_{\text{pine}} = _{\text{gum}} = _{\text{cedar}}$ (All species have the same mean bending stiffness)

Alternative hypothesis (H_A): At least one species has a different mean stiffness.

F-statistic: 32.17; Large F-value indicates that between-species variability is much greater than within-species variability.

p-value: $4.45 \times 10^{\ 1}$;Extremely small (< 0.001), so we reject H at any conventional significance level (e.g., 0.05 or 0.01).

Conclusion: There is **strong statistical evidence** that the mean bending stiffness differs significantly between timber species.

```r
# Part c) # A pairwise two-sample t-tests (with multiple comparison correction)
# Perform pairwise t-tests with Bonferroni correction
tapply(timber$stiffness, timber$species, sd) # check for group SD
```

```
##      gum    cedar     pine
## 607.0210 505.7786 506.4222
```

```r
tapply(timber$stiffness, timber$species, var) # check for group var
```

```
##      gum    cedar     pine
## 368474.5 255812.0 256463.4
```

```r
pairwise_results <- pairwise.t.test(timber$stiffness, timber$species,
                                    p.adjust.method = "bonferroni",
                                    pool.sd = FALSE, # Welch's t-test (unequal variances)
                                    paired = FALSE,  # Independent samples
                                    conf.level = 0.95)

# Print the results
print("Pairwise t-test results with Bonferroni correction:")
```

```
## [1] "Pairwise t-test results with Bonferroni correction:"
```

```r
print(pairwise_results)
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  timber$stiffness and timber$species
##
##       gum     cedar
## cedar 1       -
## pine  8.1e-08 6.0e-08
##
## P value adjustment method: bonferroni
```

Test Interpretation

Test method: Welch-adjusted pairwise t-test because the groups have unequal variances. This adjusts the degrees of freedom for each pair according to Welch's formula.

Null hypothesis (H): $_{\text{pine}} = _{\text{gum}}$ or $_{\text{pine}} = _{\text{cedar}}$ or $_{\text{gum}} = _{\text{cedar}}$ (mean bending stiffness is the same) and the *alternative hypothesis* (H_A): Means differ. We reject H if $p < 0.05$.

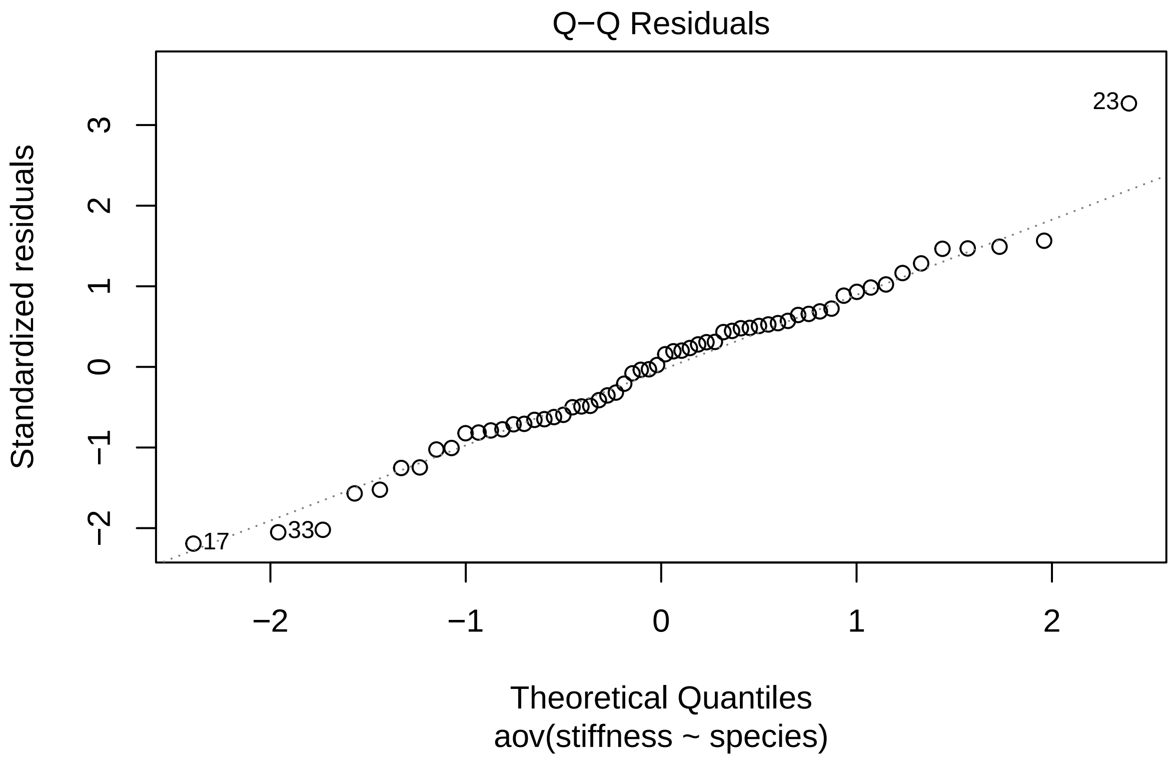
Interpretation for each pair pine vs gum: $p = 8.1 \times 10 < 0.05$ (significant) implying that mean stiffness differs between pine and gum, hence gum is stiffer than pine (looking at raw data: gum = 9500 vs pine = 8100).

pine vs cedar: $p = 6.0 \times 10 < 0.05$ (significant) implying that mean stiffness differs between pine and cedar hence cedar is stiffer than pine (cedar = 9400).

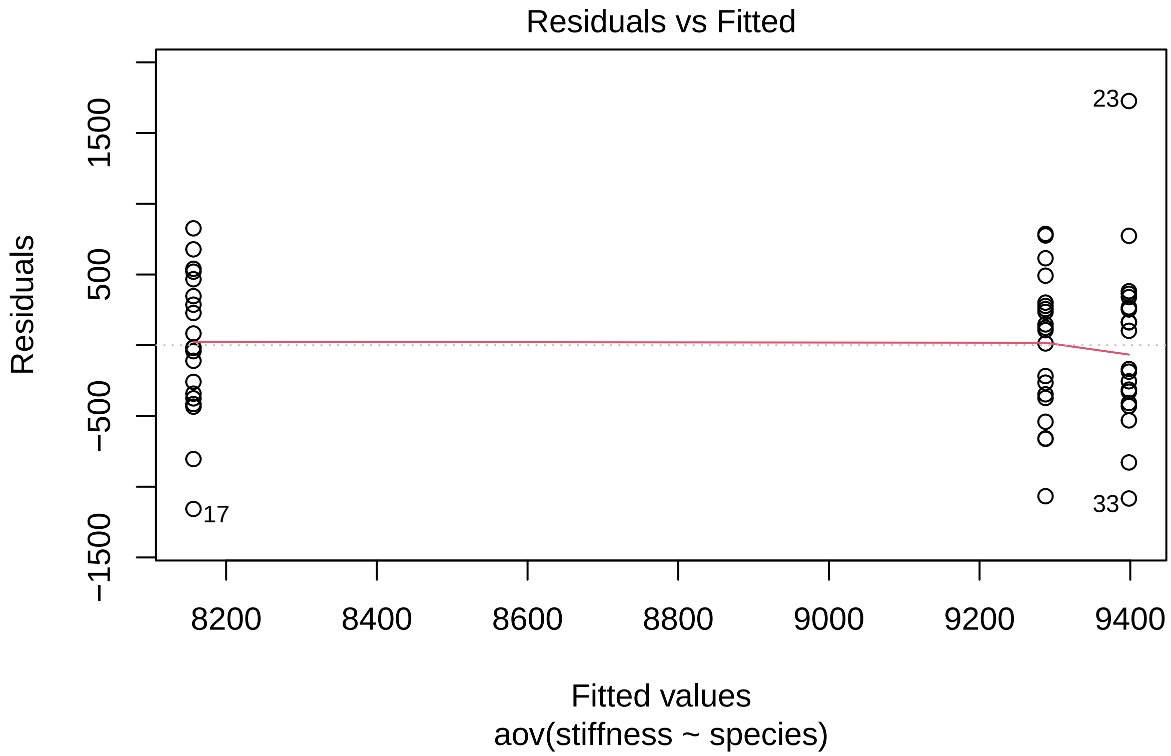
gum vs cedar: $p = 1$ (not significant) implying that there is no evidence that gum and cedar differ in mean stiffness. Their stiffness values are roughly similar (gum = 9500, cedar = 9400).

Therefore, practically, pine is the softest whereas gum and cedar have *similar* higher stiffness.

```
# Part d) ## Residual Diagnostics
plot(stiff, which = 2)
```



```
plot(stiff, which = 1)
```



From the residual plots (TA), error variance is constant and error can be expected to be zero. Errors are i.i.d (from Q-Q plot). No autocorrelation.