Analysis of Wolf River pollution

This program reads data on the relationship sampling depth and two pollutant concentrations. Find more information in the data dictionary.

This program was written by Steve Simon and Leroy Wheeler on 2024-10-23 and is placed in the public domain.

Load the tidyverse library

For most of your programs, you should load the tidyverse library. The messages and warnings are suppressed.

```
library(broom)
library(tidyverse)
```

Read the data

```
river <- read_tsv(
  file="../data/wolf-river-pollution.txt",
  col_names=TRUE,
  col_types="cnn")
names(river) <- tolower(names(river))
glimpse(river)
```

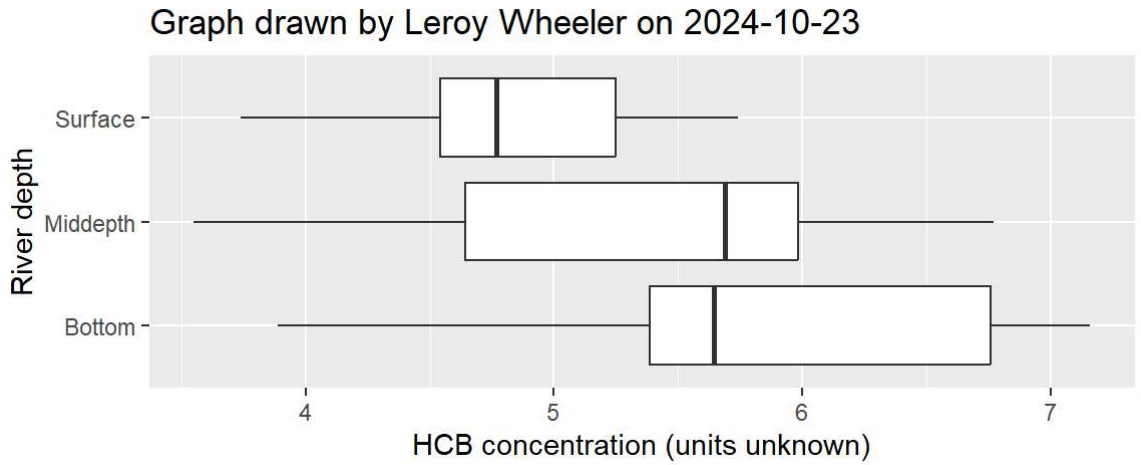
```
Rows: 30
Columns: 3
$ depth  <chr> "Surface", "Surface", "Surface", "Surface", "Surface", "Surface…
$ aldrin <dbl> 3.08, 3.58, 3.81, 4.31, 4.35, 4.40, 3.67, 5.17, 5.17, 4.35, 5.1…
$ hcb    <dbl> 3.74, 4.61, 4.00, 4.67, 4.87, 5.12, 4.52, 5.29, 5.74, 5.48, 6.0…
```

Question 1: Compare the average hcb concentrations between the surface, middepth and bottom sampling locations using analysis of variance. Be sure to include appropriate descriptive statistics and

graphs. Comment on the assumptions needed for this test, but do not conduct any alternative analyses. If there is a statistically significant difference among the three means, use the Tukey post-hoc comparison to identify where the differences lie.

Draw boxplots for hcb concentrations

```
river |>
  ggplot(aes(depth, hcb)) +
    geom_boxplot() +
    xlab("River depth") +
    ylab("HCB concentration (units unknown)") +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-23") +
    coord_flip()
```



HCB concentrations increase in the deeper parts of the river. The variation is greater in the middepth and bottom of the river and looks skewed to the left (middepth) and to the right (bottom).

Descriptive statistics of hcb measurements

```
river |>
  group_by(depth) |>
  summarize(
    hcb_mn=mean(hcb),
    hcb_sd=sd(hcb),
    n=n())
```

```
# A tibble: 3 × 4
  depth    hcb_mn hcb_sd     n
  <chr>     <dbl>  <dbl> <int>
1 Bottom     5.84  1.01     10
2 Middepth   5.33  1.11     10
3 Surface    4.80  0.631    10
```

The mid level and bottom samples have higher average concentrations and higher amounts of variability, with their standard deviation values about double that of the surface samples. This difference in variation should not be a big problem in moving forward with looking for differences among the three groups.

Analysis of variance table for hcb

```
m1 <- aov(hcb ~ depth, data=river)
tidy(m1)
```

```
# A tibble: 2 × 6
  term         df sumsq meansq statistic p.value
  <chr>     <dbl> <dbl>  <dbl>     <dbl>   <dbl>
1 depth         2  5.36  2.68       3.03  0.0649
2 Residuals    27 23.8   0.883     NA    NA
```

The F-ratio is not large enough to suggest that there is a significant difference between any of the different group samples. Additionally the p-value is greater than an alpha = 0.05. Therefore there is not enough evidence to reject the null hypothesis so we will conclude that hcb levels in the river is similar at all levels of depth.

Since there is no evidence for a significant difference in hcb in the three samples we do not conduct Pairwise tests, however I am curious to see what the TukeyHSD test tells us.

```
TukeyHSD(m1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = hcb ~ depth, data = river)

$depth
                    diff       lwr         upr     p adj
Middepth-Bottom  -0.509 -1.551101 0.533101393 0.4570575
Surface-Bottom   -1.035 -2.077101 0.007101393 0.0518462
Surface-Middepth -0.526 -1.568101 0.516101393 0.4341560
```

Tukey results are unnecessary because we failed to reject the null hypothesis. Interestingly, TuckeyHSD also confirms there are no two group means which are statistically different.

Question 2: You want to run a sample size calculation for a replication of this experiment using hcb as the outcome measure. Assume that the sample means for hcb are similar at surface and middepth, but higher at the bottom (4.8 for the surface, 4.8 for middepth, and 5.2 for the bottom). What sample size would you need to achieve 90% power at an alpha level of 0.05.Sample size calculation scenario?

Sample size calculation for hcb, R code

```r
v <- var(c(4.8, 4.8, 5.2))
power.anova.test(
    groups=3,
    n=NULL,
    between.var=v,
    within.var=1.10,
    sig.level=0.05,
    power=0.90)
```

```
     Balanced one-way analysis of variance power calculation

         groups = 3
              n = 131.4979
    between.var = 0.05333333
     within.var = 1.1
      sig.level = 0.05
          power = 0.9

NOTE: n is number in each group
```

Sample size calculation, interpretation

A sample size of 132 measurements per depth level would provide 90% power for detecting a difference between means of 4.8, 4.8, and 5.2 in hcb concentration. This assumes that the variation within groups is similar to the previous study (1.10) and an alpha level of 0.05.