Analysis of Wolf River pollution

AUTHOR
Michael Dang

PUBLISHED
October 29, 2024

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

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…
```

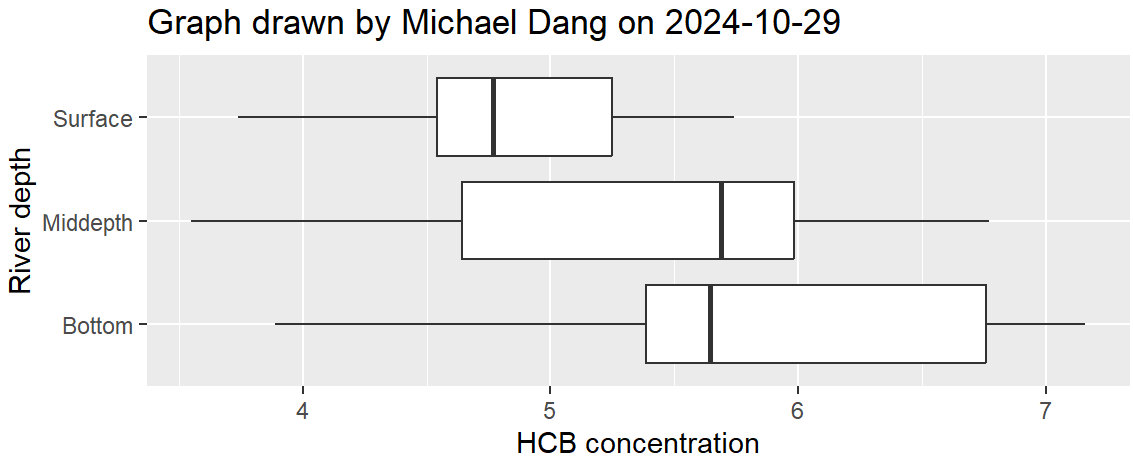
Question 1

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

```
river |>
  ggplot(aes(depth, hcb)) +
```

```
    geom_boxplot() +
    xlab("River depth") +
    ylab("HCB concentration") +
    ggtitle("Graph drawn by Michael Dang on 2024-10-29") +
    coord_flip()
```



The deeper you sample, the higher the concentration of HCB. The variation also increases as you go deeper. There are some minor deviations from normality, but nothing too serious.

Descriptive statistics

```
river |>
  group_by(depth) |>
  summarize(
    aldrin_mn=mean(hcb),
    aldrin_sd=sd(hcb),
    n=n())
```

```
# A tibble: 3 × 4
  depth    aldrin_mn aldrin_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 bottom samples have the highest average concentration and the middle have highest amount of variability.

Analysis of variance table

```
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
```

Since the p-value (0.0649) is greater than 0.05, we fail to reject the null hypothesis at the 5% significance level. This means there is not enough evidence to conclude that the mean HCB concentrations differ significantly among the three depths (surface, middepth, and bottom).

Since they are not statistically significant different among three means, hence Tukey post-hoc comparision was not used.

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, R code

```r
v <- var(c(4.8, 4.8, 5.2))
power.anova.test(
    groups=3,
    n=NULL,
    between.var=v,
    within.var=0.88,
    sig.level=0.05,
    power=0.90)
```

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

         groups = 3
              n = 105.4006
    between.var = 0.05333333
     within.var = 0.88
      sig.level = 0.05
          power = 0.9

NOTE: n is number in each group
```