Homework 1

Igor Kuivjogi Fernandes

2023-01-26

1. The slope of a linear regression line is 2.5 and the intercept is -1. What is the equation of the line?

$$y = -1 + 2.5x$$

   What would be the predicted value of Y when x = 4.3?
   For $x = 4$, $\hat{y} = -1 + 2.5(4.3) = 9.75$

2. A researcher is interested in determining if there is a difference in yield among three different cultivars. Perform the appropriate test to determine if there is a significant difference in yield among them. Use the following dataset:

```
data <- data.frame(
  Crop_Variety = c("Wheat1", "Wheat1", "Wheat2", "Wheat2", "Wheat3", "Wheat3"),
  Yield = c(3200, 2900, 2600, 2400, 2200, 2000)
)
data
```

```
##   Crop_Variety Yield
## 1       Wheat1  3200
## 2       Wheat1  2900
## 3       Wheat2  2600
## 4       Wheat2  2400
## 5       Wheat3  2200
## 6       Wheat3  2000
```

We can perform an one-way ANOVA test to check whether the difference in cultivars yield is significant. These are the null and alternative hypothesis:

$$H_0 : \text{the means are equal}$$
$$H_1 : \text{at least one mean differs}$$

```
result <- aov(Yield ~ Crop_Variety, data = data)
print(summary(result))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Crop_Variety  2 910000  455000   16.06  0.025 *
## Residuals     3  85000   28333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For a significance level of $\alpha = 0.05$, we reject the null hypothesis that the mean yield of the three cultivars are equal, because p-value $= 0.025 < 0.05$, which means that at least one mean yield differs.

3. A researcher wants to determine if there is a significant difference in growth rate before and after a treatment. He collected data on growth rate for 6 individuals before and after the treatment. Perform the appropriate test to determine if there is a significant difference in growth rate before and after the treatment. Use the following dataset:

```
data <- data.frame(
  Plant = c(1, 2, 3, 4, 5, 6),
  Treatment = c("A", "A", "A", "B", "B", "B"),
  Before = c(10, 8, 12, 11, 9, 15),
  After = c(12, 9, 14, 15, 11, 17)
)
data
```

```
##   Plant Treatment Before After
## 1     1         A     10    12
## 2     2         A      8     9
## 3     3         A     12    14
## 4     4         B     11    15
## 5     5         B      9    11
## 6     6         B     15    17
```

We can use a paired two-sided t-test. We have to used a paired version because the sample is the same, and we have matches before and after a treatment. The two-sided over another alternative is preferred because we are only interested to check whether there's a difference between the two means (not whether one is greater or lesser then the other). These are the null and alternative hypothesis:

$$H_0 : \text{the difference between the means is } 0$$
$$H_1 : \text{the difference between the means is not equal to } 0$$

```
result <- t.test(data$Before, data$After, paired = T, alternative = "two.sided")
print(result)
```

```
##
##  Paired t-test
##
## data:  data$Before and data$After
## t = -5.398, df = 5, p-value = 0.002947
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.198463 -1.134870
## sample estimates:
## mean difference
##       -2.166667
```

As the p-value $= 0.002947 < 0.05$, using a significance level of $\alpha = 0.05$, we reject the null hypothesis that the difference between the means is 0, i.e., the growth rate changed after applying the treatment.

4. Standardize (make it a standard normal distribution) the following data and include the new values below.

```r
x <- c(9.58,  6.64,  6.75,  5.26, 19.07, 10.32, 17.62,  6.32, 14.75, 18.36,
       11.97, 18.87,  6.01, 10.51, 12.41, 17.09, 18.12, 11.84, 7.63, 14.26)
z <- (x - mean(x)) / sd(x)
print(z)
```

```
##  [1] -0.53513178 -1.14281330 -1.12007691 -1.42805156  1.42639800 -0.38217793
##  [7]  1.12669113 -1.20895550  0.53347823  1.27964498 -0.04113218  1.38505912
## [13] -1.27303076 -0.34290600  0.04981335  1.01714311  1.23003833 -0.06800246
## [19] -0.93818585  0.43219797
```

5. Explain what each term is and give an example:

a) Pseudo-replication
Is when you use the same group of people or things repeated times to collect information, rather then collecting data from different groups each time. It inflates the sample size, leading to incorrect inferential statistical tests.
Example: a researcher randomly selects 4 plots of land, applies a new fertilizer to all the plots, and measures the yield of each plot. This is a pseudo-replication, because all the plots used the same fertilizer (treatment), which means that the measurements are not independent from each other.

b) True replication
Is the smallest experimental unit to which a treatment is independently applied.
Example: a researcher randomly selects 4 plots of land, applies a different fertilizer (treatment) to each plot, and measures the yield of each plot. This is the case of a true replication because the researcher accounted for independent treatments for each group of samples.

c) Experimental unit
Is the smallest entity that can be randomly assigned to a different treatment condition.
Example: if a researcher randomly selects 4 plots of land, applies a different fertilizer (treatment) to each plot, and measures the yield of each plot, then each plot would be a experimental unit, because in this case is the smallest entity in which you would apply a different treatment.

d) Observational unit
Is the unit observed (measured) by the researcher in the study.
Example: if a researcher is studying the impact of a new fertilizer on the crop yield, the observational unit could be a group of plants.

e) Dependent variable
It is the variable of interest being studied in the experiment, that might be affected by other variables (the independent variables).
Example: a researcher is studying how a fertilizer (a independent variable X_1) and the quantity of nitrogen in the soil (another independent variable X_2) affects the crop yield (the dependent variable Y).

6. A researcher is interested in determining if there is a significant difference in yield between an experimental fertilizer and a control fertilizer. He collected data on yield for 10 plots treated with the experimental fertilizer and 10 plots treated with the control fertilizer. The mean yield for the experimental fertilizer is 8 with a standard deviation of 2 and the mean yield for the control fertilizer is 7 with a standard deviation also 2. Determine the 95% confidence interval for the difference in means and interpret the results.

We can use a t-statistic to build a confidence interval for the difference in means.
Denote A as the experimental fertilizer and B as the control fertilizer:

```r
xbar_A <- 8
xbar_B <- 7
s_A <- s_B <- 2
n_A <- n_B <- 10

# pooled estimate of the standard deviation is used because both s_B and s_B are equal
sP <- sqrt((((n_A - 1) * s_A ^ 2) + ((n_B - 1) * s_B ^ 2)) / (n_A + n_B - 2))

# standard error between difference of means
se_diff <- sP * sqrt((1 / n_A) + (1 / n_B))

# use the t-statistic with p = 0.05 / 2
ic <- c()
ic[1] <- (xbar_A - xbar_B) + (qt(p = 0.025, df = n_A + n_B - 2) * se_diff)
ic[2] <- (xbar_A - xbar_B) - (qt(p = 0.025, df = n_A + n_B - 2) * se_diff)
ic
```

```
## [1] -0.8791218  2.8791218
```

7. Answer the following:

a) What are the dimensions of D and B in the following equation? $\underset{?\times?}{D} = \underset{3\times4}{A} \times \underset{?\times?}{B} \times \underset{10\times10}{C}$

```r
A <- matrix(nrow = 3, ncol = 4)
B <- matrix(nrow = 4, ncol = 10)
C <- matrix(nrow = 10, ncol = 10)
D <- A %*% B %*% C
print(dim(D))
```

```
## [1]  3 10
```

b) Which row and column would the item $b_{1,3} = 24$ go on matrix $\mathbf{B}$?
c) What are the dimensions of A'A and AA'?

8. Define the following matrices:

a) diagonal matrix;
b) identity matrix (What R function do we use to create it?)
c) upper or lower triangular matrix;
d) square matrix

9. Given the following matrices:

```
A = [1 2 3]        B = [ 7    8]         C = [4   7  2]     k = 10
    [4 5 6]            [ 9   10]             [5  8  4]
                       [11 12]               [6  9  5]
```

Calculate: a) A x B b) Inverse of C c) A B (Kronecker product of A and B) d) Transpose of B e) kA

10. Represent the following system of equations with matrix notation. Also show how to create the resulting matrix with R.

$$3x + 2y = 12 \; 4x - y = 2$$