HW2

Math 189

Friday of Week 2, 04/08/2022

1. (a) Run the following R commands:

```
# install the packages if needed by using
# install.packages("...")
library(tidyr)
library(readr)
library(tidytuesdayR)
urlRemote <- 'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/'
pathGithub <- 'data/2020/2020-07-28/'
fileName <- 'penguins.csv'
penguins <- paste0(urlRemote, pathGithub, fileName) %>% read.csv(header = TRUE)
dfr <- drop_na(as.data.frame(penguins))
head(dfr)
```

   (b) Use R commands to report the number of rows and number of columns in `dfr`.

2. The data in Question 1 comes from Dr. Kristen Gorman by way of the palmerpenguins R package by Dr. Kristen Gorman, Dr. Allison Horst, and Dr. Alison Hill. The details of all of the variables is as follows:

variable	class	description
species	integer	Penguin species (Adelie, Gentoo, Chinstrap)
island	integer	Island where recorded (Biscoe, Dream, Torgersen)
bill_length_mm	double	Bill length in millimeters (also known as culmen length)
bill_depth_mm	double	Bill depth in millimeters (also known as culmen depth)
flipper_length_mm	integer	Flipper length in mm
body_mass_g	integer	Body mass in grams
sex	integer	sex of the animal
year	integer	year recorded

   Let `X <- dfr[,3:6]`. Find the mean vector, covariance matrix and correlation matrix of **X**. What are the meanings of the elements in variance-covariance matrix and correlation matrix?

3. Let A be the correlation matrix you obtained in Question 2. Use R to calculate the following results.

   (a) Find $2A$.

   (b) Choose an integer as your own random seed to replace the number 1 in the following code chuck. Define

```
set.seed(1) # replace 1 by your own choice
B <- matrix(rnorm(16),nrow=4)
```

   Find $C = B^T B$. Is C symmetric?

   (c) Choose two nonzero real numbers a and b, and find $aA + bB$.

(d) Find the eigenvalues, eigenvectors, and the square root matrix of A.

4. Let X_1, X_2, X_3, X_4 denote the variables `bill length`, `bill depth`, `flipper length` and `body mass`, respectively. Suppose that we introduce two new variables: $Y_1 = 3X_1 + 2X_2$ and $Y_2 = X_2 + X_3 + X_4$. Let $\mathbf{Y}$ be the dataset recording variables Y_1 and Y_2 of the same individuals as in $\mathbf{X}$. Find the mean vector and covariance matrix of $\mathbf{Y}$.

5. Let X be a general population, and let $X_1, \ldots, X_n$ be a simple random sample from X. Define the *sample loss function* as

$$L(a) = \frac{1}{n} \sum_{i=1}^{n} (X_i - a)^2.$$

   (a) Find $\hat{a}$ that minimizes $L(a)$. That is, find

$$\hat{a} = \arg\min L(a).$$

   (b) Plug $\hat{a}$ into $L(a)$, and what is your value of $L(\hat{a})$?

   (c) Discuss that whether $\hat{a}$ is an unbiased estimator of the population mean?

   (d) Discuss that whether $L(\hat{a})$ is an unbiased estimator of the population variance?