Introductions

- Name
- Year
- Previous stats courses
- Make sure to sign in on the google form (linked here)

Goals

- Learn relevant R skills for the week
- See similar examples to the homework
- Learn something about the world

Linear algebra and matrices in R

Let

$$\mathbf{a} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 2 \\ -2 & 3 & -2 \end{bmatrix} \mathbf{f} = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

1. How do you multiply matrices? What about scalar multiplication? Find $4\mathbf{d}$. Why is $\mathbf{ab}$ not defined? What should we change to be able to multiply them?

2. How do you find the determinant of a matrix? What is the determinant of $\mathbf{d}$?

3. How do you find the inverse of a matrix? When does the inverse of a matrix exist? What is the inverse of $\mathbf{f}$?

Some useful commands:

- `rep(k, n)` gives a vector `n` long of `k`
- `cbind(c1, c2, ...)` makes a matrix with columns `c1`, `c2`, ...
- `rbind(r1, r2, ...)` makes a matrix with rows `r1`, `r2`, ...
- `matrix(k, nrow, ncol)` makes a matrix with all entries `k`
- `diag(x)` gets the diagonal of a matrix
- `%*%` does matrix multiplication
- `t(x)` gives the transpose of `x`
- `det(x)` gives the determinant of `x`
- `solve(x)` gives the matrix inverse of `x`

4. What should the following code print?

```r
a = cbind(c(1/3, 1/3, 1/3))
b = cbind(c(2, 3, 4))
mu = cbind(rep(4,3))
d = rbind(c(1, 2, 0), c(3, -1, 2), c(-2, 3, -2))
f = matrix(1, nrow = 3, ncol = 3)
diag(f) <- 3
```

```
d * 4 # Scalar multiplication

a %*% b
# TODO: Correct code here

a %*% d
# TODO: Correct code here

# Determinant
det(d)

solve(d)
# TODO: Find the inverse of another matrix that is invertible

# TODO: Show that the inverse times itself is the identity
```

5. What does the following code represent if $\mathbf{b}$ is a data vector, $\boldsymbol{\mu}$ is a mean vector, and $\mathbf{f}$ is a covariance matrix?

```
t(a) %*% (b-mu) %*% (t(a) %*% f %*% a)^(-1/2)
```

```
##            [,1]
## [1,] -0.7745967
```

Distribution of the sample mean with and without covariance

1. If we have $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, 1)$, what is the distribution of $\bar{X}$? If $n = 50$, what is its variance?
2. If we have $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, 1)$, but each is correlated with correlation ρ with its neighbors (and 0 otherwise), what is the distribution of $\bar{X}$? If $n = 50$ and $\rho = 0.5$, what is its variance? What about if $\rho = 0.2$?
3. If you want your sample mean to have a variance equal to the variance of the sample mean of n uncorrelated observations when you have a correlation of ρ, what n' do you need for your correlated samples? Find the exact answer and an approximation for large n.

```
set.seed(139)

# Import MASS for mvrnorm
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

```
# Number of samples
n <- 50
```

4. Write a function called `moving_mean` that returns the mean of the first n elements of x.

```r
# Gets the mean of the vector x through index n
moving_mean <- function(x, n) {
  # TODO
}

# Show convergence with no correlation
x <- rnorm(n, 0, 1)
means = vector(length = n)
for (i in 1:n) {
  means[i] <- moving_mean(x, i)
}
plot(1:n, means, ylim = c(-3,3))

# Correlation (Why could this not be above 0.5? E.g. what would be wrong if it was 1?)
p = 0.5
```

5. Create a matrix filled with 0s except for a diagonal of 1 and 1-off diagonals of ρ.

```r
# Creates covariance matrix with correlation between adjacent samples
Sigma = # TODO

# Show convergence with correlation
x <- mvrnorm(n = 1, rep(0, n), Sigma)
means = vector(length = n)
for (i in 1:n) {
  means[i] <- moving_mean(x, i)
}
plot(1:n, means, ylim = c(-3,3))
```

6. Fill in the outputs table so the first column is uncorrelated means and the second is correlated means.

```r
# Simulate many times to get the variance of the sample mean
nsim = 10000
outputs <- matrix(nrow = nsim, ncol = 2)
for (i in 1:nsim) {
  outputs[i,1] # TODO
  outputs[i,2] # TODO
}

# No correlation
hist(outputs[,1], xlim = c(-1,1))
var(outputs[,1])

# Correlation
hist(outputs[,2], xlim = c(-1,1))
var(outputs[,2])
```

Data exploration for country demographics

This section will deal with a data set of country-level statistics from this source. We'll go over the following things:

- Summary statistics
- Overlaid histogram
- Box plot
- Scatter plot
- Two-way table

```
# Read in the data
countries <- read.csv("data/countries.csv", check.names = F)
```

1. Calculate the following summary statistics for the `Population` variable: sample mean, sample standard deviation, min, median, max, and the 1st and 3rd quartiles. Also calculate the proportion of countries with less than 10 million people.

```
# TODO: Summary statistics

# TODO: Proportion of countries with less than 10 million people
```

2. Split the countries into two groups: those with less than 10 million people and those with more than 10 million people. Use summary statistics and graphics to explore whether there is evidence of a difference in land area between the two groups. Comment on the results without performing a formal hypothesis test. (Because the data are very right skewed, it will help to take the log of both the population and the area; just make sure to set your 10 million threshold before taking the log!)

```
# TODO: Split countries and get summary statistics

# TODO: Add a column called under10 with an indicator of whether the country has under 10 million peopl

library(ggplot2)
ggplot(countries, aes(x=`Area (sq. mi.)`, fill=under10)) +
  geom_histogram(alpha=0.4, position="identity") +
  xlab("Area") +
  theme_bw()

# TODO: Add a column to countries called logArea and plot an overlaid histogram

# TODO: Add a box plot showing the same data

ggplot(countries, aes(x=log(Population), y=logArea)) +
  geom_point() +
  xlab("Log Population") +
  ylab("Log Area") +
  theme_bw()
```

3. Find the number of countries with under and over 10 million people by region. Does there seem to be a difference between regions?

```
# TODO: Make a table of regions vs over/under 10 million people
```