# Week 1 Quiz Material

When copy and pasting from a code block, or from your local `R` session, be sure to include all digits for any numeric answer. Also, do no modify the default digits option in the code blocks or your local `R` session.

# Practice

## Exercise 1

```
# starter
x = 1:100
```

Calculate

$$\sum_{i=1}^{n} \ln(x_i).$$

That is, sum the log of each element of `x`.

```
# solution
sum(log(x))
```

```
## [1] 363.7394
```

## Exercise 2

```
# starter
set.seed(42)
a_vector = rpois(250, lambda = 6)
```

How many of the elements of `a_vector` are greater than or equal to 5? (Notice were using two functions `set.seed()`, and `rpois()` to create this vector. We will discuss these at length when we begin to discuss probability.) Be sure to run the two lines in order, otherwise your vector will not contain the expected elements.

```
# solution
sum(a_vector >= 5)
```

```
## [1] 181
```

Because the elements in the vector are discrete, `sum(a_vector >= 5)` and `sum(a_vector > 4)` will both give the correct answer.

## Exercise 3

```
# starter
x = 1:100
```

Create a new vector `y`, which adds 5 to the elements stored in odd indices of `x` and subtracts 10 from the elements stored even indices of `x`. Calculate the standard deviation of this new vector.

```
# solution
z = c(5, -10)
y = x + z
sd(y)
```

```
## [1] 29.8481
```

Possible hint: If you are using a for loop, instead try creating a second vector and then use vectorized operations.

# Exercise 4

```
# starter
quiz_list = list(
  x = c(1, 2),
  y = "Hello Quiz Taker",
  z = "z"
)
```

Which of the following would return the third element of the list `quiz_list`?

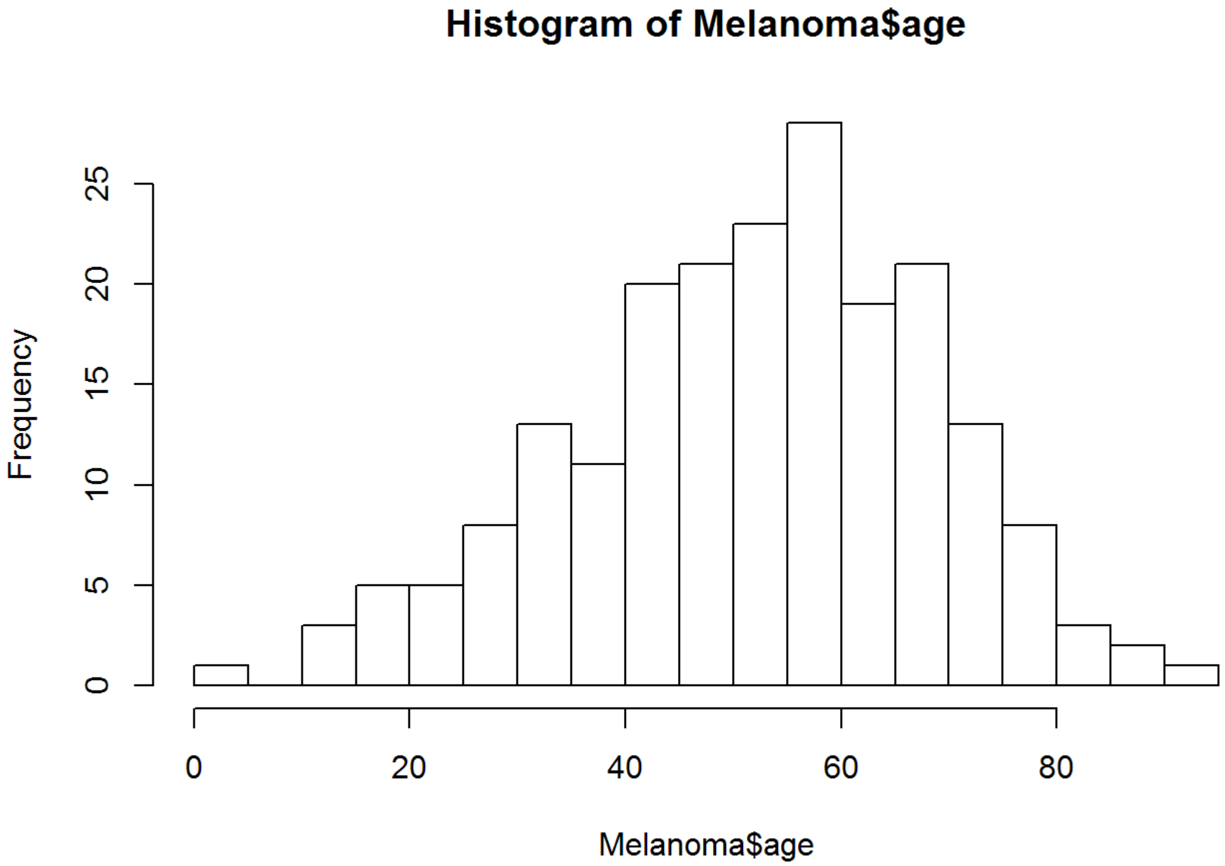
```
quiz_list[3]
quiz_list[[3]] # correct
quiz_list["3"]
quiz_list$z # correct
quiz_list$3
```

# Exercise 5

```
# starter
library(MASS)
```

Create a histogram of `age` in the `Melanoma` dataset from the `MASS` package. How would you describe this data?

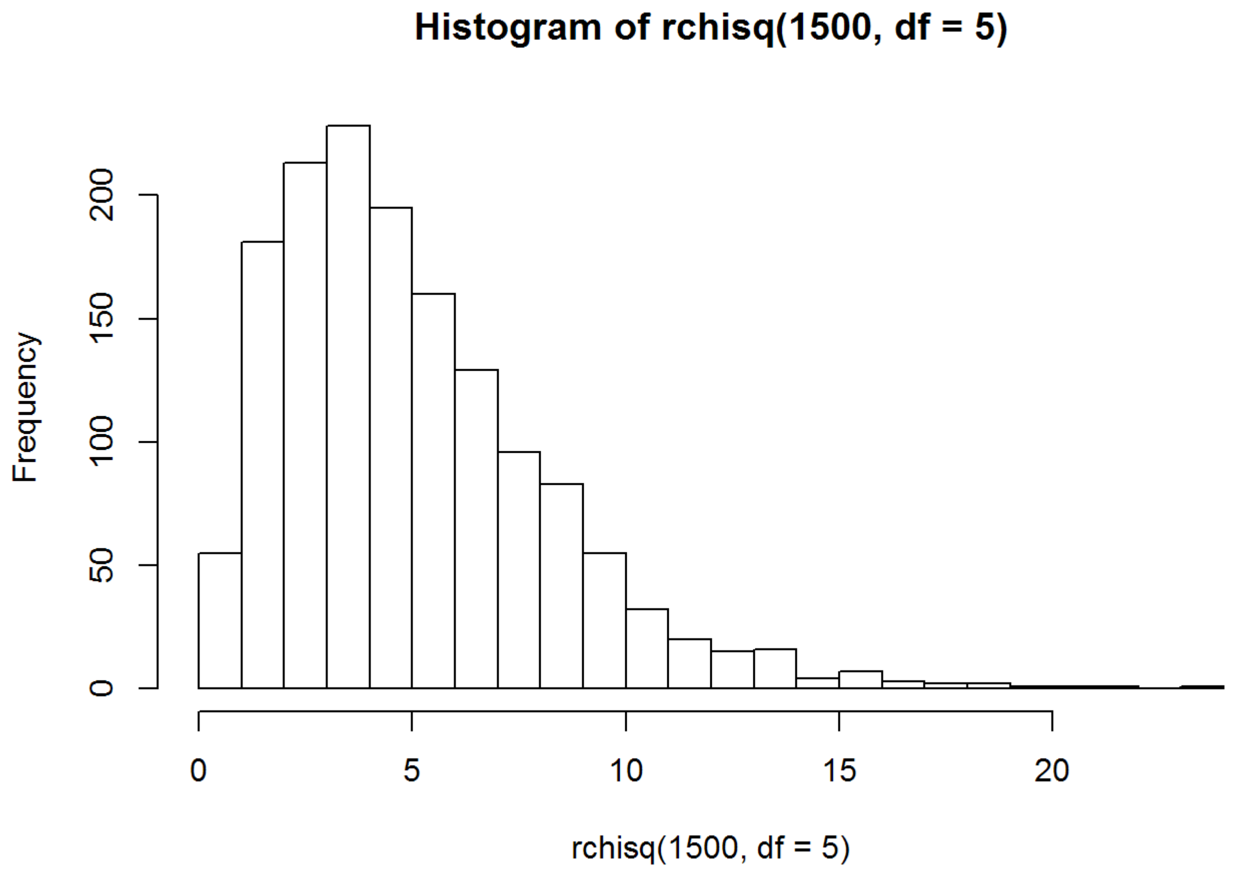
```
# solution
hist(Melanoma$age, breaks = 20)
```



- **Large left skew**
- **Slight left skew**
- Symmetric
- Slight right skew
- Large right skew

Possible hint: For an example of a right skew, run:

```
hist(rchisq(1500, df = 5), breaks = 20)
```



# Graded

## Exercise 1

How many individuals in the `Melanoma` dataset from the `MASS` package died from a melanoma?

```
# solution
library(MASS)
table(Melanoma$status)
```

```
##
##   1   2   3
##  57 134  14
```

Answer: 57

## Exercise 2

What is the average age of individuals in the `Melanoma` dataset from the `MASS` package who are alive?

```
# solution
library(MASS)
mean(subset(Melanoma, status == 2)$age)
```

```
## [1] 50.00746
```

# Exercise 3

Which animal in the `mammals` dataset from the `MASS` package has the largest brain weight relative to its body weight? (That is, the largest brain weight to body weight ratio.)

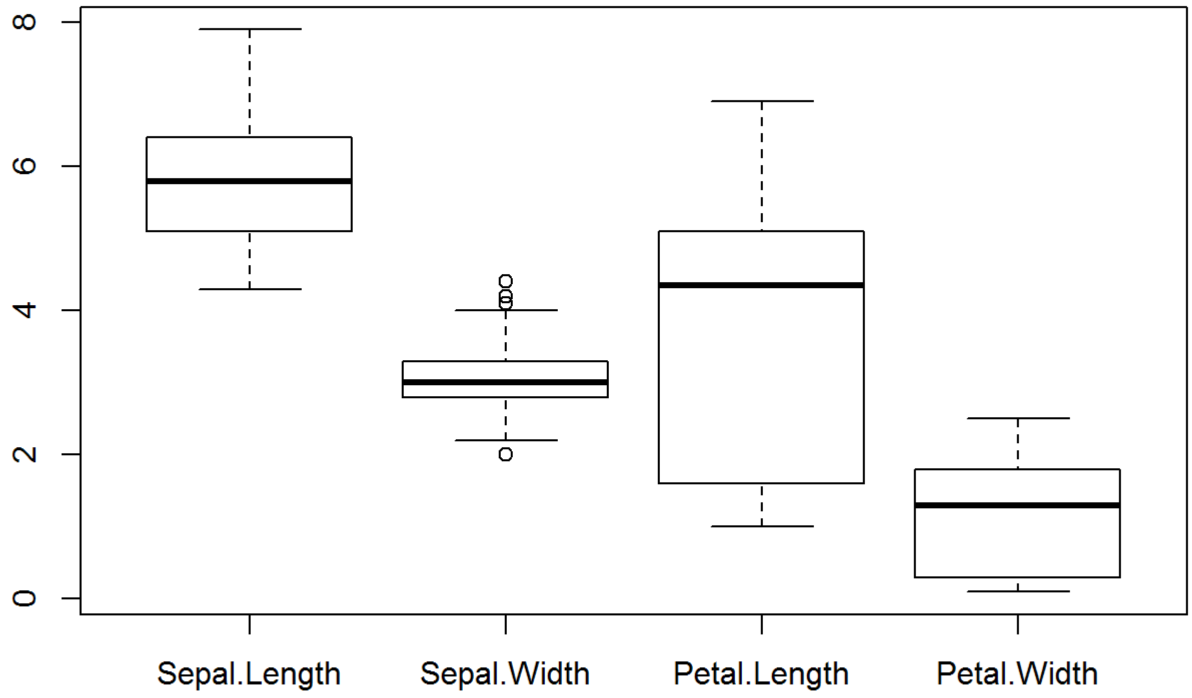
```
# solution
library(MASS)
rownames(mammals[which.max(mammals$brain / mammals$body), ])
```

```
## [1] "Ground squirrel"
```

# Exercise 4

Create side-by-side boxplots for each of the numeric variables in the `iris` dataset. To do so, simply supply the usual function with a dataframe of only the numeric variales of the dataset. Based on this plot, which variable is the most variable? Calculate the standard deviation of this variable.

```
# solution
boxplot(iris[, -5])
```



```
sd(iris$Petal.Length)
```

```
## [1] 1.765298
```

# Exercise 5

```
# preamble
set.seed(42)
z = list(
  round(rnorm(n = 25, 0, 5), 2),
  c(1, 1, 2, 3, 5, 8),
  sample(30)
)
```

The above code block has access to a list stored in the variable `z` .

Calculate the sum of:

- The minimum first element of `z`
- The maximum of the second element of `z`
- The mean of the third element of `z`

```
# solution
min(z[[1]]) + max(z[[2]]) + mean(z[[3]])
```

```
## [1] 10.22
```

# Exercise 6

Where were the measurements taken in the `airquality` dataset?

```
# solution
?airquality
```

- Chicago, IL
- Los Angeles, CA
- **New York, NY**
- Champaign, IL
- Paris, France

# Exercise 7

Using the `airquality` dataset, what is the average wind speed in May ?

```
# solution
mean(subset(airquality, Month == 5)$Wind)
```

```
## [1] 11.62258
```

# Exercise 8

Using the `airquality` dataset, what is the average ozone measurement? Hint: read the documentation of any function that returns an unexpected result. You will likely find a solution to the issue.

```
# solution
mean(airquality$Ozone, na.rm = TRUE)
```

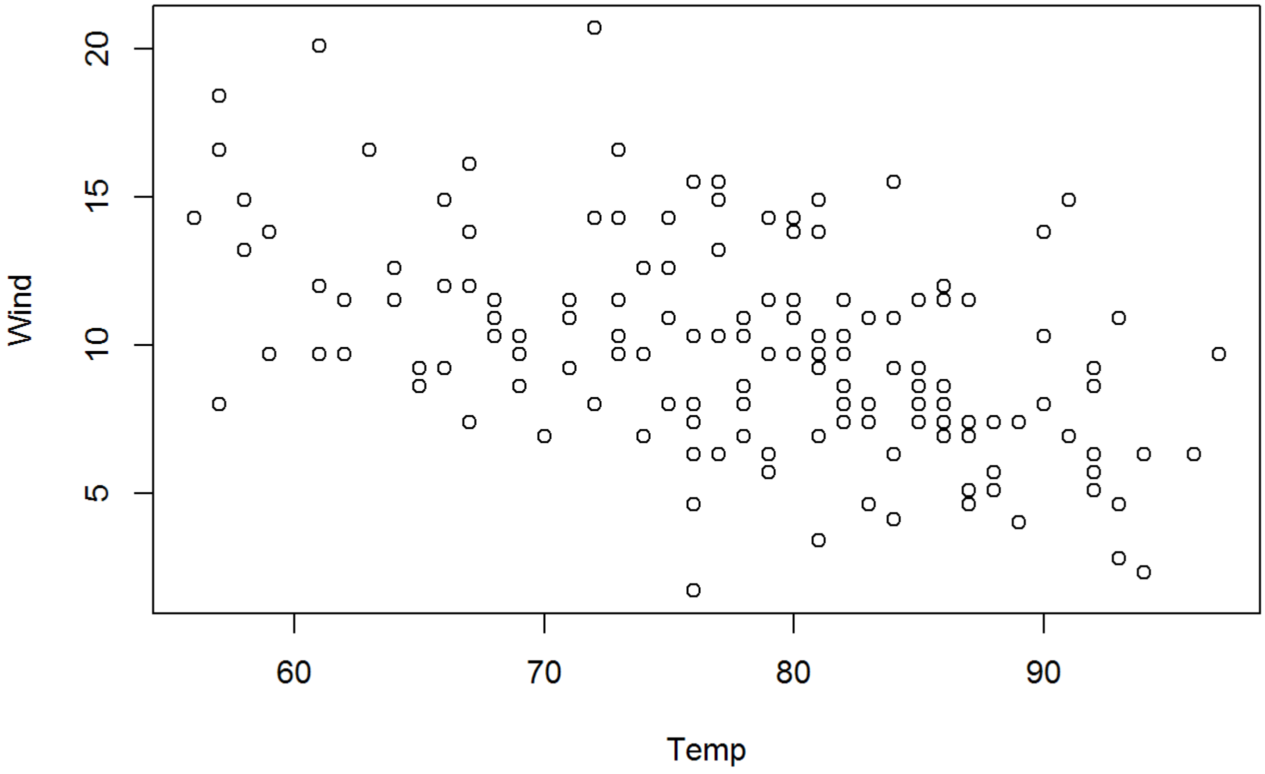
```
## [1] 42.12931
```

# Exercise 9

Using the `airquality` dataset, create a scatter plot to compare windspeed and temperature. Based on this plot, you believe that:

- Wind speed and temperature have no relationship.
- As temperature increases, wind speed increases.
- **As temperature increases, wind speed decreases**

```
# solution
plot(Wind ~ Temp, data = airquality)
```



# Exercise 10

```
# starter
set.seed(1337)
x = rnorm(10000)
```

What proportion of the elements of x are larger than 2 in magnitude? Be sure to run the two lines in order, otherwise your vector will not contain the expected elements.

```
# solution
mean(abs(x) > 2)
```

```
## [1] 0.0444
```

# Exercise 11

```
# starter
set.seed(42)
x = rnorm(100, mean = 0, sd = 10)
mean(f(input = x)) - f()
```

Write a function called `f` that has a single argument `input` with a default value of `42` which is assumed to be a vector of numeric valves. The function should output a vector that is `input` but with any negative values replaced with `0`.

Hint: The `ifelse()` function could be useful here. Note that all three arguments to `ifelse()` are vectors.

Run your function followed by the three lines given. Submit the output as your answer.

```
# solution
f = function(input = 42) {
  ifelse(input > 0, input, 0)
}

set.seed(42)
x = rnorm(100, mean = 0, sd = 10)
mean(f(input = x)) - f()
```

```
## [1] -37.70725
```

# Exercise 12

```
# starter
set.seed(42)
y  = 5 * x0 + x1 + rnorm(n = 30, mean = 0 , sd = 1)
```

Create three vectors `x0`, `x1`, and `y`. Each should have a length of `30` and store the following:

`x0` : Each element should be the value 1 `x1` : The first 30 square numbers, starting from 1 (so 1, 4, 9, etc.) `y` : The result of running the given code, after creating the other two vectors

Report the mean of the values stored in `y` .

```
# solution
x0 = rep(1, 30)
x1 = (1:30) ^ 2
set.seed(42)
y  = 5 * x0 + x1 + rnorm(n = 30, mean = 0 , sd = 1)
mean(y)
```

```
## [1] 320.2353
```

# Exercise 13

**(Continued from Exercise 12)** Create a matrix `X` with columns `x0` and `x1` . Report the sum of the elements in rows 17 and 19.

```
# solution
X = cbind(x0, x1)
sum(X[c(17, 19), ])
```

```
## [1] 652
```

# Exercise 14

**(Continued from Exercises 12 and 13)** Use matrix operations to create a new matrix `beta_hat` defined as follows. Report the sum of the values stored in this matrix.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

```
# solution
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
sum(beta_hat)
```

```
## [1] 6.427899
```

# Exercise 15

**(Continued from Exercises 12, 13, and 14)** Create a new variable `y_hat` which stores the result of the matrix operation,

$$\hat{y} = X\hat{\beta}.$$

The result will be a $30 \times 1$ matrix. Perform and report the result of the following operation,

$$\sum_{i=1}^{30}(y_i - \hat{y}_i)^2.$$

```
# solution
y_hat = X %*% beta_hat
sum((y - y_hat) ^ 2)
```

```
## [1] 42.67698
```

```
crossprod(y - y_hat, y - y_hat)
```

```
##          [,1]
## [1,] 42.67698
```