# Unit 2 Practice Questions with Solutions

1. Consider the following 12 observations on students' heights (in inches):

   $64, 65, 62, 70, 68, 66, 65, 66, 69, 71, 68, 72$

   What type of data is this and what would be an appropriate plot to display the data, respectively?

   ○ qualitative data, histogram

   ○ count data, pie chart

   √ **quantitative data, boxplot**

   ○ categorical data, bar graph

   ○ time series data, time series plot

2. Refer to Question 3 of the Unit 1 Practice Questions.

   (a) Write the R code that makes a contingency table showing how many animals of each type in `animalDF` weigh below 5 pounds or weigh at least 5 pounds.

   **Solution:**

```
weightCat = character(length(animalDF$weight))
for(i in 1:length(animalDF$weight)){
   if(animalDF$weight[i] < 5){
      weightCat[i] = "light"
   } else{
      weightCat[i] = "heavy"
   }
}
table(weightCat,animalDF$animal)
```

   Alternatively:

```
table(animalDF$weight < 5, animalDF$animal)
```

   The reason this works is because `animalDF$weight < 5` creates a vector of TRUEs/FALSEs, which can be considered categorical data.

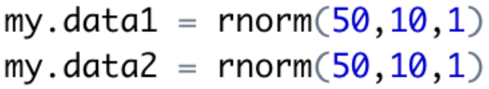
   (b) Name three graphical summaries and one numerical summary that will tell us if the distribution of the weights of the chickens is approximately symmetric or skewed.

   **Solution:**

   Graphical summaries: any 3 of: histogram, boxplot, QQ plot, violin plot, sina plot, kernel density plot

   Numerical summaries: any 1 of: moment coefficient of skewness, five number summary

3. Consider the following randomly generated data:



```
my.data1 = rnorm(50,10,1)
my.data2 = rnorm(50,10,1)
```

(a) Write R code that makes a bar chart showing how many **my.data1** values are less than 10 and how many are greater than 10. Use the *names.arg* argument of the function that makes a bar chart to label the bars.

**Solution:**

The data in **my.data1** is quantitative. In order to make a bar chart, we need to convert it to categorical data. If we check if the values are less than 10 or greater than 10, we end up with TRUEs and FALSEs, which are a type of categorical data.

```
counts = c(length(which(my.data1 < 10)),length(which(my.data1 > 10)))
barplot(counts,names.arg=c("<10",">10"))
```

**Alternate solution:**

```
counts = c(sum(my.data1<10),sum(my.data1>10))
barplot(counts,names.arg=c("<10",">10"))
```

The reason this alternate solution works is because, for example, **my.data1** $< 10$ will return a vector of TRUEs and FALSEs. If you sum that vector, the TRUEs are treated as 1s and the FALSEs are treated as 0s and you get the number of **my.data1** values that are less than 10.

(b) Consider the 50 values in **my.data1** and the 50 values in **my.data2** as 50 ordered pairs (like $(x, y)$ points on a coordinate grid). Write R code that makes a contingency table showing:

- the number of ordered pairs where both values are less than 10
- the number of ordered pairs where both values are greater than 10
- the number of ordered pairs where the **my.data1** value is less than 10 and the **my.data2** value is greater than 10
- the number of ordered pairs where the **my.data1** value is greater than 10 and the **my.data2** value is less than 10

**Solution:**

We need to create a contingency table using the `table()` function, which requires passing in two vectors with categorical data. We need to create these two categorical data vectors from the quantitative data in **my.data1** and **my.data2**.

```
my.data3 = numeric(50)
for(i in 1:50){
   if(my.data1[i] < 10){
      my.data3[i] = 0
   } else{
      my.data3[i] = 1
   }
}
my.data4 = numericd(50)
for(i in 1:50){
   if(my.data2[i] < 10){
      my.data4[i] = 0
   } else{
      my.data4[i] = 1
   }
}
table(my.data3,my.data4)
```

**Alternate solution:**

We need to create a contingency table using the table() function, which requires passing in two vectors with categorical data. A vector of TRUEs and FALSEs is categorical data.

```
table(my.data1 < 10, my.data2 < 10)
```

4. (a) A mosaic plot displays the distribution of what type of data?
   **Solution:**

   Categorical data

   (b) Does a mosaic plot show the distribution of one variable, two variables, or more than two variables?
   **Solution:**

   Two variables

5. If we have a quantitative data vector `x`, explain what the second line of code produces after the histogram of `x` is plotted, without using the word "density" in your response.

```
hist(x)
lines(density(x))
```

**Solution:**

The second line of code plots an estimate of the population distribution from which the sample data was drawn.

6. Consider a quantitative data set. Which two plots would display the distribution of the data?

   ○ bar chart and pie chart

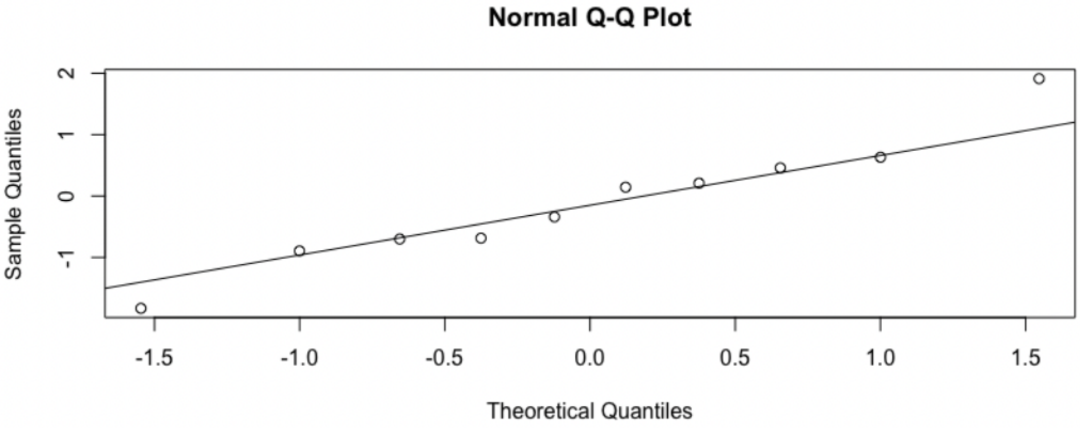
   ○ time series plot and QQ plot

   √ **box plot and histogram**

   ○ box plot and bar chart

   ○ QQ plot and bar chart

7. Consider the following QQ plot produced in R:



   What is an accurate observation about the data based on this plot?

   √ **The data are approximately normally distributed.**

   ○ The data are positively correlated.

   ○ The data likely has an excess kurtosis of approximately 3.

   ○ The data are positively skewed.

   ○ All of the above.

8. We know that the median of (1,6,3), a vector with an odd number of values, is 3 and the median of (2,8,4,6), a vector with an even number of values, is 5. If there are an odd number of values, the median is in position $\frac{n+1}{2}$ of the sorted data vector. If there are an even number of data values, we go to position $\frac{n+1}{2}$ of the sorted data vector and then take the average of the corresponding data values.

   Of course, R has a function called `median()` that can calculate the median of a dataset. Create your own R function called `myMedian()` that takes a vector of data values and returns the median of the dataset, *without* using the `median()` function.

   Hint: we can use modular arithmetic to determine if there are an odd or even number of data values in a vector `x`. If `x%%2` equals 1, then there are an odd number of data values in `x`. If `x%%2` equals 0, then there are an even number of data values in `x`.

```
myMedian = function(x){
  length = length(x)
  medianpos = (length + 1)/2
  sortedx = sort(x)
  if(length %% 2 == 0){
    ans = mean(c(sortedx[medianpos-0.5],sortedx[medianpos+0.5]))
  } else{
    ans = sortedx[medianpos]
  }
  ans
}
```

   Or: `medianpos-0.5` can be replaced by `length/2`; `medianpos+0.5` can be replaced by `length/2+1`.

9. Consider the following three equivalent formulas to calculate correlation:

$$r = \frac{1}{(n-1)s_x s_y}\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})] \tag{1}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{2}$$

$$\text{where } SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2, SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)}} \tag{3}$$

(a) Show how formulas (1) and (2) are equivalent.

**Solution:**

Starting with formula (1):

$$\begin{aligned}
r &= \frac{1}{(n-1)s_x s_y}\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})] \\
&= \frac{1}{(n-1)\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}}\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})] \\
&= \frac{\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}
\end{aligned}$$

(b) Show how formulas (2) and (3) are equivalent.

**Solution:**

In formula (2):

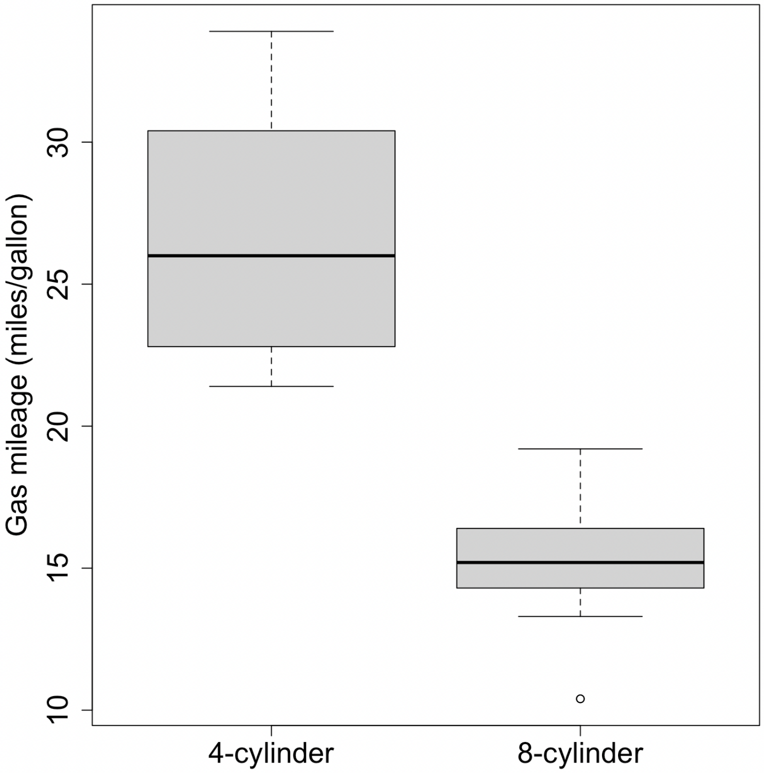
$$\begin{aligned}
SS_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 \\
&= \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}
\end{aligned}$$

It can similarly be shown that $SS_{yy} = \sum_{i=1}^{n} y_i^2 - \dfrac{(\sum_{i=1}^{n} y_i)^2}{n}$.

Also:

$$\begin{aligned}
SS_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum_{i=1}^{n}(x_iy_i - x_i\overline{y} - \overline{x}y_i + \overline{xy}) \\
&= \sum_{i=1}^{n} x_iy_i - \overline{y}n\overline{x} - \overline{x}n\overline{y} + n\overline{xy} \\
&= \sum_{i=1}^{n} x_iy_i - n\overline{xy} \\
&= \sum_{i=1}^{n} x_iy_i - n \cdot \frac{\sum_{i=1}^{n} x_i}{n} \cdot \frac{\sum_{i=1}^{n} y_i}{n} \\
&= \sum_{i=1}^{n} x_iy_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}
\end{aligned}$$

10. Consider the below side-by-side boxplots of gas mileage (in miles per gallon) for 4-cylinder vs. 8-cylinder cars in the mtcars dataset.



Which of the following are true?

(I) The median gas mileage for 4-cylinder cars is higher than the median for 8-cylinder cars.

(II) Gas mileages for 4-cylinder cars are less variable than gas mileages for 8-cylinder cars.

(III) The moment coefficient of skewness for both distributions is likely negative.

○ none

√ **(I) only**

○ (I) and (II) only

○ (I) and (III) only

○ (I), (II) and (III)

11. For this question, choose ALL of the answers that are correct. Consider a data set consisting of information on students, including their height. If one was interested in determining the shape of the distribution of heights, which measures would be useful?

    ○ mean

    ✓ **skewness**

    ○ correlation

    ✓ **kurtosis**

    ○ autocorrelation

12. Describe what is meant by lag-2 autocorrelation.

    **Solution:**

    The lag-2 autocorrelation is used for time series data. It is the correlation between a variable and itself shifted by two time units.

13. (a) What is the benefit of using a violin plot over a boxplot?

    **Solution:** In a violin plot, you can see the shape of the entire distribution of data values, which is not shown by a boxplot.

    (b) When is a contour plot useful to use?

    **Solution:** A contour plot is useful for showing 3 dimensions on a 2-dimensional scatterplot. The contours show the concentration/density of $z$ values at varying values of $x$ and $y$.