Homework #3

Due: Tuesday, October 12 @ 6pm

Problem 1:

Recall the set of measurements from HW1:

52	16	180	1	199	8	3	23	156	63
808	25	5	554	85	1	64	52	7	192

a. Write an R expression to generate a vector of these values and assign it to the variable x
b. Read the help page for the `rep()` function. Using `rep()` and `c()`, write an R expression to generate those values in any order and assign them to `y`. Show `y` after the assignment.
c. Read the help pages for `any()` and `all()`, and briefly describe what they do.
d. What do you expect to get from `all(y==x)`, and why? Check your intuition in R.
e. Suppose you wanted to see if two vectors contained exactly the same values, *regardless of the order they were in*. How might you go about doing that? Write an R expression to test `x` and `y` this way.

Problem 2:

Using the same data as above,

a. Compute:
   - The mean of `x`
   - The median of `x`
   - The sample standard deviation of `x`
   - The mean and sample SD of `2x`
   - The mean and sample SD of `x + 10`
   - The mean and sample SD of `2x + 10`
   - The mean and sample SD of `2(x + 10)`
b. Explore the help pages and online materials to figure out how to plot:
   - A histogram of `x`
   - A boxplot of `x`
c. Consider now the $log_2(x)$ for this data. Write an R expression that computes it assign it to a new variable of your choosing.
d. Plot a histogram of your new $log_2(x)$ variable. How does it compare to the histogram you got for `x`?
e. Suppose we added two additional observations to `x`, both of which were exactly equal to the mean of `x` (as obtained in part (a) above). Write an R expression to have `x` include those two additional values. Then compute the new mean and SD, as you did in HW1. Are they what you expected?

Problem 3:

In this problem, we'll explore how R deals with missing data. Suppose you had a vector `y <- c(1,1,2,3,4,10)`:

a. Write an expression to set the element of `y` that is equal to 10 to `NA`.
b. Imagine the second element of `y` was erroneous. Give two ways one might get rid of it.
c. Show `y`. Without using R, what do you expect the mean of `y` to be?
d. What does `mean(y)` give? How does this compare to your expectation above?
e. Read the help page for `mean()` and give an expression for the mean of the non-missing values of y.
f. Write an expression to test whether *all* elements of `y` are greater than 1.
g. Write an expression to test whether *any* element of `y` is `NA`.
h. Write an expression to count the number of elements of `y` that are not `NA`.

Problem 4:

We will continue to use y in from the previous problem in this exercise.

a. Suppose you were to take many, many random samples from the non-`NA` elements of `y` (with replacement). On average, what fraction of them would you expect to be > 2?
b. Write an expression to take a sample, with replacement, of size 20 from the non-`NA` elements of `y`
c. How many of them did you expect to be > 2? How many actually were > 2?

Problem 5:

As we discussed in class, R has a number of probability distribution functions built in. You can see the list of them with ?distributions . Here, we'll use the functions for the normal distribution, abbreviated `*norm` (i.e. `pnorm()`, `dnorm()`, `qnorm()`, and `rnorm()`).

Remember: If you don't specify mean or sigma when you call these functions, it assumes a standard normal with `mean=0` and `sigma=1` by default. Hence, `rnorm(10)` will get you 10 random numbers from a a standard normal.

Let's practice using these by computing the following. Give both the R code you used and the numerical value in your answers. Be sure to think about what you get - do the results seem reasonable (e.g., no probabilities > 1, values that "make sense" given the means & SD's you're putting in, etc.).

a. What is the probability that `x ~ N(10,2)` will be ≤ 10?
b. What is the probability that `x ~ N(-1, 1)` will be greater than 1.3?
c. What is the probability that `x ~ N(1,1)` will be more extreme than ± 2 (i.e. greater than 2 or less than -2)?
d. What is the probability that `x ~ N(0,3)` falls between 2 and 4? (*Hint: consider the total area under the curve and ask where x doesn't fall.*)
e. Assuming `x ~ N(-3,2)`, what is the **q** such that half the area under the curve lies to the right of q?
f. Assuming `x` is normally distributed with mean 0 and SD=0.3, what is **q** such that $P(x \geq q) = 0.05$?
g. Consider a **z score**. What is **q** such that the probability that $z \geq q$ OR $z \leq (-q)$ is 0.05?
h. Assuming `x ~ N(0, 1)`, for what value **q** is 50% of the under the curve in a band between -q and +q?