PSYC 7014, Week 4 stats lab: Normality and Probability

For this homework we will using `R` to run simulations of data. It's critically important that you take a look at this week's walkthroughs (especially 1,3,4) In order to complete the homework. To help, I will provide some code templates

Part 1: Assessing Normality

Using a continuous variable from your lab data, perform the following:

1. use `psych::describe()` and report the mean, median, skew, and kurtosis
2. Plot a distribution of this variable. Does the plot look roughly normal? How does what you see relate to the skew and kurtosis values you just obtained
3. Create a QQ-plot to assess the normality of your distribution
4. What does the `shapiro.test()` say about this function?

Part 2: Assessing probability

1. Use `pnorm` to determine the probability of obtaining the following z-scores **or more extreme**. By this I mean that if **Z** is positive, what is the pobability of a score ≥ Z. If **Z** is negative, what is the prbability of a score ≤ **Z**. If using `pnorm` to get your answers remeber that if Z is positive you need to set `lower.tail=FALSE`
    - Z = .5
    - Z = -1.4
    - z = 2.5
    - Z = -3
2. Assuming a sample with a mean of 125 and an standard deviation of 11, use `pnorm` to determing the probability of obtaining the following scores or more extreme. Remember that for values **above** the mean you need to set `lower.tail=FALSE`
    - 105
    - 95
    - 152
    - 73

Part 3: Simulating a sampling distribution of means

Use `rnorm` to create a population of 50000 members with mean ≈ 35.2 and sd ≈ 6.8. From this population create a **sampling distribution of means**. To create this distribution sample `N` members of the population, resampling with replacement `i` many times.

1. Create a histogram of your sampling distribution of means.
2. Perform steps 1 and 2 for each of the following `N` (sample size), `simulations` scenarios:

- for `simulations` = 100; `N` = 10, `N` = 30, `N` = 100
- for `simulations` = 1000; `N` = 10, `N` = 30, `N` = 100

  How does the shape of sampling distribution of means change when you increase `N` (the size of each sample)? How does this relate to the **standard error of this distribution**? Does the shape of the histograms seem to be affected by increasing the number of simulations.

BONUS: The Central Limit Theorm

Using this code modified from Walkthrough 1, generate a skewed distribution

```
set.seed(2020)
skewed_population <- SimDesign::rValeMaurelli(10000,
                                              mean=25,
                                              sigma=5,
                                              skew=1.7,
                                              kurt=3.4) %>% as.vector()
```

1. What is the value of skew for this distribution?
2. Create a **sampling distribution of means** from this `skewed_population` by running 20000 simulations taking 100 samples for each simulation.
3. Generate a histogram of your **sampling distribution of means**
4. How does this exercise demontrate the **Central Limit Theorem**?