# HW2 - Analyzing BRFSS.csv Dataset

**Statistics**

1. Write a program to simulate tossing a fair coin for 100 times and count the number of heads. Repeat this simulation 10**5 times to obtain a distribution of the head count and plot the histogram as well as CDF. Label your plots clearly.
2. Use the binomial distribution CDF (use scipy.stats.binom.cdf) to estimate the probability of having NO MORE THAN k heads out of 100 tosses, where k = 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Do these probabilities agree with the numbers of head counts you obtained in 3a? (Plot the head counts you obtained from the simulation results in 3a against the probabilities from your theoretical calculation here. Plot in loglog scale is probably needed to visualize small probabilities.)
3. Make a normal probability plot (thinkstats ch 4.4) to show that this distribution is close to a normal distribution with mean 50 and standard deviation 5.
4. Use normal distribution approximation to calculate the cumulative probabilities that you were asked to calculate in 3b, and compare the two results using a loglog plot. (Hint: If head count follows a normal distribution with mean = 50 and std = 5, a head count of 40 is equivalent to z-score = -2, and the corresponding CDF can be calculated using scipy.stats.norm.cdf.)

**Data analysis**

In this exercise we will be analyzing the BRFSS weight vs height data. Download data and code skeleton from course website. The code skeleton contains code to load the data and generate a numpy array object. The five columns in the numpy array represent: age, current_weight (kg), weight_a_year_ago (kg), height (cm), and gender, where gender == 1 represents male and 2 represents female.

1. Use your code from HW1 to produce a summary statistics graph on current_weight, weight_a_year_ago, and height.
2. Define weight_change = (current_weight – weight_a_year_ago). Calculate correlation between weight_change and the following variables, and determine which one is most correlated (regardless of sign of correlation) with weight_change. Use scatter plot to support your conclusion.
    1. current_weight
    2. weight_a_year_ago
    3. age
3. Calculate and compare the mean and SEM (standard error of the mean) for the weight_change of male and female. Use t-test to test whether there is a significant difference between the weight_change of male and female.
4. Randomly split the subjects into two groups of roughly equal sizes, and use t-test to test whether there is a significant difference between the weight_change of the two groups. Repeat the process 1000 times and plot the distribution of the

-log10(p-value) of the t-test results. What can you say about the difference between male and female in terms of their weight_change? (Consider both the p-value and the absolute differences between the two means.)

5. Define weight_height _ratio as current_weight/height. Use t-test to test whether there is a significant difference between the weight_height_ratio of male and female. Also, repeat the analysis you did in 4d, but replace weight_change with weight_height_ratio in your analysis.