Professor John Romley
PPD303 Statistics for Policy, Planning and Development
Fall 2020

Problem Set 4

Probability

1. (35 points) If you draw an M&M candy at random from a bag that contains 50 candies, the candy you draw will have one of six colors. The probability of drawing each color depends on the proportion of each color among all candies made.
Here are the probabilities of each color for a randomly chosen M&M:

Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.3	0.2	0.2	0.1	0.1	?

(a) What must be the probability of drawing a blue candy? (5 pts)

(b) What is the probability of NOT drawing a red candy? (5 pts)

(c) What is the probability that an M&M is either red, yellow, or orange? (5 pts)

(d) A bag originally contains 50 candies. You draw two M&Ms from the bag without replacing them back into the bag. What is the probability that both M&Ms drawn are Yellow? (5 pts)

(e) Suppose you are back to 50 M&Ms with the same probabilities as given in the prompt. You draw two M&Ms: one Red first and Brown second, <u>without</u> replacing them back into the bag. What is the *conditional probability* that Brown candy was drawn second—given that the Red was drawn first? (5 pts)

(f) Let's go back to the bag of 50 M&Ms. I steal all blue M&Ms from the bag. With the remaining M&Ms, what is the probability that you draw a yellow M&M first and NOT draw a green M&M second—given that you put the first M&M back to the bag before drawing the second? [Hint: Work with actual number of M&Ms in the bag] (10 pts)

<u>Law of Large Numbers, Central Limit Theorem, and Confidence Intervals</u>

2. (15 points) In an exercise, your Professor generated random numbers in Excel. The mean is supposed to be 0.5 because the numbers are supposed to be spread at random between 0 and 1. I asked the software to generate samples of 100 random numbers repeatedly. Here are the sample means $\bar{x}$ for 50 samples of size 100:

0.532	0.450	0.481	0.508	0.510	0.530	0.499	0.461	0.543	0.490
0.497	0.552	0.473	0.425	0.449	0.507	0.472	0.438	0.527	0.536
0.492	0.484	0.498	0.536	0.492	0.483	0.529	0.490	0.548	0.439
0.473	0.516	0.534	0.540	0.525	0.540	0.464	0.507	0.483	0.436
0.497	0.493	0.458	0.527	0.458	0.51	0.498	0.480	0.479	0.499

The sampling distribution of $\bar{x}$ is the distribution of the means from all possible samples. We actually have the means from 50 samples.

(a) Make a histogram of these 50 observations. Use bins of width 0.02, starting at 0.41. I suggest using Excel's bar chart function.
(b) Write down the Law of Large Numbers and the Central Limit Theorem, in your own words.
(c) Compute the mean and the median of the sample. Use this information and the histogram in (a). Does the distribution appear to be roughly normal, as the central limit theorem says will happen for large enough samples?

3. (20 points) Here are the IQ test scores of 31 seventh-grade girls in a Midwest school district:

114	100	104	89	102	91	114	114	103	105	
108	130	120	132	111	128	118	119	86	72	
111	103	74	112	107	103	98	96	112	112	93

(a) We expect the distribution of IQ scores to be close to normal. Make a histogram of the distribution of these 31 scores. Use bins of width 5, starting at 65; I suggest using Excel's bar chart function. Does your histogram show outliers?
(b) Using a calculator (or Excel), find the mean and standard deviation of these scores.
(c) Treat the 31 girls as an SRS (simple random sample) of all middle-school girls in the school district. Calculate the 95% confidence interval for the mean score in the population.
(d) In fact, the scores are those of all seventh-grade girls in one of the several schools in the district. Explain carefully why we cannot trust the confidence interval from (c).

Hypotheses Testing

4. (10 points) Experiments on learning in animals sometimes measure how long it takes mice to find their way through a maze. The mean time is 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of several mice takes with a noise as stimulus. What are the null hypothesis H_0 and alternative hypothesis H_a?

5. (20 points) The mean IQ for the entire population in any age group is supposed to be 100. Treat the IQ scores in Question 3 above as if they were an SRS (simple random sample) from all middle-school girls in this district. Do the scores provide good evidence that the mean IQ of this population is not 100? Go through steps a – c below to figure this out. Use s to estimate σ.

(a) What is the null hypothesis and the alternative hypothesis?

(b) Compute the z-score and the p-value.

(c) Interpret your results:

i) Do you reject your null hypothesis?

ii) Are your results statistically significant at the $\alpha = 0.025$ level?

iii) What does the p-value mean?