Name: _____ SID: _____

Collaborators: _____

Week 3 Problem Set

The normal curve and sampling distributions of averages and proportions
PHW142

You must put your name and SID at the top of the page.

This problem set focuses on the skills needed to use the normal curve for both populations and sampling distributions.

For every problem, briefly explain your logic and show all important steps.

What does this mean?

Show the key formulas and substitute the specific values from the problem into the formulas, show the z value if needed, include the R function you used and its results or a screenshot of the OOMPH Stat results, in addition to the final answer.

For the normal curve problems, you may use the OOMPH online normal calculator, or the R functions `pnorm` and `qnorm`. I strongly suggest not using Table B from Baldi and Moore.

Sketches of areas under normal curves are really helpful, but there's no easy way to get hand-drawn figures into a document like this except by scanning. I encourage you to make them to work through the questions, but you do not need to submit them.

For binomial calculations, use the R `pbinom` function.

Please submit your work as a single PDF file.

When you upload your problem set solution, the problem set key will be unlocked for you. Because the problem set key will be released when you upload your solution, you may only upload once.

No exceptions.

This problem set is worth 22 points.

Significant digits:

If you look at the examples in Baldi and Moore, they round their z scores to 2 digits after the decimal place, so that students may use Table B more easily. With OOMPH Stat and R's pnorm function, there is no need to do that. If needed, use at least 3 digits after the decimal place in your z scores. For normal curve areas, it's conventional to use 4 digits after the decimal place.

Some further notes on formatting your problem set submission:

You may write out the names of symbols in words: mu, sigma, x-bar, p-hat and so on.

You can also search Google with "[math symbol] copy and paste" to copy and paste the character into the PDF.

weights of full-term babies

adapted from Baldi and Moore exercise 11.36 in both 3rd and 4th editions

1. **(5 points)** In the exercises in practice problems, Baldi and Moore gave us the information that the distribution of birth weight in grams for infants born at full term is approximately follows a normal curve with mean 3350 grams and standard deviation 440 grams.

   1.1 What is the probability that a randomly chosen fill-term birth baby is low birth weight, less than 2500 grams? **1 point**

   1.2 What is the probability that a randomly chosen full-term birth baby would have very high birth weight (macrosomia) defined as birth weight over 4,000 grams? **1 point**

1.3 What is the probability that a randomly chosen full-term birth baby would have a birth weight greater than 2500 grams but less than 4000 grams? **1 point**

1.4 If we know 25 full-term babies were born on a given day, and we assume that the births are independent events, what is the probability that none of them have risks due to low birth weight or very high birthweight (macrosomia)? **2 points**

arsenic in blood

Baldi and Moore exercise 11.40 in the 3rd edition and 11.35 in the 4th edition

2. **(2 points)** Arsenic is a compound that occurs naturally in very low concentrations. Arsenic blood concentrations in healthy adults are normally distributed with a mean $\mu = 3.2$ micrograms per deciliter (μg/dl) and standard deviation $\sigma = 1.5$ micrograms per deciliter(μg/dl).

   What is the range of arsenic blood concentrations corresponding to the middle 90% of healthy adults? (Calculate the lower and the upper endpoints of the interval.)

lightning strikes

exercises 13.28 and 13.30 in both editions of Baldi and Moore

3. **(5 points)** The number of lightning strikes on a square kilometer of open ground in a year has mean 6 and standard deviation 2.4. These values are typical of much of the United States. The National Lightning Detection Network uses automatic sensors to watch for lightning in a sample of 25 randomly chosen square kilometers.

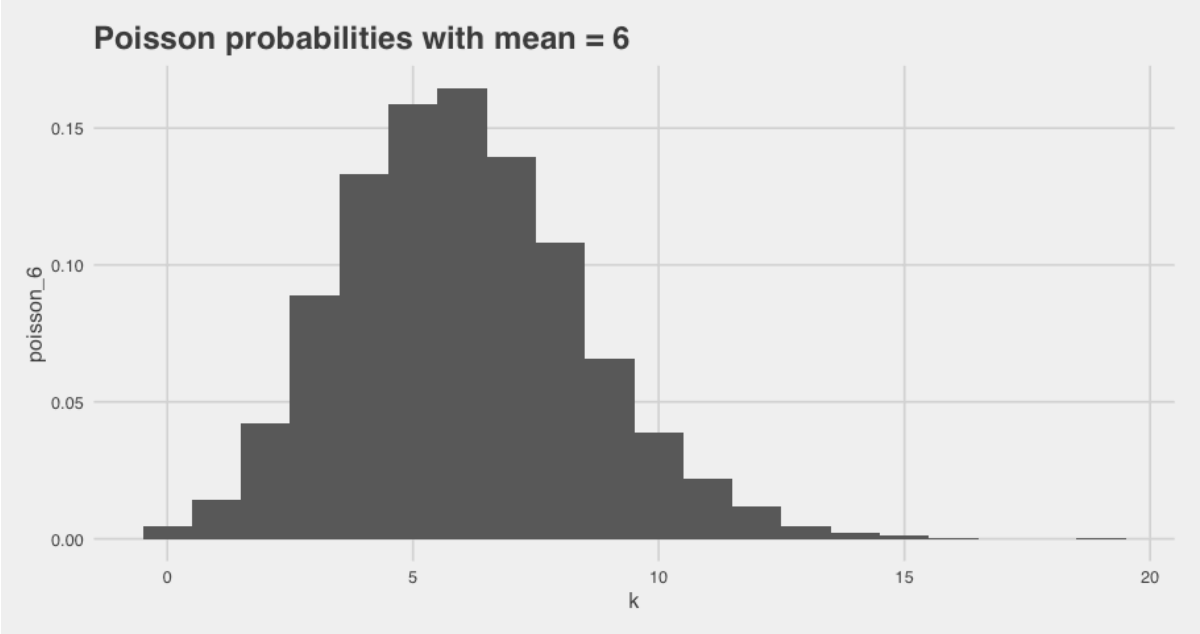
   A background note:

   If we were asking questions about the number of lightning strikes in a year in a single area, we would use the Poisson distribution. One of the features of the Poisson distribution is that the mean and the variance are the same. With a mean of 6 and a standard deviation of 2.4, which is very close to the square root of 6 ($\sqrt{6} = 2.45$), this scenario might fit the Poisson model assumptions.

   Because we are working with the average count for $n = 25$ areas, the normal curve could be used to answer questions about $\bar{x}$, provided the sampling distribution of $\bar{x}$ is approximately normal curve.

   With n only 25, the sampling distribution of $\bar{x}$ is approximately normal curve only if the distribution of the counts is not strongly skewed or plagued by outliers.

   I wish Baldi and Moore had said more about this, but it will be OK to use the normal curve for $\bar{x}$ even though the sample size is only 25, because the spike plot of the probabilities of Poisson counts with mean = 6 looks like this, at least in the range from 0 to 15.



3.1 What are the mean and standard deviation of $\bar{x}$, the mean number of strikes per square kilometer? **1 point**

3.2 What is the probability that the average number of lightning strikes per square kilometer per year for 25 randomly chosen square kilometers is 4.8 or less? **2 points**

3.3 What size sample would be needed to have the standard deviation of $\bar{x}$ equal to .20? **2 points**

STDs

exercise 13.42 in the 4th edition of Baldi and Moore

4. **(5 points)** To track epidemics, the CDC requires physicians to report all cases of important transmissible diseases. In 2014, for instance, a total of 350,062 cases of the sexually transmitted disease (STD) gonorrhea were officially reported, 53% of which were individuals in their 20s. Assume that the value 53% remains the same every year. Researchers plan to take a random sample of 400 diagnosed cases of gonorrhea to study the risk factors associated with the disease. Call $\hat{p}$ the proportion of cases in the sample corresponding to individuals in their 20s.

   4.1 What are the mean and standard deviation of the sampling distribution of $\hat{p}$ in random samples of size 400? **1 point**

4.2 Explain why, under these conditions, the sampling distribution of $\hat{p}$ is approximately normal. **1 point**

4.3 Use the normal approximation to find the probability that no more than half of the 400 gonorrhea cases sampled would be from individuals in their 20s. **1 point**

4.4 Use the binomial distribution to find the probability that no more than 200 of the 400 gonorrhea cases sampled would be from individuals in their 20s, and compare to your answer to 4.4. **2 points**

does caffeine affect the taste of cola?

5. **(5 points)** Suppose that an investigator hypothesized that consumers cannot really tell the difference between 2 versions of a cola drink that are identical except one naturally contains caffeine and the other does not. If the investigator is correct, then if we give a person a sample of each cola, tell them that one of the 2 identical-looking samples has contains caffeine and the other does not and ask them which one they think has the caffeine, they should correctly identify the one containing caffeine with probability 1/2.

   We will assume that the investigator's hypothesis is correct, meaning that the population proportion really is .5 for this question.

   The investigator carries out this experiment with n = 250 subjects.

   5.1 Explain why the sampling distribution of $\hat{p}$ is approximately normal. **1 point**

   5.2 What is the expected value of $\hat{p}$? What is the standard deviation of $\hat{p}$? **1 point**

5.3 What percent of all possible samples will give $\hat{p}$ values between .45 and .55 ?

5.4 Suppose the investigators found that 58% of their participants correctly identified the cola sample as the one that contained caffeine. Explain why these results support the claim that people really can tell the difference between the cola with and without caffeine. **2 points**