Question 5.

In lab, we tried to show that the sample standard deviation is biased to under-estimate the population standard deviation when we divide by N rather than N – 1. Briefly and in your own words, explain why this bias occurs.

(Hint: The answer here is very much related to your answer for Question 4.)

[This is much longer than your answer needs to be.]
Broadly, we can show why you need to divide by N-1 by starting with the formula for the standard deviation in the population:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

However, in a sample we do not know what μ is, so we need to estimate it using $\bar{x}$. We also know that the mean for any set of numbers will produce the smallest sum of squared errors. Thus, in the best-case scenario, if $\bar{x} = \mu$, then they would produce the same sum of squared errors. And if $\bar{x} \neq \mu$ then the sum of squared errors will always be smaller for the sample mean than for the population mean.

Because this numerator is going to be smaller when we use the sample mean, then we need to divide by N-1 in order to proportionally shrink the denominator. Thus, when dividing by N-1, we can ensure that the sample standard deviation is proportional to the population standard deviation:

$$\sigma \approx s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}}$$

##
For those of you who are so inclined, we can also do a formal proof for why it is N-1 specifically (as opposed to N-0.5 or some other number).

First, lets consider the expected difference between the population variance, σ^2, and the sample variance if we divide by N, called s^2_{bias}.

$$E[\sigma^2 - s^2_{bias}] = E[\frac{1}{n}\Sigma(x_i - \mu)^2 - \frac{1}{n}\Sigma(x_i - \bar{x})^2]$$

The squared terms can be re-written as:

$$= E[\frac{1}{n}\Sigma((x_i^2 - 2x_i\mu + \mu^2)) - \frac{1}{n}\Sigma((x_i^2 - 2x_i\bar{x} + \bar{x}^2))]$$

From there, and with some work, this can be reduced to:

$$= E[\frac{1}{n}\Sigma((\mu^2 - \bar{x}^2 + 2x_i(\bar{x} - \mu))]$$

As $\mu^2 - \bar{x}^2$ would be added to each term in the summation, we effectively have $(n\mu^2 - n\bar{x}^2)/n$, so these terms can be moved outside of the summation:

$$= E[\mu^2 - \bar{x}^2 + \frac{1}{n}\Sigma((2x_i(\bar{x} - \mu))]$$

$$= E[\mu^2 - \bar{x}^2 + 2(\bar{x} - \mu)\bar{x}]$$

$$= E[\mu^2 - 2\bar{x}\mu + \bar{x}^2]$$

$$= E[(\bar{x} - \mu)^2]$$

And the expected difference between the sample mean and the population mean is the variance of the sampling distribution, $var(\bar{x})$, which is

$$= \frac{\sigma^2}{n}$$

So the expected value of the biased estimator would be:

$$E[s^2_{bias}] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Thus, an unbiased estimator would be:

$$s^2_{unbiased} = \frac{n}{n-1}s^2_{bias}$$