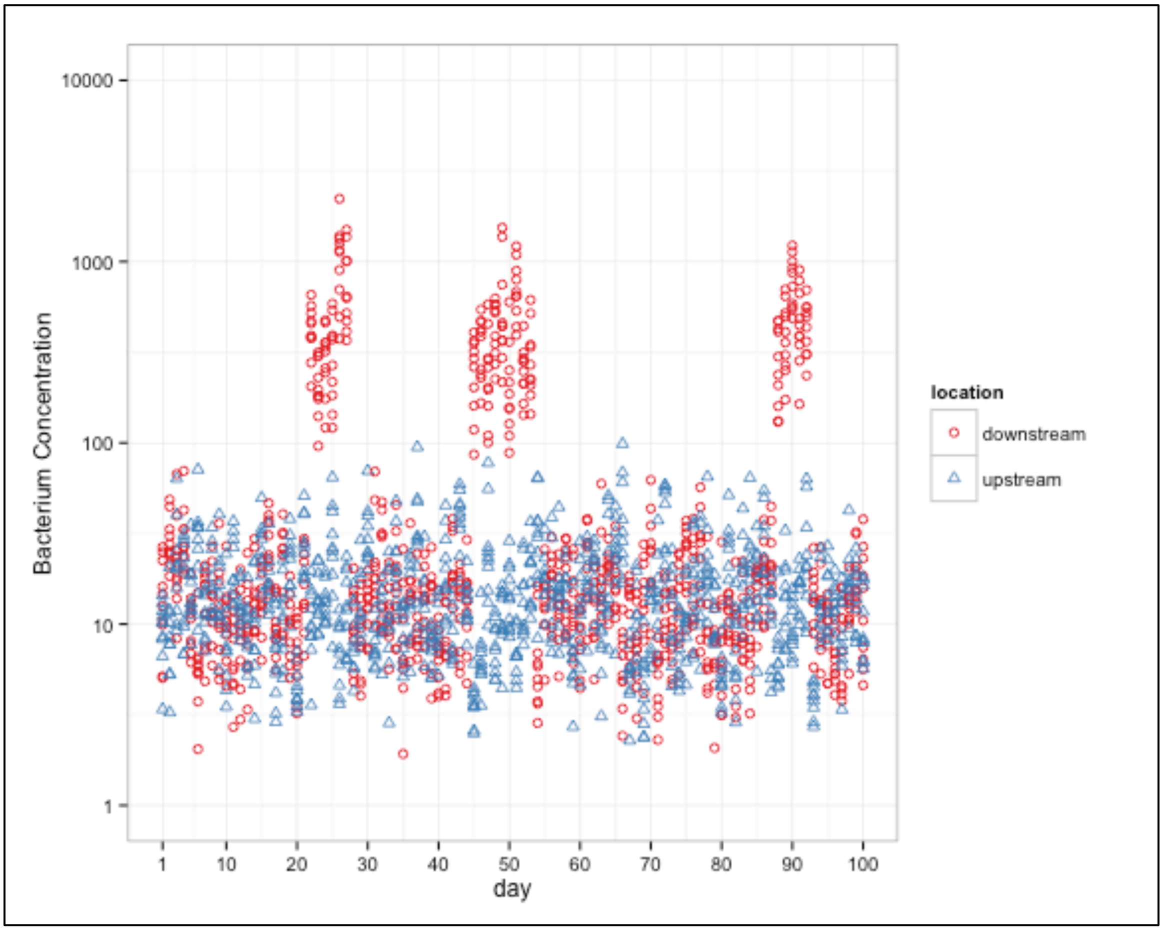
Name:

(please write your name on all pages)

Question 1

You have recently taken over an EPA lab that monitors a river for wastewater discharge from a sewage plant. The current protocol is to take twenty samples per day of river water: ten from upstream of the plant, and ten from downstream of the plant. The concentration of bacterium X (which is associated with nasty intestinal disorders) is measured for each sample, and a t-test is performed to see if the downstream level of the bacterium differs significantly from the upstream level. If the p-value is less than 0.05, an inspector is sent to the plant to look for engineering problems.

The plot below shows data from recent tests, and the analysis for day 50 is also given. Inspection records show confirmed discharge of contaminated wastewater from three periods (days 22-27, days 45-53 and days 88-92).



```
> downstream
 [1] 360.38315 153.73575 215.82799 598.96846 127.22812 156.50622 186.06700
 [8] 109.37437 251.37803  88.09115
> upstream
 [1] 10.630916  8.515942  9.161750 21.436415 17.580860 28.767846 20.503948
 [8] 13.694587  9.538456 14.163520
> t.test(downstream, upstream)

	Welch Two Sample t-test

data:  downstream and upstream
t = 4.3104, df = 9.033, p-value = 0.001944
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  99.54599 319.16721
sample estimates:
mean of x mean of y
224.75602  15.39942
```

Comment critically on the current procedures. Is there anything you would change?

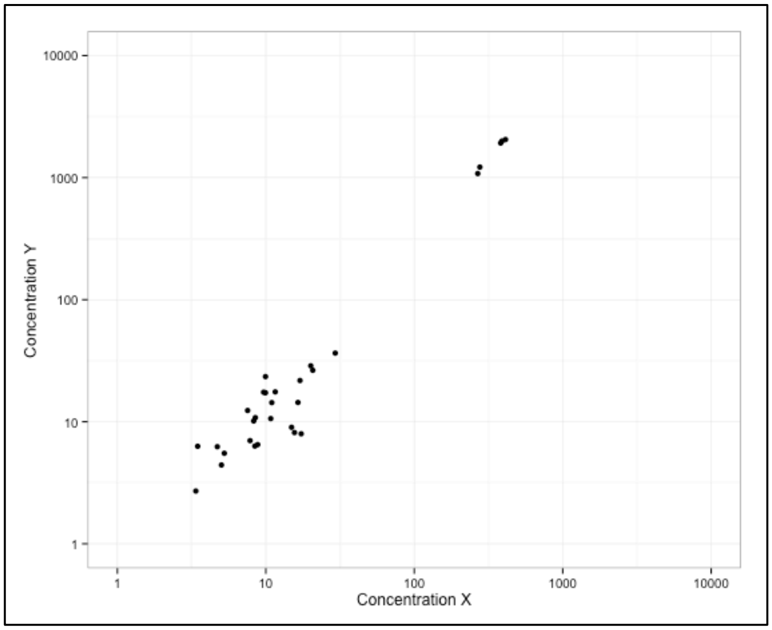
The current cost of processing each sample is $25. The instrument company offers a new kit that will produce more accurate measurements; they claim that the standard deviation of the instrument’s readings on a control sample is reduced by a factor of 10. However, each new kit costs $50. The salesman suggests that by using this kit, you’ll be able to reduce the number of samples you collect, and ultimately save money. Would you use this new kit? Why or why not?

Question 2

Continuing from the previous question, it has recently been discovered that bacterium X is in fact harmless, and that the true cause of illness is bacterium Y. Unfortunately, the protocol for measuring the concentration of bacterium Y takes several days, and can't be used to detect the discharge of contaminated wastewater in a timely manner.

Your helpful lab manager has decided to check if the concentrations of X and Y are related. He randomly selected 30 stored water samples from the last 100 days, and tested for the concentration of Y. The results of a regression, as well as a plot of the data, are shown below.

```
> cor.test.m <- cor.test(bacterium.X, bacterium.Y)
> print(cor.test.m)

	Pearson's product-moment correlation

data:  bacterium.X and bacterium.Y
t = 62.3987, df = 28, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9924108 0.9983165
sample estimates:
      cor
0.9964236

> cor.test.m$estimate^2
      cor
0.9928601
```



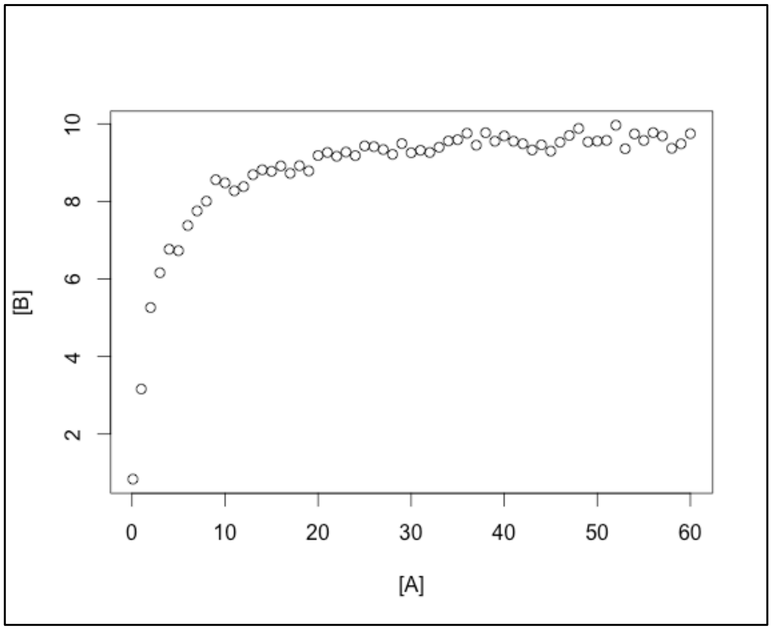
Your lab manager points out that the r^2 computed from the correlation test is 0.99, so it is OK to assume that the concentration of X is equal to the concentration of Y.

Comment on your lab manager's analysis and interpretation. Would you accept the suggestion, or propose something different? Explain your reasoning.

Question 3

It has been well established empirically in your lab that the equilibrium concentration of the activated form of your favorite enzyme, B, is related to the concentration of A by a classic Michaelis-Menten relationship:

$$[B] = \frac{V_m[A]}{K_A + [A]}$$



This fit of this model to the data is excellent, with the well established parameters being V_m = 10, and K_A=2.

A recent paper by a competitor suggests that the relation may be indirect, via an intermediate enzyme, the active form of which is I. It proposes the following relations:

$$[I] = \frac{I_m[A]}{K_I + [A]} \qquad [B] = \frac{J_m[I]}{K_J + [I]}$$

Combining these relations yields…

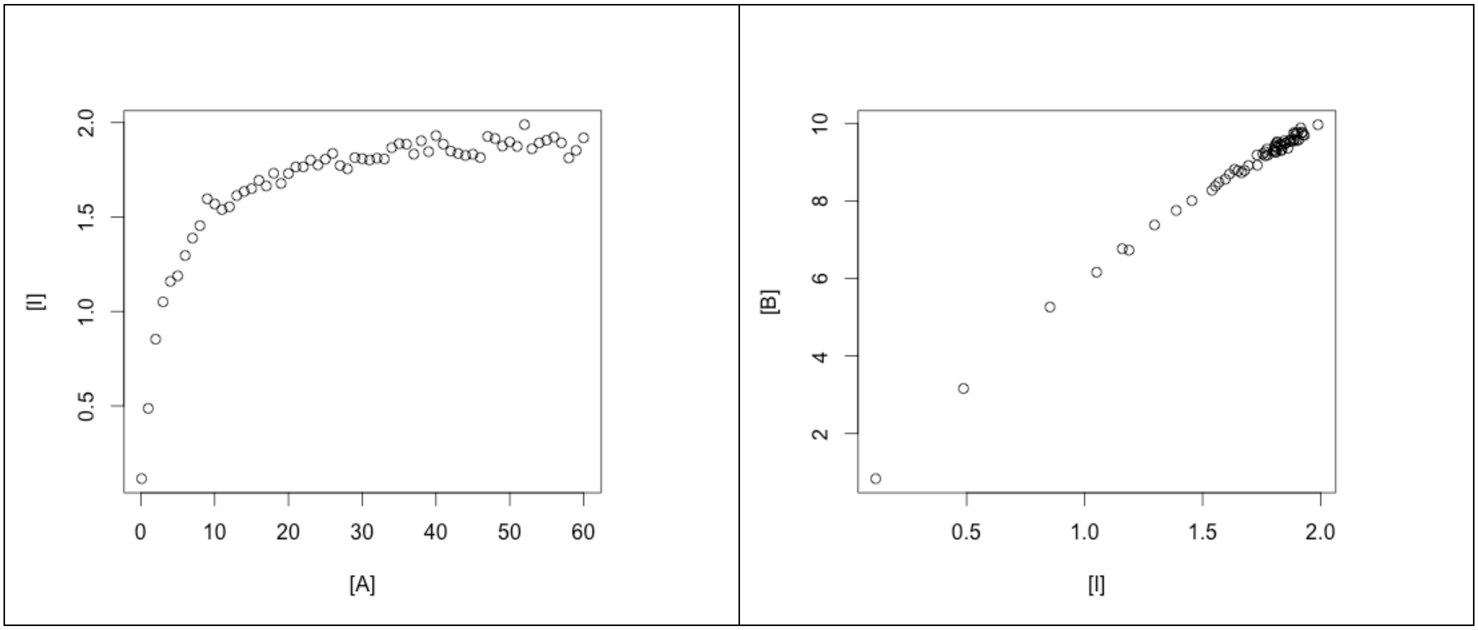
$$[B] = \frac{J_m \left\{\frac{I_m[A]}{K_I + [A]}\right\}}{K_J + \left\{\frac{I_m[A]}{K_I + [A]}\right\}}$$

After learning from you how to use the `nls` function in R, your lab mate has been trying for days to fit this four-parameter model to the data, but can never seem to get it to work, even though fitting the same data to the well-established two-parameter model works perfectly. A typical attempt is shown here:

```
> m1 <- nls(b ~ Vmax.fit * a / (Ka.fit + a), start=list(Vmax.fit=Vmax, Ka.fit=Ka))
> summary(m1)

Formula: b ~ Vmax.fit * a/(Ka.fit + a)

Parameters:
         Estimate Std. Error t value Pr(>|t|)
Vmax.fit  9.99115    0.03698  270.20   <2e-16 ***
Ka.fit    1.98154    0.06615   29.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.1741 on 59 degrees of freedom

Number of iterations to convergence: 2
Achieved convergence tolerance: 1.518e-06

> m4 <- nls(b ~ Jmax.fit * (Imax.fit * a / (Ki.fit + a)) / (Kj.fit + (Imax.fit * a / (Ki.fit + a))),
+           start=list(Jmax.fit=Jmax, Kj.fit=Kj, Imax.fit=Imax, Ki.fit=Ki))
Error in nlsModel(formula, mf, start, wts) :
  singular gradient matrix at initial parameter estimates
```

What advice would you give your lab mate about fitting the four-parameter model? Explain your reasons.

Frustrated with all this fitting, your lab mate has returned to the bench and collected data measuring how [I] varies with [A], and how [B] varies with [I], as shown below.



Your lab mate has been able to fit the [I] ~ [A] data, but didn't bother with the data in the right-hand plot since it didn't have the classic saturation curve representative of the Michaelis-Menten equation:

```
> m2 <- nls(i ~ Imax.fit * a / (Ki.fit + a), start=list(Imax.fit=Imax, Ki.fit=Ki))
> summary(m2)

Formula: i ~ Imax.fit * a/(Ki.fit + a)

Parameters:
         Estimate Std. Error t value Pr(>|t|)
Imax.fit 1.994664   0.009327  213.86   <2e-16 ***
Ki.fit   2.944995   0.100377   29.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03855 on 59 degrees of freedom

Number of iterations to convergence: 2
Achieved convergence tolerance: 5.019e-06
```

How would you respond to your lab mate's comments about the [B] ~ [I] plot?

Given all of the information above, can you estimate J_m and K_J?

Question 4

Give one advantage and one disadvantage of using a non-parametric test over a parametric one?

Give an example of a statistical test that is achieved by simulation. Discuss the advantages and disadvantages of computing p-values by simulation.

Bonus Question

Approximately how many base pairs are there in the human genome?

[extra space if you need it]

[extra space if you need it]

[extra space if you need it]