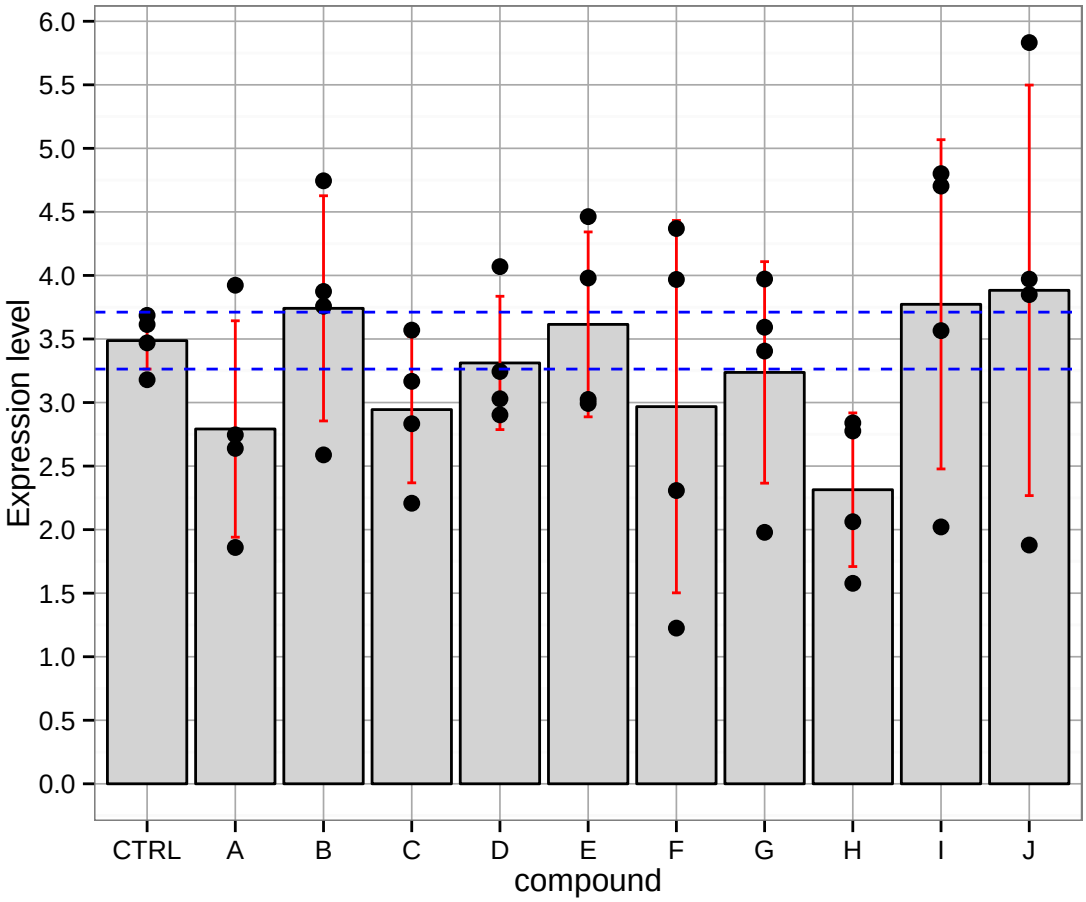
Question 1

Your lab is looking for a compound that suppresses the expression of EGFR. You have received the results of a small molecule screening study, which suggested ten candidate compounds (A-J). A postdoc in your lab was tasked with validating these compounds. To do this, he measured the expression of EGFR in a control cell culture treated with a known inert compound, and with each of the ten compounds. Each condition was represented by four biological replicates (raw data shown as points).

The postdoc presented the figure below at lab meeting. He explained that compounds A, C, F, G, and H were validated, because their measured mean (shown as the height of each bar) is outside of the range of error bars for the control case (indicated by the dotted blue lines). Of course, you ask what exactly the error bars are, and he replies that they span the range of $mean \pm 1SD$.



Comment critically on this postdoc's line of reasoning.

Based on the figure, which compounds would you conclude are suppressors of EGFR? Justify your answer.

Based on the figure, which compounds would you conclude have no effect on the expression level of EGFR? Justify your answer.

Outline the steps you would propose to analyze the data that your labmate has collected.

Question 2

Continuing from the previous question, **produce a rough estimate of the standard deviation of EGFR expression level measurements in this experimental setup?** Explain your reasoning.

Estimate how large of a difference between the control and any one compound this experiment reasonably had a chance of uncovering? (Assume power = 50% and $\alpha = 0.05$) Show your work.

[remember, power calculations are rough estimates, so feel free to round off inconvenient numbers]

Question 3

You have been appointed the student representative to the University's Faculty Recruitment Committee, and have been asked to help evaluate applicants' publication records in order to identify candidates who are leaders in their field. To assess this, the committee has decided to compare the impact factors of the 'best' paper published by each candidate. However, since papers that are published when fields are 'hot' tend to have higher impact factors, the committee has decided to determine the percentile ranking of each representative paper from those in the same field in the same year.

For example, Dr. Smith published a paper in Vision (an ophthalmology journal) in 2009 that received an impact factor of 4.32. To evaluate this paper, you took a random sampling of 99 other ophthalmology papers published in 2009, and got these impact factors:

```
##  [1]  0.10  0.18  0.22  0.23  0.24  0.25  0.29  0.33  0.36  0.37  0.40
## [12]  0.41  0.42  0.48  0.49  0.50  0.51  0.58  0.62  0.64  0.69  0.72
## [23]  0.73  0.77  0.81  0.82  0.84  0.85  0.89  0.89  0.95  1.13  1.20
## [34]  1.25  1.33  1.36  1.38  1.39  1.39  1.47  1.62  1.64  1.64  1.66
## [45]  1.67  1.68  1.68  1.69  1.71  1.71  1.73  1.79  1.81  1.93  1.94
## [56]  1.99  2.09  2.14  2.32  2.38  2.39  2.44  2.47  2.51  2.52  2.54
## [67]  2.59  2.77  2.77  2.78  2.95  3.07  3.22  3.42  3.46  3.51  3.66
## [78]  3.82  3.97  4.04  4.14  4.56  4.65  4.67  4.74  4.90  5.51  5.62
## [89]  5.93  6.32  7.15  7.95  8.49  9.13  9.86 10.45 10.69 11.41 11.65
```

What percentile rank would you assign to Dr. Smith's paper? Explain your reasoning.

One of the committee members points out that your assignment is only an estimate, and, to facilitate fair comparisons of candidates, you should compute a 95% confidence interval for each of your estimates. **How would you propose to compute the 95% CI of your percentile estimate?** *[Note: you don't need to compute the CI here, just explain how you would do so.]*

Your University pays for every impact factor retrieved from the bibliometric database, and your committee has received very limited funding to support its efforts, so the committee would prefer methods that require less data to be collected.

One of the candidates works in a very obscure field, for which you could find only nine other papers published in the same year as her representative publication. **How would you adjust your procedure for computing the CI of the percentile ranking for this case?** Explain your reasoning.

Question 4

A student from the local high school is interning in your lab for the summer. Your PI has given him some kinetic data for an enzymatic reaction, and asked him to fit the data to a Hill model.

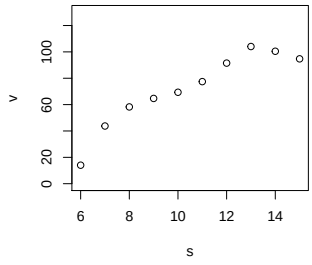
When reporting his results, the student claims that fitting to a simple quadratic ($v = a \cdot s^2 + b \cdot s + c$) yields a better model. He even computed AICs, and found that the AIC of the quadratic model was lower. Furthermore, he claims that the quadratic model is preferred because it comes from a linear regression, and notes that he initially had some trouble getting the Hill model to converge.

You ask to see the R code for his analysis, and he shows you the following:

```
d <- data.frame(s, v)
d # show the data
```

```
##     s         v
## 1   6  14.04087
## 2   7  43.75442
## 3   8  58.24108
## 4   9  64.67959
## 5  10  69.37014
## 6  11  77.44074
## 7  12  91.49114
## 8  13 104.06287
## 9  14 100.47581
## 10 15  94.72098
```

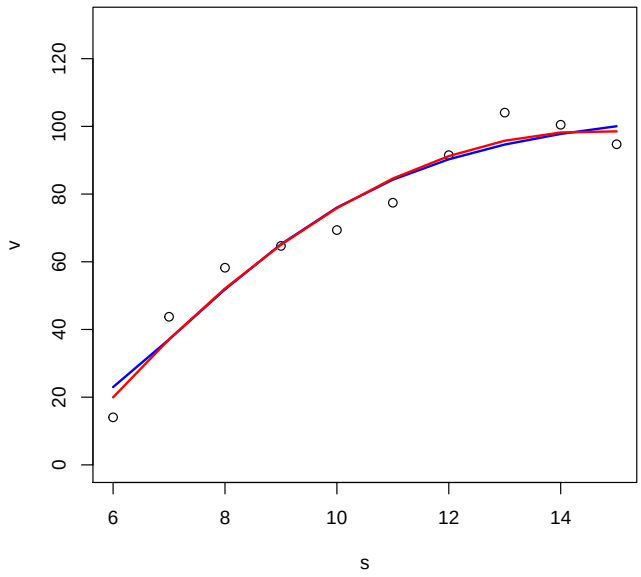
```
# plot the raw data
plot(v ~ s, data = d, xlim = c(6, 15), ylim = c(0, 130))
```




```
# fit models and plot regressed curves
m.hill <- nls(v ~ Vmax * s ^ n / (Km ^ n + s ^ n),
                    start = list(Vmax = 100, Km = 9,n = 4),
                    data = d)
```

```
lines(s, predict(m.hill, newdata = data.frame(s = s)),
                  col = "blue", lwd = 2)

m.poly <- lm(v ~ I(s^2) + s, data = d)
lines(s, predict(m.poly, newdata = data.frame(s = s)),
                  col = "red", lwd = 2)
```




```
#compare the models
summary(m.hill)

##
## Formula: v ~ Vmax * s^n/(Km^n + s^n)
##
## Parameters:
##      Estimate Std. Error t value Pr(>|t|)
## Vmax 107.2918    10.6311  10.092 2.01e-05 ***
## Km     8.1274     0.5294  15.353 1.20e-06 ***
## n      4.2822     1.0894   3.931  0.00567 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.394 on 7 degrees of freedom
##
## Number of iterations to convergence: 7
## Achieved convergence tolerance: 1.696e-06
```

```
summary(m.poly)
```

```
##
## Call:
## lm(formula = v ~ I(s^2) + s, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1478 -5.3776 -0.0385  5.1794  8.3063
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -126.8921    30.0679  -4.220  0.00394 **
## I(s^2)        -1.0493     0.2846  -3.687  0.00779 **
## s             30.7684     6.0197   5.111  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.539 on 7 degrees of freedom
## Multiple R-squared:  0.9583,Adjusted R-squared:  0.9463
## F-statistic: 80.35 on 2 and 7 DF,  p-value: 1.486e-05
```

```
c(aic.hill = AIC(m.hill), aic.poly = AIC(m.poly))
```

```
## aic.hill aic.poly
## 72.82431 70.36909
```

Critically evaluate the intern's analysis. What would you suggest be done next?

[extra space if you need it]

[extra space if you need it]

[extra space if you need it]

[extra space if you need it]

[extra space if you need it]