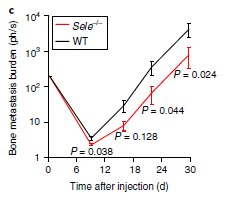
**Week12 Assignment Due November 20 (20 points plus 5 Extra credit points)**

Multiple comparisons, 2x2 table, correlation and simple linear regression



**Esposito M. et al. Nature Cell Biol (2019) 21, 627-639**

**Fig. 1 |** E-selectin is critical for bone but not lung metastasis. **c**, BLI quantification of the bone metastasis burden following intracardiac injection of the BM2 cell line into WT or Sele−/− SCID mice. Two-sided Mann–Whitney U tests; n = 12 mice per group.

The figure above appears to be data from a repeated measures study on the same mice that monitored the development of bone metastases over time. Instead of doing a two-way RM ANOVA (or mixed model), the authors choose to compare WT and Sele knockout cells at four different time points using multiple Mann-Whitney tests. This opens them up to the criticism of multiple testing.

Q1. Apply a Bonferroni correction to the data they reported in the figure above. Which comparisons remain significant after you take multiple comparisons into account? Include the adjusted p-value threshold in your response. (2 pts)

Given that there are 4 comparisons, the new threshold p-value would be 0.05/4 = 0.0125. Given this new significance threshold, none of the comparisons remain significant (all are >0.015).

Health officials were worried that people in village in Malawi may have been exposed to cholera when they found out sewage from an upstream village that had a cholera outbreak had accidentally spilled into the river. They did what they could to prevent the disease from developing, but some people ended up getting sick despite their efforts. They were concerned that children might be at increased risk due to the fact they were found playing in the river despite warnings from their parents and the health officials. The table below shows the number of people who got sick and those who didn’t in the village stratified by age groups.

Cholera Age group Count

No Adults 1006

No Children 1209

Yes Adults 21

Yes Children 45

Q2. Is this a prospective or case-control study and what measure of effect should you use to see if children are at increased risk for cholera? (1 pt)

Prospective, RR

Q3. Fill in the 2x2 table below to do your analysis to answer their research question. Label the column at rows with the appropriate variable and category to do the analysis in Prism to answer their research question (2 pts) pay attention to the order of the variables, it is different from Chi-square.

Variable:

\_\_\_\_\_\_ \_\_\_\_\_\_

Variable: \_\_\_\_\_\_ #: #:

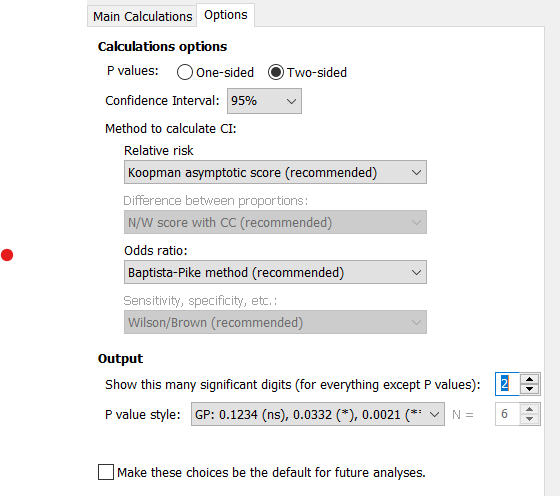
\_\_\_\_\_\_ #: #:

Variable: Cholera

YES NO

Variable: Age group Child #:45 #:1209

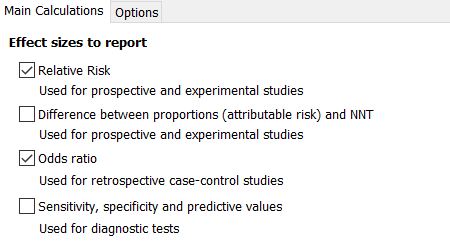
Adult #:21 #:1006

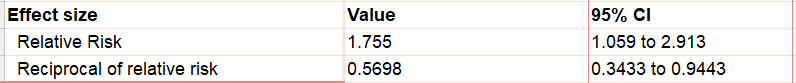


You can change the decimal in the options, but do not change other things.

Q4. Use Prism to calculate the measure of effect (and 95%CI) you chose in Q3. Write up your results calculating your results to 2 decimal points, and explain your results. Include if you reject or fail to reject the null hypothesis

based on the 95%CI. (2 pts)







RR= 1.76 (95%CI: 1.06, 2.91). (*RR could be 1.75 depending if they rounded up or down*). Children are at 1.76 times (or 76%) increased risk for getting cholera compared to adults. Because the 95%CI do not include 1.0, we reject the null hypothesis that the RRpopulation=1.0.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

A researcher wants to validate an ELISA test for the detection of *Giardia duodenalis*, a parasite of the intestinal tract. The test uses rabbit polyclonal antibodies to detect *G. duodenalis* antigen in fecal material. The results of her study are below.

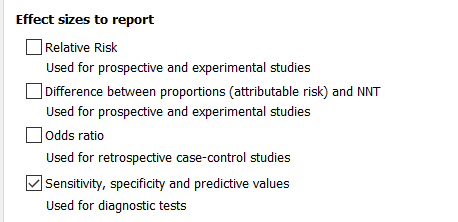
Parasitological diagnosis of *G. duodenalis*

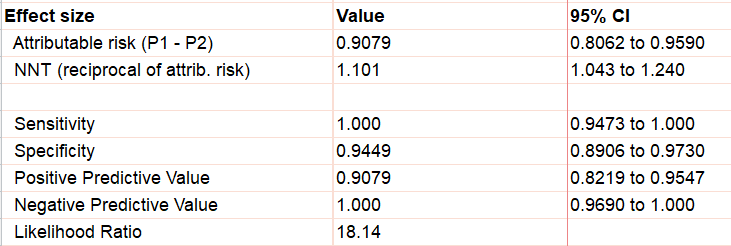
Positive Negative

ELISA test Positive 69 7

Negative 0 120

Q5. Use Prism to calculate the values in the table below: (1 pt)





Value (to 3 decimal points)

Sensitivity 1.000 or 100%

Specificity 0.945

Pos Predictive value 0.907

Neg Predictive value 1.000

Q6. Interpret the sensitivity (complete the sentence below). (1 pt)

Given that you have a *G. duodenalis* infection, you will have a positive test 100% of the time.

Q7. Interpret the positive predictive value (complete the sentence below). (1 pt)

Given that you have a positive ELIZA test, there is a 90.7% chance that you have a *G. duodenalis* infection

**Extra credit**

EC1. Fill in a 2x2 table based on the information below. Make sure to label the columns and rows to set up calculations for sensitivity, specificity, PPV, NPV to test if a positive test is associated with disease. (2 EC pts)

There are 200 people in the study. The prevalence of disease is 20%

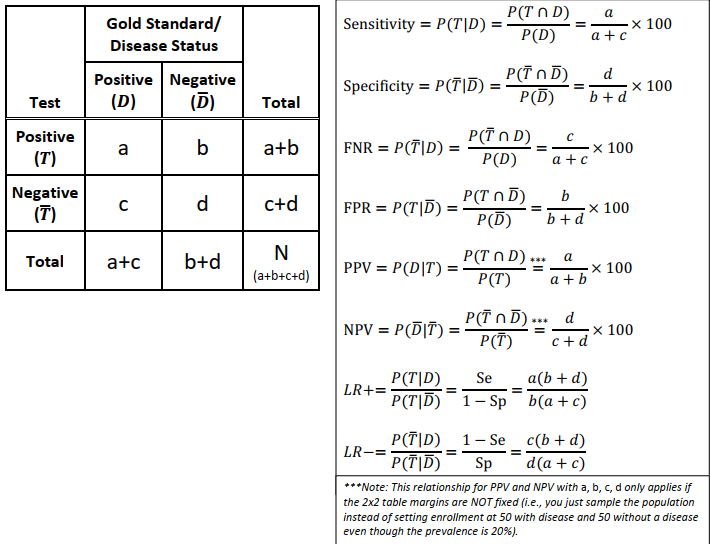
The specificity of the test is 80% The sensitivity of the test is 90%

Variable:

\_\_\_\_\_\_ \_\_\_\_\_\_

Variable: \_\_\_\_\_\_ #: #:

\_\_\_\_\_\_ #: #:

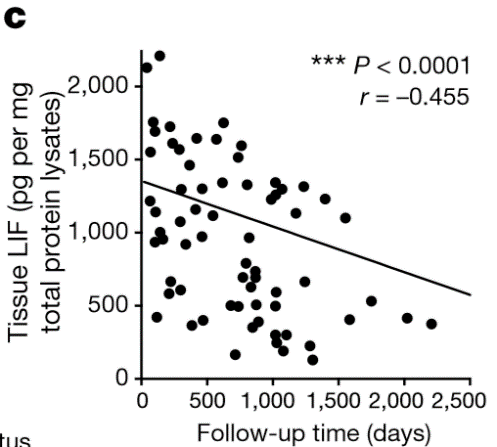


|  |  |  |  |
| --- | --- | --- | --- |
|  | Disease + | Disease - |  |
| Test+ | 36 | 32 | 68 |
| Test- | 4 | 128 | 132 |
|  | 40 | 160 | 200 |

EC2. Calculate and interpret the negative predictive value from the data you entered in to the table above. (2 EC pts)

NPV = 128/132 = 0.97 or 97%

If you have a negative test, there is 97% chance you do not have the disease.

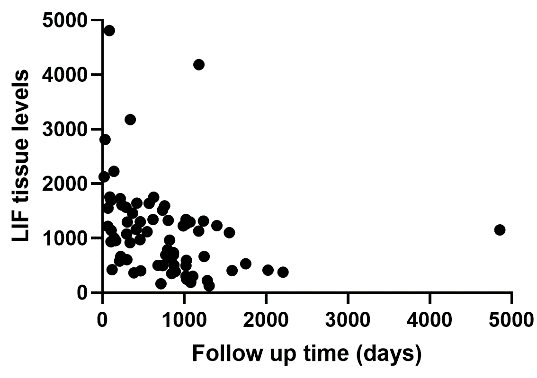


Shi Y, et al. Nature (2019) 569:131-137

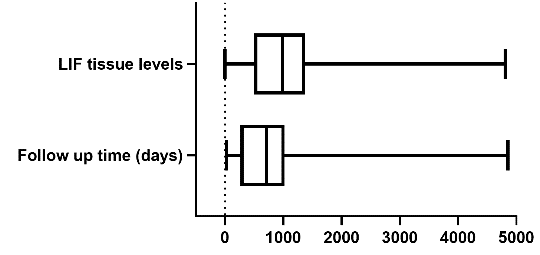
Fig. 5 | LIF can be a biomarker for PDAC monitoring. a–c, LIF levels in human pancreatic tissues quantified by ELISA (a), and their correlation with tumour differentiation status (b) and overall survival (c). n = 9 (normal tissue); n = 5 (chronic pancreatitis); n = 77 (PDAC). CP, chronic pancreatitis. Statistical significance was determined by nonparametric Spearman correlation test (c).

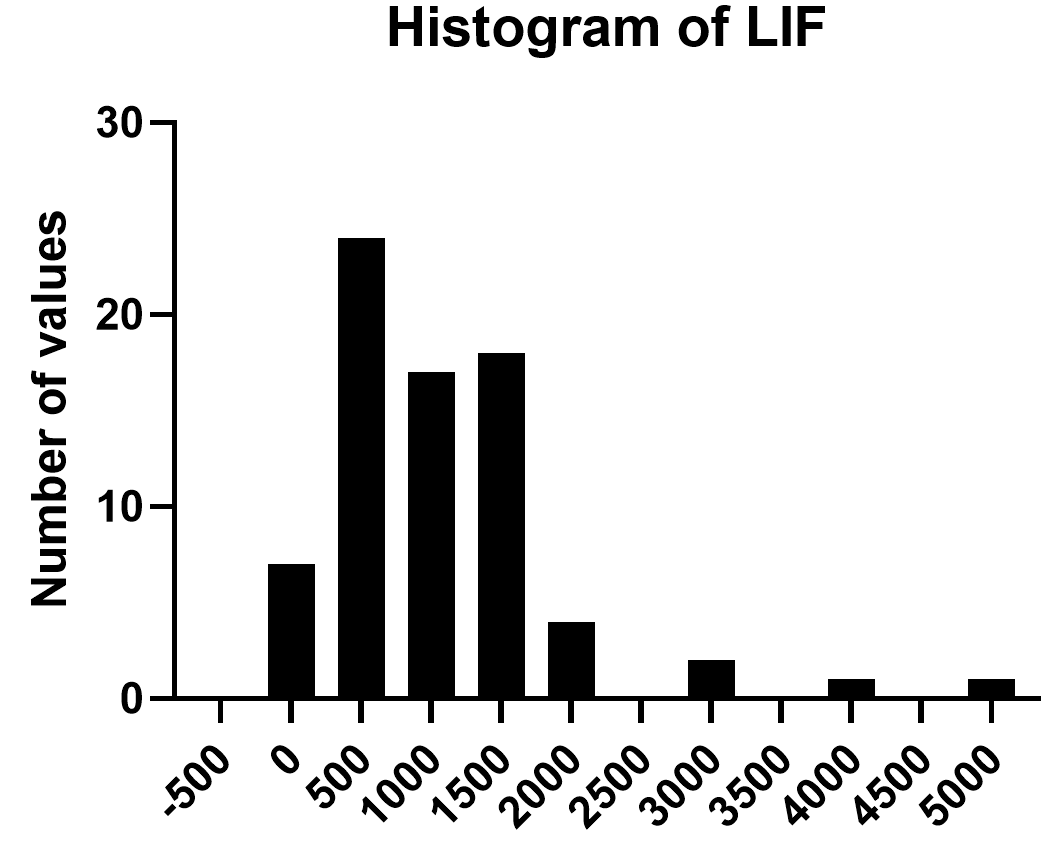
Q8. Why did the authors use the Spearman rather than the Pearson correlation coefficient? Plot the data yourself and see if you can find a reason why. (1 pt)

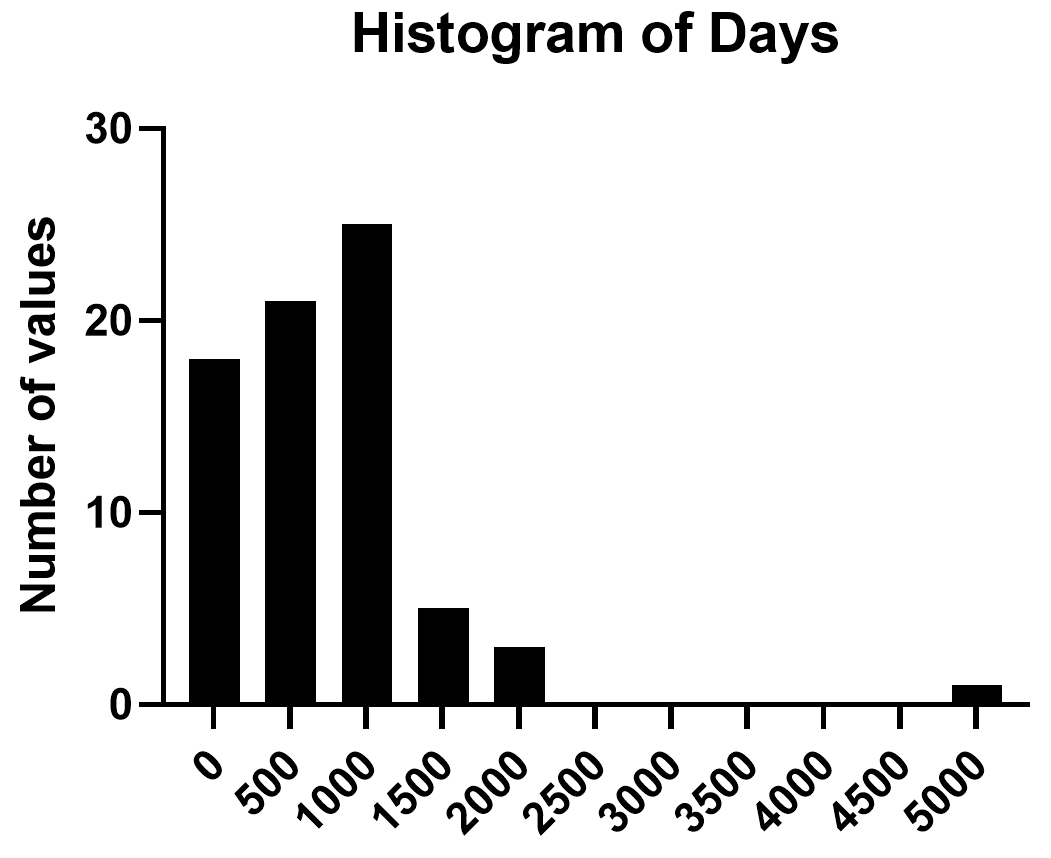
The graph they put in the paper had truncated axes that did not show all data. There are some potential outliers.

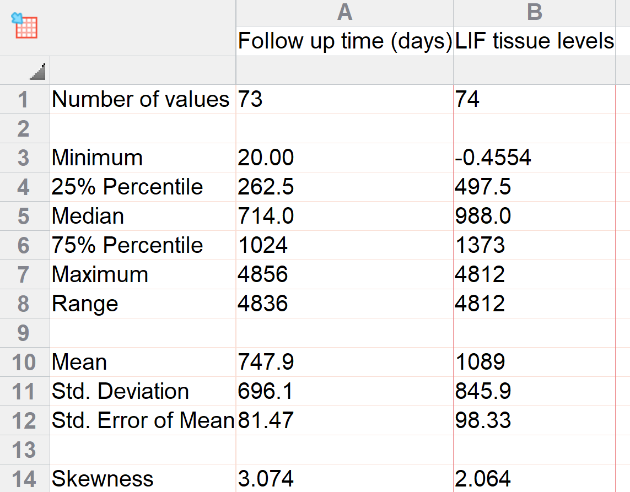


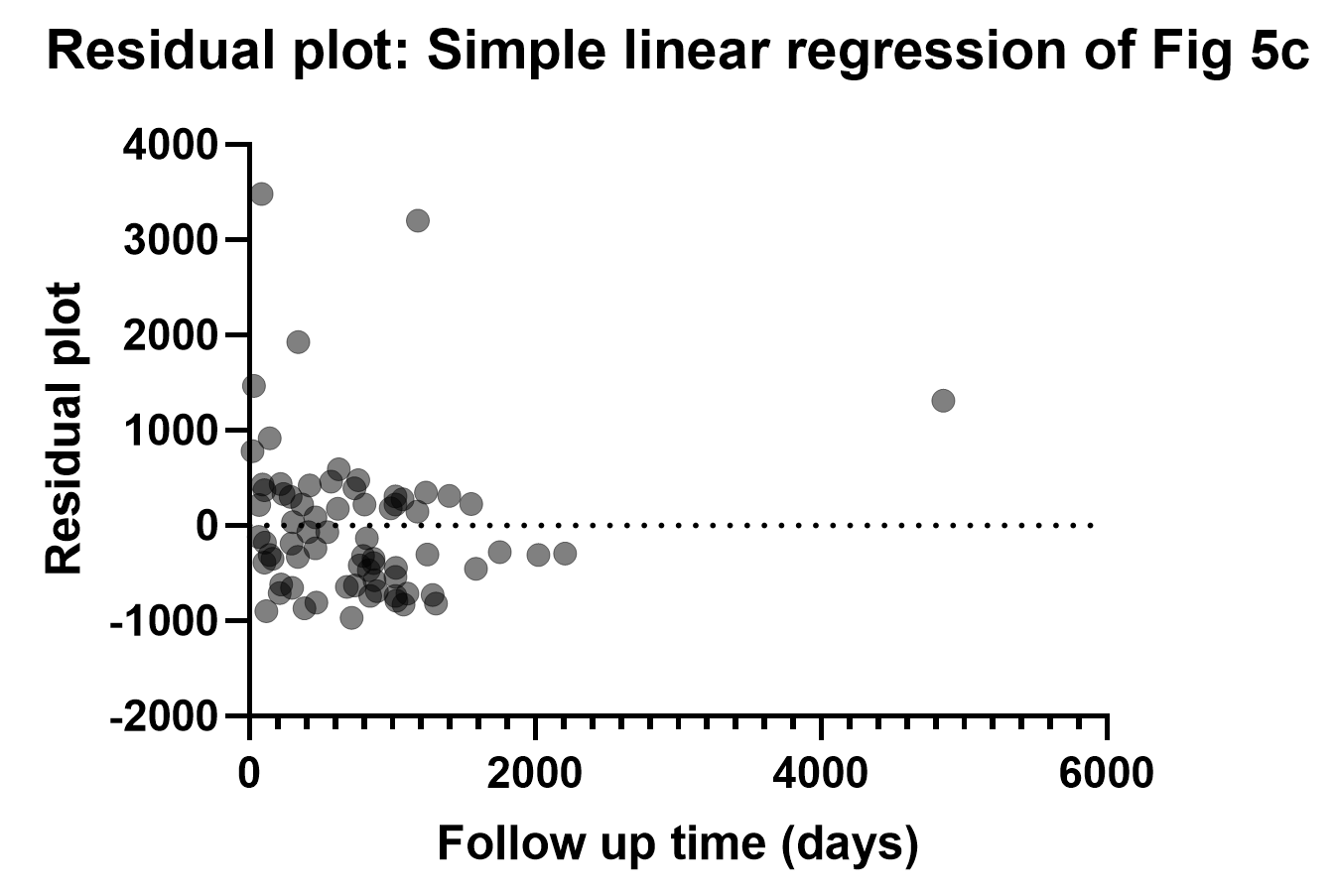
Q9. Check the assumptions for doing a Pearson correlation coefficient. I have provided the appropriate graphs and data. I have filled in part of the table for you below. (1.5 pts)











The variables must be continuous **YES**

The variables must be approximately normally distributed

No outliers

Homoscedasticity along the range of x values

There is a linear relationship between the two variables **? – maybe monotonic?**

Every data point must be in pairs **YES**

The variables must be continuous **YES**

The variables must be approximately normally distributed **NO**

No outliers **NO – Outliers are present**

Homoscedasticity along the range of x values **NO**

There is a linear relationship between the two variables **? – maybe monotonic?**

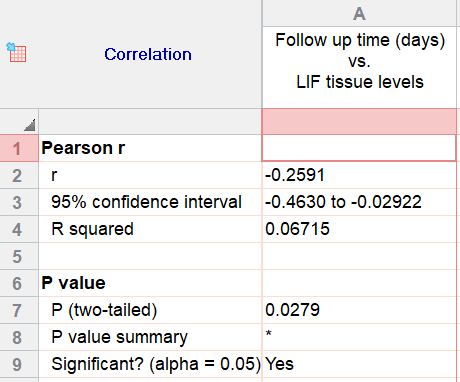
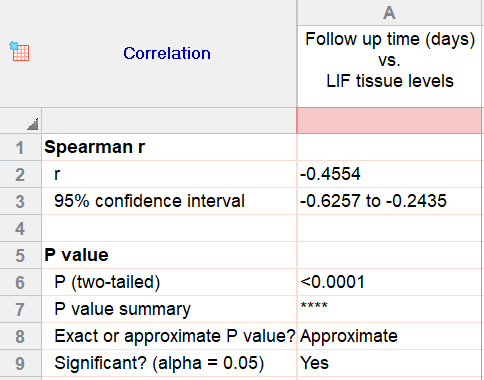
Every data point must be in pairs **YES**

Q10. Calculate both Spearman and Pearson correlation coefficients from the data. Fill in the table below (values to 3 decimal points). (1.5 pts)

r 95%CI p R2

Pearson -0.463 to -0.029

Spearman ---

r 95%CI p R2

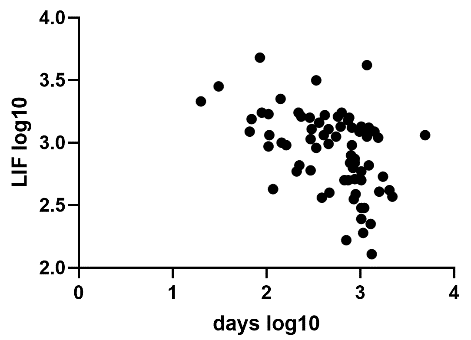
Pearson -0.259 -0.463 to -0.029 0.028 0.067

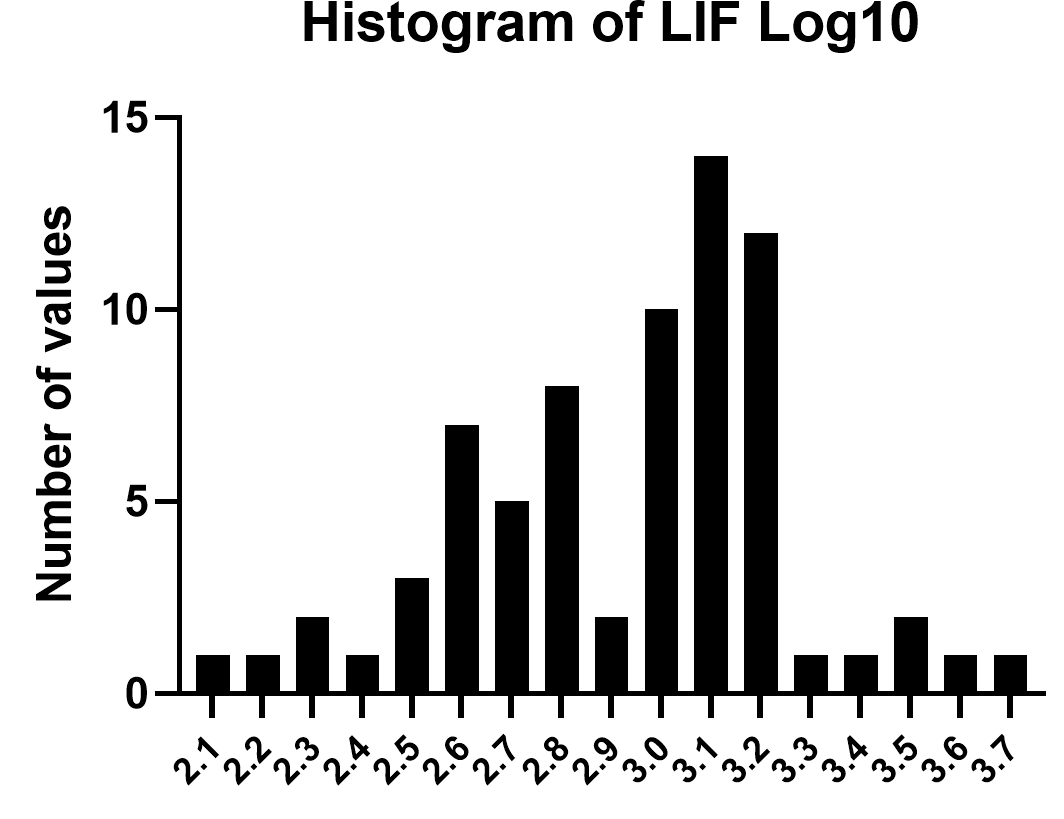
Spearman -0.455 -0.626 to -0.243 <0.0001 ---

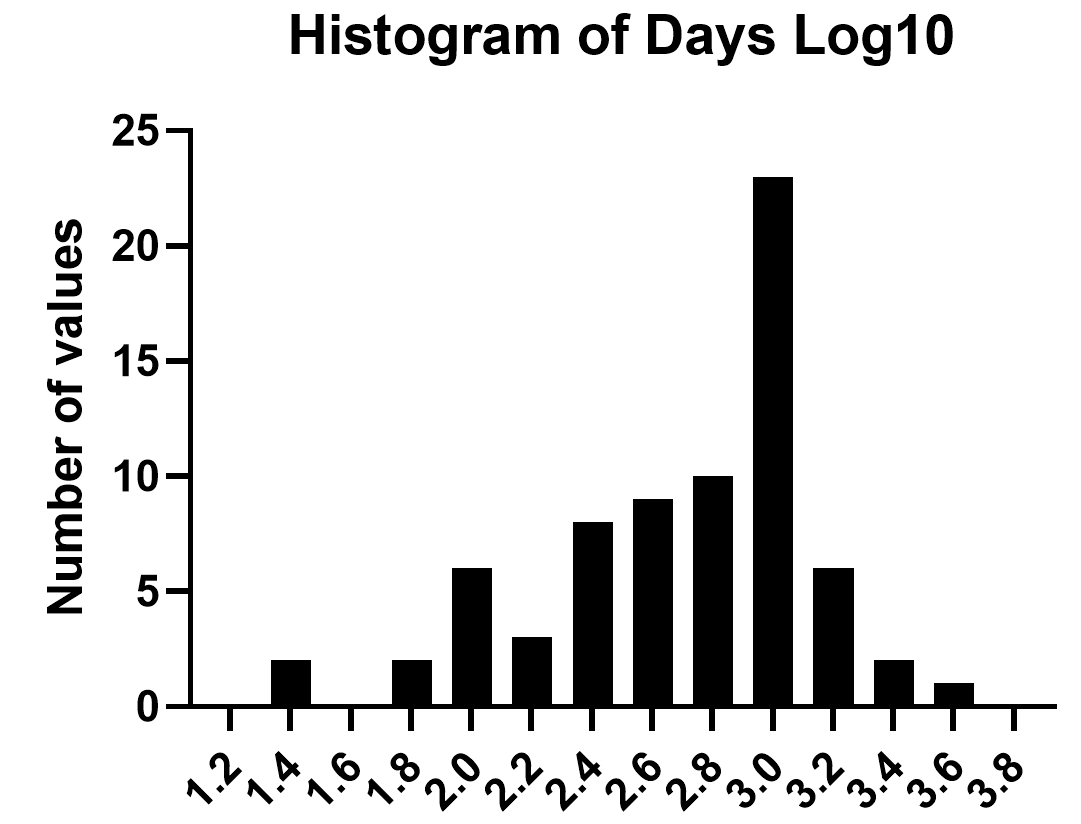
Q11. Based on your answers to the questions above, do you think the authors were correct in reporting the Spearman correlation coefficient? (1 pt)

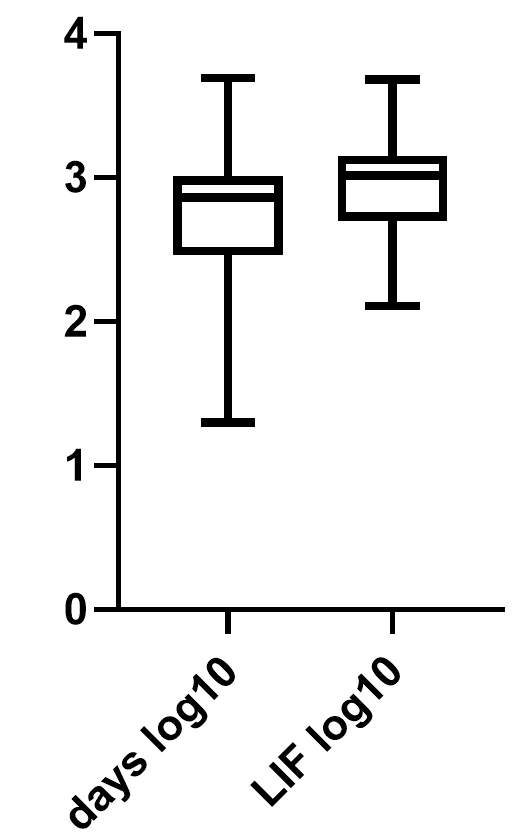
Yes, the data did not meet the assumptions of the Pearson test.

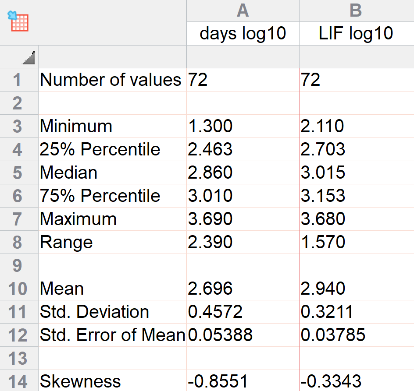
Q12. The authors think that follow-up time day may predict LIF tissue levels. They decided to do a simple linear regression to help test this. We know there is already a problem with at least one of the assumptions for a simple linear regression, specifically that the Y variable (LIF) is not normally distributed. We will try transforming the data (log10) to see if the data will meet the test assumptions for a linear regression. I have done some of the work for you. I first tried transforming only the Y variable, but decided I needed to transform the X variable as well. I completed the assumptions table below. Your job is to create and copy the residual homoscedasticity graph and Q-Q plot of the residuals and copy them below. (2 pts)











YES/NO

Normally distributed continuous dependent (Y) variable **YES symmetrical (some skew but both neg.)**

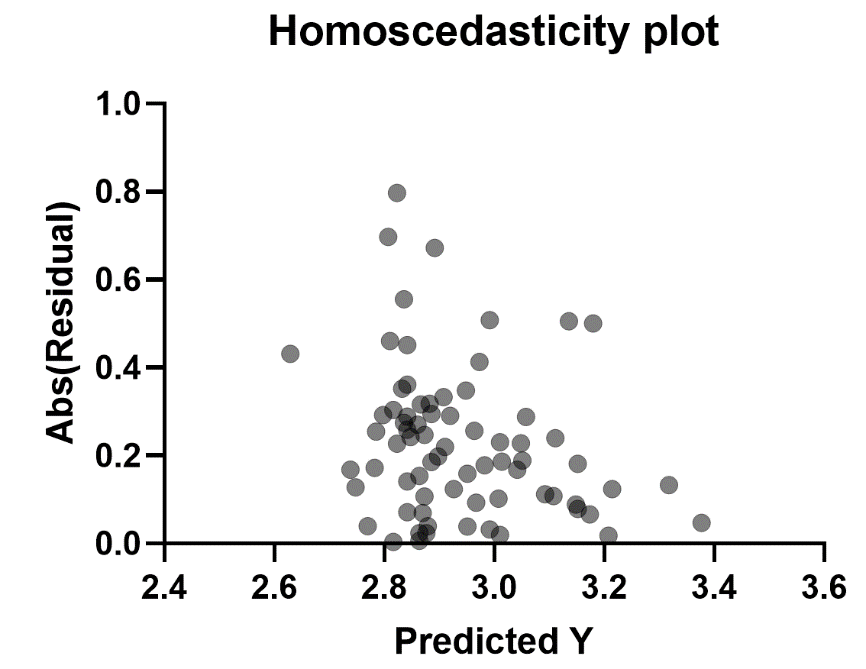
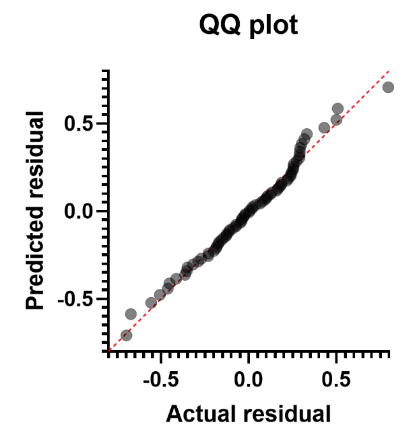
X has a linear relationship with Y **YES it appears so**

Independence **YES**

Residuals are normally distributed **YES**

The variance of Y at every value of X is the same **YES, although not perfect but OK**

Graphs they should create



Q13. Perform a simple linear regression on the log10 transformed data. Is there a negative or positive relationship between days of follow-up and LIF tissue levels? (1 pt)

Negative

Q14. Do you reject or fail to reject the null hypothesis? (1 pt)

Reject

Q15. Write the equation for the linear regression line. (1 pt)

Y= -0.3130\*X + 3.784

Extra credit.

EC 3. What would the actual (not Log10 transformed) LIF tissue level be at 222 days based on your equation in Q19? Limit decimal points to 2. (1 EC pt)

Start by taking the log10 of 222 so we can plug the appropriate number into the equation.

log10(222)=2.35

Y = -(0.31)(2.35) + 3.78 = 3.05

To transform back to a natural number Y = 103.05 = 1,122.02