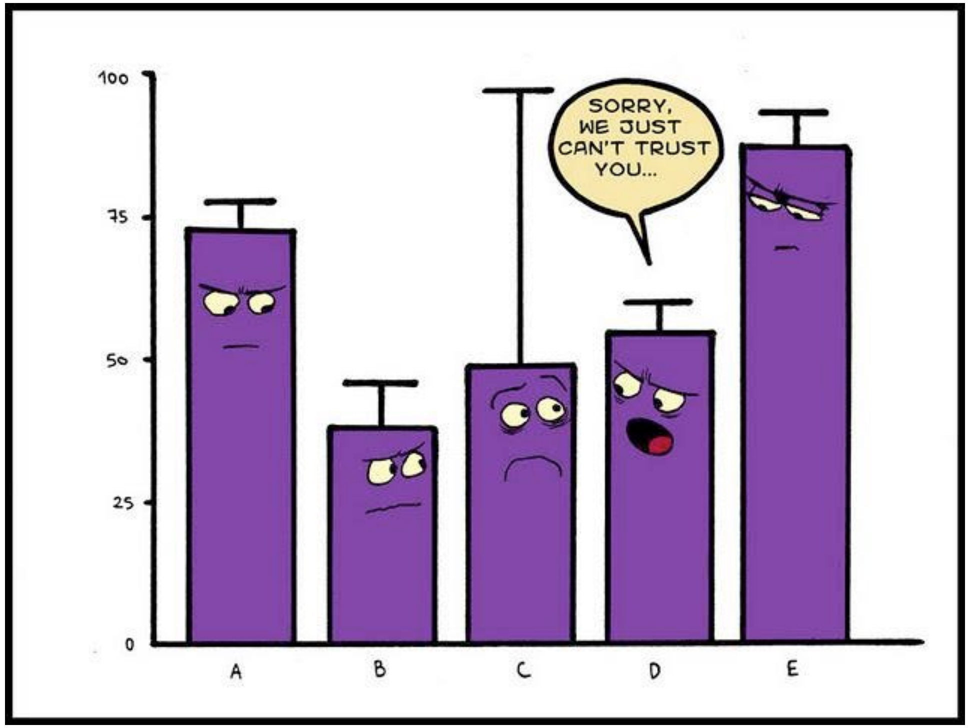
M1D8:
Examine experimental data using statistical methods

1. Prelab discussion
2. Apply statistical analyses to data sets
3. Start Data summary!



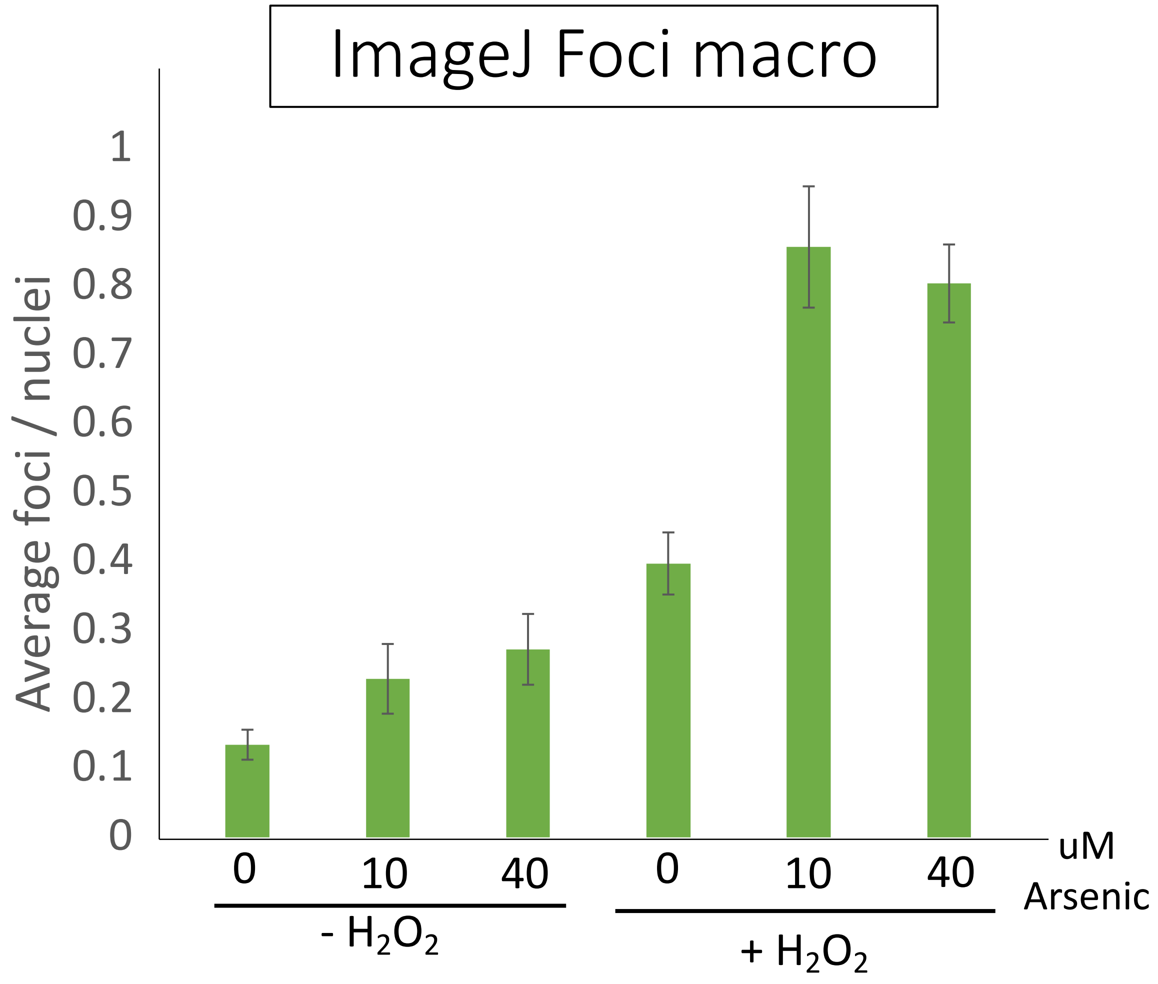
Why did we just learn about machine learning and Fisher's exact P value?

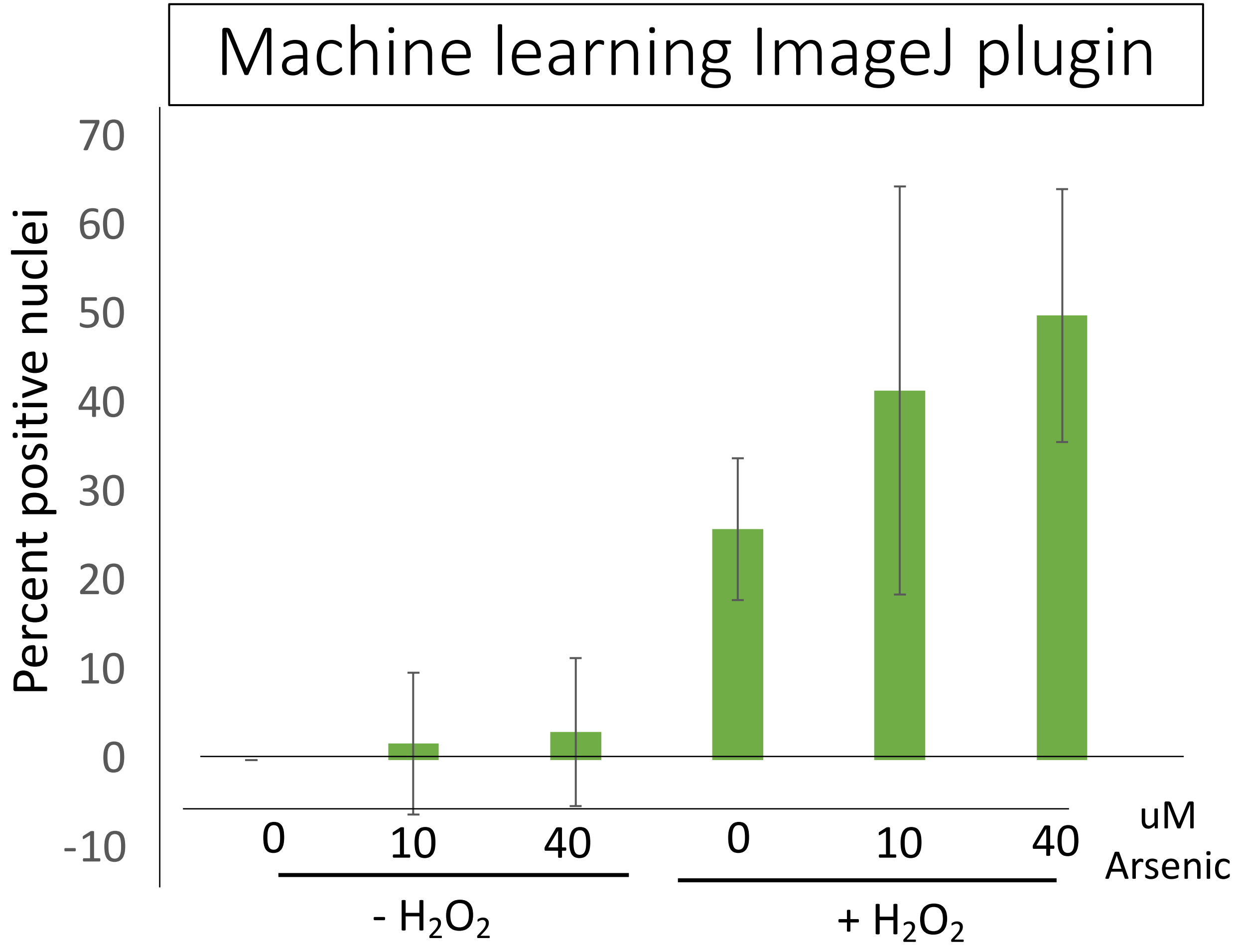
- Goal of Dushan's lecture: show you how you can expand a study using the same yH2AX output as you used

- As you generate larger amounts of data, you want an expanded analysis capable of:
  - Processing a large number of complex images
  - Identifying features of potentially complex data
  - Use statistical tests appropriate for the data

- How does this relate to the data you yH2AX images you analyzed?

Comparison between analysis methods





What do you notice when comparing different analyses of the same data set?

What method is best for analyzing / reporting your data?

- Best analysis type can depend on type of data generated

- Machine learning analysis is great for large data sets:
  - Screening bacteria subtypes in a large colony
  - Examining yH2AX positive cells in a multi-cell slice of tissue

- But what if you are looking for nuanced differences?

- What changes would we want to make in our experiment to make machine learning more useful?

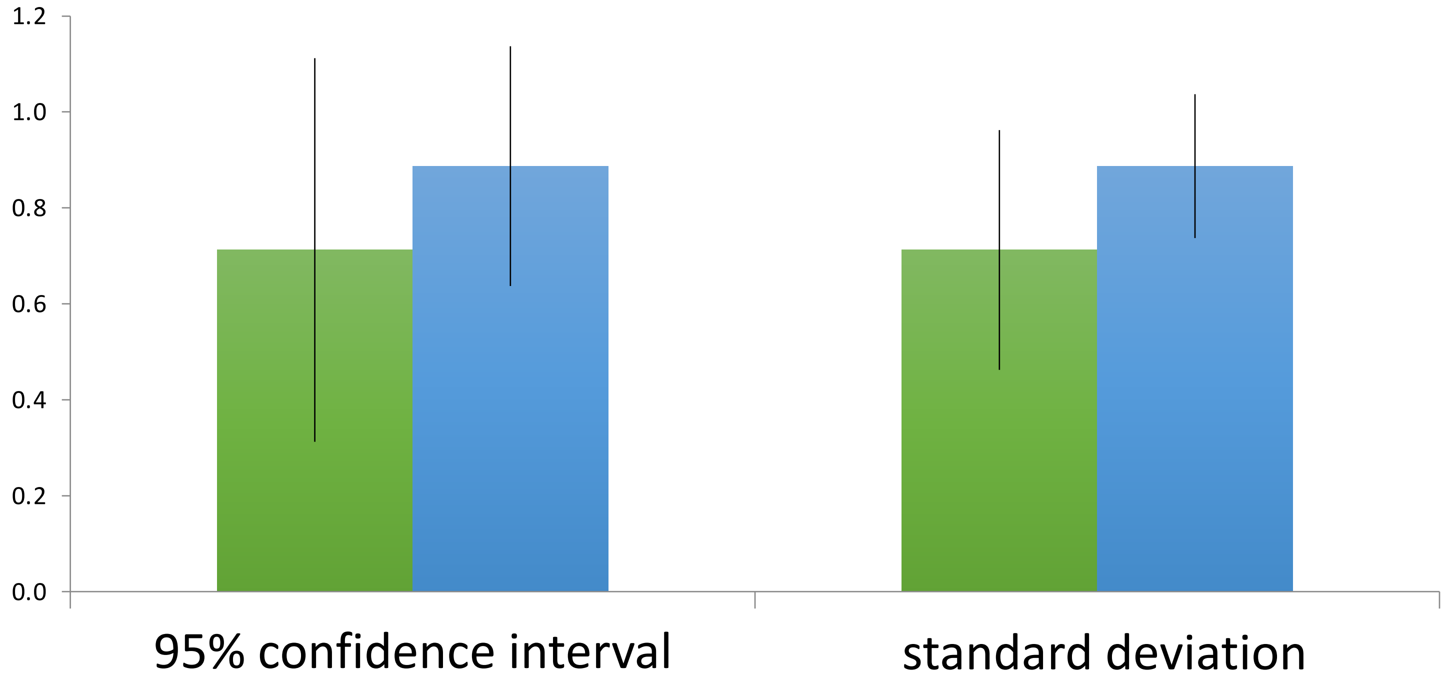
How will you evaluate and interpret your data?

1. Do you think there is noise (or variation) in your data?

2. Do your data support that there is a difference between the populations / treatments?

Confidence intervals show variance in data

- At 95% confidence interval, there is a 95% chance that the true mean is within the defined range
- Error bars used to represent variance

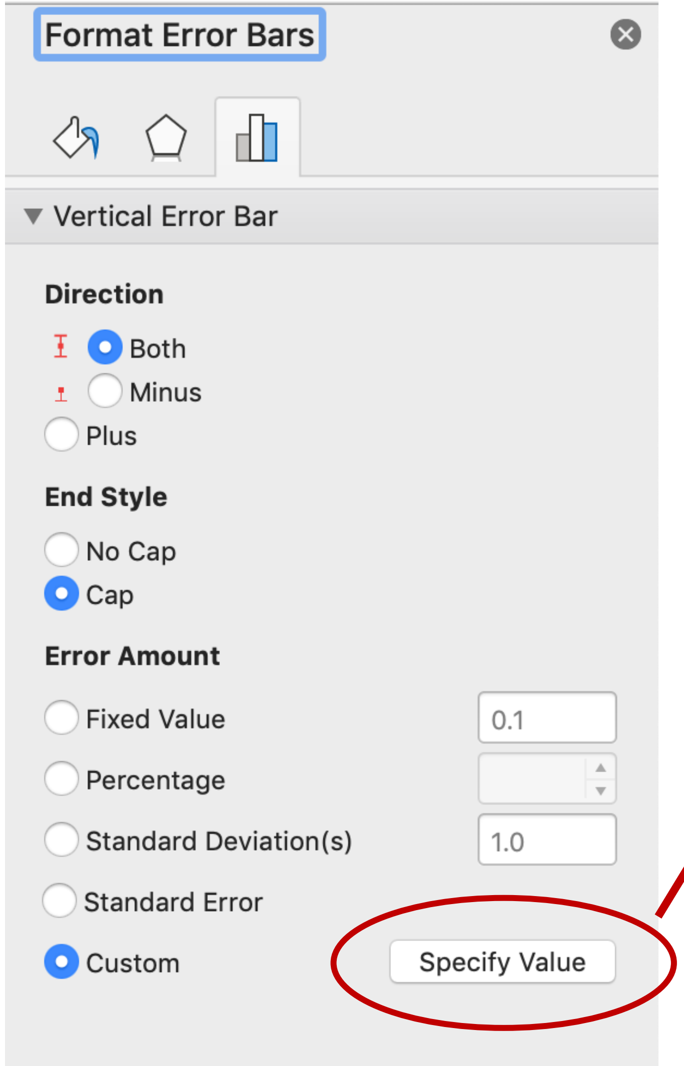


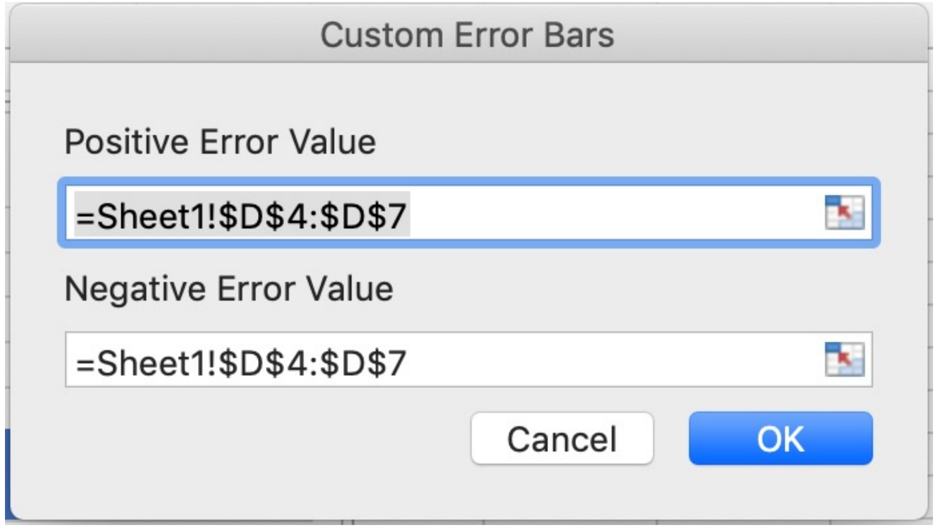
Calculating confidence interval in Excel

= CONFIDENCE(confidence level, standard dev., size)

- Confidence level:
- Standard deviation:
- Size:

How do you customize error bars in Excel?

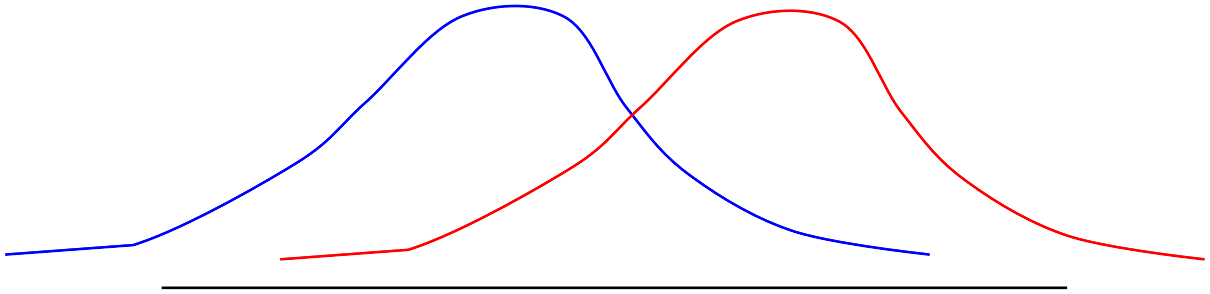


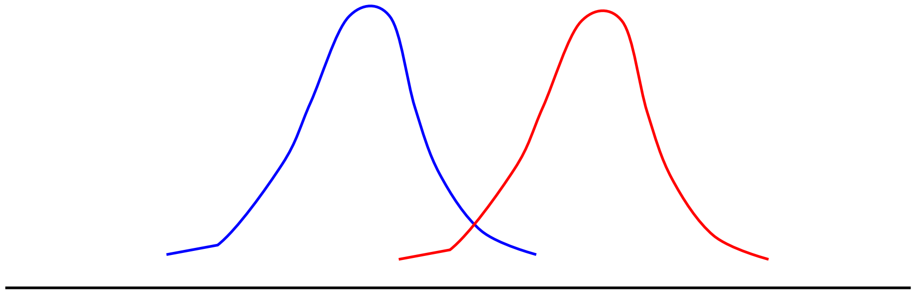


Enter value calculated for confidence level as custom error bars

Student's *t*-test determines if populations are significantly different

- Assume data follows *t*-distribution
- At $p < 0.05$, there is less than a 5% chance that populations are the same (95% chance that populations are different)
- Examines signal (means) : noise (variance) ratio


high noise


low noise

Calculating Student's *t* in Excel

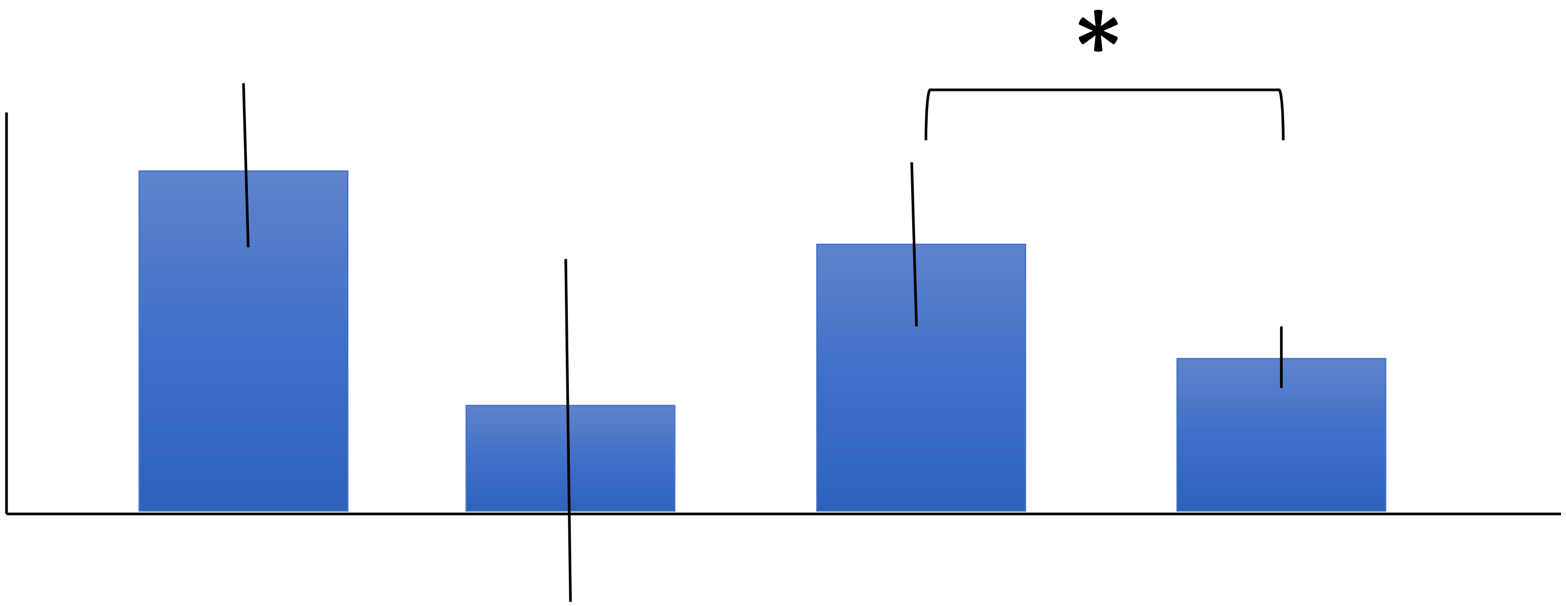
P = TTEST (array1,array2,2,3)

- Arrays:

- 2 = two-tailed test:

- 3 = population variances not assumed:

How will you use statistics in your data analysis?



- Student’s t-test can only be used to compare two populations
- What if data are not significant? Can data be almost significant?

For today…

- Apply statistics to evaluate your data
- **Use extra time to get a head start on your Data summary!**

For M2D1…

- Review Mod2 overview and M2D1 page