Heavy use of equations impedes communication among biologists

Proceedings of the National Academy of Sciences 2012 vol. 109

Tim W. Fawcett & Andrew D. Higginson

Abstract

Most research in biology is empirical, yet empirical studies rely fundamentally on theoretical work for generating testable predictions and interpreting observations. Despite this interdependence, many empirical studies build largely on other empirical studies with little direct reference to relevant theory, suggesting a failure of communication that may hinder scientific progress. To investigate the extent of this problem, we analyzed how the use of mathematical equations affects the scientific impact of studies in ecology and evolution. The density of equations in an article has a significant negative impact on citation rates, with papers receiving 28% fewer citations overall for each additional equation per page in the main text. Long, equation-dense papers tend to be more frequently cited by other theoretical papers, but this increase is outweighed by a sharp drop in citations from nontheoretical papers (35% fewer citations for each additional equation per page in the main text). In contrast, equations presented in an accompanying appendix do not lessen a paper's impact. Our analysis suggests possible strategies for enhancing the presentation of mathematical models to facilitate progress in disciplines that rely on the tight integration of theoretical and empirical work.

Quantitative Thinking in the Life Sciences

October 10rd – Linking probability, mathematical functions and data

Today

- Going over my homework
  - standard error and standard deviation revisited
  - t-test degrees of freedom
- Simple mathematical relationships and probability
- Assignment # 6
- More R fun!
  - Assignment 5 R code

Housekeeping

About that error terminology

- Sample standard deviation is a measure of the variability in your sampled population
- Standard error is an estimate of how well you measured a value (mean, intercept, slope, etc)

Standard deviation

- Sample standard deviation is a measure of the variability in your population

- True SD (uses n) is so frequently unknown that programs that calculate SD will assume that you are looking for the sample standard deviation (uses n - 1)

Number	Data, x	x - μ	(x - μ)^2
1	0.971	-1.175	1.381
2	0.348	-1.798	3.233
3	0.526	-1.620	2.624
4	4.014	1.868	3.489
5	4.871	2.725	7.426
Mean, μ	2.146		
Sum			18.153
Sum/(n-1)			4.538
Sqrt(sum/(n-1))			2.130
StDev (Excel)	2.130		

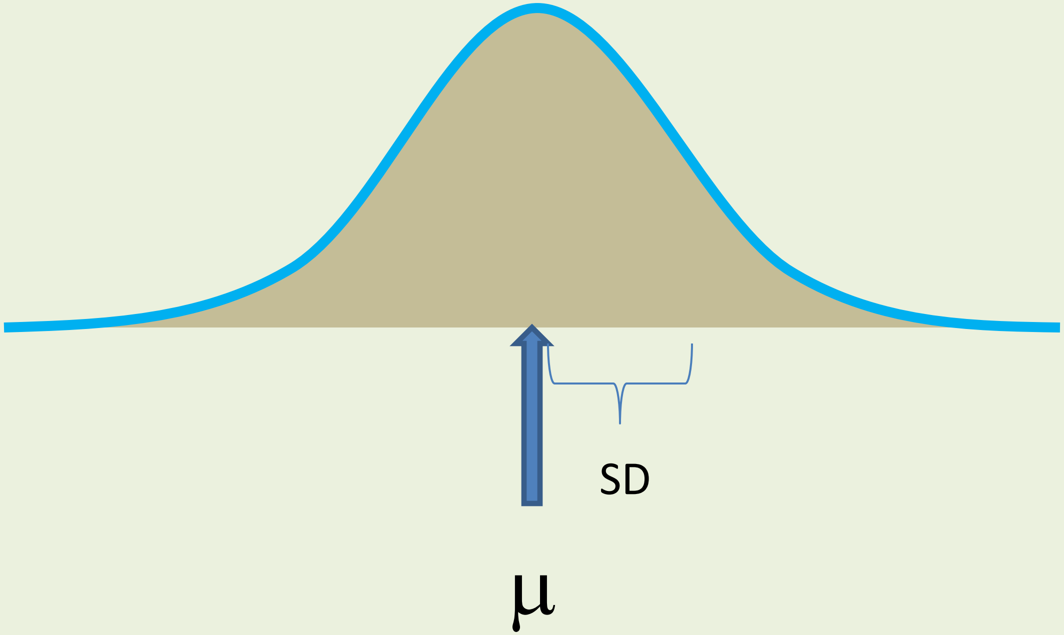
An intuitive explanation of the n-1 correction – sort of

You are given the following puzzle:

- This morning, the left hemisphere of Scott’s brain logically calculates that it is functioning like it is 25 years old
- The right hemisphere of Scott’s brain feels like it is X years old
- On average, Scott’s brain is acting like it is 50 years old
- How old does the right hemisphere of Scott’s brain feel?

If we know the mean, μ, we have n-1 independent samples. If we were given n-1 data values and the mean, we could calculate the last datum value.

Standard deviation



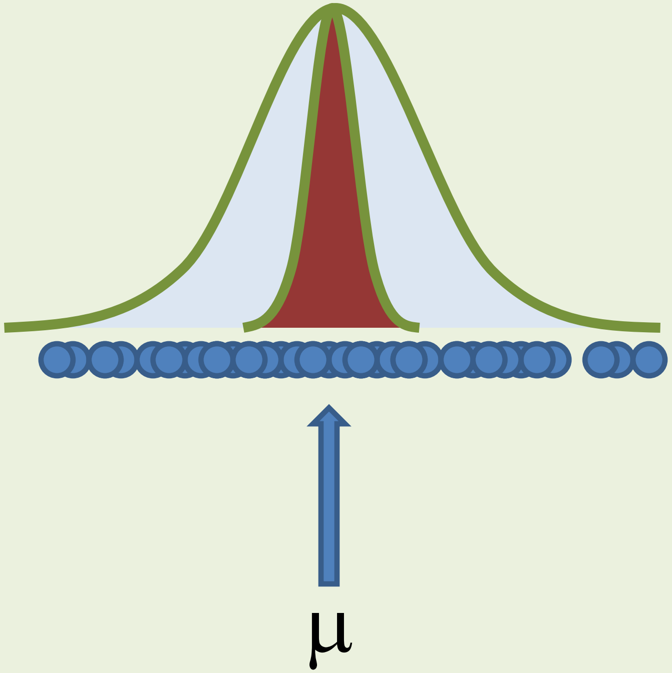
From wikipedia

- The **standard error of the mean** (SEM) is the standard deviation of the sample-mean's estimate of a population mean

Standard error:

How well do we measure μ

How confident are we in our measurement of μ



Terminology

- Standard error: how well is a value measured
- Standard deviation: how much variation was measured in the system

Probability and a couple simple statistical tests

- testing for differences between populations
- testing for a linear effect

Degrees of freedom

- DF are how much information is available for the statistical test to use to calculate the probability that the null is true

n ↑ leads to df ↑

model complexity ↑ leads to ↓ df

t-tests!

(Students t, Pairwise t-tests, Welch's t-test)

- t-tests are typically used for:
  - Test a null hypothesis that the means of two normally distributed populations are equal
  - Test that a population have a mean value (specified as your null hypothesis). Example – Terry used this for testing his card experiment null hypothesis
  - Paired or repeated measures test (collect data from something twice and see if the data differ). Example – test a population's weight, add three pre-class beers to population's diet, then test population's weight at the end of the semester.

Last week's t-test (Welch's)

Assumes variances between populations may be unequal.

Degrees of freedom are calculated as:

$$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1 - 1)} + \frac{s_2^4}{N_2^2(N_2 - 1)}}$$

where s_i^2 is the sample variance of population i

and N_i is the population size of i

Weird df numbers actually kind of cool!

$$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1 - 1)} + \frac{s_2^4}{N_2^2(N_2 - 1)}}$$

$$v = \frac{\left(\cancel{\frac{s_1^2}{N_1}} + \frac{s_2^2}{N_2}\right)^2}{\cancel{\frac{s_1^4}{N_1^2(N_1 - 1)}} + \frac{s_2^4}{N_2^2(N_2 - 1)}}$$

$$v = \frac{\left(\frac{s_2^2}{N_2}\right)^2}{\frac{s_2^4}{N_2^2(N_2 - 1)}}$$

$$v = \frac{\frac{s_2^4}{N_2^2}}{\frac{s_2^4}{N_2^2(N_2 - 1)}}$$

$$v = \frac{\frac{1}{1}}{\frac{1}{1 * (N_2 - 1)}}$$

$$v = \frac{1}{\frac{1}{(N_2 - 1)}}$$

$$v = (N_2 - 1)$$

$$v = 9$$

- t-tests are typically used for:
  - Test a null hypothesis that the means of two normally distributed populations are equal
  - **Test that a population have a mean value (specified as your null hypothesis).** Example – Terry used this for testing his card experiment null hypothesis
  - Paired or repeated measures test (collect data from something twice and see if the data differ). Example – test a population's weight, add three pre-class beers to population's diet, then test population's weight at the end of the semester.

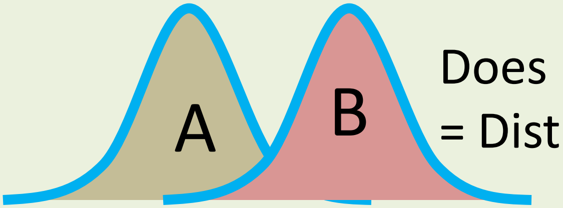
```
> t.test(native2,mu =50)
# cultivated2 all equal to 50

One Sample t-test
data:  native2
t = 3, df = 9, p-value = 0.01496
alternative hypothesis: true mean
is not equal to 50
```

```
Welch Two Sample t-test
data:  native2 and cultivated2
t = 3, df = 9, p-value = 0.01496
alternative hypothesis: true difference in
means is not equal to 0
```

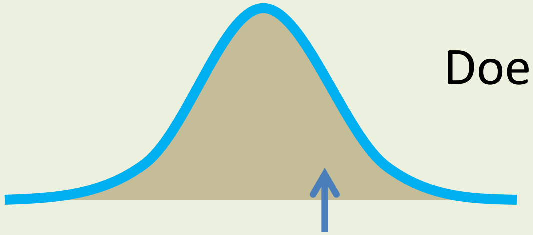
t-test will allow us to test

Test a null hypothesis that the means of two normally distributed populations are equal



Does Distribution A = Distribution B?

Test that a population have a mean value (specified as your null hypothesis).



Does $\mu = 5$?

Paired or repeated measures test (collect data from something twice and see if the data differ).

Miticide Trial Data

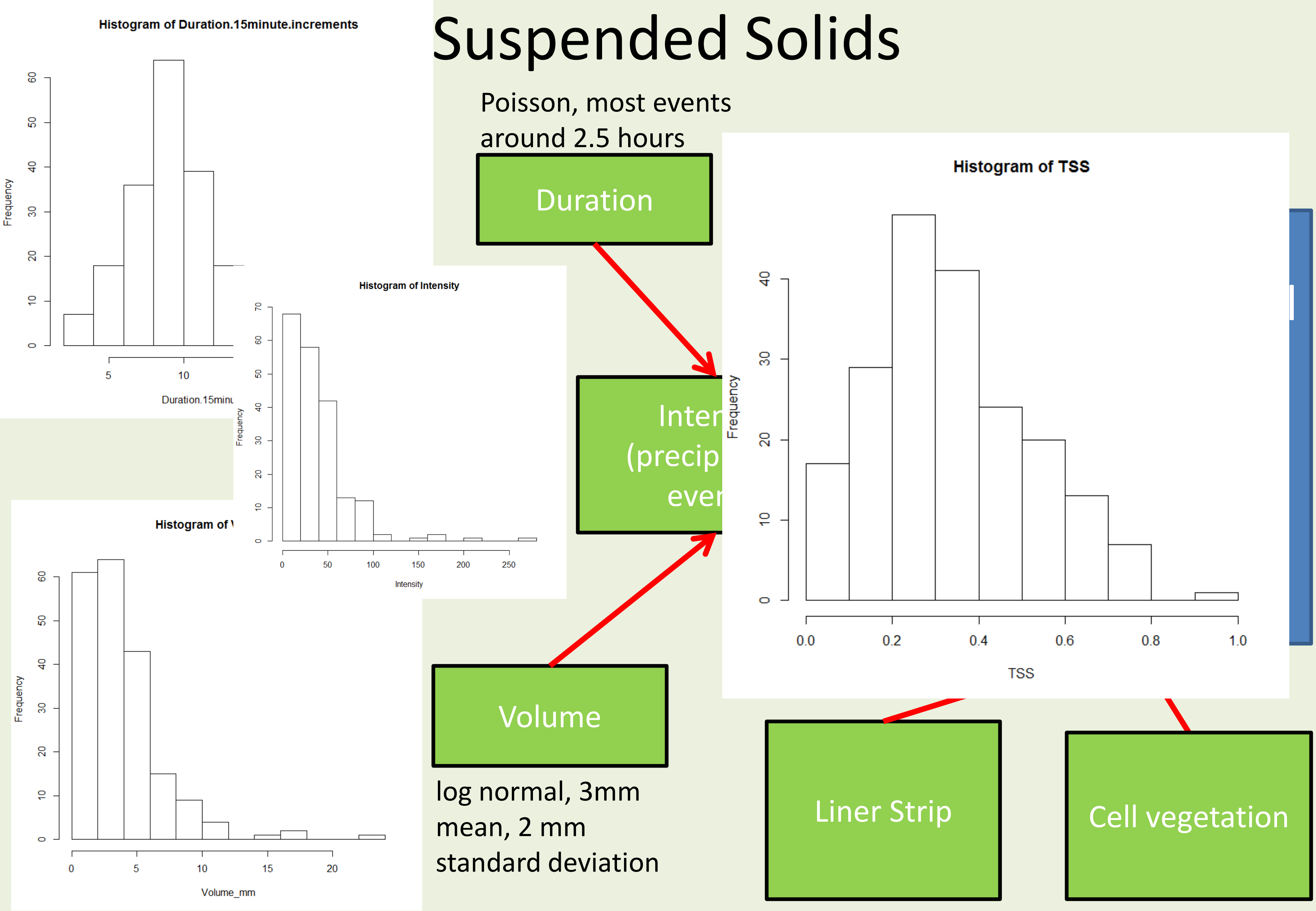
Mites/Plant	Before	After
Corn plot 1	0.500	22.967
Corn plot 2	10.657	29.364
Corn plot 3	43.469	15.972
Corn plot 4	7.045	7.683
Corn plot 5	9.626	10.089
Corn plot 6	18.534	14.059
Corn plot 7	34.237	23.093
Corn plot 8	38.291	28.351
Corn plot 9	11.959	4.898
Corn plot 10	1.582	13.964

Does Treatment A change the population?

Assignment # 6

- Assignment # 6 is due on Oct 17th
- Worth 50 points
- Part 1: Simulation
  - Using the provided functions for distributions, take a first pass at simulating data for each of your components where you will be taking data. Assume that data will be measured perfectly (no measurement error).
  - Write up in manuscript form for a few of the components. That is, introduce the system (you can self-plagiarize but make it clean), describe how you will sample (or already sampled) components (Methods section), describe your simulation inputs, include output plots. Discuss in brief.
- Part 2: Chapter 7 R code found on my website
  - Distribution exercises and examples for use in simulation will be in this chapter

Testing a Bioretention systems: Total Suspended Solids



Poisson, most events around 2.5 hours

Duration

Inter
(precip
ever

Volume

log normal, 3mm mean, 2 mm standard deviation

Liner Strip

Cell vegetation

Chapters 4-6 R code review

- To R we go!
  - Dropbox\\Quantitative Thinking\\Oct 10 notes_assignment 5 r code.R