THE STANDARD ERROR OF THE LAB SCIENTIST

. . . and other common statistical misconceptions in the scientific literature.

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Some Quotes

Instead of an outline, here are some quotes from scientific publications that we will have a closer look at:

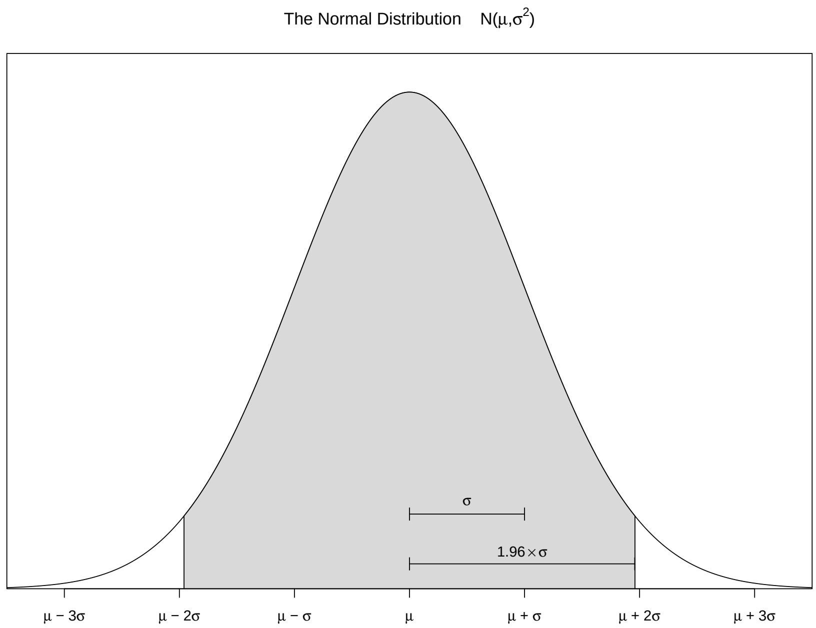
“ . . . 95% confidence intervals (mean plus minus two standard deviations) . . . ”

“ . . . the model predicted the data well (correlation coefficient R^2 = 0.85) . . . ”

“ . . . we used the jackknife to estimate the error for future predictions . . . ”

Quote #1

" . . . 95% confidence intervals (mean plus minus two standard deviations) . . . "



Parameters and Statistics

A statistic is a numerical quantity derived from a sample to estimate an unknown parameter that describes some feature of the entire population.

For example, assume that the measurements taken in an experiment follow a normal distribution $X \sim N(\mu, \sigma^2)$, and assume that we carry out n independent experiments, i. e. let $X_1, \ldots, X_n$ be a random sample from X.

$\longrightarrow$ μ is the unknown population mean (a parameter).

$\bar{X} = \sum_i X_i/n$ is the sample mean (a statistic).

$\longrightarrow$ σ is the standard deviation of X.

$\hat{\sigma} = \sqrt{S^2/(n-1)}$ is the sample standard deviation, where $S^2 = \sum_i (X_i - \bar{X})^2$.

Note that $\hat{\sigma}$ is not a standard error!

The Standard Error

The standard error of a statistic is the standard deviation of its sampling distribution.

For example: $X \sim N(\mu, \sigma^2)$, hence $\bar{X} \sim N(\mu, \sigma^2/n)$, and therefore the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$.

A standard error itself is a parameter, not a statistic!

As the standard deviation of X is often unknown, so is the standard error of $\bar{X}$, but in practice we can estimate it, for example by $\widehat{\text{se}}(\bar{X}) = \hat{\sigma}/\sqrt{n}$.

In general, the standard error depends on the sample size: the larger the sample size, the smaller the standard error.

This means that the term *standard deviation* in " ... 95% confidence intervals (mean plus minus two standard deviations) ... " better be referring to the sampling distribution, not the population.

But what about that factor 2?

Confidence Intervals

If the standard deviation σ of X is known, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

We can obtain a 95% confidence interval for the population mean μ as

$$I \;=\; [\bar{X} - z_{0.975} \times \sigma/\sqrt{n}\;;\;\bar{X} + z_{0.975} \times \sigma/\sqrt{n}]$$

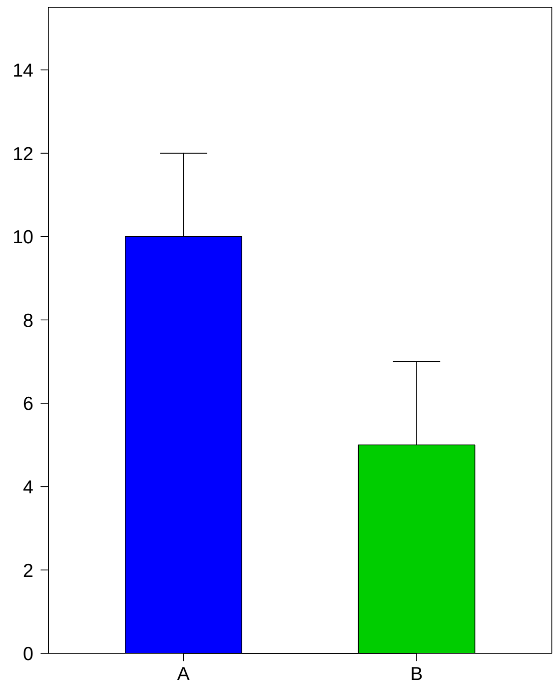
If σ is unknown and we have to estimate it from the data as well, then $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$.

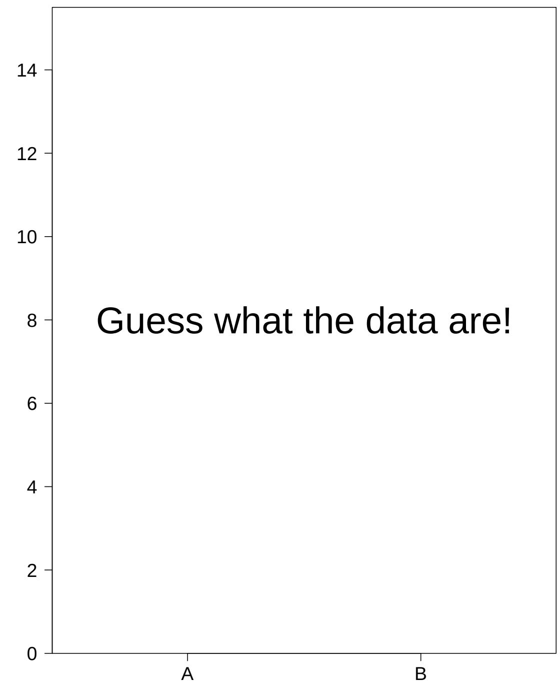
The 95% confidence interval for μ is now

$$I \;=\; [\bar{X} - t^{n-1}_{0.975} \times \hat{\sigma}/\sqrt{n}\;;\;\bar{X} + t^{n-1}_{0.975} \times \hat{\sigma}/\sqrt{n}]$$

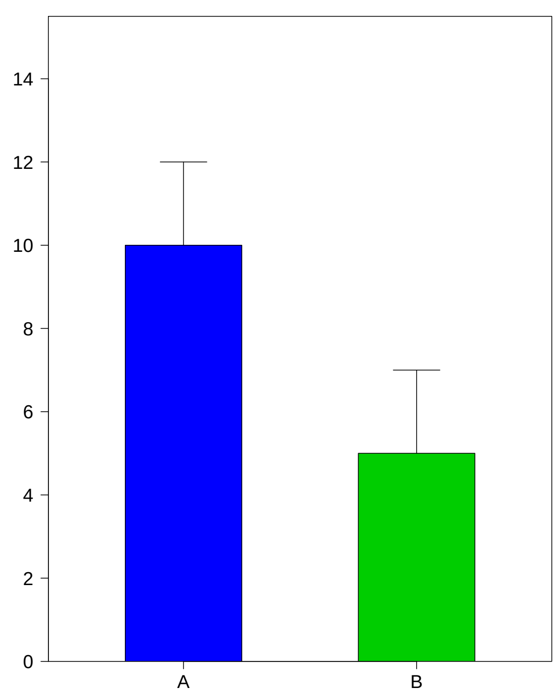
n	3	4	5	6	7	8	9	10
$t^{n-1}_{0.975}$	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26

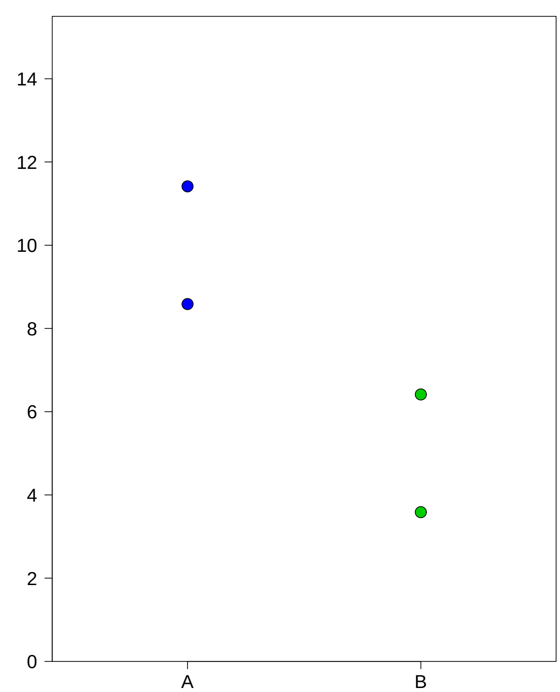
Plotting Data (Confusion Part 1)



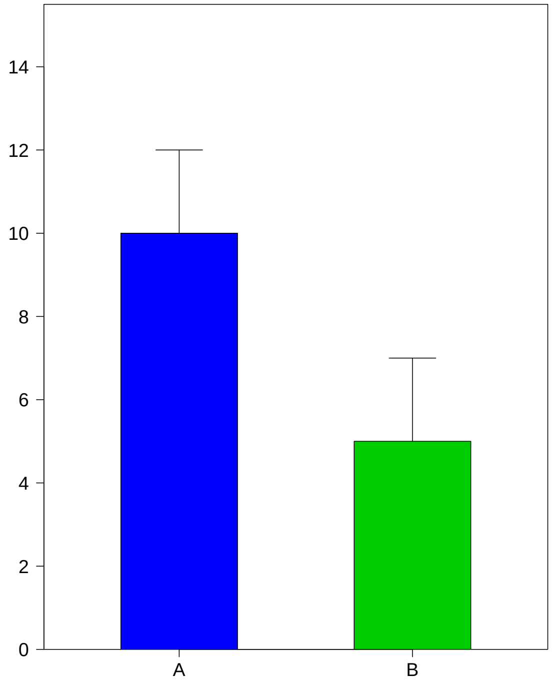


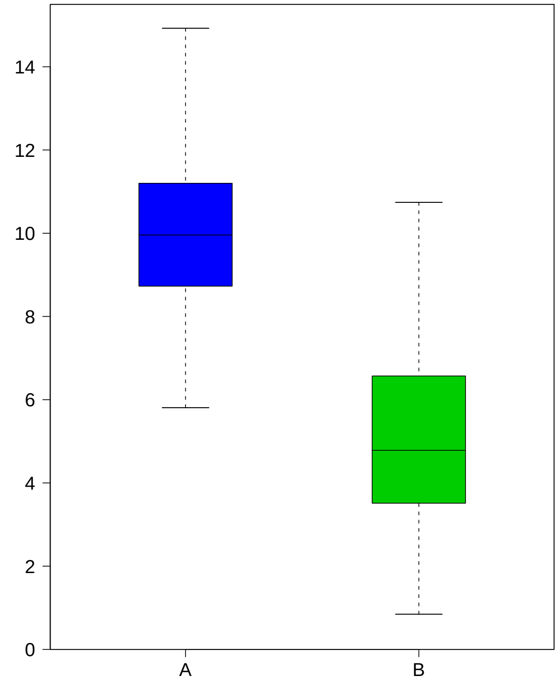
Plotting Data (Confusion Part 1)



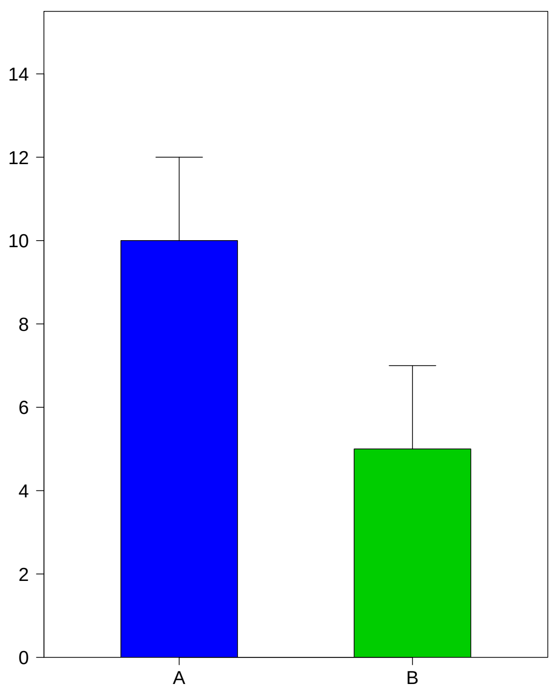


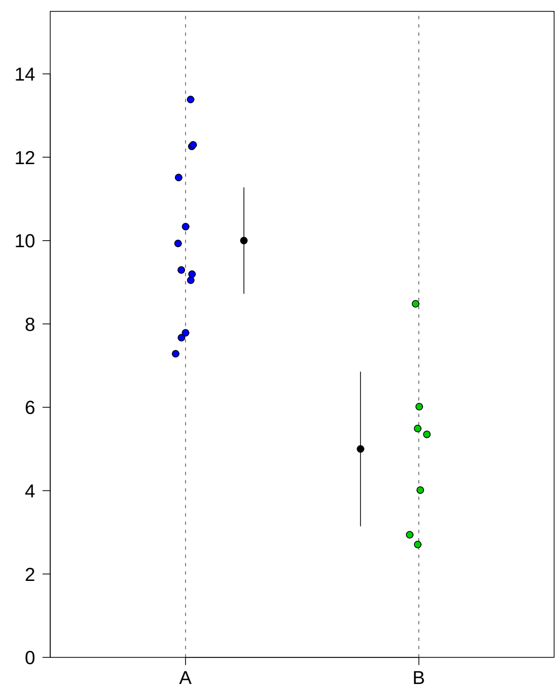
Plotting Data (Confusion Part 1)





Plotting Data (Confusion Part 1)





Reporting Uncertainty (Confusion Part 2)

- Results are frequently reported in the form 'mean plus minus standard error', such as 7.4 ($\pm$1.3).
- What is reported as the (estimated) standard error is often the sample standard deviation ($\hat{\sigma}$, not $\hat{\sigma}/\sqrt{n}$).
- The plus/minus notation can also mislead readers to believe 7.4 ($\pm$1.3) is a confidence interval.
- To allow others to correctly quantify uncertainty, it is also necessary to report the number of experiments that have been performed (for the t-quantile and to calculate an estimate for the standard error, if necessary).

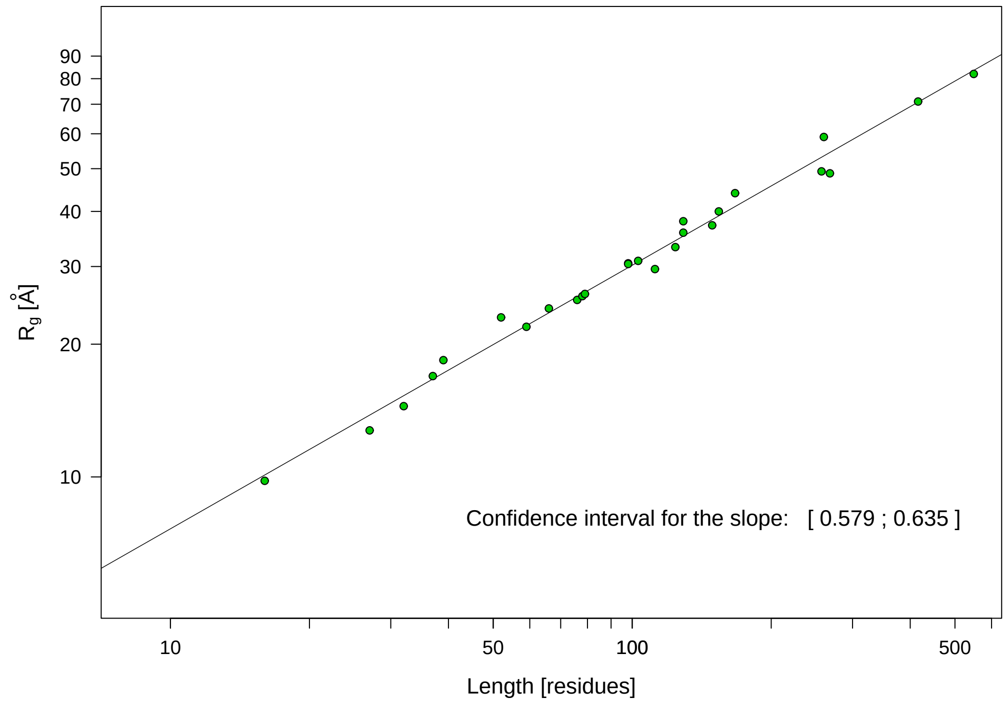
An Example

Do chemically denatured proteins behave as random coils?

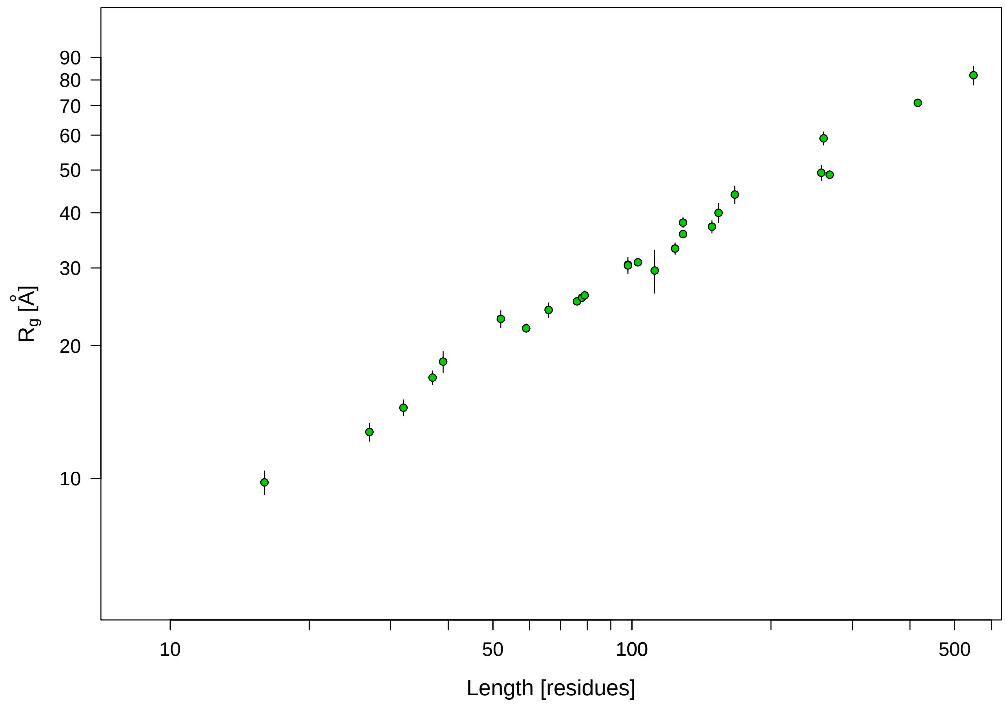
- The radius of gyration R_g of a protein is defined as the root mean square distance from each atom of the protein to their centroid.
- For an ideal (infinitely thin) random-coil chain in a solvent, the average radius of gyration of a random coil is a simple function of its length n: $R_g \propto n^{0.5}$.
- For an excluded volume polymer (a polymer with non-zero thickness and non-trivial interactions between monomers) in a solvent, the average radius of gyration, we have $R_g \propto n^{0.588}$ (Flory 1953).

$\longrightarrow$ The radius of gyration can be measured using small angle x-ray scattering.

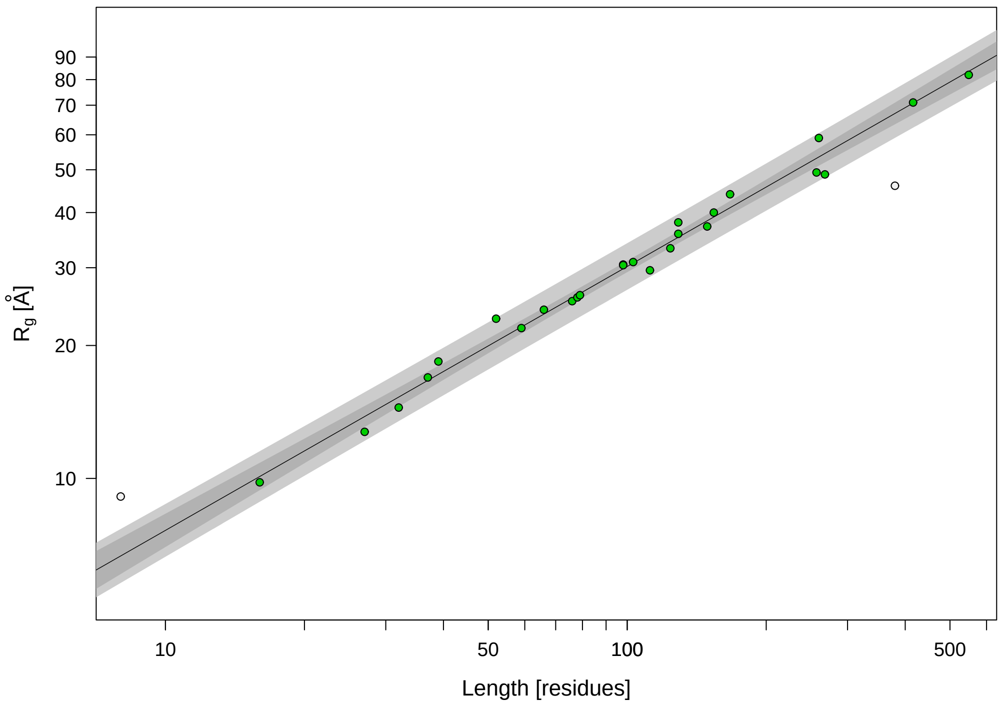
An Example



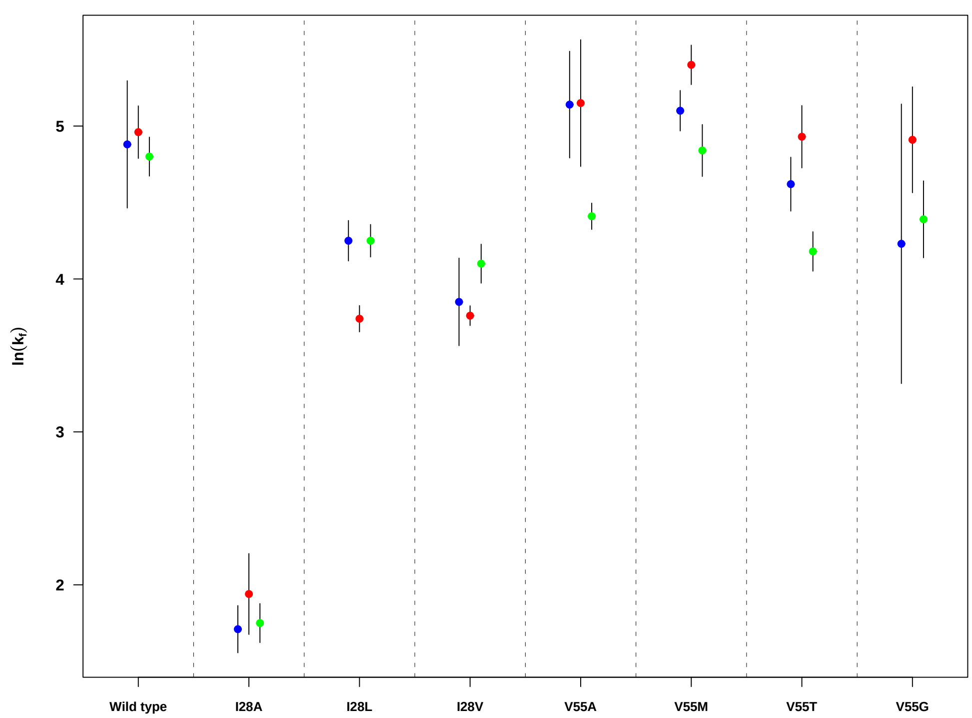
An Example



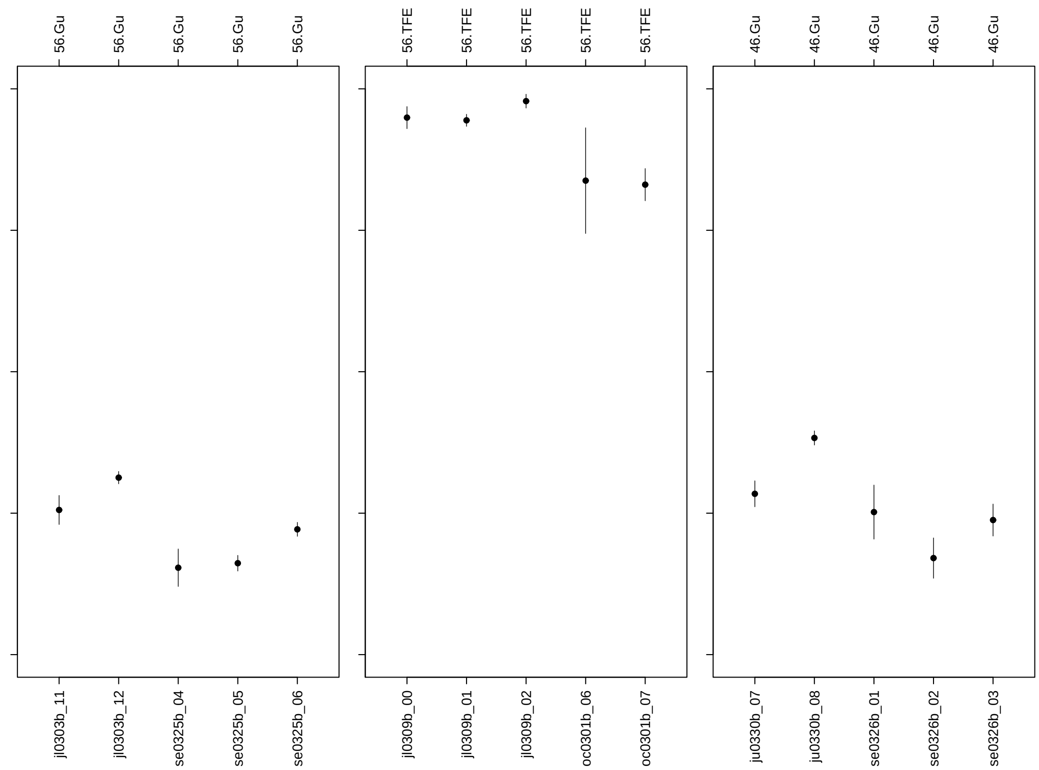
Variability



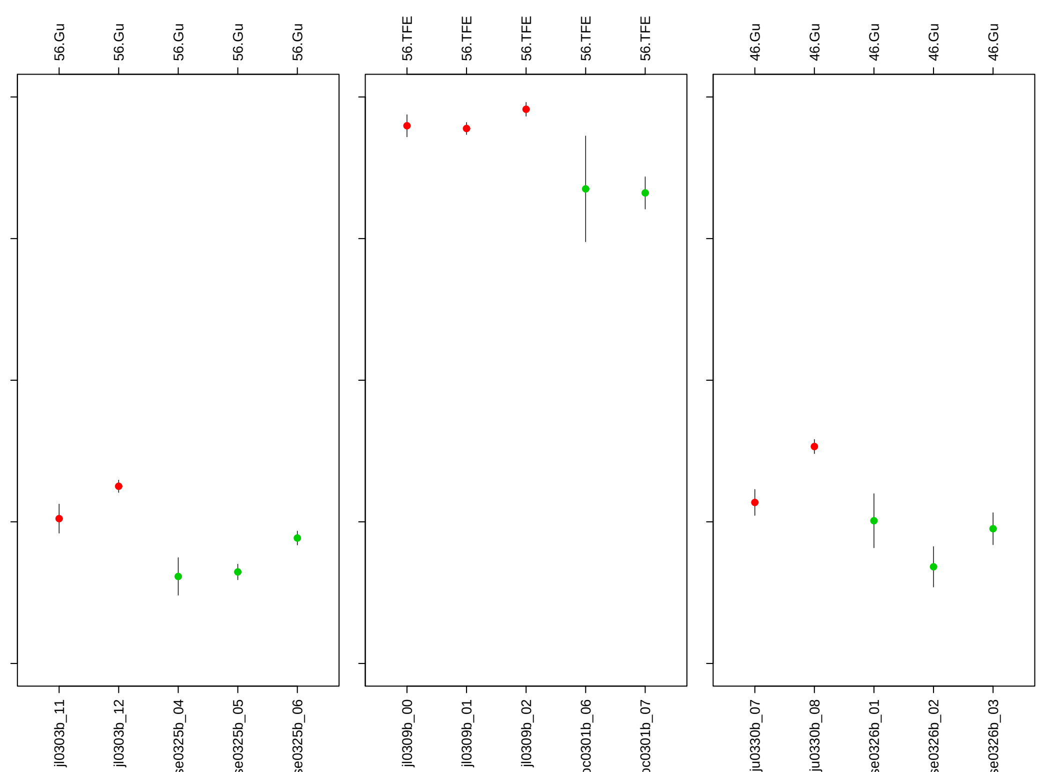
Variance Components



Variance Components



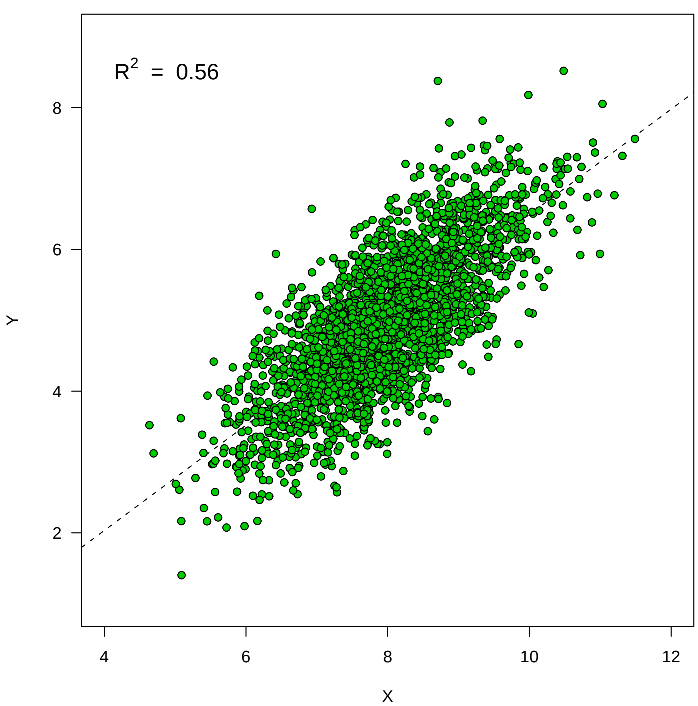
Variance Components



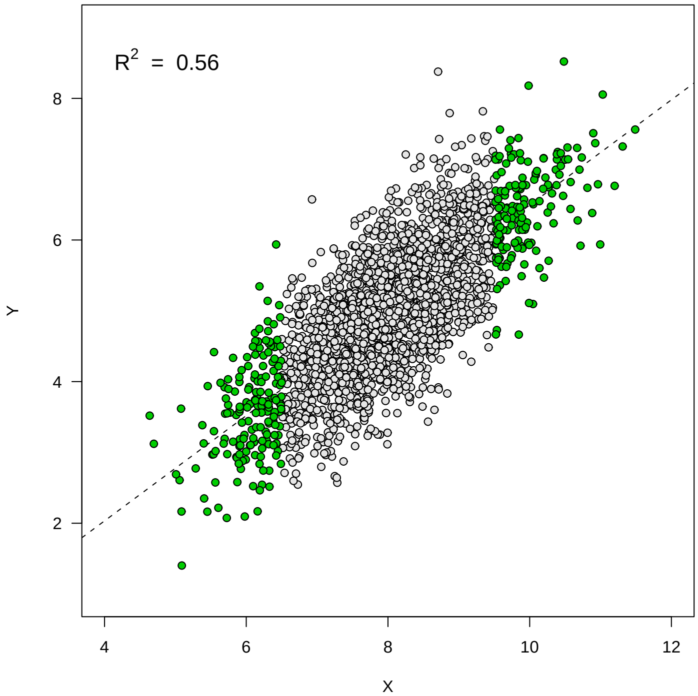
Quote #2

" . . . the model predicted the data well (correlation coefficient $R^2 = 0.85$) . . . "

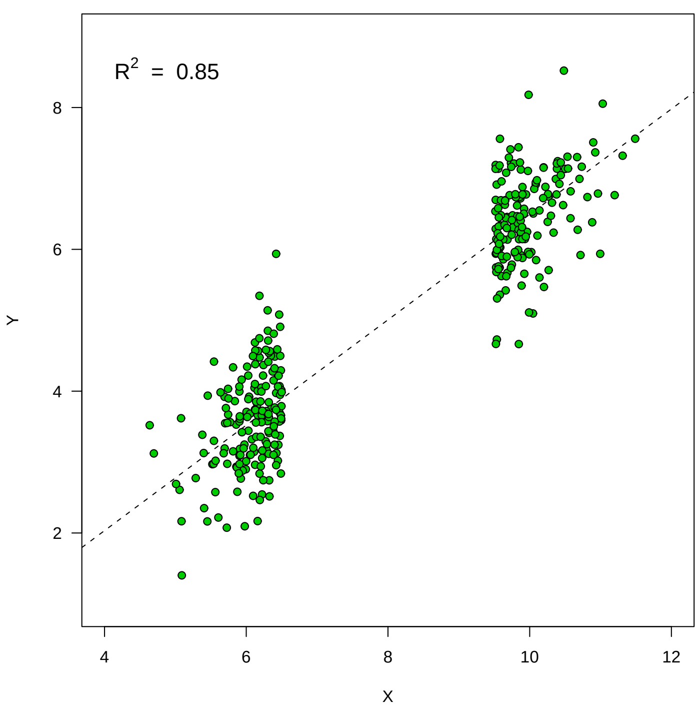
Correlation



Correlation



Correlation



Correlation vs Regression

- In a correlation setting we try to determine whether two random variables vary together (covary).
- There is no ordering between those variables, and we do not try to explain one of the variables as a function of the other.

- In regression settings we describe the dependence of one variable on the other variable.
- There is an ordering of the variables, often called the dependent variable and the independent variable.

Correlation vs Regression

The correlation coefficient of two jointly distributed random variables X and Y is defined as

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where cov(X,Y) is the covariance between X and Y, and σ_X and σ_Y are their respective standard deviations.

If X and Y follow a bivariate normal distribution with correlation ρ

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right)$$

then

$$y_i | x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

where $\beta_0 = \mu_Y - \beta_1\mu_X,\ \beta_1 = \rho\sigma_Y/\sigma_X,$ and $\sigma^2 = \sigma_Y^2(1-\rho^2)$.

Some Comments

- The sample (multiple) correlation coefficient in a regression setting is defined as the correlation between the observed values Y and the fitted values $\hat{Y}$ from the regression model: R = cor($Y, \hat{Y}$)

- R^2 is called the coefficient of determination: it is equal to the proportion of the variability in Y explained by the regression model.

- The notion "the higher R^2, the better the model" is simply wrong.

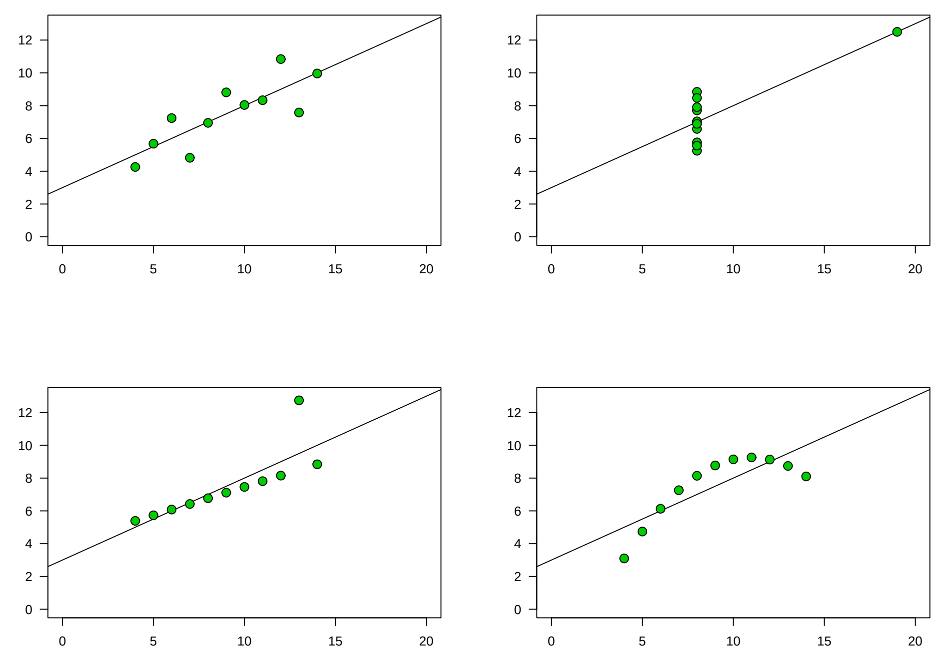
- Assuming we have an intercept in the (linear regression) model, the more predictors we include in the model, the higher R^2.

- However, there is a test for "significant" reductions in R^2 (there is a one-to-one correspondence to the usual t and F statistics).

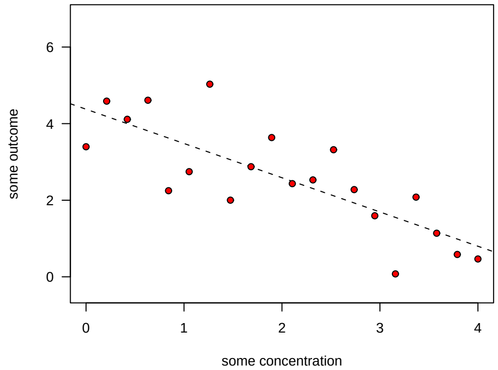
- R^2 tells us nothing about model violations.

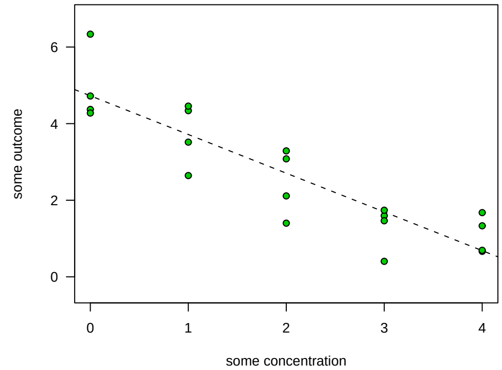
Model Fit

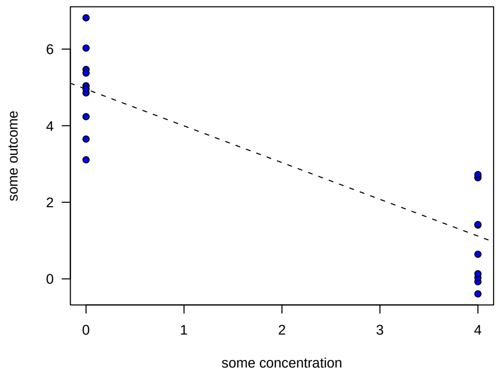
$\hat{\beta}_0 = 3.0$, $\hat{\beta}_1 = 0.5$, p-value (slope) = 0.002, $R^2 = 0.67$, RSE = 1.24 (9 df).



Experimental Design







Standard error ratios
for the slope:

1.65 : 1.41 : 1.00

In Conclusion: A Few Suggestions

- Take Karl Broman's course "Statistics for Laboratory Scientists" (140.615/616).
- For your analysis, use tools that help you understand the data, and try to get an idea what all that statistical output from your program means.
- Avoid "black boxes" as much as possible. Plot the data.
- For more complicated quantitative projects, adopt a biostatistician.
- Keep recruiting people like Ray and Matthew. Cheers!