Bootstrapping: resampling with replacement.

Original *n* data points	{ 1	2	3	4	5	6	…	50 }
Sample 1	{ 1	4	6	9	10	11	…	50 }
Sample 2	{ 1	2	3	4	4	5	…	49 }
Sample 3	{ 1	1	3	4	4	8	…	50 }
…							…	
Sample 1,000	{ 1	2	3	3	3	3	…	49 }

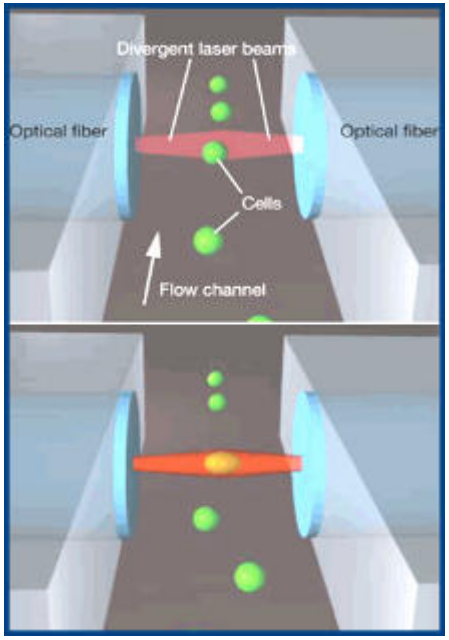
Each "bootstrapped sample" is the same size as the original set.

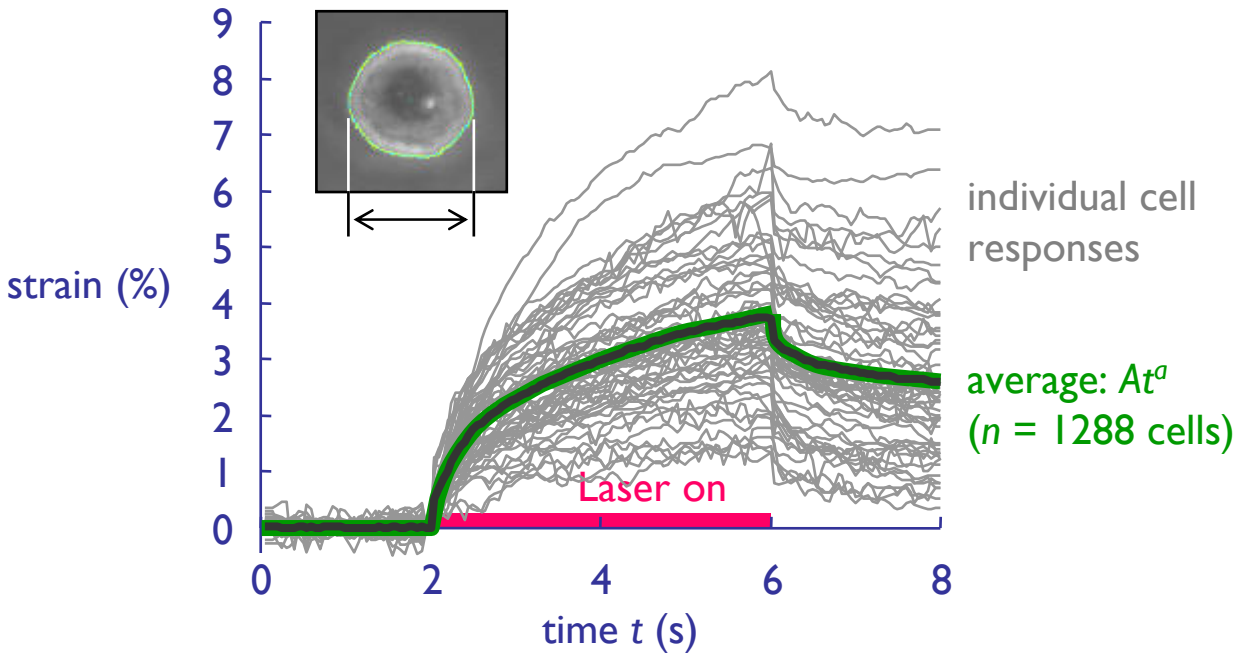
The principle of bootstrapping is that the population looks like the sample—but much larger.

Bootstrapping: resampling with replacement.

- A Monte Carlo technique for use when parametric approaches are not possible (such as when underlying distributions are unknown, or closed-form error terms aren't available for your application).
- Example: you use your entire data set to estimate a parameter; now what is the error of this parameter?
- More generally, how certain are we of the parameters we estimate and the conclusions we draw?
- With unlimited resources, we would repeat our studies until the underlying distribution and the variation of every parameter is perfectly understood.
- The principle of bootstrapping is that, with limited resources, the best estimate of future data is a resampled set from existing data.

Bootstrapping example 1: estimating the error in a fitted parameter.





Background: Living cells can be deformed by laser pressure, producing an average power-law rheological response (strain = At^a) that identifies the cell as intermediate between an elastic solid (a = 0) and a viscous fluid (a = 1).

Limitation: The best fit is a = 0.34, but with what error? 0.1? 0.0000001? (If another population exhibited a = 0.36, for example, could we conclude that the two groups are significantly different?)

Adapted from Guck et al. *Biophys J* 88(5) (2005), Maloney and Van Vliet et al., *Biophys J* 100(1) (2010)

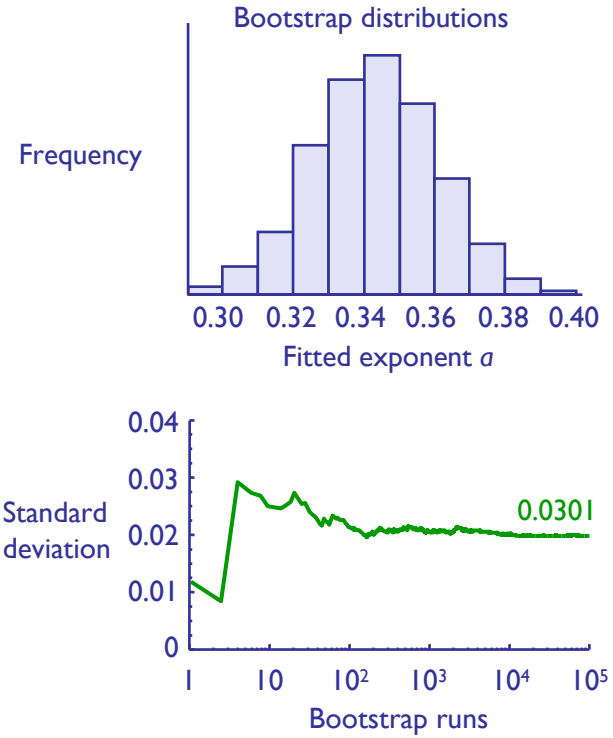
Bootstrapping example 1: estimating the error in a fitted parameter.

		Fitted exponent
Original *n* cells	{ 1 2 3 4 5 6 … 1288 }	0.341
Sample 1	{ 1 4 6 9 10 11 … 1288 }	0.388
Sample 2	{ 1 2 3 4 4 5 … 1288 }	0.382
Sample 3	{ 1 1 3 4 4 8 … 1288 }	0.334
…	…	…
Sample 1,000	{ 1 2 3 3 3 3 … 1287 }	0.358

Average of *N* bootstrapped outputs: 0.344

SD of outputs: 0.0301

Estimated SE of fitted exponent (SD x $\sqrt{n/(n\text{-}1)}$): 0.0301



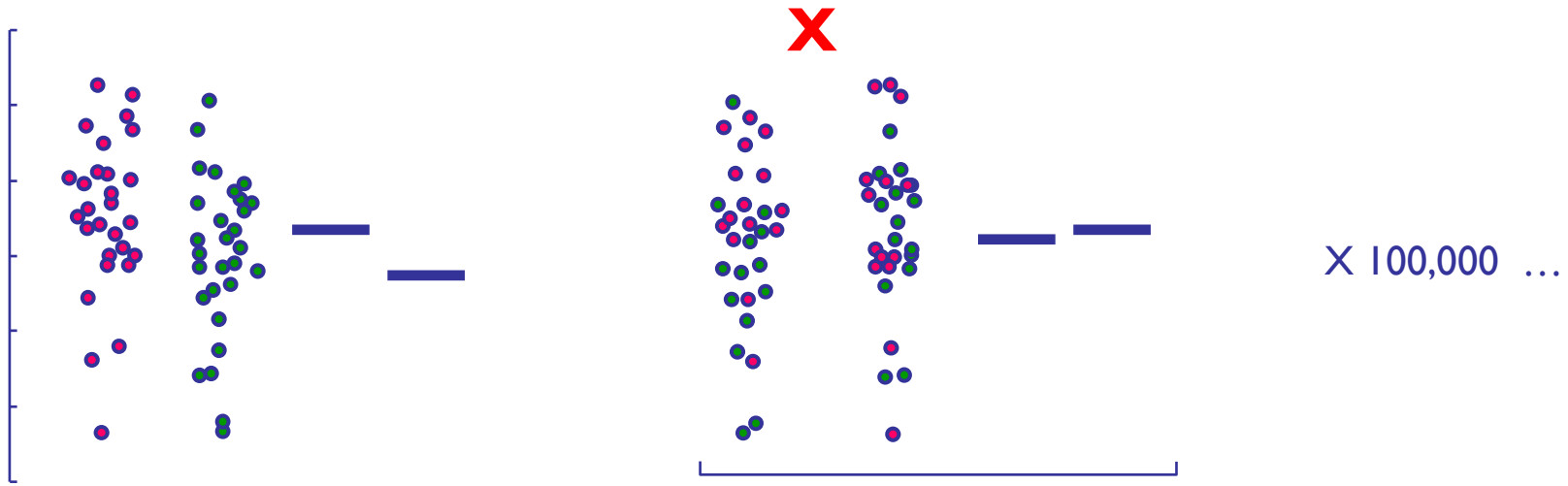
Bootstrapping solution: sample with replacement from the original data set (some values are repeated, some omitted).

Recalculate our parameter of interest and repeat thousands of times; the output changes slightly with each run.

The standard deviation of the collection of bootstrapped outputs, multiplied by $\sqrt{[n/(n\text{-}1)]}$, is an estimator of the true standard error of the true output.

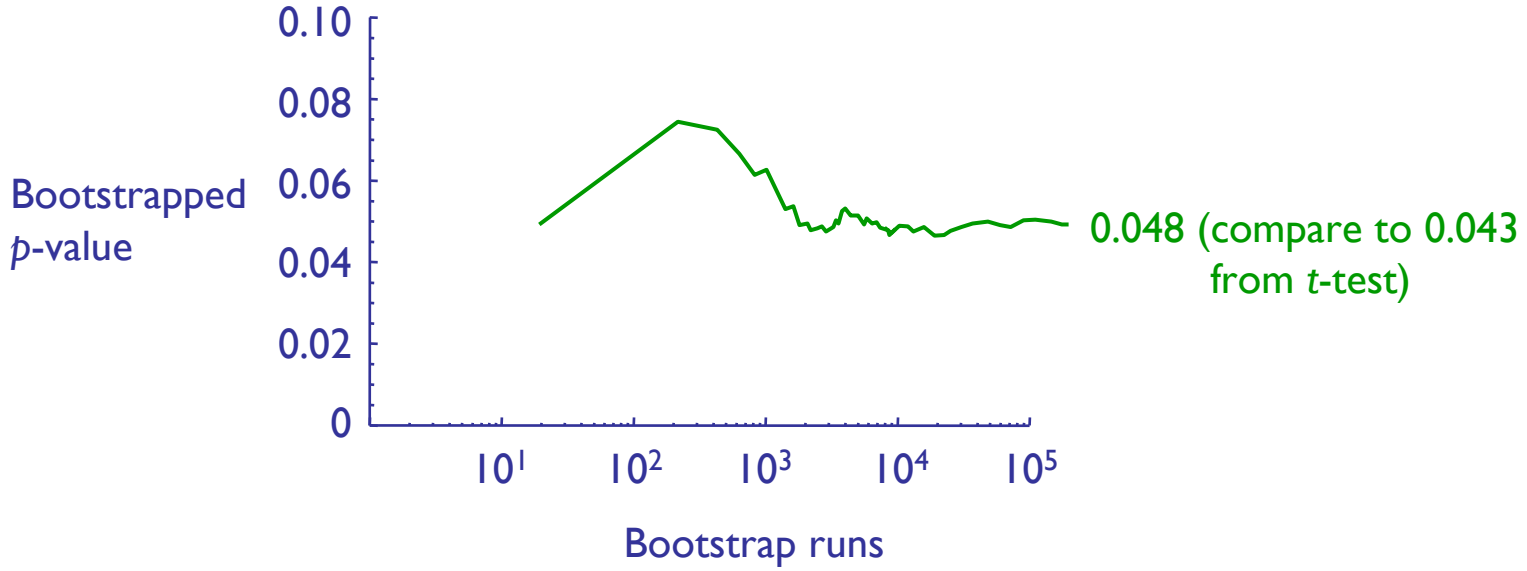
Adapted from Maloney and Van Vliet et al., *Biophys J* 100(1) (2010)

Bootstrapping example 2: hypothesis testing.

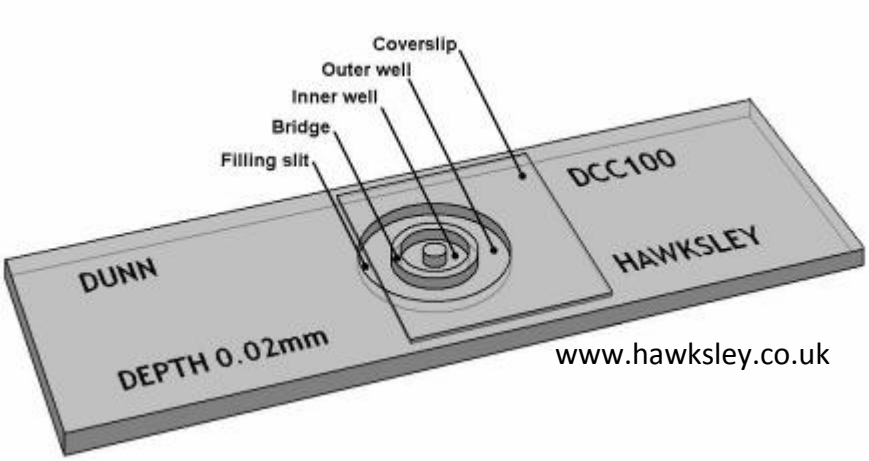


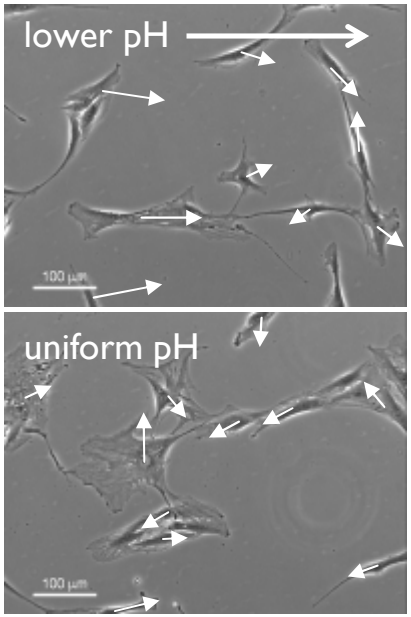
Original data and means

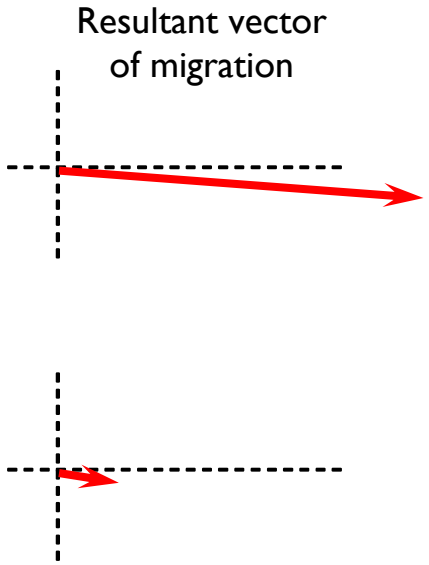
New groups are resampled from all the original data merged together: how typical is it for the means to be as different as they were?



Bootstrapping example 3: hypothesis testing with unknown distribution.





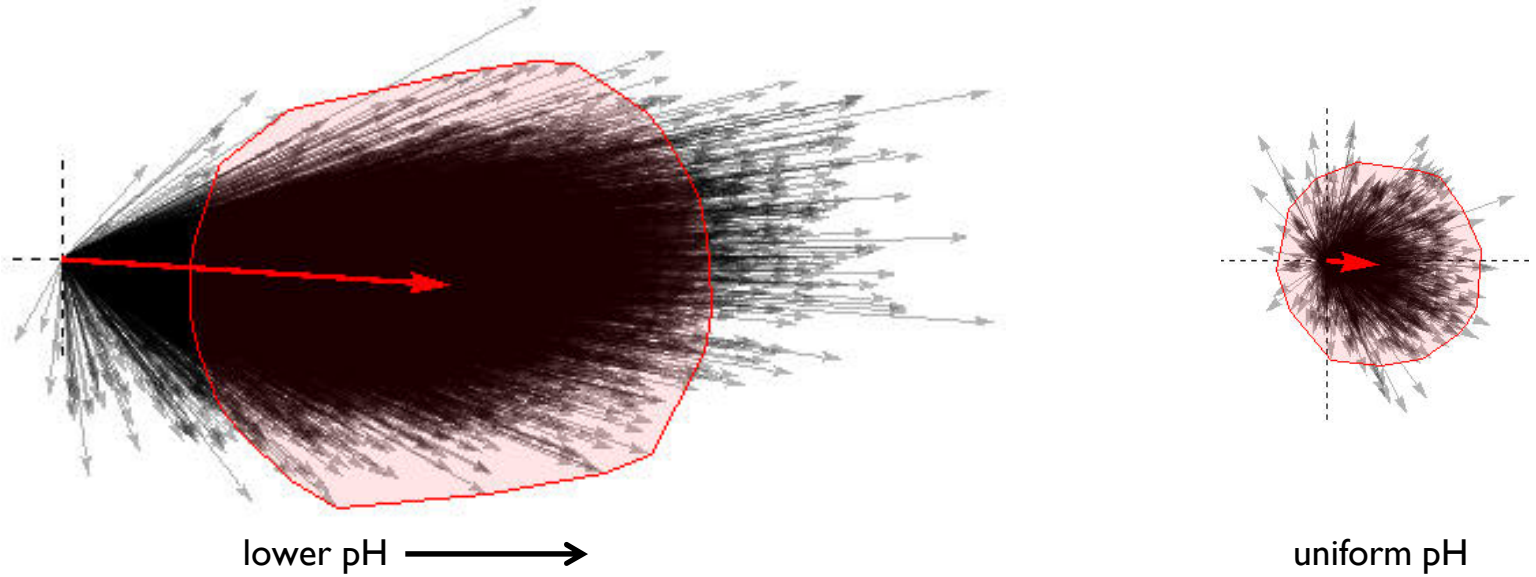


Background: Cells attach to their surroundings in part via transmembrane integrin molecules, which are hypothesized to adhere better in acidic (pH < 7.4) environments. It follows that in a pH gradient, an attached cell would migrate towards the acidic side because the trailing edge of the cell would detach more easily than the leading edge.

Limitation: Experimental limitations restrict data set size, so it is not clear whether the observed migration vectors or their mean are Gaussian. Is migration directed in either the pH gradient or the control?

Courtesy of R K Paradise

Bootstrapping example 3: more hypothesis testing with unknown distribution.

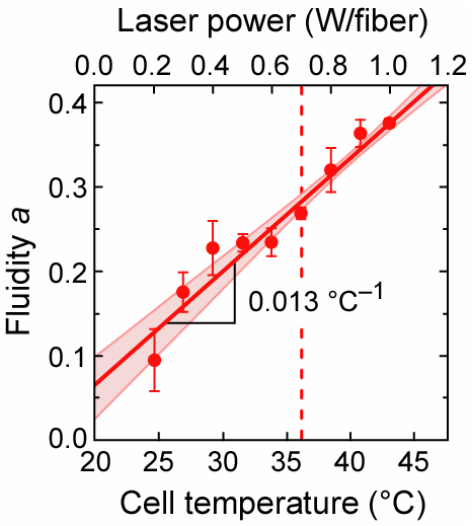


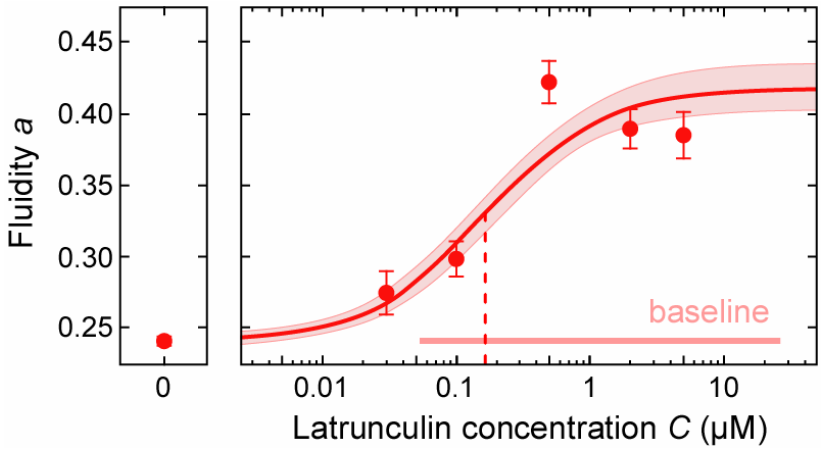
Bootstrapping solution: over thousands of iterations, sample with replacement from the original data set and calculate a total resultant migration vector.

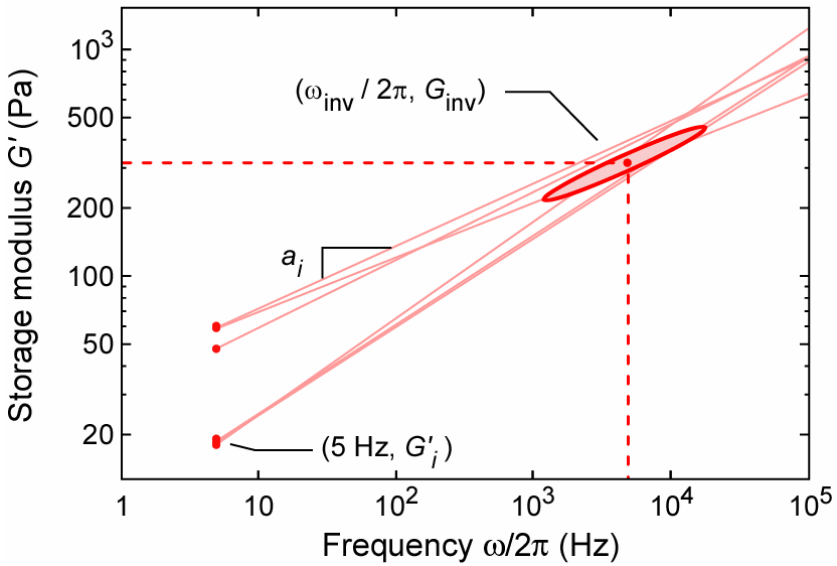
Drop the 5% most extreme vectors to obtain a 95% confidence region that will converge with sufficient iteration.

If this region doesn't contain the origin, we can conclude that migration is directed. If it does, we can conclude that the appearance of directed migration could easily be due to chance.

Bootstrapping example 4: 95% confidence regions.







Maloney and Van Vliet et al., *Soft Matter* 10(40) (2014)

Caveats and possible pitfalls:

- Small data sets can be problematic; consensus is that $n > 10–30$ is recommended.
- If the bootstrapped distribution is skewed, things get more complicated. (Actually a good thing: bootstrapping can help estimate parameter-fitting bias. Consult references.)
- Quality is only as good as the quality of the original experiment.

Summary recommendation:

Let's move away from limited, assumption-containing parametric functions like $se = \sqrt{\Sigma_i (x_i - \bar{x})^2 / ((n-1)n)}$ and start *routinely* using available computational power to study the real uncertainty in the parameters we estimate and conclusions we draw.

References

Efron, *An Introduction to the Bootstrap*;
Chernick, *Bootstrap Methods*;
Manly, *Randomization, Bootstrap, and Monte Carlo Methods in Biology*;
Shalizi, *The Bootstrap*.

MatLab: *bootstrapped_sample* = **randsample**(*data*,**length**(*data*),**true**)
Mathematica: *bootstrapped_sample* = **RandomChoice**[*data*,**Length**[*data*]]

These slides will be available at john.maloney.org
(also accessible through Van Vliet Group website).

Questions?