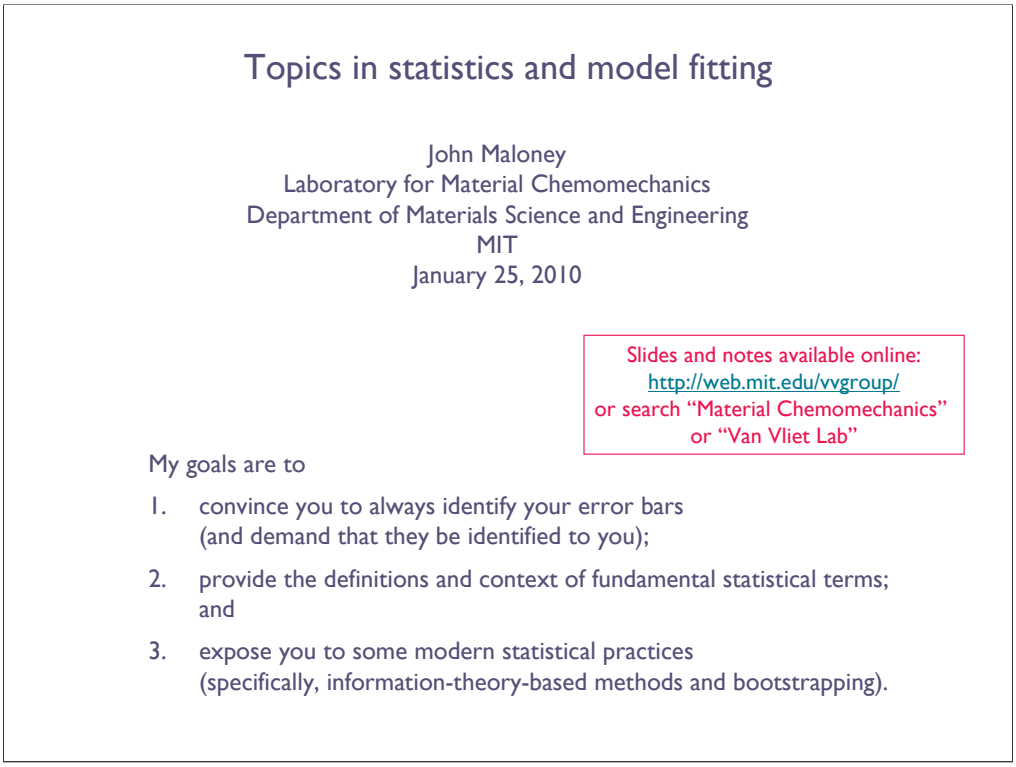
THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
STORY OF HUMANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY ◊ IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING ◊
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

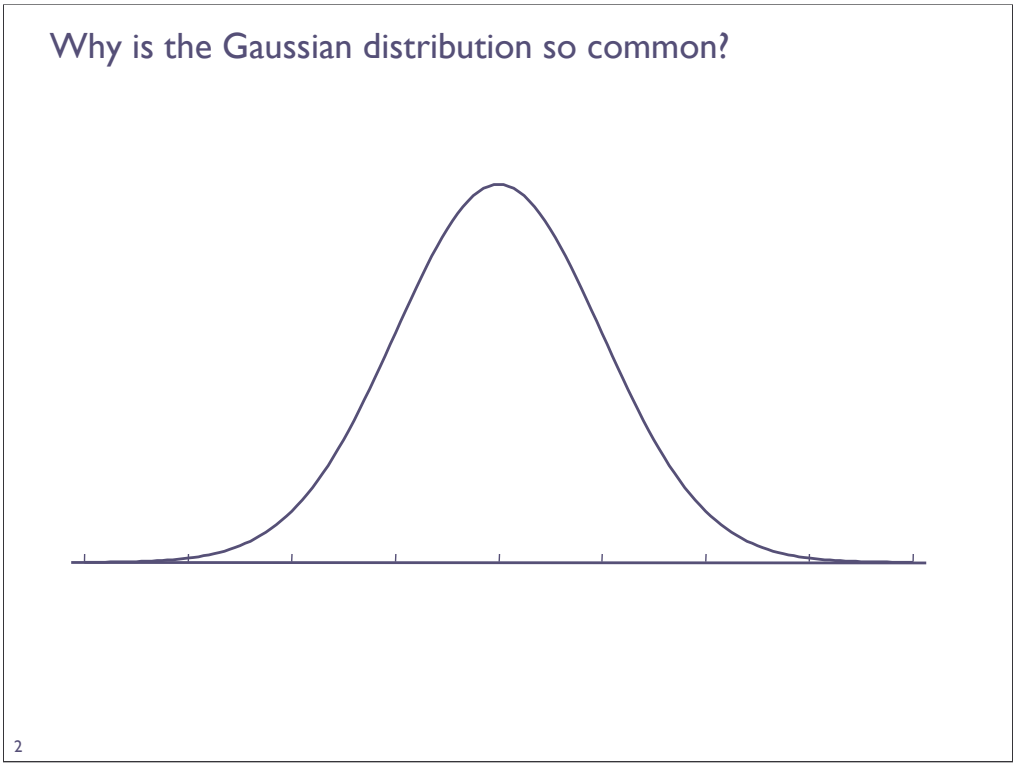
(after W. J. Youden, via E. Tufte)



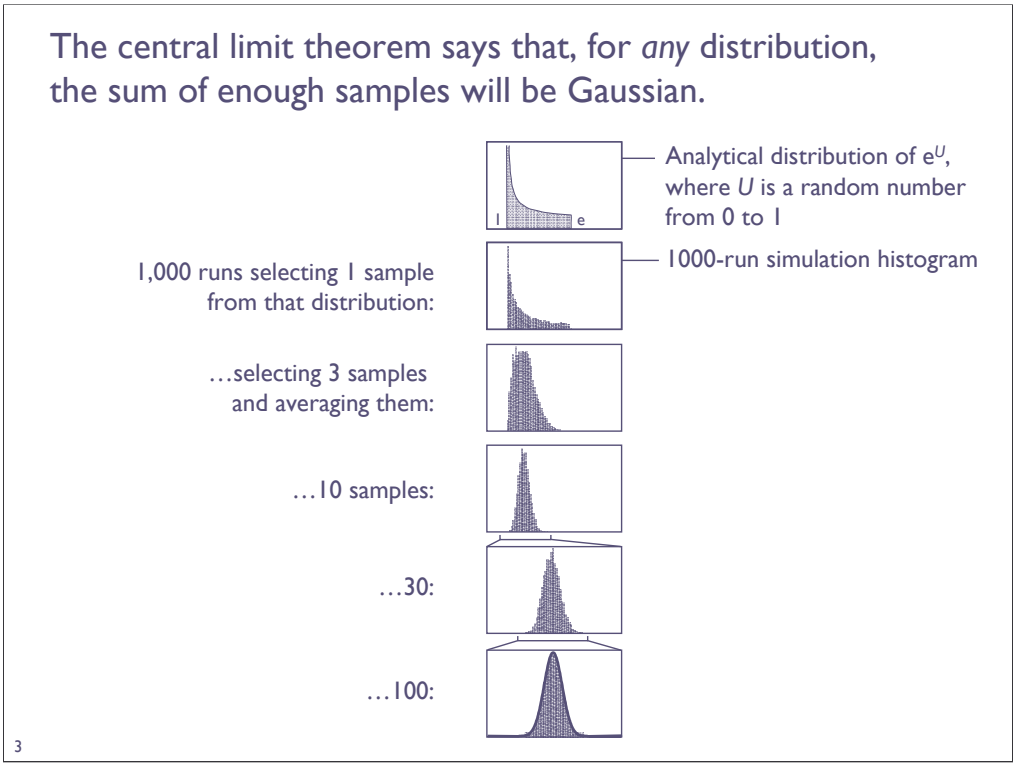
The theme of this talk is making decisions (specifically, research conclusions) in the face of uncertainty.

Before entering the materials science graduate program at MIT, my field was mechanical engineering and microfabrication. When I changed fields to materials science and cell biology, and began to research the mechanics and behavior of live tissue cells, I was astonished at the tremendous uncertainty that researchers in the life sciences encounter when working with complex living organisms.

Over the years, I've developed opinions on good practices in research. In terms of drawing well-founded conclusions from data, I think researchers are off to a great start if they (1) know how to relate uncertainty accurately to others, (2) know the basic ideas and nomenclature of statistical testing, and (3) know one or two clever ways to analyze data. So that's exactly what this talk covers.



First, we have to address one question: why is the Gaussian distribution – the "bell-shaped curve," also known as $P(x) \propto \exp(-x^2)$ – so common? Many statistical tests assume that the input data is Gaussian (i.e., that the data is drawn from some true Gaussian distribution existing in nature). Many introductory statistical texts never discuss anything else. It's even called the "normal" distribution. Why?

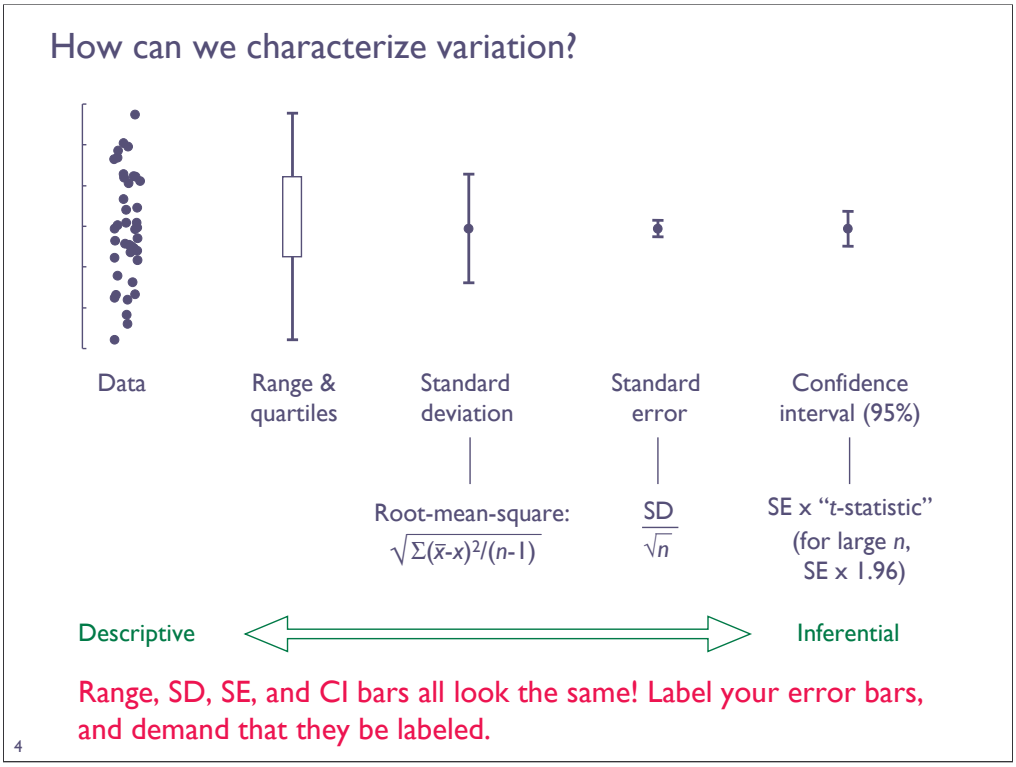


The ubiquity of Gaussian distributions arises from the so-called central limit theorem, which says that the sum (and therefore the average) of enough samples will be Gaussian, regardless of the distribution from which the samples are drawn. This is an amazing result! (The usefulness of this result is unfortunately tempered by the frequency that it is assumed without much basis.)

Before we look at an example, a note of explanation: the first box above contains a theoretical probability distribution: a graph of *expected* frequency (*y*-axis) that one will observe certain values of a certain parameter (*x*-axis). The remaining boxes contain histograms, which display the *actual* number of observations (here, via simulation), grouped into bins.

Now the example: consider the exponential function applied to a random number between zero and one. This is an asymmetric, bounded distribution, but a symmetric, smooth bell-shaped curve emerges after a dozen or so samples are averaged together. By the time we have accumulated one hundred samples, the Gaussian distribution is an excellent fit to the average.

The implication is that, when dealing with the average of a large enough number of samples, we can assume the data to be Gaussian-distributed and base our statistical reasoning and tests on this assumption. Most of the rest of this talk will assume that parameters are Gaussian-distributed (except for a discussion of the bootstrapping technique near the end).



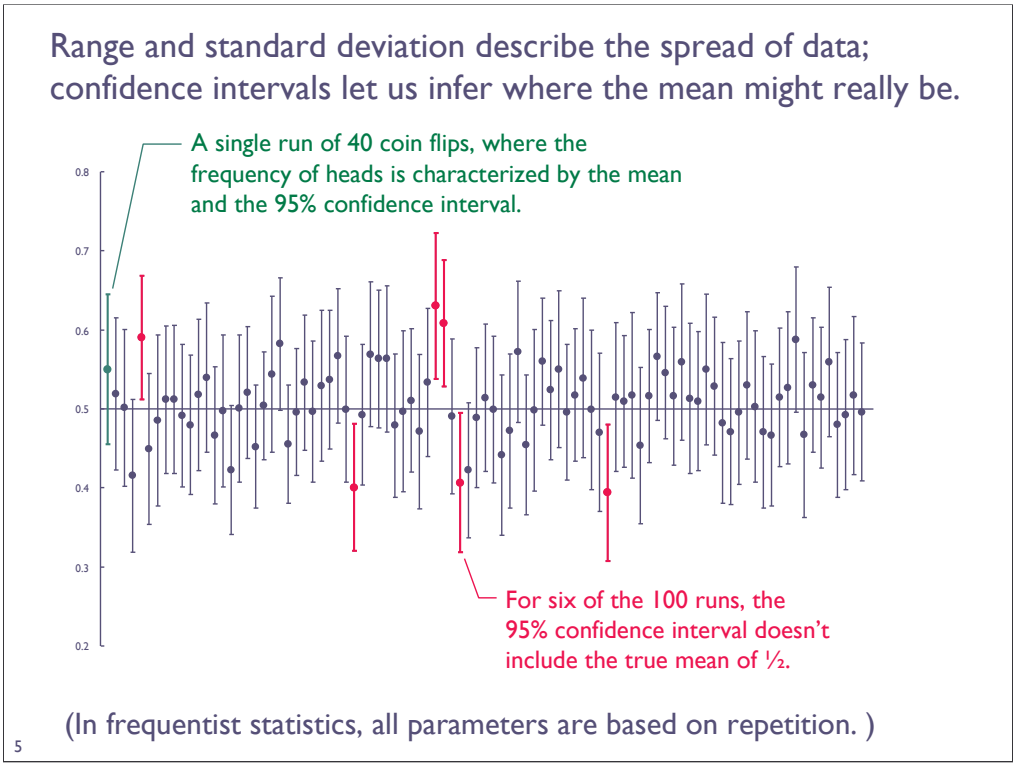
When conducting experiments that produce continuous values, we never observe the exact same number over and over. Instead, we observe a distribution of values. There are multiple ways to characterize this variation.

The complete data (here, simulated Gaussian data) is the most descriptive. But large data sets are unwieldy, and we prefer something more succinct. We might therefore choose to report variation in the form of a range.

Another metric, the standard deviation (SD), is defined as the square root of the average squared deviation of the data from the mean. Compared to the range, it is less sensitive to single extreme outliers. (The reason that the average in the SD is calculated by dividing by n-1 instead of n is discussed on the last slide.)

The standard error (SE) is calculated from the SD and, assuming a Gaussian distribution, represents the standard deviation not of the distribution but of the *mean* (note this difference between SD and SE). From this parameter we can begin to infer properties of the true mean.

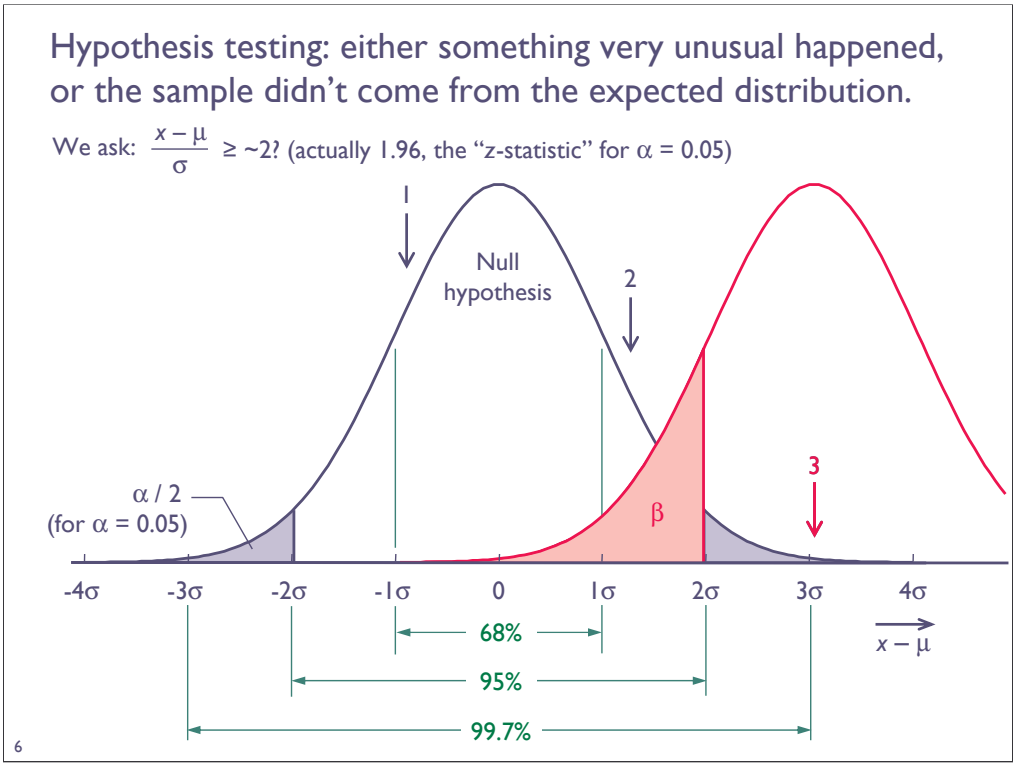
The confidence interval, another inferential statistic, is defined and discussed on the next slide; for now, let's just note that all these metrics of variation are graphically denoted by error bars. If these error bars aren't accompanied by an identifying label, what they signify is a mystery to the audience.



Confidence intervals (CIs) are defined in the context of *frequentist* statistics. Frequentist statistics are based on repeated equivalent experiments, in which we accumulate sample data in the hopes of estimating the properties of a true distribution.

Here's how CIs are defined: ninety-five percent of all 95% confidence intervals contain the true mean. (We can use any percent value we like, but 95% is especially common.) If we flipped a coin forty times while scoring heads as 1 and tails as 0, and repeated this experiment many times, the resulting 95% CIs (calculated according to the equations on the last slide) would very often contain the true mean of 0.5. But about 5% of the time, they wouldn't. If we didn't know the true mean (and we often don't when running experiments), we'd have no idea whether we were in the 95% group or the 5% group. Would using 99% CIs be better? They would more often contain the true mean, but they would also be larger and therefore less useful.

Note that a mistaken understanding of CIs would lead us to say that there's a 95% *chance* that a particular single CI contains the true mean. This isn't correct; as shown in the coin toss data above, a given CI either does or does not contain the true mean. Frequentist statistics do not describe degrees of belief. (An alternate school of thought, Bayesian statistics, does incorporate and quantify belief and is discussed later.)



When dealing with Gaussian-distributed data, it's handy to remember that the majority of the data lie within one standard deviation, the vast majority of the data within two, and essentially all the data within three.

Let's assume we've characterized some process and want to change the input conditions to see if an output value also changes. (This is really the key issue of this whole talk: in research, we want to see what parameters are relevant in natural processes. In development, we want to control and optimize engineering processes.) We'll sample this output once and compare it to the status quo.

A single measurement at location "1" or "2" would be unremarkable. If we measured output value "3" above, though, we'd have to conclude that either the process was unchanged (call this the null hypothesis) and something very unusual happened, or that the process was changed (call this the alternate hypothesis) and we've discovered some observable effect of changing the input. We can quantify this decision-making process by setting an arbitrary limit (here, about two standard deviations from either side of the mean). If the output is more extreme than this threshold, we reject the null hypothesis.

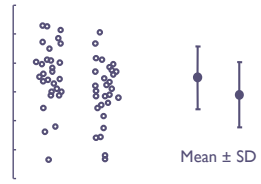
Two errors are possible, though: the "type I" error of excess credulity (α), where we reject the null hypothesis even though it applied and something unusual *did* happen, and the "type II" error of excess skepticism (β), where we fail to reject the null hypothesis because the measured output wasn't very extreme… even though the process was changed, and an alternate distribution applied.

Hypothesis testing enables decisions in the face of uncertainty.

We ask: $\frac{x-\mu}{\sigma} \geq$ ~2? (actually 1.96, the "*z*-statistic" for $\alpha = 0.05$)

$\frac{\text{signal}}{\text{noise}} \geq$ cutoff value?

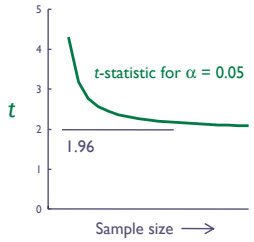
Often we want to compare two groups:



Mean ± SD

$\frac{\bar{x}_1 - \bar{x}_2}{S} \geq t$?

Pooled standard error



When there are more than two groups, we use ANOVA, a generalized *t*-test. (Why can't we just perform multiple *t*-tests?)

We use stricter *post hoc* tests for comparing data sets that catch our eye. (Why can't we just use the regular *t*-test?)

7

One way of thinking about hypothesis testing is in terms of a signal-to-noise ratio. Is the signal of a possible effect noticeable in the midst of system noise? For a single data point drawn from a Gaussian distribution with mean μ and standard deviation σ, we simply check whether the distance from the mean, normalized to the standard deviation, is larger than 1.96. This is the "*z*-test," with the *z*-statistic of 1.96 being the number of standard deviations (left and right) that encloses 95% of the data.

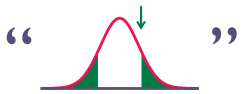
Often we want to compare two groups whose true means and standard deviations are unknown. We can estimate these parameters from the data itself, but the threshold value that marks a statistically significant difference is more demanding (larger than 1.96), especially when the sample sizes are small. In other words, the smaller the samples, the less we know, the higher the effective noise, and the higher the necessary signal to convince us that an effect exists. This test is the widely used "*t*-test," invented by William Gosset and published under the pseudonym "Student" (thus, "Student's *t*-test").

A couple caveats: (1) Each *t*-test has a built in type I error rate of α. If we were to collect several data sets and perform many *t*-tests to compare each possible pair, we're more likely to conclude that a data set is statistically significantly different, even if it isn't. This is no good. Analysis of variance (ANOVA) is a generalized approach for comparing more than two data sets correctly. (2) It is assumed that the data sets are independent, so we can't select extreme results that catch our eye after performing the experiment. If we could, then I could select the most opposed confidence intervals on slide 5, compare those two runs, and conclude that the coin is statistically significantly different from itself.

What is a *p*-value? (And what is it not?)

A *p*-value is the expected frequency, in many repeated experiments, of observing at least as extreme a result, given that the null hypothesis is true.

Briefly, $p = P$ (data | null hypothesis).



It is not the probability of a false positive (this is α, the level of significance), or the probability of a false negative (this is β, where $1-\beta$ is called the test's power).

It is not the probability that the null hypothesis is true (consider flipping 11/20 heads; $p = 0.41$).

If greater than α, it does not signify that there is no effect.

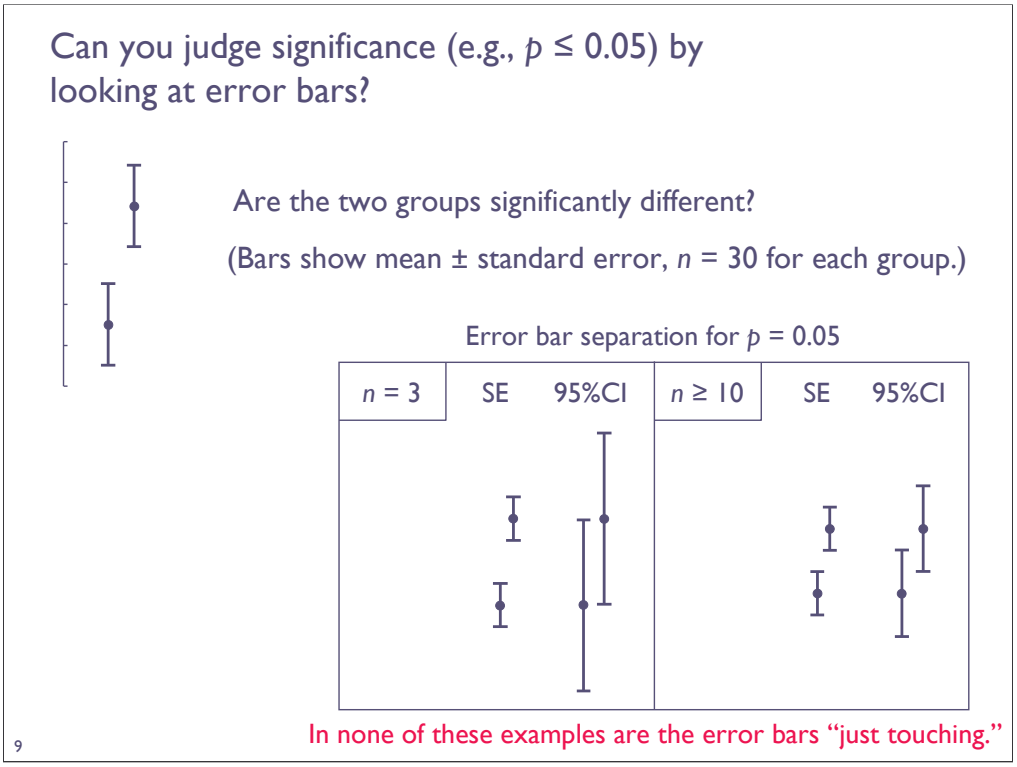
It does not give information about the scientific importance of the effect.

8

Like the confidence interval, the *p*-value is grounded in frequentist statistics. It can only be interpreted in the context of equivalent repeated experiments. It is calculated under a (somewhat odd) fundamental assumption: that the null hypothesis *is* true. Like a *reductio ad absurdum* argument, the null hypothesis is proposed in the hope of rejecting it. The key idea of frequentist hypothesis testing is this: if the null hypothesis generally fails to predict results as extreme as our data (or more extreme), then we should reject the null hypothesis. The *p*-value is the test output that we compare to the significance level α to decide to reject or not reject.

Unfortunately, there are many ways to misinterpret *p*-values, all of which can be found in the literature. It is particularly common, for example, to see conclusions that factor *X* has "no effect" on result *Y* just because the researcher obtained a $p > 0.05$. The logical error here is hopefully obvious.

Researchers have also been known to confuse statistical significance with scientific importance. Prof. David Lykken wisely noted that "[t]he value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied."



Some people try to identify statistical significance of a purported effect by checking if the error bars are "just touching." As far as I can tell, this method doesn't have much basis. (Of course, if the errors bars aren't labeled, it's especially pointless.)

The paradigm of hypothesis testing isn't flawless.

The level of significance α is an arbitrary value (e.g., 0.05) that separates publishable results from unpublishable results.

The null hypothesis is often known (and usually hoped) to be false.

Confidence intervals and *p*-values aren't what they're often interpreted to be: the *p*-value is P(data | null hypothesis), *not* P(hypothesis) or P(hypothesis | data).

The alternative hypothesis isn't even necessarily evaluated.

It would be very unusual (e.g., a 1-in-100 chance) to observe at least as extreme a result as X, given that the null hypothesis is true.

But we have observed result X.

Therefore, with a *p*-value of 0.01, we reject the null hypothesis at a significance level of $\alpha = 0.05$.

It would be extremely unusual (a chance of 535 in 300M) for a particular American to be a member of Congress.

Person X is a member of Congress.

Therefore, we reject the null hypothesis that person X is American, due to a *p*-value of 0.000002. So significant!

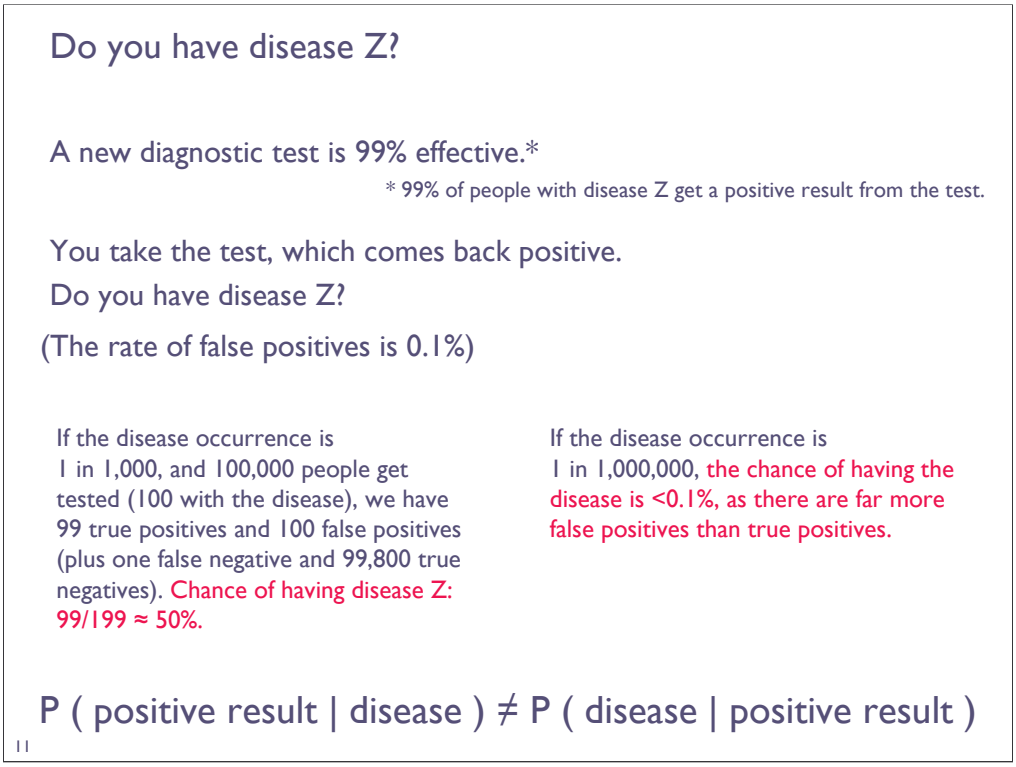
What went wrong?

10

(after Cohen)

There are several peculiarities and limitations of frequentist hypothesis testing and *p*-values. One notable weakness is that the null hypothesis can end up being rejected without even considering the likelihood of alternate hypotheses.

The two examples given here demonstrate first a standard, and then a ludicrous, conclusion reached through standard hypothesis testing. Illogical results like this one are used to motivate the study and use of an alternative school of statistical thought: Bayesian statistics.



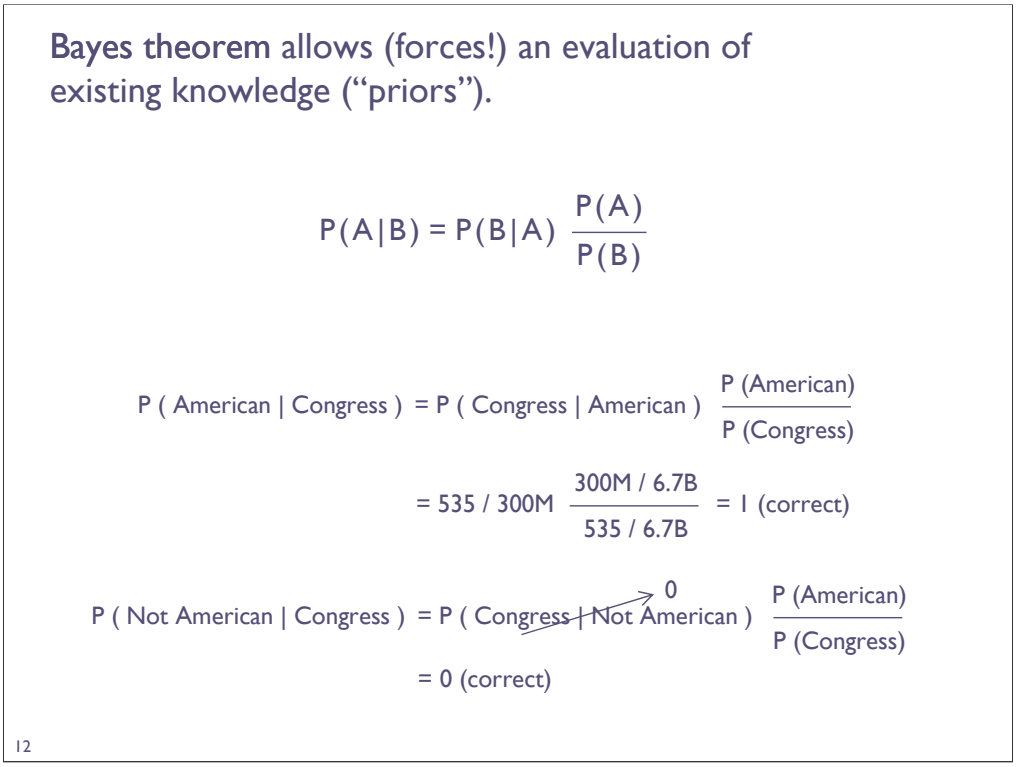
This is another example (really, the canonical example) used to highlight problems with frequentist hypothesis testing and to motivate Bayes' Theorem, which is described on the next slide.

Let's consider the scenario of receiving a positive result on a diagnostic test for disease Z. We might think: since this diagnostic test seems so effective, and false positives so rare, let's take the positive result as proof that we have the disease. We can even quantify our decision as $p = 0.001$. So significant!

In the nomenclature of hypothesis testing, we've rejected the null hypothesis (which is that we're disease-free) due to the unlikelihood of getting a positive result if the null hypothesis holds. Unfortunately for our reasoning, the alternative hypothesis (which is that we have the disease) might be even less likely than the chance of getting a false positive. Whoops! (Or hooray, since we're probably disease-free.)

The bottom line is that, counterintuitively, a disease that is "99% effective" (but note how this is defined) and with a very low false positive right might still give the wrong answer almost all the time. As patients, we'd prefer to know
P (disease | positive result) as opposed to P (positive result | disease), which is of more use to the test designers and manufacturers.

The more general concept is P (data | hypothesis) ≠ P (hypothesis | data).

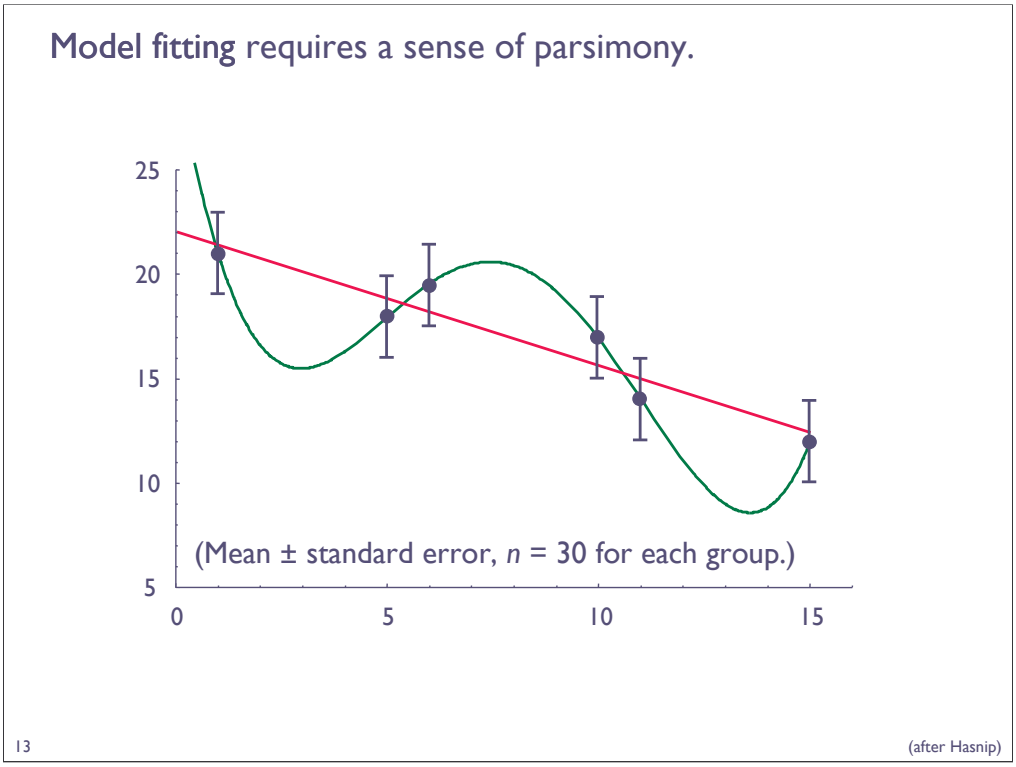


Bayes' Theorem, shown above, relates P (A | B) to P (B | A); hopefully the value of such a relationship is evident after the last two examples.

Bayes' Theorem is one of the tools of Bayesian statistics, in which we first think carefully about what is already known (called the prior distribution), then we collect data, then we adjust our prior beliefs in light of new data (the output is called the posterior distribution). Note that the Bayesian approach gets the American congressman problem right, under the condition that we enter additional information: P (American) and P (Congress member).

Unlike frequentist statistics, Bayesian statistics allow us to talk about a degree of belief. It is totally acceptable to speak in terms of the probability of a single event that cannot be repeated. Another contrast: Frequentists regard the distributions as real; that is, there exist physical laws in nature that we may approximate with our limited experimental data. Bayesians regard the experimental data as the only thing that is real; the physical law is merely a model to be fit.

A limitation to Bayesian statistics is that our prior knowledge is subjective and could even vary from person to person, and is therefore often subject to vigorous debate. But frequentist statistics can be subjective too, for example in our choice of the alternative hypothesis that is duly accepted whenever the null hypothesis is rejected.

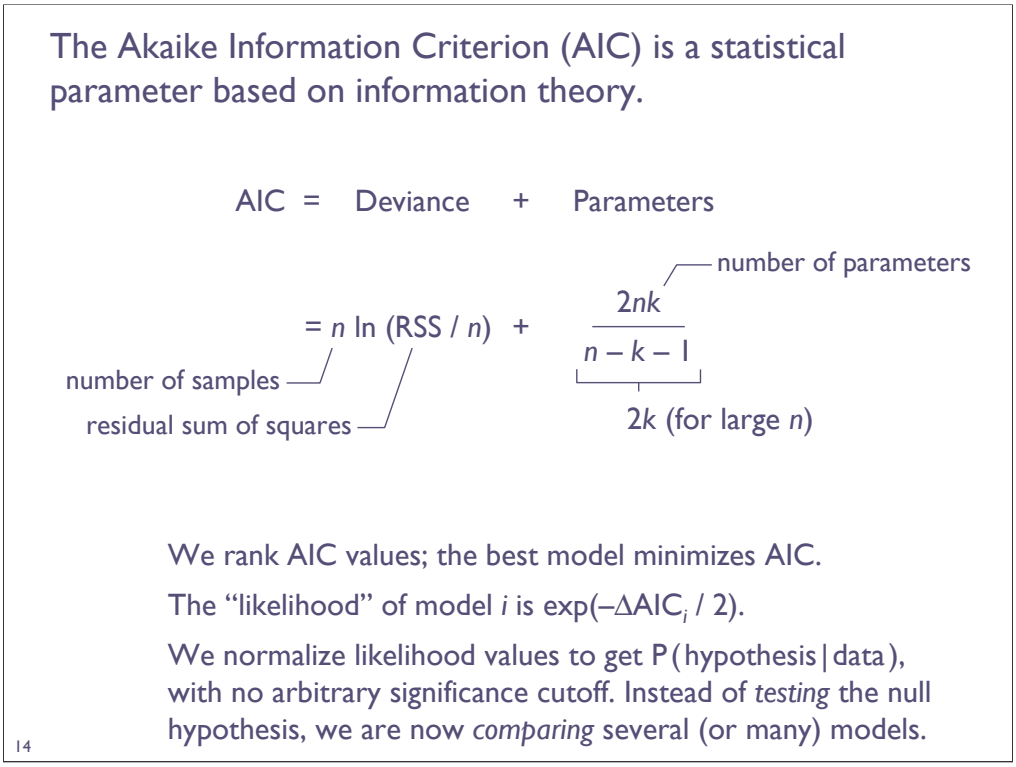


Now that we've explored the difference between calculating data (given the null hypothesis) and evaluating a hypothesis (given the data), let's look at a concrete example of choosing a suitable model (a type of hypothesis), given a set of data.

Which is the better model, the green curve or the red line? The green curve fits the data perfectly, while the red line misses every average. If we were to collect a new set of data points tomorrow, though, would the green curve still be such a great fit? Given the error bars, one might suspect that the green curve is fitting quite a bit of noise, not just the signal, and that the straight red line is the better model because it uses fewer variables yet still seems to capture the underlying behavior of the system.

We should be parsimonious with our fitted variables. As William of Ockham noted: let us not needlessly accumulate explanations. Applied to research conclusions, "Ockham's Razor" metaphorically slices away superfluous parameters.

But how do we do this rigorously? How do we assign a penalty to excess variables, for example?

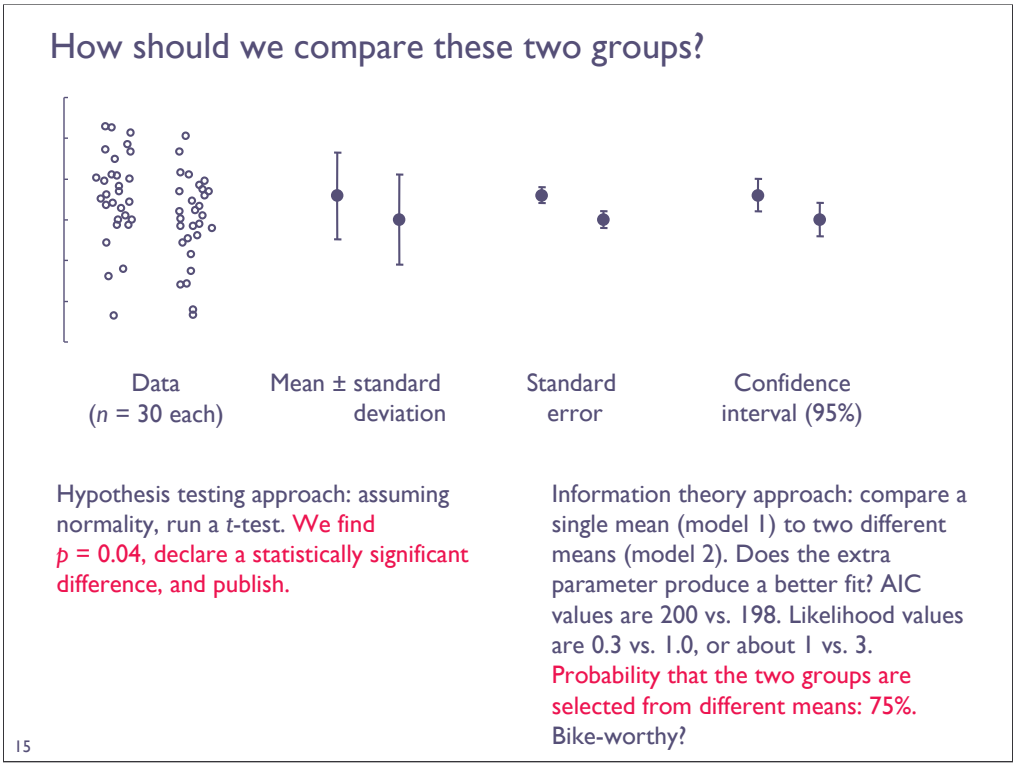


About forty years ago, Hirotsugu Akaike developed a way to compare models and hypothesis by using some heavy-duty information theory mathematics. His metric, the AIC, rewards good fits with a minimum of variables by including penalty terms for the deviance of the model and also the complexity of the model. By minimizing the AIC of each model, we hope to find a happy medium between accuracy and simplicity.

In AIC ranking, there is no privileged hypothesis such as the null hypothesis and the alternative hypothesis. In addition, we can evaluate as many models as we like simultaneously. Some simple mathematics converts the AIC difference between models to a "likelihood" value that has been compared to lottery tickets; just as holding more tickets gives you a better chance of winning the lottery, a higher likelihood value marks a model that is more likely to fit current and future data.

We can normalize each likelihood value by the total sum to obtain a probability, P (hypothesis | data). This is usually the figure of merit we want, the one that is not available through hypothesis testing.

Data analysis software such as Mathematica and Origin now allow the user to compare models by minimum AIC value, and also by related criteria (such as the BIC, the Bayes Information Criterion) that work a similar way but use slightly different penalty terms.



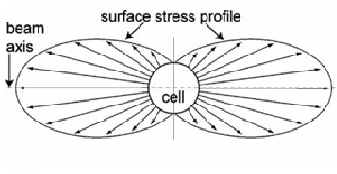
We return to the question of whether two groups came from the same distribution. Perhaps these are strength measurements of a weld prepared two different ways; perhaps they are growth measurements of cell cultures grown with two different nutrients. Whatever the experimental system, we're back at the pivotal research question: is the difference in the sample means due to chance, or a real effect?

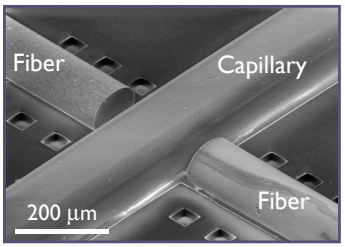
The frequentist approach is to conduct a *t*-test, assuming that the data are sufficiently normal. It is totally unremarkable for a *p*-value of 0.04 to be published and used as evidence for a purported effect.

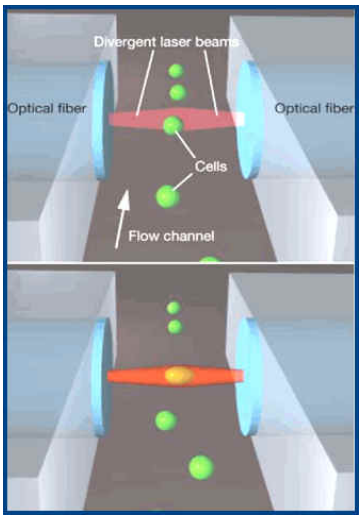
Earlier we noted that the *p*-value is the frequency we'd see at least this much difference in the means, assuming there was no effect (i.e., *p* equals the probability of the data given the null hypothesis of no effect), and how it would be nice to be able to calculate rather the probability of an effect given the data. We realize that hypothesis testing is itself a model fitting problem: should we be parsimonious and describe both data sets by a single mean, or do the data sets differ sufficiently to warrant using two means and thus concluding that an effect exists? The AIC can be applied to such a question.

Under the AIC approach, we compare likelihoods of the two hypotheses and calculate a 75% chance that an effect exists. Is this enough to publish? I think it's enough to justify repeating the experiment.

How should we model photon-induced tissue cell deformation?







We hypothesize that mesenchymal stem cells can be distinguished and characterized by their mechanical properties as evaluated by optical stretching.

16

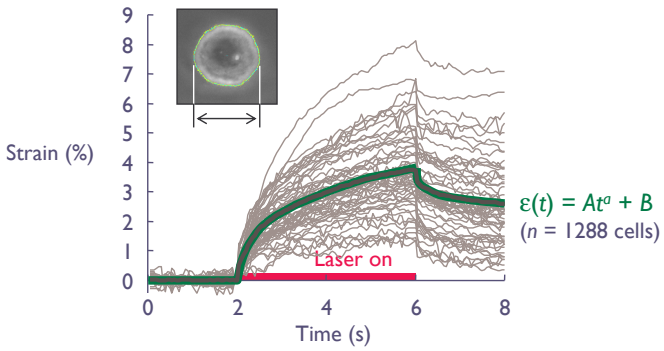
Guck J et al. *Biophys J* 88(5) (2005)

I research tissue cell mechanics: how far cells can "feel" into their surrounding environment and how single cells behave as deformable materials. To probe the deformability of individual cells, I use a technique called optical stretching, in which twin infrared laser beams are aimed at single cells floating in suspension. The photonic pressure alone is enough to deform the cells, which stretch about 10% over a few seconds after a step increase in laser power, recovering partially when the laser power is reduced again.

The non-contact, high-throughput nature of this technique (developed by Jochen Guck, Cambridge University) gives it great promise in diagnosing cell-altering diseases, identifying sub-populations of heterogeneous cell populations such as bone-marrow-derived mesenchymal stem cells, and studying the contributions of cytoskeletal components to the mechanical behavior of single cells.

For this example, we are interested in finding the uncertainty of an fitted parameter in an empirical model of cell deformability under load.

"Bootstrapping": a technique for estimating variance.



Best fit: a = 0.22 (but with what error?)

In bootstrapping we sample with replacement from our original data set (some values are repeated, some omitted).

We recalculate our parameter of interest and repeat hundreds or thousands of times; the output changes slightly with each run.

The standard deviation of the collection of bootstrapped outputs is a good estimate of the true standard error of the true output.

17

We are looking at the time-dependent elongation of a suspended cell stretched by photonic pressure. Although each individual cell's response is too noisy to fit a model accurately, we can average the responses together to get a smoother curve. This data can then be fit to a governing equation, which is a power law in this case. (The finding of a power-law exponent $a = 0.22$ is especially interesting to us, since other groups have found $a \approx 0.2$ when investigating attached cells with other techniques. At this point, this exponent appears to be a conserved, universal feature of tissue cells.)

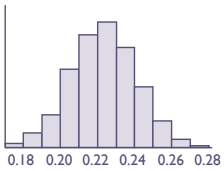
The problem lies in estimating the uncertainty of the fitted exponent. Is it 0.22 ± 0.1? 0.22 ± 0.00001? We've already used all the data to get the fitted value; there's nothing left that we can use to extract a standard error. Or is there?

The bootstrap technique is perfect for this problem. The central concept of bootstrapping is based on resampling: *Lacking additional data, we* simulate *additional data by resampling existing data with replacement*. This resampled data turns out to be remarkably valuable. (In a sense, we're pulling ourselves up by our bootstraps, as the saying goes, by acquiring extra information from our existing data).

We're going to repeat the resampling-with-replacement process many times (this is shown in detail on the next slide). Each time, we'll calculate our parameter of interest. A major result of bootstrap theory is that the standard deviation of all the bootstrapped outputs (multiplied by a correction factor close to unity) is a good estimate of the standard error of that parameter of interest.

Bootstrap example (cont'd): resampling and results.

		Fitted exponent
Original *n* cells	{ 1 2 3 4 5 6 … 1288 }	0.221
Sample 1	{ 1 4 6 9 10 11 … 1288 }	0.268
Sample 2	{ 1 2 3 4 4 5 … 1288 }	0.262
Sample 3	{ 1 1 3 4 4 8 … 1288 }	0.214
…	…	…
Sample 1,000	{ 1 2 3 3 3 3 … 1287 }	0.238



Average of *N* bootstrapped outputs: 0.224

SD of outputs: 0.0201

Estimated SE of fitted exponent (SD x $\sqrt{n/(n-1)}$: 0.0201

In bootstrapping we sample with replacement from our original data set (some values are repeated, some omitted).

We recalculate our parameter of interest and repeat hundreds or thousands of times; the output changes slightly with each run.

The standard deviation of the collection of bootstrapped outputs is a good estimate of the true standard error of the true output.

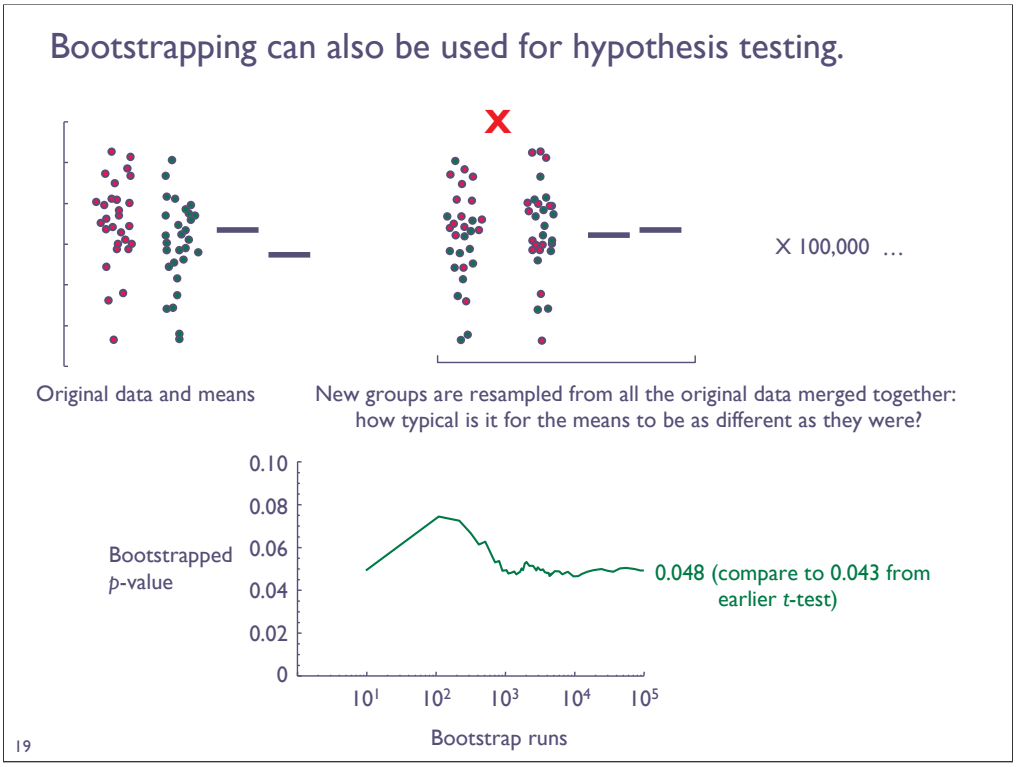
18

Here are the details of the bootstrap process, applied to the problem of identifying the variance of the fitted exponent of the power-law model.

In sampling with replacement, some data are omitted and some repeated. Here the data are not single values, but rather entire strain vs. time responses for each of 1288 cells.

We use the bootstrapping result that the standard deviation of the bootstrapped outputs is a good estimate of the standard error of the parameter of interest. We just need to multiply the standard deviation by a correction factor (the square root of $n/(n-1)$, where n is the data set size), which is essentially unity here.

Our results are: (1) The distribution of *a* appears to be Gaussian (this is not a necessity for bootstrapping; in fact, an advantage of bootstrapping it that it is amenable to analyzing non-Gaussian distributions.); (2) The average bootstrapped *a* matches the original *a* (0.224 vs. 0.221), which indicates a lack of bias (bias is not covered in this talk but is discussed in all introductory bootstrap references); (3) The standard deviation in the exponent is 0.0201, which is what we wanted to find; (4) Moving further, we could sort the boostrapped outputs, look at the 2.5% and 97.5% percentiles, and conclude that the 95% confidence interval for *a* is [0.18, 0.26]. We have thus identified a confidence interval by Monte Carlo methods alone, without having to assume a Gaussian distribution or use a *z*-statistic.



Bootstrapping can also be used for hypothesis testing, giving us yet another way to compare two groups to see if the difference in their means is likely or not to be due to chance. Let's consider the two groups that scored a *p*-value of 0.04 earlier.

Once again, there's a parallel between hypothesis testing and model fitting. The null hypothesis is that the true mean are identical; the alternative hypothesis is that they're different. Equivalently, we consider one model in which the groups came from the same distribution, and another in which they came from different distributions. The second model fits better, but requires an additional parameter (the second mean).

The bootstrapping approach here is to merge all the data and repeatedly draw two groups from the merged data. If the means are more different that they were in the original data, score that run a "1." Otherwise, score it a zero. After numerous runs, the normalized score represents a bootstrapped *p*-value. It is literally a score of how likely the differences in the original data sets could be due just to chance.

The bootstrapped *p*-value converges after 10^3 to 10^4 runs and agrees well with the *t*-test *p*-value calculated earlier. (Note that we're free to perform bootstrapping on distributions that aren't necessarily Gaussian, and textbooks on bootstrap theory cover how unusual these distribution can get before even the bootstrap approach breaks down.) For now, let's just note that we have an additional tool at our disposal to investigate how likely a research result is due to chance and hopefully to improve our decision-making process.

Standard deviation:
why use $\sqrt{\Sigma(\bar{x} - x)^2 / (n - 1)}$ instead of $\sqrt{\Sigma(\bar{x} - x)^2 / n}$?

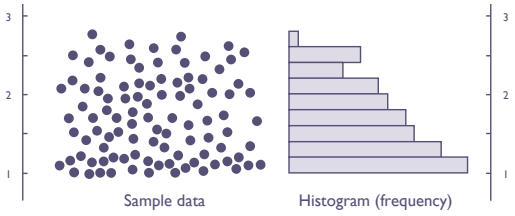
Although we generally don't know a population's true standard deviation, we can estimate it from a sample. The second, simpler equation above (with *n* in the denominator) provides a pretty good estimator of the population standard deviation (σ), especially for large sample sizes; however, it generally underestimates σ. It is a *biased estimator* (unlike the sample mean, which is an unbiased estimator of the population mean (μ)).

Here's one way to see why: typically somewhere out in the population, not included in our sample, are large positive and negative values. These extreme values tend to cancel each other out when calculating averages, so they don't cause much of a difference between the sample and population means (over the long run, they cause no difference at all). However, extreme values, positive or negative, *always* increase standard deviation calculations because of the squared term. Our sample estimator is therefore too small as a result of missing them. We correct for this bias by using ($n - 1$) in the denominator instead of *n*, thus increasing slightly the value of our estimator.

Bootstrap example: description and data.

The bootstrap method is a way to estimate the uncertainty in any function of a set of sample data. In this example, we will let the function be simply the mean or the standard deviation of our sample, and the data will be drawn from a known distribution to let us check the answers later. The real power of the bootstrap method, however, comes in applying it to complex functions of limited data from unknown (especially non-Gaussian) distributions. *Lacking additional data, we* simulate *additional data by resampling existing data*. This resampled data turns out to be remarkably valuable.

Our sample data consists of 100 values drawn from a clearly non-Gaussian distribution. The data range from approximately 1.0 to 2.7.



In the bootstrap method, we repeatedly sample 100 values from this data set *with replacement*. Nearly always, some values will be repeated and some omitted. We calculate the function(s) of interest from each sample, and repeat hundreds or thousands of times.

Bootstrap example (cont'd): resampling and results.

		Mean	SD
Original data	{1.27, 1.74, 2.13, 2.18, 1.94, 1.37, 2.21, …, 1.23}	1.66	0.469
Sample 1	{1.13, 1.50, 2.48, 1.23, 1.04, 2.48, 1.28, …, 1.67}	1.70	0.466
Sample 2	{1.46, 1.09, 1.04, 1.79, 1.76, 1.25, 1.06, …, 2.40}	1.72	0.494
Sample 3	{1.88, 1.14, 1.12, 2.14, 1.50, 1.74, 1.07, …, 1.59}	1.56	0.420
…	…	…	…
Sample 10,000	{1.59, 1.42, 1.94, 1.28, 1.36, 1.42, 1.62, …, 1.91}	1.66	0.496

A key result of bootstrap theory is that the standard deviation of a collection of bootstrapped function values (e.g., mean or standard deviation) is an estimator for the standard error of the same function value of the original sample.

Average of means: 1.66

SD of means: 0.0468

95% CI for the true mean: [1.57, 1.75]

Average of standard deviations: 0.465

SD of standard deviations: 0.0237

95% CI for the true standard deviation: [0.420, 0.495]

Corroboration 1 (see next slide)

Corroboration 2

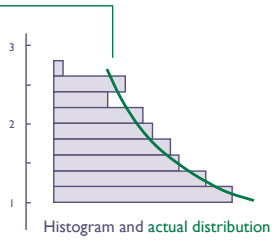
Bootstrap example (cont'd): conclusions.

The bootstrapped function values converged to within 1% within 7,000 iterations.

Because of the large number of samples and because I know the distribution from which the samples were taken, we can corroborate some of the bootstrap estimates with analytical calculations.

Corroboration 1: With 100 samples, we know from the CLT that the mean is Gaussian-distributed. Therefore, we could have estimated the standard error of the mean from the sample standard deviation divided by $\sqrt{n}$ (= $\sqrt{100}$), giving 0.0469. The bootstrapped estimate is 0.0468!

Corroboration 2: The distribution I used was $e^{U[0,1]}$, the exponential function of a random number between 0 and 1. I calculated the mean to be $(e - 1) = 1.72$ and (with great effort) the standard deviation to be $[(e - 1)(3 - e)/2]^{1/2} = 0.492$. The respective bootstrapped 95% confidence intervals ([1.57, 1.75] and [0.420, 0.495]) include these values.



References

Typeset normal law: Youden, *Experimentation and Measurement*; Tufte, *The Visual Display of Quantitative Information*.

Error bars: Cumming et al., "Error bars in experimental biology"; Belia et al., "Researchers misunderstand confidence intervals and standard error bars"; Vaux, "Error message"; McDonald, *Handbook of Biological Statistics*.

Weaknesses of hypothesis testing: Harlow et al., *What If There Were No Significance Tests?*; Cohen, "The Earth is round ($p < .05$)."

Parsimony and model fitting: Hasnip, "Mathematical Modelling."

AIC: Anderson, *Model Based Inference in the Life Sciences: A Primer on Evidence*; Wagenmakers and Farrell, "AIC model selection using Akaike weights"; Anderson et al., "Null hypothesis testing: problems, prevalence, and an alternative"; Motulsky and Christopoulos, *Fitting Models to Biological Data Using Linear and Nonlinear Regression*.

Bootstrap: Chernick, *Bootstrap Methods*; Manly, *Randomization, Bootstrap, and Monte Carlo Methods in Biology*.

General: van Belle, *Statistical Rules of Thumb*; Ambrosius et al., *Topics in Biostatistics*.