Experimental Design, the Art of Scientific Measurement, and the NIH Policy on Enhancing Reproducibility in Research

Emory University Pediatric K-Club

Pete Lollar, MD
Aflac Cancer and Blood Disorders Center
November 8, 2021

The “Reproducibility Crisis”

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine | www.plosmedicine.org

August 2005 | Volume 2 | Issue 8 | e124

THE WALL STREET JOURNAL April 7, 2017

The Breakdown in Biomedical Research

Contaminated samples, faulty studies and inadequate training have created a crisis in laboratories and industry, slowing the quest for new treatments and cures

NIH plans to enhance reproducibility

| NATURE | VOL 505 | 30 JANUARY 2014

Publications in 2020 with “reproducibility” in the title: 899

NIH Policy : Enhancing Reproducibility through Rigor and Transparency

Guidance: Rigor and Reproducibility in Grant Applications

Learn how to address rigor and reproducibility in your grant application and discover what reviewers are looking for as they evaluate the application for scientific merit.

Scientific rigor - the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results.

Transparency - the accessibility of information

https://grants.nih.gov/policy/reproducibility/guidance.htm

What is Reproducibility?

What does research reproducibility mean?

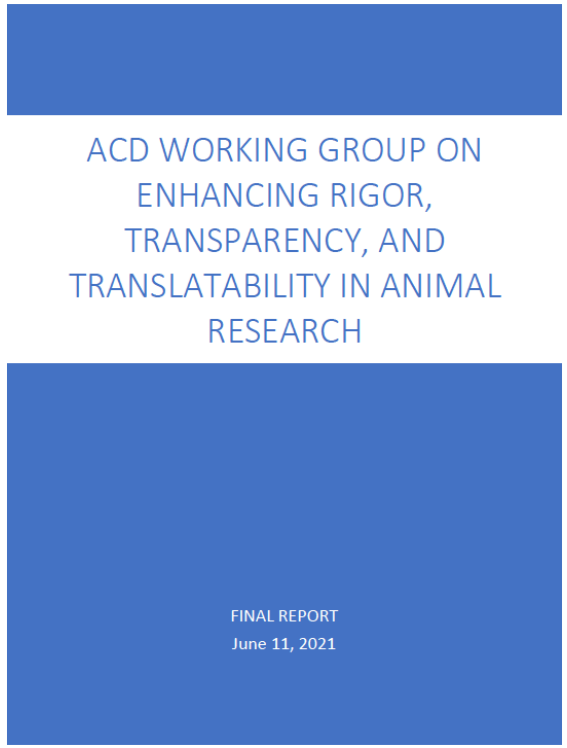
Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

www.ScienceTranslationalMedicine.org 1 June 2016 Vol 8 Issue 341

Methods reproducibility - Sufficient detail about procedures and data is provided so that the same procedures can be *exactly* repeated. Similar to the definition of "transparency".

Results reproducibility - The *same results* are obtained when an independent study with procedures as close to the original study as possible is conducted. Similar to some definitions of "replicability."

Inferential reproducibility - Qualitatively similar conclusions are drawn from either an replication or reanalysis of the original study. This is part of the process by which a scientific field decides which research claims or effects are to be accepted.



Advisory Committee to the Director of the NIH (ACD) 2021

What is Reproducibility?

"What scientists are … really concerned about when they debate research reproducibility is the truth of research claims".

Goodman SN et al. *Science Translational Medicine* 2016

Truth claims are qualitative statements based on inferential reproducibility, which in turn is based on the reproducibility of methods and results.

Examples:

The earth and the planets revolve around the sun.

mRNA encoding SARS-CoV-2 vaccines prevent COVID-19.

The chemical formula of water is H_2O.

The SARS-CoV-2 virus originated at the Wuhan Institute of Virology

Some Causes of Irreproducible Results

NIH plans to enhance reproducibility

| NATURE | VOL 505 | 30 JANUARY 2014

- Poor experimental design
- Poor training of researchers in experimental design, including statistics
- Poor transparency
- Unpublished negative results
- Overemphasis on making provocative statements rather than presenting technical details
- Overvaluation by academic centers, funding agencies and publishers of high-profile research
- Financial interests

Collins FS & Tabek LA *Nature* 2014

Elements of Poor Experimental Design

- Small sample size
- Small effect size
- Lack of blinding
- Lack of randomization
- Failure to account for effects of sex differences
- Testing too many variables
- Overreliance on unproven analytical procedures

Collins FS & Tabek LA *Nature* 2014
Ioannidis JP *PloS Medicine* 2005

What is Scientific Rigor?

NIH definition - the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results.

In other words, do good science.

But how?

What is Science?

Philosophers, on the outside looking in, disagree on the nature of the scientific method and what constitutes scientific truth.

"Although science has been miraculously successful ..... this strange fact cannot be explained."

Karl Popper *Objective Knowledge* 1972

Today's Presentation

Sources of irreproducibility		Keys to reproducibility
Random error	→	Use sound statistical inference when necessary
Systematic error	→	Strive for accuracy
Population differences	→	Use good models
Self-deception	→	Aspire to scientific humility
Fraud	→	Be honest

Combating Self-Deception with Scientific Humility

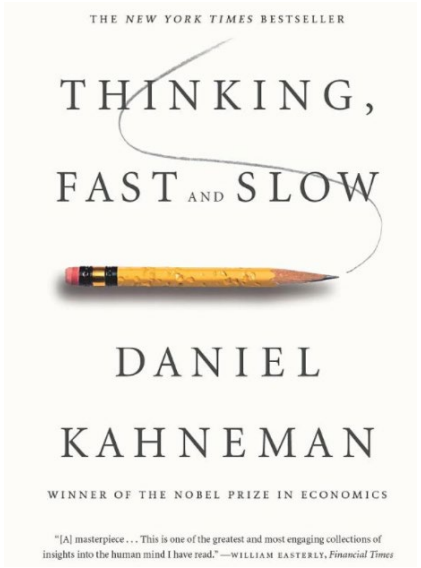
FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION.
BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

NATURE | VOL 526 | 8 OCTOBER 2015

Hardwired for survival

Team of investigators 80,000 BCE



nature human behaviour
PERSPECTIVE
https://doi.org/10.1038/s41562-021-01203-8
Check for updates

Aspiring to greater intellectual humility in science

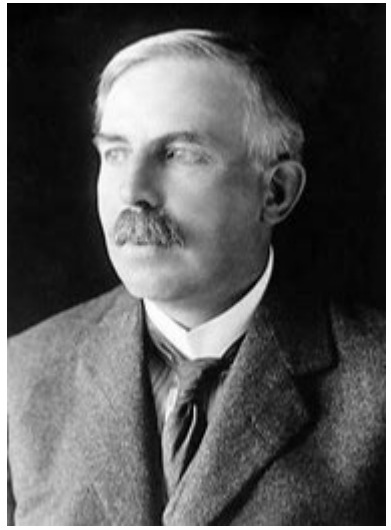
Rink Hoekstra [1,4] and Simine Vazire [2,3,4]

Reproducibility – The Role of Statistical Inference

“To understand God’s thoughts, we must study statistics, for these are the measure of his purpose.”
Florence Nightingale (1820 – 1910)

Pioneer in epidemiological research
Developed methods for the graphical representation of data

"If your experiment needs statistics, you ought to have done a better experiment."
Ernest Rutherford (1871-1937)



Discovered alpha and beta radiation and the atomic nucleus

The Case for Rutherford

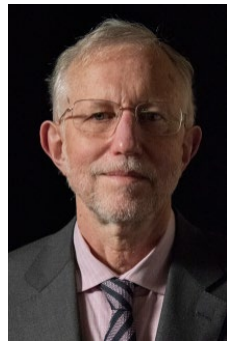
Jennifer Doudna & Emmanuelle Charpentier
Nobel Prize in Chemistry in 2020

A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity

Martin Jinek,[1,2]* Krzysztof Chylinski,[3,4]* Ines Fonfara,[4] Michael Hauer,[2]†
Jennifer A. Doudna,[1,2,5,6]‡ Emmanuelle Charpentier[4]‡

17 AUGUST 2012 VOL 337 **SCIENCE** www.sciencemag.org

No statistical analysis



Charles Rice
Nobel Prize in Physiology or Medicine in 2020

Transmission of Hepatitis C by Intrahepatic Inoculation with Transcribed RNA

Alexander A. Kolykhalov, Eugene V. Agapov, Keril J. Blight,
Kathleen Mihalik, Stephen M. Feinstone, Charles M. Rice*

SCIENCE • VOL. 277 • 25 JULY 1997 • www.sciencemag.org

No statistical analysis

The Case for Rutherford – Statistics and Transparency

How transparent is this?

Neutralising antibody activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 vaccination

*Emma C Wall, Mary Wu, Ruth Harvey, Gavin Kelly, Scott Warchal, Chelsea Sawyer, Rodney Daniels, Philip Hobson, Emine Hatipoglu, Yenting Ngai, Saira Hussain, Jerome Nicod, Robert Goldstone, Karen Ambrose, Steve Hindmarsh, Rupert Beale, Andrew Riddell, Steve Gamblin, Michael Howell, George Kassiotis, Vincenzo Libri, Bryan Williams, Charles Swanton, Sonia Gandhi, *David LV Bauer*

david.bauer@crick.ac.uk

Francis Crick Institute, London NW1 1AT, UK (ECW, MW, RH, GKe, SW, CSa, RD, PH, SHu, JN, RG, KA, SHi, RB, AR, SGam, MH, GKa, CSw, SGan, DLVB); National Institute for Health Research (NIHR) University College London Hospitals (UCLH) Biomedical Research Centre, London, UK (ECW, RB, VL, BW); NIHR UCLH Clinical Research Facility, London, UK (ECW, RB, VL, BW); University College London, London, UK (EH, YN, VL, BW, CSw, SGan); Department of Infectious Disease, St Mary's Hospital, Imperial College London, London, UK (GKa)

www.thelancet.com Vol 397 June 19, 2021

"Statistical significance of the difference in median viral neutralisation IC50 values between different strains was performed using a paired Wilcoxon Ranked sum test

The 95% confidence interval for the difference in median viral neutralisation IC50 between different strains was determined using bootstrap statistics, implemented in R using the *bootstrap* and *boot.ci* functions, using the *type="basic"* argument to avoid any assumptions about the normality of the data.

Correlation analysis of neutralizing antibody titres between different virus strains and between neutralizing antibody titre and age and BMI was carried out using Spearman's test using the *cor.test* function in R

Comparisons of neutralising antibody response by sex were carried out using a paired Wilcoxon Ranked sum test.

p-values reported have not been corrected for multiple testing.

Analysis of stratified neutralising antibody responses by strain for each dose of vaccine was carried out using ordered logistical regression using the *lrm* function of the Regression Modeling Strategies ('*rms*') package in R, using the formula IC50 ~Strain or IC50 ~Strain*Age , and *p*-values were calculated using the Wald Chi-Square test.

Analysis of variance was carried out using the *anova* function in R."

The Case for Nightingale

Scientific endeavors are characterized by unavoidable uncertainty.

Statistical methods are procedures for making decisions in the face of uncertainty.

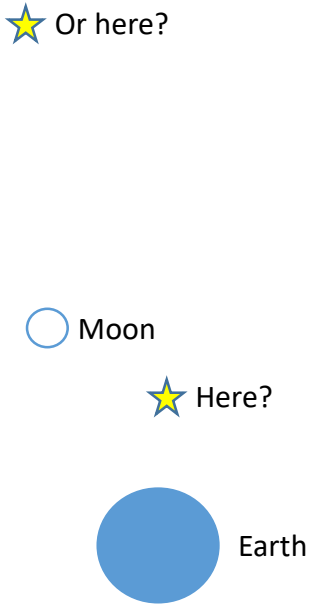
Experimental Error and Biological Variation – A History

“Usually more knowledge can be gained by watching the construction of something than by inspecting the polished finished product”.

Walter Moore

The New Star of 1572

- A new star (now called a supernova) appeared in November 1572 and disappeared 16 months later.
- The prevailing dogma was that new celestial events occurred between the earth and the moon.
- An alternative hypothesis was that the new star was far away, among the “fixed stars”.
- The two competing hypotheses could be tested by measuring the position of the new star in the sky relative to other celestial objects.
- Twelve researchers measured the position of the new star and 12 different results were obtained.



Hald, A *Int. Statistical Rev.* 1986

Location of the New Star of 1572

Galileo did a literature search …

	Observer	Altitude of the pole, x
1a	Tycho	55°58′
1b	Tycho	
2a	Camerarius	52°24′
2b	Camerarius	
2c	Camerarius	
3	Peucer	51°54′
4	The Landgrave	51°18′
5	Reinhold	51°18′
6	Busch	51°10′
7	Gemma	50°50′
8	Ursinus	49°24′
9a	Hainzel	48°22′
9b	Hainzel	
9c	Hainzel	
10	Hagek	48°22′
11	Muñoz	39°30′
12	Maurolycus	38°30′

Dialogue Concerning the Two Chief World Systems, 1632

Galileo's Generalizations of Experimental Error

- There is one number which gives the distance of the star, the true distance.
- All observations are encumbered with errors.
- Observations are distributed symmetrically about the true value.
- Small errors occur more frequently than large errors.

This type of error is now called random error.

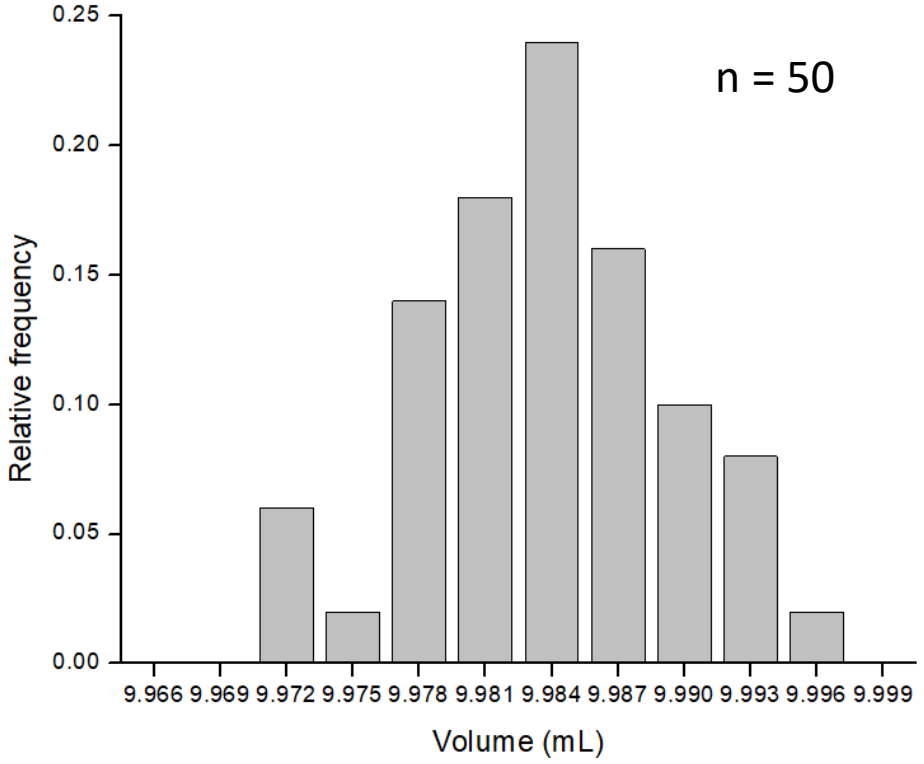
Hald A *Int. Statistical Rev.* 1986

Random Error – An Example

Calibration of a 10 mL pipet

Procedure

- Dispense a nominal 10 mL volume of H_2O
- Weigh the discharged volume
- Use the density of water to calculate the volume.



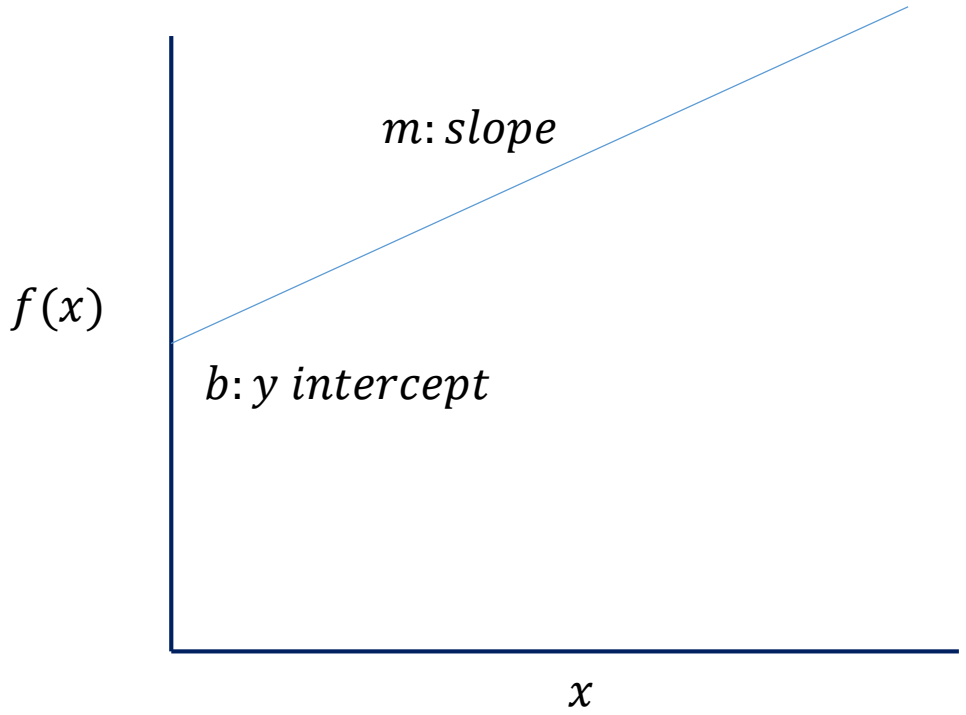
- All the measurements are encumbered with errors.
- The measurements are distributed symmetrically
- Small errors occur more frequently than large errors

Crouch SR & Holler FJ *Applications of Microsoft Excel in Analytical Chemistry*, 2nd Ed.

Math Review - Functions

Parameters

$$y = f(x) = mx + b$$



The Normal Distribution

Gauss

The normal density function was derived by Carl Gauss in 1809.

Assumptions:

1. Observations are distributed symmetrically about the true value.
2. Small errors occur more frequently than large errors.

The best estimate of the true value of the quantity being measured is the mean.

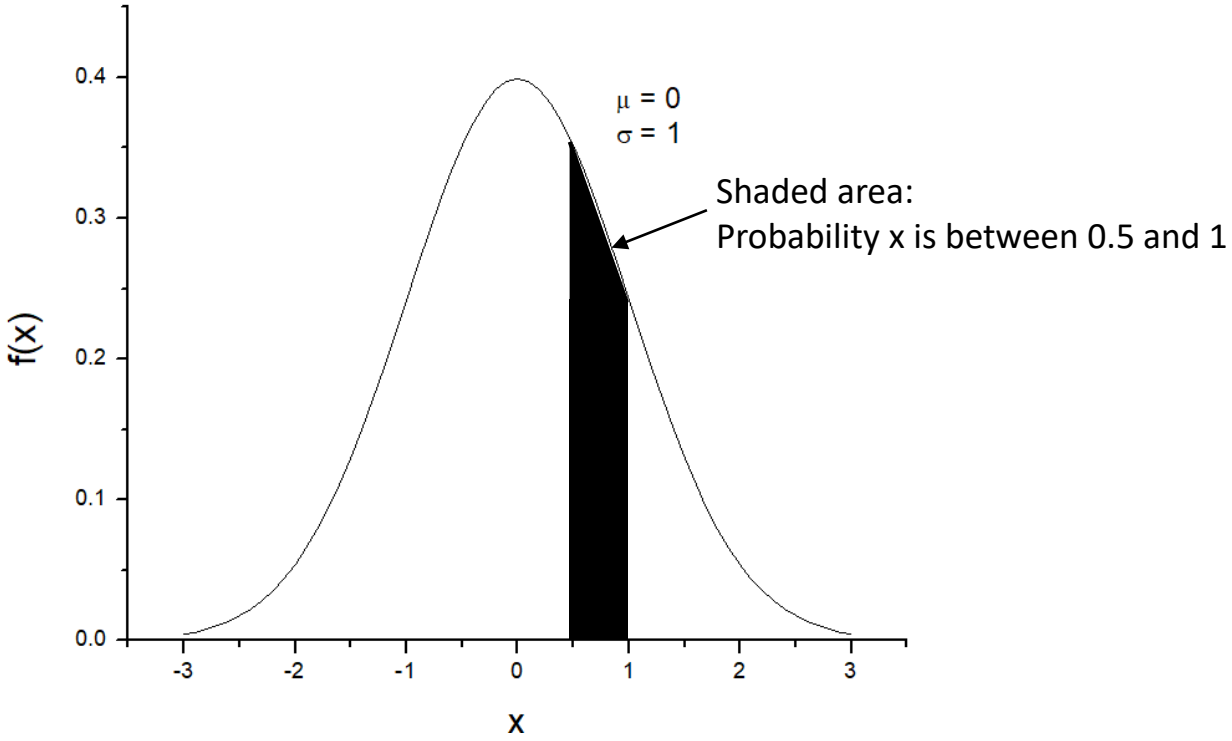
The probability x occurs between two values is the area under the curve between those values.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameters:

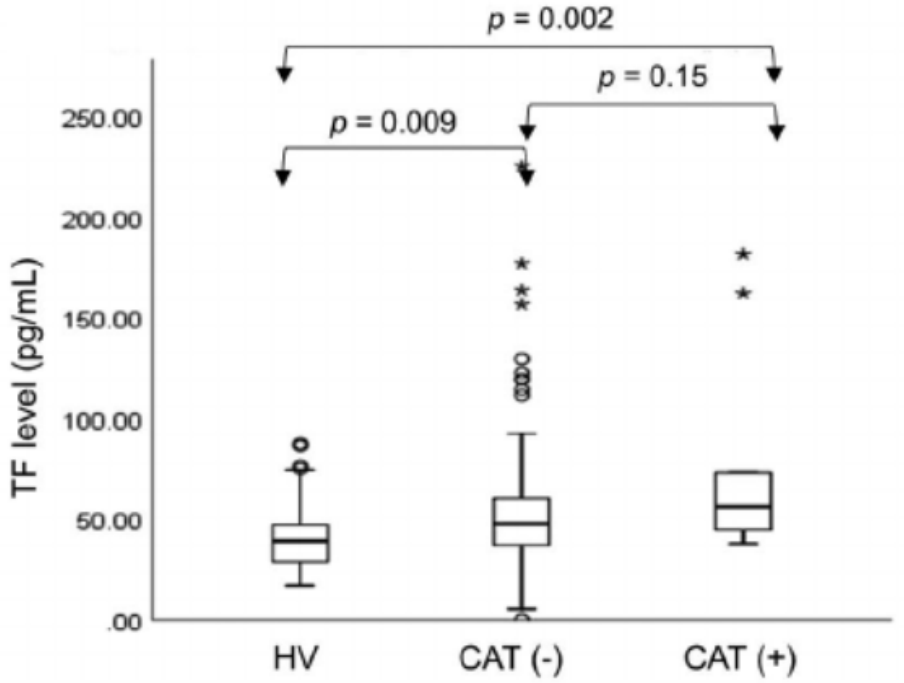
μ Mean

σ Standard deviation



Probability statements regarding the normal distribution and distributions based on the normal distribution form the basis for the “frequentist” approach to statistical inference.

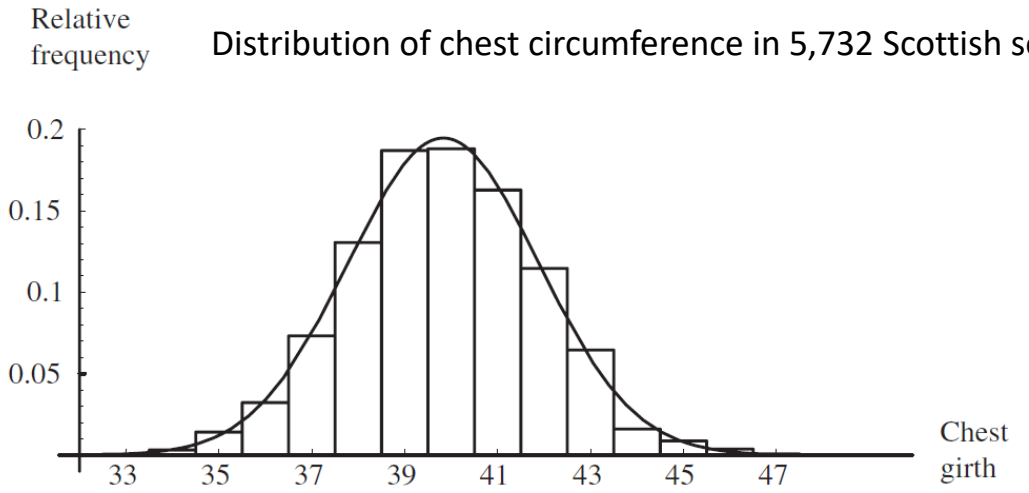
Test for Statistical Significance - Those Ubiquitous Asterisks



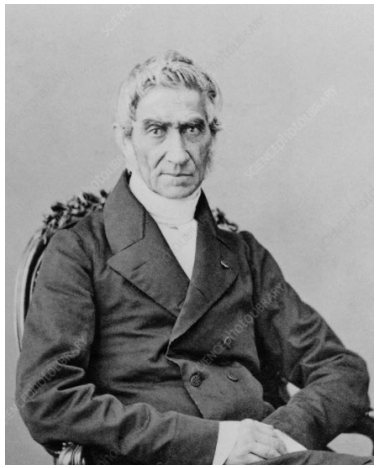
Kobayashi S et al. *Cancer Science* 2021

The Normal Distribution of a Biological Variable

In 1846, Adolphe Quetelet reported that the variation of male chest circumference could be described by a normal distribution.



Distribution of chest circumference in 5,732 Scottish soldiers



Quetelet

Stahl S *Mathematics Magazine* 2006

Why Do Measurements and Biological Variables Tend to Be Normally Distributed?

A measurement is the aggregate of a large number of elementary “errors”

- Electrical noise in instruments
- Pipetting errors
- Combinations of gene expression driving biological traits
- Etc.

The central limit theorem of statistics

> The sample mean of a random variable approaches a normal distribution as the sample size approaches infinity, even if the random variable itself is not normally distributed.

Random Error and Statistical Inference

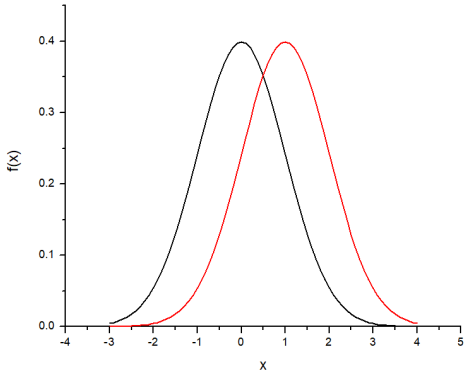
- Scientific truth claims are often based on analysis of random errors.
- This analysis forms the basis of hypothesis testing with p-values or confidence limits.
- The quantitative application of error analysis falls into three broad categories:

1) Making point estimates and assigning confidence limits

E.g., Establishing normal ranges in healthy populations

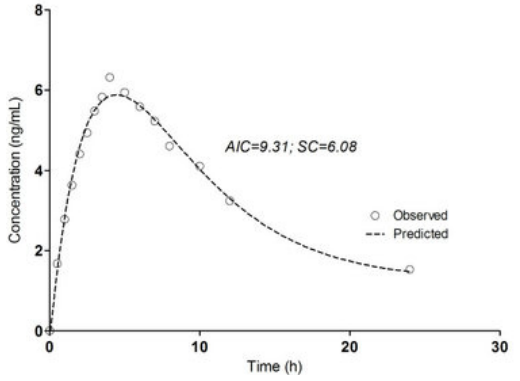
2) Comparing of means of different populations

E.g., Measuring the effects of two anti-hypertensive medications on blood pressure



3) Assessing quantitative models

E.g.,
Fitting drug clearance data to pharmacokinetic models
Fitting antibody – antigen binding measurements to mass action models to estimate binding affinity
Fitting X-ray scattering data to determine crystal structures of molecules

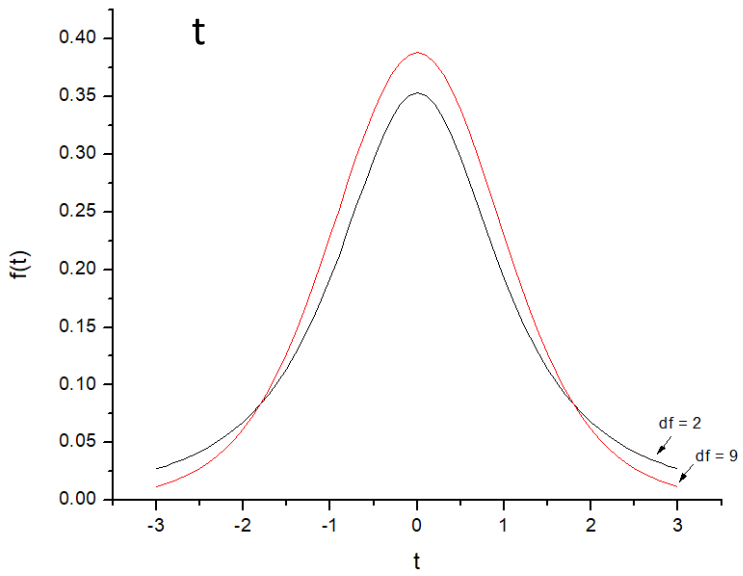


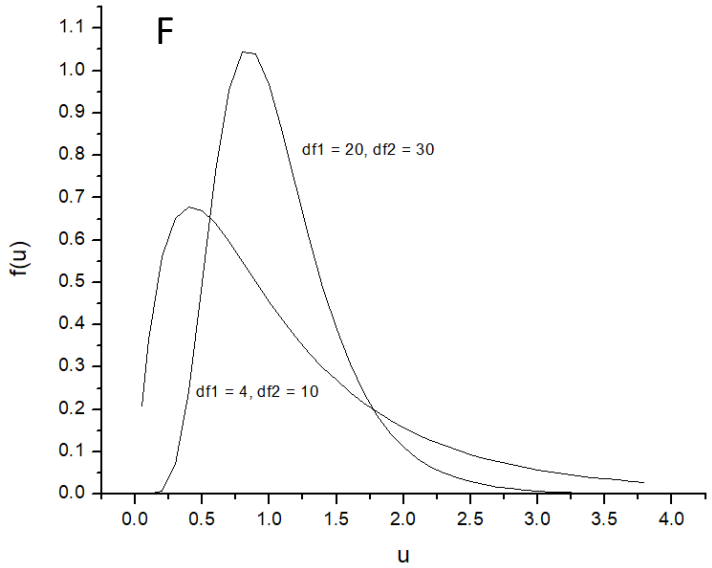
Commonly Used Distributions in Statistical Inference

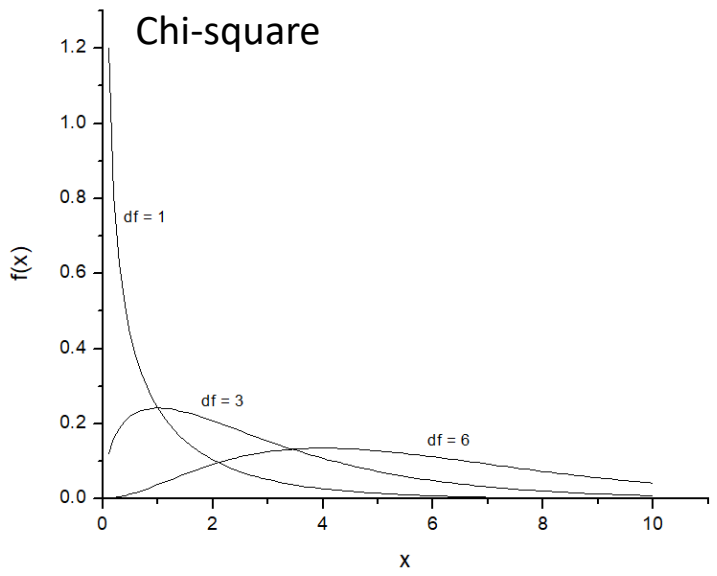
Student's t distribution – point estimates/confidence limits, comparison of two population means

F distribution – comparison of several population means (analysis of variance)

Chi-square distribution – testing goodness of fit of models, contingency tests



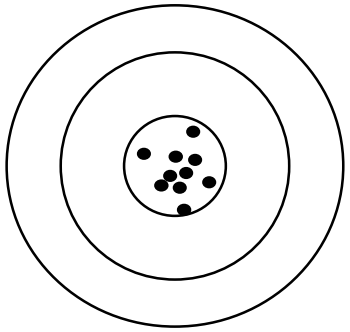




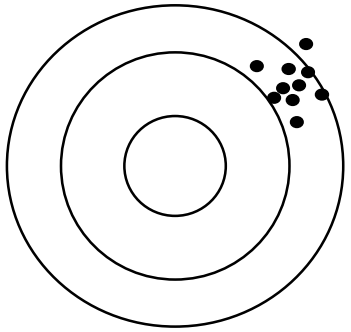
These distributions assume sampling is from a normally distributed population

Non-parametric statistics are applied when the underlying distribution is not normal.

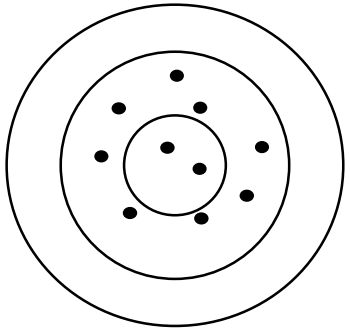
Random *versus* Systematic Error



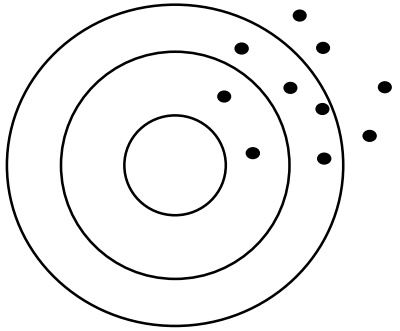
Random: small
Systematic: small
Precise and accurate



Random: small
Systematic: large
Precise, not accurate



Random: large
Systematic: small
Not precise, but accurate



Random: large
Systematic: large
Not precise, not accurate

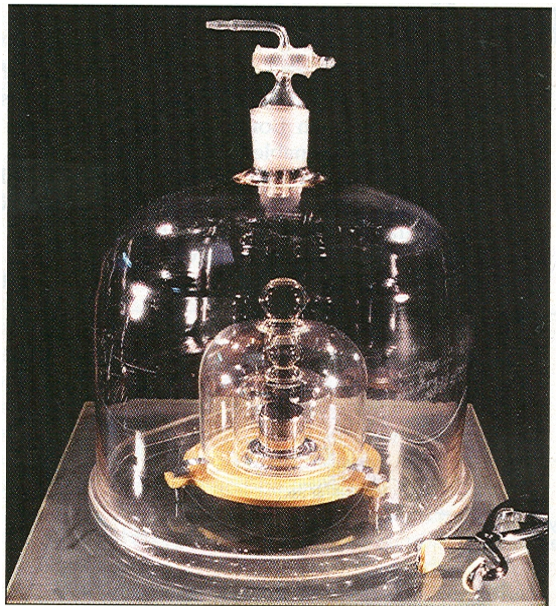
Taylor JR *Introduction to Error Analysis*, 2nd Ed.

Accuracy and Reference Standards

"The evaluation of accuracy in the absence of an absolutely known reference standard is reduced to an educated guess."
Mandel J, *The Statistical Analysis of Experimental Data*

Examples of reference standards:

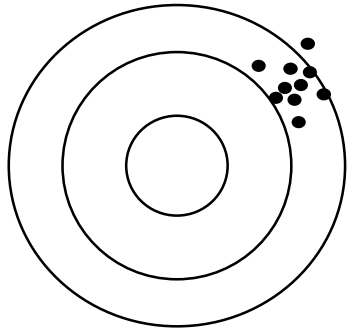
"Le Grand K", Paris
Standard of Mass from 1889 to 2019



U.S. National Institute of Standards & Technology
Calcium Carbonate pH Standard

Systematic Error

- Systematic errors are hard to detect and evaluate.
- In contrast to random errors, there are no mathematical models for systematic errors.
- Many, if not most measurements, are dominated by systematic errors
- Identifying and reducing systematic errors requires experience and judgment.
  - Attempt to identify and estimate the size of possible sources
  - Attempt to compare the results of independent measuring processes
    - E.g., measure mRNA levels by QT-PCR and Northern blotting



There are many causes of systematic error

Examples:

- You multiplied by the number of cells, but should have divided.
- An assay variable is pH-dependent and you used an incorrectly calibrated pH meter.
- You measured weight loss or gain in an animal experiment and forgot to zero the balance.
- You used tap water in your cell culture experiment.

Confidence Limits

Suppose a group of 4 mice is subjected an experimental diet and weight gain is measured after 3 months.

Weight (g)
+1.25
+1.26
+1.28
+1.29

What can we say about the average weight gain in the population of mice represented by this sample?

Assuming weight gain in the population is normally distributed, Student's t distribution provides an estimate of the population mean and confidence limits for this estimate.

Result: 1.27 ± 0.03 g (mean ± 95% confidence limits)

- Confidence limits pertain to the precision and random error, not the accuracy and systematic error, of measurements.
- Confidence limits should be used, if possible, instead of "±" or standard deviation.

Confidence Limits – A Tricky Concept

Correct:

Based on this single experiment, the claim is made that 1.27 ± 0.03 g contains the population mean. If for each such experiment the same claim were made for the interval corresponding to that experiment, then 95% of such claims would be true in the long run.

Incorrect:

The probability is 0.95 that the average weight gain in this population is 1.27 ± 0.03 g.

Hoel PG *Introduction to Mathematical Statistics*, 4th Ed.

What is a Population?

In statistics, the totality of all possible outcomes is called the population of outcomes.

We can never know the parameters of a population, we can only estimate them.

Misunderstanding or overinterpreting the generalizability of the population being sampled is arguably more common than random or systemic error in the development of false scientific claims.

What is the Population?

Mice are sensitive to minor changes in food, bedding and light exposure.

REPRODUCIBILITY

A mouse's house may ruin studies

Environmental factors lie behind many irreproducible rodent experiments.

264 | NATURE | VOL 530 | 18 FEBRUARY 2016

Mouse populations vary from facility to facility.

In a phase III clinical trial of a drug, only patients who meet the study entry criteria and give informed consent are enrolled. How does this population differ from the population that will receive the drug if it is approved?

The "heterogeneity curse" – inter-individual differences may invalidate the assumption that clinical trial outcomes can be judged by significance testing of average values.

Hanin L *BMC Medical Research Methodology* 2017

Bayesian Inference

Does a subject with this EKG finding have coronary artery disease?

Bayesian Inference

The probability of disease given a positive test for that disease depends on the prevalence of the disease in the population being studied.

This prevalence is called the prior probability.

Bayesians versus Frequentists

There is a fundamental disagreement between frequentists and Bayesians regarding the definition of probability.

For frequentists, probability only has meaning in terms of a limiting case of repeated measurements.

For Bayesians, probabilities are fundamentally related to their own knowledge about an event. They often proceed with only a guess about the prior probability.

Developing Statistical Literacy
Master of Science in Clinical Research Courses at Emory

- Introduction to Clinical and Translational Research
- Analytic Methods for Clinical and Translational Research I
- Analytic Methods for Clinical and Translational Research II
- Introduction to Biostatistics
- Data Management
- Analysis of Clinical Research Data
- Fundamentals of Bioinformatics
- Clinical Trial Design
- Big Data to Knowledge (BD2K) in Clinical and Translational Research
- Advanced Data Management in R

www.actsi.org/training/ms-in-clinical-research/mscr-courses.html

Statistical Training for Laboratory Based Investigators at Emory

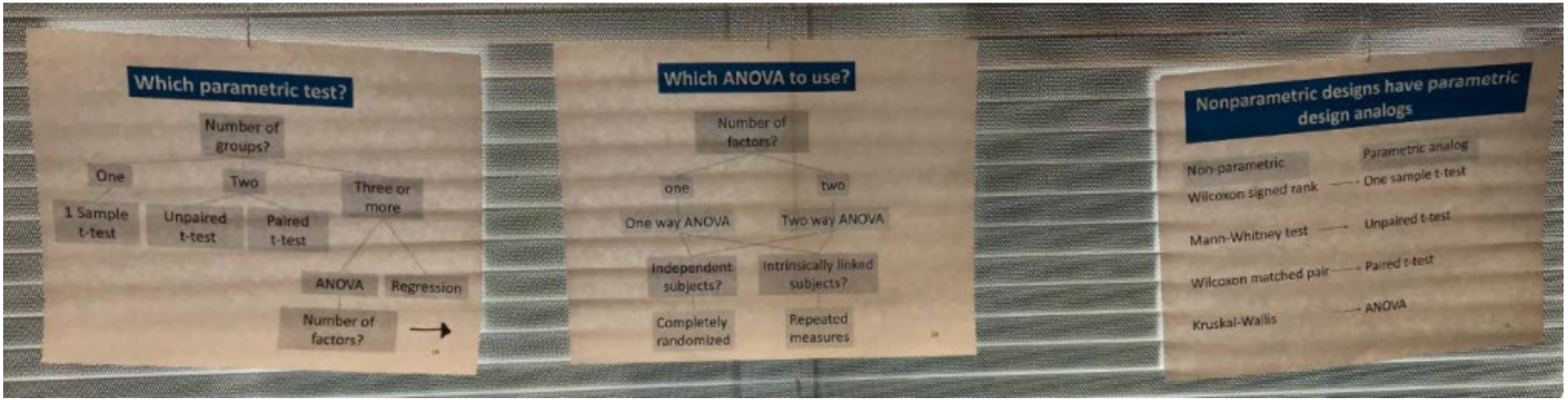
Using Statistical Methods

Use statistical methods as a tool you have no hope of understanding

Don’t look under the hood!

We don’t need to know automobile mechanics, we need to need to know how to use a car as a tool.

Developing Statistical Literacy



Photograph of the window blinds, 4th Floor, Emory Children's Center

Developing Statistical Literacy

The Journal of Biological Chemistry - Instructions to Authors:

“Resources and guides on statistical analyses may be found on the GraphPad website.”

http://jbcresources.asbmb.org/collecting-and-presenting-data

Developing Statistical Literacy

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

GAISE College Report ASA Revision Committee

http://www.amstat.org/education/gaise

JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

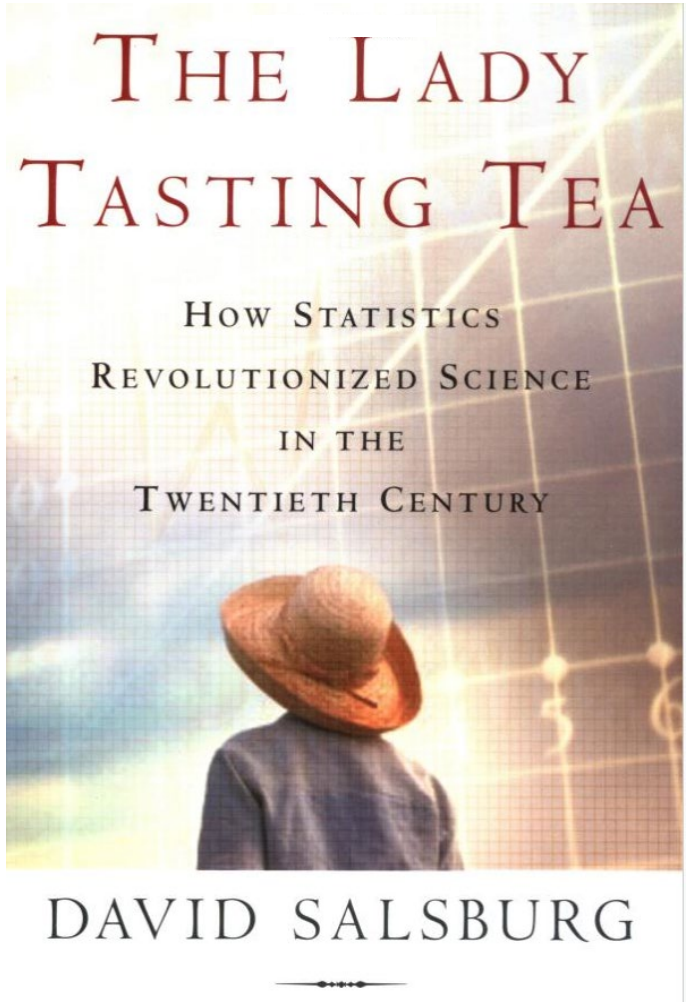
A Statistics Curriculum for the Undergraduate Chemistry Major

Nicholas E. Schlotter*

Department of Chemistry, Hamline University, St. Paul, Minnesota 55104, United States

J. Chem. Educ. 2013, 90, 51−55

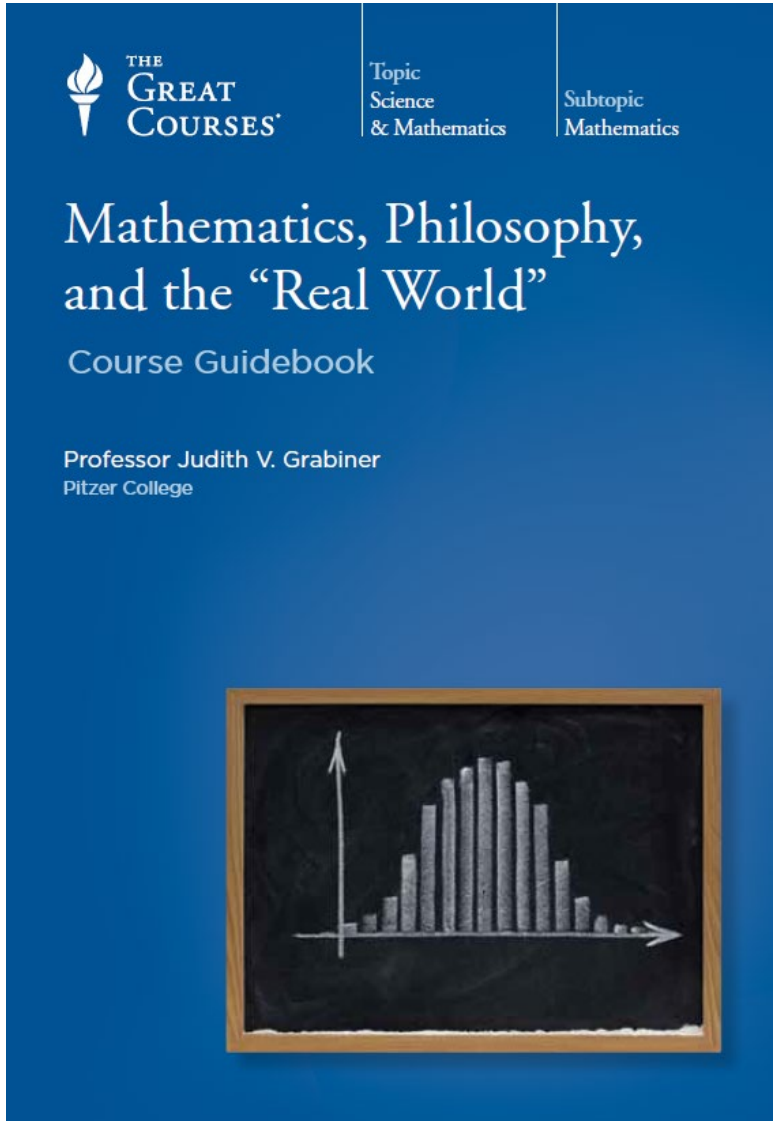
Developing Statistical Literacy



Stories about:

Cox
Fisher
Gosset ("Student")
Kolmogorov
Mann & Whitney
Mantel
Markov
Neyman
Pearson, E.
Pearson, K.
Tukey
Wilcoxon
and others

Developing Statistical Literacy



Summary – The Yin and Yang of Scientific Research

Most research findings are false	Some research findings are true
We need the truth, now	It will all come out in the wash.
Requirements for rigor should be established	Rigor requirements are stifling, restrictive and expensive
Literature lasts a long time – be careful what you publish	Perfect is the enemy of the good - Anthony Fauci
Statistics are the measure of God's purpose	Incisive experiments do not require statistics
Study the mathematical foundations of statistical inference	Do what Prism tells you to do
The path to enlightenment is through your local biostatistician	Controversies exist in statistics
In medicine, the problem comes to you	In research, you go to the problem
Attack important problems and develop the tools to do so	Find problems within an established scientific discipline