

Authors: The Anesthesia & Analgesia Statistical Guide for Authors is being written. This PowerPoint, “Common Statistical Errors”, is a placeholder to provide information on the contents of the Common Statistical Errors” document that will be posted, hopefully by September, 2007.

Mistake 1: Not having an hypothesis

- Statistical tests are used to distinguish between two hypotheses:
 - Null hypothesis: two groups are the same
 - Experimental hypothesis: two groups are different
 - Start with a basic hypothesis in mind
 - Avoid fishing expeditions except to generate hypotheses

Mistake 2: Having a zillion hypotheses

- An experiment where you plan to gather a truckload of data, and analyze the hell out of it.
 - “There’s got to be a pony in here somewhere.”

Data Dredging

- The digits of π are random:
 - 3.1415926535897932384626433832795028841
97169399375105820974944592307816406286
2089986280348253421170679
- But, in the first 100 digits, there is a sequence of 10 consecutive even numbers.
- The odds state that 10 consecutive even numbers should show up only once in every 2^{10} (1024) numbers.

Data Dredging: π

- The next sequence of 10 even numbers occurs at 1279, and is a sequence of 14 numbers, which should only occur once in every 16384 numbers.
- You can always find “non-random” patterns in random sequences. That is the very nature of randomness.

Problems With Multiple Comparisons

- Bonferroni Correction (very conservative)
- Used for the “global null” hypothesis, when multiple independent tests signify success
 1. Nausea < control at 6 hours OR
 2. Vomiting < control at 6 hours OR
 3. Rescue < control at 6 hours OR
 4. Nausea < control at 12 hours OR
 5. Vomiting < control at 12 hours OR...

Correct For Multiple Comparisons

- Choose test appropriate for design:
 - all possible comparisons (Newman-Keuls)
 - comparison to control (Dunnett)
 - specific comparisons (Bonferroni)
- If aim is to show “no difference”, consider not correcting

Mistake 3: No Power Analysis

- A power analysis is intended to help YOU avoid wasting your time by studying too few, or too many, patients.
 - This is a sample size calculation.
- There is no point in doing a power analysis after you have done a study. Particularly if you have a positive result. You have whatever you have.

Post Hoc Power Analyses

Reviewer: Your study came close to achieving statistical significance ($P = 0.06$). Please do a power analysis.

My reply: Why bother? The power analysis would surely demonstrate that my sample was *slightly* too small!

It may be useful to guide future study design, so you can do it as a public service.

Mistake 4: Not Graphing Your Data

- If a finding is real, there should be some way of presenting the data to make it clear visually
- Before you start with any statistical tests, examine the raw data
- You may identify incorrectly entered data points, points with the wrong units, etc.

Mistake 5: Confusing statistical significance with clinical significance

- If a reviewer asks for a power analysis, the reviewer is actually asking for a confidence limits, but may not know it.
- Confidence limits help you understand the difference between clinical and statistical significance

Confidence Limits

- The 95% boundaries of the finding of the study
- The important question is the relationship between
 - confidence limits
 - 0 (*not interesting!*)
 - and a clinically important difference

Confidence Limits

L — X — U



Clinically
Significant
Difference

Confidence Limits as “Power Analysis”

L — X — U



We can say with 95% probability that the difference between the groups is between L and U.

Scenario 1: Worse Possible Outcome



↑
0

↑
D (clinically significant difference)

The most probable location of the difference is X. ALL OTHER THINGS BEING EQUAL, then it would still make sense to go with the group that gave the better clinical result.

Scenario 2: Ideal Negative Study

L — X — U

↑
0

↑
D

No significant difference from 0, and can rule out clinically significant benefit at $p < 0.05$.

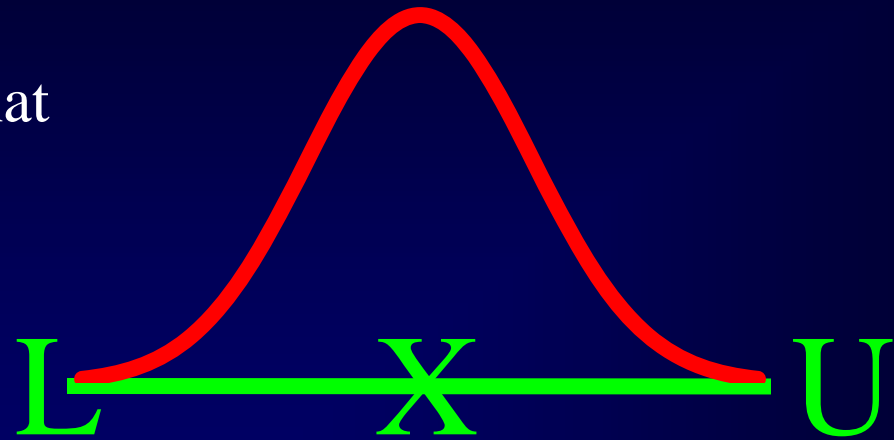
Scenario 3: Statistical vs. Clinical Significance

L — X — U

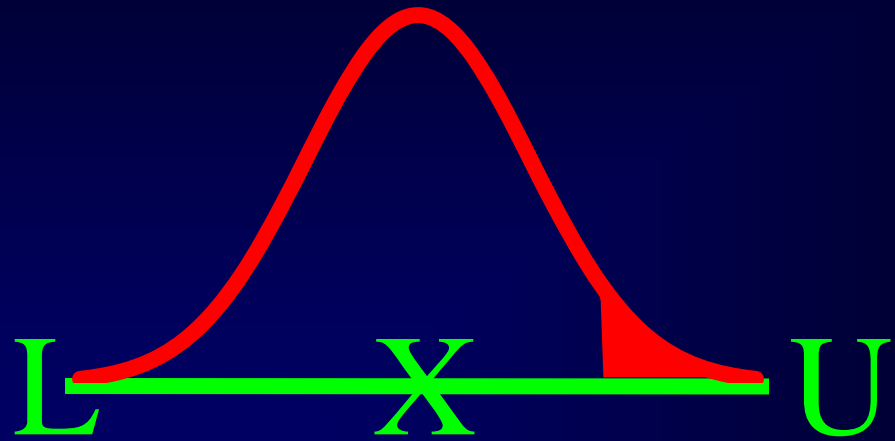


Most studies would trumpet these results showing that the groups are truly different. True! However, the study has not shown that the difference is clinically significant.

What is the probability that
the difference is greater
than D?



The probability is the
integral.



0

D

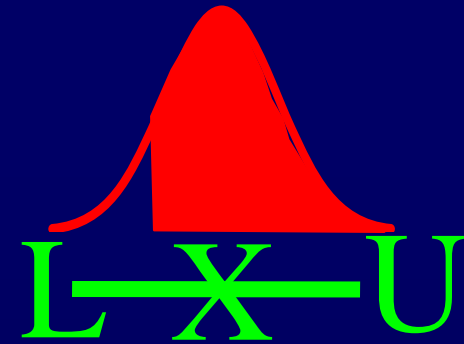
Scenario 4: Statistical vs. Clinical Significance

$L - X - U$



Same question as before. Most studies would call this a real success! $P \ll 0.05$ that the groups are different, and the mean (X) is greater than D . However, X is not $> D$ at $p < 0.05$!

Scenario 4: Statistical vs. Clinical Significance



↑
0

↑
D

The integral calculates the probability that the difference exceeds D.

Scenario 5: Statistically Significant, Clinically Insignificance

~~L-X-U~~

↑
0

↑
D

Statistically significant difference, at $P < 0.05$.

Clinically insignificant difference, at $P < 0.05$.

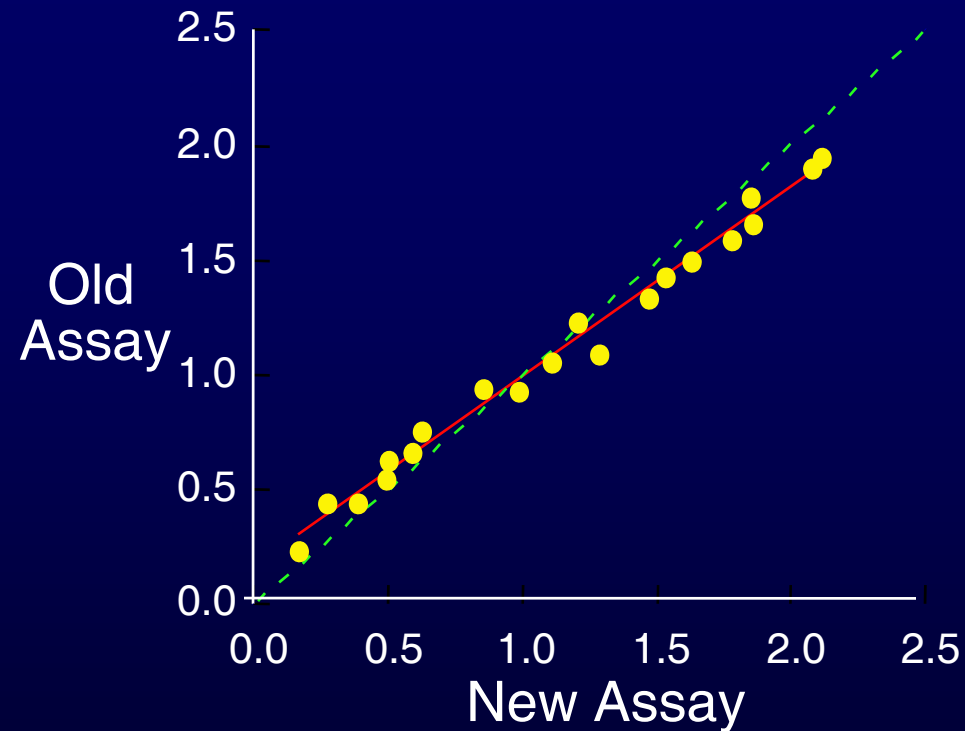
Scenario 6: Clinically Significant Result

~~L-X-U~~

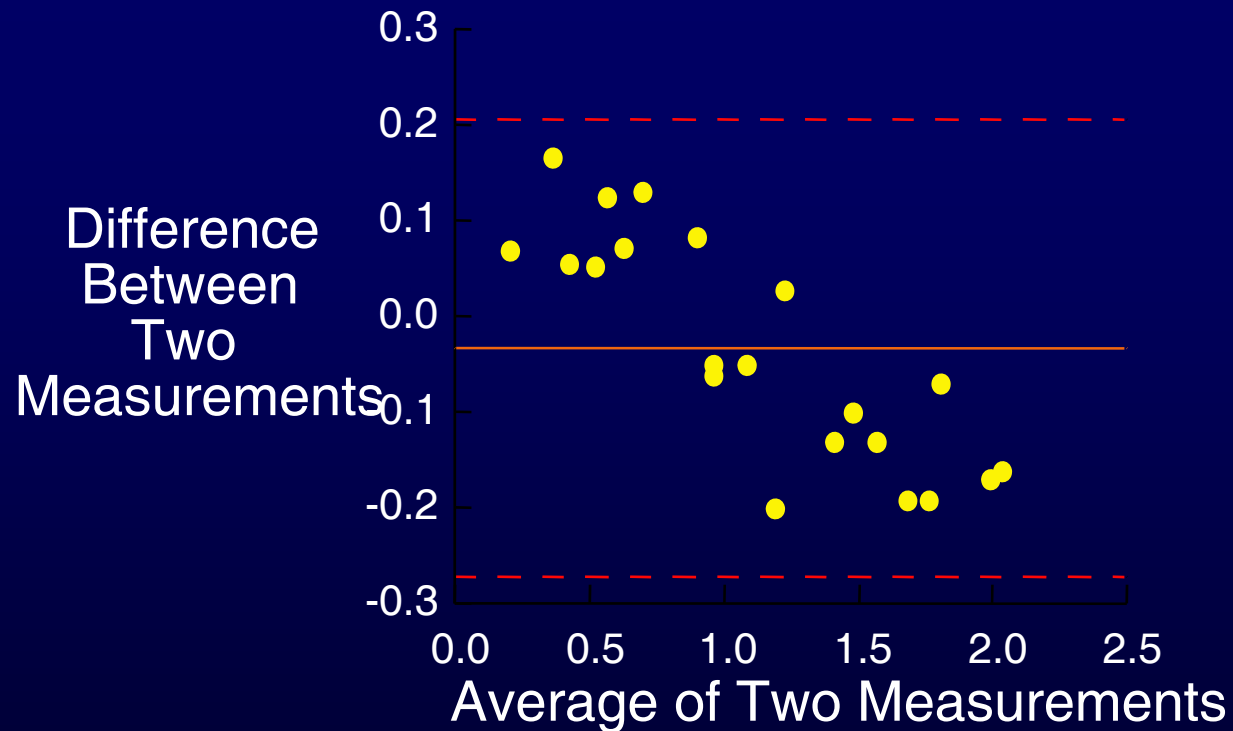
↑ ↑
0 D

Clinically significant
difference, at $p < 0.05$.
The ideal result!

Mistake 6: Inappropriate use of regression



Preferred Method: Bland-Altman Plot



Bland , Altman . Lancet i:307-310, 1986

Mistake 7:

Reporting meaningless values

- Values obtained in a study:
1, 2, 3, 4, 5, 6, 7, 8, 1000000
- Mean: 111111
Only meaningful if data are normally distributed
- Solution: Report median

Example

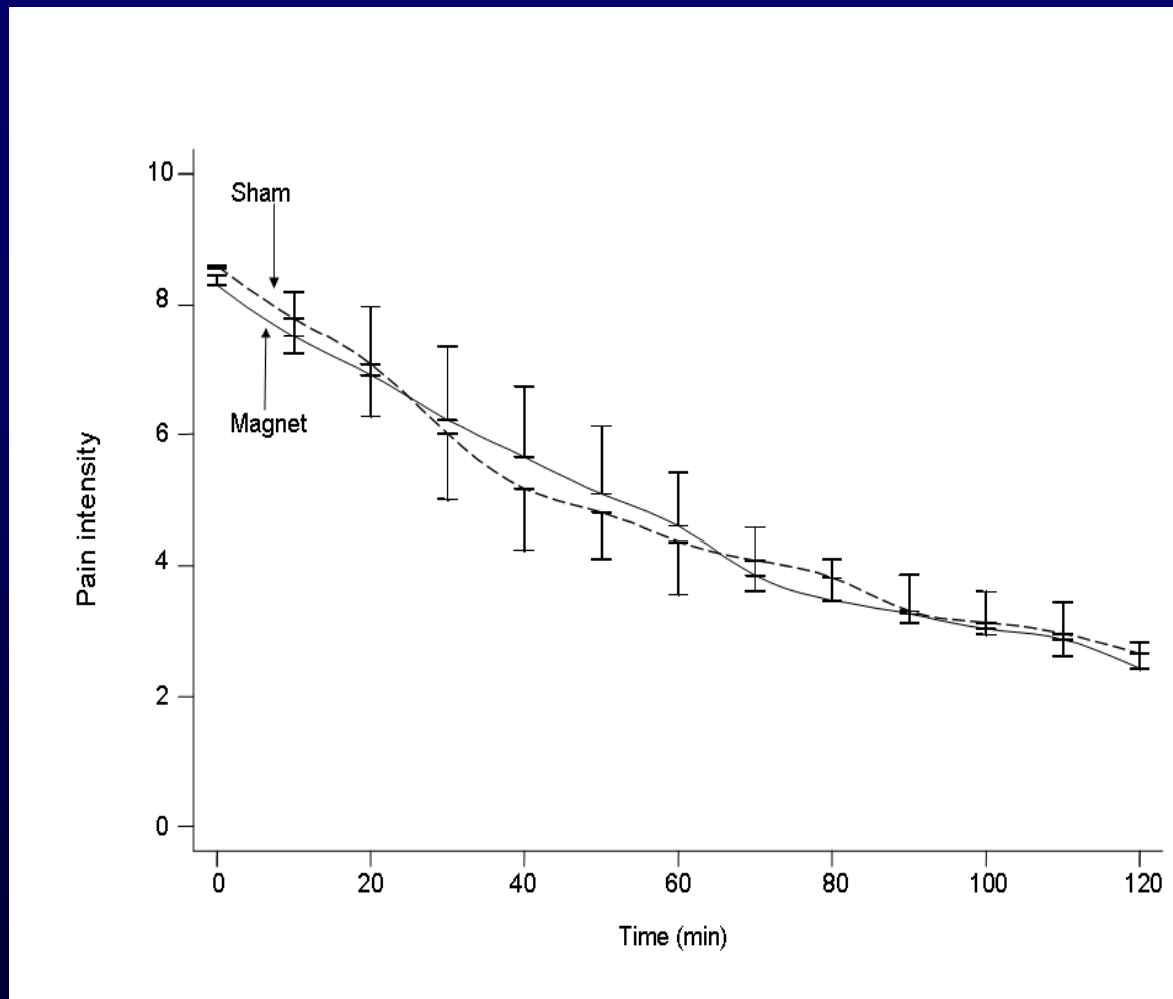
- Trans-carotid emboli post aortic cross-clamping
- Mean \pm SD reported as 100,000 \pm 300,000
- Nine patients had < 10 emboli; one had 1,000,000
- Is the mean a useful descriptor?

Mistake 8:

Confusing SD and SE

- If you are describing a characteristic of a population (height, weight, age), the reader will want to know the variability of that characteristic
 - Standard deviation describes this
- If you comparing the mean of two populations, then the reader needs to know the variability of the mean
 - Standard error describes this
- Report statistic appropriate to the question being asked

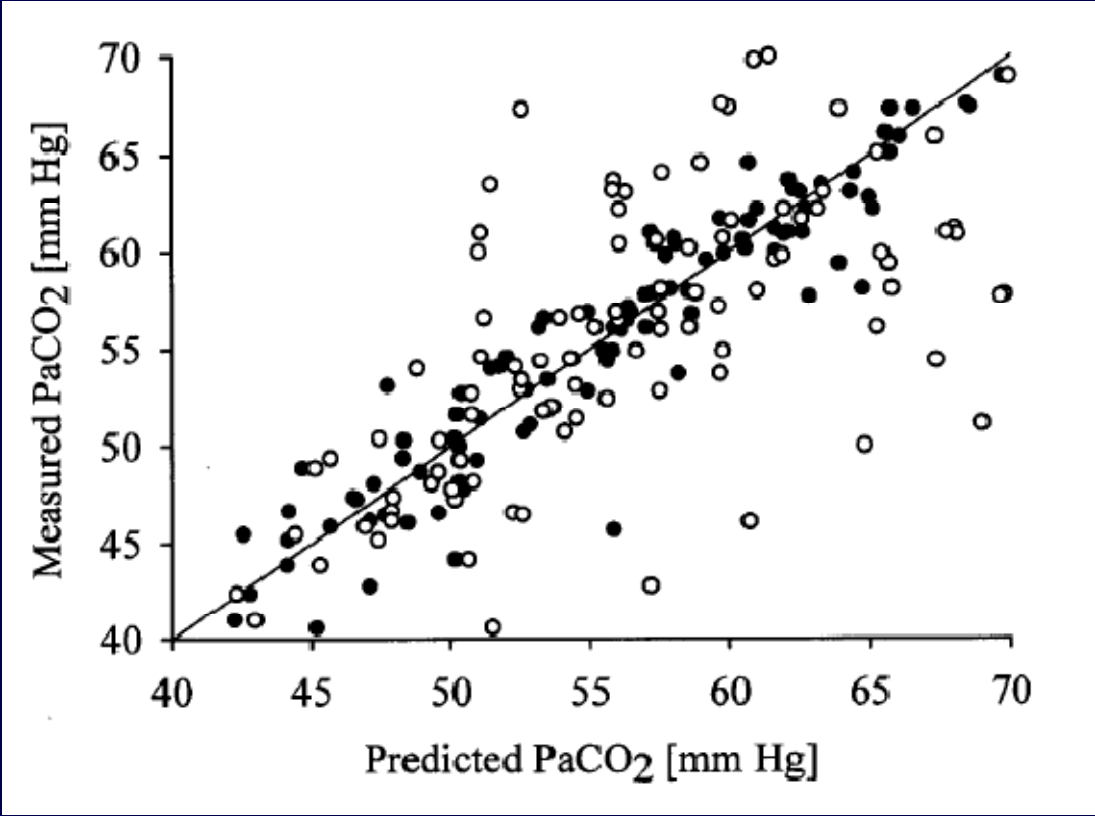
Magnetic Analgesia



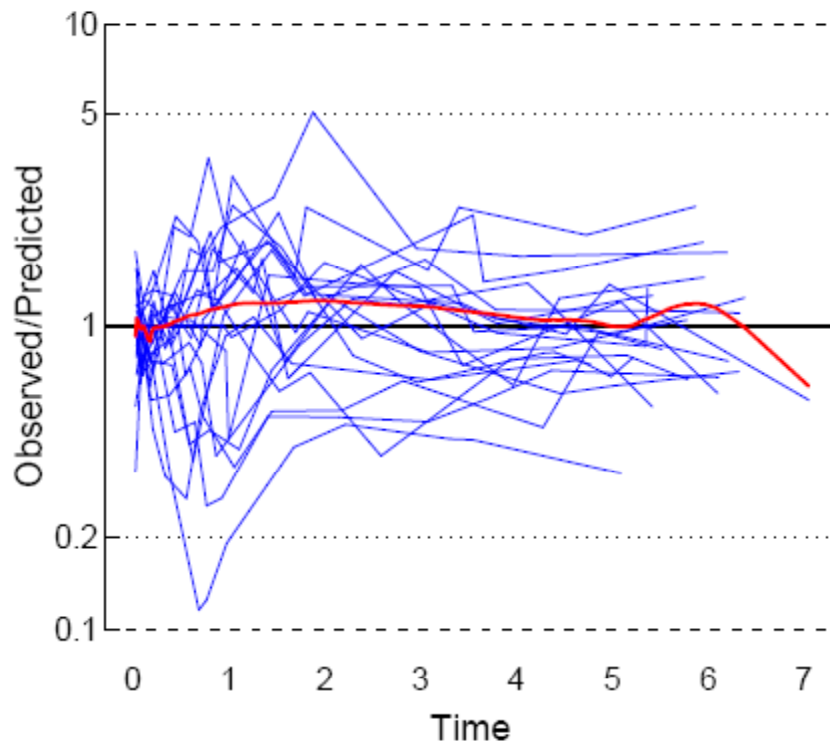
Mistake 9:

No Analysis of Goodness of Fit

- If you create a model, show how well it fits the data
- Best/Median/Worst case examples
- Measured (Y) vs. Predicted (X)
- Errors vs. time

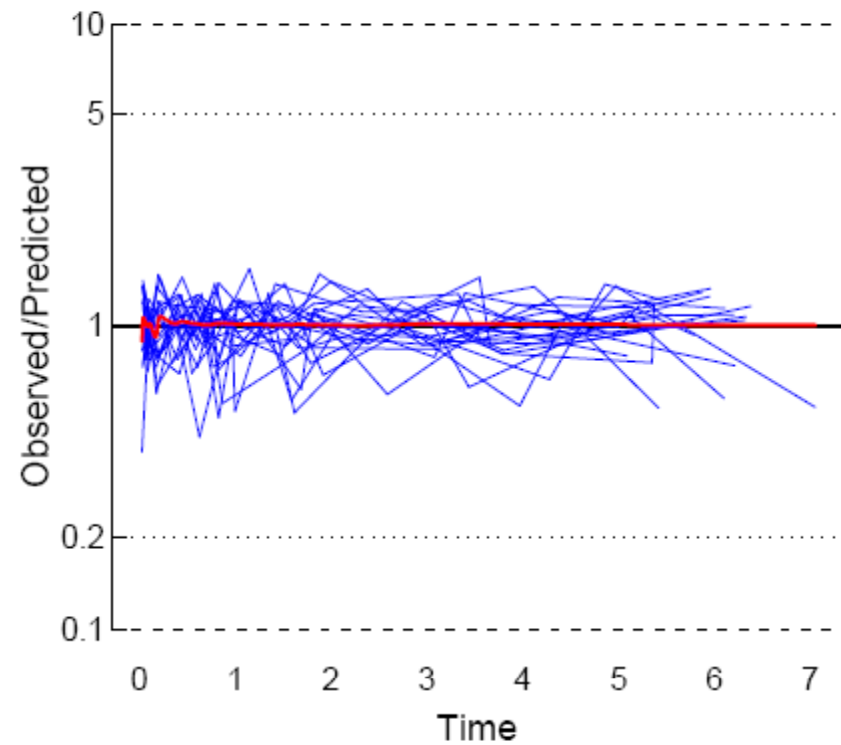


Population



MDPE = -0.047
MDAPE = 0.310

Post Hoc



MDPE = -0.008
MDAPE = 0.158

Mistake 10:

Failure to consider prior probability

- The prior probability has an enormous influence on the posterior probability that a claim is true.

Reverend Thomas Bayes

- 1701-1761
- Presbyterian minister and mathematician
- Published his theorem for probability 2 years after his death.
- His intentions remain a mystery.



Bayes Theorem

Posterior

Prior

Quality of the
Evidence

$$p(H | E, I) = p(H | I) \times \frac{p(E | H, I)}{p(E | I)}$$

Probability of
hypothesis H, given
evidence E, in the
setting I

Probability of
H in this
setting

Probability of
the evidence,
were H true in
this setting

Bayesian Statistics

- The prior probability that something is true has a huge influence on the posterior (i.e., after the fact) probability
- Extraordinary claims demand extraordinary proof

Mistake 11: Using Overly Complex Statistics

- Most issues can be addressed with simple statistics
- Beware of complicated statistics

Mistake 12: Inappropriate Parametric Statistics

	Group 1	Group 2
	1	10
	2	100
	3	1000
	4	10000
Mean:	2.5 ± 1.3	2777.5 ± 4836

Parametric: $p=0.27$

NP: $p = 0.027$

Mistake 13:

Doing Statistics Yourself

- Find a statistician you can work with, particularly for complex problems.
 - Anesthesia & Analgesia is proactive in helping authors get statistics correct.



Two Statistical Approaches

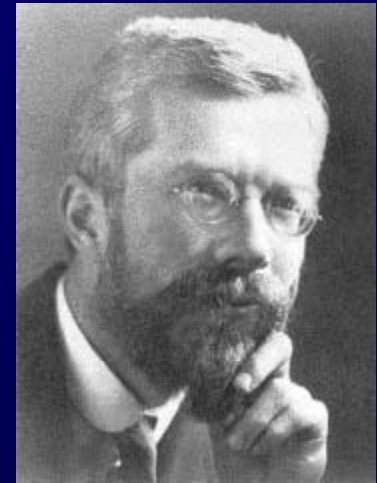
- Fisherian
- Bayesian

Fisherian Inference

- Frequentist approach
 - Probabilities are calculated by counting event frequency
- Distributions in populations are calculated from distributions in samples
- Basis of most parametric and non-parametric statistics
- Unintuitive

Sir Ronald A Fisher

- 1890-1962
- Primary interest was statistics as it applied to genetics and evolution
- Pioneered Analysis of Variance, Maximum Likelihood, F distribution,
- Co-developed t test with Gosset



Probability That Second Trial Would Find $P < 0.05$ Effect if Second Trial Were Identical to First Trial

<u>P Value in First Trial</u>	<u>Probability of $P < 0.05$ in Second Trial</u>
0.10	37%
0.05	57%
0.01	73%
0.005	80%
0.001	91%

Bayesian Statistics

- Nothing is known with certainty
- Even our measures have error
 - How many patients in a nausea study actually had nausea?
 - How accurate are the VAS scores used in pain studies?
 - Do you believe every blood pressure you measure in the OR? Every pulse oximeter reading? Every BIS?