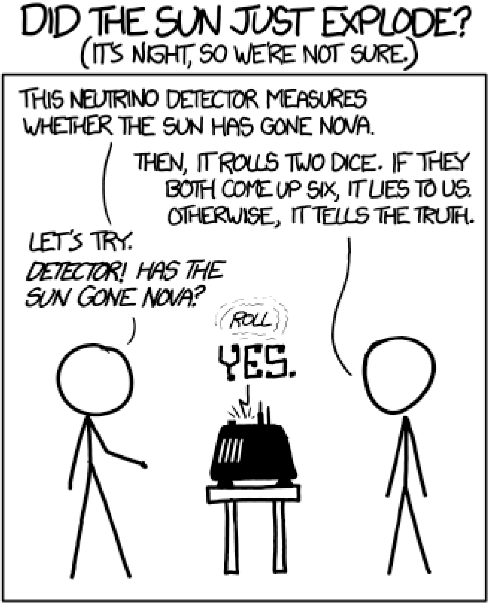
Problem-specific Modeling and Statistics Principles and Examples

Dan Ruan

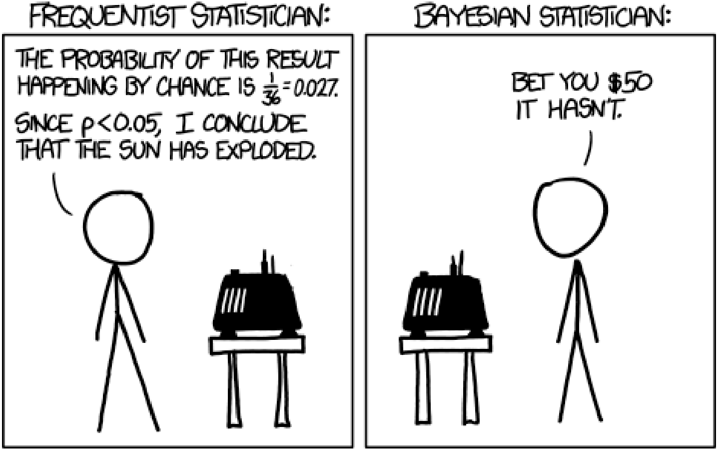
University of California, Los Angeles

AAPM Annual Meeting 2013

Outline



comic source: http://xkcd.com/1132/



frequentist $P(D|H)$ Bayesian $P(H|D)$

comic source: http://xkcd.com/1132/

Outline

1. Following the last talk: Bayesian
2. Lies, damned lies and statistics
   - Causation vs Correlation
   - Post Hoc
   - Average vs. spread
   - Problem with Estimation
   - Statistical significance
3. Something special
   - The special role of Poisson
   - What is so special about Gaussian?
4. Summary

Lies, damned lies and statistics

A little philosophical discussion:

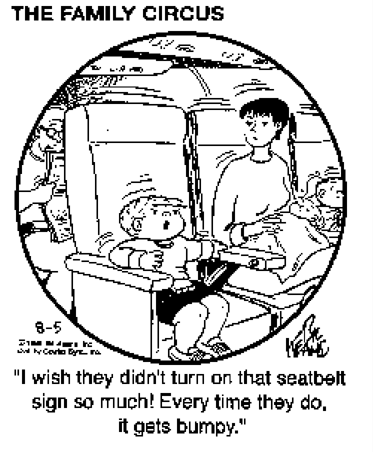
- Aim of scientific communication: prove a certain truth or convince a group of people of a specific viewpoint
- Mathematics is often the authoritative communication tool used for explanations.
- Statistics is the primary tool used in supporting evidence.
- Prevalance in practice vs. relatively low level of formal (re)training $\Rightarrow$
  - for some to manipulate presentation of data or stat methods themselves to sell a particular agenda
  - subject to incidental misuse

"I can see only one thing missing in your report, you've overlooked the problem we're trying to solve."

Image from Morris Business Cartoons

Causation vs Correlation

- Correlation: two variables occur together as observed naturally
- Causation: a first and a second event, where the 2nd event is understood to be a consequence of the first.



For any correlated events A and B, the followings are possible

- $A \rightarrow B$
- $B \rightarrow A$
- $C \rightarrow A$, $C \rightarrow B$
- no connection, coincidental correlation.

Post Hoc

A fallacy involving a special case of correlation

- A occurs before B
- Conclude: A is the cause of B

Example:

Rooster crows (A) before dawn occurs (B), therefore rooster (A) caused dawn (B) to occur.

Post Hoc

A fallacy involving a special case of correlation

- A occurs before B
- Conclude: A is the cause of B

Example:

The anti-vaccination argument:

Children receive their 12-month vaccination , eg., MMR (A) and then develop autism symptoms(B), it would be tempting to ascribe the symptoms to the vaccination.

- Psychologically, people incline to do so for a related children, looking for a cause outside genetic predisposition.
- This instinct-based association often has negative ramifications of understanding data presented to us.

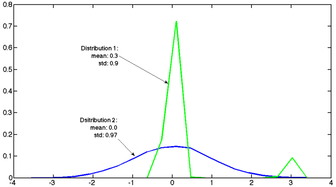
Story line behind this particular anti-vaccination panic:

- 1998 study by Dr. Andrew Wakefield, misrepresented or altered the medical histories of all 12 of the patients
- anti-vaccination panic:
    - Vaccination rates dropped sharply in Britain after its publication, falling as low as 80% by 2004
    - In the United States, more cases of measles were reported in 2008 than in any other year since 1997
- May 2011, Britain stripped Wakefield of his medical license in May
- "Meanwhile, the damage to public health continues, fueled by unbalanced media reporting and an ineffective response from government, researchers, journals and the medical profession," BMJ states in an editorial accompanying the work.

Question to audience: how are we to identify/avoid this?

Average vs. Spread

- Generally in biostats/epidemiology: a fallacy of claiming that the average (say, intelligence or prevalence of crime) is different a particular group A compared with another group B, therefore, it is reasonable to treat individuals in group A as if they, say, had lower intelligence or higher prevalence of crime.
- In Med. Phys. relates to measurement discrepancy.



- Distribution 1: $0.9\mathcal{N}(0, 0.1) + 0.1\mathcal{N}(3, 0.1)$
- Distribution 2: $\mathcal{N}(0, 1)$

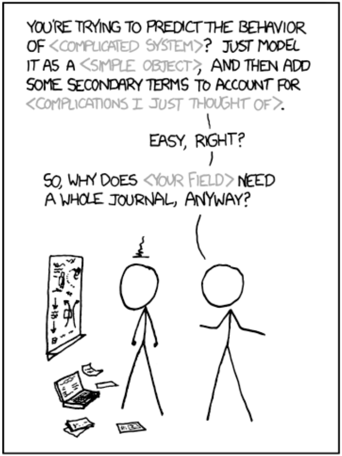
- A higher mean does not necessarily mean a "higher" group
- Even a similar spread may still insufficient to make such claims.

What should we examine, then?

- Most of the time, we may assume certain distributional forms, though we may not have access to the parameters.
- What to examine? How do we characterize

Sufficient Statistics: roughly, given a set X of independent identically distributed data conditioned on an unknown parameter θ , a sufficient statistic is a function $T(X)$ whose value contains all the information needed to compute any estimate of the parameter.

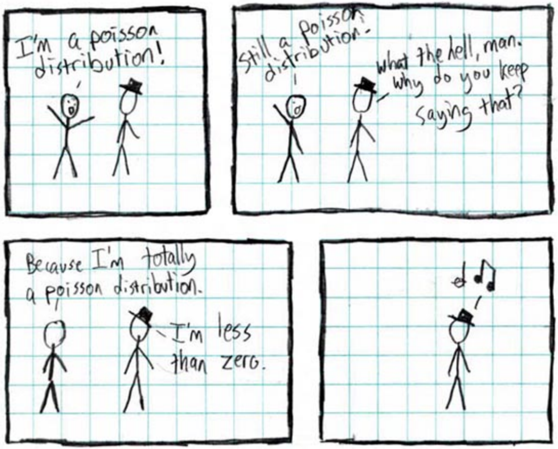
- Bernoulli distribution (failure/success trial),
  - $B(n,p)$, probability of K successes in the experiment $B(n,p)$: $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$.
  - $T(X) = \sum_{i+1}^{N} X_i$
- Uniform distribution
  - uniform prob $\frac{1}{\beta-\alpha}$ within interval $[\alpha, \beta]$, zero otherwise.
  - $T(X) = (\min_i X_i, \max_i X_i)$.
- Poisson distribution
  - Pois(λ): $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
  - $T(X) = \sum_{i+1}^{N} X_i$
- Normal distribution
  - For normal distribution with expected value μ and known finite variance σ^2,
  - $T(X) = \bar{X} = \frac{1}{n} \sum_{i+1}^{N} X_i$ for μ.
- ...



LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

Often tempted to transform the data (round up, quantize, etc).
Physicists are good at simplifying problem
we use all kinds of expansions and approximations!
please validate assumptions for your estimation and approximation!

Need to be careful about the *domain of operation* - problem specific!

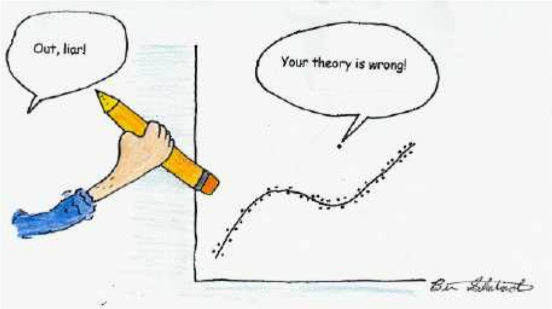


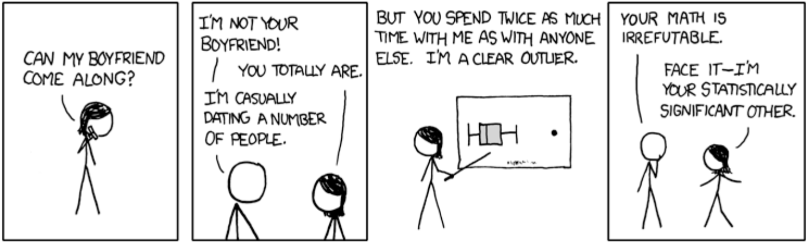
Question to audience: can additive noise to images really be Gaussian?

http://math.sfsu.edu/beck/

Statistical Significance: An assessment of whether the observations made are part of a pattern rather than randomness. *Statistical significance does not necessarily mean important or meaningful in the practical sense.* It means that you are x% sure that your result is accurate.

For example, testing at an alpha level of 0.05 indicates that if your observations fall under the extreme 5%, you are 95% sure that your result is accurate with a 5% chance that it was due to randomness.





Question to audience: is the use of "statistically significant other" appropriate, albeit funny, in this case? What test would we use?

comic source: http://xkcd.com/1132/

Outline

- The special role of Poisson is mostly governed by the modeling and phenomena encountered in medical physics: the common encounter of "counting process".
- For Poisson, a single parameter λ dictates both mean and variance.
- Large number of counts $\Rightarrow$ larger λ
-
- $\Rightarrow$ larger signal to noise ratio.

- The special role of Poisson is mostly governed by the modeling and phenomena encountered in medical physics: the common encounter of "counting process".
- For Poisson, a single parameter λ dictates both mean and variance.
- Large number of counts $\Rightarrow$ larger λ
- $\Rightarrow$ larger standard deviation $\sqrt{\lambda}$
- $\Rightarrow$ larger signal to noise ratio.

Central Limit Theorem

Given certain conditions, the mean of a sufficiently large number of iterates of independent random variables, each with a well-defined mean and well-defined variance, will be approximately normally distributed.

Rice, John (1995), Mathematical Statistics and Data Analysis.

(Loading movie...)

Video source: Galtonboard/galtonrett simulation or bean machine

Settings

Let $\{X_1, \dots, X_n\}$ be a random sample of size n that is, a sequence of independent and identically distributed random variables drawn from distributions of expected values given by μ and *finite* variances given by σ^2. Suppose we are interested in the sample average

$$S_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

Law of Large Number

The sample averages S_n converge in probability and almost surely to the expected value μ as $n \to \infty$.

Question to audience: what are the corresponding quantities in the bean machine experiment? What are X_i and S_n respectively? and what are we trying to observe?

Classic CLT:

- Classic CLT states that as n gets larger, the distribution of the difference between the sample average S_n and its limit μ, when multiplied by the factor $\sqrt{n}$, approximates the normal distribution with mean 0 and variance σ^2. For large enough n, the distribution of S_n is close to the normal distribution with mean μ and variance σ^2/n.
- The usefulness of the theorem is that the distribution of $\sqrt{n}(S_n - \mu)$ approaches normality *regardless of the shape of the distribution of the individual X_is.*

Question to audience: again, how are we to interpret this in the context of bean machine? What is the parent probability?

Outline

To be aware of

- model mismatch: mixture Gaussian for Poisson in image reconstruction...
- imperfection in data acquisition, data and modeling itself
- error introduced by approximations
  - quantization, observation collapsing (lost of phase information etc)
  - for convenience - to utilize analytical forms, to make problem tractable.
- the need to properly interpret the above and the subsequent analysis $\Rightarrow$ error propagation analysis and reporting $\rightarrow$ Dr. Cui.