Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

Inference

- Inference involves drawing conclusions from evidence
  - In logic, the evidence is a set of premises
  - In data analysis, the evidence is a set of data
  - In statistics, the evidence is a sample from a population
    - A population is assumed to have a distribution
    - The sample is assumed to be random (There are ways around that)
    - The population may be the same size as the sample
- There are two historical approaches to statistical inference
  - Frequentist
  - Bayesian
- There are many widespread abuses of statistical inference
  - We cherry pick our results (scientists, journals, reporters, …)
  - We didn't have a big enough sample to detect a real difference
  - We think a large sample guarantees accuracy (the bigger the better)

Inference

Deductive (top down)

All men are mortal. (premise)
Apollo is a man. (premise)
Therefore, Apollo is mortal. (conclusion)

The conclusion is guaranteed if premises are true

Abductive

Bill and Jane had a fight and stopped seeing each other
I just saw Bill and Jane having coffee together
I conclude they are friends again

The conclusion is not guaranteed even if premise(s) are true

Inductive (bottom up)

All of the swans we have seen are white.
Therefore, all swans are white.

The conclusion is not guaranteed even if premise(s) are true
There exist black swans (also blue lobsters)

Mathematical proofs are deductive

Data-analytic inference tends to be abductive

Statistical inference tends to be inductive

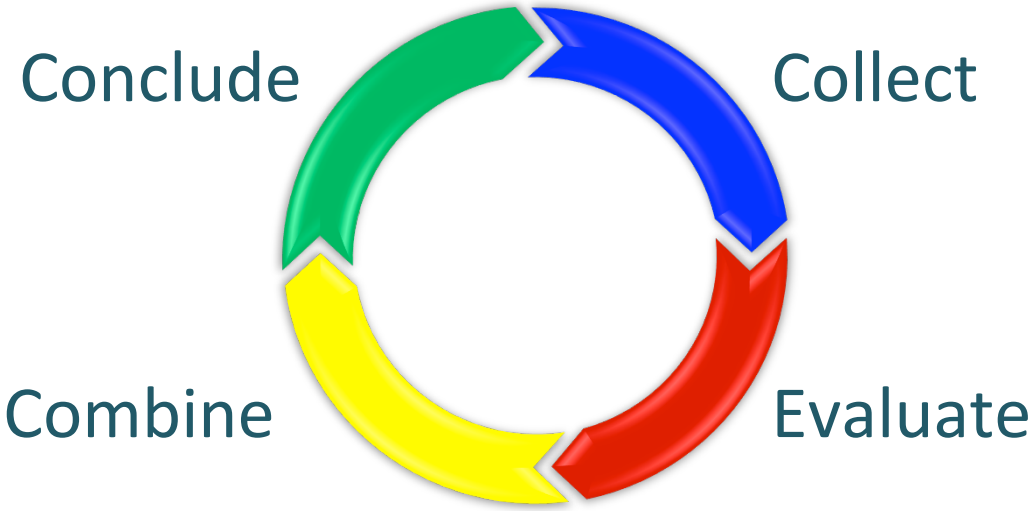
Abductive and inductive inference necessarily involve risk

Inference

Data Analytic Inference

Works like a legal argument

- Collect evidence
- Evaluate the believability of each piece of evidence
- Combine evidence
- Draw a conclusion

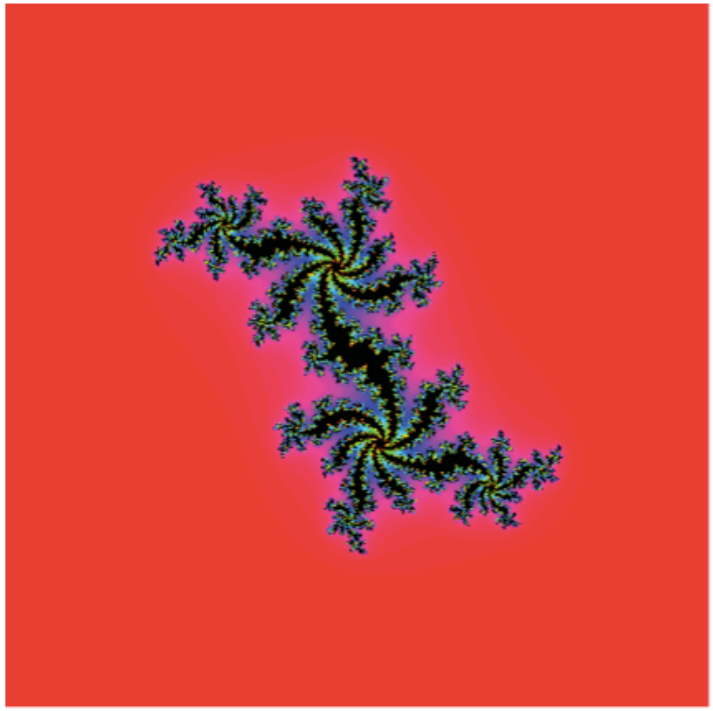


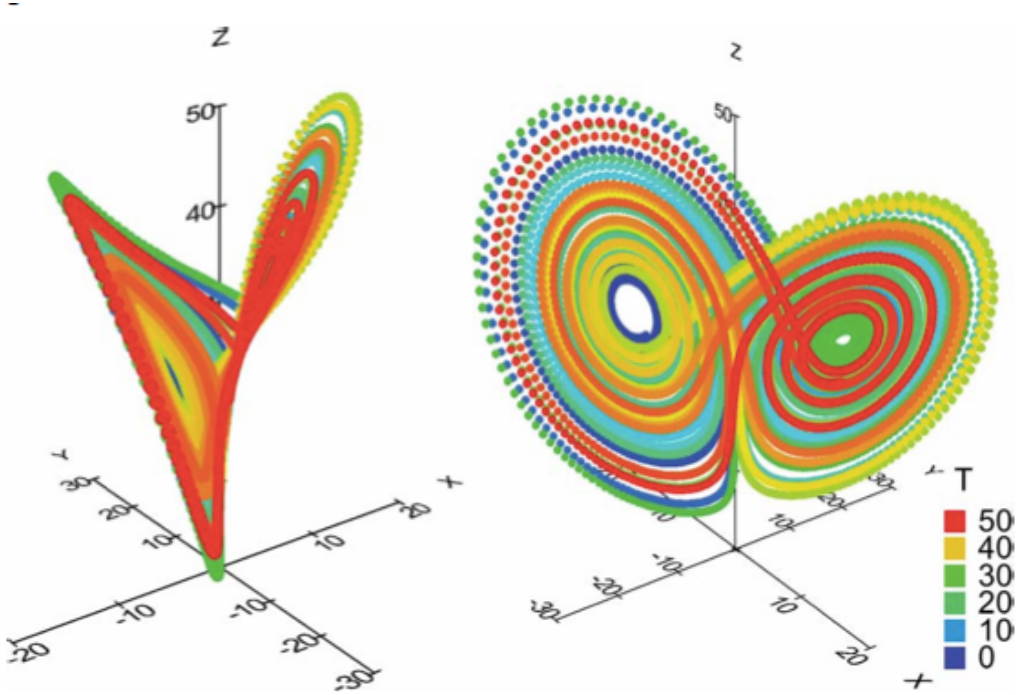
Inference

Data Analytic Inference

When data analysis is sufficient (without needing statistics)

The data are deterministic





Inference

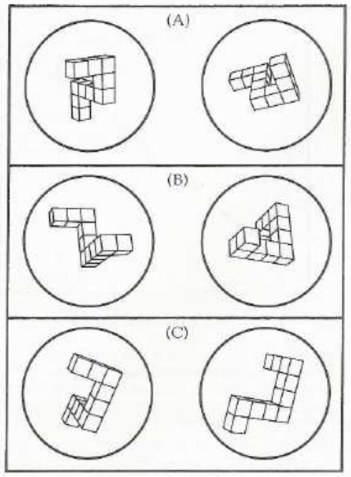
Data Analytic Inference

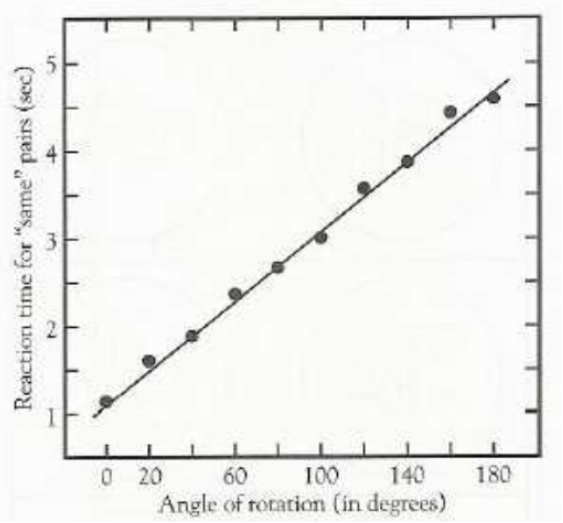
When data analysis is sufficient

Error distribution is ignorable (Berkson's Intraocular Traumatic Test)

Shepard and Metzler mental rotation angle beautifully predicted reaction time

The time it took to identify whether two figures were the same



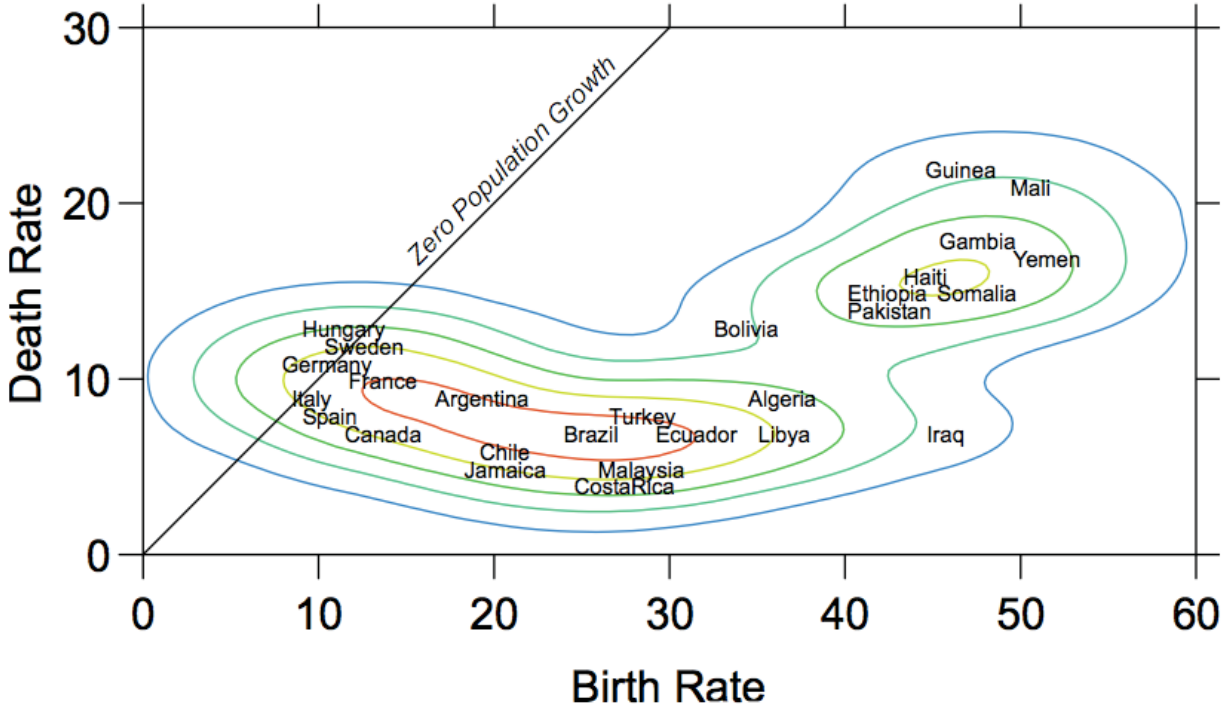


Shepard & Metzler (1971)

Inference

Data Analytic Inference

When a graph tells the whole story

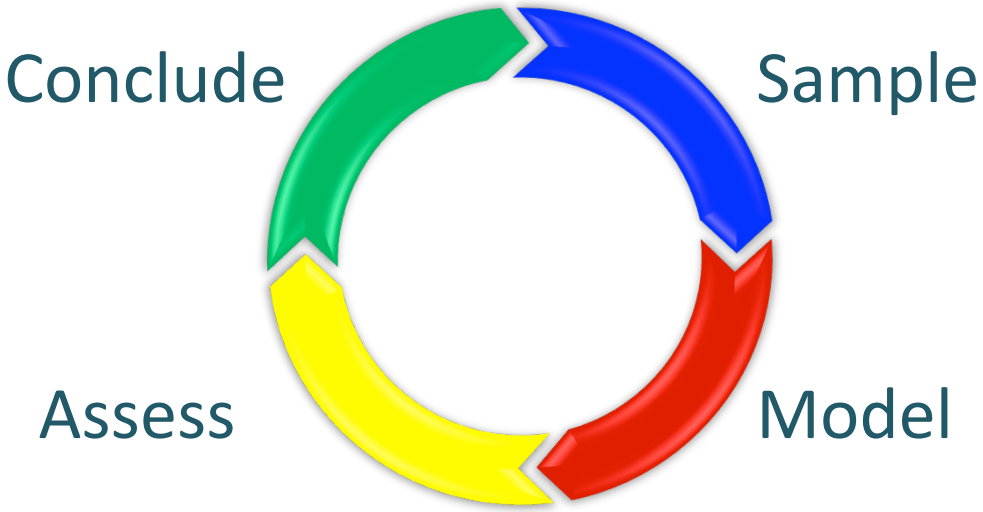


Inference

Statistical Inference

Based on probability distributions

- Sample data
- Apply a probability model to the data
- Assess the adequacy of the model
- Draw conclusion with resulting level of confidence



Inference

Inferring Parameters of a Distribution via Maximum Likelihood

We have a sample

- We know (assume) it is a simple random sample (SRS) from a population
  - SRS: every possible sample of size *n* has an equal probability of being selected
- Each observation sampled is independent of the others sampled
- We know (assume) the probability distribution representing the population

What is not a random sample?

- Every other case, record, instance, number in the phone book, etc.
- First n cases
- Any method that fails to consider every possible case
- Persi Diaconis tossing a coin (he can toss heads every time)
- Human-generated random numbers (people can't imitate randomness)
- Pseudo-random numbers
  - Actually, good algorithms produce numbers indistinguishable from truly random
  - So we use them and hold our breath

Inference

Inferring Parameters of a Distribution via Maximum Likelihood

How do we infer the parameter(s) of that probability distribution?

The likelihood that θ is our parameter value, based on our sample information is:

$$L(\theta; x_1, \ldots, x_n) = P(x_1, \ldots, x_n; \theta)$$

The likelihood is the probability of observing our sample values based on different values of θ

- It is **not** a probability density function (its mass or the area under it is not 1)

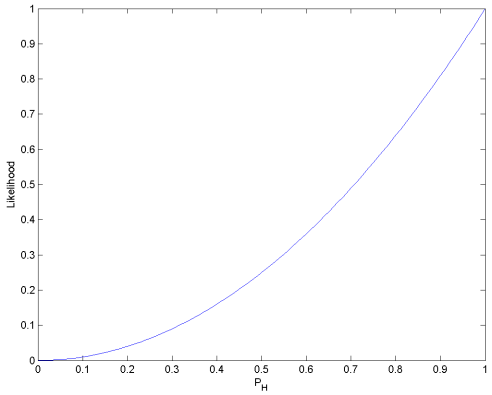
We are going to maximize this likelihood in order to estimate θ

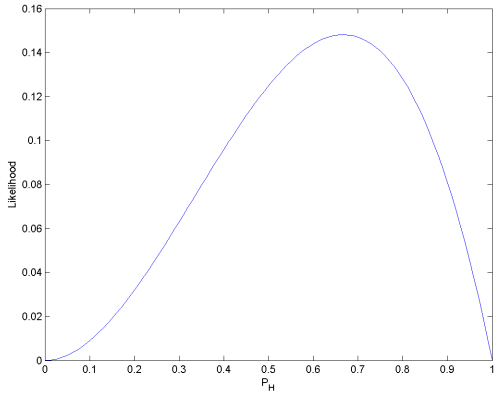
- Because we want our estimate to be the most likely value to have generated our data

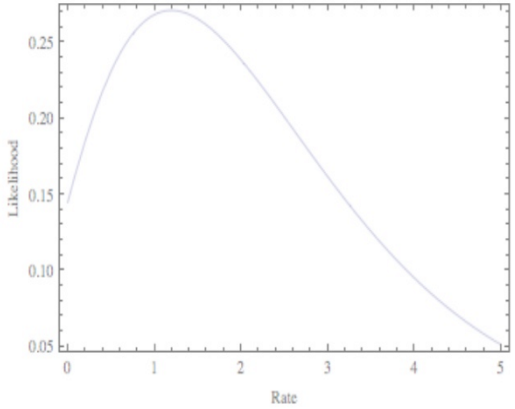
Inference

Inferring Parameters of a Distribution via Maximum Likelihood

Likelihood functions are not probability density functions







Inference

Inferring Parameters of a Distribution via Maximum Likelihood

Given the product rule for independent events,

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

The product function is rather awkward, so we log the likelihood

$$\begin{aligned} l(\theta; x_1, \ldots, x_n) &= \log\left[L(\theta; x_1, \ldots, x_n)\right] \\ &= \sum_{i=1}^{n} \log f(x_1, \ldots, x_n; \theta) \end{aligned}$$

Maximizing the log-likelihood is equivalent to maximizing the likelihood

So, all we need is $f(.)$ for a given probability distribution

Problems:

- A closed-form solution may not exist
  - In that case, we have to use numerical optimization
- In other cases, there may be no maximum
  - In that case, we are hosed

Inference

Maximum Likelihood Estimates for Normal Distribution

$N(\mu, \sigma^2)$ has two parameters

Density is

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

Likelihood is product of densities

$$L(\mu, \sigma^2; x_1, \ldots, x_n) = (2\pi\sigma^2)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

And log-likelihood is

$$l(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Inference

Maximum Likelihood Estimates for Normal Distribution

Maximizing the log-likelihood

$$\max_{\mu,\sigma^2} l(\mu, \sigma^2; x_1, \ldots, x_n)$$

requires

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2; x_1, \ldots, x_n) = 0$$
$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2; x_1, \ldots, x_n) = 0$$

The respective partial derivatives are

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right)$$

and,

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{1}{2\sigma^2} \left[\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n \right]$$

Inference

Maximum Likelihood Estimates for Normal Distribution

Maximizing the log-likelihood with respect to μ

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

implies (because σ^2 cannot be 0)

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and maximizing it with respect to σ^2

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - n\right] = 0$$

implies

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Inference

Likelihood Ratio Tests

Let L_1^{max} be the maximum value of the likelihood for a given full model

Let L_0^{max} be the maximum value of the likelihood for a restricted model

A restricted model is one where the values of some of the parameters are fixed

These fixed values may be null (set to zero) or some other value

Then

$$\chi_k^2 = -2 \log \left(\frac{L_0^{max}}{L_1^{max}} \right)$$ (the difference between two log-likelihoods)

has a chi-square distribution with *k* degrees of freedom

k is the difference between the number of parameters in the full vs the restricted

(The *Wald Test* is a type of LR test)

Assumptions

Models must be *nested*

The test is asymptotic (*n* must be large)

This last assumption is widely abused

Inference

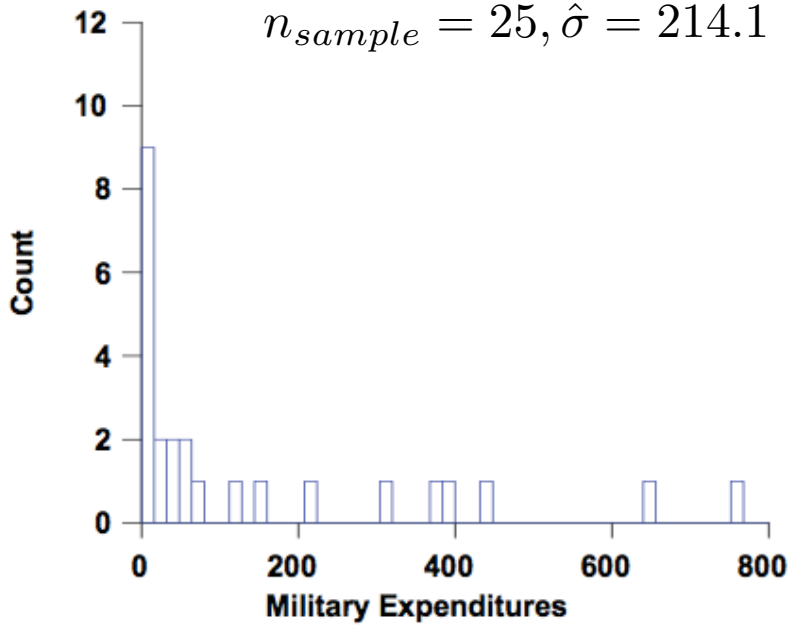
- Inferring Parameters of a Distribution via the Bootstrap
  - Efron (1981)
  - Sample with replacement from a sample
  - Compute estimate of a parameter from this bootstrap sample
  - Do this lots of times (say, 1000)
  - Histogram the bootstrap parameter estimates
  - Compute sample statistics on histogram
    - sample mean, sd
    - fractiles
    - confidence intervals
  - Or, smooth the histogram before computing statistics

  - Efron and others have proofs for why this works
    - Not as effective for skewed distributions
    - Not as effective for dependent observations

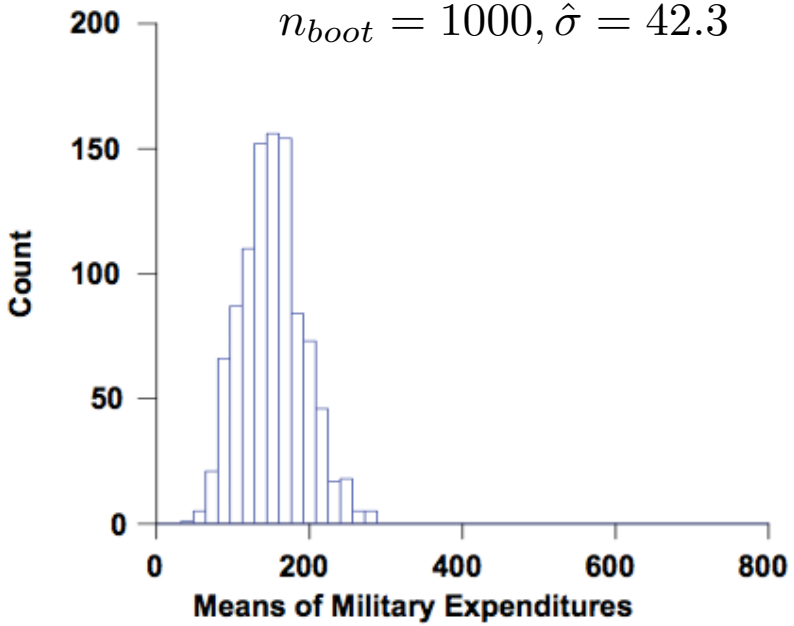
Inference

Inferring Parameters of a Distribution via the Bootstrap

$n_{sample} = 25, \hat{\sigma} = 214.1$



$n_{boot} = 1000, \hat{\sigma} = 42.3$

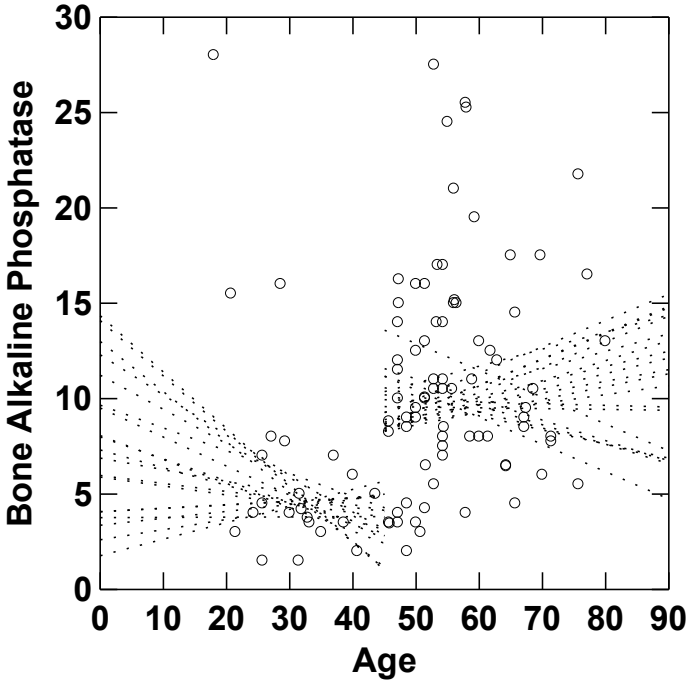


$\sigma_\mu = \frac{\sigma}{\sqrt{n}}$ central limit theorem $42.82 = \frac{214.1}{\sqrt{25}}$ pretty close!

Inference

Inferring Parameters of a Distribution via the Bootstrap

20 Bootstrap estimates of robust piecewise regression



Inference

Confidence Intervals

An interval $I_\theta = [l(x), u(x)]$ such that $P(\theta \in I_\theta) = 1 - \alpha$

For interval on normal mean, we use $I_\mu = (\hat{\mu} - z_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + z_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}})$

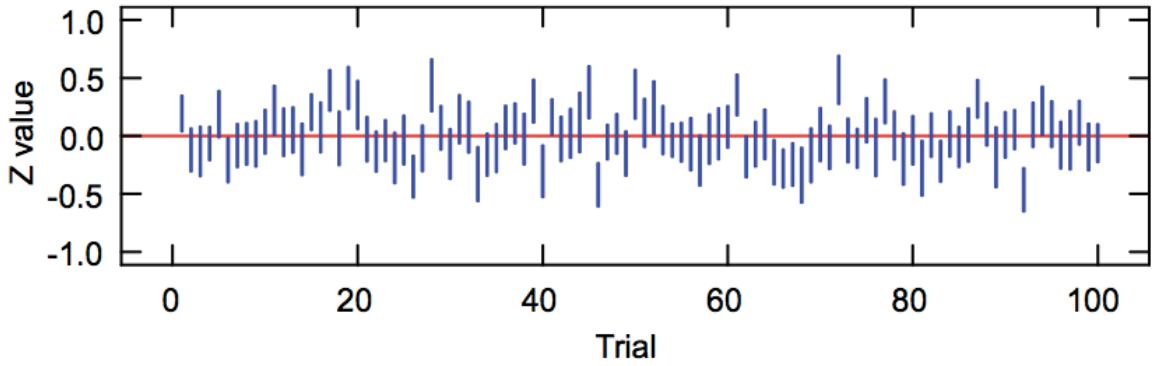
Computation based on likelihood $P(x|\theta)$, where θ is a fixed value

$P(\theta \in I_\theta)$ is based on a collection of intervals, not this one

Wrong to say, "There is a 95% chance θ lies in *this* interval"

It either does or doesn't (the *interval* is a random variable, not θ)

Say instead, "there is a 95% chance that when I compute a confidence interval on a sample from this population, the true value of θ will fall within it"



Inference

Why confidence is not probability

Let $x_1, x_2 \sim U(\theta - 1, \theta + 1)$

There is a 25% chance that both x_1 and x_2 will lie below θ

There is a 25% chance that both x_1 and x_2 will lie above θ

Therefore, there is a 50% chance that θ will lie between them

Then $(y_1 = \min[x1, x2], y_2 = \max[x1, x2])$ is a 50% confidence interval

However, when $y_2 - y_1 > 1$, it MUST contain θ, even though

(y_1, y_2) is a confidence interval

In other words, confidence intervals are not betworthy

Hartigan proved this argument for other distributions (e.g., Normal)

Thanks to Jerry Dallal for distilling Hartigan's argument

Inference

Credible Intervals

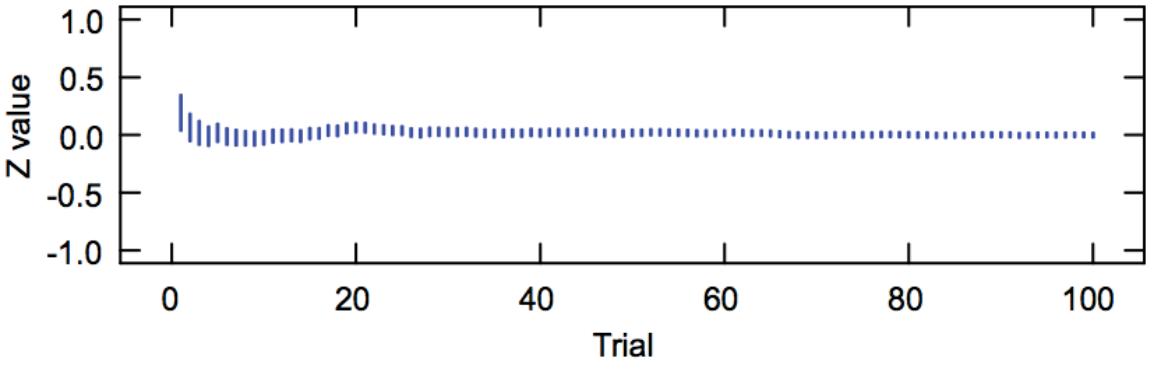
An interval such that $P(l(\theta) \leq \theta \leq u(\theta)) = 1 - \alpha$

Computation based on posterior $P(\theta|x) \sim P(x|\theta)P(\theta)$

θ is fixed, but we are uncertain about its value, so we use $P(\theta)$ prior

$P(l(\theta) \leq \theta \leq u(\theta))$ is based on observed data

"Given our observed data, there is a 95% chance that the true value of θ falls within this credible interval"



Inference

Hypothesis testing

Protecting against false positives

Construct Null Hypothesis H_0 (usually, that a result is due to chance)

State rule for rejecting H_0

Compute likelihood of observed result under H_0

Draw a conclusion based on decision rule

Inference

Hypothesis testing

Protecting against false positives – the first significance test

> *An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes*. By Dr. John Arbuthnot, Physician in Ordinary to her Majesty, and Fellow of the College of Physicians and the Royal Society

There seems no more probable Cause to be assigned in Physics for this Equality of the Births, than that in our 'first Parents Seed there were at first formed an equal Number of both Sexes.

[…] From hence it follows, that Polygamy is contrary to the Law of Nature and Justice, and to the Propagation of the Human Race; for where Males and Females are in equal number, if one Man take Twenty Wives, Nineteen Men must live in Celibacy, which is repugnant to the Design of Nature; nor is it probable that Twenty Women will be so well impregnated by one Man as by Twenty.

$$P(\text{exactly equal numbers of Males and Females}) = \binom{n}{n/2}\left(\frac{1}{2}\right)^n$$

Inference

Hypothesis testing

The Lady Tasting Tea (Fisher)

The lady claimed she could tell whether milk or tea was first added to the cup

Fisher gave her 8 cups, 4 of each type, in random order

The table on the left below shows the lady's responses

The tables on the right show all possible tables, given the margins of 4 cups

There is only a 1 in 70 chance that the lady could have guessed all 8 correctly

Answer

Truth	Milk First	Tea First	Total
Milk First	4	0	**4**
Tea First	0	4	**4**
Total	**4**	**4**	**8**

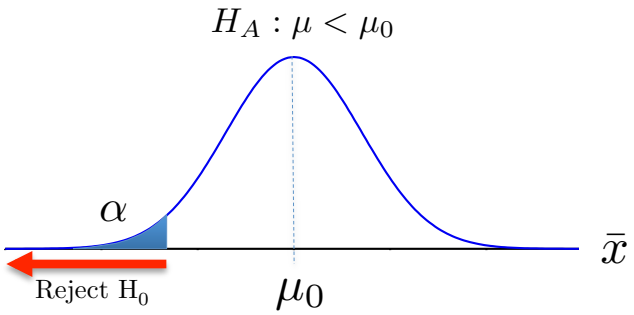
0	4	1	3	2	2	3	1	4	0	
4	0	3	1	2	2	1	3	0	4	
1		16		36		16		1		Total 70
.014		.229		.514		.229		.014		1

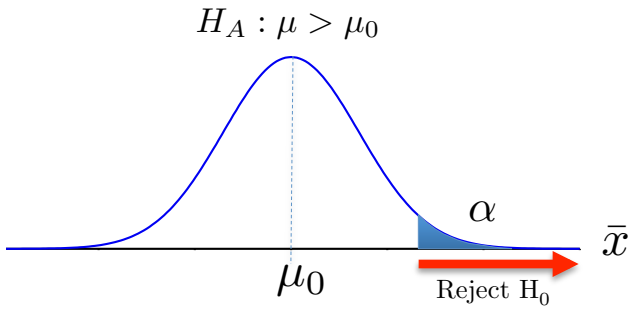
Inference

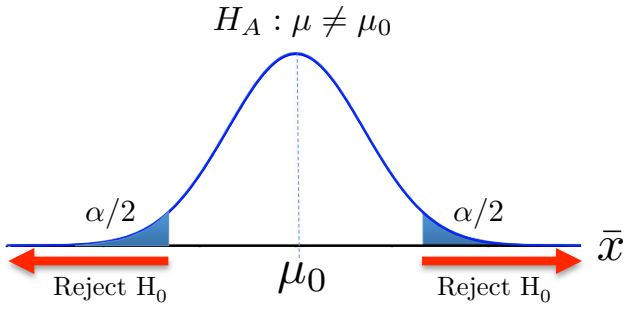
Hypothesis testing

one-tailed and two-tailed tests

Tukey once pointed out that the two-sided null hypothesis that a parameter is zero is really saying that we don't know its sign.





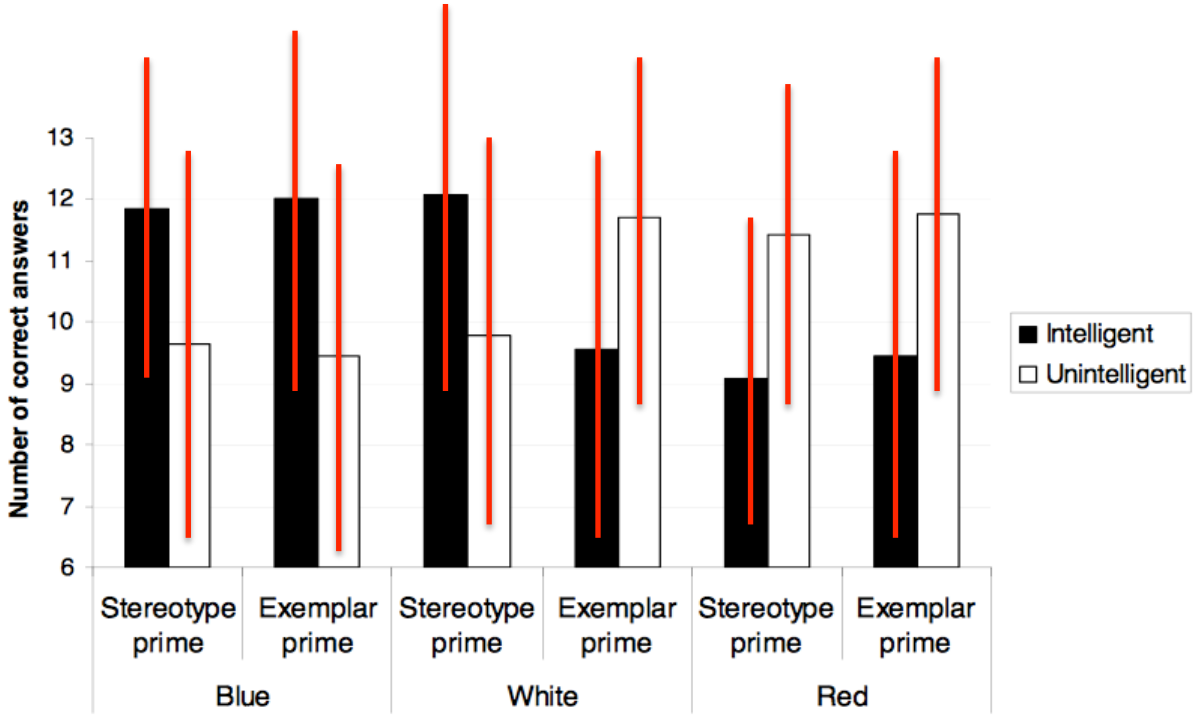


A one-sided test

Here is a graph of the results of an experiment by Smeeters & Liu

The bars show the mean number of correct answers

The red lines (my addition) show the standard deviations

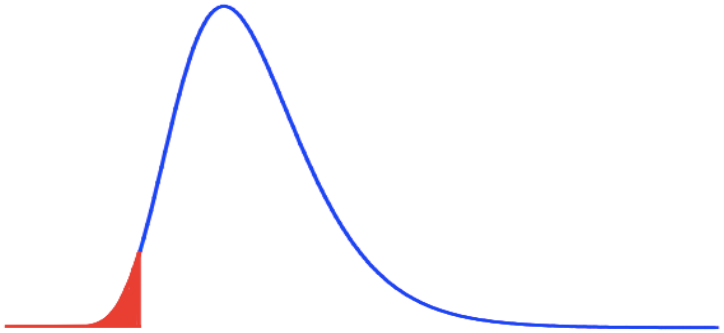


Smeeters & Liu (2011) *JEXP*

Thanks to Uri Simonsohn and Richard Gill

A one-sided test

- Contrast black vs. white bars over each level
  - Compute an ANOVA on black vs. white bars
- If *F* value is very small, be very suspicious
- Because between-groups variation is too small relative to within
- The *lower* tail of the *F* distribution yields our *p* value
  - For the Smeeters and Liu article, *F* was so small that it revealed fraud
  - Smeeters was forced to resign from the university



Inference

Hypothesis testing

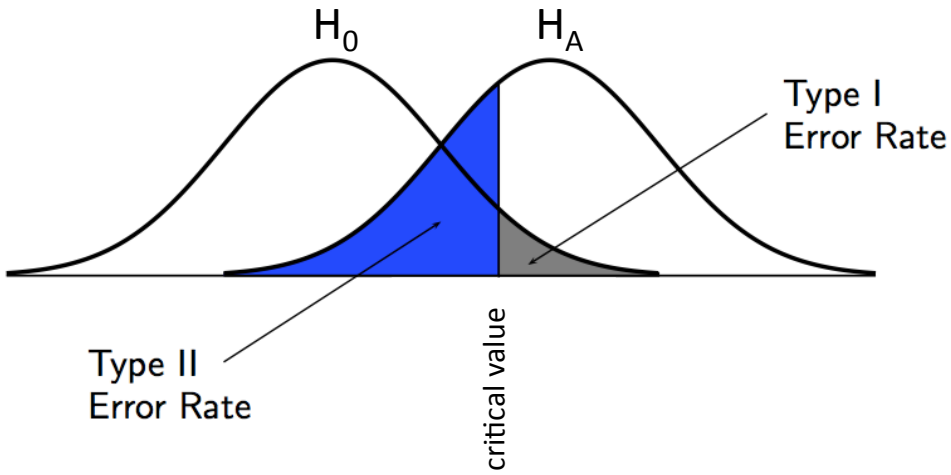
Neyman-Pearson procedure (Jerzy Neyman and Egon Pearson)

- Construct Null Hypothesis H_0 (ordinarily, that a sample result is due to chance)
- Construct Alternate Hypothesis H_A (ordinarily, that a sample result is not H_0)
- State criterion for rejecting H_0
- Compute test statistic
- Make a decision based on test statistic
- Like Fisher's method, this is a falsification procedure
- But it allows us to determine the power of the test

Inference

Hypothesis testing

Type I and Type II errors (power analysis)



Conclusion

Null Hypothesis	Accept* H_0	Reject H_0
True	True Negative	Type I error (α) False Positive
False	Type II error (β) False Negative	True Positive

*Fail to reject

Inference

Hypothesis testing

Multiple Tests (*m* tests)

Per Comparison Error Rate (PCER)

Uncorrected single tests

$$P(FD_k > 0) \leq \alpha \; \forall \; 1 \leq k \leq m$$

$$\rightarrow \alpha_k^* = \alpha$$

Family-Wise Error Rate (FWER)

e.g., Bonferroni

$$P(FD > 0) \leq \alpha$$

$$P(\bigcup_{k=1}^{m} E_k) \leq \sum_{k=1}^{m} P(E_k)$$

$$\rightarrow \alpha_k^* = \alpha / m$$

False Discovery Rate (FDR)

Benjamini-Hochberg

$$E\left[\frac{FD}{D}\right] \leq \alpha$$

$$\rightarrow \alpha_k^* = \alpha \frac{k}{m}$$

	Retain H_0	**Reject H_0**
H_0 True	True Nondiscovery	False Discovery
H_0 False	False Nondiscovery	True Discovery

Inference

Hypothesis testing

Multiple Tests

False Discovery Rate (FDR)

4 false discoveries out of 10 rejected null hypotheses is a more serious error than

20 false discoveries out of 100 rejected null hypotheses

Assumption is that tests are independent

Although, this can be relaxed somewhat

Doesn't depend on distribution, only *p* values from tests

Inference

Hypothesis testing

Multiple Tests

FDR Plot

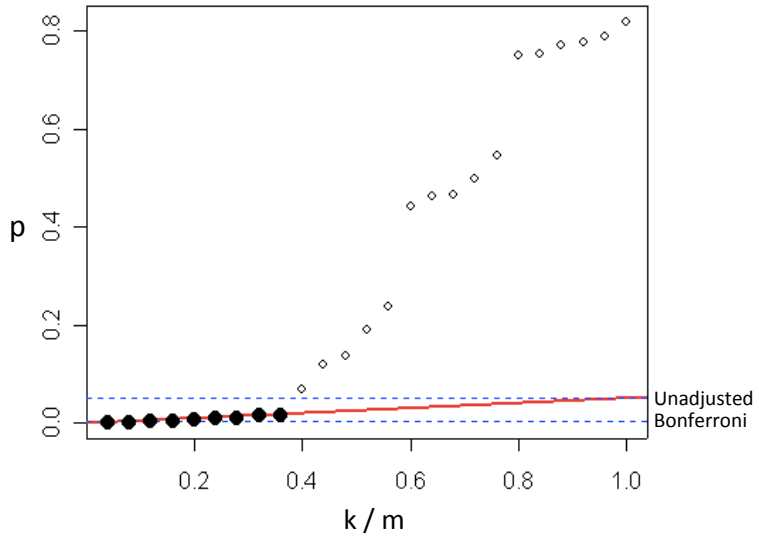
Sort the *p*-values from largest to smallest

Plot the ordered *p*-values on the *y*-axis versus *k/m* on the *x*-axis

Superimpose a line that passes through the origin and has slope α

Any *p*-value that falls on or below this line corresponds to a significant result

m=25, α = .05



Original graph by Jack Weiss, UNC

Inference

- Hypothesis testing
  - Covert multiple tests
    - Texas sharpshooter fallacy
      - A Texan fires some gunshots at the side of a barn
      - He paints a target centered on the biggest cluster of hits
      - He then claims to be a sharpshooter

M. Feychting and M. Alhbom (1992). Magnetic fields and cancer in children residing near Swedish high-voltage power lines. *American Journal of Epidemiology, 138,* 467-481.

- Surveyed everyone living within 300 meters of high-voltage power lines from 1960 through 1985.
- Looked for statistically significant increases in relative risk (against baseline) of over 800 illnesses.
- Found that there was a significant relative risk of childhood leukemia for those living near power lines.
- The number of illnesses considered was so large, however, that there was high probability that the increased risk of at least one illness would appear statistically significant by chance alone.
- Subsequent studies failed to show any links between power lines and childhood leukemia.

Inference

- Hypothesis testing
  - Highly significant *p*-value doesn't mean effect is large or strong or influential
  - Statistical significance does not imply practical significance
    - Practical significance (importance) depends on meaning, not chance

Inference

- Hypothesis testing
  - H_0 doesn't mean zero value of the parameter
  - H_0 can involve any value
    - Zero value used for "nil hypothesis" instead of "null hypothesis"
    - Nil hypothesis is absurd, of course

Inference

- Hypothesis testing
  - Failure to reject H_0 doesn't prove it
    - Can increase *n* enough to make almost any H_0 false
      - This is a trick used in ESP research (see Duke studies)
    - WRONG: *p*=0.05 means "the probability of the null hypothesis not being true is 95%"
    - WRONG: *p* = .06 means "the average effect size (d=0.04) is not different from 0."

Inference

- Hypothesis testing
  - Confidence intervals are not a cure for NHST problems
    - They come out of same calculations for NHST, although they convey more information

Inference

- Hypothesis testing
  - *p*= .05 is not sacred
    - Fisher thought of *p* values as quantifying evidence against an hypothesis
    - He picked .05 for a cutoff for most practical problems

Inference

- Hypothesis testing
  - Falsification (Popper) is wrong
    - We build evidence for a theory
    - There is no such thing as a critical experiment (see Kuhn)
    - But a **theory** that cannot be falsified is suspect
      - Try to disprove Freud's Oedipus Complex
      - Freud wouldn't accept any evidence to the contrary

Inference

Hypothesis testing

P-values are not the main problem

- **fraud**
- journal **selection bias** in favor of “significant” results and against “non-significant” replication (the winner’s curse)
- Publishing p-values without **supporting information** (effect sizes, confidence intervals, …)
- failure to control false discovery rate in **multiple tests**
- **small samples** (low power, Tversky and Kahneman’s “Law of Small Numbers")
- **large samples**("more than 100,000 women diagnosed from 1988 to 2011 with DCIS”; wow! that must make this study trustworthy)
- **convenience samples** (“we studied depression by giving a questionnaire to sophomore psychology students”)
- **experimenter bias** (failure to use double blind and other controls when available)
- **cherry picking** (uncontrolled model selection, stepwise regression, …)
- **promiscuous mining** -- for one of the most egregious examples, see

http://googlecloudplatform.blogspot.com/2014/08/correlating-patterns-of-world-history-with-bigquery.html

Inference

Hypothesis testing

P-values are not the main problem

- **pollution from money** (covert support by tobacco, soft-drink, chemical, energy, food and drug companies, …)
- overly **complex statistical models** in place of simple alternatives (LISREL, HLM, BUGS, …)
- **misuse of statistical concepts** in interpreting results
- relying on **statistical bloggers** or Wikipedia for advice (the madness of crowds)
- **15 minutes of fame** (h-index, citations, awards, keynotes, TED talks, and other factors driving excess publications, premature media publicity, and inattention to detail)
- inflation in **tenure** requirements (10 publications a year? Are you kidding?)
- pressure to get **grants**
- **ignorant media reporters** (failure to understand the basics of causation, probability and inference -- coffee causes cancer, cancer causes coffee)
- **plagiarism** (copy someone else's study without understanding the statistics and data analysis)

Inference

The Bayesian objection

Likelihood principle (Leonard Jimmie Savage)

After *x* is observed, all relevant experimental information is contained in the likelihood function for the observed *x*. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other.

Suppose *X* is the number of heads in 12 flips of a fair coin and *Y* is the number of flips needed to get 3 heads.

A frequentist tests the result that *X* = 3 against a Binomial, with resulting *p* = .073.

But she tests the result that *Y* = 12 against a Negative Binomial, with *p* = .0327.

The data are the same in both circumstances, but the experiments differ

The difference between observing *X* = 3 and observing *Y* = 12 lies not in the actual data, but merely in the design of the experiment. In the first case, one has decided in advance to try 12 flips. In the second, one has decided to keep flipping until 3 successes are observed. Bayesians say the inference about θ should be the same because the two likelihoods are proportional to each other.

$$L(\theta) \propto p^3(1-p)^9$$

Inference

Hypothesis testing

Bayesian inference

Identify prior distribution $P(H)$ on hypothesis parameters

Specify parameter values of prior distribution

Compute likelihood $P(E|H)$ on evidence given hypothesis

Compute marginal likelihood $P(E)$, which averages over parameters of interest

Compute posterior distribution through Bayes' theorem

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Graphically display the posterior distribution

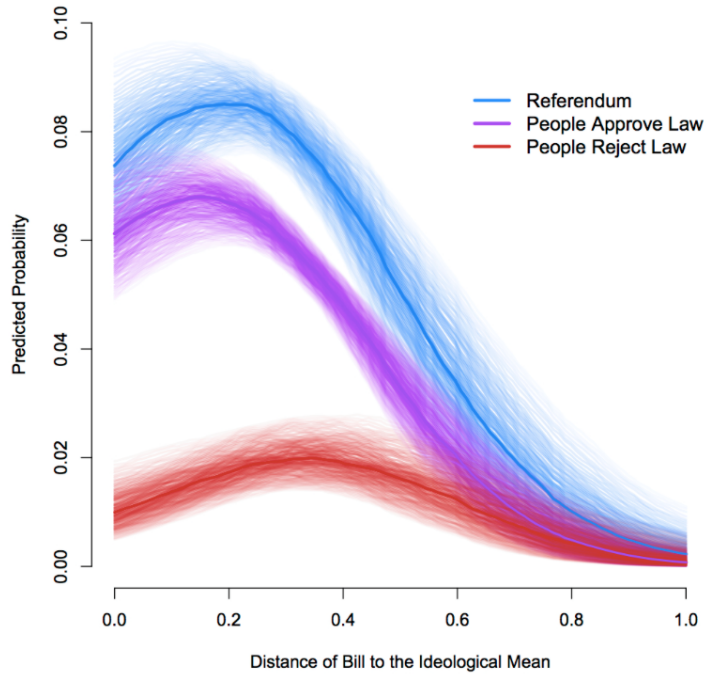
Or, compute credible intervals and other statistics characterizing posterior

The posterior can be used as a prior with new data (Bayesian updating)

Inference

Hypothesis testing

Bayesian visualization

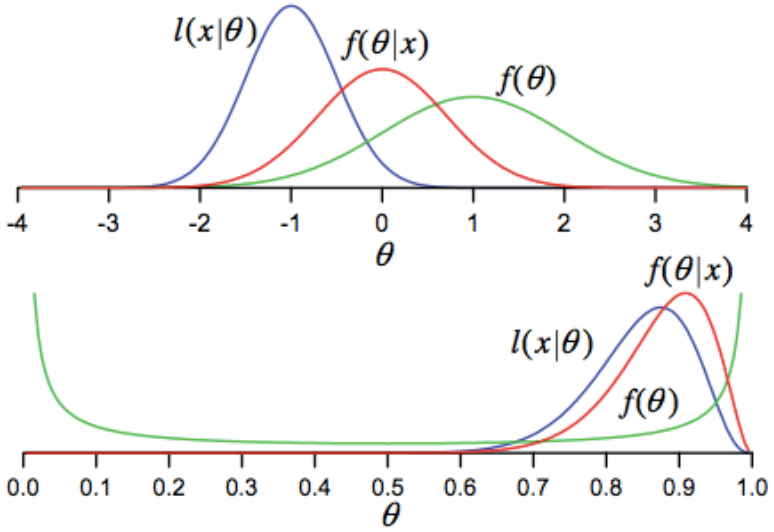


Andrewgelman.com

Inference

Jerry Dallal: How to annoy a Bayesian

With a lot of work, it may be possible to elicit a one dimensional prior.



Inference

Jerry Dallal: How to annoy a Bayesian

With a lot of work, it may be possible to elicit a one dimensional prior.

There *may* be some circumstances where, with a WHOLE lot of work, it is possible to elicit a two-dimensional prior.

Inference

Jerry Dallal: How to annoy a Bayesian

With a lot of work, it may be possible to elicit a one dimensional prior.

There *may* be some circumstances where, with a WHOLE lot of work, it is possible to elicit a two-dimensional prior.

NO ONE can specify a three-dimensional prior!

Persi Diaconis:

It's very hard to put meaningful priors on high-dimensional real problems. And the choices can really make a difference.

Inference

Jerry Dallal: How to annoy a Bayesian

Even if one were to assume that a multivariate normal distribution were appropriate, not only are there means and standard deviations to estimate, but there are also correlations (or covariances) to estimate.

Eliciting opinions about covariances is likely to be much more difficult that eliciting opinions about means and standard deviations, and the actual values matter…***a lot***!

Inference

Jerry Dallal: How to annoy a Bayesian

If the prior is sharp relative to the likelihood, the posterior distribution will look like the prior.

If the likelihood is sharp relative to the prior, the posterior distribution will look like the likelihood.

If neither the prior nor the likelihood is sharp relative to the other, the posterior distribution will be a mix of the two.

Inference

Jerry Dallal: How to annoy a Bayesian

If the prior is sharp relative to the likelihood, the posterior distribution will look like the prior.

Why are we doing this study?

Inference

Jerry Dallal: How to annoy a Bayesian

If the likelihood is sharp relative to the prior, the posterior distribution will look like the likelihood.

…and the results will be the same as from a frequentist analysis!

This is scary. It says that frequentist results that could be mistaken for probability statements really are probability statements!

Inference

Jerry Dallal: How to annoy a Bayesian

If neither the prior nor the likelihood is sharp relative to the other, the posterior distribution will be a mix of the two.

Be afraid! Be very afraid!!

(See next slide…)

Inference

Jerry Dallal: How to annoy a Bayesian

Whose prior?

- Sponsors
- Special Interest Groups*
- Investigators
- Reviewers
- Policy Makers
- Consumers

**If you are my friend, you will do your best to avoid using the word "stakeholder" in my presence, says Jerry.*

Inference

Jerry Dallal: How to annoy a Bayesian

Bayes methods do not allow for *surprise!*

This is by definition. A prior distribution reflects belief about expectations.

Inference

Jerry Dallal: How to annoy a Bayesian

Something that is impossible under the prior distribution MUST be impossible under the posterior distribution. Okay, nothing is impossible, so we'll withhold a bit of prior probability to spread around (…but how much?)

This doesn't solve the problem. Something that is unexpected under the prior must still be rare under the posterior unless there is a HUGE amount of data, or it really wasn't all that unexpected.

Inference

Who is right?

Inference

- Who is right?
  - Both
    - Statistical methodology is not going to satisfy philosophers any time soon
      - (Nobody seems to satisfy a philosopher)
    - Frequentist statistical procedures have proved effective in practice
      - Drug trials (including "best of breed" H_0)
      - Frequentists do use Bayes' theorem
      - Frequentists do take prior knowledge into account when they design experiments
        - Especially when computing power
    - Bayesian procedures have brought plausibility to social science research
    - Frequentist and Bayesian analyses often come up with similar results
    - And when they don't, disagreements can lead to progress
      - Science is all about controversy
      - So is the law
      - So are politics
      - So is religion
      - So is life
      - Oh, never mind…

Inference

Problems with Frequentist Inference

Sir David Cox (a Frequentist)

"… I felt, for instance, that various aspects of the Neyman-Pearson theory -- choose alpha, choose a critical region, reject or accept the null hypothesis -- give a rigid procedure, that this isn't the way to do science."

"Neyman talked a lot about inductive rules of behavior and it seemed to me he took the view that the only thing that you could ever say is if you follow this procedure again and again, then 95% of the time something will happen; that you couldn't say anything about a particular instance. Now, I don't think that's how he actually used statistical methods when it came to applications; he took a much more flexible way.

But even apart from that, you can say, is this notion of 5% or 95% region -- is this just an explanation of what a 95% confidence interval would mean? A sort of hypothetical explanation, if you were to do so and so, such and such would happen? Or is it an instruction on how to do science? It seems to me okay as the first, in fact very good as the first, terrible as the second."

Inference

Problems with Bayesian inference

John Hartigan (a Bayesian)

"The degree to which I am a Bayesian is directly proportional to the distance I am from the computer center."

"Most Bayesians reject frequentist ideas as being outrageous and the next thing that comes out of their mouths is, 'Let's use these Bayesian techniques that match up with frequentist methods.' Therefore, although they object to frequentist philosophy, they follow frequentist practice."

Inference

- Inference is worthless without data analysis
  - No experiment in the real world has i.i.d. trials
  - Small random samples do not insure *ceteris paribus*
  - Matching (blocking) doesn't either
  - Experiments in the real world are difficult to replicate
  - All data are contaminated
    - Mixtures
    - Outliers
  - Distributions change over time
  - Models are sensitive (even robust models)
  - Time/space dependencies are more harmful than outliers
  - There are no good tests of distributional shape
    - Tests for normality are dubious; don't waste your time
    - Tests for multinormality are worthless
    - There is no substitute for looking at your data
  - But ignore the rules of inference and Nature will bite you

Inference

References

Barry, Daniel (2005). A Conversation with John Hartigan. *Statistical Science, 20*(4), 418–430.

Hacking, Ian (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hawkins, D. (1980). *Identification of Outliers*. New York: Chapman and Hall.

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medicine*.

Reid, Nancy (1994). A Conversation with Sir David Cox. *Statistical Science, 9*(3), 439-455.

Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Stigler, Stephen (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594604.