The Perils & Promise of Statistics With Large Data Sets & Complicated Models

Bayesian-Frequentist Cross-Fertilization

Tom Loredo

Dept. of Astronomy, Cornell University

http://www.astro.cornell.edu/staff/loredo/

Lecture 2

- A fundamental connection: Decision theory & optimal frequentist procedures
- Conditioning on data:
  - Estimation & ancillary statistics
  - Valid significance levels for testing
- Accounting for size of hypothesis space:
  - Nuisance parameters
  - Shrinkage
  - Multiple testing
- Summary & references

$\mathcal{B}$–$\mathcal{F}$ Relationships, I

$\mathcal{B}$ is both more general and more narrow than $\mathcal{F}$

- More general: Can (in princ.) contemplate p of *anything*, not just "random variables"
- More narrow: No freedom in how data appear in inferences—always through the *likelihood*, $\mathcal{L}_M(\theta) \equiv p(D_{\text{obs}}|\theta, M)$

$\mathcal{F}$ can always base a procedure on a $\mathcal{B}$ calculation

$\mathcal{B}$–$\mathcal{F}$ Relationships, II

Basic parameter estimation

Flat/reference priors $\rightarrow$ $\mathcal{B}$ credible regions are approximate $\mathcal{F}$ confidence regions (often better approximation than conventional $\mathcal{F}$ approximations)

Model testing

$\mathcal{F}$ and $\mathcal{B}$ testing have qualitatively different behavior

Decision

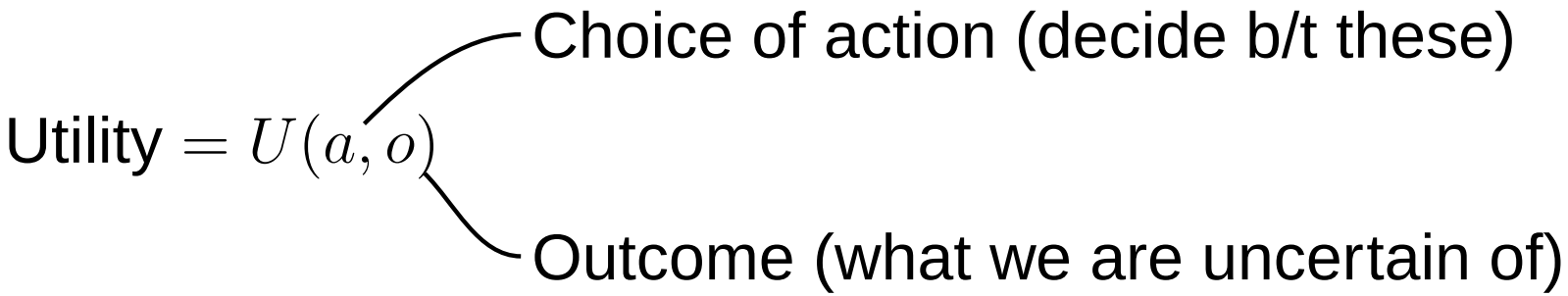
Acting Amidst Uncertainty

Decisions depend on consequences

Might bet on an improbable outcome provided the payoff is large if it occurs and the loss is small if it doesn't.

Utility and loss functions

Compare consequences via *utility* quantifying the benefits of a decision, or via *loss* quantifying costs.



Loss $L(a, o) = U_{\max} - U(a, o)$

Frequentist Decision Theory

Consider a *rule* $a(D)$ that chooses a particular action when D is observed.

Central quantity: *Risk*

$$R(o) = \sum_D p(D|o) L[a(D), o]$$

Seek rules with small risks for anticipated outcomes

Admissable rules: $a(D)$ is admissiable if there is no other rule that is at least as good for all o, and better for at least one o

Frequentist Inference and Decision

Inference: Frequentist calibration

In repeated practical use of a statistical procedure, the long-run average *actual* accuracy should not be less than (and ideally should equal) the long-run average *reported* accuracy (procedures should be calibrated).

Many procedures/rules can be created that are calibrated this way. How to choose among them?

Decision: Optimal Rules

- Devise a family of rules with desired performance
- Specify a loss function
- Find the rule with the “best” risk

Optimal $\mathcal{F}$ inference and decision are inseparable

Bayesian Decision Theory

We are uncertain of what the outcome will be

$\rightarrow$ average:

$$EU(a) = \sum_{\text{outcomes}} P(o|\ldots)\, U(a, o)$$

The best action maximizes the expected utility:

$$\hat{a} = \arg\max_{a} EU(a)$$

I.e., minimize expected loss

Inference and decision are distinct in $\mathcal{B}$—can report inferences without making decisions.

Well-Posed Inference Problems

Well-posed problems have unique solutions.

Both approaches require specification of models giving $p(D|\ldots)$. They differ in what *else* is needed.

Frequentist

- Primary measure of performance (bias, coverage, α)
- Family of procedures providing desired performance
- Loss function comparing different procedures with same primary measure (squared error, interval size, power)

Bayesian

- Information specifying priors

Example—Parameter Estimation

Estimate a normal mean, μ, from N observations, $x = \{x_i\}$; σ known

$\mathcal{F}$: Point and interval estimates

- $\bar{x}$ is best linear unbiased estimator (*BLUE*; squared-error loss)
- $I(x) = [\bar{x} \pm \sigma/\sqrt{N}]$ is shortest 68.3% *confidence interval*: $p(I(x) \text{ covers } \mu | \mu) = 0.683$

Example—Testing

Is $\mu = 0$ (M_0, "null hypothesis) or $\mu \neq 0$?

$\mathcal{F}$: Neyman-Pearson Significance test

- Procedure: Accept M_0 if $-2\sigma/\sqrt{N} < \bar{x} < 2\sigma/\sqrt{N}$
- Type I error probability $\alpha =$ "false alarm rate" $=$ 5%
- Uniformly most powerful (UMP) test (has smallest Type II error rate for any test with $\alpha = 0.05$ against $\mu \neq 0$ Normal alternatives)

$\mathcal{B}$–$\mathcal{F}$ Relationships, III

Wald's Complete Class Theorem

Admissable decision rules are Bayes rules!

Little impact on $\mathcal{F}$ practice—an admissable rule can be "worse" than an inadmissable rule:

- Admissable: $\hat{\mu} = 5$, *always*
- Inadmissable: $\hat{\mu} = (x_1 + x_N)/2$

Wald's theorem can eliminate many bad rules, but not all

Suggests $\mathcal{F}$ approach: Study $\mathcal{F}$ performance of classes of Bayes rules.

Suggests $\mathcal{B}$ approach: Identify "good" priors by studying frequentist performance of Bayes rules

Conditioning

Don't average over all possible data

Motivating example

Two instruments available for measuring μ; one adds Gaussian noise with $\sigma = 1$, the second with $\sigma = 10$.

Make a fair coin toss and let $A = 1$ or 2 denote which experiment is chosen. Then make the measurement. The sampling distribution $p(A, x|\mu)$ is a 50–50 mixture of the two normals in its x dependence—broader than $\sigma = 1$ but narrower than $\sigma = 10$.

Shouldn't we just use the noise distribution for the experiment we actually did?

Conditioning and Ancillarity

$p(A|\mu)$ is *independent* of μ: A is *ancillary* for μ. Fisher first suggested *conditioning on ancillaries*.

Cauchy example

Draw two samples from the Cauchy distribution with location (median) μ

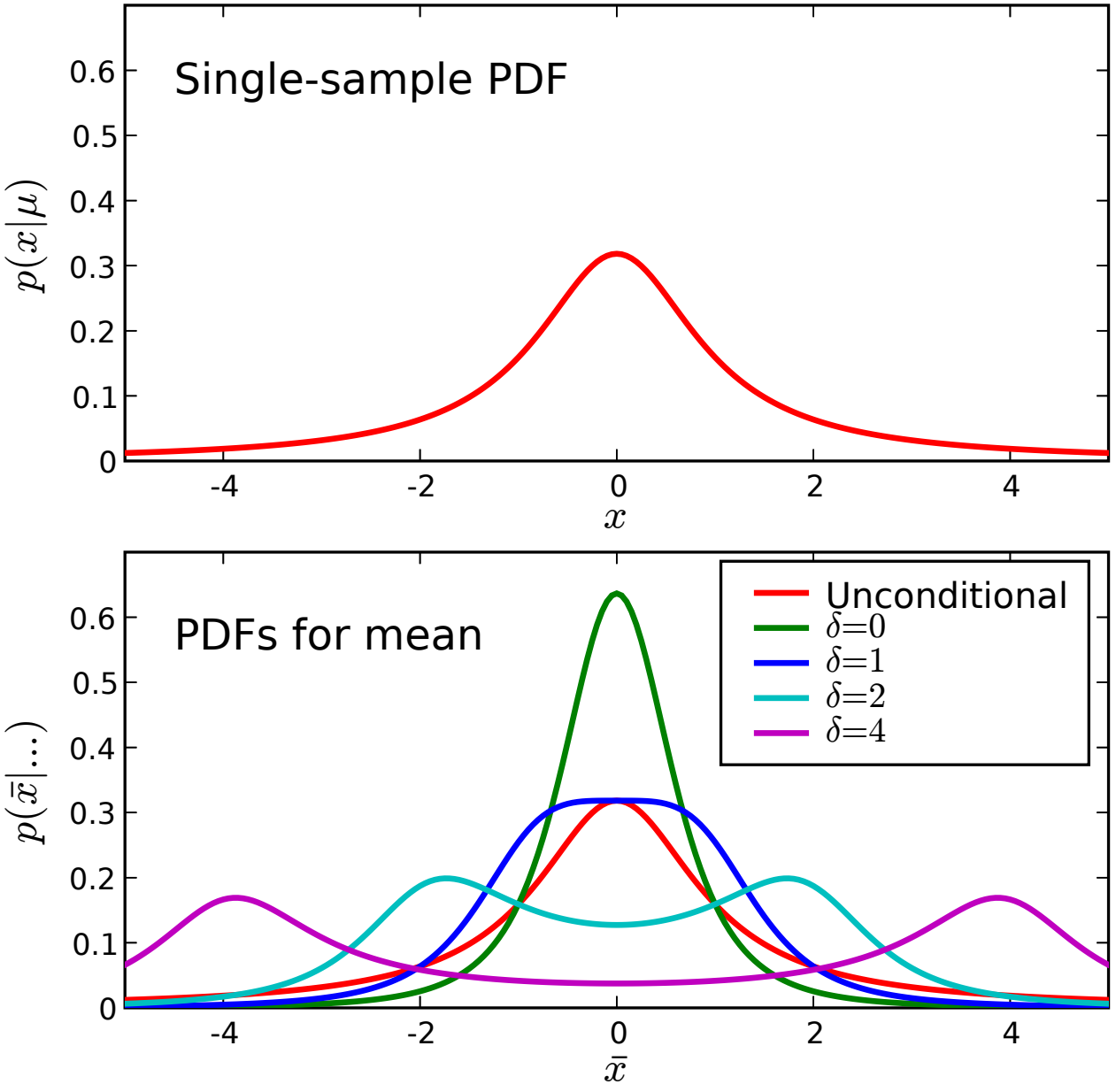
$$p(x_i|\mu) = \frac{1}{\pi}\,\frac{1}{1+(x-\mu)^2}$$

Straightforward calculation shows that $p(\bar{x}|\mu)$ is the same function! Is the mean really no better than either separate datum?

When we know (x_1, x_2), we know not only $\bar{x}$ but also $\delta = (x_2 - x_1)/2$. $p(\delta|\mu)$ is independent of μ, but

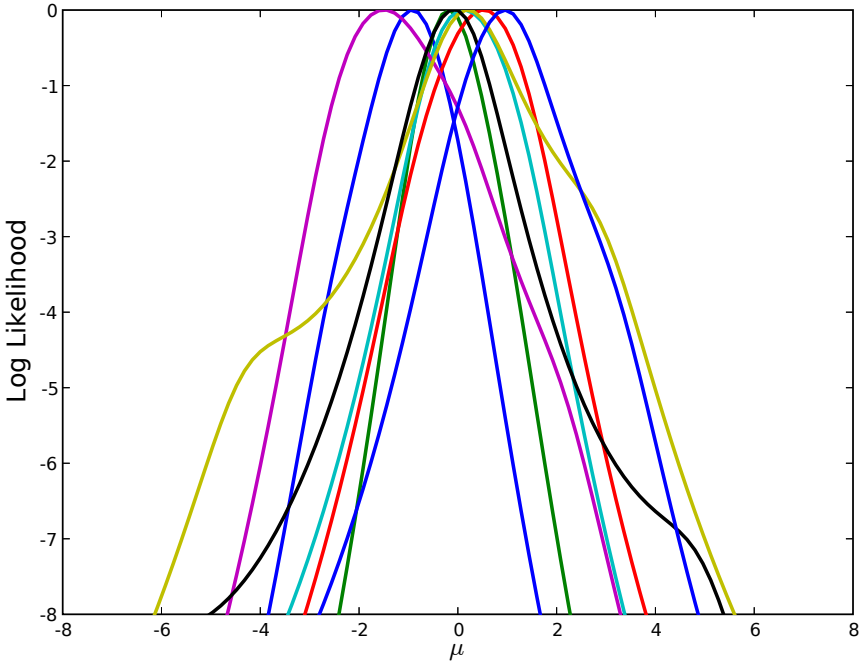
$$p(x_1, x_2|\mu) \propto p(\bar{x}, \delta|\mu) = p(\delta)p(\bar{x}|\delta, \mu)$$

The spread of the measurements determines the width of the $\bar{x}$ distribution.

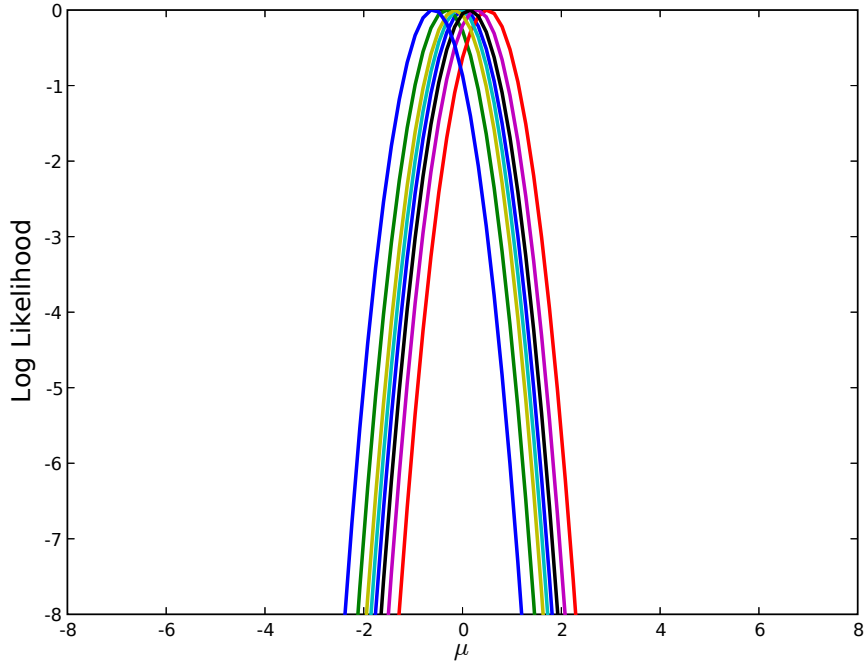

Single-sample PDF
$p(x|\mu)$
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-2
0
2
4
x
PDFs for mean
$p(\bar{x}|...)$
Unconditional
δ=0
δ=1
δ=2
δ=4
$\bar{x}$

Likelihoods for 8 samples of size $N = 5$

Cauchy



Normal



Ancillarity and relevance/precision

Ancillary statistics can provide a measure of the relevance or precision of the available data.

Good ones can be hard to identify, or hard to use if complicated.

$\mathcal{B}$ fully conditions on D_{obs}, so it automatically conditions on ancillaries when they exist.

Conditional estimation in physics

Poisson counting experiments with background: conventional intervals can have poor coverage or extend to negative signals

Improved intervals condition: use ensemble with background counts $\leq$ observed total counts (Woodroofe et al.)

Significance Tests

Almost no one does N-P testing!

Reason: The α you report must be specified *before* you analyze the data in order for α to really be the error rate.

$\rightarrow$ If you decide to reject null at $\alpha = 0.05$, that is the quantification of the evidence both for a 2σ detection and a 10σ detection.

Fisher's p-value

Fisher recommended reporting the p-value ("significance level")—the α for a test that would have just rejected the null.

Shades of conditioning/ancillarity: N-P tests seem to ignore information about the strength of the evidence in the actually observed data. p-values are an attempt to account for this.

Problem: Though the p-value is an α level, it is one chosen *based on the data*. This corrupts it as a measure of the error rate.

Significance level is not a valid error rate!

A Simple Significance Test

Model: $x_i = \mu + \epsilon_i$, ($i = 1$ to n) $\epsilon_i \sim N(0, \sigma^2)$

Null hypothesis, H_0: $\mu = \mu_0 = 0$

Test statistic:

$$t(x) = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$$

Actual Error Rate

Collect the α values from a large number of tests in situations where the truth eventually became known, and determine how often H_0 is true at various α levels.

- Suppose that, overall, H_0 was true half of the time.
- Focus on the subset with $t \approx 2$ (say, $[1.95, 2.05]$ so $\alpha \in [.04, .05]$), so that H_0 was rejected at the 0.05 level.
- Find out how many times in that subset H_0 turned out to be true.
- Do the same for other significance levels.

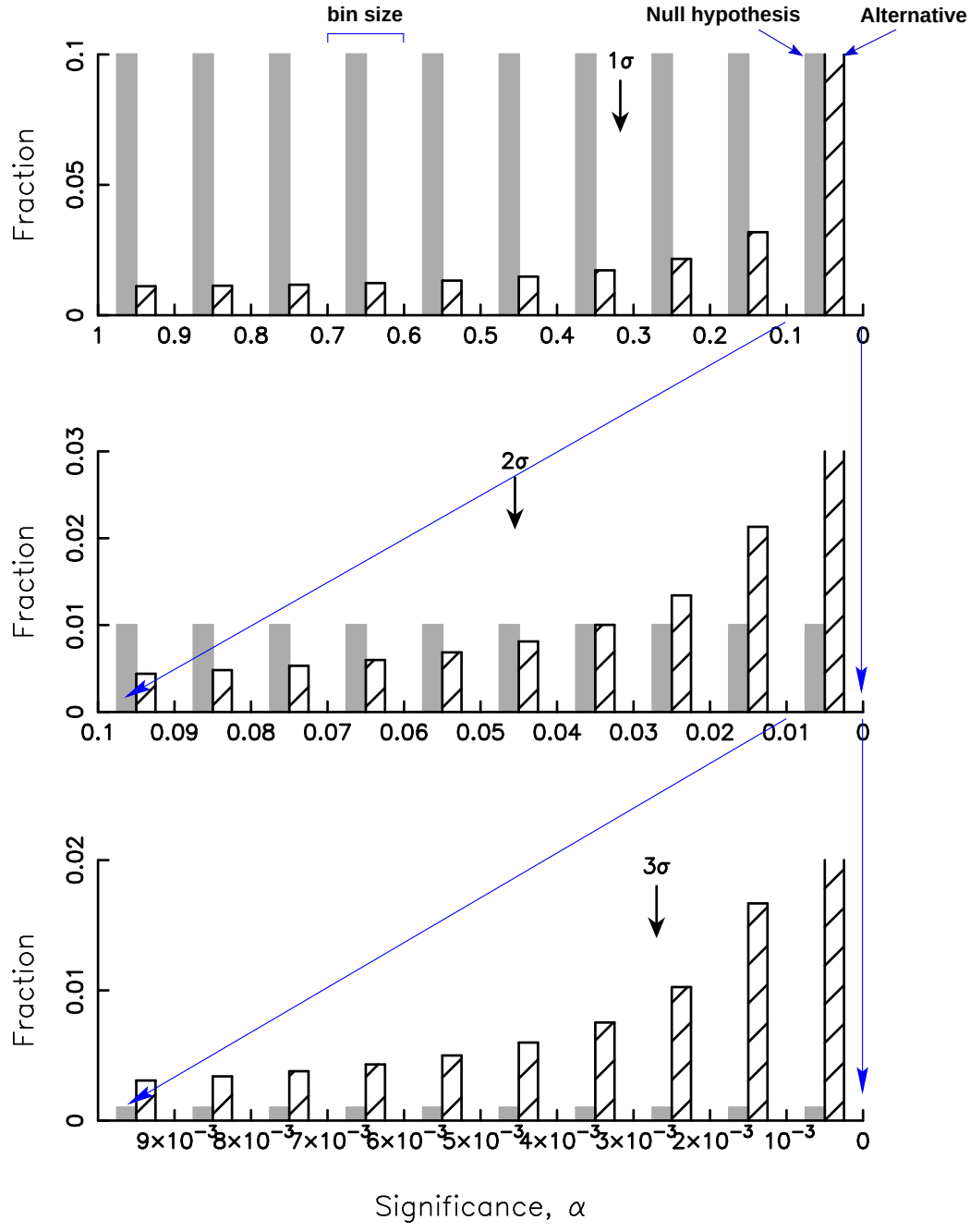
A Monte Carlo experiment:

- Choose $\mu = 0$ OR $\mu \sim N(0, 4\sigma^2)$ with a fair coin flip
- Simulate $n = 20$ data, $x_i \sim N(\mu, \sigma^2)$
- Calculate $t_{\rm obs} = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$ and $\alpha(t_{\rm obs}) = P(t > t_{\rm obs}|\mu = 0)$
- Bin $\alpha(t)$ separately for each hypothesis; repeat

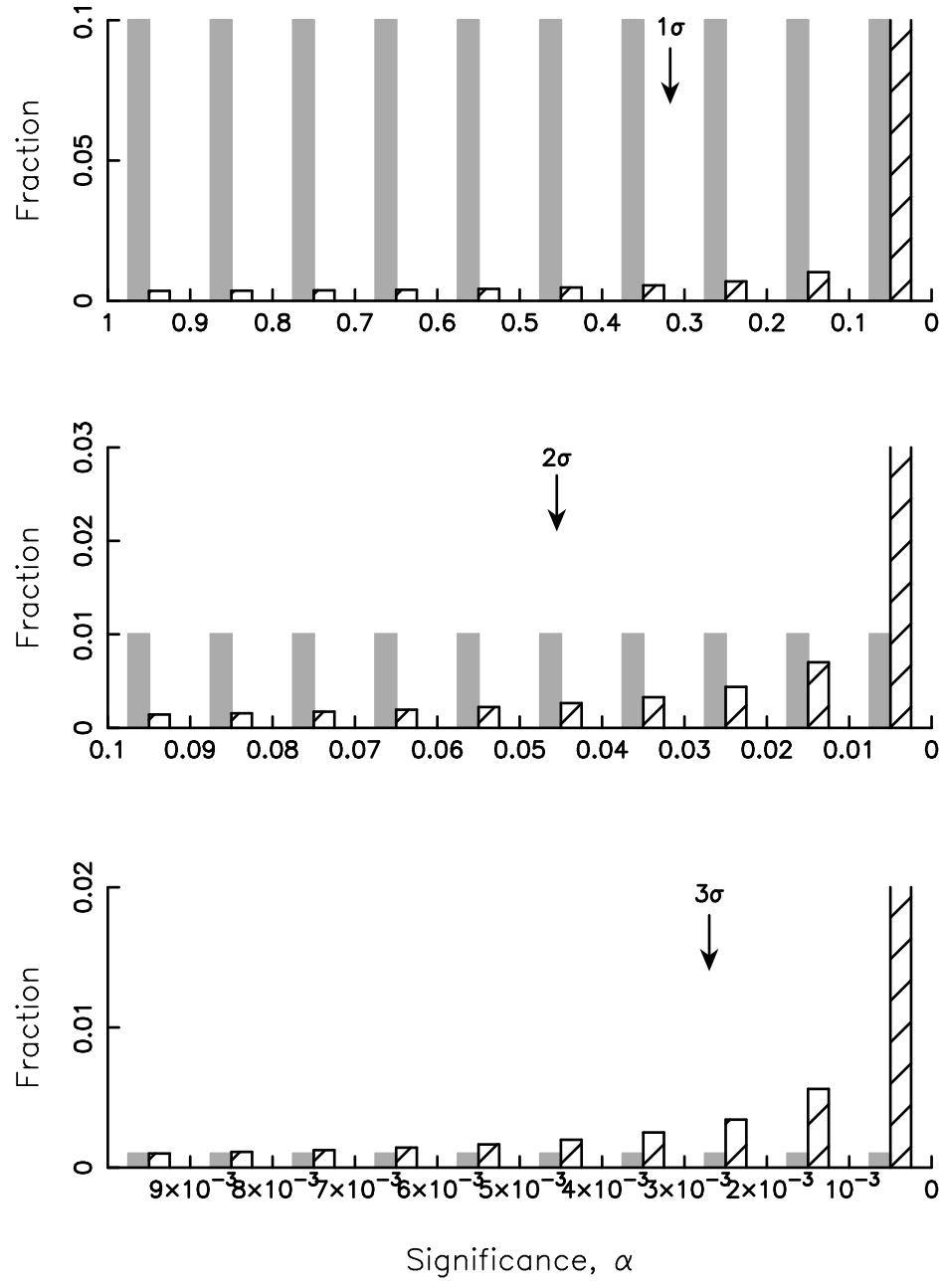
Compare how often the two hypotheses produce data with a 2– or 3–σ effect.

(Jim Berger's Java Applet will do this in real time in your web browser!)

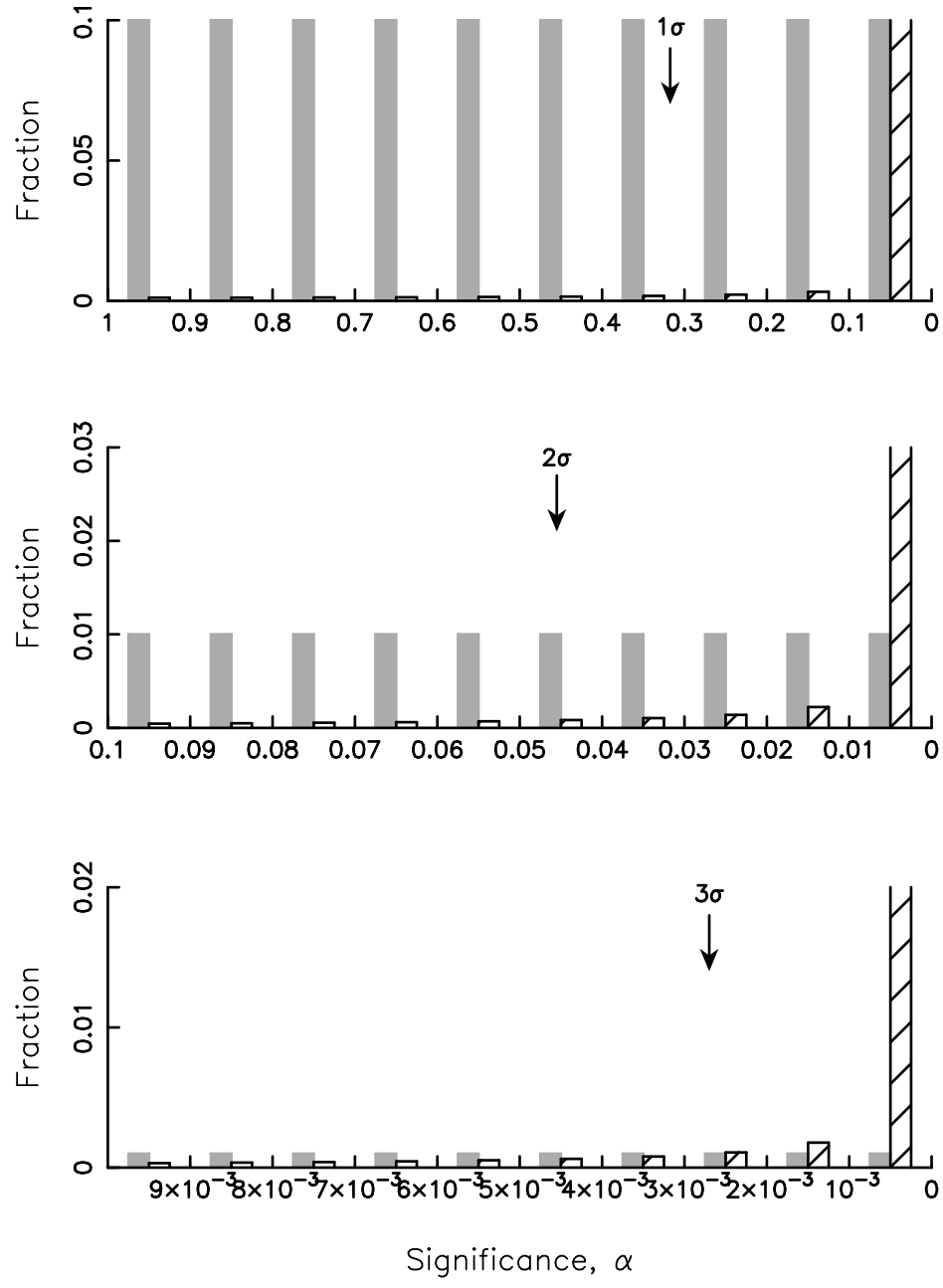
Significance Level Frequencies, $n = 20$



Significance Level Frequencies, $n = 200$



Significance Level Frequencies, $n = 2000$



What about another μ prior?

- For data sets with H_0 rejected at $\alpha \approx 0.05$, H_0 will be true *at least* 23% of the time (and typically close to 50%). (Edwards et al. 1963; Berger and Selke 1987)
- At $\alpha \approx 0.01$, H_0 will be true *at least* 7% of the time (and typically close to 15%).

What about a different "true" null frequency?

- If the null is initially true 90% of the time (as has been estimated in some disciplines), for data producing $\alpha \approx 0.05$, the null is true at least 72% of the time, and typically over 90%.

In addition . . .

- At a fixed α, the proportion of the time H_0 is falsely rejected grows as $\sqrt{n}$. (Jeffreys 1939; Lindley 1957)
- Similar results hold generically; e.g., for χ^2. (Delampady & Berger 1990)

Significance is not an easily interpretable measure of the weight of evidence against the null.

- Significance does not accurately measure how often the null will be wrongly rejected among similar data sets.
- The “obvious” interpretation overestimates the evidence.
- For fixed significance, the weight of the evidence decreases with increasing sample size. (N-P recommended decreasing α with N, but no rule is available.)
- Results above are robust w.r.t. choice of priors.

Conditional Testing

Recent school of thought tries to adapt conditioning/ancillarity ideas from parameter estimation to testing.

Basic idea: Test statistic $T(D)$ is a 1-dimensional summary of the full data set measuring the strength of the evidence in D for/against hypotheses. Evaluate *conditional error probabilities* using ensemble of all D *consistent with observed* T *value*.

$$\begin{aligned} \alpha(T) &= p(\text{reject } M_0 | T, M_0) \\ \beta(T) &= p(\text{accept } M_0 | T, M_1) \end{aligned}$$

The critical T is found by calculating the usual p-values, p_0 based on the null, p_1 on the alternative:

- If $p_0 \leq p_1$ reject M_0, report error probability $\alpha(T)$
- If $p_0 > p_1$ accept M_0, report error probability $\beta(T)$

Bayesian connections

For the Gaussian example,

$$B \equiv \frac{p(\{x_i\}|H_1)}{p(\{x_i\}|H_0)} = \frac{p(\alpha_{\rm obs}|H_1)}{p(\alpha_{\rm obs}|H_0)}$$

$\rightarrow B$ is just the ratio calculated in the Monte Carlo!

Also known that $\alpha(T)/\beta(T) = B$ (for simple hypotheses)

- Research is extending results to composite hypotheses
- Goodness-of-fit: Combine conditioning with *averaging over parameter uncertainty* improves accuracy of p-values over "plug-in" p-values.

Hypothesis Space Size

Nuisance Parameters

Almost always $\theta = \{\phi, \eta\}$ where ϕ is a subset of interest, and η are uninteresting but necessary for modelling the data.

Profile likelihood

Motivated by maximum likelihood, calculate

$$\mathcal{L}_p(\theta) = \max_{\eta} \mathcal{L}(\theta, \eta)$$

Use this like a “normal” likelihood.

Unfortunately resulting estimates can be biased, unphysical, or inconsistent; confidence intervals are too small.

Bayesian approach

No freedom—probability theory dictates use of marginal

$$\begin{aligned} p(\theta|D_{\text{obs}}) &= \int d\eta\; p(\theta,\eta|D_{\text{obs}}) \\ &\propto \int d\eta\; p(\theta)p(\eta)\mathcal{L}(\theta,\eta) \\ &\approx p(\theta)\mathcal{L}(\theta,\hat{\eta}_\theta)\frac{\delta\eta}{\Delta\eta} \end{aligned}$$

Accounts for volume of allowed part of η-space

Modern $\mathcal{F}$ approaches

- Asymptotic adjustment of profile likelihood

$$\mathcal{L}_a(\theta) = \mathcal{L}_p(\theta) \times |I_{\eta\eta}(\theta)|^{-1/2} \times \ldots$$

  with $I_{\eta\eta} =$ information matrix for η

- Evaluate $\mathcal{F}$ properties of marginal for "default" η priors

Both improve considerably on $\mathcal{L}_p$, but are difficult.

Recognizes importance of *volume in param space*.

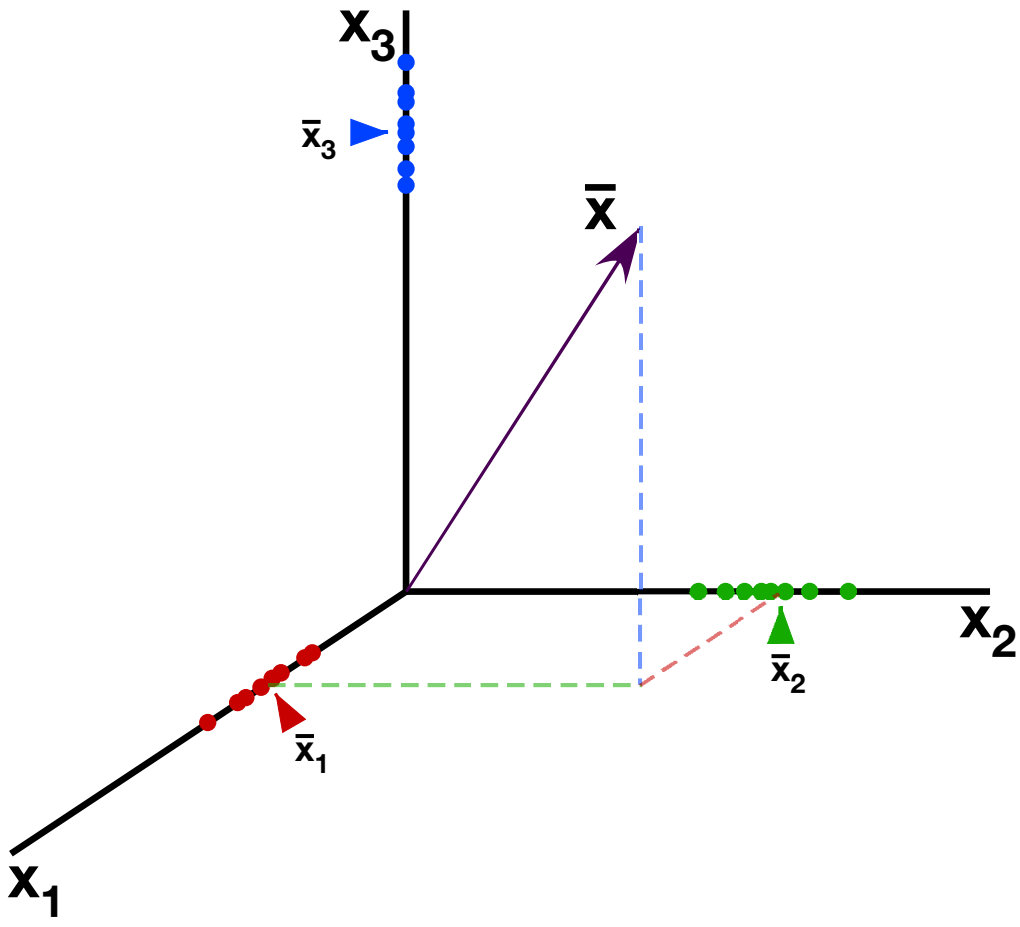
Shrinkage Estimators

Jointly estimate N normal means, μ_i, from N independent unit-variance samples x_i.

Naive estimates

- BLUE/maximum likelihood estimate is $\hat{\mu} = x$
- C% joint confidence region is sphere centered at x with radius given by C% point of χ^2_p dist'n
- Uniform prior $\rightarrow$ similar results for $\mathcal{B}$

The BLUE/Least Squares Estimator



Shrinkage estimators (Stein, James-Stein)

$\hat{\mu} = \bar{x}$ is inadmissable for squared-error loss (distance from truth) if $p > 2$!

$$\tilde{\mu} = \left[1 - \frac{p-2}{x^2}\right] \times x$$

has smaller expected squared error for all μ, despite its significant bias (toward the origin).

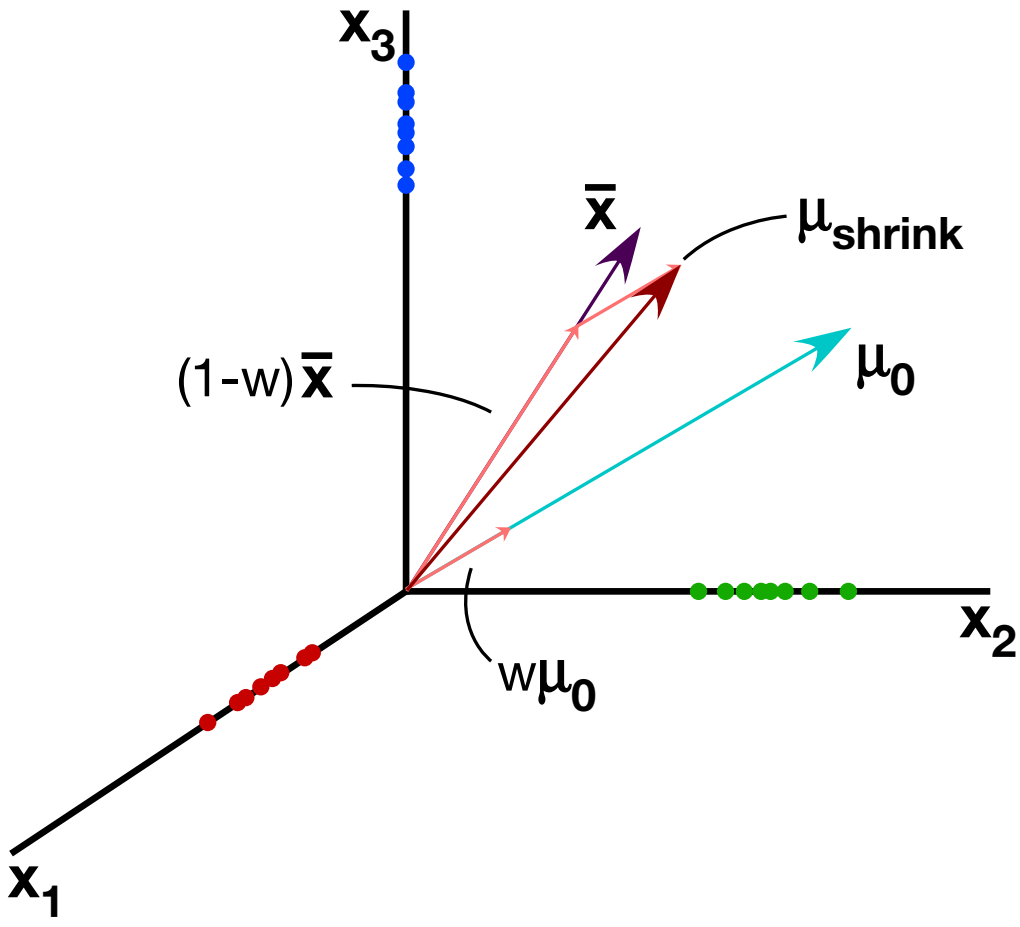
Same is true if we "shrink" toward *any* point μ_0:

$$\tilde{\mu} = w\mu_0 + (1-w) \times x$$

with w decreasing with dispersion of data around μ_0.

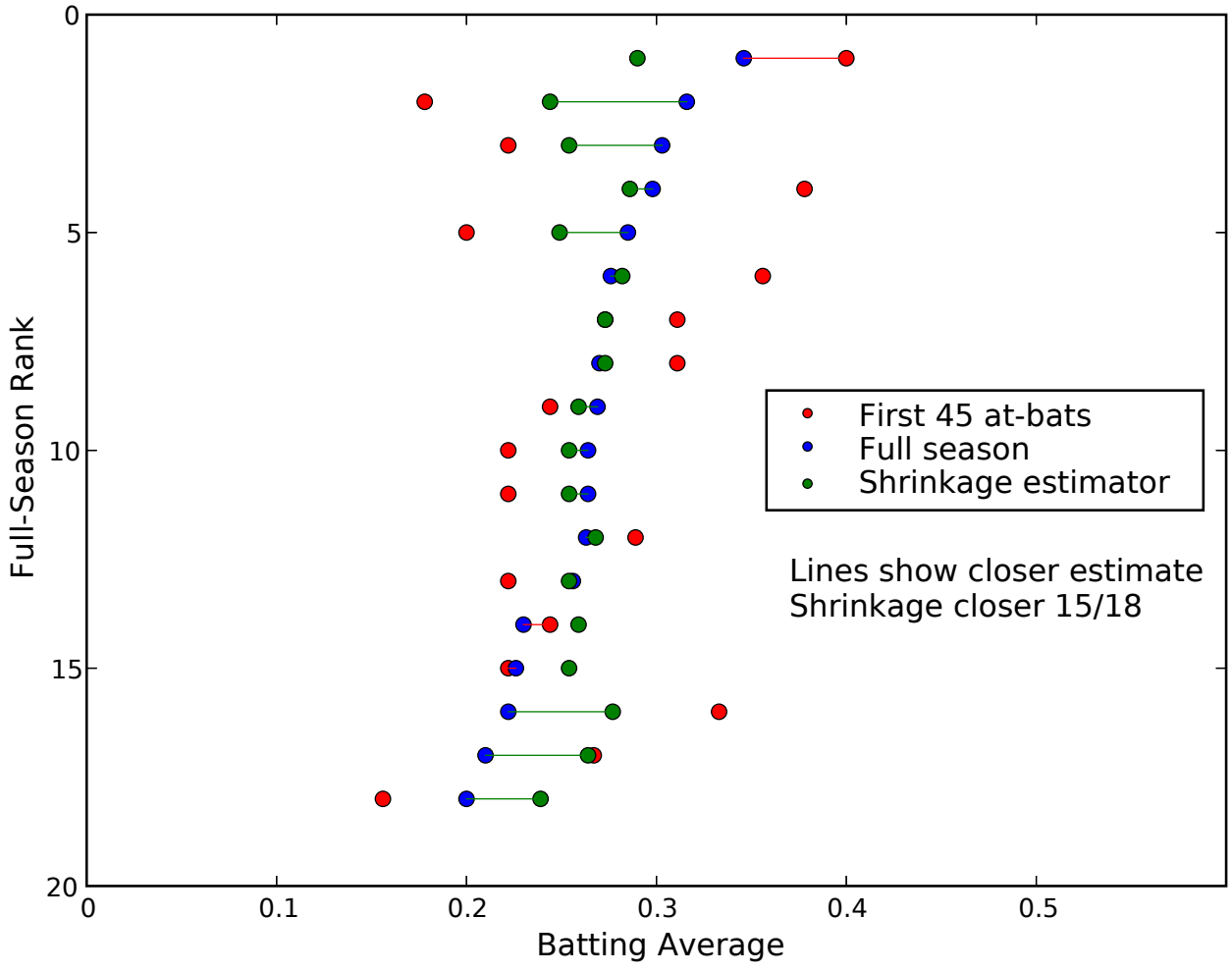
Risk is substantially reduced only if μ_0 near truth.

Shrinkage Estimators



1977 Batting Averages

Efron & Morris



Understanding shrinkage

Suppose the μ_i are related: If we were to learn the value of some of them, we'd expect the others to be like it.

Suggests a *hierarchical* Bayesian model—a prior relating the μ_i that itself has free parameters:

$$p(\mu_i|s, M) \propto N(\mu_0, s^2) \quad \text{for all } i$$

with a broad (e.g., flat, log-flat) prior on μ_0, s.

The posterior mean μ exhibits "tuned" shrinkage!

$\mathcal{F}$ study of this effect is based on *empirical Bayes*—hierarchical Bayes with $s = \hat{s}$ (integrate over μ_i but not s).

Lessons

Univariate intuition can be misleading, even about so basic a notion as the desirability of zero bias. *Biased* estimators are not *bad* estimators!

Inference with squared error loss implicitly presumes relationships among the μ_i that can be used to improve inference—the implications of a loss function are subtle!

Volume in *parameter* space is important, even for $\mathcal{F}$—the loss function provides a metric in parameter space.

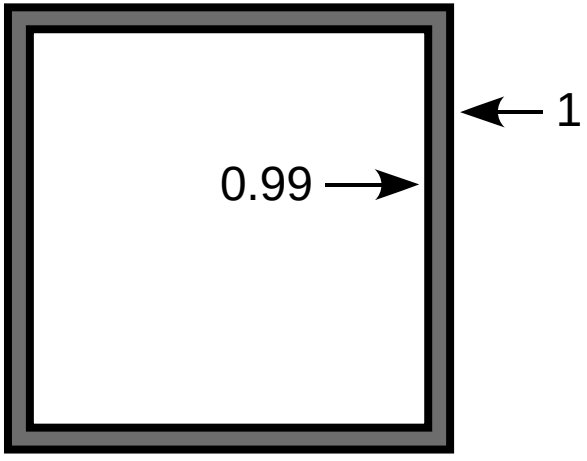
Current research explores how different relationships and priors shrink estimates.

Main applications: signal processing (wavelet analysis of images and time series), economic forecasting, epidemeology. . .

Curse of Dimensionality

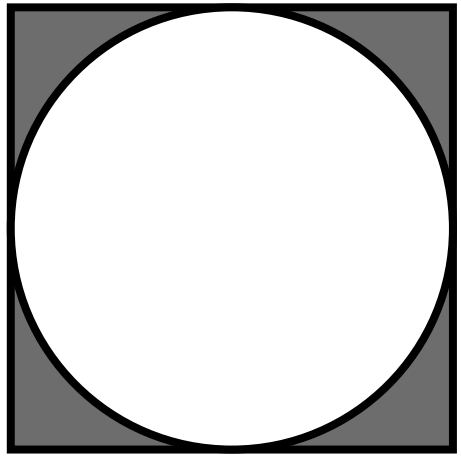
In large spaces, almost everything may be “surprising” → must carefully consider how geometries of large sample or parameter spaces influence inferences

Skin Fraction



n	f
2	0.02
10	0.1
100	0.63

Corner Fraction



n	f
2	0.21
6	0.92
10	0.998

Multiple Testing

Often we perform *many* hypothesis tests (i.e., to detect multiple signals). We will find many "detections" at a fixed α due to multiplicity.

	Detection Result:		
Source Present	*Negative*	*Positive*	Total
H_0: No	T_-	F_+	ν_0
H_1: Yes	F_-	T_+	ν_1
Total	N_-	N_+	N

$T_- = \#$ *T*rue nondetections, $F_+ = \#$ *F*alse detections, etc.

Only the Ns are known
(N_- and N_+ only after analysis)

Controlling Error Rates

For a single test, we set threshold to control the *error rate*.

Multiple tests have a variety of possible error rates; threshold will depend on which we want to control.

Two extremes

- *Per-Comparison (PCER)* — Control $\langle F_+ \rangle / N$
  I.e., ignore multiplicty
- *Family-wise (FWER)* — Control $P(F_+ \geq 1)$
  This requires *Bonferroni correction*: For N tests, pick threshold so α is reduced by M.

Compromise: False Discovery Rate (FDR)

- Control $\sim F_+/N_+$ — Fraction of all discoveries that are false

Benjamin Hochberg FDR

Threshold must be set before you get the data $\rightarrow$ neither F_+ nor N_+ are known.

BH showed a simple algorithm can control

$$\begin{aligned} FDR &= E\left[\frac{F_+}{\max(N_+,1)}\right] \\ &= E\left[\frac{F_+}{N_+} \mid N_+ > 0\right] P(N_+ > 0) \end{aligned}$$

To assure $\langle FDR \rangle \leq \alpha$:

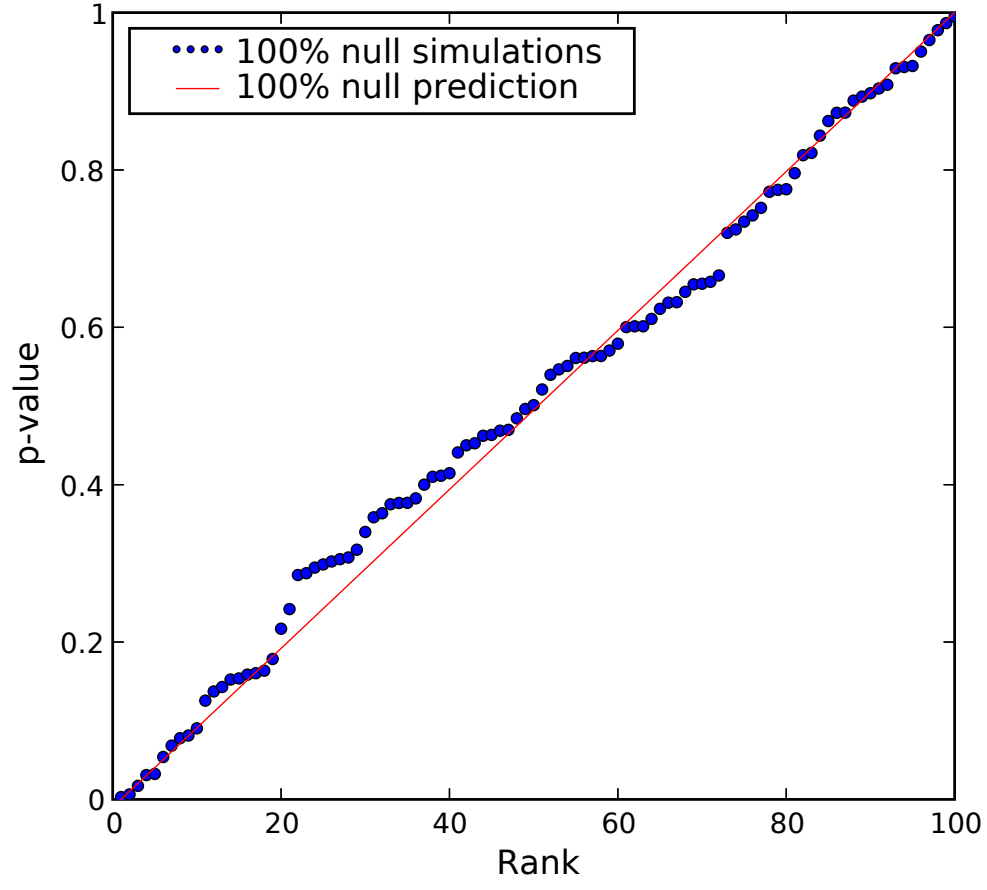
- List p-values in increasing order, $P_1, \ldots, P_N$
- Find $d = \max\left(j \ : \ P_j < \frac{j\alpha}{N}\right)$
- Reject null for tests with $P < P_d$

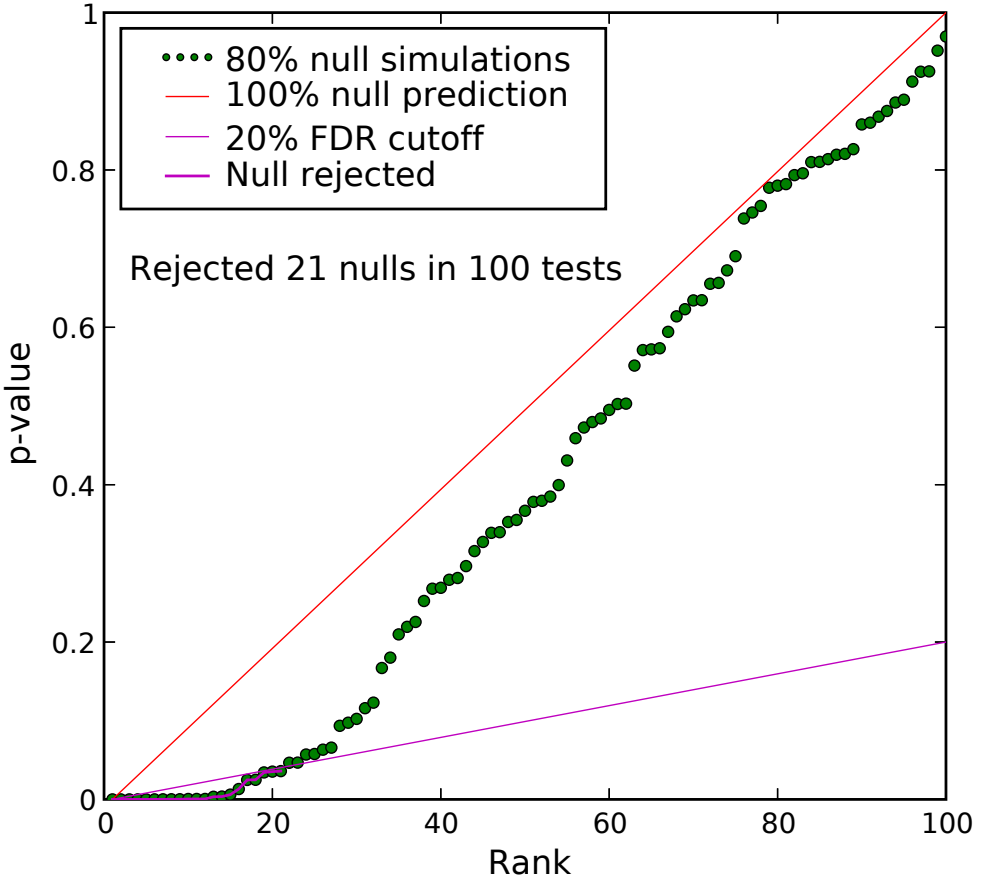
Works without specifying the alternative!

Simple FDR Example: Two Gaussians

M_0: $x_i \sim N(0,1)$ $\qquad\qquad$ M_1: $x_i \sim N(1,1)$

100 tests with sample size $\nu = 10$





Challenges for FDR

- Controls *expectation* of F_+/N_+; research on providing confidence bands
- Alternative: Prop'n of false discoveries $\langle F_+ \rangle / \langle N_+ \rangle$; has better decision-theoretic justification
- Alternative: Expected (Type I) Error Rate $\langle F_+ \rangle / N$
- When are *any* of these appropriate? E.g., for targeted pulsar searches (with a single expected period) Bonferroni seems correct.
- Does yes/no detection really address the ultimate questions? (Classification uncertainty)

Bayesian Multiple Testing

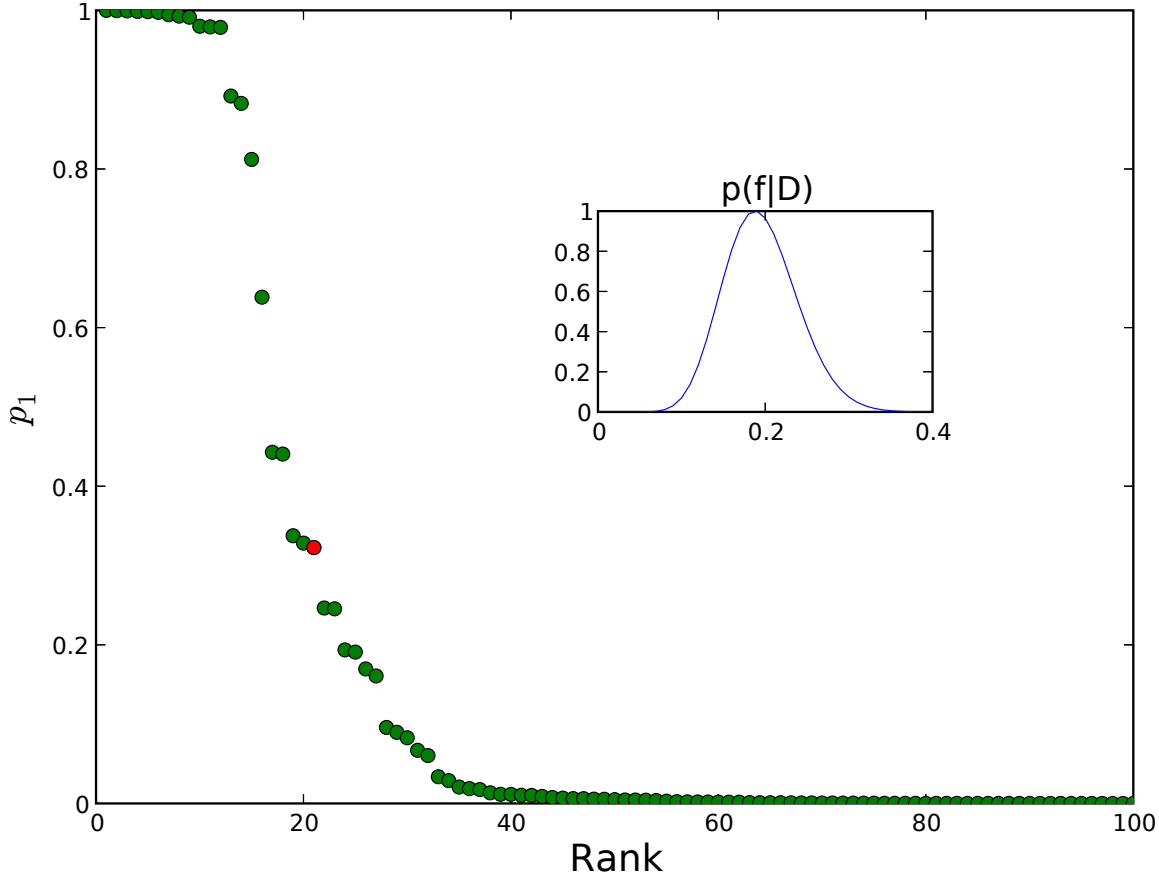
If you can specify the alternatives Bayesian calculations correct for multiplicity automatically and appropriately via model-averaging and Occam's razor.

This is accomplished without "hard classification"—Bayes factors assign weights for each candidate detection.

Two Gaussians: Bayesian Classification

Take $f =$ fraction from M_1 to be an unknown parameter.

Estimate f and probability that each sample is from M_0 or M_1.



Summary

Statistics “beyond the textbook” is subtle and challenging, but the challenges are being met.

$\mathcal{F}$ and $\mathcal{B}$ approach the challenges in very different ways, but there is intriguing cross-fertilization/convergence:

- The major $\mathcal{F}$ developments borrow two key ideas from $\mathcal{B}$:
  - Conditioning
  - Accounting for size of hypothesis space
- The thorny problem of priors (esp. in many dimensions) are being addressed by analyzing $\mathcal{F}$ performance of $\mathcal{B}$ procedures

There is still plenty left to do!

References

Background:

- Casella & Berger (2001): “Estimation: Point and Interval” and “Hypothesis Testing in Statistics,” *International Encyclopedia of the Social and Behavioral Sciences*. Available at `http://www.west.asu.edu/rlberge1/papers.html`.
  Brief overviews of the essential ideas of frequentist statistics; exceptionally clear.
- C. Glymour et al. (1997): “Statistical Themes and Lessons for Data Mining,” *Data Mining and Knowledge Discovery*, **1**, 11–28. `http://tinyurl.com/9kldd`
  An influential article by leading experts on statistical aspects of data mining, highlighting key issues arising in mining large data sets.

Decision theory and Bayesian/Frequentist foundations & relationships:

- Michael Newton: “On Statistical Decision Theory,” course handout for Stat 775 at U. Wisconsin-Madison, at `http://www.stat.wisc.edu/~newton/st775/`; 4 pp summary of essentials of frequentist decision theory, admissibility of Bayes.
- Berger & Bayarri (2004): “The interplay between Bayesian and frequentist analysis” (*Statistical Science*, **19**)
  `http://www.isds.duke.edu/~berger/papers/interplay.html`
- Ed Jaynes (2003): *Probability Theory: The Logic of Science*; very clear discussion of decision theory in Ch. 13, 14, targeted to physical scientists & engineers.

Ancillary statistics, conditional estimation:

- Ed Jaynes (1976): “Confidence Intervals vs. Bayesian Intervals,” reprint available as Paper 32 at `http://bayes.wustl.edu/etj/node1.html`
- Nancy Reid (1995): “The Roles of Conditioning in Inference,” *Stat. Sci.*, **10**, 138–199.
- Brad Efron (2003): “Bayesians, Frequentists, and Physicists”
  From PhyStat2003:
  `http://www.slac.stanford.edu/econf/C030908/proceedings.html`

Significance test controversy, conditional testing:

- Understanding P-values:
  `http://www.isds.duke.edu/~berger/p-values.html`
  Jim Berger’s web site with links to other web sites covering issues with significance tests, and Berger’s p-value Java applet.
- Berger & Bayarri (2003), cited above.

Nuisance parameters:

- Berger, Liseo, & Wolpert (1999): “Integrated Likelihood Methods for Eliminating Nuisance Parameters” (*Statistical Science*, **14**, with discussion) `http://www.isds.duke.edu/~berger/papers/brunero.html`
- Nancy Reid (2003): “Likelihood inference in the presence of nuisance parameters” (PhyStat2003, URL above)

Shrinkage:

- Stein (1962): “Confidence sets for the mean of a multivariate normal distribution,” *JRSS*, **B24**, 265–296. See Lindley’s discussion for a Bayesian viewpoint.
- Lindley & Smith (1972): “Bayes Estimates for the Linear Model,” *JRSS*, **B72**, 1–41. Connection between shrinkage and hierarchical Bayes models.
- Efron & Morris (1973): “Combining Possibly Related Estimation Problems,” *JRSS*, **B35**, 379–421. Shrinkage and empirical Bayes; see esp. the discussion.
- Brad Efron (1975): “Biased Versus Unbiased Estimation,” *Adv. in Math.*, **16**, 259–277. Early, practical review.

Shrinkage, cont'd:

- Jim Berger: "The Stein Effect," entry in *Encyclopedia of Statistical Sciences*.
- Stephen Stigler (1990): "A Galtonian Perspective on Shrinkage Estimators," *Stat. Sci.*, **5**, 147-155.
- Brad Efron (2003): "Bayesians, Frequentists, and Physicists" (PhyStat2003)

Multiple testing:

- Alex Lewin: "Multiple Testing," slides available at `http://www.bgx.org.uk/alex/LewinMultTest2005.pdf`; nice tutorial.
- Hopkins et al. (2002): "A new source detection algorithm using the false-discovery rate," *ApJ*, **123**, 1086–1094.
- Genovese & Wasserman (2003): "Bayesian and frequentist multiple testing" (*Bayesian Statistics 7*); tech. report 764 at `http://www.stat.cmu.edu/`.
- Brad Efron (2005): "Bayesians, Frequentists, and Scientists" (*JASA*, **100**, 1–5)
- Berger & Bayarri (2004), cited above, section 5; see also their ISBA 2004 talk, `http://isba.mat.puc.cl/abstract/down.php?file=1081633131_multcomp.pdf`