

TELLING THE TRUTH WITH STATISTICS  
F. JAMES, ACADEMIC TRAINING 10-14 JAN, 2008

①

# PRINCIPLES AND PROBLEMS IN STATISTICS

NOT A COOK BOOK

PLAN: COMPARISON OF

BAYESIAN



FREQUENTIST

APPROACHES

ALL BASED ON DEFINITION OF

PROBABILITY:

- "MATHEMATICAL"

- "FREQUENTIST"

- "BAYESIAN"

# "MATHEMATICAL" PROBABILITY

(2)

SATISFIES AXIOMS: (KOLMOGOROV)

1.  $0 \leq P(A) \leq 1$

2.  $P(\Omega) = 1$

3.  $P(A \cup B) = P(A) + P(B)$  if  $(A \cap B) = \emptyset$

- ABSTRACT PROBABILITY

- DOES NOT NECESSARILY CORRESPOND TO ANYTHING IN THE REAL WORLD.

- MANY THEOREMS - ONE OF THEM DEFINES CONDITIONAL PROBABILITY:

$$P(A|B) P(B) = P(B|A) P(A)$$

↑  
(A, GIVEN B)

THIS IS KNOWN AS BAYES' THEOREM

# FREQUENTIST PROBABILITY (3)

EVENT A      EXAMPLE:  $\Lambda^0 \rightarrow p\pi^-$   
EVENT SPACE  $\Omega$       EXAMPLE:  $\Lambda^0 \rightarrow \text{anything}$   
 $N(A)$ ,  $N(\Omega)$       NUMBER OF OBSERVED EVENTS

$$P(A) = \lim_{N(\Omega) \rightarrow \infty} \frac{N(A)}{N(\Omega)}$$

## LIMITING FREQUENCY DEFINITION

SOME OBSERVATIONS ABOUT FREQUENTIST PROB.

1. IT IS NOT NECESSARY TO EVALUATE THE LIMITING FREQUENCY UNLESS YOU NEED THE NUMERICAL VALUE.  
MUST ONLY BE POSSIBLE IN PRINCIPLE.

2. THIS IS THE PROBABILITY OF QUANTUM MECHANICS

$$P(\underline{x} \in \underline{D}) = \int_{\underline{D}} \psi^2 dx$$

BUT IT EXISTED BEFORE Q.M.

3. RESTRICTED TO REPENTABLE EXPERIMENTS

4. IN GENERAL, AN INTRINSIC PROPERTY OF THE SYSTEM. [EXAMPLES: B.R. OF  $\Lambda^0 \rightarrow p\pi^-$ ]

5. SOME BAYESIANS (DEFINETTI & CO.) DENY THE EXISTENCE OF FREQ. PROB.

# BAYESIAN PROBABILITY

④

MORE GENERAL, APPLIES TO NON-REPEATABLE EXPTS.

1.  $P(\text{It will rain tomorrow})$
2.  $P(\text{I will die before a certain date})$
3.  $P(M_{\text{Higgs}} > 95 \text{ GeV})$

DAVID HUME (1748): "DEGREE OF BELIEF"

THOMAS BAYES (1763): BAYES THEOREM

BRUNO DE FINETTI (1936): THE FAIR BET.

---

SOME OBSERVATIONS ABOUT BAYESIAN PROB.

1. " $P(E)$  IS NOT AN INTRINSIC PROPERTY OF  $E$ , BUT DEPENDS ON THE STATE OF INFORMATION AVAILABLE TO WHOEVER EVALUATES  $P(E)$ "  
[DIAGOSTINI YELLOW REPORT 99-03]
2. IT IS ALWAYS CONDITIONAL ON SOME THING, SO IT SHOULD BE WRITTEN  $P(E|I)$
3. OFTEN USED IN EVERYDAY REASONING  
[EX.: "HE IS PROBABLY LYING."]
4. OFTEN SUBJECTIVE: CANNOT BE VERIFIED OR FALSIFIED IN MOST CASES.

# BAYESIAN VS. FREQUENTIST (5)

## PROBABILITY OF DRAWING A RED BALL FROM A MATHEMATICAL URN

$N(R)$  = Number of Red Balls in Urn  
 $N(W)$  = Number of White Balls in Urn  
 $N_T = N(R) + N(W)$

GIVEN	FREQUENTIST $P_2(R)$ **	BAYESIAN $P_1(R)$ **
$N(R) \geq 0$ , $N(W) \geq 0$ , $N_T > 0$	$P_2(R) = \frac{N(R)}{N_T}$ = unknown	$P_1(R) = \frac{1}{2}$ * PASCAL'S PRINCIPLE OF "INSUFFICIENT REASON"
100 BALLS HAVE BEEN DRAWN, 40 WERE RED	$P_2(R) = 0.40 \pm .05$	$P_1(R) = 0.405$ *
1 BALL HAS BEEN DRAWN, IT WAS RED	$P_2(R) \neq 0$	$P_1(R) = 0.6$ *
$N(R) = N(W) = 1$	$P_2(R) = \frac{1}{2}$	$P_1(R) = \frac{1}{2}$

\* - Light red, dark red?

\* - "typical" Bayesian value

\*\* - Notation  $P_1, P_2$  due to Carnap, Seidenfeld

# DECISION THEORY [GREATLY SIMPLIFIED]

EXAMPLE DECISION: WHETHER OR NOT TO TAKE AN UMBRELLA TO WORK TOMORROW

OBSERVABLE SPACE  $\Theta$  :  $\begin{cases} R = \text{it rains tomorrow} \\ \bar{R} = \text{no rain} \end{cases}$

DECISION SPACE  $\mathcal{D}$  :  $\begin{cases} u = \text{take umbrella} \\ \bar{u} = \text{do not take it} \end{cases}$

LOSS FUNCTION  $\mathcal{L}(\mathcal{D}, \Theta)$  :

	$\bar{R}$	$R$
$u$	1	1
$\bar{u}$	0	3

DECISION RULE:

1. BAYESIAN DECISION RULE: MINIMIZE EXPECTED LOSS

$$E(\mathcal{L})_u = 1 \cdot P(\bar{R}) + 1 \cdot P(R) = 1$$

$$E(\mathcal{L})_{\bar{u}} = 0 \cdot P(\bar{R}) + 3 \cdot P(R) = 3 \cdot P(R)$$

$\Rightarrow$  TAKE UMBRELLA IF  $P(R) > \frac{1}{3}$

2. MINIMAX DECISION RULE: MINIMIZE MAXIMUM LOSS

$\Rightarrow$  ALWAYS TAKE UMBRELLA

# BAYES' THEOREM (1763)

$$P(A|B) P(B) = P(B|A) P(A)$$

EXAMPLE: PARTICLE DETECTOR DESIGNED TO IDENTIFY PROTONS HAS THE PROPERTY:

$$P(\text{hit} | p) = 0.90$$

$$P(\text{hit} | \pi) = 0.05$$

NOW AN UNKNOWN PARTICLE GOES THROUGH THE DETECTOR AND PRODUCES A HIT.

WHAT IS THE PROBABILITY THAT IT WAS A PROTON?

$$P(p|\text{hit}) = P(\text{hit}|p) \frac{P(p)}{P(\text{hit})}$$

$$\begin{aligned} P(\text{hit}) &= P(\text{hit}|p) P(p) + P(\text{hit}|\pi) P(\pi) \\ &= 0.9 P(p) + 0.05 P(\pi) \end{aligned}$$

$$P(p|\text{hit}) = \frac{0.9 P(p)}{0.9 P(p) + 0.05 P(\pi)}$$

# BAYESIAN USE OF BAYES' THEOREM TO ESTIMATE PARAMETERS $\mu$

$$P(\mu/\text{data}) = \frac{P(\text{data}/\mu) \cdot P(\mu)}{P(\text{data})}$$

$$P(\text{data}) = \int_{\mu} P(\text{data}/\mu) P(\mu) d\mu \quad \text{is only a normalization constant}$$

$$P(\mu/\text{data}) = P(\text{data}/\mu) \cdot P(\mu)$$

↑  
POSTERIOR DENSITY  
(p.d.f.)

↑  
PRIOR DENSITY  
(p.d.f.)

↑  
LIKELIHOOD FUNCTION  
(function of  $\mu$ )

NOTE: NONE OF THESE P's ARE  
PROBABILITIES !!

p.d.f. = Probability density function,  
must be integrated between 2 values  
of  $\mu$  to obtain a probability

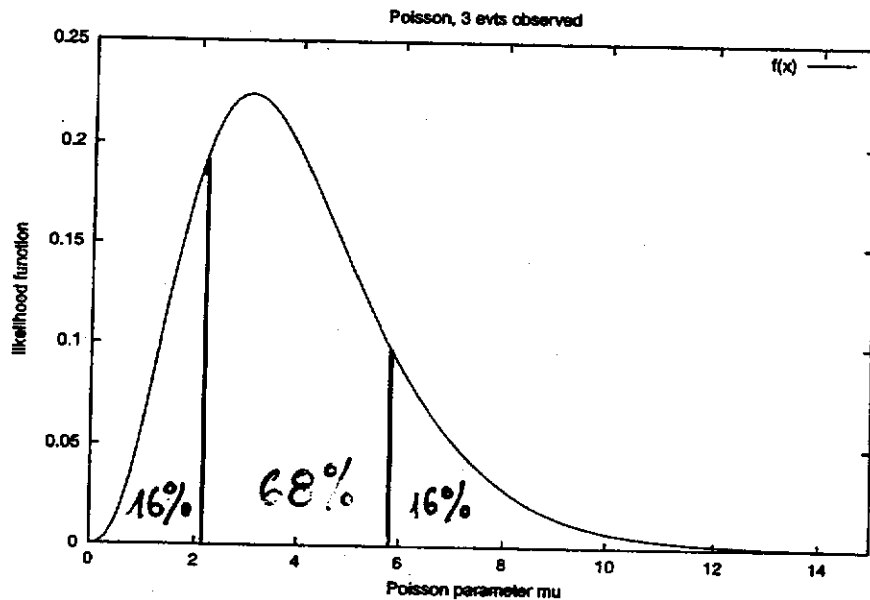
L.F. was a probability density  
before we evaluated it at the  
observed data



# EXAMPLE OF BAYESIAN ESTIMATION OF PARAMETER $\mu$ OF POISSON DISTR.

ASSUME WE HAVE OBSERVED 3 EVTS.

$$\begin{aligned} P(\mu|3) &= P(3|\mu) \cdot P(\mu) \\ &= \frac{e^{-\mu} \mu^3}{3!} \cdot P(\mu) \end{aligned}$$



$$P(2.09 < \mu < 5.92) = 68\%$$

CENTRAL CONFIDENCE INTERVAL

(10)

# SUMMARY

## THE BAYESIAN METHOD

1. PROVIDES THE BEST KNOWN  
DECISION RULE
2. PROVIDES AN ELEGANT WAY  
TO UPDATE PREVIOUS BELIEFS  
WITH NEW INFORMATION

I'm definitely aware of Bayesian approaches to statistics, and am generally sympathetic to them as a way of quantifying the beliefs (subjective probability distribution) of a person and of explaining how those beliefs should be rationally modified in the face of new evidence.

Alan Sokal, 1999

3. DOES NOT PROVIDE A CREDIBLE  
WAY TO SUMMARIZE RESULTS  
INDEPENDENTLY OF ANY  
PREVIOUS BELIEFS & KNOWLEDGE.

# MAJOR BAYESIAN PROBLEM:

PRIOR DENSITY  $P(\mu)$

(1) SUBJECTIVITY:

$P(\mu)$  GIVES PHYSICIST'S PRIOR "FEELINGS"  
ABOUT POSSIBLE VALUES OF  $\mu$ .

DIFFERENT PHYSICISTS

→ DIFFERENT FEELINGS

→ DIFFERENT RESULTS  
QUOTED FOR THE SAME  
EXPERIMENTAL DATA!

(2) HOW TO EXPRESS TOTAL IGNORANCE?

$$P(\mu) = \text{constant?} \quad (0 \leq \mu < \infty)$$

⇓

$$P(0 \leq \mu \leq 10) = P(2000 \leq \mu \leq 2010)$$

# TOTAL IGNORANCE (continued)

(12)

## PROPERTIES OF THE UNIFORM PRIOR

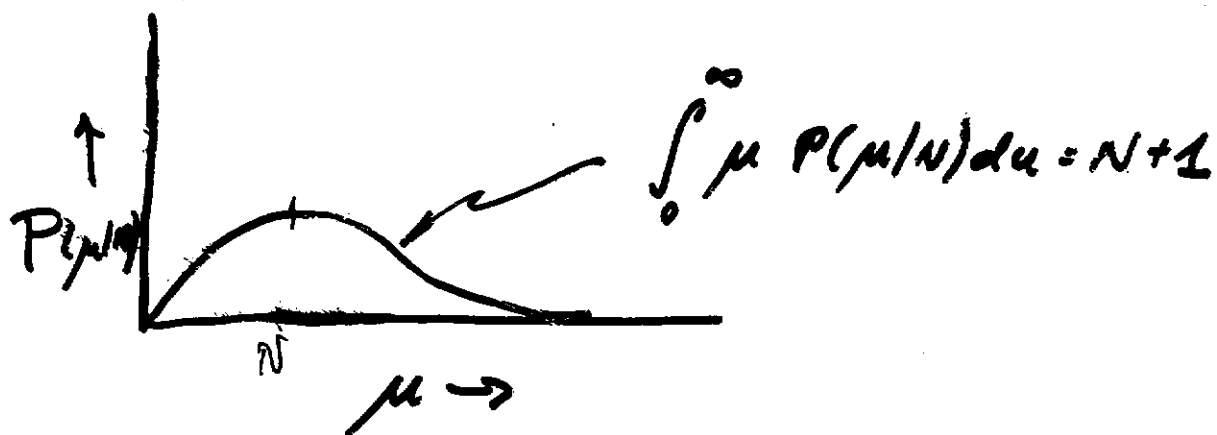
IF  $N$  EVENTS ARE OBSERVED,  
THE POSTERIOR DENSITY GIVES EXPECTATION

$$E(\mu) = N + 1$$

for Poisson:

$$P(N|\mu) = \frac{e^{-\mu} \mu^N}{N!}$$

THAT IS, YOU GET:



YOU MIGHT PREFER  $E(\mu) = N$

SINCE, FOR POISSON DISTRIBUTION

$$E(N) = \mu$$

$$[E N P(N/\mu) = \mu]$$

# TOTAL IGNORANCE (still continued)

(13)

## THE ARGUMENT OF JEFFREYS

### SCALE INVARIANCE

SUPPOSE TWO EXPERIMENTERS MEASURE THE PROTON DELAY RATE  $\theta$ .

THEY EXPECT IN TIME  $t$  :

$$P(N/\theta) = \frac{(\theta t)^N e^{-\theta t}}{N!}$$

BUT THEY USE DIFFERENT UNITS OF TIME,  
RELATED BY  $t_1 = g \cdot t_2$

$$\text{THEREFORE } \theta_2 = g \cdot \theta_1$$

THEIR PRIOR DENSITIES ARE :  $\begin{cases} f_1(\theta_1) \\ f_2(\theta_2) \end{cases}$

BUT BOTH WISH TO EXPRESS THE SAME  
TOTAL IGNORANCE, SO FUNCTIONAL FORM  
MUST BE SAME :  $f_1(\theta) = f_2(\theta)$

$$\therefore \boxed{f(\theta) = g \cdot f(g\theta)}$$

$$\therefore \boxed{f(\theta) = \text{const.} / \theta}$$

SCALE-  
INVARIANT  
PRIOR DENSITY

still TOTAL IGNORANCE

PHYSICAL EVIDENCE FOR  $1/x$ .

ARE OBJECTS IN NATURE  
DISTRIBUTED LIKE  $1/x$  ?

SIZES OF LAKES  
PLANTS  
ANIMALS  
MOLECULES

THE DISTRIBUTION OF FIRST DIGITS

LOGARITHM TABLES + SLIDE RULES

---

HAVE WE FINALLY FOUND

ABSOLUTE TRUTH ?

MORE TOTAL IGNORANCE

PROPERTIES OF THE

$\frac{1}{\mu}$  PRIOR

USING  $\frac{1}{\mu}$ :

$$E(\mu) = N$$

THIS IS MUCH NICER THAN THE  
UNIFORM PRIOR, EXCEPT FOR

$$N=0$$

IF NO EVENTS ARE OBSERVED,  
THEN BAYES +  $\frac{1}{\mu}$  SAYS:

THERE IS ZERO PROBABILITY<sub>1</sub> THAT  
 $\mu_{\text{TRUE}}$  COULD BE ANYTHING  
DIFFERENT FROM ZERO!

THINGS ARE NOT LOOKING GOOD  
FOR ABSOLUTE TRUTH!

# END OF TOTAL IGNORANCE

IT WOULD APPEAR THAT:

UNIFORM PRIOR  $\Rightarrow$  FAVOURS LARGE VALUES OF  $\mu$  TOO MUCH

$1/\mu$  PRIOR  $\Rightarrow$  FAVOURS SMALL VALUES TOO MUCH.

SO WHAT ABOUT:

(1)  $1/\sqrt{\mu}$  ?

(2)  $1/\mu$  WITH CUTOFFS TO AVOID ?  
THE INFRARED CATASTROPHE.

(3) SOMETHING ELSE ?



# CURRENT THINKING ABOUT NON-INFORMATIVE PRIORS (N.P.)

KASS + WASSERMAN (1996): N.P. ARE CHOSEN  
BY PUBLIC AGREEMENT, LIKE UNITS SYSTEMS.

J. BERNARDO (1979): REFERENCE PRIORS  
BERGER & BERNARDO (1979)

BOX & TIAO (1973) } N.P. HAVE MINIMAL EFFECT  
BERNARDO & SMITH (1994) } RELATIVE TO THE DATA.

J. BERNARDO (1997) "N.P. DO NOT EXIST."

PERICCHI + WALLEY (1991): LARGE CLASSES OF  
PRIOR DISTRIBUTIONS ARE NEEDED



THE DOMINANT BAYESIAN SCHOOL  
IS NOW DE FINETTI (SUBJECTIVE)

G. D'AGOSTINI (1999):

"OVERCOMING PRIORS ANXIETY" TELLS US  
WE SHOULD NOT BE ASHAMED OF SUBJECTIVISM.

# THE DEVELOPMENT OF "CLASSICAL STATISTICS"

13

KARL PEARSON - CHI-SQUARE TEST  
(1900)

↕ (BIOLOGISTS) -  
DARWIN?

R. A. FISHER - MAXIMUM LIKELIHOOD

EXACT FREQUENTIST THEORY

J. NEYMAN - CONFIDENCE  
INTERVALS

+ E. G. PEARSON - HYPOTHESIS TEST

+ A. WALD - DECISION THEORY

BY ~1935, THE "NEW STATISTICS"  
HAD COMPLETELY REPLACED BAYESIAN  
THINKING.

NEO-BAYESIAN MOVEMENT

B. DE FINETTI, J. BERNARDO, J. BERGER

PHYSICISTS: H. JEFFREYS, E. T. JAYNES

TIME  
↓

# R. A. FISHER

## ■ MOTIVATION, PLAN:

- WE NEED A NEW THEORY

- MUST ELIMINATE

"PRIOR PROBABILITIES"

THEY ARE ARBITRARY

SUBJECTIVE

DO NOT DISTINGUISH BETWEEN

$\left\{ \begin{array}{l} P = \text{KNOWN} \\ P = \text{NOT KNOWN} \end{array} \right.$

## ■ STRATEGY: - USE LIKELIHOOD

- DEFINE & EVALUATE PROPERTIES

- REDUCE PROBLEMS TO STANDARD FORM.

## ■ SUCCESSES:

- DEVELOPED TECHNIQUES EVERYONE USES

- ESTABLISHED CRITERIA FOR NEW TECHNIQUES

## ■ FAILURES:

- HE GOT CONFIDENCE INTERVALS WRONG

- THEORY WAS "AD HOC"

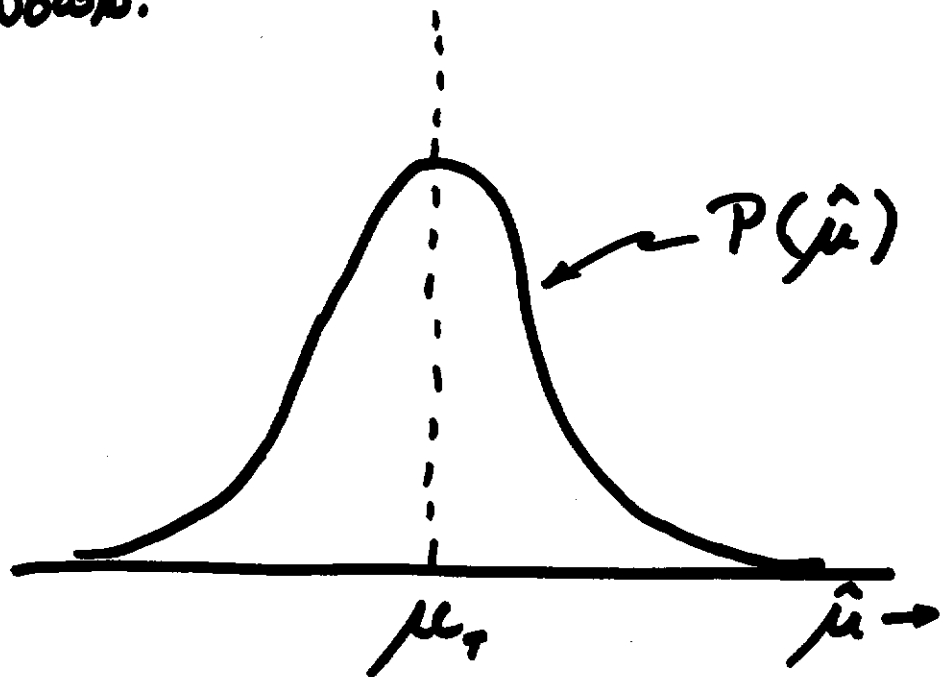
- THE EXACT THEORY WOULD BE GIVEN BY NEYMAN, PEARSON, WALD, OTHERS.  $\uparrow$  E.G.

# "CLASSICAL" PARAMETER ESTIMATION (R.A.F.)

AN ESTIMATOR  $\hat{E}$  IS A METHOD OR ALGORITHM TO CALCULATE AN ESTIMATE  $\hat{\mu}$  OF A PARAMETER WHOSE TRUE VALUE IS  $\mu_T$ .

$$\hat{\mu} = \hat{E}(\text{data})$$

SINCE THE DATA IS RANDOM, WE KNOW ONLY  $P(\text{data} | \mu)$ , AND  $\hat{\mu}$  IS A RANDOM VARIABLE WHOSE DISTRIBUTION  $P(\hat{\mu})$  IS KNOWN IF  $\mu_T$  IS KNOWN.



NOW WE CAN CALCULATE EXPECTATIONS  
LIKE  $E(\hat{\mu}) = \int \hat{\mu} P(\hat{\mu}) d\hat{\mu}$  depends on  $\hat{E}$

# PROPERTIES OF ESTIMATOR $\hat{E}$ (2)

1.  $\hat{E}$  IS CONSISTENT IF

$$\lim_{N \rightarrow \infty} \hat{\mu}_N \rightarrow \mu_T$$

$N$  is the number of events

2. THE BIAS OF  $\hat{E}$  IS

$$B(\hat{\mu}) = E_N(\hat{\mu}) - \mu_T$$

3. THE VARIANCE OF  $\hat{E}$  IS

$$V(\hat{\mu}) = E_N [\hat{\mu} - E_N(\hat{\mu})]^2$$

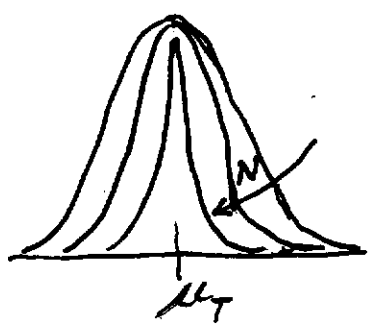
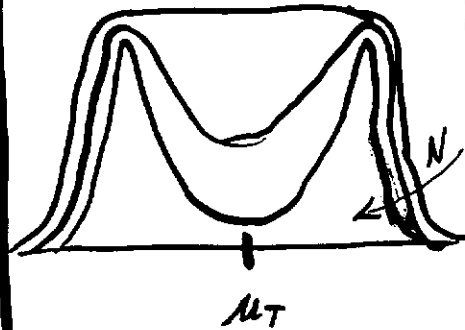
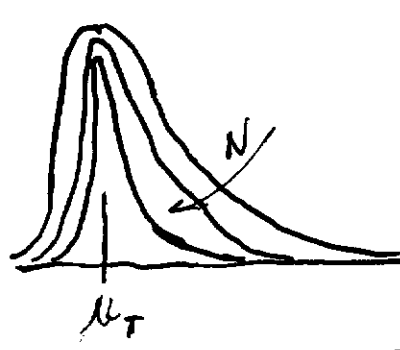
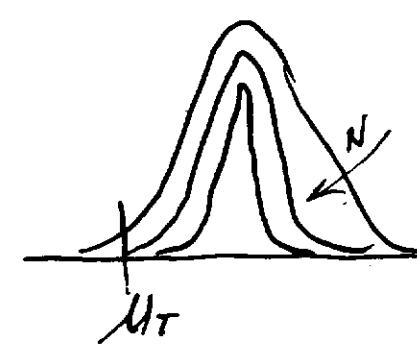
AND THE EFFICIENCY OF  $\hat{E}$  IS

$$Eff(\hat{\mu}) = \frac{V_{min}}{V(\hat{\mu})}$$

WHERE  $V_{min}$  IS THE SMALLEST POSSIBLE VARIANCE OF ANY ESTIMATOR.

# CONSISTENCY & BIAS

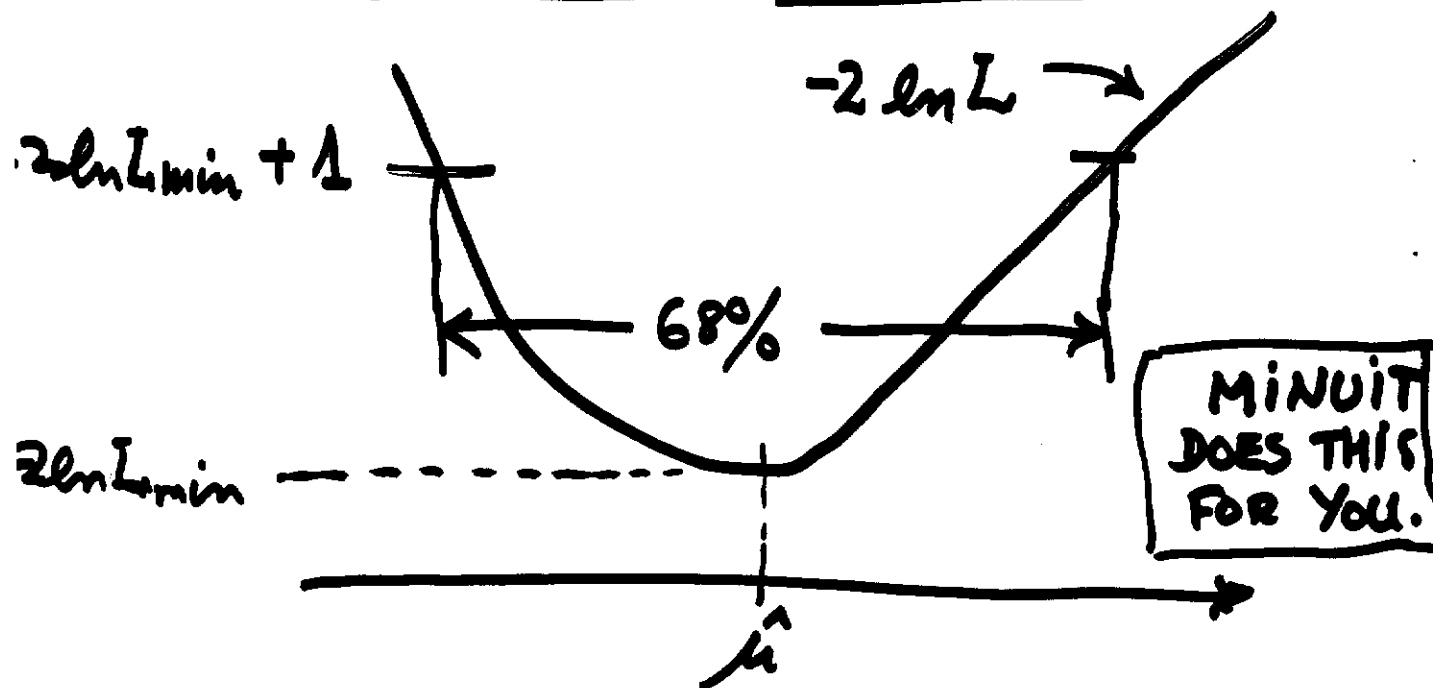
$P(\hat{\mu}_T | \mu_T)$

	CONSISTENT	INCONSISTENT
UNBIASED		
BIASED		

AS A GENERAL RULE, AS  $N$  INCREASES:

- STATISTICAL ERROR DECREASES  $\propto \frac{1}{\sqrt{N}}$
- STATISTICAL BIAS DECREASES  $\propto \frac{1}{N}$
- SYSTEMATIC BIAS (ERROR) IS CONSTANT

# THE MAXIMUM LIKELIHOOD METHOD



METHOD IS ASYMPTOTICALLY OPTIMAL  
UNDER RATHER GENERAL CONDITIONS:

1.  $\hat{\mu}$  is invariant:  $f(\hat{\mu}) = f(\mu)$
2. CONSISTENT
3. EFFICIENT

HOWEVER - SMALL-SAMPLE  
PROPERTIES  
MAY BE POOR!

SOLUTION: EXACT CONF. INTERVALS.  
[J. NEYMAN]

# CONFIDENCE INTERVALS à la FISHER

FROM THE MAX. LIK. METHOD,  
THE CONFIDENCE LEVEL (68%, 90%)  
IS OBTAINED BY TRANSFORMING  
OR REDUCING THE PROBLEM TO  
THE GAUSSIAN CASE

$$P(\hat{\mu}) \propto e^{-\frac{(\mu_T - \hat{\mu})^2}{2\sigma^2}}$$

WHERE THE SYMMETRY  $\mu_T \leftrightarrow \hat{\mu}$

MAKES OBJECTIVE INFERENCE POSSIBLE

FISHER KNEW THIS WAS ONLY  
APPROXIMATE, BUT HIS ATTEMPT TO  
CONSTRUCT AN EXACT THEORY  
("FIDUCIAL INTERVALS") WAS NOT  
SUCCESSFUL.

WE NEED A THEORY WHICH IS

A. CONSTRUCTIVE

B. EXACT

C. USES ONLY "FORWARD"  
PROBABILITIES.

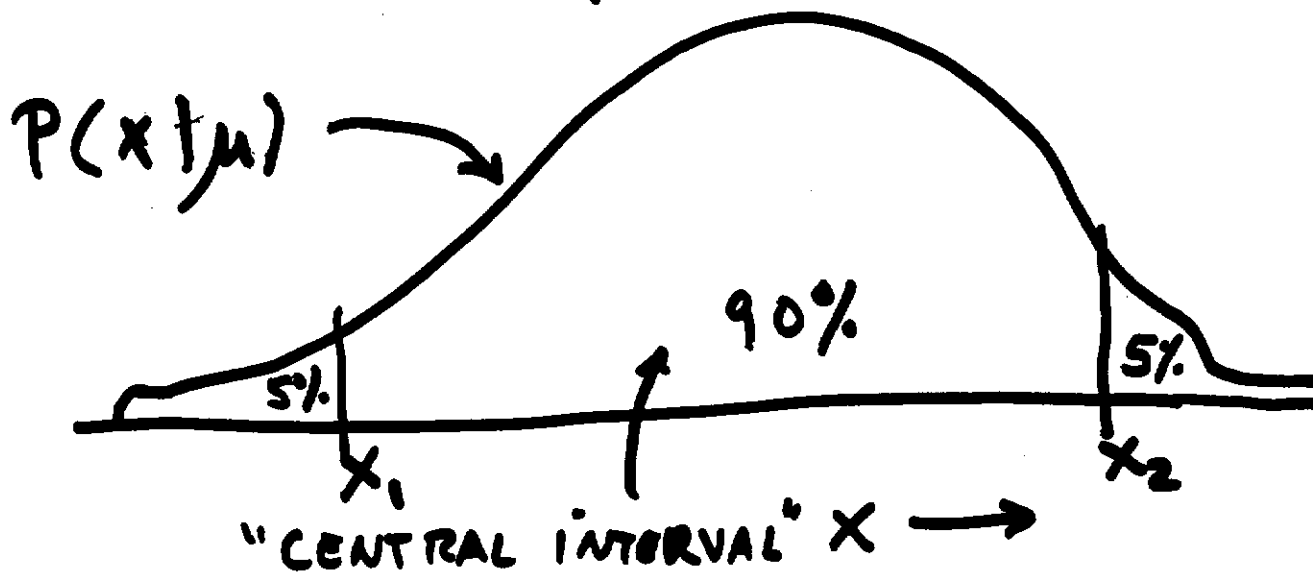


## NEYMAN'S «GROSSE FUGE»

data  $X$  is random (but known)  
true value  $\mu_T$  is fixed (but unknown)

$P(X|\mu)dx$  is known for any  $\mu$   
 $\therefore$  FOR ANY VALUE OF  $\mu$ , CAN FIND  $(X_1, X_2)$

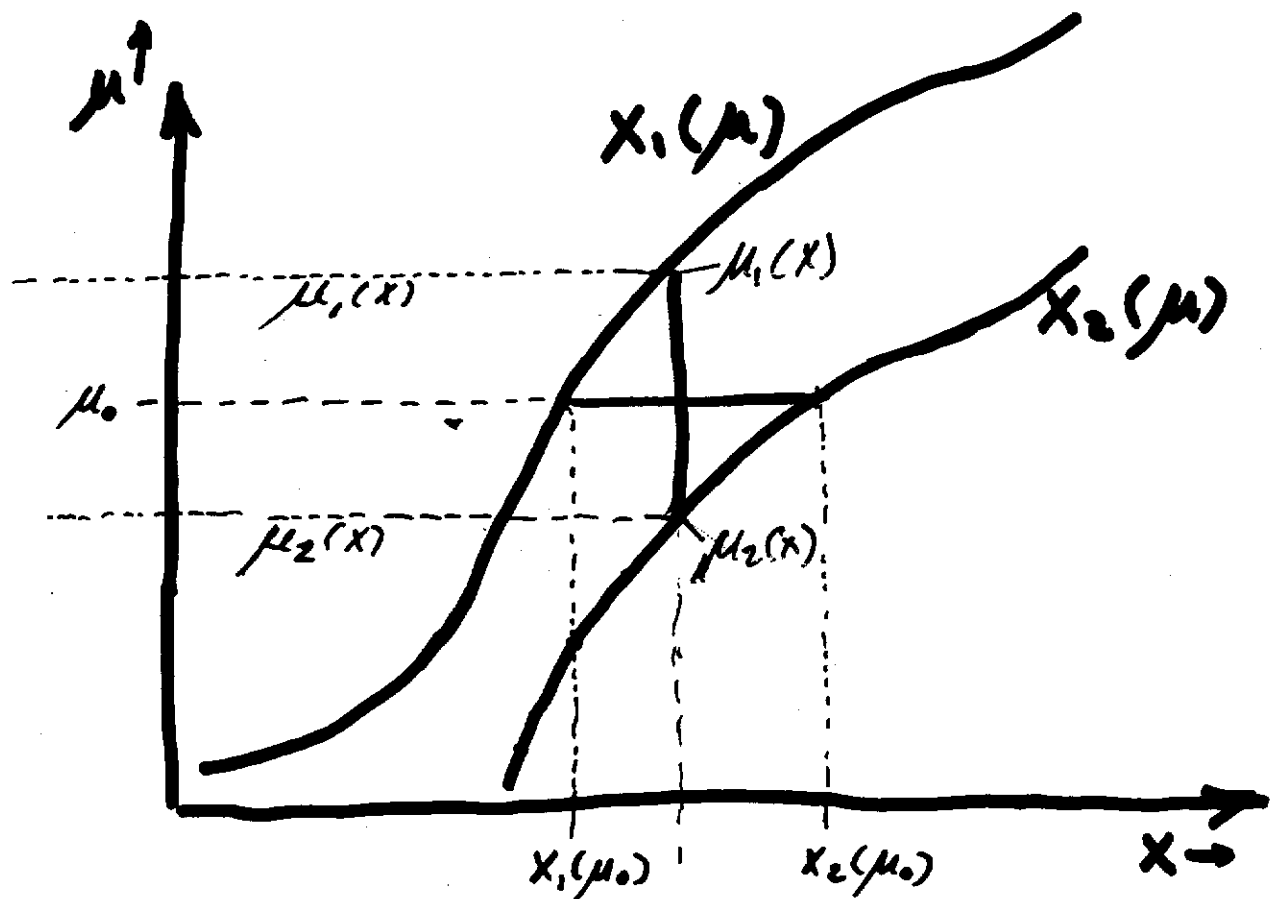
SUCH THAT  $\int_{X_1}^{X_2} P(X|\mu)dx = 90\%$



SO WE CAN ESTABLISH TWO CURVES  
 $X_1(\mu)$  ,  $X_2(\mu)$

DEFINING A ZONE OF 90% PROBABILITY  
OF OBSERVING DATA  $X$ .

# NEYMAN CONSTRUCTION OF CONFIDENCE INTERVALS



CONSIDER SOME VALUE OF  $\mu = \mu_0$

FOR ALL VALUES OF  $x$ :  $x_1(\mu_0) < x < x_2(\mu_0)$

IT HAPPENS THAT:  $\mu_2(x) < \mu_0 < \mu_1(x)$

SINCE THIS IS TRUE FOR ANY  $\mu_0$ , IT HOLDS FOR  $\mu_{TRUE}$ .

$$P[x_1(\mu_T) < x < x_2(\mu_T)] = P[\mu_2(x) < \mu_T < \mu_1(x)]$$

= 90% BY CONSTRUCTION

$$\therefore P(\mu_2 < \mu_T < \mu_1) = 90\%$$

COVERAGE

26

# MAJOR SUCCESS OF THE NEYMAN CONSTRUCTION

PRODUCES A FREQUENTIST  $P$

$$P(\mu_- < \mu_T < \mu_+) \geq 0.9$$

- BY A CONSTRUCTIVE ALGORITHM
- WITHOUT ANY ASSUMPTIONS ABOUT PRIOR DISTRIBUTIONS OF  $\mu$ .
- NO MATTER WHAT  $\mu_T$

NOTE: THE COMPLETE SPECIFICATION  
AND JUSTIFICATION OF THE NEYMAN  
CONSTRUCTION IS NOT WIDELY TAUGHT.  
IT IS OFTEN MISUNDERSTOOD.

above and thinner below the point  $\alpha(\theta'_1, E')$  of its intersection with the hyperplane  $G(\theta'_1)$ . The confidence interval  $\delta(E'')$  corresponding to another sample point,  $E''$ , is not cut by  $G(\theta'_1)$  and is situated entirely above this hyperplane.

Now denote by  $A(\theta'_1)$  the set of all points  $\alpha(\theta'_1, E)$  in  $G(\theta'_1)$  in which this hyperplane cuts one or the other of the confidence intervals  $\delta(E)$ , corresponding to any sample point. It is easily seen that the coordinate  $\theta_1$  of any point belonging to  $A(\theta'_1)$  is equal to  $\theta'_1$  and that the remaining coordinates  $x_1, x_2, \dots, x_n$  satisfy the inequalities

$$\theta(E) \leq \theta'_1 \leq \bar{\theta}(E). \quad (24)$$

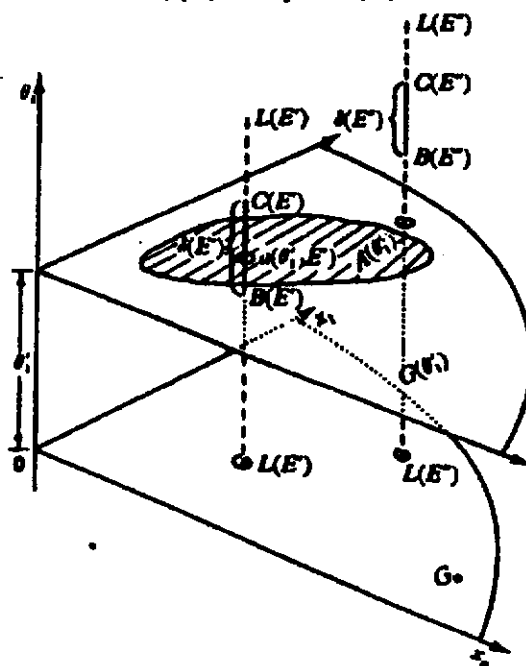


Fig. 1. The general space  $G$ .

In many particular problems it is found that the set of points  $A(\theta_1)$  thus defined is filling up a region. Because of this  $A(\theta_1)$  will be called the region of acceptance corresponding to the fixed value of  $\theta_1 = \theta'_1$ .

It may not seem obvious that the region of acceptance  $A(\theta_1)$  as defined above must exist (contain points) for any value of  $\theta_1$ . In fact, it may seem possible that for certain values of  $\theta_1$  the hyperplane  $G(\theta_1)$  may not cut any of the intervals  $\delta(E)$ . It will, however, be seen below that this is impossible.

As mentioned above, the coordinates  $x_1, x_2, \dots, x_n$  of any sample point  $E$  determine in the space  $G$  the straight line  $L(E)$  parallel to the axis of  $\theta_1$ . If this line crosses the hyperplane  $G(\theta_1)$  in a point belonging to  $A(\theta_1)$  it will be convenient to say that  $E$  falls within  $A(\theta_1)$ .

If for a given sample point  $E$  the lower and the upper estimates satisfy the inequalities  $\theta(E) \leq \theta'_1 \leq \bar{\theta}(E)$ , where  $\theta'_1$  is any value of  $\theta_1$ , then it will be convenient to describe the situation by saying that the confidence interval  $\delta(E)$  covers  $\theta'_1$ . This will be denoted by  $\delta(E)C\theta'_1$ .

The conception and properties of the regions of acceptance are exceedingly important from the point of view of the theory given below. We shall therefore discuss them in detail proving separately a few propositions, however simple they may seem to be.

# MAJOR PROBLEMS

## IN THE FREQUENTIST METHOD

### 1. DISCRETENESS

ALSO BAYES

FOR POISSON, BINOMIAL, ETC.

$$\int_{t_1}^{t_2} f(t/\theta) dt \rightarrow \sum_{n_1}^{n_2} P(n_i/\theta)$$

### 2. ARBITRARINESS OF INTERVAL

ALSO BAYES

MANY  $(t_1, t_2)$  SATISFY  $\int_{t_1}^{t_2} f(t/\theta) dt = \beta$

### 3. PHYSICAL LIMITS ON $\theta$

BAYES SOLVES THIS

### 4. COVERAGE GLOBALLY

BAYES ?

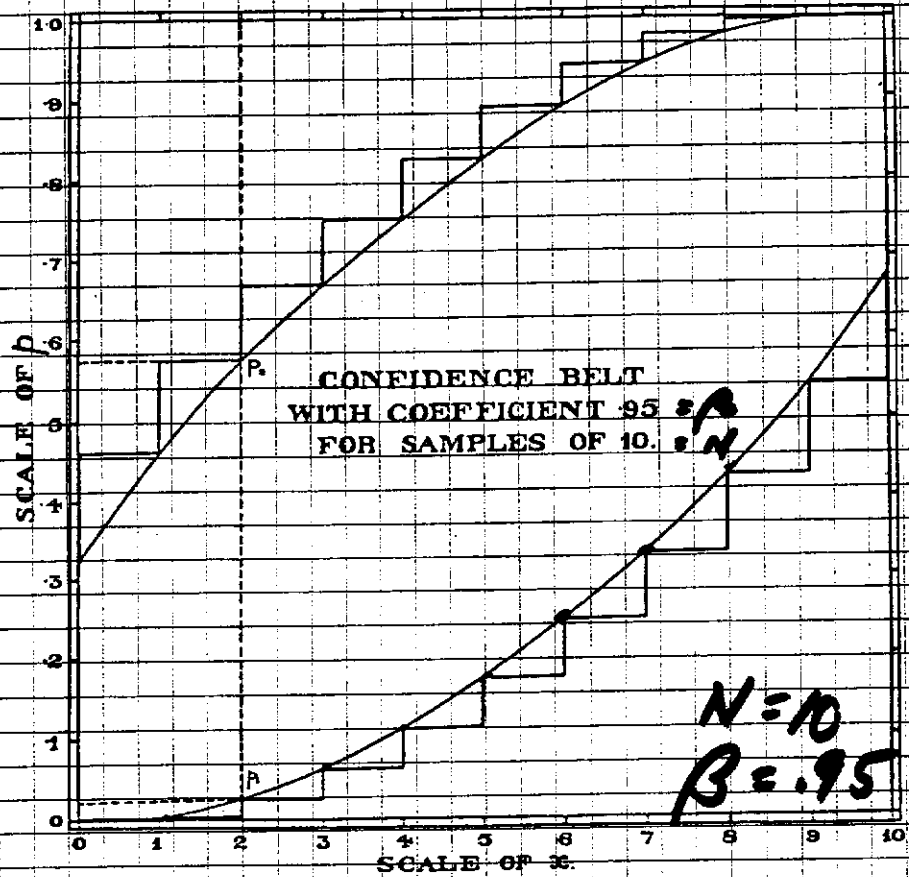
MUST NOT CHOOSE { UPPER LIMIT  
OR  
2-SIDED INTERVAL }

AFTER SEEING DATA.

### 5. NUISANCE PARAMETERS

BAYES ?

EXAMPLE: CLOPPER + PEARSON  
 BIOMETRIKA 26, 404 (1934)  
 CONFIDENCE INTERVALS FOR  
 THE BINOMIAL PARAMETER  $p$



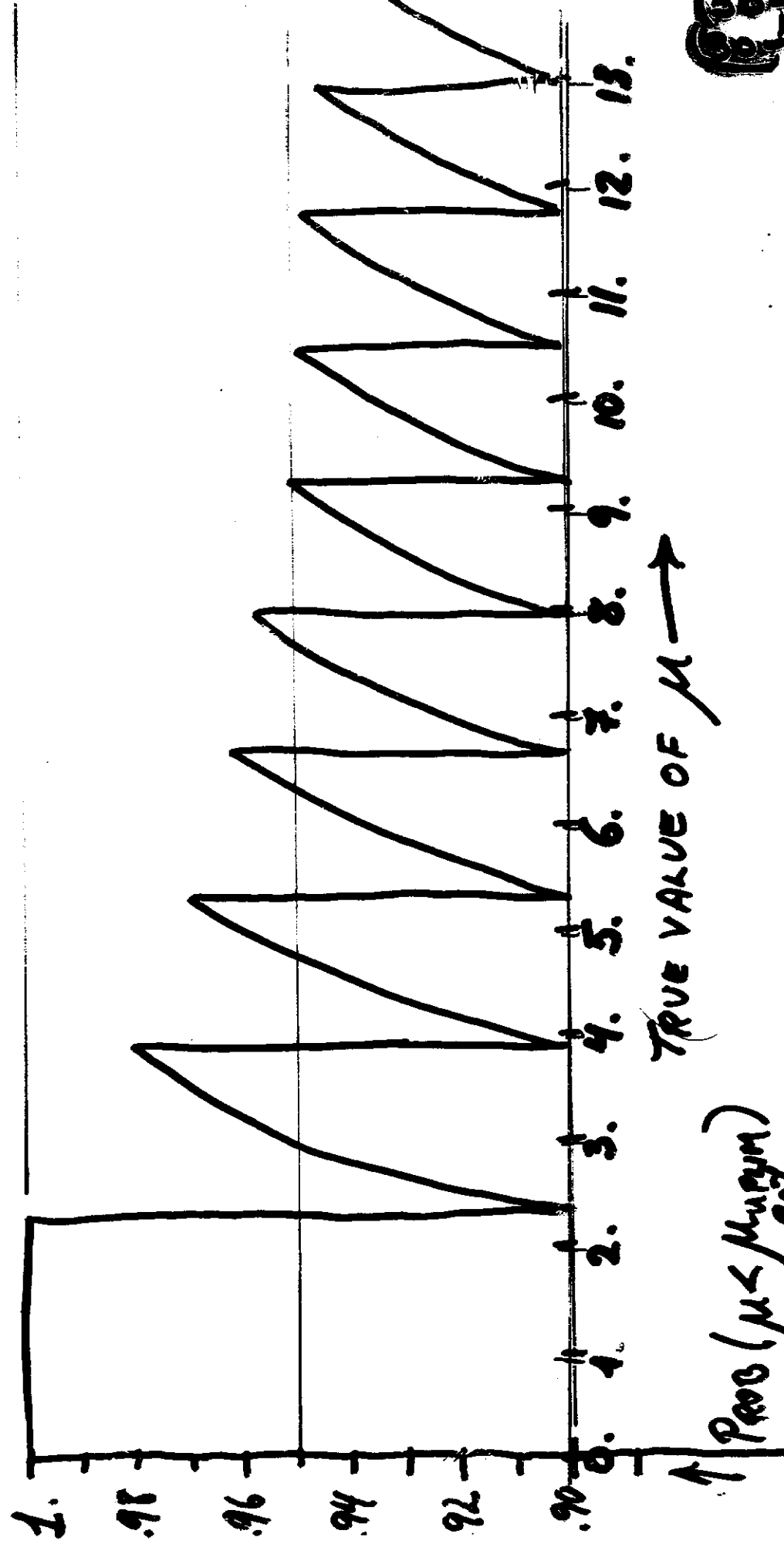
BINOMIAL PROBABILITY:

$$P(x) = \binom{N}{x} p^x (1-p)^{N-x}$$

IF WE OBSERVE  $2/10$

$$95\% = (.03 < p < .57)$$

(OVER-) COVERAGE OF FREQUENTIST 90%  
UPPER LIMITS FOR SMALL POISSON SIGNALS



# BACKGROUND IN POISSON PROCESS

## 1. BACKGROUND EXPECTATION IS KNOWN

1 A. OBSERVE 10 EVTS.  
EXPECT 3 BKGD.

1 B. OBSERVE 0 EVTS.  
EXPECT 3 BKGD.

## 2. EXPECTATION MEASURED WITH SOME UNCERTAINTY

$$\mu_B = 3.1 \pm 1.2$$

THIS IS CALLED A "NUISANCE PARAMETER"



TABLE OF 90% UPPER LIMITS  
FOR POISSON DISTRIBUTION  
WITH BACKGROUND  
PURE FREQUENTIST METHOD  $\alpha = 0.1$

		$N_T = 0$	1	2	3	4
$M_B =$	0	2.30	3.89	5.32	6.68	7.99
	0.5	1.80	3.39	4.82	6.18	7.49
	1.0	1.30	2.89	4.32	5.58	6.99
	2.0	0.30	1.89	3.32	4.68	5.99
	3.0	-0.70	0.89	2.32	3.68	4.99

↑  
 THINGS AGAIN LOOKING BAD

# TABLE OF 90% UPPER LIMITS

(MUST BE NORMALIZED BY "LUMINOSITY")

$\alpha = 0.1$

## BAYESIAN WITH UNIFORM PRIOR

$N_T =$	0	1	2	3	4
$\mu_B = 0$	2.30	3.89	5.32	6.68	7.99
0.5	2.30	3.50	4.83	6.17	7.49
1.0	2.30	3.26	4.44	5.71	6.99
2.0	2.30	3.00	3.87	4.92	6.08
3.0	2.30	2.83	3.52	4.37	5.35

## BAYESIAN WITH $1/\mu$ PRIOR

$\alpha = 0.1$

$N_T =$	0	1	2	3	4
$\mu_B = 0$	0.00	2.30	3.89	5.32	6.68
0.5	0.00	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00	0.00
2.0	0.00	0.00	0.00	0.00	0.00

# THE "CORRECTED FREQUENTIST" METHOD

(3)

- (1) TAKE ACCOUNT OF ALL WAYS IN WHICH  $N_T$  CAN BE DIVIDED INTO  $N_S + N_B$  ( $N_T + 1$  combinations)
- (2) CALCULATE PROB (each combination)
- (3) WORK ONLY WITH  $N_S$  and  $\text{Prob}(N_S)$

[STEP (2) REQUIRES KNOWING  $\mu_S$  (AS WELL AS  $\mu_B$ )

## METHOD:

(1) (NO BACKGROUND)

DEGREE OF CONFIDENCE  $\rightarrow D(\hat{\mu}_S | N_S) = \sum_{n > N_S} e^{-\hat{\mu}_S} \frac{\hat{\mu}_S^n}{n!} = \Lambda(N_S | \hat{\mu}_S)$

(ORDINARY "FREQUENTIST" METHOD)

(2) WITH BG:

$\rightarrow D(\hat{\mu}_S | N_T) = \sum_{N_S} P(N_S | \hat{\mu}_B, \mu_B) \Lambda(N_S, \hat{\mu}_S)$

WEIGHTED AVERAGE OVER POSSIBLE VALUES OF  $N_S$

# "CORRECTED" FREQUENTIST (continued) (35)

ABOVE FORMULA GIVES

- DEGREE OF CONFIDENCE ( $\alpha$ )  
AS FUNCTION OF "UPPER LIMIT"  $\hat{\mu}_s$   
INVERTING (NUMERICALLY), GET  $\hat{\mu}(\alpha)$

## CORRECTED FREQUENTIST UPPER LIMITS

$N_T =$	0	1	2	3	4
$\mu_B = 0.$	2.30	3.89	5.32	6.68	7.99
0.5	2.30	3.76	5.13	6.46	7.75
1.0	2.30	3.65	4.96	6.24	7.49
2.0	2.30	3.46	4.66	5.86	7.05
3.0	2.30	3.30	4.38	5.49	6.63

$\alpha = 0.1$   
(90%)

THIS IS CALLED "CONDITIONING"

# Unified approach to the classical statistical analysis of small signals

Gary J. Feldman<sup>\*</sup>

*Department of Physics, Harvard University, Cambridge, Massachusetts 02138*

Robert D. Cousins<sup>†</sup>

*Department of Physics and Astronomy, University of California, Los Angeles, California 90095*

(Received 21 November 1997; published 6 March 1998)

We give a classical confidence belt construction which unifies the treatment of upper confidence limits for null results and two-sided confidence intervals for non-null results. The unified treatment solves a problem (apparently not previously recognized) that the choice of upper limit or two-sided intervals leads to intervals which are not confidence intervals if the choice is based on the data. We apply the construction to two related problems which have recently been a battleground between classical and Bayesian statistics: Poisson processes with background and Gaussian errors with a bounded physical region. In contrast with the usual classical construction for upper limits, our construction avoids unphysical confidence intervals. In contrast with some popular Bayesian intervals, our intervals eliminate conservatism (frequentist coverage greater than the stated confidence) in the Gaussian case and reduce it to a level dictated by discreteness in the Poisson case. We generalize the method in order to apply it to analysis of experiments searching for neutrino oscillations. We show that this technique both gives correct coverage and is powerful, while other classical techniques that have been used by neutrino oscillation search experiments fail one or both of these criteria.

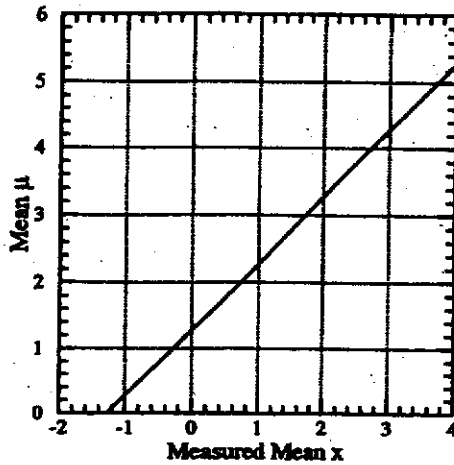
[S0556-2821(98)00109-X]

PACS number(s): 06.20.Dk, 14.60.Pq

## I. INTRODUCTION

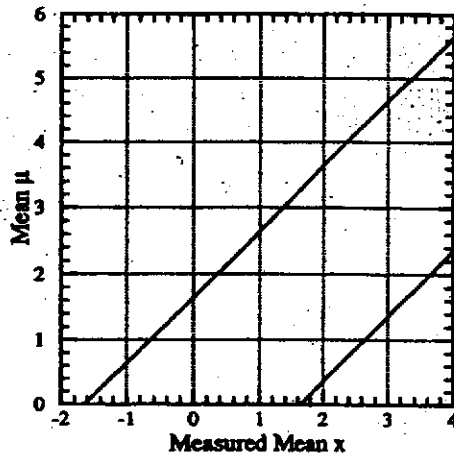
decide whether to consult confidence interval tables for upper limits or for central confidence intervals.

# NEYMAN CONFIDENCE INTERVALS FOR GAUSSIAN MEAN. NEAR A PHYSICAL BOUNDARY



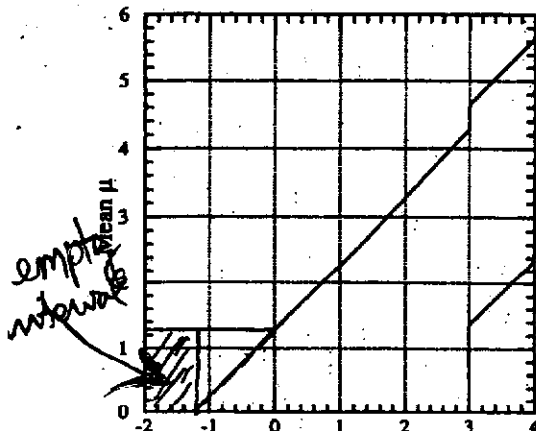
90% UPPER LIMITS

FIG. 2. Standard confidence belt for 90% C.L. upper limits for the mean of a Gaussian, in units of the rms deviation. The second line in the belt is at  $x = +\infty$ .



90% 2-SIDED  
INTERVALS

FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

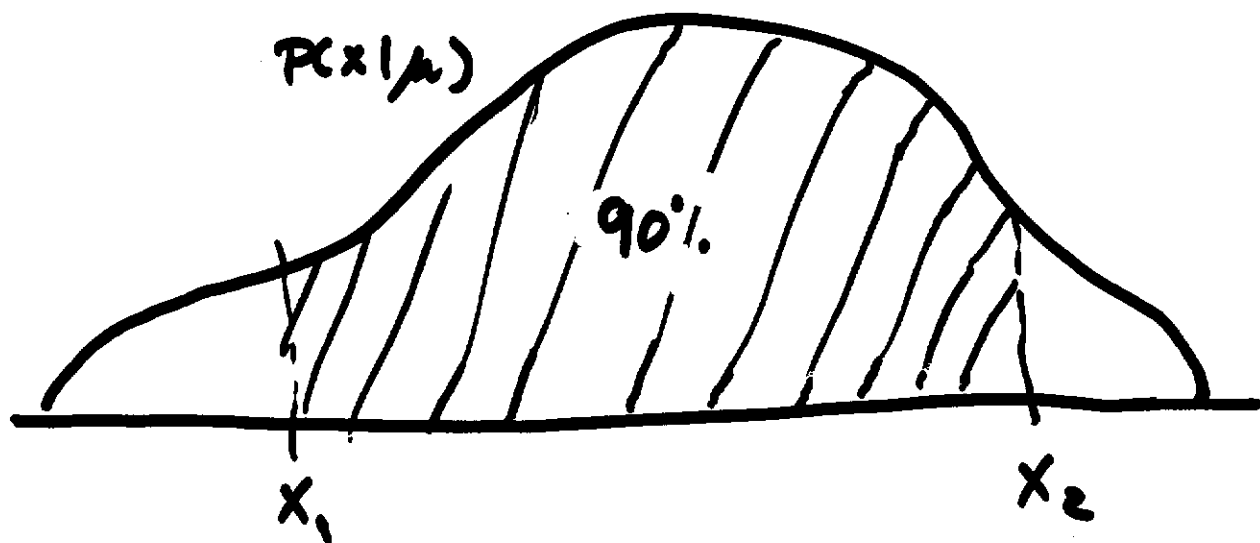


FLIP-FLOPPING

QUOTE U.L. IF  $x < 3\sigma$

2-SIDED IF  $x > 3\sigma$

# THE "ORDERING PRINCIPLE"



THERE IS FREEDOM TO CHOOSE  $(x_1, x_2)$  SUCH THAT  $\int_{x_1}^{x_2} P(x|\mu) dx = .9$

THE MOST GENERAL WAY TO RESOLVE THIS AMBIGUITY IS TO SPECIFY:

THE ORDER IN WHICH INFINITESIMAL ELEMENTS WILL BE INCLUDED (OR EXCLUDED) IN THE INTERVAL.

ORDERING PRINCIPLE EXAMPLE  
MEAN BACKGROUND  $b = 3.0$

SIGNAL MEAN  $\mu = 0.5$

FIND  $[N_1, N_2]$  SUCH THAT  $P(N \in [N_1, N_2]) = 0.9$

TABLES

TABLE I. Illustrative calculations in the confidence belt construction for signal mean  $\mu$  in the presence of known mean background  $b = 3.0$ . Here we find the acceptance interval for  $\mu = 0.5$ .

$n$	$P(n \mu)$	$\mu_{best}$	$P(n \mu_{best})$	$R$	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5		
2	0.185	0.	0.224	0.826	3	✓	✓
3	0.216	0.	0.224	0.963	2	✓	✓
4	0.189	1.	0.195	0.966	1	✓	✓
5	0.132	2.	0.175	0.753	4	✓	✓
6	0.077	3.	0.161	0.480	7	✓	✓
7	0.039	4.	0.149	0.259		✓	✓
8	0.017	5.	0.140	0.121		✓	
9	0.007	6.	0.132	0.050		✓	
10	0.002	7.	0.125	0.018		✓	
11	0.001	8.	0.119	0.006		✓	

FOR EACH  $N$ , FIND  $\mu_{best} = \text{BEST FIT } \mu \geq 0$

$$\mu_{best} = \text{MAX}(0, N-b)$$

LET  $R = \frac{P(N|\mu)}{P(N|\mu_{best})}$  ; ORDER USING  $R$ .



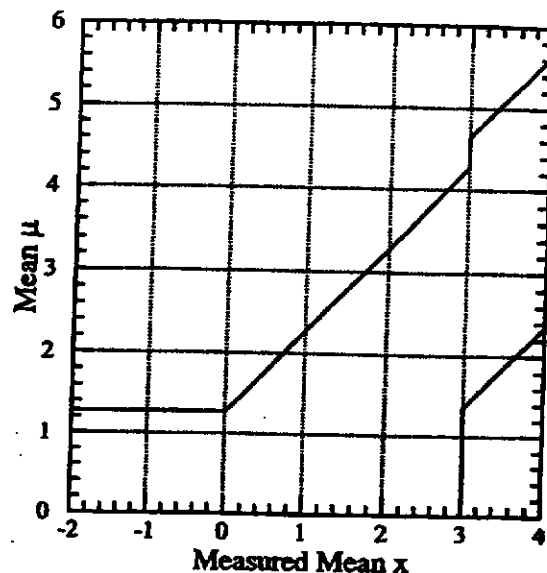


FIG. 4. Plot of confidence belts implicitly used for 90% C.I. confidence intervals (vertical intervals between the belts) quoted by flip-flopping physicist X, described in the text. They are not valid confidence belts, since they can cover the true value at a frequency less than the stated confidence level. For  $1.36 < \mu < 4.28$ , the coverage (probability contained in the horizontal acceptance interval) is 85%.

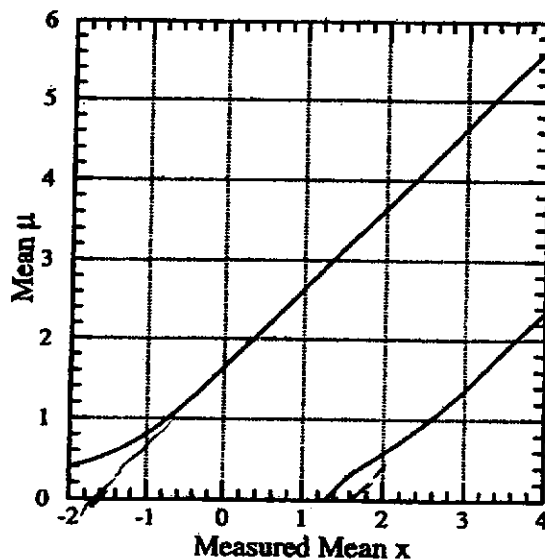


FIG. 10. Plot of our 90% confidence intervals for the mean of a Gaussian, constrained to be non-negative, described in the text.

# BAYESIAN UNIFORM

$N_T = 0$	1	2	3	
$\mu_B = 0.$	2.30	3.89	5.32	6.68
1.0	2.30	3.26	4.44	5.71
2.0	2.30	3.00	3.87	4.92
3.0	2.30	2.83	3.52	4.37
4.0	2.30	2.73	3.30	3.95

$1/\mu$

$N_T = 0$	1	2	3	
$\mu_B = 0.$	0.	2.30	3.89	5.32
1.0				
2.0				
3.0				
4.0				

0.00

# FREQUENTIST PURE

$N_T = 0$	1	2	3	
$\mu_B = 0.$	2.30	3.89	5.32	6.68
1.0	1.30	2.89	4.32	5.68
2.0	0.30	1.89	3.32	4.68
3.0	neg.	0.89	2.32	3.68
4.0	neg.	neg.	1.32	2.68

"CORRECTED"

$N_T = 0$	1	2	3	
$\mu_B = 0.$	2.30	3.89	5.32	6.68
1.0	2.30	3.65	4.96	6.24
2.0	2.30	3.46	4.66	5.86
3.0	2.30	3.30	4.38	5.49
4.0	2.30	3.18	4.16	5.19

"FELDMAN/KOUSINS"

$N_T = 0$	1	2	3	
$\mu_B = 0.$	2.44	$\frac{4.36}{0.11}$	$\frac{5.91}{0.53}$	$\frac{7.42}{1.10}$
1.0	1.61	3.36	4.91	$\frac{6.42}{0.10}$
2.0	1.26	2.53	3.91	5.42
3.0	1.08	1.88	3.04	4.42
4.0	1.01	1.39	2.33	3.53

$$\mu_B = 1.0$$

OVERSATISFACTION PROB ( $\mu < \hat{\mu}_{90\%}$ )

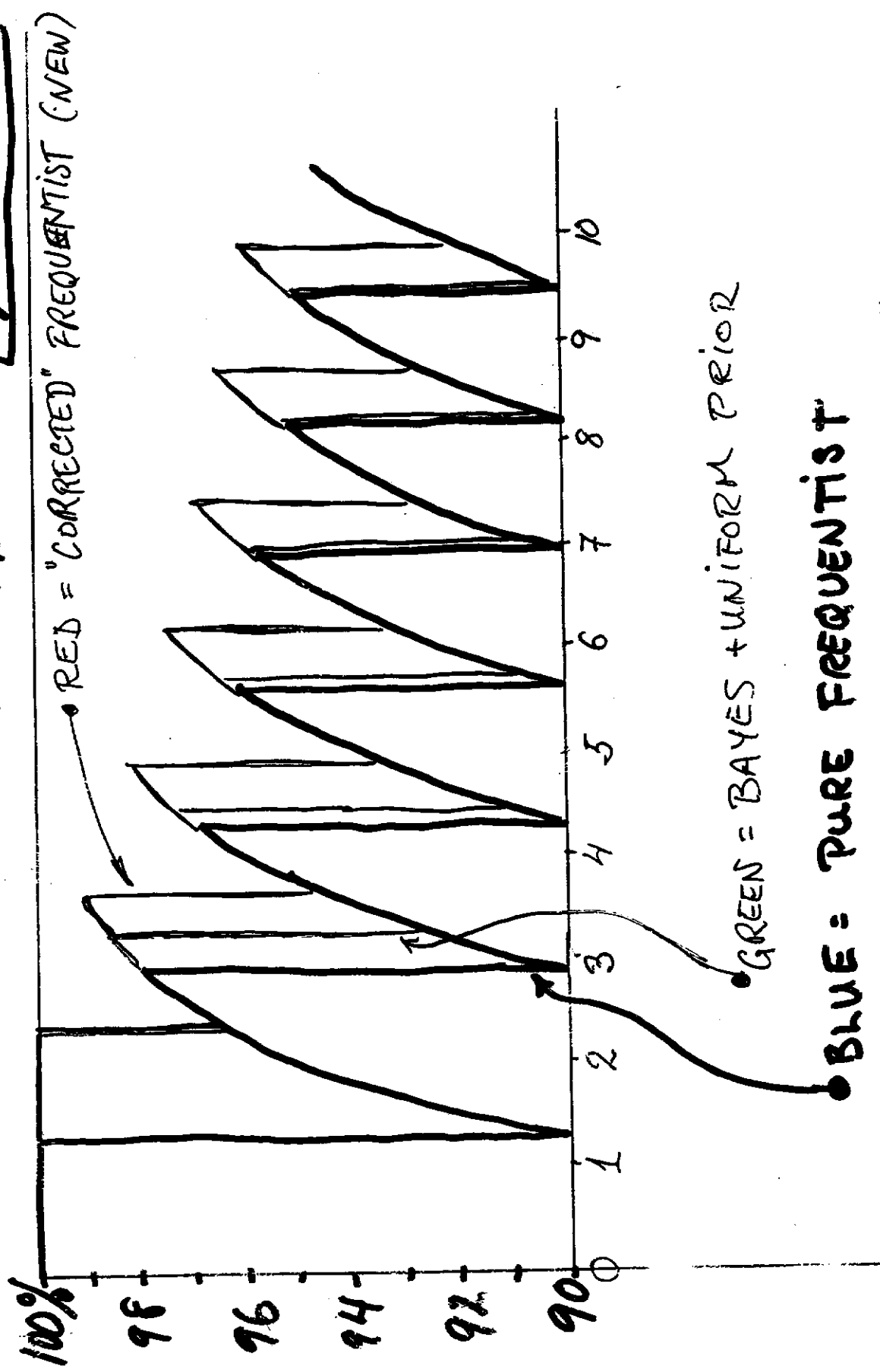


Table 1. 68% C.L. confidence intervals  $(\mu_1, \mu_2)$  for the mean of a Poisson distribution, based on the single observation  $n_0=3$ , calculated by various methods.

Method	Prior	Defining equation(s)	Interval	Length	Coverage?
Root-mean-square deviation	...	$n_0 \pm \sqrt{n_0}$	(1.27, 4.73)	3.46	no
Classical central	...	Eqs. (6) and (7)	(1.37, 5.92)	4.55	yes
Classical shortest	...	Method of Crow and Gardner <sup>a</sup>	(1.29, 5.25)	3.96	yes
Likelihood ratio	...	Eq. (9)	(1.58, 5.08)	3.50	no
Bayesian central	1	Eqs. (16) and (17)	(2.09, 5.92)	3.83	no
Bayesian shortest	1	Eq. (16); minimum $\mu_2 - \mu_1$	(1.55, 5.15)	3.60	no
Bayesian equal $\pm$	1	Eq. (16); $\hat{\mu} - \mu_1 = \mu_2 - \hat{\mu}$	(1.15, 4.85)	3.70	no
Bayesian central	$1/\mu_1$	Eqs. (16) and (17)	(1.37, 4.64)	3.27	no
Bayesian shortest	$1/\mu_1$	Eq. (16); minimum $\mu_2 - \mu_1$	(0.86, 3.85)	2.99	no
Bayesian equal $\pm$	$1/\mu_1$	Eq. (16); $\hat{\mu} - \mu_1 = \mu_2 - \hat{\mu}$	(1.36, 4.63)	3.27	no

<sup>a</sup>Reference 31.

## COVERAGE IN PRACTICE

- IN A REAL EXPERIMENT, WE USE MANY APPROXIMATIONS, AND MAY WANT TO VERIFY WHETHER COVERAGE IS STILL OBTAINED
- CALCULATE COVERAGE BY M.C.
  - GENERATE M.C. DATA WITH  $\mu_T$  KNOWN
  - FIT THE DATA, OBTAINING C.I. FOR  $\mu$ .
  - REPEAT MANY TIMES, NOTING HOW OFTEN  
 $(\mu_- < \mu_T < \mu_+)$
- PRACTICAL PROBLEM:  
 YOU HAVE TO KNOW HOW YOU WOULD HANDLE EVERY M.C. DATA SAMPLE IF IT WERE REAL DATA.

# PROPERTIES OF UPPER LIMITS FOR POISSON $\mu$ WITH BACKGROUND $b$ .

	BAYES		FREQ.		
	$1/\mu$	UNI FORM	NAIVE	F/C	J
1. CONSISTENT	NO	—	—	—	—
2. SCALE-INVARIANT	—	NO	—	—	—
3. GENERALLY INVARIANT	NO	NO	—	—	—
4. COVERS	NO	—	—	—	—
5. OPTIMALLY COVERS	NO	NO	—	—	NO
6. EXACTLY COVERS	NO	NO	NO	NO	NO
7. NON-NULL INTERVALS	NO	—	NO	—	—
8. IND'P'T OF $b$ WHEN $N=0$	NO	—	NO	NO	—
9. GOOD MEASURE OF SENSITIV.					
10. CAN BE COMBINED W OTHERS					

these are the important properties,  
and NO METHOD is good for these!

# HYPOTHESIS (THEORY) - TESTING

## DEFINITIONS:

SIMPLE HYPOTHESIS - COMPLETELY SPECIFIED  
COMPOSITE HYPOTHESIS - MAY HAVE SOME UNKNOWN PARAM'S.

$$H_0 \longleftrightarrow H_1$$

"NULL"

"ALTERNATIVE"

TEST STATISTIC  $X$  } - SOME FUNCTION OF  
IS POINT IN SPACE  $W$  } THE OBSERVATIONS,  
TO BE USED IN TEST.

CRITICAL REGION  $w$  } SET OF  $X$  FOR  
IN SPACE  $W$  } WHICH WE  
REJECT  $H_0$

ACCEPTANCE REGION  $W-w$  } SET OF  $X$  FOR  
WHICH WE  
WILL ACCEPT  $H_0$

$$P(X \in w | H_0) = \alpha$$

THE LEVEL OF SIGNIFICANCE

$$P(X \in w | H_1) = 1 - \beta$$

THE POWER OF THE TEST

$$P(X \in W-w | H_1) = \beta$$

	<u><math>H_0</math> TRUE</u>	<u><math>H_1</math> TRUE</u>
ACCEPT $H_0$	PROB. = $1 - \alpha$ <div>GOOD ACCEPTANCE</div>	PROB. = $\beta$ ERROR OF THE SECOND KIND CONTAMINATION
ACCEPT $H_1$	PROB. = $\alpha$ ERROR OF THE FIRST KIND LOSS	PROB. = $1 - \beta$ <div>GOOD REJECTION</div>

EXAMPLE: TO STUDY ELASTIC pp SCATTERING, SEPARATE EVENTS OF TYPE

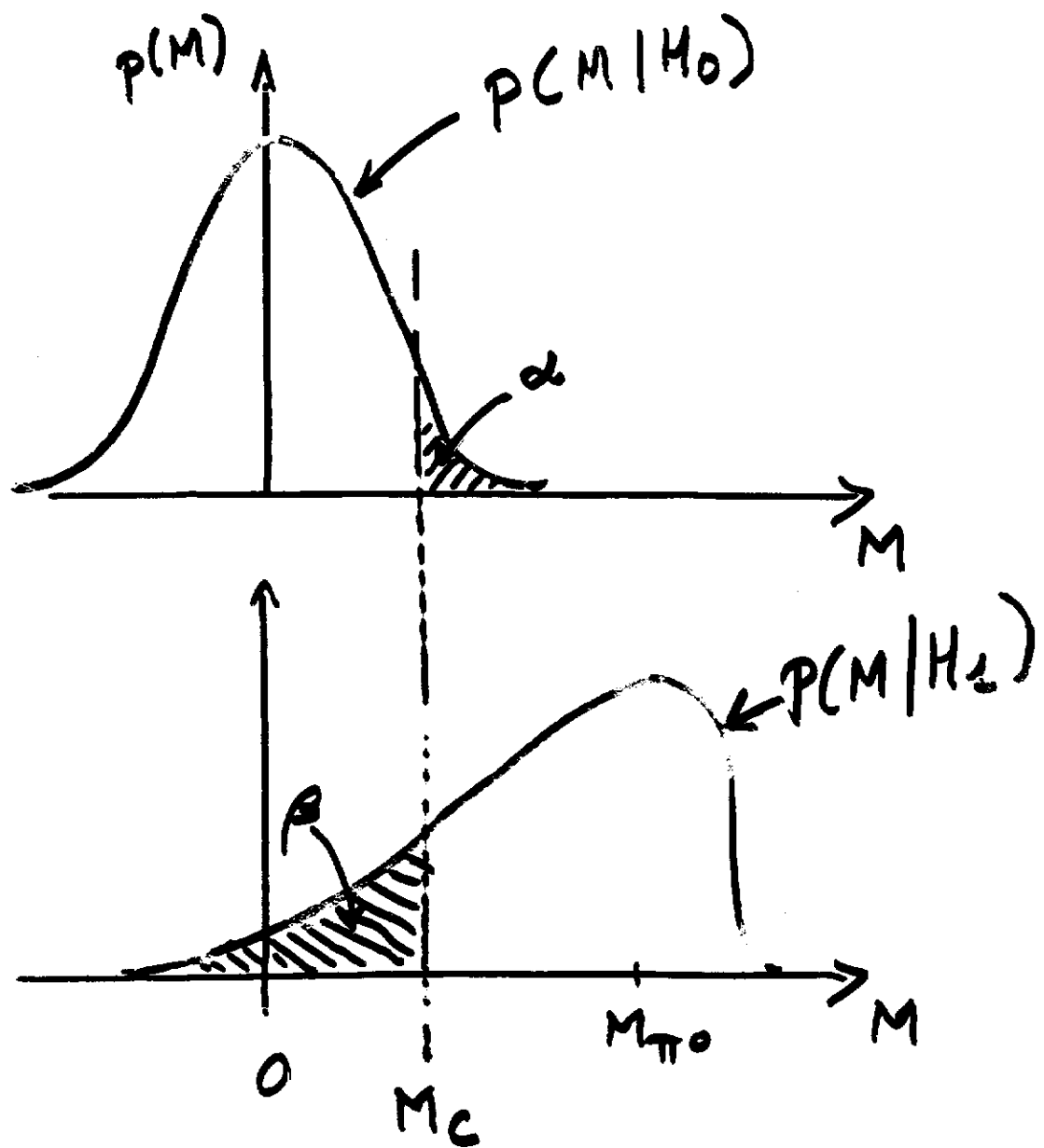
$pp \rightarrow pp$   $H_0$

FROM

$pp \rightarrow pp\pi^0$   $H_1$

TO DO THIS, WE NEED A TEST STATISTIC, SUCH AS THE MISSING MASS.

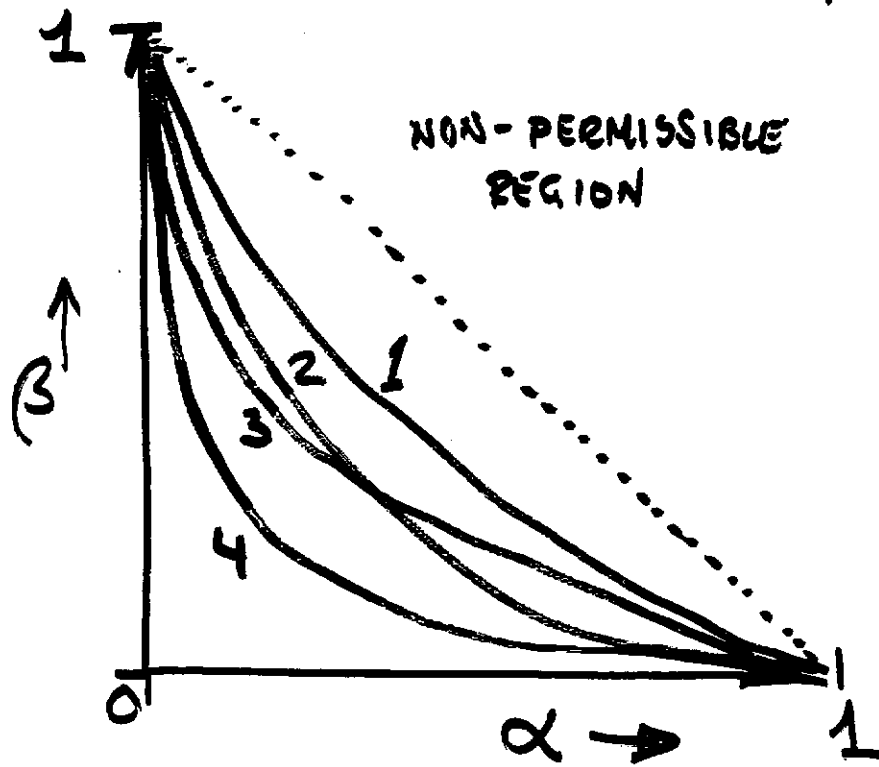




WE CAN CHOOSE ANY  $\alpha$ ,  
 BUT THEN  $\beta$  IS DETERMINED.  
 IN ORDER TO IMPROVE BOTH  
 $\alpha$  AND  $\beta$ , WE CAN LOOK FOR A  
 BETTER TEST STATISTIC -  
 MISSING ENERGY? MOMENTUM?

# TRADING-OFF

$$\alpha \leftrightarrow \beta$$



TEST #4 IS ALWAYS BEST  
 #2 IS SOMETIMES BETTER THAN #3  
 TEST #1 IS ALWAYS WORST

# THE NEYMAN-PEARSON TEST

H  
5

WE LOOK FOR THE

MOST POWERFUL TEST FOR GIVEN  $\alpha$

$\Rightarrow$  BEST CRITICAL REGION IN X-SPACE

critical region  $\rightarrow$   $\int_{W_\alpha} f(\underline{x} | \theta_0) d\underline{x} = \alpha$

$\Rightarrow$  FIND REGION  $W_\alpha$  TO MINIMIZE  $\beta$

$$1 - \beta = \int_{W_\alpha} f(\underline{x} | \theta_1) d\underline{x}$$

$$= \int_{W_\alpha} \left[ \frac{f(\underline{x} | \theta_1)}{f(\underline{x} | \theta_0)} \right] f(\underline{x} | \theta_0) d\underline{x}$$

$$= \text{Expectation of } \left[ \quad \right]_{\theta = \theta_0}$$

H  
6

THE MOST POWERFUL TEST  
FOR 2 SIMPLE HYPOTHESES  
CONSISTS IN CHOOSING THE  
CRITICAL REGION SUCH  
THAT THE LIKELIHOOD RATIO  
IS LARGER AT ALL POINTS IN  $W_K$   
THAN OUTSIDE  $W_K$ .

THE COMPUTATION MAY BE  
DIFFICULT SINCE  $\underline{X}$  IS  
IN GENERAL MULTIDIMENSIONAL

THE SECRET IS TO FIND A  
ONE-DIMENSIONAL  
TEST STATISTIC  $X_t(\underline{X})$

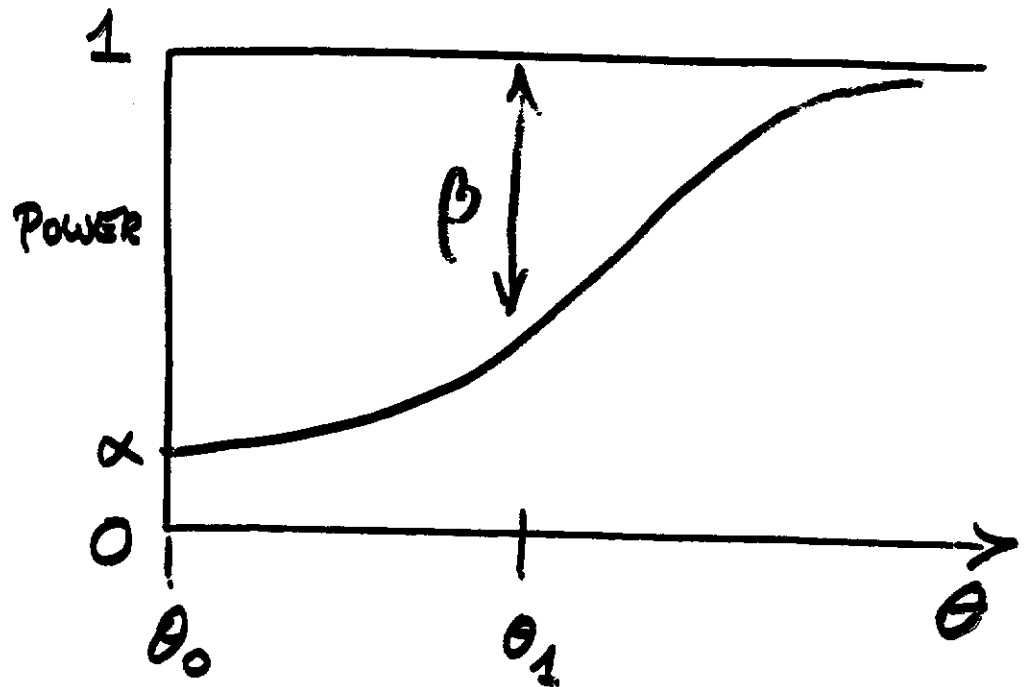
BUT IT WILL NOT IN GENERAL BE  
AS POWERFUL AS THE  
NEYMAN-PEARSON TEST.

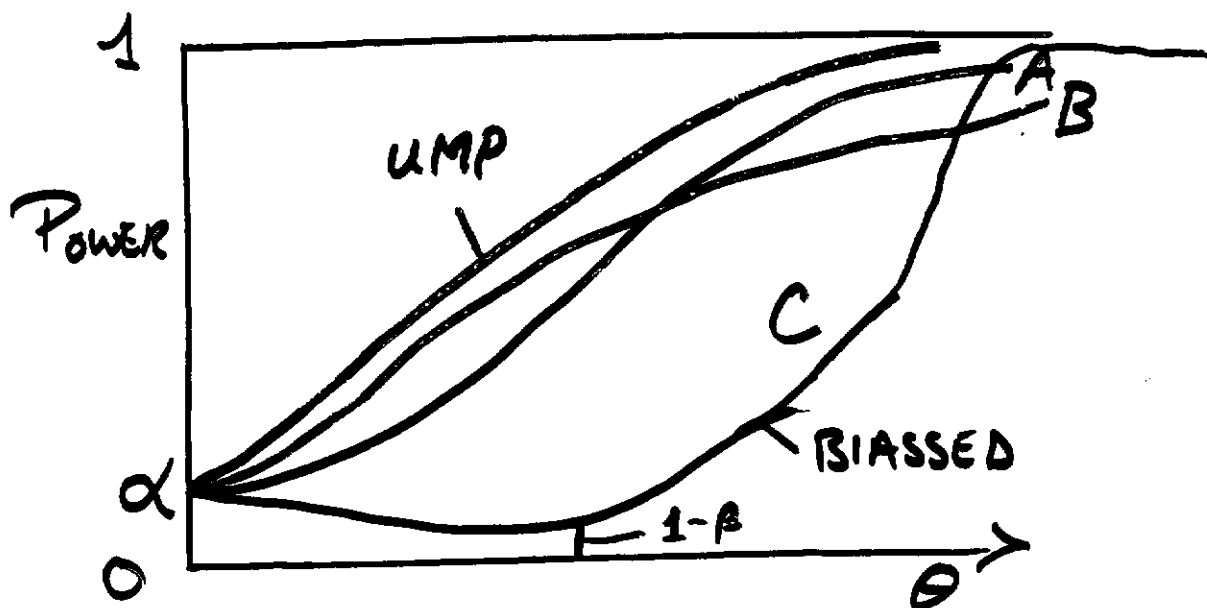
# COMPARISON OF TESTS

CONSIDER  $H_0 : \theta = \theta_0$

$H_1 : \theta = \theta_1$

AND CONSIDER THE POWER  
OF A TEST AS A FUNCTION  
OF  $\theta_0 - \theta_1$





UMP = UNIFORMLY MOST POWERFUL

TEST A IS MORE POWERFUL THAN  
TEST B ONLY FOR LARGE VALUES  
OF  $\theta_1$

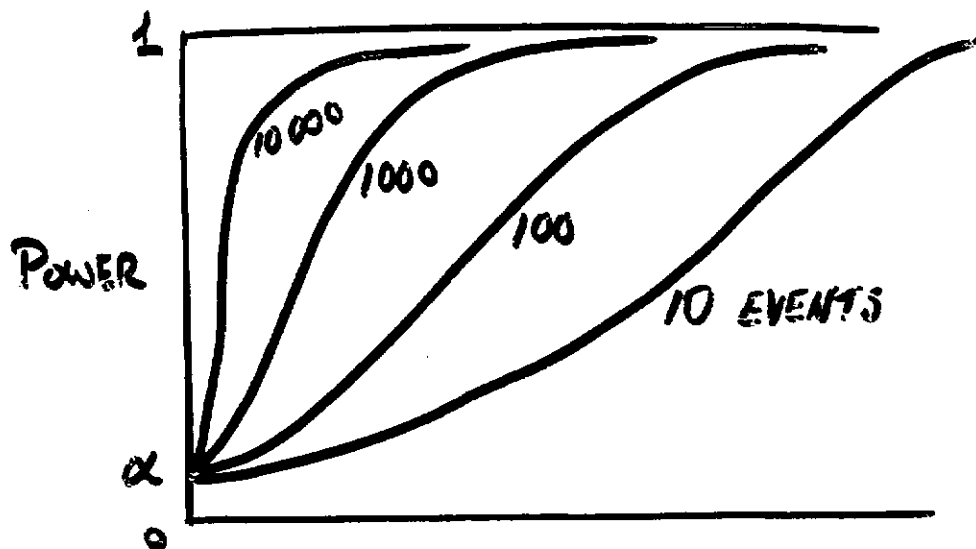
TEST C IS BIASSED BECAUSE

$$1 - \beta < \alpha \quad \text{FOR SOME } \theta_1$$

THAT MEANS WE ARE MORE LIKELY TO  
ACCEPT  $H_0$  WHEN IT IS FALSE  
THAN WHEN IT IS TRUE.

(BIAS USUALLY BAD, BUT COULD BE  
USED FOR SOME VALUES OF  $\theta_1$ )

H  
9  
THE CONSISTENCY OF A TEST  
DEPENDS ON ITS POWER  
AS THE AMOUNT OF DATA INCREASES



ALL CURVES REFER TO THE  
SAME TEST, BUT FOR  
DIFFERENT AMOUNTS OF DATA.

A TEST IS CONSISTENT IF:

$$\lim_{N \rightarrow \infty} P(\underline{X} \in W_\alpha / H_1) = 1$$

# FOR COMPOSITE HYPOTHESES

THE THEORY IS NOT SO CLEAR  
OPTIMAL TECHNIQUES ARE KNOWN  
ONLY IN ASYMPTOTIC LIMIT  
AND MAY BE QUITE BAD FOR SMALL  
SAMPLES OF DATA.

EXAMPLES OF COMPOSITE HYPOTH...

$$H_0 : \theta_1 = a \quad \theta_2 = b$$

$$H_1 : \theta_1 \neq a \quad \theta_2 \neq b$$

$$H_0 : \theta_1 = a \quad \theta_2 \text{ unspecified}$$

$$H_1 : \theta_1 = b \quad \theta_2 \text{ unspecified}$$

THE MOST POWERFUL TOOL

IS THE

MAXIMUM LIKELIHOOD RATIO



## LIKELIHOOD FUNCTION:

$$L(\underline{\theta}) = \prod_{i=1}^N f(\underline{x}_i | \underline{\theta})$$

↑ observations      ↑ parameters

## MAX LIKELIHOOD RATIO:

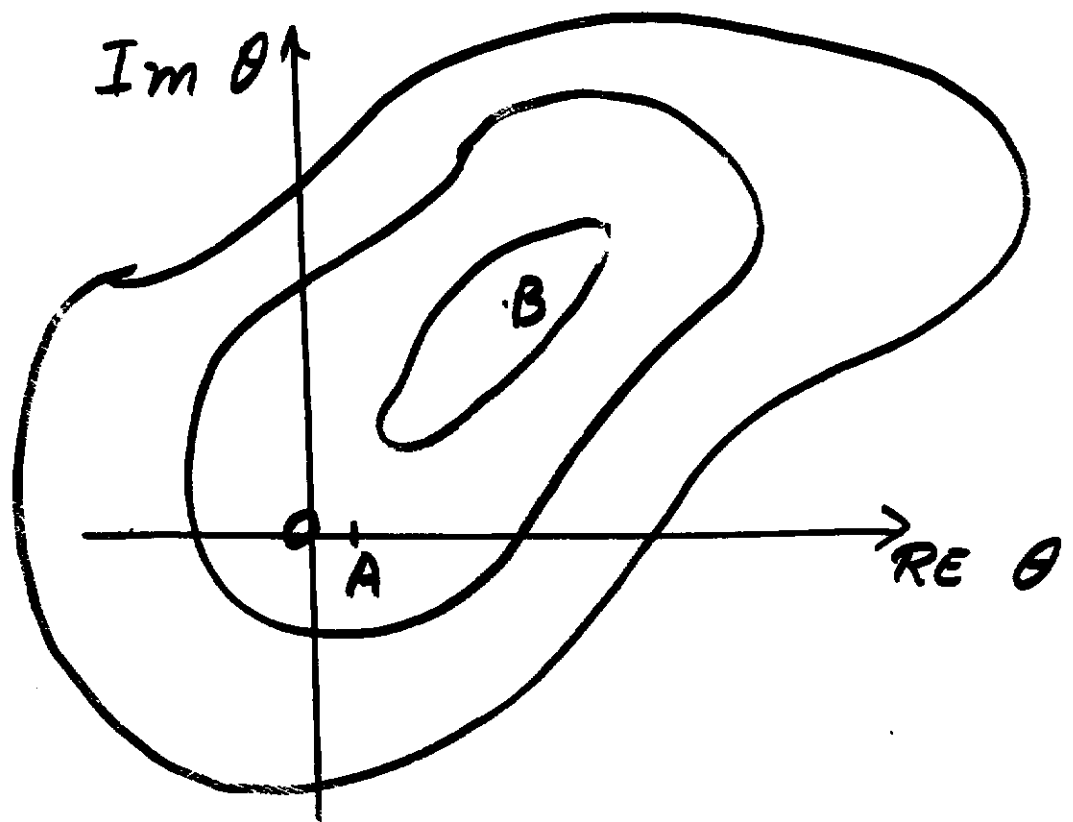
$$\lambda = \frac{\max L(\theta | H_0)}{\max L(\theta | H_1)}$$

WHERE  $H_1$  HAS  $r$  MORE FREE PARAMETERS THAN  $H_0$ ,

$$-2 \ln \lambda \rightarrow \chi^2(r) \Big|_{H_0 \text{ TRUE}}^{\text{FOR LARGE } N}$$

SO THAT ONE CAN LOOK UP THE SIGNIFICANCE OF THE EXTRA PARAMETER(S) IN A TABLE OF CHI-SQUARE.

EXAMPLE : TEST OF WHETHER  
A COMPLEX PARAMETER  $= 0$ .



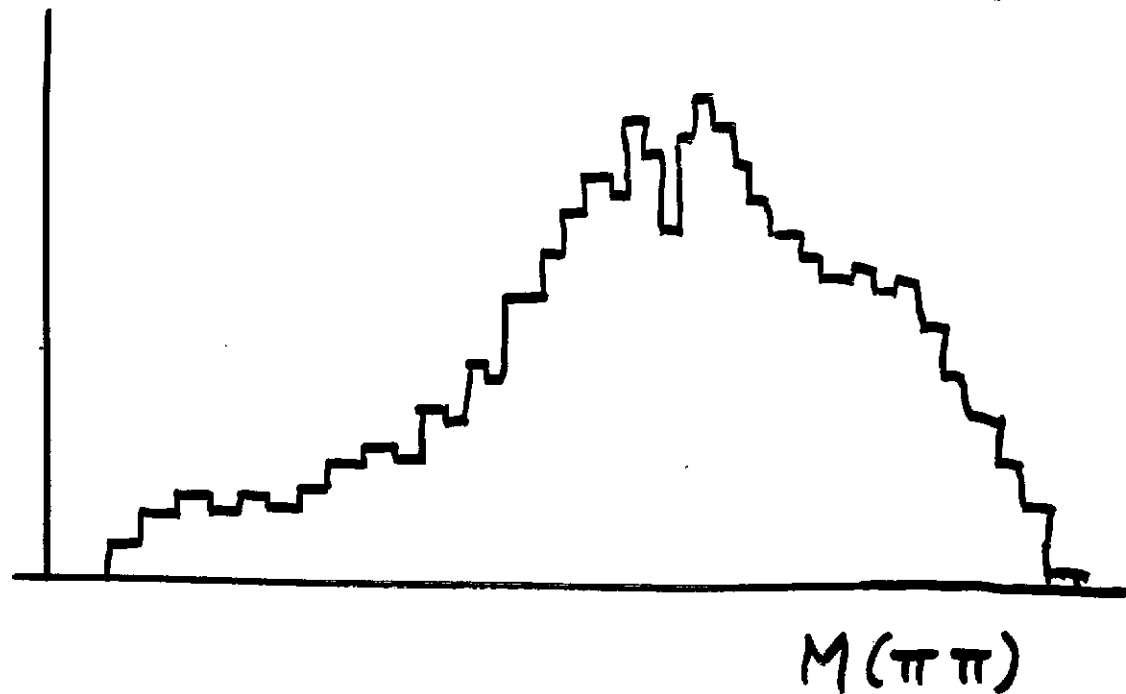
TEST THAT  $\theta = 0$  :

$$-2 \ln \left[ \frac{L(0)}{L(B)} \right] \sim \chi^2(2)$$

TEST THAT  $\theta$  IS REAL :

$$-2 \ln \left[ \frac{L(A)}{L(B)} \right] \sim \chi^2(1)$$

# A FAMOUS BIASED TEST: THE SPLITTING OF THE $A_2$



$H_0$ : THERE IS ONE PEAK

$H_1$ : THERE ARE TWO PEAKS

IT TURNS OUT THAT FOR THIS CASE  
THE ASYMPTOTIC LIMIT WHERE

$$\lambda \rightarrow \chi^2$$


IS VERY HIGH, AND FOR SMALL  $N$ ,  
YOU CAN ALWAYS FIND 2 PEAKS.

SOLUTION: MONTE CARLO.

# GOODNESS-OF-FIT

TEST  $H_0$  AGAINST  $\left\{ \begin{array}{l} \text{ALL} \\ \text{OTHER} \\ \text{POSSIBILITIES} \end{array} \right.$

$$P(X \in W_\alpha \mid H_0) = \alpha$$

  
SIZE OF TEST  
OR  
LEVEL OF CONFIDENCE

WE CAN NO LONGER MEASURE  
THE POWER OF A TEST  $(1-\beta)$   
BECAUSE THERE IS NO  
ALTERNATIVE HYPOTHESIS  $H_1$

CANNOT GIVE EVIDENCE FOR  $H_0$   
CAN ONLY GIVE EVIDENCE AGAINST  
THERE WILL ALWAYS BE SOME OTHER  
HYPOTHESIS THAT WILL FIT THE DATA  
BETTER THAN  $H_0$ .

# CHI-SQUARE

H  
14

UNFORTUNATE WORD -

THREE DIFFERENT MEANINGS

① A MATHEMATICAL FUNCTION

$$\chi^2(t, N) = \frac{\frac{1}{2} \left( \frac{t}{2} \right)^{\frac{N}{2} - 1} e^{-t/2}}{\Gamma(N/2)}$$

IT IS THE EXPECTED DISTRIBUTION<sub>(DENSITY)</sub>  
OF A SUM OF SQUARES OF  
GAUSSIAN-DISTRIBUTED VARIABLES

$$t = \sum_{i=1}^N X_i^2, \quad X \text{ GAUSSIAN}$$

# [MEANINGS OF CHI-SQUARE (CONT.)]

H  
15

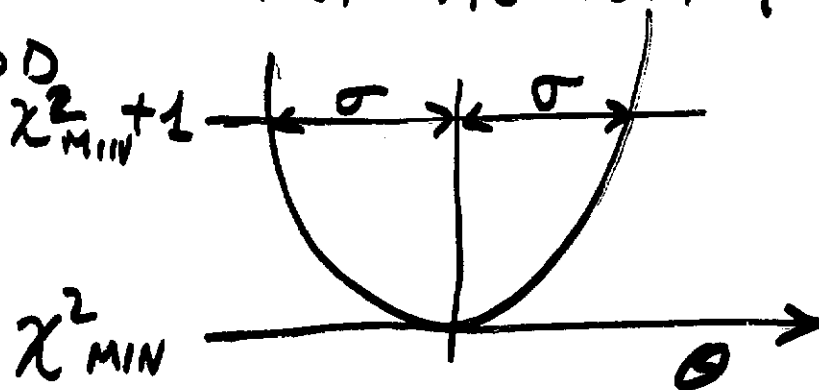
② A FUNCTION OF THE DATA, WHICH IS MINIMIZED, TO ESTIMATE PARAMETERS

$$\chi^2(\underline{\theta}) = \sum_{i=1}^N \left( \frac{Y_i - f(x_i, \underline{\theta})}{\sigma_i} \right)^2$$

ASYMPTOTICALLY

$$\chi^2(\underline{\theta}) \rightarrow -2 \ln L(\underline{\theta})$$

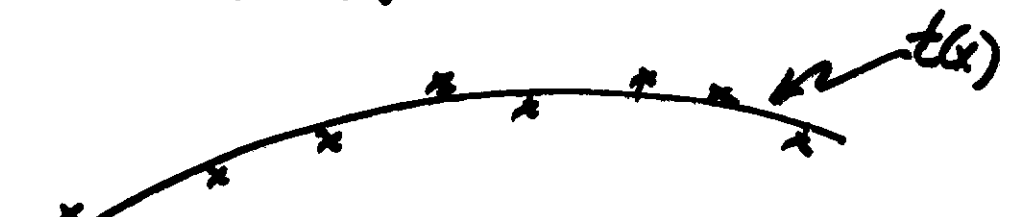
ESTIMATES OF PARAMETERS AND UNCERTAINTY INTERVALS CAN BE OBTAINED EXACTLY AS WITH LIKELIHOOD



③ A FUNCTION OF THE DATA  
(SIMILAR TO ②, BUT NO PARAM'S)  
USED TO TEST FOR  
GOODNESS-OF-FIT.

TWO MOST COMMON TYPES OF APPLICATION:

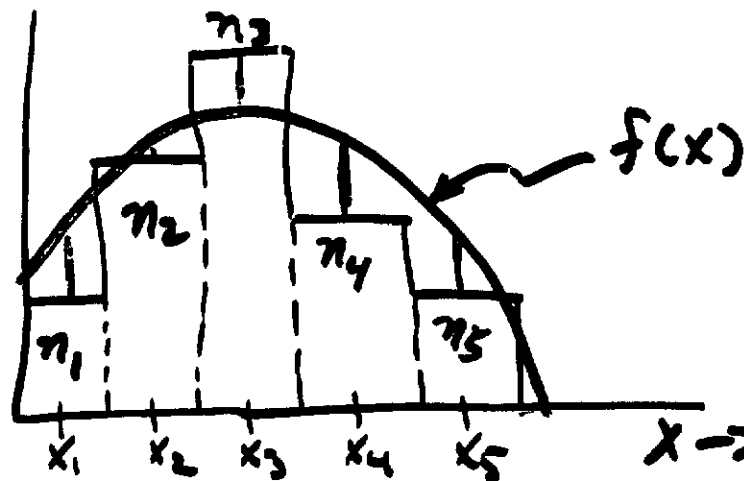
A. TRACK RECOGNITION



$$\chi^2_8 = \sum_{i=1}^8 \left( \frac{Y_i - t_i}{\sigma_i} \right)^2 \longleftrightarrow \chi^2_1(8)$$

IF THE MEASUREMENTS  $Y_i$   
ARE GAUSSIAN WITH STD. DEV =  $\sigma_i$ ,  
THEN THE SUM OF SQUARES SHOULD  
BE DISTRIBUTED LIKE  $\chi^2$  WITH  
8 DEGREES OF FREEDOM (8 MEASURED POINTS)

### 3B TESTING A MODEL BY FIT TO A HISTOGRAM



THE NUMBER OF EVENTS IN ONE BIN  
IS MULTINOMIAL-DISTRIBUTED

IF OVERALL NORMALIZATION IS IMPOSED  
OR POISSON-DISTRIBUTED IF NOT.

EVENTS IN DIFFERENT BINS ARE  
INDEPENDENT (APART FROM NORMALIZATION)

$$\chi^2 = \sum_{i=1}^5 \left( \frac{f(x_i) - n_i}{\sqrt{n_i}} \right)^2$$

SHOULD  
BE AN  
INTEGRAL  
OVER BIN

EVENT NUMBERS NEARLY GAUSSIAN  
→  $\chi^2$  NEARLY  $\chi^2_1$



USUALLY, PHYSICISTS DO  
BOTH ESTIMATION AND TESTING  
WITH THE SAME  $\chi^2$ -FUNCTION

FIRST - MINIMIZE  $\chi^2$  TO FIND  
BEST VALUE(S) OF PARAMETER(S)

THEN - USE VALUE OF  $\chi^2_{\min}$   
FOR TEST.

BUT NOW IT SHOULD BE  
DISTRIBUTED LIKE

$$\chi^2(N - p - 1)$$

↑                      ↑                      ↗  
BINS                      FREE                      NORMALIZATION  
                                 PARAMETERS                      CONSTRAINT

MAY BE!

MAJOR PROBLEM: BINNING  
(SOME INFORMATION MUST BE LOST!)  
MUST HAVE ENOUGH EVENTS/BIN

SO THE GAUSSIAN APPROXIMATION  
IS VALID. - BUT IF  
TOO FEW BINS, LOSE POWER  
TO TEST HYPOTHESIS.

### OPTIMUM BINNING:

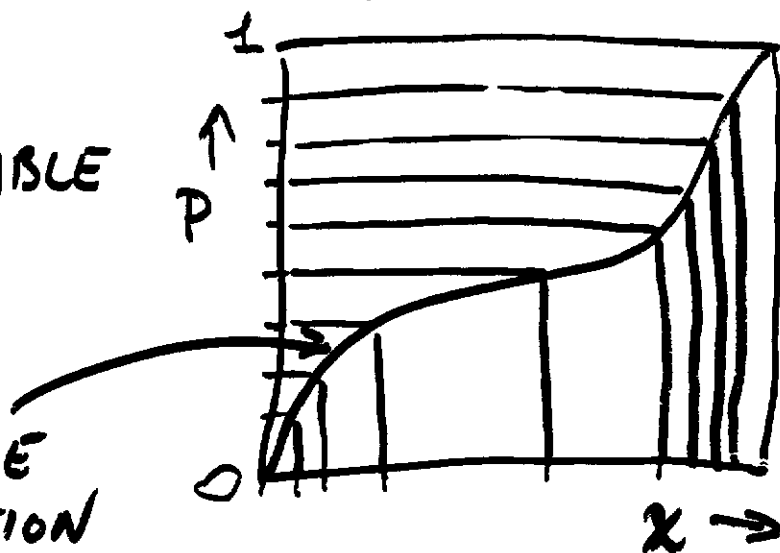
- ① CHOOSE  
EQUIPROBABLE  
BINS

CUMULATIVE  
DISTRIBUTION  
FUNCTION

$$F = \int f(x) dx$$

- ② MAKE EVENTS/BINS  $\approx 20$

- ③ IN ANY CASE, NOT LESS THAN  
10 (5?) EVENTS IN ANY BIN.



H  
19  
HOW WOULD ONE GO ABOUT  
TRYING TO FIND OPTIMUM  
BIN SIZE?

MAXIMIZE LOCAL POWER.

MAKE AN INFINITESIMAL

PERTURBATION TO  $H_0$ ,

CALCULATE POWER WITH RESPECT  
TO PERTURBED HYPOTHESIS,

AS A FUNCTION OF BIN SIZE.

ANSWER DEPENDS ON SHAPE OF PERTURB.

---

GOOD THING ABOUT  $\chi^2$  TEST:

(ASYMPTOTICALLY)

DISTRIBUTION-FREE

CONFIDENCE LEVEL  $\alpha$  DEPENDS  
ONLY ON NUMBER OF BINS,

NOT ON  $H_0$

CAN USE ONE  $\chi^2$  TABLE  
FOR ALL OF TIME

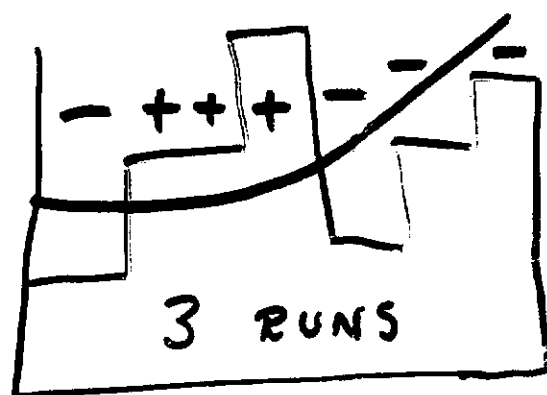
# THE RUNS TEST

IN  $\chi^2$  TEST THE SIGNS(!) OF THE DEVIATIONS ARE NOT USED.

THE RUNS TEST USES THIS INFORMATION UNDER  $H_0$ , ALL PATTERNS OF SIGNS ARE EQUALLY PROBABLE.

A RUN IS A SEQUENCE OF DEVIATIONS OF THE SAME SIGN.

THE PROBABILITY OF ANY NUMBER OF RUNS CAN BE CALCULATED EXACTLY.



$$E(R) = 1 + \frac{2MN}{M+N}$$

NUMBER OF + DEV.  $\swarrow$   
NEG. DEV.  $\nwarrow$

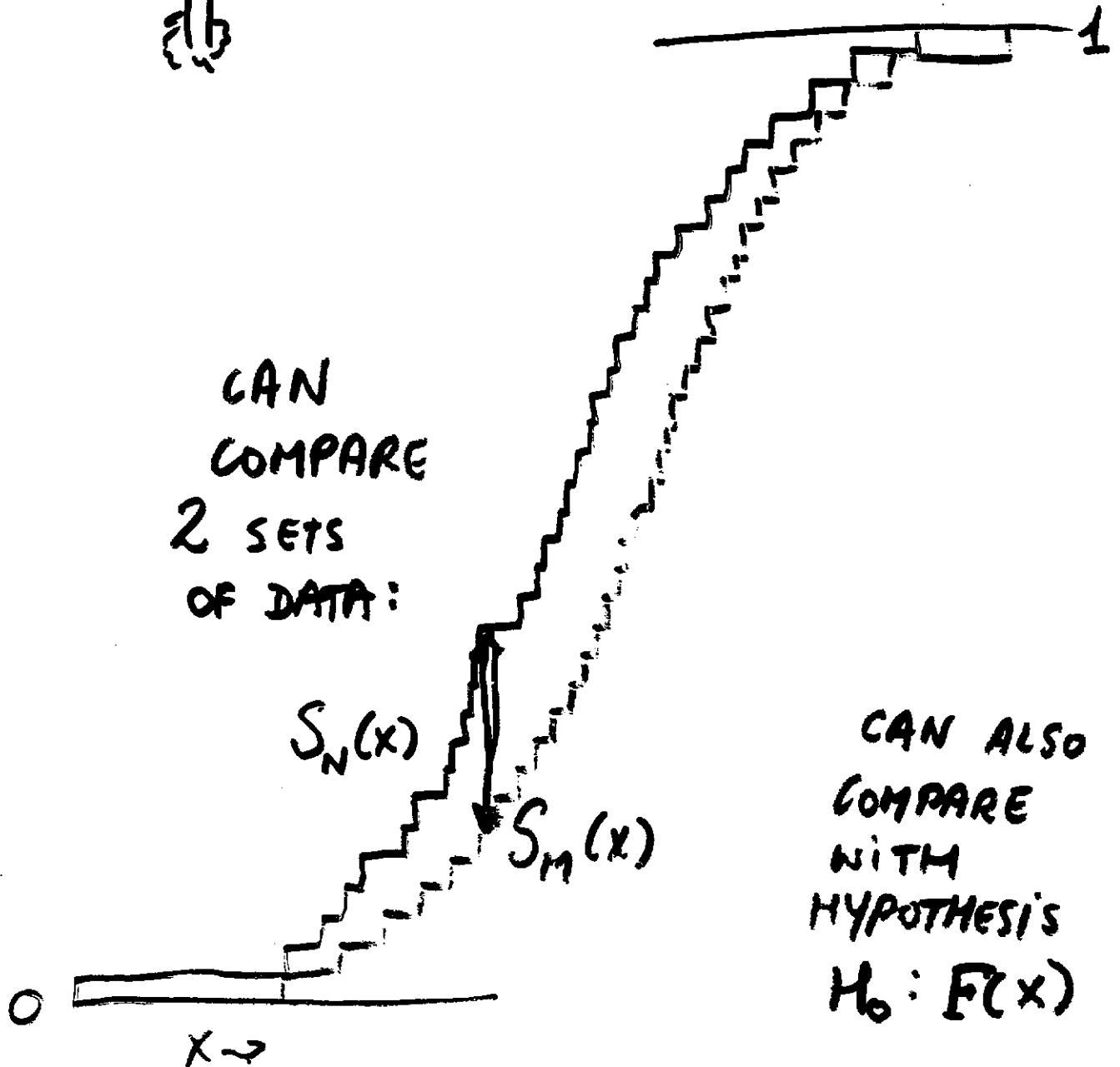
NUMBER OF RUNS  $\nearrow$

$$V(R) = \frac{2MN(2MN - M - N)}{(M+N)^2(M+N+1)}$$

# ORDER STATISTICS (VERY USEFUL)

= THE CUMULATIVE DISTRIBUTION OF  
THE DATA.

NO BINNING



2 TESTS BASED ON DIFFERENT  
MEASURES OF THE "DISTANCE"  
BETWEEN CUMULATIVE DISTRIBUTIONS:

① SMIRNOV-CRAMER-VON MISES

$$W^2 = \int_{-\infty}^{\infty} [S_N(x) - F(x)]^2 f(x) dx$$

DISTRIBUTION-FREE  
FOR ALL  $N$

$$E(NW^2) = \frac{1}{6}$$

$\alpha$	$NW^2$
0.1	0.347
0.05	0.461
0.01	0.743

② KOLMOGOROV-SMIRNOV

$$D_N = \max_{all x} |S_N(x) - F(x)|$$

PROBKL

CRITICAL VALUES

KNOWN ONLY

FOR 'LARGE  $N$ ',

BUT IN PRACTICE

$N > 10$  SEEMS VERY GOOD.

$\alpha$	$\sqrt{N} D_N$
0.1	1.22
0.05	1.36
0.01	1.63

## COMBINING TESTS:

H  
23

CAN WE MAKE  $> 1$  DIFFERENT  
TESTS ON THE SAME DATA?

YES, IF: ① TESTS ARE INDEPENDENT

$$[P(t_1, t_2 | H_0) = P(t_1 | H_0) \cdot P(t_2 | H_0)]$$

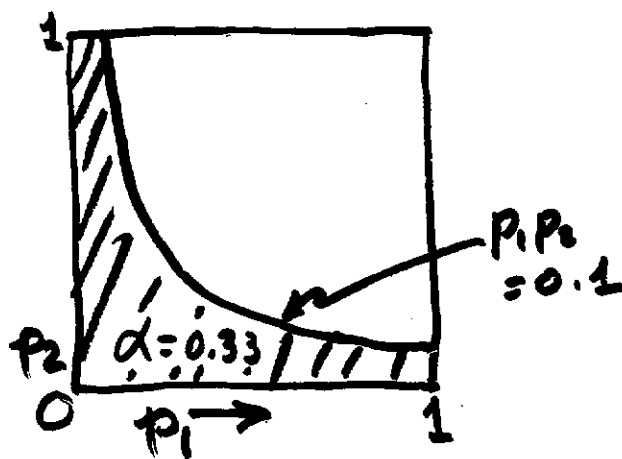
RUNS AND CHI-SQUARE TESTS (ASYMPT.) INDEPEND.

② YOU KNOW HOW TO COMBINE THE RESULTS.

$$\alpha = \alpha_1, \alpha_2 ? \quad \text{WRONG.}$$

$$\alpha =$$

$$\underline{\alpha_1 \alpha_2 (1 - \ln \alpha_1 \alpha_2)}$$



FOR DISCRETE TESTS (E.G. RUNS TEST)  
A SLIGHTLY MODIFIED METHOD  
MUST BE USED

# IMPORTANT ADVANCES IN FREQUENTIST STATISTICS SINCE 1950

- ROBUST ESTIMATION 1964 →

P. HUBER & OTHERS

- EXPLORATORY DATA ANALYSIS, <sup>1980 →</sup> DATA MINING <sup>1990 →</sup>  
J. TUKEY, J. FRIEDMAN

- BOOTSTRAP - 1982

B. EFRON



Table 8.5

Asymptotic efficiencies of location estimators

Distribution	Sample median	Sample mean	Sample midrange
Normal	0.64	1.00	0.00
Uniform	0.00	0.00	1.00
Cauchy	0.82	0.00	0.00
Double-exponential	1.00	0.50	0.00

(For asymptotic variances, see Table 8.6)

ROBUST ESTIMATION FINDS THE  
BEST COMPROMISE AMONG ALL THE ABOVE

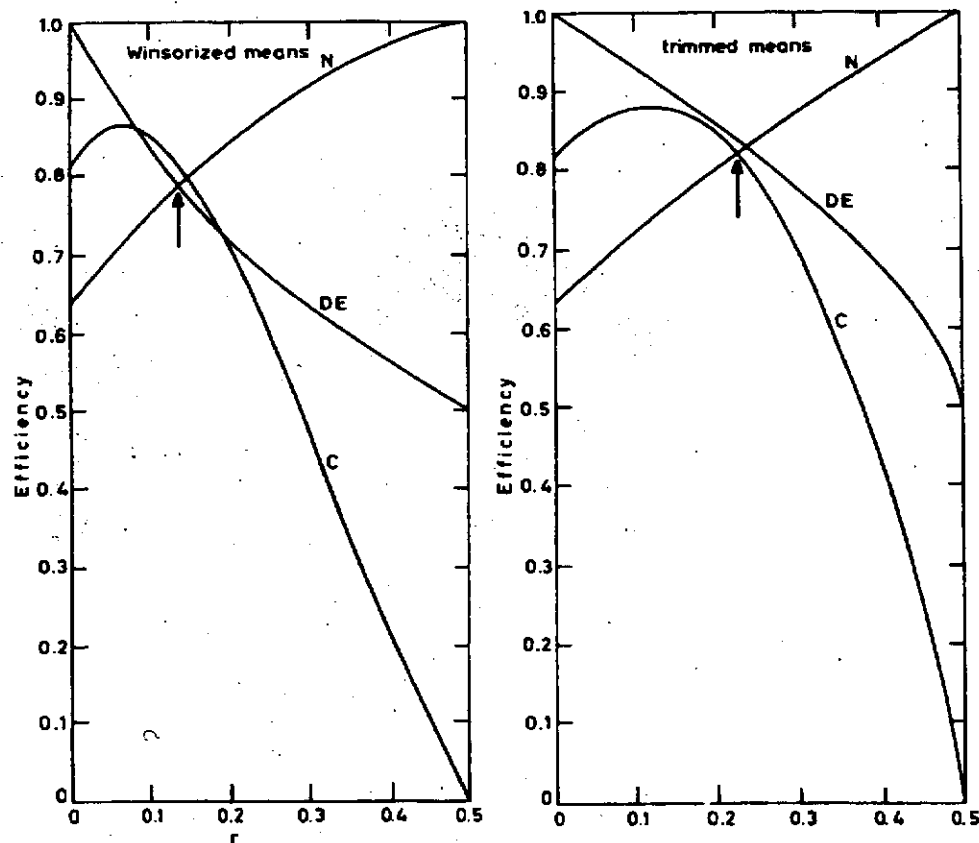


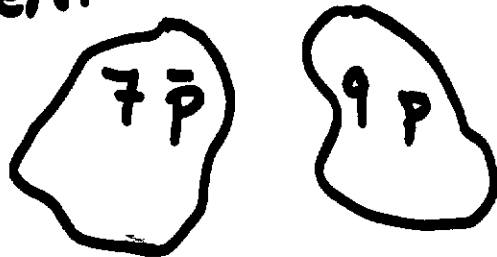
Fig. 8.3 Asymptotic efficiencies of trimmed and Winsorized means for normal (N), double-exponential (DE) and Cauchy (C) distributions.

# HYPOTHESIS TESTING BY RE-SAMPLING

## "BOOTSTRAP" (B. EFRON)

### TESTING FOR SMALL SAMPLES FROM UNKNOWN DISTRIBUTIONS

PROBLEM:



IS THE ENERGY  
DISTR. OF THE  $\bar{p}$   
THE SAME AS THE  $p$

OLD APPROACH:  $t = \frac{\langle E_{\bar{p}} \rangle - \langle E_p \rangle}{\sigma_{\bar{p}p}}$

[EFFICIENT IF ENERGY DISTR. ARE GAUSSIAN]

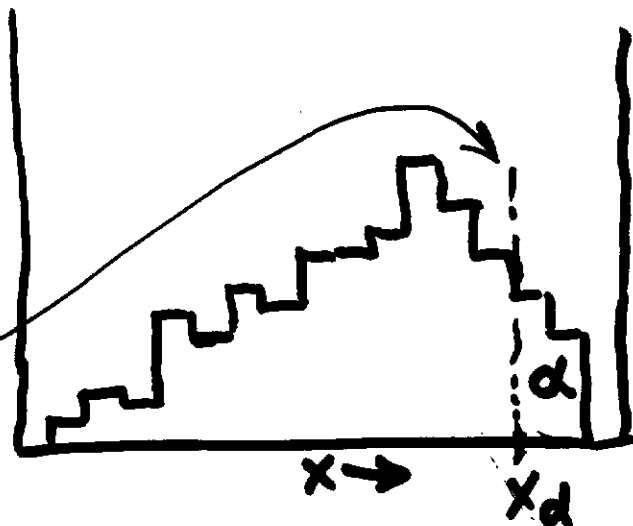
BOOTSTRAP: USE THE DATA ITSELF.

- PUT ALL 16 ENERGIES INTO ONE SAMPLE
- CONSIDER ALL WAYS OF DIVIDING  $16 \rightarrow 7+9$   
(THERE ARE  $\frac{16!}{7!9!} = 11,440$  DIFFERENT WAYS)

- FOR EACH COMBINATION,

CALCULATE  
 $\langle E_{\bar{p}} \rangle_i - \langle E_p \rangle_i$

ONE OF THESE  $x_i = x_d$   
IS THE REAL DATA SAMPLE.



# BAYESIAN HYPOTHESIS TESTING.

$$P(H_0 | \text{data}) = \frac{P(\text{data} | H_0) \cdot P(H_0)}{P(\text{data})}$$

can be written as

$$\sum_i P(\text{data} | H_i) \cdot P(H_i)$$

THIS WORKS FOR PARAMETER ESTIMATION

BUT IF  $H_0$  IS A HYPOTHESIS, THIS CANNOT BE NORMALIZED TO ANYTHING MEANINGFUL.

$\therefore$  BAYESIANS CAN ONLY COMPARE HYPOTHESES

$$\frac{P(H_0 | \text{data})}{P(H_1 | \text{data})} = \frac{P(\text{data} | H_0) P(H_0)}{P(\text{data} | H_1) P(H_1)}$$

THESE ARE LIKELIHOOD FUNCTIONS

AND THIS IS CALLED THE "BAYES FACTOR"

	<u>2 HYPOTH.</u>	<u>1 HYPOTH.</u>
FREQUENTIST	LOSS AND CONTAMINATION	ONLY LOSS
BAYES	RATIO OF PROBABILITIES	NOTHING.

# BAYESIAN HYPOTH. TESTS (2)

1. MUST HAVE 2 HYPOTHESES
2. MUST PUT IN PRIOR PROBABILITIES
3. CAN HANDLE COMPOSITE HYPOTHESES EASILY
  - JUST INTEGRATE OVER  $P(\mu)$
4. CANNOT CALCULATE LOSS OR CONTAMINATION (BUT THEY EXIST!) BECAUSE: DO NOT "ACCEPT" OR "REJECT" HYPOTHESES.
5. USES ONLY OBSERVED DATA (AND PRIOR P.)
  - LIKELIHOOD PRINCIPLE
  - STOPPING RULE

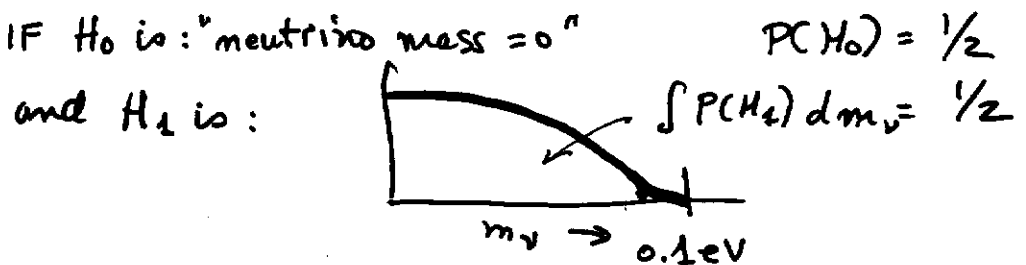
## HOW TO TEST GOODNESS-OF-FIT?

- MUST CREATE AN ALTERNATIVE HYPOTHESIS.
- USUALLY LOOKS SOMETHING LIKE:

IF  $H_0$  is: "neutrino mass = 0"

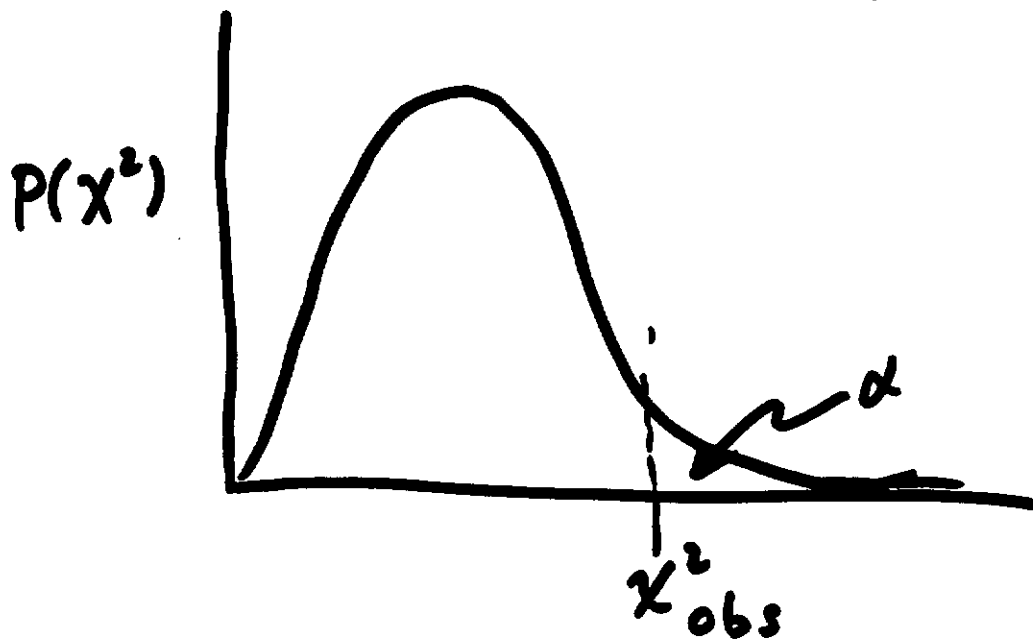
$$P(H_0) = 1/2$$

and  $H_1$  is:



- SO YOU HAVE A "POINT NULL" HYPOTHESIS AND A CONTINUOUS ALTERNATIVE. YOU NEED ONE POINT-PRIOR AND ONE PRIOR DISTR.

# RECALL FREQUENTIST $\chi^2$ TEST



$$P(\chi^2 > \chi_{obs}^2 | H_0) = \alpha$$

1. LARGE VALUES OF  $\chi^2$  IMPLY "BAD" FIT.
2.  $P(\chi_{obs}^2)$  IS MEANINGLESS  
(ONLY A DENSITY)

ONLY  $\int_A^B P(\chi^2) d\chi^2$  IS A PROBABILITY

3. BAYESIAN CANNOT USE THIS  
BECAUSE IT DEPENDS ON THE  
PROBABILITY OF DATA NOT  
OBSERVED.

BOTH FREQUENTISTS AND BAYESIANS  
WOULD LIKE TO ASK THE QUESTION:

WHAT IS THE PROBABILITY OF  
OBTAINING EXACTLY THE DATA  
I OBSERVED IF  $H_0$  IS TRUE?

BUT THE ANSWER IS ALWAYS ZERO  
IF THE DATA ARE CONTINUOUS BECAUSE  
 $P(X | H_0)$  IS ONLY A DENSITY.

- FREQUENTISTS SOLVE THIS PROBLEM  
BY FINDING A FINITE CRITICAL VOLUME  
WITH A FINITE PROBABILITY
- BAYESIANS CANNOT DO THAT  
(DATA NOT OBSERVED)

THINK OF A PROBABILITY DENSITY FUNCTION (P.D.F.)  
TRANSFORMED INTO ITS "NATURAL" OR "NORMALIZED" VARIABLE



THEN IT IS CLEAR THAT NO VALUE OF  $X^2$   
IS "MORE PROBABLE" THAN ANOTHER!

B. EFRON\*

Originally a talk delivered at a conference on Bayesian statistics, this article attempts to answer the following question: why is most scientific data analysis carried out in a non-Bayesian framework? The argument consists mainly of some practical examples of data analysis, in which the Bayesian approach is difficult but Fisherian/frequentist solutions are relatively easy. There is a brief discussion of objectivity in statistical analyses and of the difficulties of achieving objectivity within a Bayesian framework. The article ends with a list of practical advantages of Fisherian/frequentist methods, which so far seem to have outweighed the philosophical superiority of Bayesianism.

KEY WORDS: Fisherian inference; Frequentist theory; Neyman-Pearson-Wald; Objectivity.

## 3. FISHERIAN STATISTICS

In its inferential aspects Fisherian statistics lies closer to Bayes than to NPW in one crucial way: the assumption that there is a *correct* inference in any given situation. For example, if  $x_1, x_2, \dots, x_{20}$  is a random sample from a Cauchy distribution with unknown center  $\theta$ ,

$$f_{\theta}(x_i) = \frac{1}{\pi[1 + (x_i - \theta)^2]}$$

then in the absence of prior knowledge about  $\theta$  the correct 95% central confidence interval for  $\theta$  is, to a good approximation,

$$\hat{\theta} \pm 1.96 / \sqrt{-\bar{I}_{\hat{\theta}}}$$

... hood estimator (MLE) and log-likelihood function (lly) equally good ap

## 1. INTRODUCTION

The title is a reasonable question to ask on at least two counts. First of all, everyone used to be a Bayesian. Laplace wholeheartedly endorsed Bayes's formulation of the inference problem, and most 19th-century scientists followed suit. This included Gauss, whose statistical work is usually presented in frequentist terms.

A second and more important point is the cogency of the Bayesian argument. Modern statisticians, following the lead of Savage and de Finetti, have advanced powerful theoretical reasons for preferring Bayesian inference. A byproduct of this work is a disturbing catalogue of inconsistencies in the frequentist point of view.

Nevertheless, everyone is not a Bayesian. The current era is the first century in which statistics has been widely used for scientific reporting, and in fact, 20th-century statistics is mainly non-Bayesian. [Lindley (1975) predicts a change for the 21st!] What has happened?

The title is a reasonable question to ask on at least two counts. First of all, everyone used to be a Bayesian. Laplace wholeheartedly endorsed Bayes's formulation of the inference problem, and most 19th-century scientists followed suit. This included Gauss, whose statistical work is usually presented in frequentist terms.

Th of two Kiefer decision frequentists. markable degree

1920 and 1935. NPW began 1933, asymptoting in the 1950s, though there have continued to be significant advances such as Stein estimation, empirical Bayes, and robustness theory.

Working together in rather uneasy alliance, Fisher and NPW dominate current theory and practice, with Fisherian ideas particularly prevalent in applied statistics. I am going to try to explain why.

\*B. Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305.

optimal inference. In short, he does not have to think a lot about the specific situation in order to get on toward its solution.

Bayesian theory requires a great deal of thought about the given situation to apply sensibly. This is seen clearly in the efforts of Novick (1973), Kadane, Dickey, Winkler, Smith, and Peters (1980), and many others to at least partially automate Bayesian inference. All of this thinking is admirable in principle, but not necessarily in day-to-day practice. The same objection applies to some aspects of

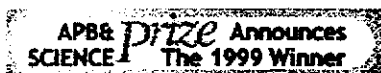
), is not correct

For any given reduction to an nce is obvious. king inferences al model

(1)

ch reductions: ibutions, and inventions, all something like superseded an rson's method s.)

rried out in a natic nature of ation is the orig- ly applicable and ation, the working od in an automatic (scented hands) of going



BUY REAGENTS ONLINE!

Science  
MAGAZINE

HOME

HELP

SEARCH

ARCHIVE

SUBSCRIPTIONS

FEEDBACK

ORDER an article

SIGN IN

\* || Sign In as Member || FAQ

STATISTICS:

## Bayes Offers a 'New' Way to Make Sense of Numbers

David Malakoff

**A 236-year-old approach to statistics is making a comeback, as its ability to factor in hunches as well as hard data finds applications from pharmaceuticals to fisheries**

- ▶ Summary of this Article
- ▶ Similar articles found in:  
SCIENCE Online
- ▶ Search Medline for articles by:  
Malakoff, D.
- ▶ Alert me when:  
new articles cite this article

▶ Collections under which this article appears:  
Computers/Mathematics

After 15 years, environmental researcher Kenneth Reckhow can still feel the sting of rejection. As a young scientist appearing before an Environmental Protection Agency review panel, Reckhow was eager to discuss his idea for using an unorthodox statistical approach in a water-quality study. But before he could say a word, an influential member of the panel unleashed a rhetorical attack that stopped him cold. "As far as he was concerned, I was a Bayesian, and Bayesian statistics were worthless." recalls Reckhow, now at Duke University in Durham, North Carolina. "The idea was dead before I even got to speak."

Reckhow is no longer an academic outcast. And the statistical approach he favors, named after an 18th century Presbyterian minister, Thomas Bayes, now receives a much warmer reception from the scientific establishment. Indeed, Bayesian statistics, which allows researchers to use everything from hunches to hard data to compute the probability that a hypothesis is correct, is experiencing a renaissance in fields of science ranging from astrophysics to genomics and in real-world applications such as testing new drugs and setting catch limits for fish. The long-dead minister is also weighing in on lawsuits and public policy decisions (see p. 1462), and is even making an appearance in consumer products. It is his ghost, for instance, that animates the perky paperclip that pops up on the screens of computers running Microsoft Office software, making Bayesian guesses about what advice the user might need. "We're in the midst of a Bayesian boom," says statistician John Geweke of the University of Iowa, Iowa City.

Advances in computers and the limitations of traditional statistical methods are part of the reason for the new popularity of this old approach. But researchers say the Bayesian approach is also appealing because it allows them to factor expertise and prior knowledge into their computations—something that traditional methods frown upon. In addition, advocates say it produces answers that are easier to understand and forces users to be explicit about biases obscured by reigning "frequentist" approaches.

To be sure, Bayesian proponents say the approach is no panacea—and the technique has detractors. Some researchers fear that because Bayesian analysis can take into account prior opinion, it could spawn less objective evaluations of experimental results. "The problem is that prior beliefs can be just plain wrong"



might bias the results.

Such prior information can be very helpful to researchers trying to discern patterns in massive data sets or in problems where many variables may be influencing an observed result. Larry Bretthorst of Washington University in St. Louis, Missouri, for instance, developed Bayesian software that improved the resolution of nuclear magnetic resonance (NMR) spectrum data—used by chemists to figure out the structure of molecules—by several orders of magnitude. It uses prior knowledge about existing NMR spectra to clarify confusing data, yielding resolution improvements that were "so startling that other researchers had a hard time believing he hadn't made a mistake," says Kevin Van Horn, an independent computer scientist in American Fork, Utah.

Genomics researchers have also become converts. "You just say 'Bayesian,' and people think you are some kind of genius." says statistician Gary Churchill of The Jackson Laboratory in Bar Harbor, Maine, who is working on ways to analyze the flood of data produced by DNA sequencing and gene expression research. Some researchers, for instance, are using what they already know about a DNA sequence to identify other sequences that have a high probability of coding for proteins that have similar functions or structures, notes Jun Liu, a statistician at Stanford University in Palo Alto, California. "No easy frequentist method can achieve this," he says. Similarly, "Bayesian has become the method of choice" in many astrophysics studies, says astrophysicist Tom Lored~~o~~ of Cornell University in Ithaca, New York. The approach has allowed users to discern weak stellar signal patterns amid cosmic background noise and take a crack at estimating the locations and strengths of mysterious gamma ray bursts.

### **Lifesaving statistics?**

In other fields, such as drug and medical device trials, Bayesian methods could have practical advantages, say advocates. Indeed, at a 2-day conference last year, the Food and Drug Administration (FDA) office that approves new devices strongly urged manufacturers to adopt Bayesian approaches, arguing that they can speed decisions and reduce costs by making trials smaller and faster.

Telba Irony, one of two Bayesian statisticians recently hired by the division, says the savings flow from two advantages of the Bayesian approach—the ability to use findings from prior trials and flexibility in reviewing results while the trial is still running. Whereas frequentist methods require trials to reach a prespecified sample size before stopping, Bayesian techniques allow statisticians to pause and review a trial to determine—based on prior experience—the probability that adding more patients will appreciably change the outcome. "You should be able to stop some trials early," she says. So far, just a handful of the 27,000 device firms regulated by FDA have taken advantage of the approach. But FDA biostatistician Larry Kessler hopes that up to 5% of device trials will be at least partly Bayesian within a few years. "We're not going to change the statistical paradigm overnight," he says. "There is still a healthy degree of skepticism out there."

Such skepticism has also limited the use of Bayesian approaches in advanced drug trials, a potentially much bigger arena. But a team led by M. D. Anderson's Berry and researchers at Pfizer Inc.'s central research center in Sandwich, England, is about to challenge that taboo. Next June, using a heavily Bayesian study design, the company plans to begin human trials aimed at finding the safest effective dose of an experimental stroke drug designed to limit damage to the brain. The trial—called a phase II dose ranging trial—will help the company decide whether to move the drug into final testing trials. "There are huge economic consequences on the line," says Pfizer statistician Andy Grieve.

The team believes that Bayesian methods will allow the company to reach conclusions using 30% fewer

# The Return of the Prodigal: Bayesian Inference For Astrophysics

THOMAS J. LOREDO

Department of Astronomy, Space Sciences Building,  
Cornell University, Ithaca, New York 14853

## ABSTRACT

Astronomers are skeptical of statistical analyses, the more so the more sophisticated they are. This has been true especially of Bayesian methods, despite the fact that such methods largely originated in the astronomical analyses of Laplace and his contemporaries in the early 1800s. I argue here that astronomers hold statistics in low regard because many astronomers are poor statisticians. Further, I argue that astronomers are poor statisticians because the frequentist methods they use have characteristics that invite statistical sloppiness when they are used by nonexperts. The Bayesian approach to statistical inference does not share these characteristics; adoption of Bayesian methods by astronomers thus promises to improve statistical practice in astronomy. I present a simplified discussion of some of the issues arising in the recent analysis of an important astrophysical data set—that provided by the Cosmic Background Explorer satellite—to illustrate some of the practical advantages of a Bayesian outlook. I offer some advice on how to educate astronomers about Bayesian methods. I conclude with a brief survey of recent applications of Bayesian methods to the analysis of astrophysical data. The breadth and number of these applications may well indicate that the time for Bayesian methods to return to the field of their origin has arrived.

---

## 1. INTRODUCTION

One could claim without too much exaggeration that statistical inference was invented because of astronomy. As noted by Stigler (1986), problems associated with reconciling discrepant observations in astronomy and geodesy motivated such legendary mathematicians and astronomers as Legendre, Laplace, and Gauss to develop the foundations of statistical inference based on probability theory. Their analyses of astronomical and geodetic problems led to such notions as the use of means to reduce uncertainty, the method of least squares, the normal distribution, the central limit theorem, and the “method of inverse probability” (inference using Bayes’s theorem). Their work was essentially Bayesian in outlook, and the first mature treatise on statistical inference—Laplace’s *Theorie Analytique des Probabilités* (Laplace 1812)—could fairly be called a Bayesian monograph.

Viewed from the present, this aspect of the early history of statistical inference is doubly ironic. First, contemporary astronomers (and physical scientists more generally) seldom receive any formal training in statistics, and frequently display a skepticism of sophisticated statistical analysis that borders on disdain. Second, until very recently, Bayesian methods *in particular* have been poorly understood and unwelcome tools among physical scientists. This has been true despite the fact that the most influential and practical Bayesian text of the first half of this century was written by a geologist and astronomer, Sir Harold Jeffreys (Jeffreys 1939).

THE BAYESIAN STATISTICIANS AT THIS CONFERENCE  
REPLIED THAT IN ORDER TO USE BAYESIAN METHODS  
CORRECTLY, YOU ALSO HAVE TO KNOW WHAT YOU ARE DOING,

# Understanding Data Better with Bayesian and Global Statistical Methods

William H. Press

Harvard-Smithsonian Center for Astrophysics

April 16, 1996

## Abstract

To understand their data better, astronomers need to use statistical tools that are more advanced than traditional "freshman lab" statistics. As an illustration, the problem of combining apparently incompatible measurements of a quantity (the Hubble constant, e.g.) is presented from both the traditional, and a more sophisticated Bayesian, perspective. Explicit formulas are given for both treatments.

## 1 Introduction

Understanding data better is *always* an unsolved problem in astrophysics, although perhaps not in exactly the sense intended by the conference organizers. While other papers in this volume are more specifically directed at individual sub-areas of astrophysical theory, my contribution is intentionally more longitudinal: I hope that it is applicable to *all* the other areas surveyed.

If the spirit of this volume is to present a menu – a movable feast, *indeed* – of opportunities for thesis projects of smart second-year graduate students, then the opportunity that I would like to offer is one of voluntary self-choice: Whatever your choice of area, make the choice to live your professional life at a high level of statistical sophistication, and not at the level – basically freshman lab level – that is the unfortunate common currency of most astronomers. Thereby will we all move forward together.

# Statistical Analysis and the Illusion of Objectivity

James O. Berger

Donald A. Berry

In many scientific journals, statistical analysis is used to give the seal of objectivity to conclusions. Yet this general perception of the objectivity of statistics, and perhaps of science in general, may be misguided. Let us be careful here; objectivity is a loaded word, and the next worst thing to being a fraud is to be "nonobjective." We are not going to discuss the manner in which a scientist strives to obtain objective evidence. Rather, we will discuss whether or not it is possible to arrive at an objective conclusion based on data from an experiment.

We grant that objective data can be obtained, but we will argue that reaching sensible conclusions from statistical analysis of these data may require subjective input.

This conclusion is in no way harmful or demeaning to statistical analysis. Far from it; to acknowledge the subjectivity inherent in the interpretation of data is to recognize the central role of statistical analysis as a formal mechanism by which new evidence can be integrated with existing knowledge. Such a view of statistics as a dynamic discipline is far from the common perception of a rather dry, automatic technology for processing data.

Acknowledging the subjectivity of statistical analysis would be healthy for science as a whole for at least two reasons. The first is that the straightforward methods of subjective statistical analysis, called Bayesian analysis, yield answers which are much easier to understand than standard statistical answers, and hence much less likely to be misinterpreted. This will be dramatically illustrated in our first example.

The second reason is that even standard statistical methods turn out to be based on subjective input—input of a type that science should seek to avoid. In particular,

standard methods depend on the intentions of the investigator, including intentions about data that might have been obtained but were not. This kind of subjectivity is doubly dangerous. First, it is hidden; few researchers realize how subjective standard methods really are. Second, the subjective input arises from the producer rather than the consumer of the data—from the investigator rather than the individual scientist who reads or is told the results of the experiment.

This article is an introduction to one side of a long and ongoing fundamental debate in statistics between the subjectivists, or Bayesians, and the nonsubjectivists. The Bayesian school of statistics is named after the Reverend Thomas Bayes, who proposed the basic ideas in 1763 (1). The opposing school is actually many schools going by different names; we will use "standard statistics" as a generic name. If you have a passing familiarity with statistical ideas, they are almost certainly what we call standard.

The debate involves a number of issues in addition to that of subjectivity. A closely related concern is "conditioning" (2). Simply put, conditionalists (including Bayesians) feel that only the actual data are relevant to the inferences drawn from an experiment; in standard statistics, as suggested above, the thoughts of the investigator about data that might have been observed but were not are also deemed relevant. This important issue and its ramifications will be clarified as we proceed.

In many—perhaps most—statistical applications, the various approaches will give very similar answers. There are at least two kinds of situations, however, in which major differences of interpretation arise. The first is the testing of precise hypotheses, such as scientific theories, and the second is the analysis of accumulating data, commonly encountered in clinical trials. We will give an example of each type.

## Testing a precise hypothesis

Let us start with a simple example of testing a precise hypothesis. Suppose an experiment is conducted to study the effectiveness of vitamin C in treating the common cold, and that standard statistical analysis finds "significant evidence at the 0.05 level" that vitamin C has an effect. Such statements concerning statistical signifi-

---

*Acknowledging the role of subjectivity in the interpretation of data could open the way for more accurate and flexible statistical judgments*

---

James Berger is the Richard M. Brumfield Distinguished Professor of Statistics at Purdue University. He received his Ph.D. in mathematics from Cornell University in 1974, and has taught at Purdue since then. His research interests include Bayesian statistics and decision theory. Donald Berry is Professor and Chairman of the Department of Theoretical Statistics at the University of Minnesota. He received his Ph.D. in statistics from Yale University in 1971, and began his appointment at the University of Minnesota in that year. His research focuses on Bayesian inference, sequential decision-making, and their application to medical problems. Address for Dr. Berger: Statistics Department, Purdue University, West Lafayette, IN 47907.

# Testing a Point Null Hypothesis: The Irreconcilability of $P$ Values and Evidence

JAMES O. BERGER and THOMAS SELKE\*

The problem of testing a point null hypothesis (or a "small interval" null hypothesis) is considered. Of interest is the relationship between the  $P$  value (or observed significance level) and conditional and Bayesian measures of evidence against the null hypothesis. Although one might presume that a small  $P$  value indicates the presence of strong evidence against the null, such is not necessarily the case. Expanding on earlier work [especially Edwards, Lindman, and Savage (1963) and Dickey (1977)], it is shown that actual evidence against a null (as measured, say, by posterior probability or comparative likelihood) can differ by an order of magnitude from the  $P$  value. For instance, data that yield a  $P$  value of .05, when testing a normal mean, result in a posterior probability of the null of at least .30 for any objective prior distribution. ("Objective" here means that equal prior weight is given the two hypotheses and that the prior is symmetric and nonincreasing away from the null; other definitions of "objective" will be seen to yield qualitatively similar results.) The overall conclusion is that  $P$  values can be highly misleading measures of the evidence provided by the data against the null hypothesis.

KEY WORDS:  $P$  values; Point null hypothesis; Bayes factor; Posterior probability; Weighted likelihood ratio.

## 1. INTRODUCTION

We consider the simple situation of observing a random quantity  $X$  having density (for convenience)  $f(x | \theta)$ ,  $\theta$  being an unknown parameter assuming values in a parameter space  $\Theta \subset \mathbb{R}^1$ . It is desired to test the null hypothesis  $H_0: \theta = \theta_0$  versus the alternative hypothesis  $H_1: \theta \neq \theta_0$ , where  $\theta_0$  is a specified value of  $\theta$  corresponding to a fairly sharply defined hypothesis being tested. (Although exact point null hypotheses rarely occur, many "small interval" hypotheses can be realistically approximated by point nulls; this issue is discussed in Sec. 4.) Suppose that a classical test would be based on consideration of some test statistic  $T(X)$ , where large values of  $T(X)$  cast doubt on  $H_0$ . The  $P$  value (or observed significance level) of observed data,  $x$ , is then

$$p = \Pr_{\theta=\theta_0}(T(X) \geq T(x)).$$

*Example 1.* Suppose that  $X = (X_1, \dots, X_n)$ , where the  $X_i$  are iid  $\mathcal{N}(\theta, \sigma^2)$ ,  $\sigma^2$  known. Then the usual test statistic is

$$T(X) = \sqrt{n}(\bar{X} - \theta_0)/\sigma,$$

where  $\bar{X}$  is the sample mean, and

$$p = 2(1 - \Phi(t)),$$

where  $\Phi$  is the standard normal cdf and

$$t = T(x) = \sqrt{n}(\bar{x} - \theta_0)/\sigma.$$

We will presume that the classical approach is the report of  $p$ , rather than the report of a (pre-experimental) Ney-

man-Pearson error probability. This is because (a) most statisticians prefer use of  $P$  values, feeling it to be important to indicate how strong the evidence against  $H_0$  is (see Kiefer 1977), and (b) the alternative measures of evidence we consider are based on knowledge of  $x$  [or  $t = T(x)$ ]. [For a comparison of Neyman-Pearson error probabilities and Bayesian answers, see Dickey (1977).]

There are several well-known criticisms of testing a point null hypothesis. One is the issue of "statistical" versus "practical" significance, that one can get a very small  $p$  even when  $|\theta - \theta_0|$  is so small as to make  $\theta$  equivalent to  $\theta_0$  for practical purposes. [This issue dates back at least to Berkson (1938, 1942); see also Good (1983), Hodges and Lehmann (1954), and Solo (1984) for discussion and history.] Also well known is "Jeffreys's paradox" or "Lindley's paradox," whereby for a Bayesian analysis with a fixed prior and for values of  $t$  chosen to yield a given fixed  $p$ , the posterior probability of  $H_0$  goes to 1 as the sample size increases. [A few references are Good (1983), Jeffreys (1961), Lindley (1957), and Shafer (1982).] Both of these criticisms are dependent on large sample sizes and (to some extent) on the assumption that it is plausible for  $\theta$  to equal  $\theta_0$  exactly (more on this later).

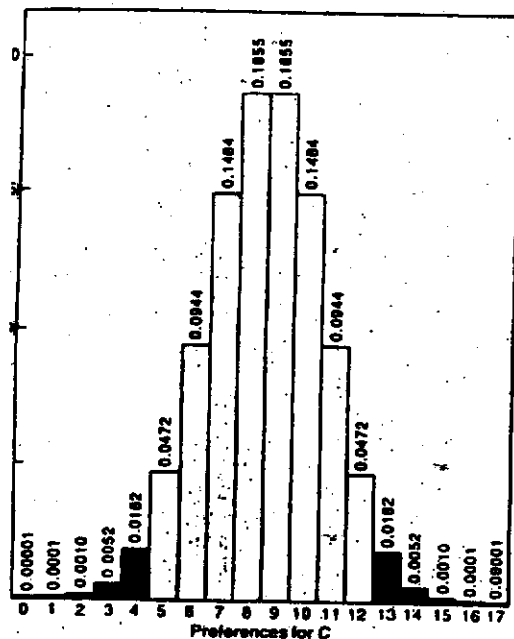
The issue we wish to discuss has nothing to do (necessarily) with large sample sizes for even exact point nulls (although large sample sizes do tend to exacerbate the conflict, the Jeffreys-Lindley paradox being the extreme illustration thereof). The issue is simply that  $p$  gives a very misleading impression as to the validity of  $H_0$ , from almost any evidentiary viewpoint.

*Example 1 (Jeffreys's Bayesian Analysis).* Consider a Bayesian who chooses the prior distribution on  $\theta$ , which gives probability  $\frac{1}{2}$  each to  $H_0$  and  $H_1$ , and spreads the mass out on  $H_1$  according to an  $\mathcal{N}(\theta_0, \sigma^2)$  density. [This prior is close to that recommended by Jeffreys (1961) for testing a point null, though he actually recommended a Cauchy form for the prior on  $H_1$ . We do not attempt to defend this choice of prior here. Particularly troubling is the choice of the scale factor  $\sigma^2$  for the prior on  $H_1$ , though it can be argued to at least provide the right "scale." See Berger (1985) for discussion and references.] It will be seen in Section 2 that the posterior probability,  $\Pr(H_0 | x)$ , of  $H_0$  is given by

$$\Pr(H_0 | x) = (1 + (1 + n)^{-1/2} \exp\{t^2/[2(1 + 1/n)]\})^{-1}, \quad (1.1)$$

some values of which are given in Table 1 for various  $n$  and  $t$  (the  $t$  being chosen to correspond to the indicated

\* James O. Berger is the Richard M. Brumfield Distinguished Professor and Thomas Sellke is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. Research was supported by National Science Foundation Grant DMS-8401996. The authors are grateful to L. Mark Berliner, Iain Johnstone, Robert Keener, Prem Puri, and Herman Rubin for suggestions or interesting arguments.



1. The hypothetical vitamin C experiment illustrates the required to arrive at a  $P$ -value. The binomial distribution of  $n$ ces for  $C$  under hypothesis  $H$ , shown here, is the same as tribution of heads that would result when a coin is tossed 17 Results more extreme than the 13 preferences for  $C$  actually ed constitute the set  $R$ , shown in color. The  $P$ -value is the ability of  $R$ , 0.049, which is calculated by adding the individual abilities of the results indicated in color.

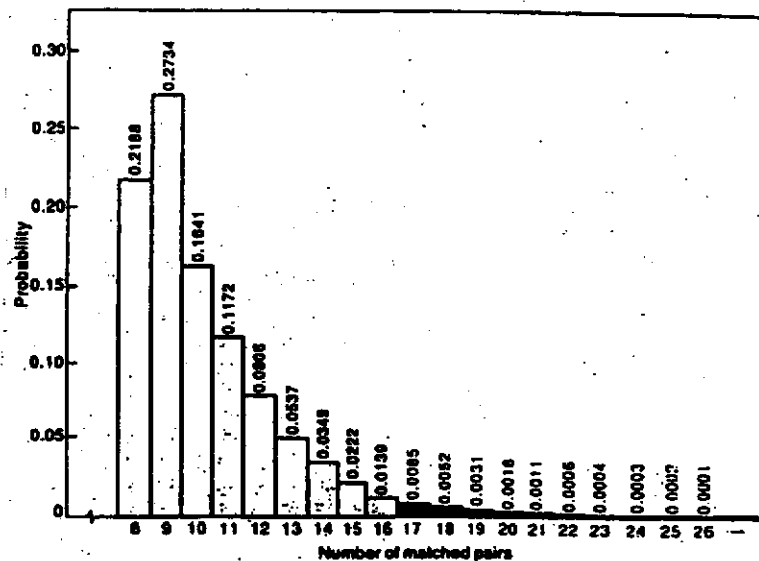


Figure 2. The effect of the intentions of the investigator on the  $P$ -value is demonstrated by this graph, which shows the probability distribution of the number of matched pairs under hypothesis  $H$  when the vitamin C experiment is designed to end as soon as at least 6 Cs and 4 Ps have been observed rather than after the treatment of 17 matched pairs. (The same distribution would result if a coin were tossed until at least 6 heads and 4 tails had been observed.) If the fourth  $P$  occurred at the seventeenth pair, the data observed—13 Cs and 4 Ps—would be the same regardless of which design the investigator had in mind when he or she stopped the experiment. However, the  $P$ -value obtained by adding the probabilities of (color) will now be 0.821 rather than the 0.049 calculated from the probabilities in Figure 1.

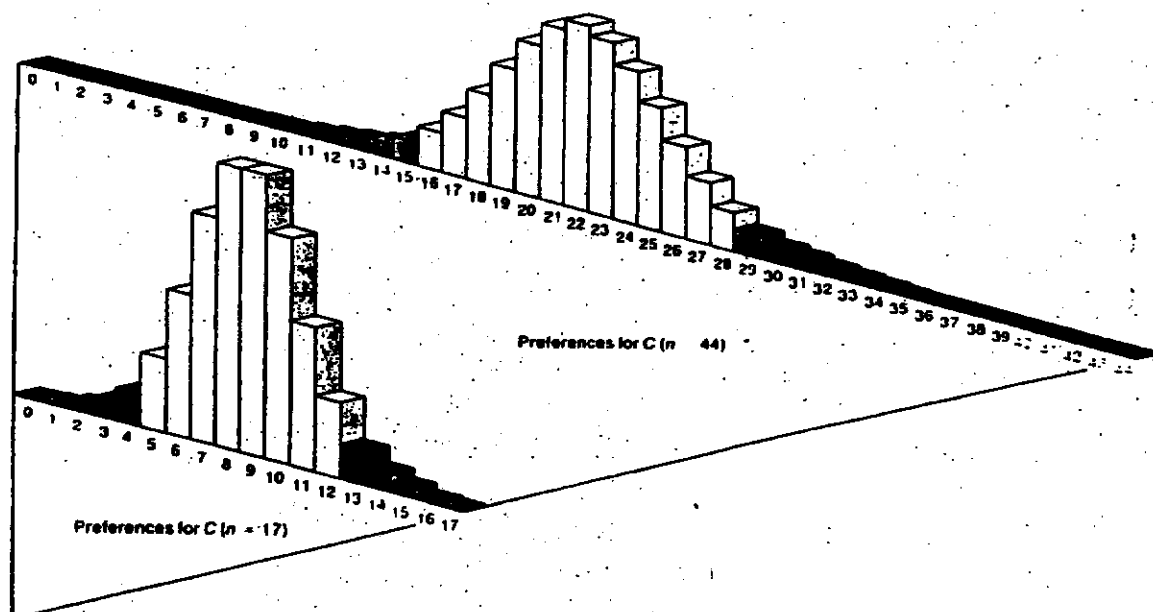


Figure 4. If the design of the vitamin C experiment is modified to allow the experiment to stop if conclusive evidence is present in the first 17 observations but otherwise to continue to 44 observations, Bayesian final probabilities would not be affected.  $P$ -values, by contrast, would reflect the fact that further testing was contemplated even if this additional testing was not carried out. Although the  $P$ -value for an experiment designed in advance to examine just 17 pairs or just 44 pairs is 0.049, the overall  $P$ -value for the two-stage design is 0.865.

# Bayesian reasoning in science

John Monson and Peter Urbach

Bayesian scientific reasoning has a sound foundation in logic and provides a unified approach to the evaluation of deterministic and statistical theories, unlike its main rivals.

Just as an uncertain world, though fortunately full of things that are not equally uncertain, is not equally uncertain to us, so we are not equally uncertain. We are fairly certain, for example, that the textbook laws of physics will remain valid for the foreseeable future, and very certain that they will remain in force through next week. We are much less certain about tomorrow's weather. We not only grade uncertainty — sometimes we measure it numerically, as in fact we do when we talk about the odds we think are merited by some predictive hypothesis or other. The more certain we are that an event will or will not take place, the longer (bigger) or shorter (smaller) the odds we are prepared to give.

Gambblers use the odds scale because odds are just the proportions in which the stakes are divided after the outcome. But it is not a very good scale on which to measure uncertainty at all. Because odds are ratios the odds scale starts at 0 and is unbounded to the right (infinite odds). An equally balanced uncertainty, corresponding to 1 on the odds scale, therefore cannot be represented as a midpoint. To symmetrize, we transform odds into the scale of probabilities, using the formula  $\text{probability } (p) = \text{odds} / (1 + \text{odds})$ , and put the probability corresponding to infinite odds equal to 1. So probabilities lie between 0 and 1 inclusive, and even-money odds now become the midpoint of the probability scale, as desired.

## Axioms

The formal theory of probability was born in the late seventeenth century in the work of Pierre Pascal, Huygens and James Bernoulli. It is summarized today in four basic laws, or axioms. The first says that the probability of any hypothesis  $h$  is a non-negative real number:  $P(h) \geq 0$ . The second says that the probability of any necessary truth  $t$  is 1:  $P(t) = 1$  (a necessary truth is one that is true whatever the world might be like; 'if it is raining then it is raining' is an example). The third, the additivity principle, says that if  $h$  and  $h'$  are mutually exclusive then the sum of their probabilities equals the probability of their disjunction: in symbols,  $P(h) + P(h') = P(h \text{ or } h')$ . The fourth, says that the conditional probability  $P(h|h')$  of  $h$  given  $h'$ , is equal to the unconditional probability  $P(h \& h')$  of the conjunction  $h$  and  $h'$ , divided by the unconditional probability  $P(h')$  of  $h'$  where that probability is positive: in symbols,  $P(h|h') = P(h \& h') / P(h')$ , where  $P(h') > 0$ .

Mathematical probability theory began life as theory of uncertainty. But in the late nineteenth century the probability axioms

became recognized also as the laws of objectively random phenomena, or objective chance. Chance turns out to play an essential role in modern science, in the theory of statistical sampling, information theory, demography, genetics, thermodynamics and quantum theory.

Our concern here is with the older idea of probability as the foundation of a theory of uncertainty. We shall show how some modern developments endorse this idea, and how they enable us to see the rules of the probability calculus as a logic of inductive inference. Suppose  $h$  is some scientific hypothesis. Experimental data can never conclusively prove that  $h$  is true, even if it is true. So you are never absolutely certain of  $h$ 's truth, only more or less. The inductive inference consists in assessing the degree of certainty warranted by the evidence. To the heirs of Bernoulli and Laplace, this means measuring the probability of  $h$  relative to data  $e$ . Scientists often estimate the probabilities of theories, but attempts to provide an objective basis for these estimates have uniformly failed.

But if the probability of a hypothesis merely reflects our own personal degree of belief in  $h$ , how can an objective logic of inductive inference be based on such probabilities? There is no paradox here. Your degrees of belief may be personal to you, but it does not follow that they are necessarily unprincipled or anarchic — in the first place, they must satisfy the axioms of probability. Of the many arguments that have been advanced to demonstrate this, the simplest is due to Frank Ramsey and Bruno de Finetti, who discovered it independently in the 1920s and 30s.

Their result is often called the Dutch book theorem. Consider a contract whereby one party agrees to exchange with the other a sum  $pS$  for the chance of receiving  $S$  if  $h$  is true and nothing if  $h$  is false.  $S$  is a non-zero sum of money or some other divisible good, called the stake. The payoffs to the first party thus are  $S - pS$  if  $h$  is true and  $-pS$  if  $h$  is false (the payoffs to the other party are of course the same with the signs reversed). The contract is tantamount to a bet in which the first party is betting on  $h$  at odds  $pS : (S - pS)$ , that is, odds  $p : (1 - p)$ . It is easy to see that  $p$  is the quantity obtained by symmetrizing the odds through the transformation  $p = \text{odds} / (1 + \text{odds})$ , we introduced earlier. There we called  $p$  a probability, but so as not to prejudice matters we shall now call it the betting quotient associated with the odds.

Suppose  $p$  is such that you deem the odds

$p : (1 - p)$  fair, in the sense that to the best of your knowledge there is no advantage to taking either side of the bet. It is customary to identify this value of  $p$  with your degree of belief in  $h$ . Now consider some arbitrary finite set of hypotheses  $h_i$ , where  $p_i$  are your corresponding fair betting quotients on the  $h_i$ . A betting strategy with respect to the  $h_i$  is a set of decisions of the form 'bet on (against)  $h_i$ ', for each  $i$ . Ramsey and de Finetti showed that if the  $p_i$  do not satisfy the probability axioms, then there are stakes  $S_i$  and a betting strategy for the  $h_i$  which must result in a certain loss for whoever follows that strategy (such a set of stakes is known as a Dutch book). Hence you cannot consistently maintain that the  $p_i$  are all fair if they do not satisfy the probability axioms.

## Proof

The proof of the Ramsey-de Finetti result uses no more than high-school algebra. Consider axiom 2 of the probability calculus, that  $P(t) = 1$  if  $t$  is a necessary truth. Suppose that  $p = P(t)$  is greater than 1. Because  $t$  is necessarily true the bettor on  $t$  will make a guaranteed loss of  $S - pS$ . If  $p$  is less than 1 then the bettor against  $t$  will make a guaranteed loss of  $S - pS$ . Either way one party or the other is guaranteed to lose, and so no value for  $p$  other than 1 can be fair. It is equally simple to see why  $P(h)$  must be non-negative, where  $h$  is any hypothesis, and only slightly less straightforward to see how to justify the remaining two axioms of the probability calculus.

An immediate consequence of the probability axioms is Bayes's theorem, which has given its name to this approach. Thomas Bayes (1702–1761) was an English Non-conformist clergyman, a gifted mathematician, and a fellow of the Royal Society. His seminal work on probability is contained in one short memoir published posthumously in 1763.

Bayes's theorem says that, for any propositions  $h$  and  $e$

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)} \quad (1)$$

In the usual applications of the theorem,  $h$  is some hypothesis and  $e$  the evidence against which it is to be evaluated.  $P(h|e)$  is the posterior probability of  $h$  on  $e$ ,  $P(h)$  is the prior probability of  $h$ , and  $P(e|h)$  is the likelihood of  $h$  on  $e$ . Equation 1 can be rewritten as:

$$P(h|e) \propto P(e|h)P(h)$$

That is, posterior probability is proportional

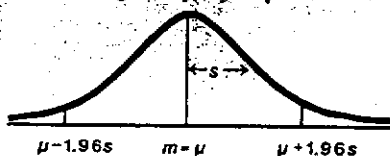
learned that if you performed significance tests repeatedly and if each time the result was significant at the 5 per cent level you acted as if believing the null hypothesis was false, you would "in the long run" be wrong on only "around 5 per cent" of the occasions.

This argument, although superficially plausible, is mistaken in every respect. First, it is simply absurd to act as if you believed a hypothesis false, when you are uncertain that it really is false. When you are uncertain that you would willingly accept a wager in which you received \$1.00 if the hypothesis was finally established to be false, and you had to forego all your worldly goods if it was finally established to be true. No reasonable person would accept such a bet, without being absolutely certain that the hypothesis in question really was false. Second, the justification is fallacious. It starts from the fact that the probability of rejecting a true null hypothesis equals the significance level; from this, it infers that if a test were repeatedly performed using a 0.05 significance level, the approximate frequency of rejecting a true null hypothesis would be 5 per cent, in the long run. But the inference is simply a nonsequitur: you cannot derive from the probability of an event occurring the approximate frequency with which the event will appear in any actual run of the experiment. Finally, the justification is irrelevant, because it refers to the supposed frequency of drawing a wrong conclusion in a sequence of possibly imaginary tests, but says nothing about the particular case at hand.

Scientists often need to know the value of physical parameter that cannot be measured directly. The task of gauging the mean height of a very large population is a simple example. In such cases, an indirect measurement must be made and this is standardly done by examining a suitably selected sample.

### Techniques

Classical statisticians have devised techniques that supposedly permit objective estimations of parameters, the principal one being that of the confidence interval, which we can explain through the example already mentioned. Let the unknown mean population height be  $\mu$ . Assume that we know the standard deviation,  $\sigma$ , of heights in the population. Now suppose a random sample of size  $n$  is drawn from the population. The mean height of people in such a sample is a random variable,  $m$ . Clearly  $m$  can take many possible values, some more probable than others. The distribution representing this situation is 'normal' and its standard deviation is given by  $\sigma/\sqrt{n}$ .



sample means against probability densities, not probabilities. The important fact for the present discussion is that the probability that  $m$  lies between two points is proportional to the area enclosed by them and the curve. Because the distribution is normal, it follows that with probability 0.95,  $-1.96\sigma \leq m - \mu \leq 1.96\sigma$ . Rearranging these inequalities gives the result that with probability 0.95,  $m - 1.96\sigma \leq \mu \leq m + 1.96\sigma$ . Suppose  $m'$  is the value of  $m$  that is actually observed in the experimental sample. Then, because we know  $\sigma$  and  $n$ , the terms  $m' - 1.96\sigma$  and  $m' + 1.96\sigma$  can be determined; the interval between them is called a 95 per cent confidence interval for  $\mu$ , and classical statisticians regard such an interval as a reasonable estimate of  $\mu$ .

The statement that such-and-such is a 95 per cent confidence interval for  $\mu$  seems objective. But what does it say? It might be imagined that a 95 per cent confidence interval corresponds to a 0.95 probability that the unknown parameter lies in the confidence range. But in the classical approach,  $\mu$  is not a random variable, and so has no probability. Nevertheless, statisticians regularly say that one can be '95 per cent confident' that the parameter lies in the confidence interval. They never say why.

In fact, there is a decisive reason why not. The confidence interval is derived from the probability distribution of sample means depicted above. This distribution gives the probabilities of all the sample means that you might have got in the experiment. It is usual to assume that all those possible samples have the same size as the actual sample. This assumption is crucial, because the shape of the distribution, and hence the width of the confidence interval, is affected by that size. Now the set of possible samples is determined by the experimenter's intention. If he had deliberately set out to sample exactly  $n$  people, the usual assumption would be justified. But suppose each time he selected a person from the population, he also tossed a fair coin; and that he planned to stop sampling as soon as the coin had produced 5 heads; or suppose the plan was simply to examine as many people as possible before lunch, or before getting bored. With any of these plans, the experimenter might still have arrived at a sample of  $n$ , but the set of possible samples would have been different, and hence, so would the confidence interval. So the degree of confidence we are invited to place in an estimate inevitably depends on the private plans of the experimenter, which is surely immensely counterintuitive.

This is the so-called stopping-rule problem. It also affects significance tests. In our earlier example, it was assumed, as it normally would be, that because the coin was tossed 20 times, all the of possible outcomes would exhibit 20 heads and/or tails. But these are the possible outcomes only if the experimenter had a premeditated plan to throw the coin 20 times. Had the plan been to

peared, he could have got just the result he did, but with a different list of unrealized, possible outcomes. Because significance is calculated by reference to these possibilities, a result could be significant if the experimenter had had one plan (or stopping rule) in mind, but not significant if it was another.

This dependence of significance tests and confidence interval estimates on the subjective, possibly unconscious intentions of the experimenter is an astonishing thing to discover at the heart of supposedly objective methodologies. It is also a most inappropriate thing to find in any methodology, for the plausibility, or cognitive value, of a hypothesis, and our rational confidence in an estimate should not depend on the contents of the experimenter's mind.

### Illusion

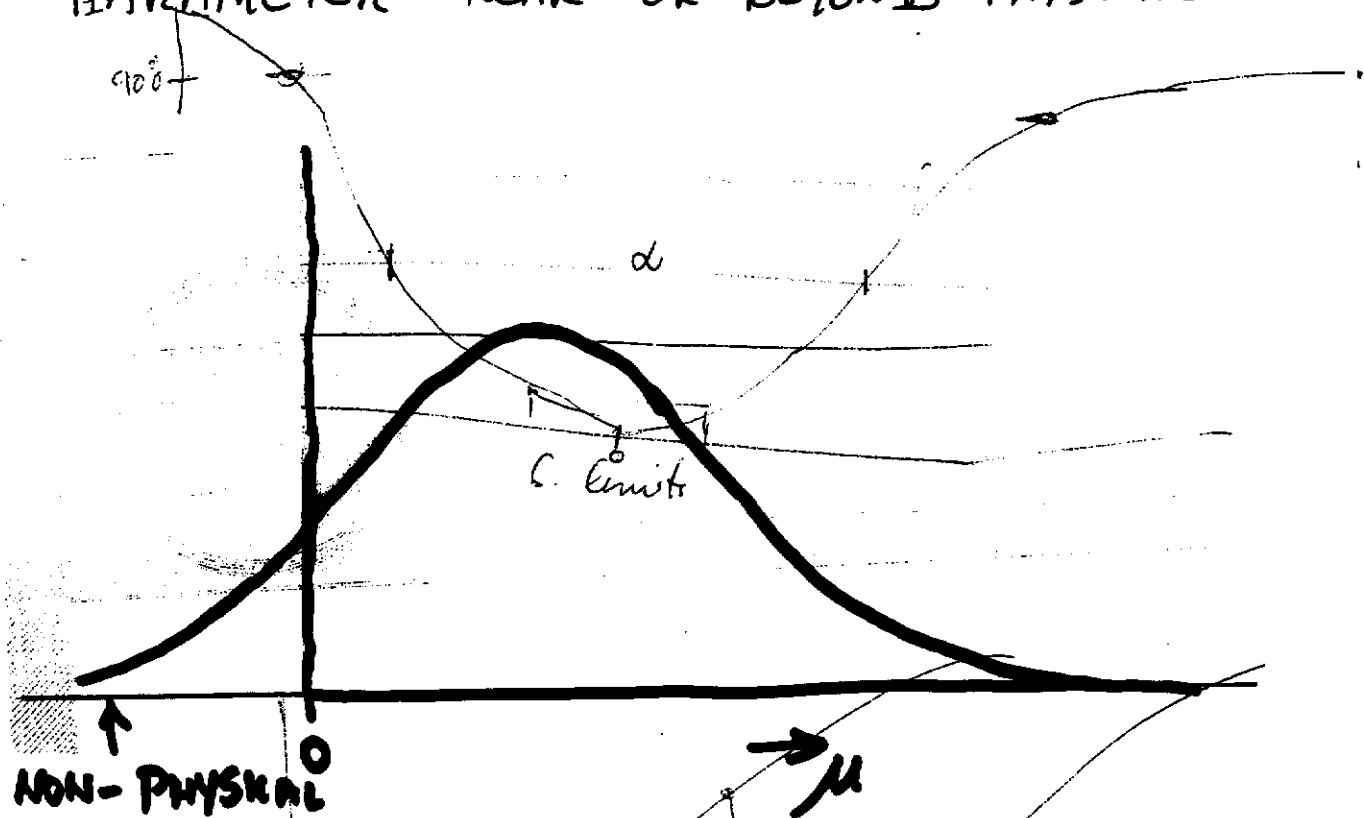
Popper's corroboration idea, and the theories of significance tests and confidence intervals were developed as supposedly objective methodologies in conscious reaction to subjective bayesianism. These methods all issue in apparently objective statements, couched in a deceptive terminology which gives the impression that an important, objective, theoretical evaluation is being achieved. But this is illusory. Corroborating a hypothesis does not strengthen it, a significant result has no significance for the truth of the null hypothesis, and a 95 per cent confidence interval has no right to impart confidence, let alone 95 per cent's worth, to an estimate.

Unlike these pseudo-objective methodologies, the bayesian approach has a solid foundation. It provides a unified approach to deterministic and statistical theories, and to questions of testing and estimation, unlike the many *ad hoc* recipes of the classical approach. It is also intuitively right. We can illustrate this by sketching the bayesian way of estimating a population mean. It starts with a distribution of subjective prior probabilities over the range of possible values of the parameter. Then, using Bayes's theorem and the sample evidence, a corresponding posterior probability is calculated. The prior probability curve typically would be very spread out, indicating considerable initial uncertainty about the parameter value, whereas the posterior probability would be concentrated in a narrow region. Then if 95 per cent of the area under the curve was enclosed between two points  $a$  and  $b$ , the bayesian estimate of the parameter would be of the form ' $\mu$  lies between  $a$  and  $b$  with probability 0.95'.

This bayesian conclusion has a clear meaning and is just the kind of conclusion people do come to. It is derived from the mean of the experimental sample alone, not the means of possible samples. Hence, it is unaffected by the experimenter's subjectively intended stopping rule, which is as it should be. Finally, the posterior distribution is very insensitive to variations in the prior



# AN AD HOC METHOD FOR ESTIMATING PARAMETER NEAR OR BEYOND PHYSICAL LIMIT



- "TRUNCATE AND RENORMALIZE"
- REASONABLE PROPERTIES, OFTEN USED
- DESCRIBED IN R.P.P. < 1998,  
SOMETIMES CALLED "THE PDG METHOD"
- IN FACT, IT IS BAYESIAN WITH  
UNIFORM PRIOR  
(THE AUTHOR DIDN'T KNOW THIS)

# SUMMARY

FREQUENTIST METHODS STILL OFFER THE ONLY WAY TO PRESENT EXPERIMENTAL RESULTS OBJECTIVELY WITH THE USUAL SCIENTIFIC MEANING.

## BUT

- BAYESIAN METHODS ARE GOOD FOR DECISION MAKING.

DO PHYSICISTS MAKE DECISIONS?

- BAYESIAN METHODS ARE GOOD FOR BETTING

DO PHYSICISTS MAKE BETS?

- BAYESIAN METHODS ARE GOOD WHEN THERE IS A PRIOR PROBABILITY OR PHASE SPACE

MAXIMUM ENTROPY METHOD

- BAYESIAN METHODS ARE A GOOD WAY TO COMBINE NEW KNOWLEDGE WITH PRIOR BELIEFS.

DO WE DO THIS?

