

BCB 731:
*Defense Against
the Dark Arts*



Everything About
Data

October 4th, 2023



Let's talk about everything

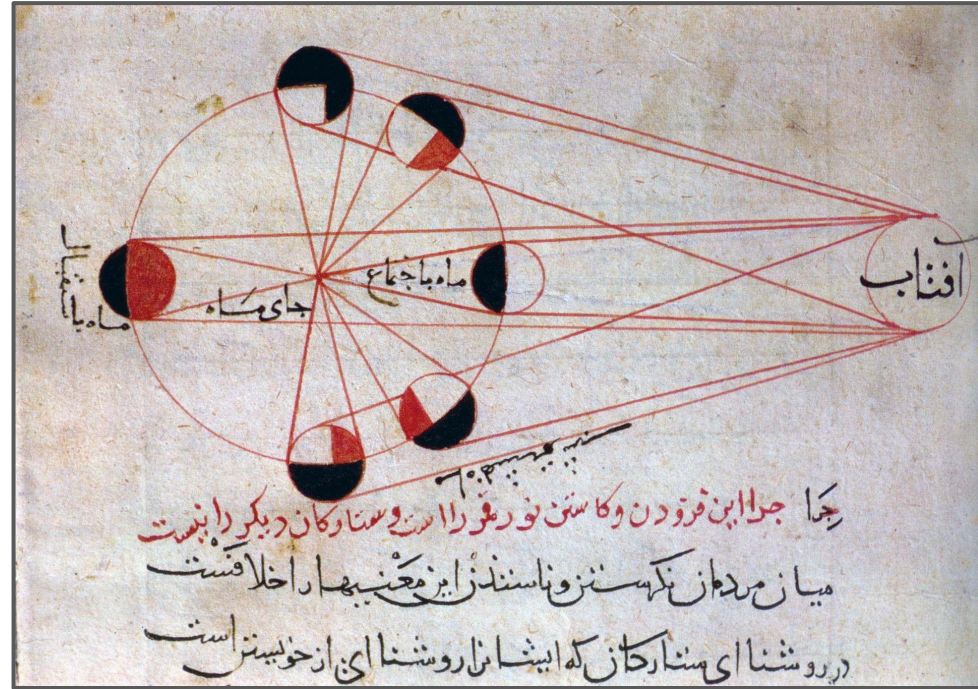
- Classical Statistics
- Exploratory Data Analysis
- Bayesian Statistics
- Classical Machine Learning
- Deep Learning
- Artificial Intelligence

...and more!

- What is knowledge?
- How do we come to know something about the world?
- What role does data play in knowledge formation?
- Do we even need statistics?

Before the “scientific revolution”

- Interplay between **observation** and strong **inductive** priors
- Abstract models more often descriptive than mathematical
- Validated by agreement with observation but also:
 - Elegance / aesthetics
 - Great “masters”
 - Theology & philosophy



Abu-Rayhan al-Biruni's *Al-Tafhim li Awa'il Sana'at al-Tanjim* (Book on the Elements of Astrology)

Modern science = empiricism

- 1600s science narrowed “natural philosophy” to:
 - Collect data
 - Build mathematical models (repeat)

“the Universe – which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics” – Galileo in *The Assayer*



Galileo Gallilei's “The Assayer”

Observations of Jupiter's Moons

20. Jan. 1610	0 **
30. Jan.	** 0 *
2. Feb.	0 ** *
3. Feb.	0 * *
3. Feb. 5.	* 0 *
4. Feb.	* 0 **
6. Feb.	** 0 *
8. Feb. H. 13.	* * * 0
10. Feb.	* * * 0 *
11.	* * 0 *

Galileo's notebook observing moons of Jupiter (interpreted as evidence for refuting geocentrism)

Statistics = seeing like a state

The Table of CASUALTIES.

The Year of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1661	1662	1663	1664	1665	1666	1667	1668	1669	1670	1671	1672	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682	1683	1684	1685	1686	1687	1688	1689	1690	1691	1692	1693	1694	1695	1696	1697	1698	1699	1700	In 20 Years																																																																																																																																																																																																			
Deaths	1315	127	134	159	184	231	278	319	359	400	440	480	520	560	600	640	680	720	760	800	840	880	920	960	1000	1040	1080	1120	1160	1200	1240	1280	1320	1360	1400	1440	1480	1520	1560	1600	1640	1680	1720	1760	1800	1840	1880	1920	1960	2000	2040	2080	2120	2160	2200	2240	2280	2320	2360	2400	2440	2480	2520	2560	2600	2640	2680	2720	2760	2800	2840	2880	2920	2960	3000	3040	3080	3120	3160	3200	3240	3280	3320	3360	3400	3440	3480	3520	3560	3600	3640	3680	3720	3760	3800	3840	3880	3920	3960	4000	4040	4080	4120	4160	4200	4240	4280	4320	4360	4400	4440	4480	4520	4560	4600	4640	4680	4720	4760	4800	4840	4880	4920	4960	5000	5040	5080	5120	5160	5200	5240	5280	5320	5360	5400	5440	5480	5520	5560	5600	5640	5680	5720	5760	5800	5840	5880	5920	5960	6000	6040	6080	6120	6160	6200	6240	6280	6320	6360	6400	6440	6480	6520	6560	6600	6640	6680	6720	6760	6800	6840	6880	6920	6960	7000	7040	7080	7120	7160	7200	7240	7280	7320	7360	7400	7440	7480	7520	7560	7600	7640	7680	7720	7760	7800	7840	7880	7920	7960	8000	8040	8080	8120	8160	8200	8240	8280	8320	8360	8400	8440	8480	8520	8560	8600	8640	8680	8720	8760	8800	8840	8880	8920	8960	9000	9040	9080	9120	9160	9200	9240	9280	9320	9360	9400	9440	9480	9520	9560	9600	9640	9680	9720	9760	9800	9840	9880	9920	9960	10000
Deaths by Smallpox	1315	127	134	159	184	231	278	319	359	400	440	480	520	560	600	640	680	720	760	800	840	880	920	960	1000	1040	1080	1120	1160	1200	1240	1280	1320	1360	1400	1440	1480	1520	1560	1600	1640	1680	1720	1760	1800	1840	1880	1920	1960	2000	2040	2080	2120	2160	2200	2240	2280	2320	2360	2400	2440	2480	2520	2560	2600	2640	2680	2720	2760	2800	2840	2880	2920	2960	3000	3040	3080	3120	3160	3200	3240	3280	3320	3360	3400	3440	3480	3520	3560	3600	3640	3680	3720	3760	3800	3840	3880	3920	3960	4000	4040	4080	4120	4160	4200	4240	4280	4320	4360	4400	4440	4480	4520	4560	4600	4640	4680	4720	4760	4800	4840	4880	4920	4960	5000	5040	5080	5120	5160	5200	5240	5280	5320	5360	5400	5440	5480	5520	5560	5600	5640	5680	5720	5760	5800	5840	5880	5920	5960	6000	6040	6080	6120	6160	6200	6240	6280	6320	6360	6400	6440	6480	6520	6560	6600	6640	6680	6720	6760	6800	6840	6880	6920	6960	7000	7040	7080	7120	7160	7200	7240	7280	7320	7360	7400	7440	7480	7520	7560	7600	7640	7680	7720	7760	7800	7840	7880	7920	7960	8000	8040	8080	8120	8160	8200	8240	8280	8320	8360	8400	8440	8480	8520	8560	8600	8640	8680	8720	8760	8800	8840	8880	8920	8960	9000	9040	9080	9120	9160	9200	9240	9280	9320	9360	9400	9440	9480	9520	9560	9600	9640	9680	9720	9760	9800	9840	9880	9920	9960	10000
Deaths by Smallpox	1315	127	134	159	184	231	278	319	359	400	440	480	520	560	600	640	680	720	760	800	840	880	920	960	1000	1040	1080	1120	1160	1200	1240	1280	1320	1360	1400	1440	1480	1520	1560	1600	1640	1680	1720	1760	1800	1840	1880	1920	1960	2000	2040	2080	2120	2160	2200	2240	2280	2320	2360	2400	2440	2480	2520	2560	2600	2640	2680	2720	2760	2800	2840	2880	2920	2960	3000	3040	3080	3120	3160	3200	3240	3280	3320	3360	3400	3440	3480	3520	3560	3600	3640	3680	3720	3760	3800	3840	3880	3920	3960	4000	4040	4080	4120	4160	4200	4240	4280	4320	4360	4400	4440	4480	4520	4560	4600	4640	4680	4720	4760	4800	4840	4880	4920	4960	5000	5040	5080	5120	5160	5200	5240	5280	5320	5360	5400	5440	5480	5520	5560	5600	5640	5680	5720	5760	5800	5840	5880	5920	5960	6000	6040	6080	6120	6160	6200	6240	6280	6320	6360	6400	6440	6480	6520	6560	6600	6640	6680	6720	6760	6800	6840	6880	6920	6960	7000	7040	7080	7120	7160	7200	7240	7280	7320	7360	7400	7440	7480	7520	7560	7600	7640	7680	7720	7760	7800	7840	7880	7920	7960	8000	8040	8080	8120	8160	8200	8240	8280	8320	8360	8400	8440	8480	8520	8560	8600	8640	8680	8720	8760	8800	8840	8880	8920	8960	9000	9040	9080	9120	9160	9200	9240	9280	9320	9360	9400	9440	9480	9520	9560	9600	9640	9680	9720	9760	9800	9840	9880	9920	9960	10000
Deaths by Smallpox	1315	127	134	159	184	231	278	319	359	400	440	480	520	560	600	640	680	720	760	800	840	880	920	960	1000	1040	1080	1120	1160	1200	1240	1280	1320	1360	1400	1440	1480	1520	1560	1600	1640	1680	1720	1760	1800	1840	1880	1920	1960	2000	2040	2080	2120	2160	2200	2240	2280	2320	2360	2400	2440	2480	2520	2560	2600	2640	2680	2720	2760	2800	2840	2880	2920	2960	3000	3040	3080	3120	3160	3200	3240	3280	3320	3360	3400	3440	3480	3520	3560	3600	3640	3680	3720	3760	3800	3840	3880	3920	3960	4000	4040	4080	4120	4160	4200	4240	4280	4320	4360	4400	4440	4480	4520	4560	4600	4640	4680	4720	4760	4800	4840	4880	4920	4960	5000	5040	5080	5120	5160	5200	5240	5280	5320	5360	5400	5440	5480	5520	5560	5600	5640	5680	5720	5760	5800	5840	5880	5920	5960	6000	6040	6080	6120	6160	6200	6240	6280	6320	6360	6400	6440	6480	6520	6560	6600	6640	6680	6720	6760	6800	6840	6880	6920	6960	7000	7040	7080	7120	7160	7200	7240	7280	7320	7360	7400	7440	7480	7520	7560	7600	7640	7680	7720	7760	7800	7840	7880	7920	7960	8000	8040	8080	8120	8160	8200	8240	8280	8320	8360	8400	8440	8480	8520	8560	8600	8640	8680	8720	8760	8800	8840	8880	8920	8960	9000	9040	9080	9120	9160	9200	9240	9280	9320	9360	9400	9440	9480	9520	9560	9600	9640	9680	9720	9760	9800	9840	9880	9920	9960	10000
Deaths by Smallpox	1315	127	134	159	184	231	278	319	359	400	440	480	520	560	600	640	680	720	760	800	840	880	920	960	1000	1040	1080	1120	1160	1200	1240	1280	1320	1360	1400	1440	1480	1520	1560	1600	1640	1680	1720	1760	1800	1840	1880	1920	1960	2000	2040	2080	2120	2160	2200	2240	2280	2320	2360	2400	2440	2480	2520	2560	2600	2640	2680	2720	2760	2800	2840	2880	2920	2960	3000	3040	3080	3120	3160	3200	3240	3280	3320	3360	3400	3440	3480	3520	3560	3600	3640	3680	3720	3760	3800	3840	3880	3920	3960	4000	4040	4080	4120	4160	4200	4240	4280	4320	4360	4400	4440	4480	4520	4560	4600	4640	4680	4720	4760	4800	4840	4880	4920	4960	5000	5040	5080	5120	5160	5200	5240	5280	5320	5360	5400	5440	5480	5520	5560	5600	5640	5680	5720	5760	5800	5840	5880	5920	5960	6000	6040	6080	6120	6160	6200	6240	6280	6320	6360	6400	6440	6480	6520	6560	6600	6640	6680	6720	6760	6800	6840	6880	6920	6960	7000	7040	7080	7120	7160	7200	7240	7280	7320	7360	7400	7440	7480	7520	7560	7600	7640	7680	7720	7760	7800	7840	7880	7920	7960	8000	8040	8080	8120	8160	8200	8240	8280	8320	8360	8400	8440	8480	8520	8560	8600	8640	8680	8720	8760	8800	8840	8880	8920	8960	9000	9040	9080	9120	9160	9200	9240	9280	9320	9360	9400	9440	9480	9520	9560	9600	9640	9680	9720	9760	9800	9840	9880	9920	9960	10000
Deaths by Smallpox	1315	127	134	159	184	231	278	319	359	400	440	480	520	560	600	640	680	720	760	800	840	880	920	960	1000	1040	1080	1120	1160	1200	1240	1280	1320	1360	1400	1440	1480	1520	1560	1600	1640	1680	1720	1760	1800	1840	1880	1920	1960	2000	2040	2080	2120	2160	2200	2240	2280	2320	2360	2400	2440	2480	2520	2560	2600	2640	268																																																																																																																																																																																							

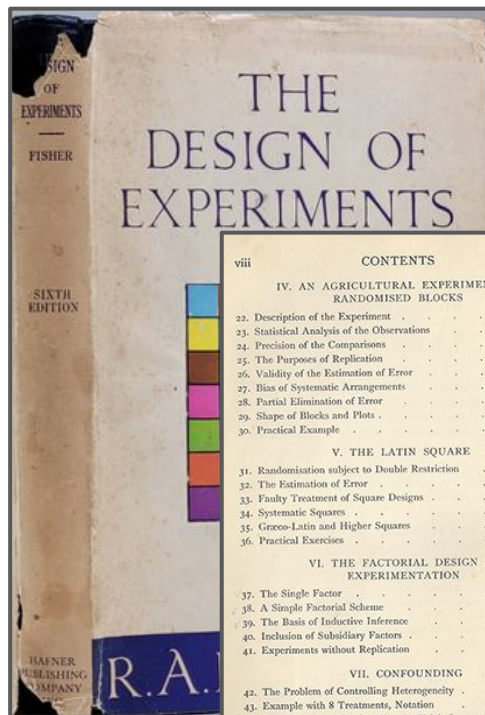
Political Arithmetick, OR, A DISCOURSE Concerning, The Extent and Value of Lands, People, Buildings; Husbandry, Manufacture, Commerce, Fishery, Artizans, Seamen, Soldiers; Publick Revenues, Interest, Taxes, Superlucration, Registries, Banks; Valuation of Men, Increasing of Seamen, of Militia's, Harbours, Situation, Shipping, Power at Sea, &c. As the same relates to every Country in general, but more particularly to the Territories of His Majesty of Great Britain, and his Neighbours of Holland, Zealand, and France.

By Sir WILLIAM PETTY, Late Fellow of the Royal Society.

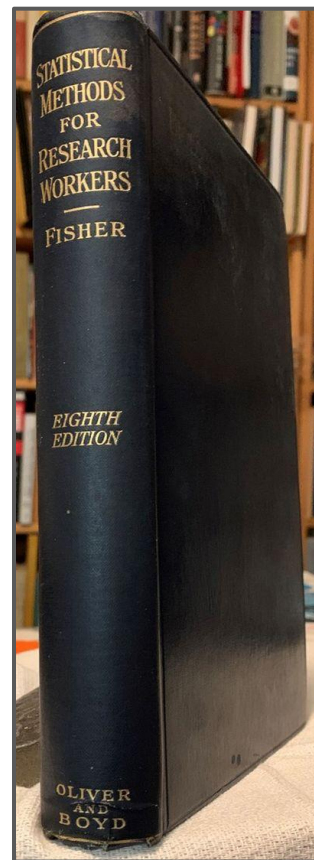
London, Printed for Robert Clavel at the Peacock, and Hen. Morlock at the Phoenix in St. Paul's Church-yard. 1691.

I n Angliantibus Tenementis Terras In Domesday	
R ex Willielmus.	xxvii. Ricard filius celestini
11. E po de recepre.	xxviii. ogerius de balli.
12. E po de confiantis.	xxix. l. ogerius de albamarle.
13. E cda clafpingborie.	xxx. l. abaild bayard.
14. E cda l. luefborie.	xxxi. l. canbul fili' braf.
15. E cda de bouchpeth.	xxxii. l. adulfus de lmeti.
16. E cda de horiane.	xxxiii. l. adulfus pangen.
17. E cda de creneburne.	xxxiiii. l. adulfus de fclgbor.
18. E cda de lubazale.	xxxv. l. adulfus de pomera.
19. E cda de hammapowen.	xxxvi. l. wald adobed.
20. E cda de mome & Mchad.	xxxvii. l. canbul filius berneni.
21. E cda de frefam de cadom.	xxxviii. l. canbul filius kulf.
22. E cda de kufm de cadom.	xxxix. l. leudius de hpana.
23. Comet hupo.	xl. l. luyedus bmo.
24. Comet mortemontis.	xli. l. niferius.
25. l. abaild uiccomen.	xlii. l. rafal.
26. l. uheld de forcanes.	xliii. o. fo filius camelen.
27. l. willam de mon.	xliiii. o. fmo de falcad.
28. l. willam clare.	xlv. v. x. horie de hron.
29. l. willam de palete.	xlvi. e. mof capellan.
30. l. willam de polte.	xlvii. e. mof filius fclgbor.
31. l. willam de oio.	xlvi. e. n. fclgbor.
32. l. walterus de clauda.	l. f. niferius.
33. l. walterus de clauda.	li. n. americus.
34. l. walterus.	lii. willyd fili' fclgbor.
35. l. walterus.	liii. Coluin fili' canbul.

1920s/30s: statistics infiltrates science



CONTENTS		CONTENTS	
viii		ix	
IV. AN AGRICULTURAL EXPERIMENT IN RANDOMISED BLOCKS		50. Interaction of Quantity and Quality 140	
22. Description of the Experiment	55	51. Resolution of Three Comparisons among Four Materials	142
23. Statistical Analysis of the Observations	57	52. An Early Example	143
24. Precision of the Comparisons	64	53. Interpretation of Results	154
25. The Purposes of Replication	66	54. An Experiment with 81 Plots	157
26. Validity of the Estimation of Error	68	IX. THE INCREASE OF PRECISION BY CONCOMITANT MEASUREMENTS. STATISTICAL CONTROL.	
27. Bias of Systematic Arrangements	71	55. Occasions suitable for Concomitant Measurements	167
28. Partial Elimination of Error	72	56. Arbitrary Corrections	173
29. Shape of Blocks and Plots	73	57. Calculation of the Adjustment	176
30. Practical Example	75	58. The Test of Significance	181
V. THE LATIN SQUARE		59. Practical Example	184
31. Randomisation subject to Double Restriction	78	X. THE GENERALISATION OF NULL HYPOTHESES. FIDUCIAL PROBABILITY	
32. The Estimation of Error	81	60. Precision regarded as Amount of Information	187
33. Faulty Treatment of Square Designs	83	61. Multiplicity of Tests of the same Hypothesis	190
34. Systematic Squares	85	62. Extension of the t Test	195
35. Greco-Latin and Higher Squares	90	63. The χ^2 Test	198
36. Practical Exercises	93	64. Wider Tests based on the Analysis of Variance	201
VI. THE FACTORIAL DESIGN IN EXPERIMENTATION		65. Comparisons with Interactions	211
37. The Single Factor	96	XI. THE MEASUREMENT OF AMOUNT OF INFORMATION IN GENERAL	
38. A Simple Factorial Scheme	98	66. Estimation in General	216
39. The Basis of Inductive Inference	106	67. Frequencies of Two Alternatives	218
40. Inclusion of Subsidiary Factors	107	68. Functional Relationships among Parameters	221
41. Experiments without Replication	111	69. The Frequency Ratio in Biological Assay	227
VII. CONFOUNDING		70. Linkage Values inferred from Frequency Ratios	230
42. The Problem of Controlling Heterogeneity	114	71. Linkage Values inferred from the Progeny of Self-fertilised or Intercrossed Heterozygotes	234
43. Example with 8 Treatments, Notation	117	72. Information as to Linkage derived from Human Families	240
44. Design suited to Confounding the Triple Interaction	119	73. The Information elicited by Different Methods of Estimation	244
45. Effect on Analysis of Variance	120	74. The Information lost in the Estimation of Error	247
46. Example with 27 Treatments	123	INDEX 251	
47. Partial Confounding	131		
VIII. SPECIAL CASES OF PARTIAL CONFOUNDING			
48. Dummy Comparisons	138		
49. Dummy Comparisons	138		

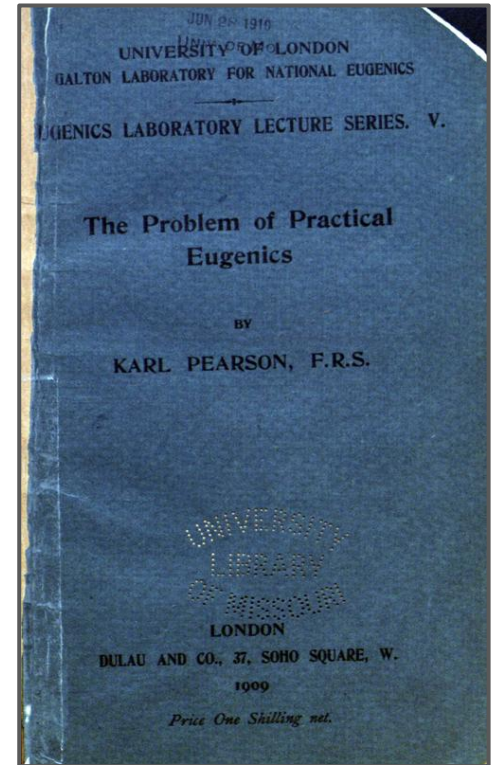


CONTENTS	
CHAP.	PAGE
EDITORS' PREFACE	v
AUTHOR'S PREFACE	vii
I. INTRODUCTORY	i
II. DIAGRAMS	27
III. DISTRIBUTIONS	43
IV. TESTS OF GOODNESS OF FIT, INDEPENDENCE AND HOMOGENEITY; WITH TABLE OF χ^2	77
V. TESTS OF SIGNIFICANCE OF MEANS, DIFFERENCES OF MEANS, AND REGRESSION COEFFICIENTS	101
VI. THE CORRELATION COEFFICIENT	138
VII. INTRAClass CORRELATIONS AND THE ANALYSIS OF VARIANCE	176
VIII. FURTHER APPLICATIONS OF THE ANALYSIS OF VARIANCE SOURCES USED FOR DATA AND METHODS	211
INDEX	233
	237
TABLES	
I. AND II. NORMAL DISTRIBUTION	} At End
III. TABLE OF χ^2	
IV. TABLE OF t	
V.A. CORRELATION COEFFICIENT—SIGNIFICANT VALUES	
V.B. CORRELATION COEFFICIENT—TRANSFORMED VALUES	
VI. TABLE OF z	
ix	

Side note: 2/4 “fathers” of statistics were very into eugenics

“...eugenics urges us to simplify our lives, and to simplify our needs; the only luxury worth having is that of a worthy human environment. We must be ready to sacrifice social success, at the call of nobler instincts.” -R. A. Fisher

“History shows me one way, and one way only, in which a high state of civilization has been produced, namely, the struggle of race with race, and the survival of the physically and mentally fitter race.” -Karl Pearson



So, what is classical statistics?

- There is some real world quantity:
 - ...how do we finite noisy measurements into a robust estimate of the “true” value?
- We have a mathematical model of reality:
 - ...how do rigorously we use finite noisy measurements to reject (or conditionally accept) the model?

Tools of classical statistics

- Experimental design
- Inference & estimators
 - consistent, unbiased, efficient
- Confidence intervals
- Statistical hypothesis testing (Neyman-Pearson)
- Null hypothesis models (Fisher)
- Their unholy marriage: NHST

Biological Question ?



Hypothesis H_0



Design Experiment



Collect Data

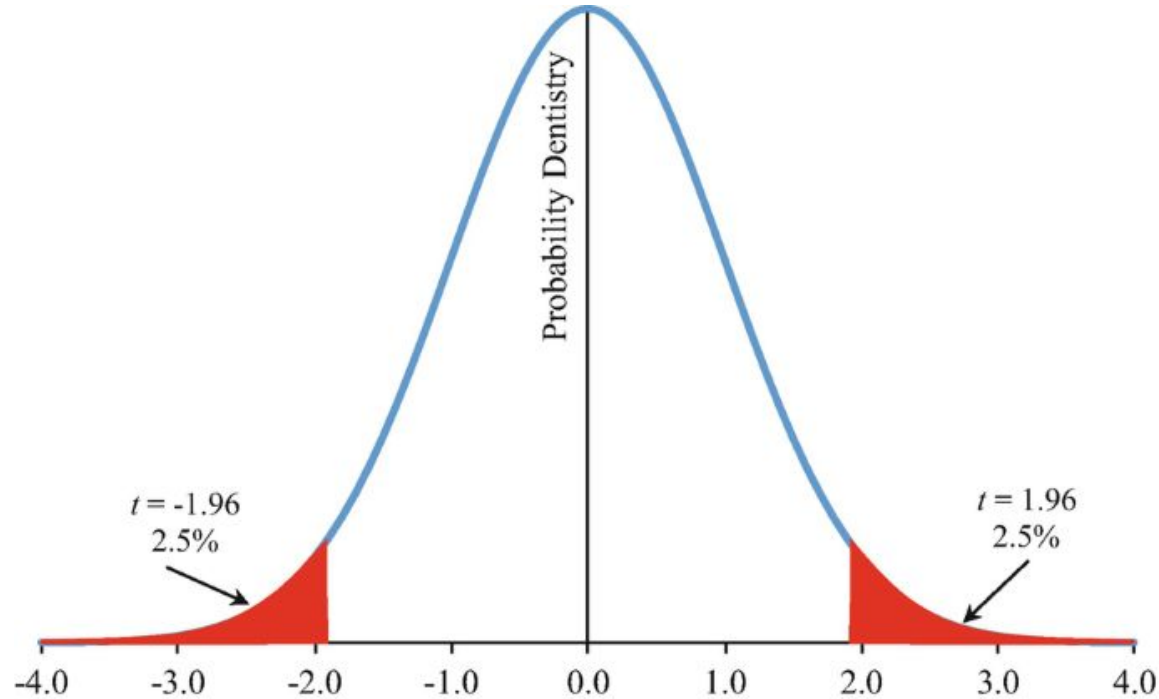


Compute p-value



Conclusion

Null Hypothesis Significance Testing



*Teaching Null Hypothesis Significance Testing (NHST) in the Health
Sciences: The Significance of Significance*

Exploratory Data Analysis: You're allowed to look at your data

Exploratory Data Analysis: Past, Present, and Future

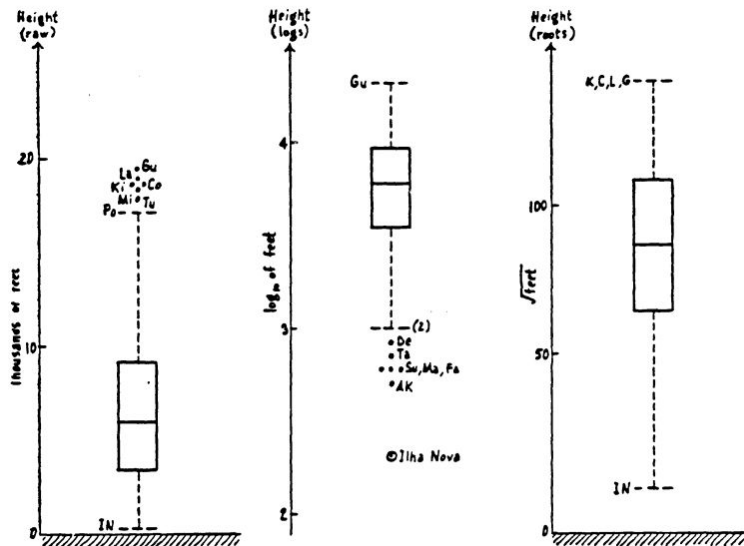
John W. Tukey¹

Technical Report No. 302

Princeton University, 408 Fine Hall, Washington Road, Princeton, NJ 08544-1000

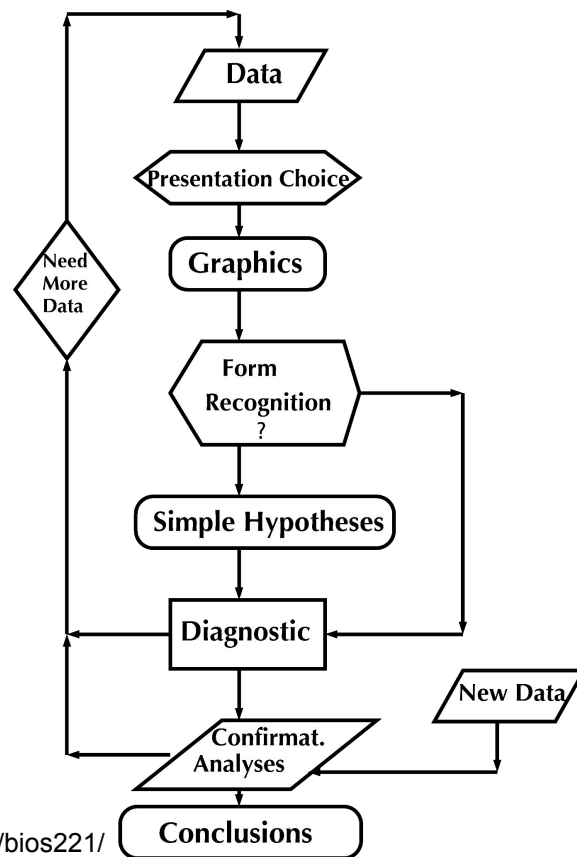
Abstract

The 1971-1977 early formulation of Exploratory Data Analysis, in terms of (a) results of some of its techniques and considerations which underlay, at various depths, the choices realized in the books. The 1991-1995 development of Exploratory Analysis of Variance, described in its simplest (two-way table) form and barely sketched in general. Discussion of the changes in apparent philosophy caused by the need to communicate more complicated things, notches, hints, the likely impact on a revised edition of Exploratory Data Analysis 1977. Dreams and targets for what might happen in 1996-2005, with emphasis on Exploratory Regression and the combined use of multiple description.



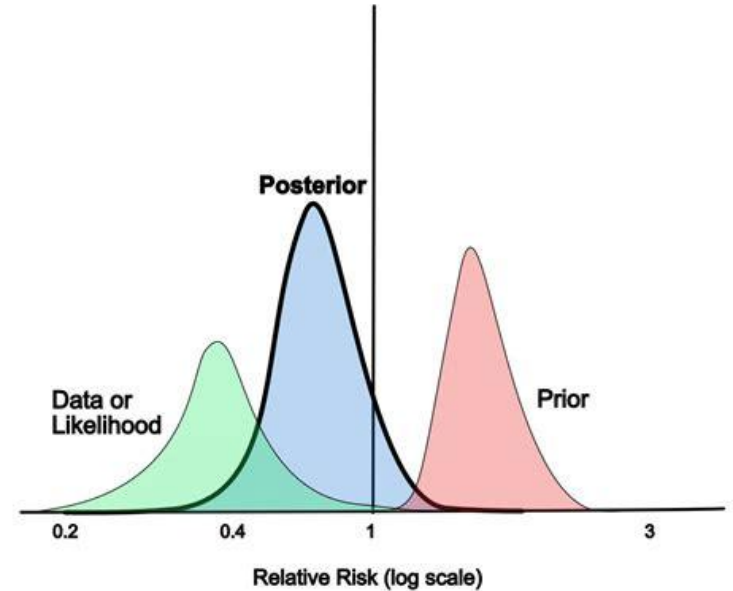
Exploratory Data Analysis: workflow

- Emphasis on visualization
- Look at residuals of your model
 - ...you can even try different models!
- Does it make sense?
- ...classical stats left for “confirmatory analyses”



Bayesian Statistics

- Why are we accepting or rejecting a hypothesis?
- Why do we think there's a single “true” value to parameters?
- Inference should give us a full probability distribution



Machine Learning

- Forget models of reality!
- ...let's just do function approximation instead.
- Data is some set of vectors x_i (can have labels y_i)
- Learn functions $f_w(x)$ parameterized by weights w
 - Clustering: $X \rightarrow \{1, \dots, k\}$
 - Regression / Classification: $X \rightarrow Y$

TO THE WHITEBOARD!

quick tour of ML / loss minimization / gradient descent / evaluation metrics / deep learning / AI

⬢ *Fin* ⬢