(Instytut Fizyki UJ, Kraków, 11 June 2019)

Evidence in quantum data

Berge Englert

with much help from many students and colleagues over many years

Centre for Quantum Technologies & Department of Physics
& MajuLab, National University of Singapore

Centre For Quantum Technologies

NUS National University of Singapore

MajuLab

Abstract

The data acquired in quantum experiments are unavoidably affected by statistical fluctuations and, therefore, the interpretation of the data must rely on methods from statistics. While p-values and confidence intervals are routinely reported, this practice is questionable. Bayesian concepts, instead, fit quite naturally to quantum data. This talk deals with various aspects of Bayesian methodology, in particular with the Bayesian notion of what constitutes evidence in favor of, or against, a hypothesis.

Quantum measurements

Reminding you of the obvious ($\equiv$ preaching to the converted):

- The data of quantum measurements are statistical in nature.
- The probabilities in quantum physics are always conditional probabilities; they are conditioned on what we know about the situation.
- The observed relative frequencies are not equal to the underlying probabilities; often the frequencies do not approximate the probabilities.
- Statistical methods are required for drawing inference from the data.

Frequentism vs Bayesian logic

Frequentism

- deals with repeated measurements;
- is appropriate when planning an experiment, when *all thinkable* data are considered — example: adaptive measurements;
- has no concept of evidence in favor of a case or against it.

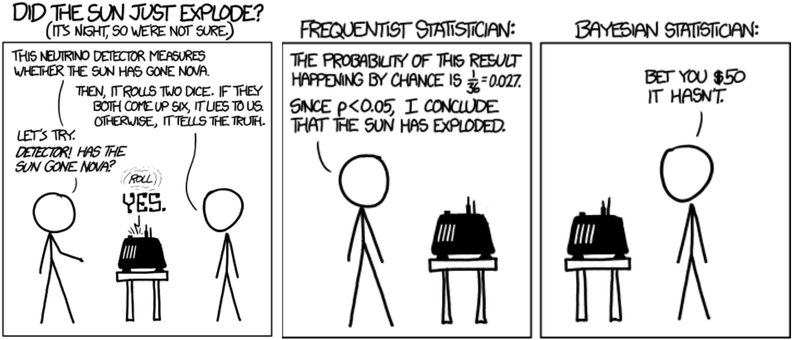
Bayesian logic

- is appropriate for *the* data actually observed;
- accounts systematically for prior knowledge;
- allows for proper marginalization of nuisance parameters;
- has a natural concept of evidence in favor or against.

Hypothesis testing: p-values (1)

P. Diaconis and B. Skyrms, *Ten Great Ideas About Chance*, Princeton UP 2018, opening sentences of the section on “Why most published research findings are false” on p. 115:

> *A cookbook frequentist hypothesis tester doesn’t have to think. He calculates a p-value. That is defined as the probability that random noise would generate a false positive.*



Source: https://xkcd.com/1132/

Hypothesis testing: p-values (2)

R. L. Wasserstein and N. A. Lazar,
The ASA's Statement on p-Values: Context, Process, and Purpose,
The American Statistician **70**, 129 (2016).

1– P-values can indicate how incompatible the data are with a specified statistical model.

2– P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3– Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4– Proper inference requires full reporting and transparency.

5– A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6– By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

See also D. J. Benjamin and 71 co-authors, *Redefine statistical significance*, Nature Human Behav. **2**, 6 (2017); V. Amrhein, S, Greenland, B. McShane, and more than 800 signatories, *Retire statistical significance*, Nature **567**, 308 (2019).

Four Experiments: Delft, Vienna, Boulder, Munich

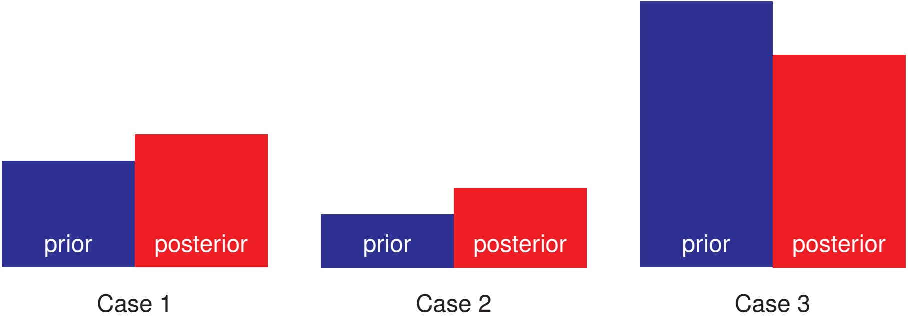
Delft ($p = 0.039$):
B. Hensen, H. Bernien, A. E. Dréau, and 16 co-authors,
Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometers, Nature (London) **526**, 682 (2015).

Vienna ($p = 3.74 \times 10^{-31}$):
M. Giustina, M. A. M. Versteegh, S. Wengerowsky, and 19 co-authors,
Significant-Loophole-Free Test of Bell's Theorem with Entangled Photons, Physical Review Letters **115**, 250401 (2015).

Boulder ($p = 2.3 \times 10^{-7}$):
L. K. Shalm, E. Meyer-Scott, B. G. Christensen, and 31 co-authors,
Strong Loophole-Free Test of Local Realism, Physical Review Letters **115**, 250402 (2015).

Munich ($p = 2.57 \times 10^{-9}$):
W. Rosenfeld, D. Burchardt, R. Garthoff, K. Redeker, N. Ortegel, M. Rau, and H. Weinfurter,
Event-Ready Bell Test Using Entangled Atoms Simultaneously Closing Detection and Locality Loopholes, Physical Review Letters **119**, 010402 (2017).

Evidence from a Bayesian perspective



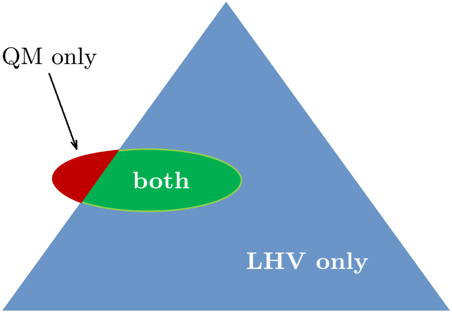
In this example, the data provide evidence in favor of cases 1 and 2 (posterior $>$ prior) and against case 3 (posterior $<$ prior).

– M. Evans, *Measuring Statistical Evidence Using Relative Belief*, Monographs on Statistics and Applied Probability, vol. 144 (CRC Press, Taylor & Francis Group, 2015)
– M. Evans and Y. Guo, *Measuring and Controlling Bias for Some Bayesian Inferences and the Relation to Frequentist Criteria*, eprint arXiv:1903.01696[math.ST]

Three regions in probability space

The probabilities of occurrence for the various detection events in the four experiments make up a high-dimensional probability space.

The probabilities permitted by QM constitute a convex set (ellipse), and those permitted by LHV constitute another (triangle), with an overlap region.



Central Question

Do the data provide evidence in favor of or against each of the three subsets?

Bayesian rules of engagement

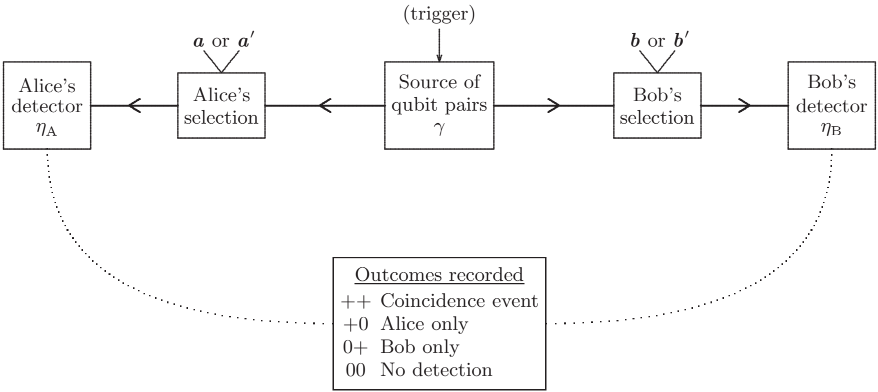
Ingredients

- The data D.
- The prior $w_0(x)$ for the parameters x.
- The likelihood $L(D|x)$ that follows from the model.
- The posterior $w_D(x) \propto L(D|x)w_0(x)$.

All ingredients should be checked against the data.

- Are the data typical? Conduct run tests; check for prior-data conflicts.
- Is the model correct? Conduct the necessary model checking.
- Is there a bias in the prior? Make sure, there isn't.

The four experiments: What they have in common



γ	trigger-signal–to–qubit-pair conversion probability
$\mathbf{a}, \mathbf{a}'$	unit vectors that specify Alice's selection, with $\mathbf{a} \cdot \mathbf{a}' = \cos\theta_A$
$\mathbf{b}, \mathbf{b}'$	unit vectors that specify Bob's selection with $\mathbf{b} \cdot \mathbf{b}' = \cos\theta_B$
η_A, η_B	Alice's and Bob's detection probabilities

Each setting $S = ab$ or ab' or $a'b$ or $a'b'$ has four probabilities: $p_{++}^{(S)}, p_{+0}^{(S)}, p_{0+}^{(S)}, p_{00}^{(S)}$.

The four experiments: How they differ

Experiment	γ	θ_{A}	θ_{B}	η_{A}	η_{B}	N
Delft	1	90°	80.6°	0.971	0.963	245
(2 runs)						228
Vienna	0.0035	64°	64°	0.786	0.762	3843698536
(3 runs)						3502784150
						9994696192
Boulder	0.0005	60.2°	60.2°	0.747	0.756	175647100
(4 runs)						886791755
						527164272
						1244205032
Munich	1	90°	90°	0.975	0.975	27885
(2 runs)						27683

$D = \left(n_{++}^{(ab)}, n_{+0}^{(ab)}, \ldots, n_{0+}^{(ab')}, \ldots, n_{00}^{(a'b')}\right)$ reports the data for one run of the experiment as a 16-element string of natural numbers, namely the 16 counts of events.
Their sum $N = n_{++}^{(ab)} + n_{+0}^{(ab)} + \cdots + n_{0+}^{(ab')} + \cdots + n_{00}^{(a'b')}$ is the total number of trigger signals (last column in the table).

Boulder experiment: Run with 5 triggers per trial

Experimental data: Recorded counts

S	$n^{(S)}_{++}$	$n^{(S)}_{+0}$	$n^{(S)}_{0+}$	$n^{(S)}_{00}$
ab	6 378	3 289	3 147	221 732 456
ab'	6 794	2 825	23 230	221 686 486
$a'b$	6 486	21 358	2 818	221 635 498
$a'b'$	106	27 562	30 000	221 603 322

Prior and posterior contents

region	prior	posterior
QM only	0.0006	0
both	0.5026	**1**
LHV only	0.4969	0

Very strong evidence in favor of “both”.
The data are inconclusive.

How does this square with the conclusion of the Boulder team who refute LHV on the basis of a p-value of 2.3×10^{-7}?

Boulder experiment: Model checking (1)

An exercise in quantum state estimation: Maximize the likelihood over the QM-permissible probabilities ($\rightarrow$ QM-MLE) and also over the LHV-permissible probabilities ($\rightarrow$ LHV-MLE).

Maximum values of the likelihood: 7.36×10^{-694} (QM) versus 5.54×10^{-40} (LHV). The data are much much more likely for LHV than for QM — by more than 650 orders of magnitude.

The actually observed data are extremely untypical; this should not happen when so many events are recorded.

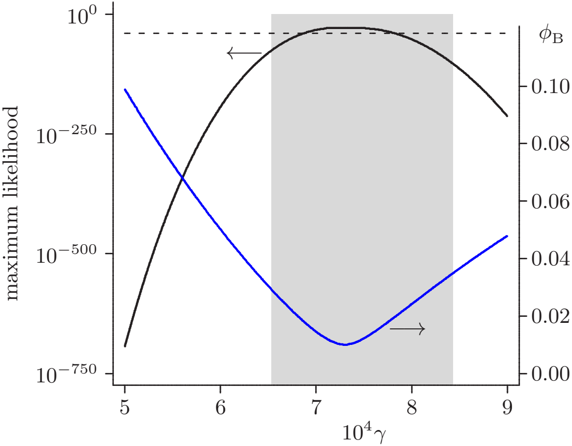
Problem with $\gamma \approx 5 \times 10^{-4}$? Check the model, estimate γ from the data.

New maximum values of the likelihood: 6.17×10^{-29} (QM) versus 5.54×10^{-40} (LHV), obtained for $\gamma = 0.000\,722$. The data are much more likely for QM than for LHV — by eleven orders of magnitude.

region	prior	posterior
QM only	0.0006	**1**
both	0.5025	0
LHV only	0.4969	0

Very strong evidence in favor of "QM only".

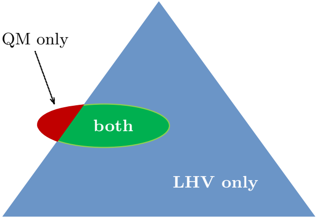
Boulder experiment: Model checking (2)



Solid black curve: QM-MLE; dashed black line: LHV-MLE;
solid blue: Bhattacharyya angle between the probabilities of the target state and the QM-MLE.

Gray range of γ values: QM-MLE probabilities violate a Bell-type inequality of the Eberhard kind.

Boulder experiment: Checking the prior for bias



10,000 mock-true probabilities in region	number of cases with evidence in favor of region: QM only	both	LHV only
QM only	8 809	1 278	0
both	0	8 365	1 635
LHV only	0	145	9 855

There is no procedural bias in favor of the "QM only" region.

The other Boulder runs and the other three experiments **tell the same story**, see Y. Gu et al., *Very strong evidence in favor of quantum mechanics and against local hidden variables from a Bayesian analysis*, Phys. Rev. A **99**, 022112 (2019).

Take-home messages

Quantum data are statistical in nature.

(You knew this already.)

When planning an experiment,
use concepts from frequentism.

Rely on Bayesian logic when drawing inference
from the actual data acquired in an experiment.

LHV are a lost cause.

(Voice from the off: They always were.)

QM vs LHV: arXiv:1808.06863 ≡ PRA **99**, 022112 (2019) & references therein.

Frequentism vs Bayes: Appendix A in arXiv:1602.05780 ≡ PRA **94**, 062112 (2016).