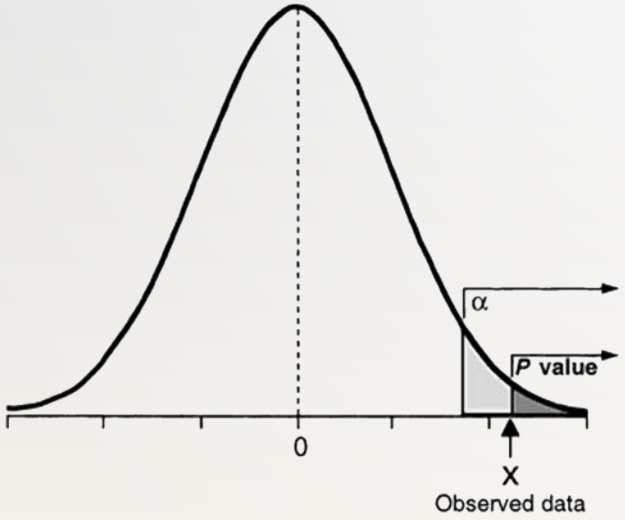
The evolution of the p-value in medical research: How science can improve Statistics



Jeffrey D. Blume, PhD

Vanderbilt University

Nashville, TN USA

Baruch Spinoza

- 1632-1677
- Philosopher
- Rationalist
- Hebrew grammar

Gottfried Leibniz

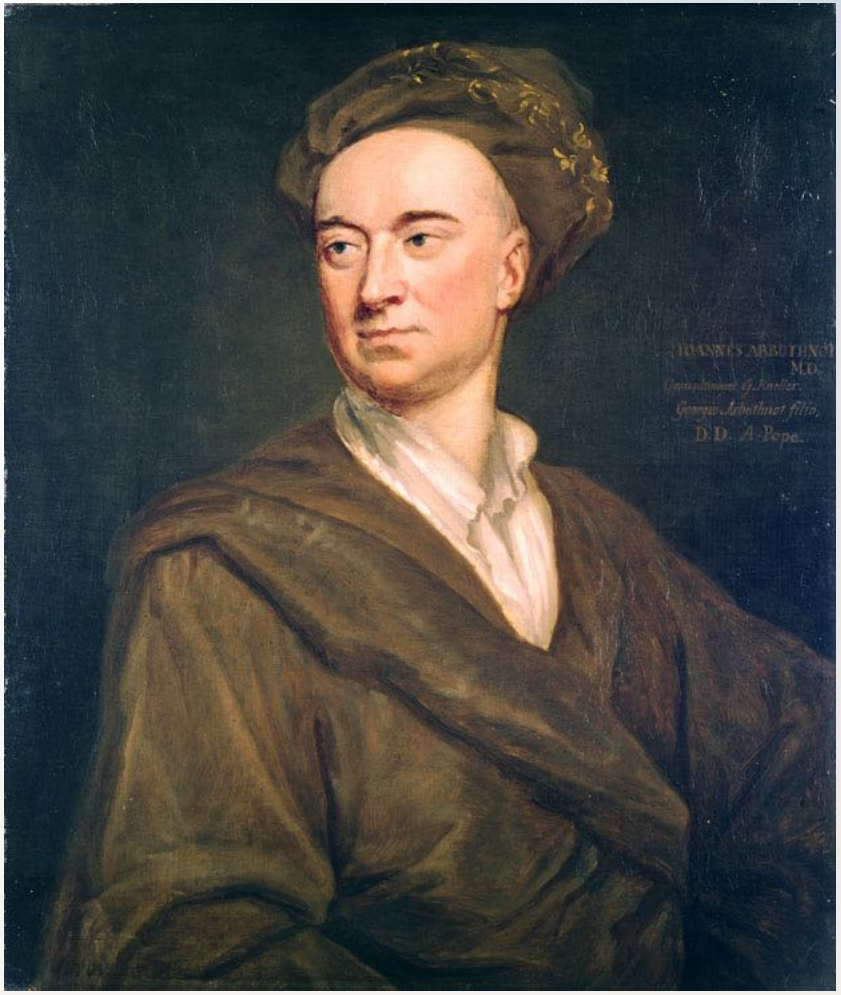
- 1646-1716
- Philosopher
- Rationalist
- Mathematician

John Arbuthnot

- 1667-1735
- Episc. cleric
- Demographer
- Physician
- Queen Anne's Medic

John Arbuthnot

- 1667-1735
- Scottish / English Cleric (Episcopalian)
- Demographer, Physician, Mathematician
- Credited with first published description of the reasoning behind statistical tests



Arbuthnot in 1710

- *An Argument for Divine Providence Taken From the Constant Regularity of the Births of Both Sexes*
- Studied register of London births from 1629 to 1710
- Noticed more males were born each year in the register (82 consecutive years)
- Was this ‘art’ (Divine Providence) or ‘chance’ (randomness)?

II. *An Argument for Divine Providence, taken from the conſtant Regularity obſerv'd in the Births of both Sexes. By Dr.* John Arbuthnott, *Phyſitian in Ordinary to Her Majeſty, and Fellow of the College of Phyſitians and the Royal Society.*

Anno.	Chriſtened. Males.	Females.	Anno.	Chriſtened. Males.	Females.
1629	5218	4683	1648	3363	3181
30	4858	4457	49	3079	2746
31	4422	4102	50	2890	2722
32	4994	4590	51	3231	2840
33	5158	4839	52	3220	2908
34	5035	4820	53	3196	2959
35	5106	4928	54	3441	3179
36	4917	4605	55	3655	3349
37	4703	4457	56	3668	3382
38	5359	4952	57	3396	3289
39	5366	4784	58	3157	3013
40	5518	5332	59	3209	2781
41	5470	5200	60	3724	3247
42	5460	4910	61	4748	4107
43	4793	4617	62	5216	4803
44	4107	3997	63	5411	4881
45	4047	3919	64	6041	5681
46	3768	3395	65	5114	4858
47	3796	3536	66	4678	4319

Anno.	Chriſtened. Males.	Females.	Anno.	Chriſtened. Males.	Females.
1667	5616	5322	1689	7604	7167
68	6073	5560	90	7909	7302
69	6506	5829	91	7662	7392
70	6278	5719	92	7602	7316
71	6449	6061	93	7676	7483
72	6443	6120	94	6985	6647
73	6073	5822	95	7263	6713
74	6113	5738	96	7632	7229
75	6058	5717	97	8062	7767
76	6552	5847	98	8426	7626
77	6423	6203	99	7911	7452
78	6568	6033	1700	7578	7061
79	6247	6041	1701	8102	7514
80	6548	6299	1702	8031	7656
81	6822	6533	1703	7765	7683
82	6909	6744	1704	6113	5738
83	7577	7158	1705	8366	7779
84	7575	7127	1706	7952	7417
85	7484	7246	1707	8379	7687
86	7575	7119	1708	8239	7623
87	7737	7214	1709	7840	7380
88	7487	7101	1710	7640	7288

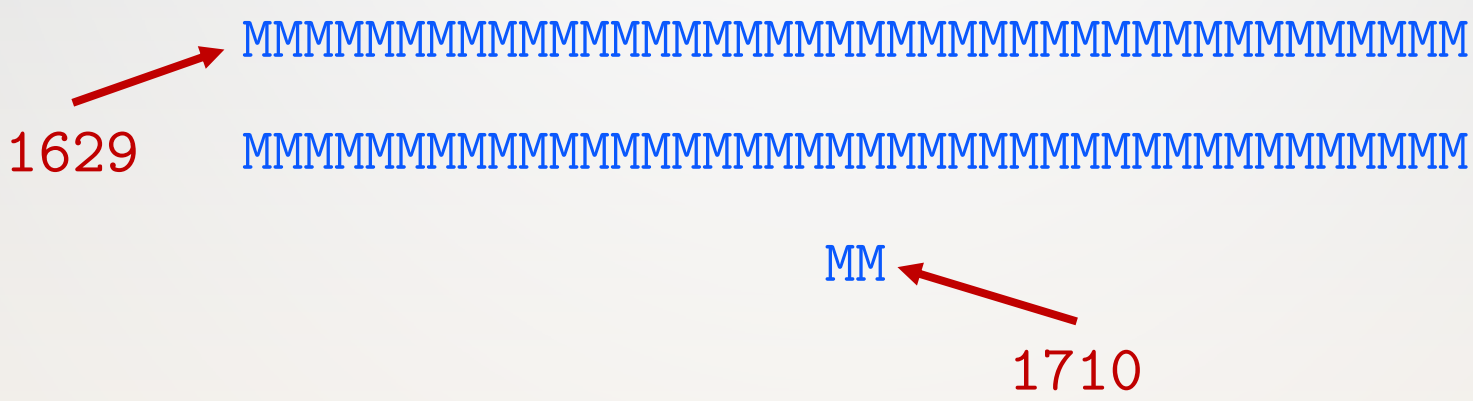
Arbuthnot Data

- Let

  M = Male births in year >= Female births in year

  F = Male births in year < Female births in year

- Arbuthnot observed 82 M's:



The null hypothesis

- Was this ‘art’ (Divine Providence) or ‘chance’ (randomness)?

- Null hypothesis set as Random Chance:

  P(male birth) = 0.5

  Very specific claim!

Arbuthnot in 1710

- He argued: if there was an even chance for male and female births, the distribution of births would be like outcomes from tossing a fair coin.

- He assumed: Births are independent events

- Now, if P(male birth) = 0.5 then P(male year) = 0.5

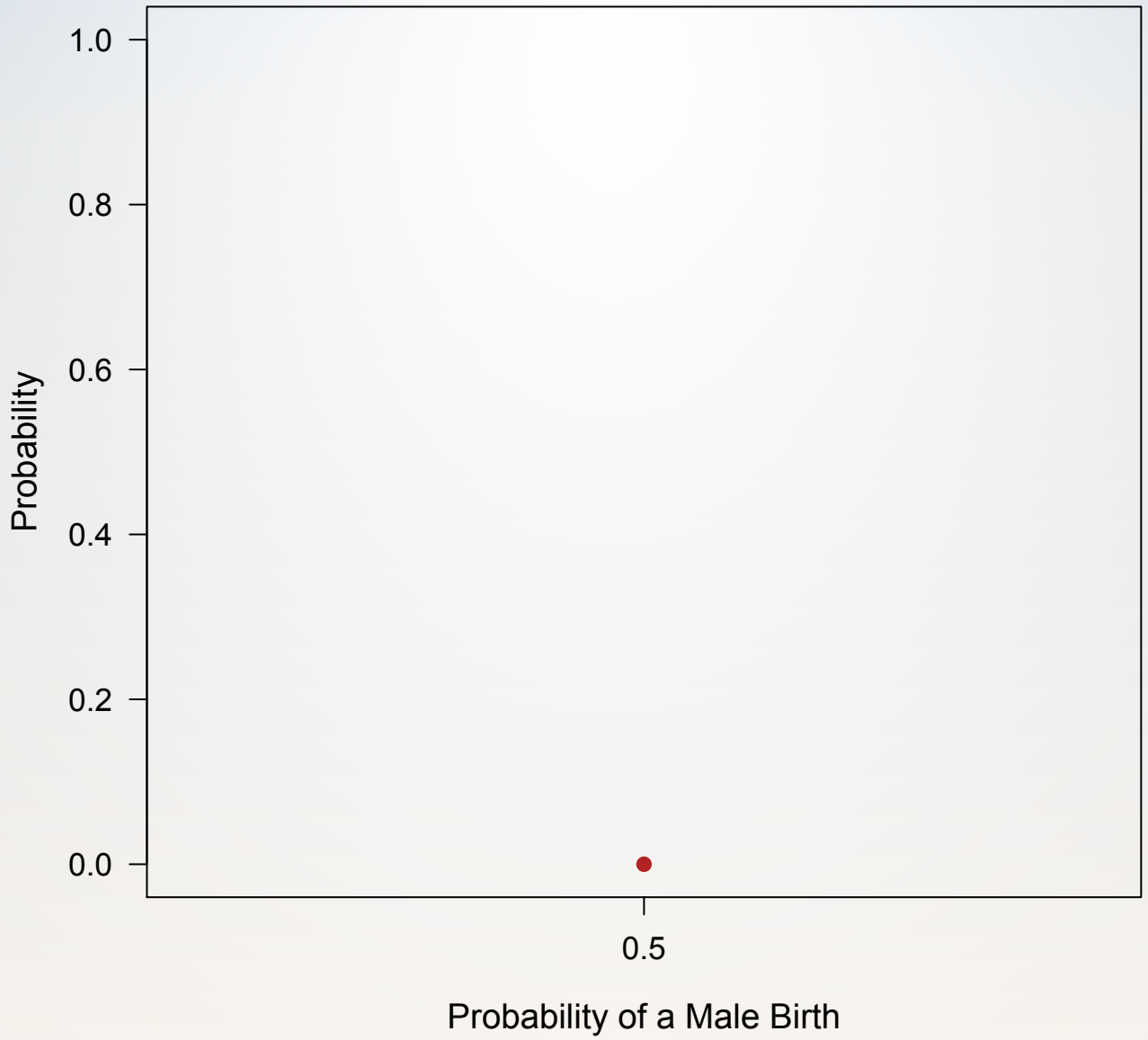
- So P(82 consec. male years) = $(1/2)^{82} = 2.068 \times 10^{-25}$

Arbuthnot in 1710

- He concluded: because P(82 consecutive males years) is very very small "*it follows that it is Art, not Chance, that governs*" the distribution of sexes.

- First use of statistical test reasoning (sign test)


- First to ignore clinical significance

Spinoza

Blume, Nov 2019

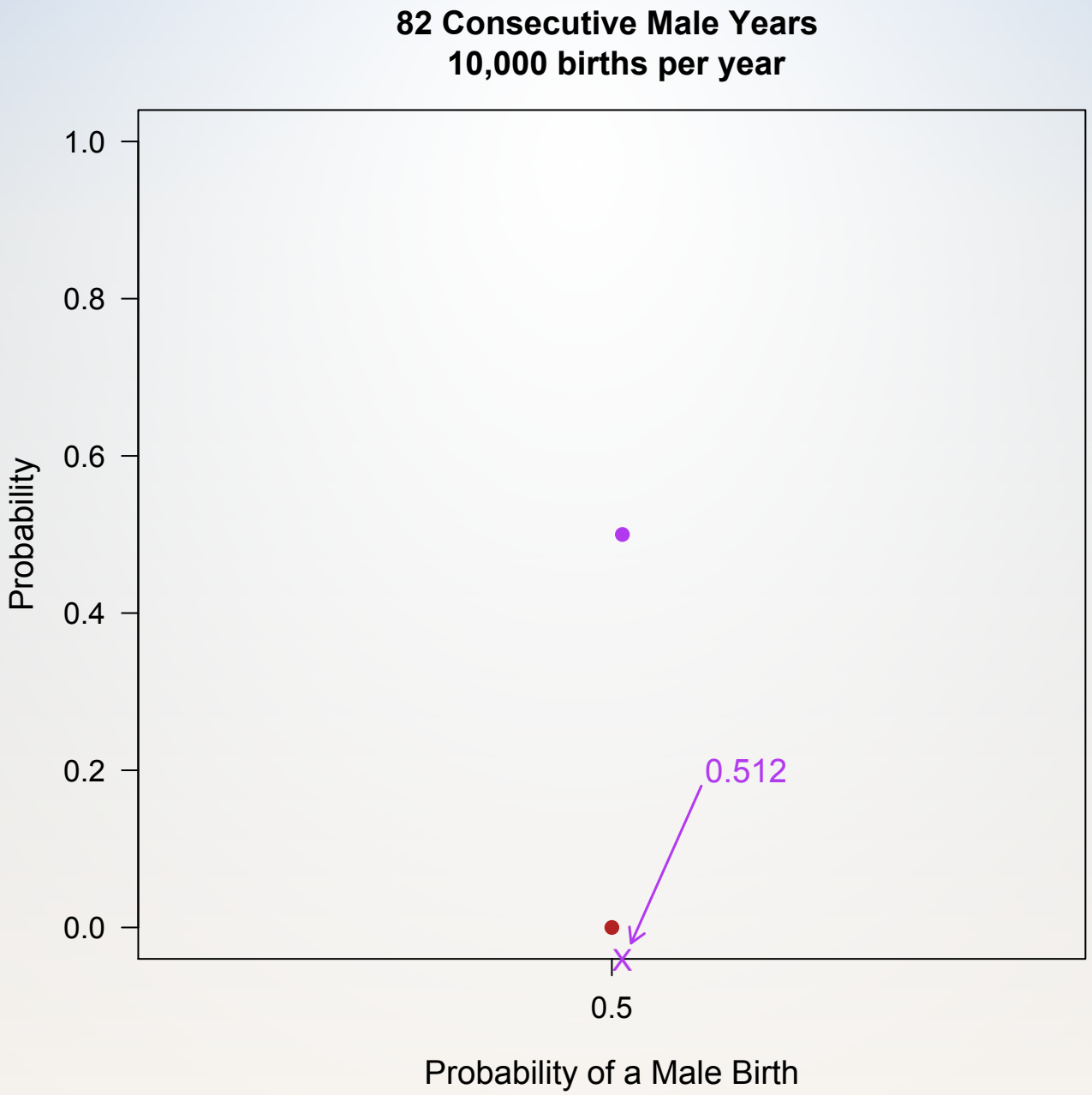
More on Arbuthnot's test

- Assuming 10,000 births per year...

- If P(male birth) = 0.512 then P(male year) = 0.99159

  Then P(82 consecutive males years) = $(0.99159)^{82}$ = 0.5

  Now, 82 consecutive males years is just as likely as not



Spinoza

Blume, Nov 2019

More on Arbuthnot's test

- Assuming 10,000 births per year...

- If P(male birth) = 0.512 then P(male year) = 0.99159

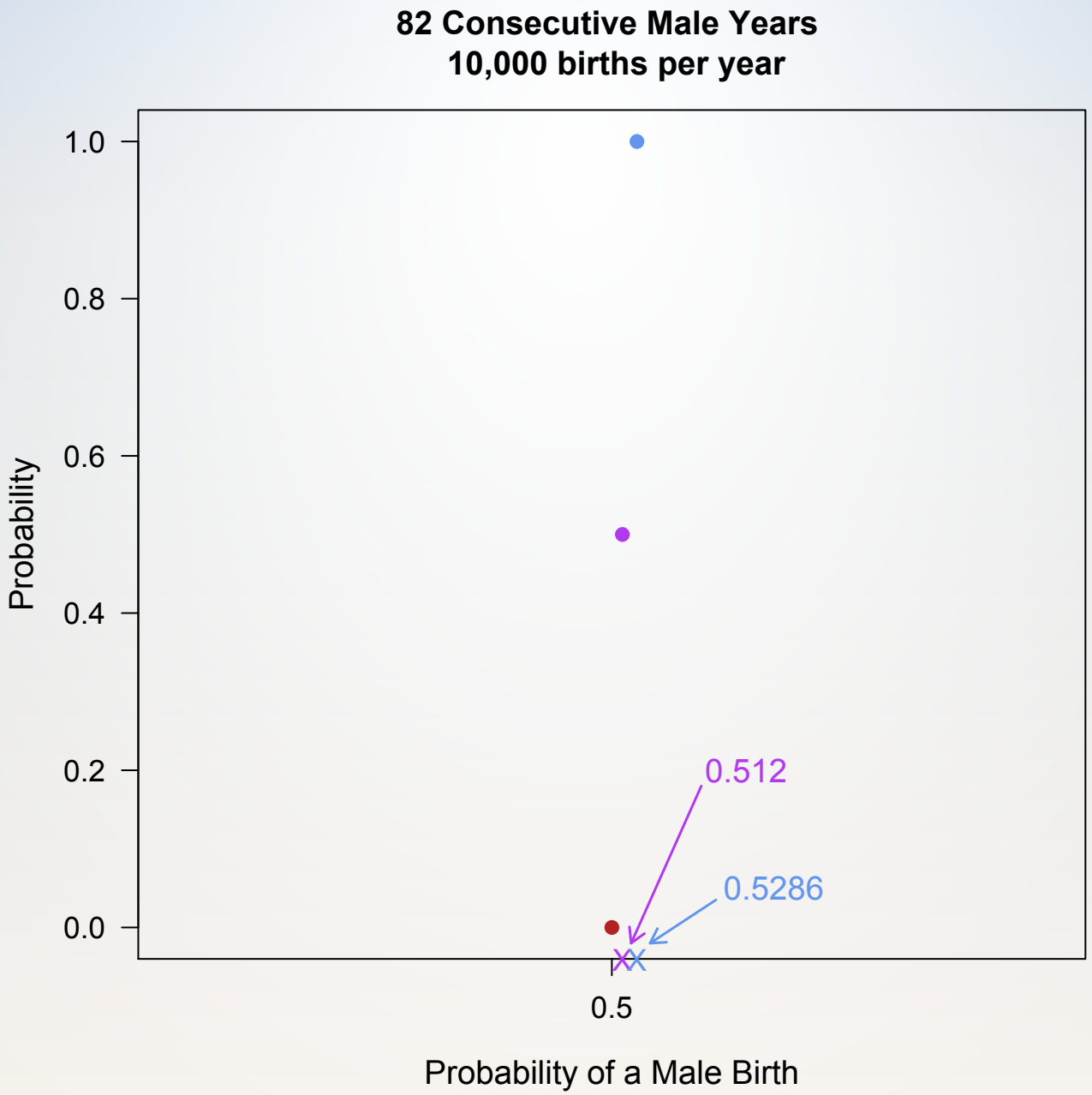
  Then P(82 consecutive males years) = $(0.99159)^{82}$ = 0.5

  Now, 82 consecutive males years is just as likely as not

- If P(male birth) = 0.528 then P(male year) = 1

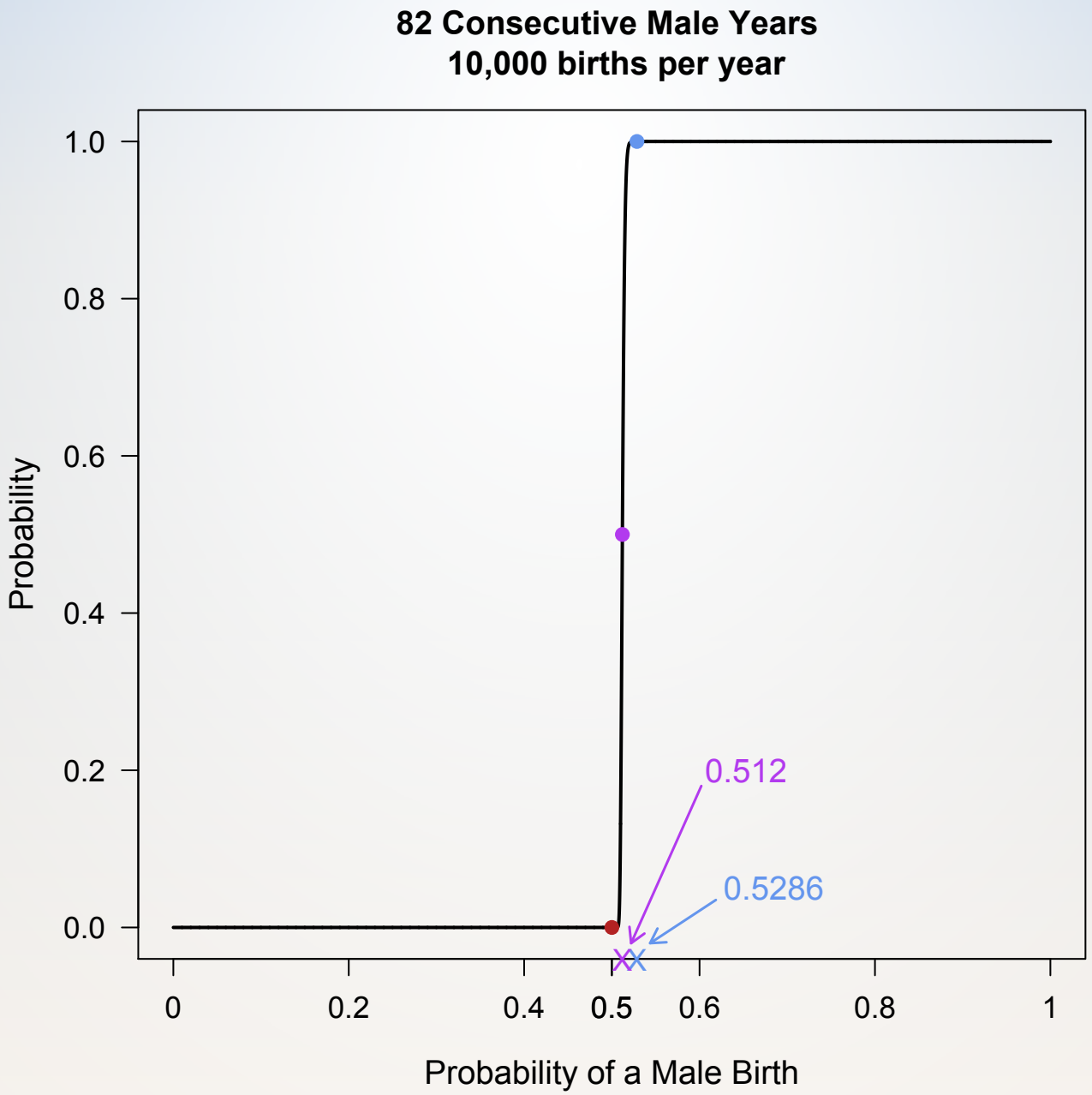
  Then P(82 consecutive males years) = $(1)^{82}$ = 1

  Now, 82 consecutive males years is a certainty



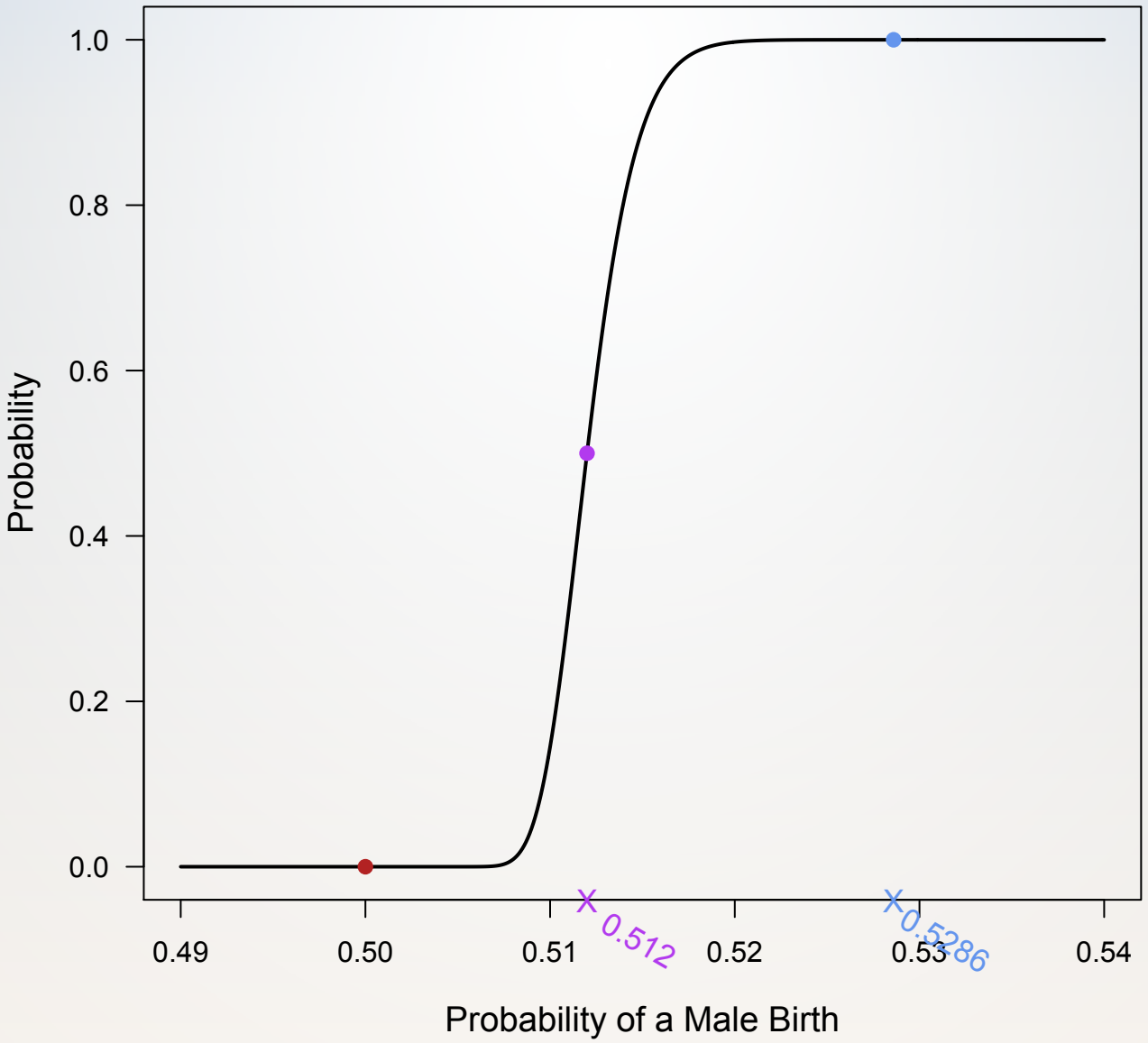
Spinoza

Blume, Nov 2019



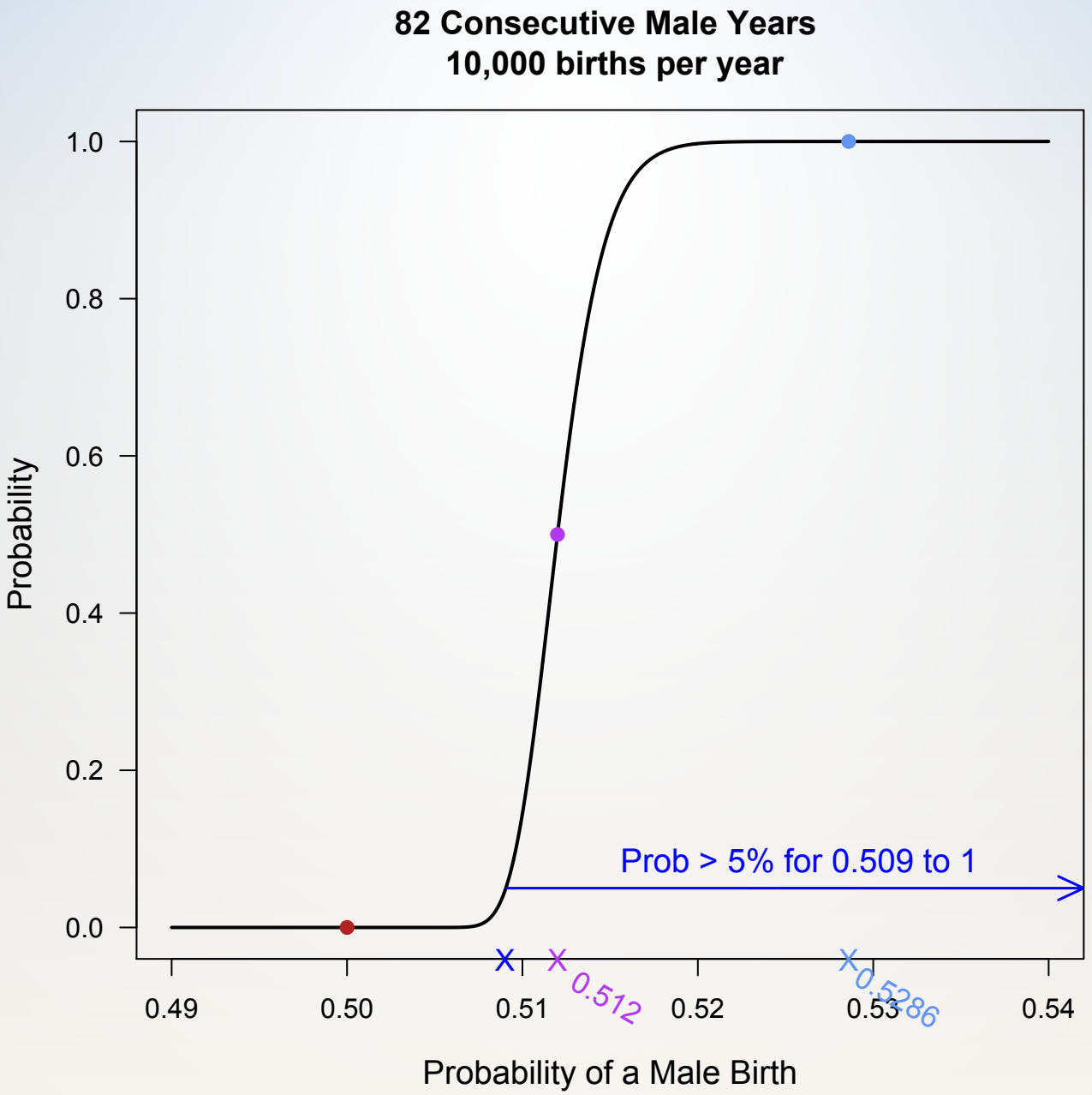
Spinoza

Blume, Nov 2019



Spinoza

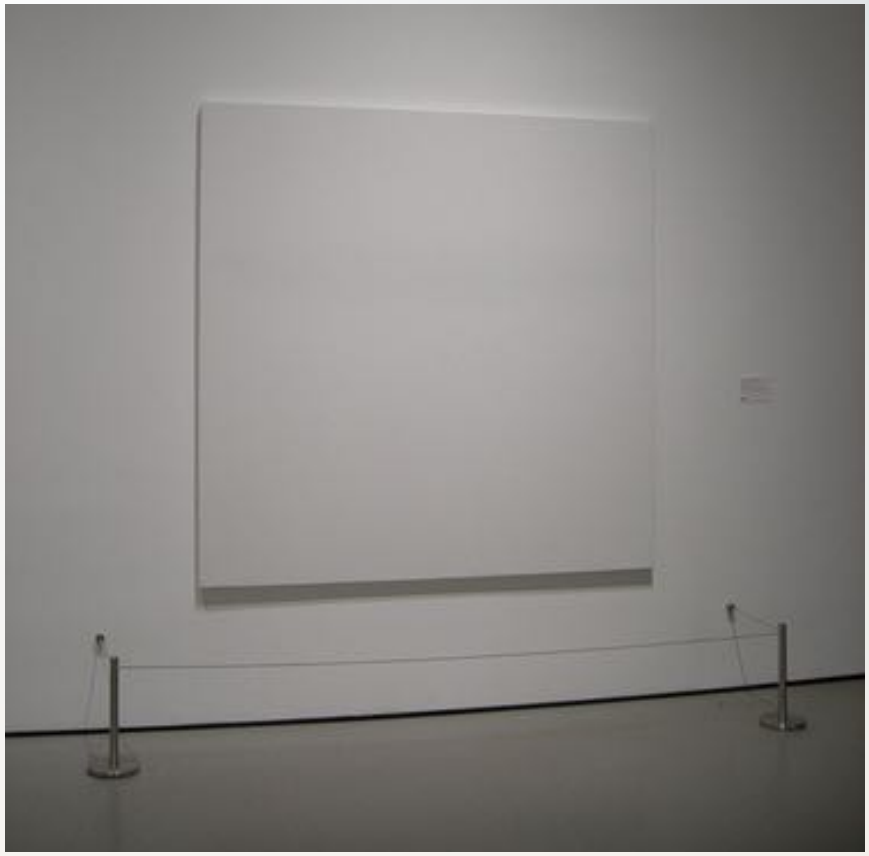
Blume, Nov 2019


82 Consecutive Male Years
10,000 births per year
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Prob > 5% for 0.509 to 1
0.512
0.5286
0.49
0.50
0.51
0.52
0.53
0.54
Probability of a Male Birth

Scale is Important

- The data are consistent with a chance of male birth being greater than 0.509.

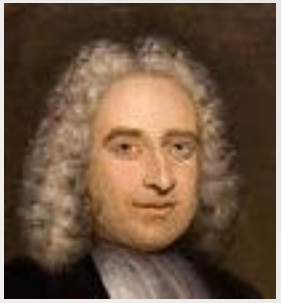
- “Art, not Chance...”
- Modern Art?

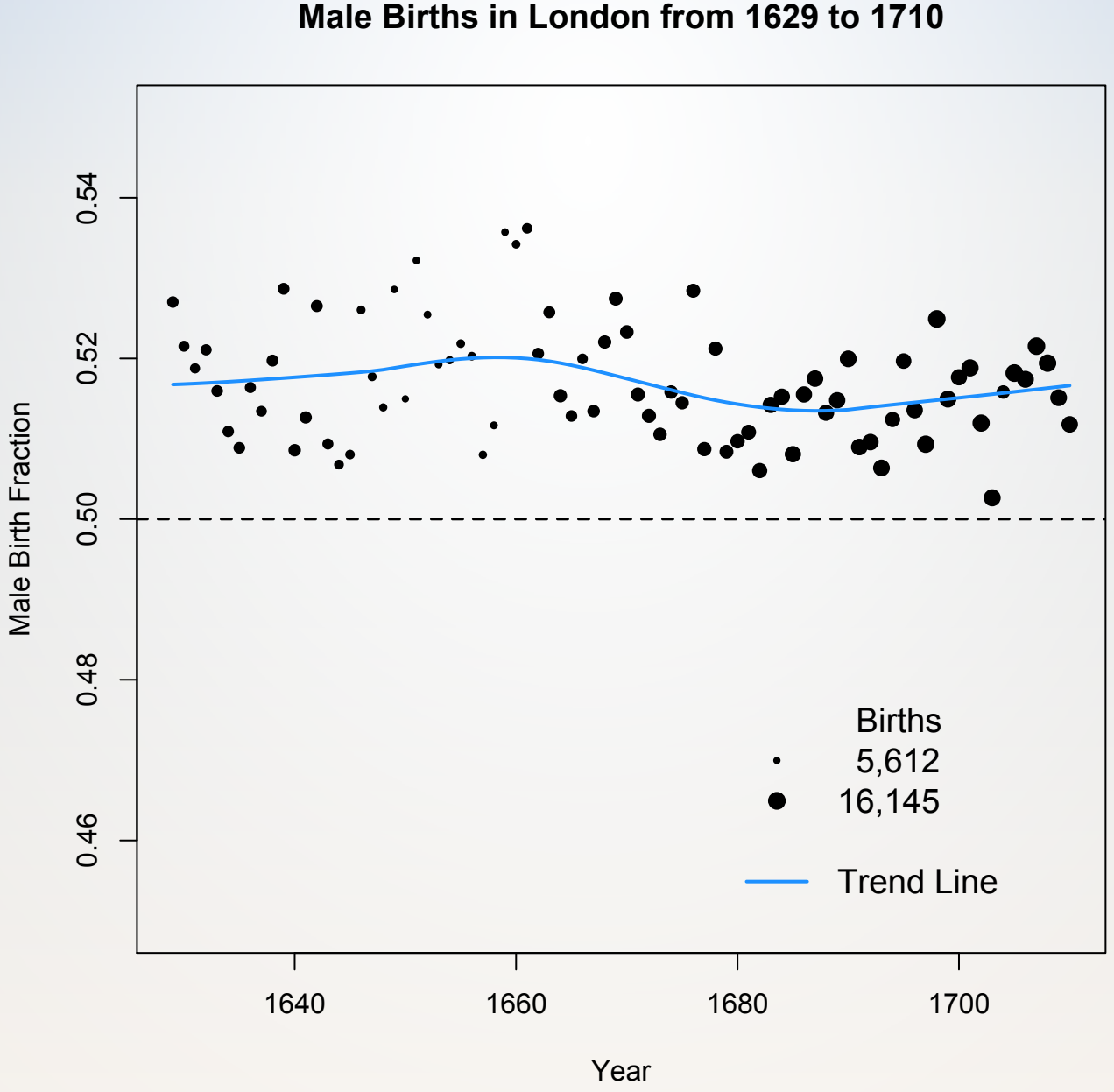


Discussion Ensued

- Subsequent discussions of his analysis led to many technical discoveries over time.

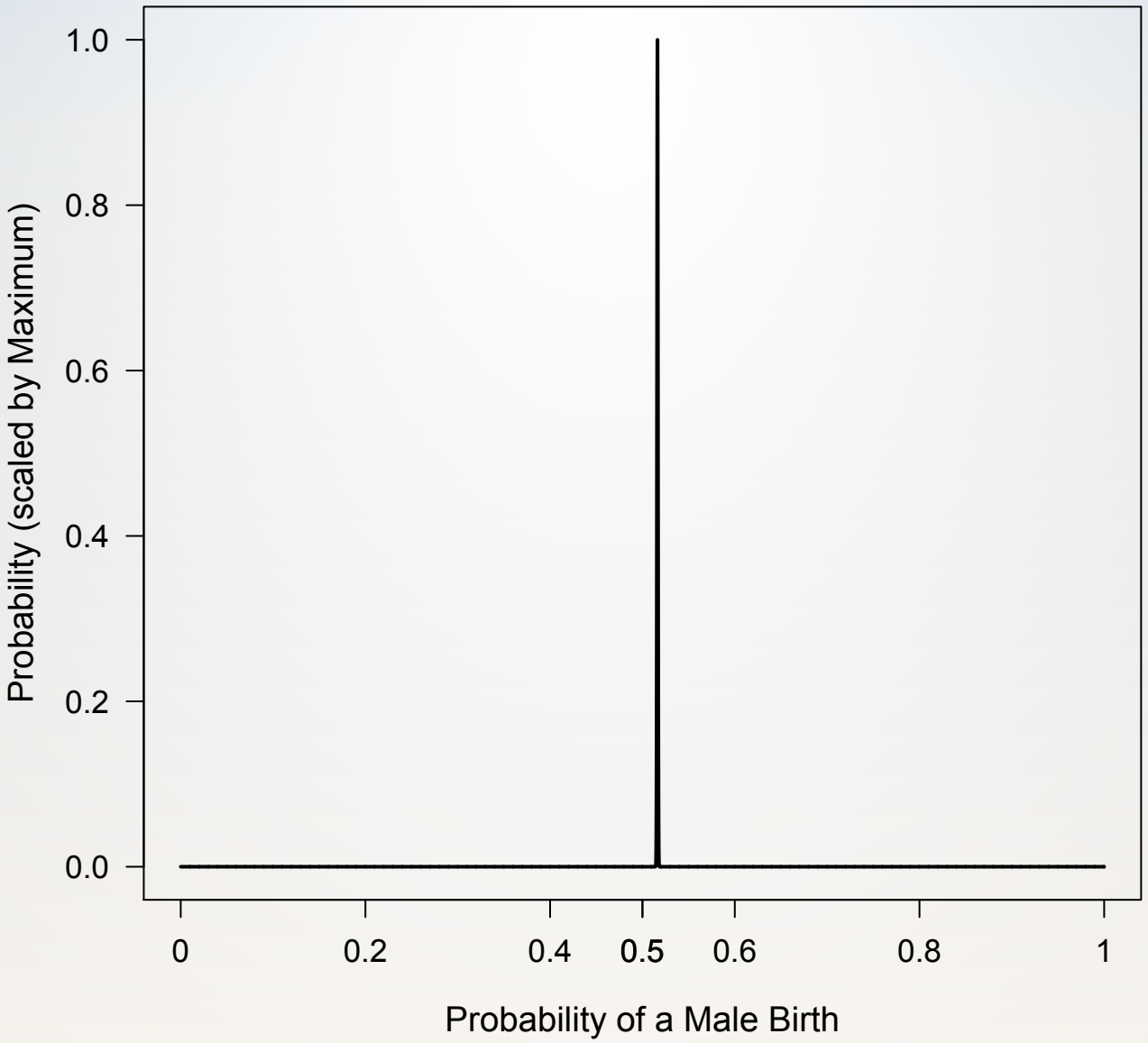
- Major contemporary discussants:
  - **Niklaus II Bernoulli** (1687-1759), Swiss Mathematician and member of famous Bernoulli family.
  - **Willem (or Guillaume) ‘sGravesande** (1688-1742), a Dutch scientist who in 1718 became Prof. of Mathematics at the University of Leiden.
  - **Bernard Nieuwentijt** (1654-1718), a Dutch physician and mathematician.





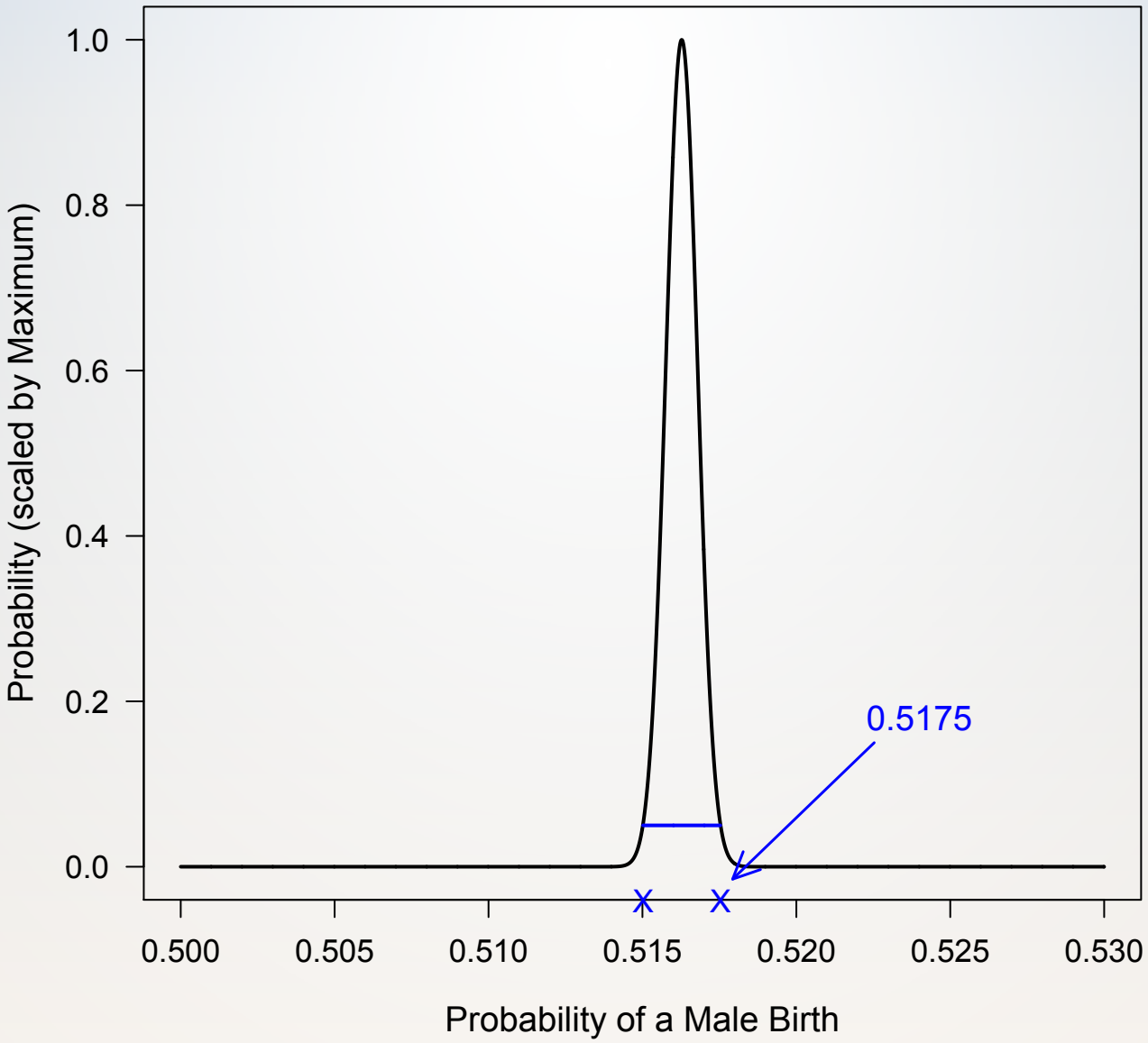
Spinoza

Blume, Nov 2019



Spinoza

Blume, Nov 2019



Spinoza

Blume, Nov 2019

Sex ratios at birth in Europe and North America (Grech et al. 2002)

Region/country	Male:female ratio (95% CI
Central Europe (40-55°):	0.5140 (0.5139 to 0.5141)
Austria	0.5132 (0.5128 to 0.5136)
Belgium	0.5141 (0.5137 to 0.5145)
Czech Republic	0.5135 (0.5127 to 0.5142)
France	0.5124 (0.5122 to 0.5125)
Hungary	0.5153 (0.5150 to 0.5157)
Ireland	0.5141 (0.5136 to 0.5147)
Luxembourg	0.5145 (0.5124 to 0.5165)
Netherlands	0.5130 (0.5126 to 0.5133)
Poland	0.5158 (0.5156 to 0.5160)
Romania	0.5138 (0.5135 to 0.5140)
Switzerland	0.5131 (0.5127 to 0.5136)
United Kingdom	0.5140 (0.5138 to 0.5141)
Germany	0.5144 (0.5143 to 0.5146)
Central Europe and Mediterranean (35-55°)	0.5142 (0.5142 to 0.5143)

Established Testing Elements

- Set a Null hypothesis

- Compute the probability of observing the data (or data more extreme) assuming the null hypothesis is true

- Examine that probability
  - If the probability is small, ...
  - If the probability is large, ...

. . .

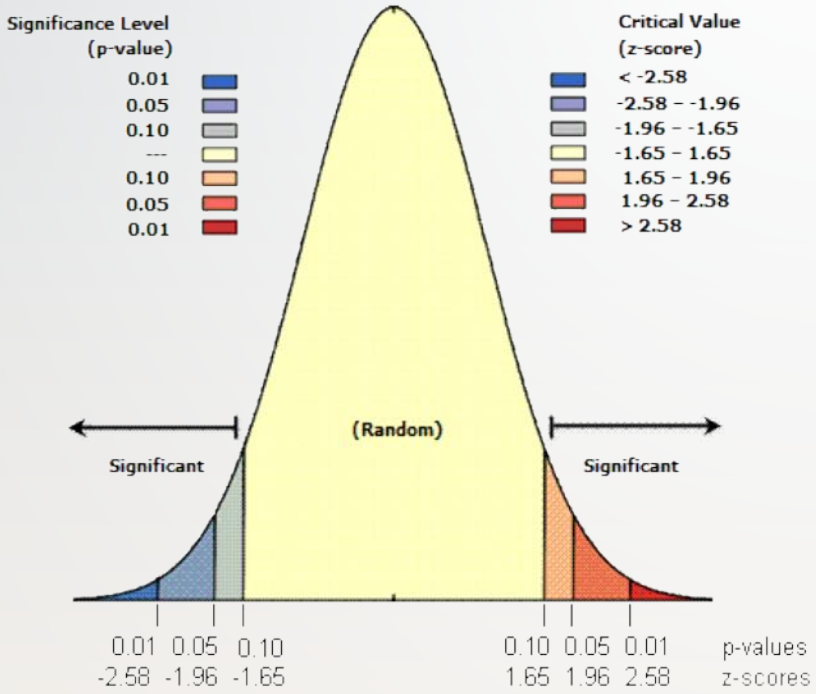

...and this is where we put the non-significant results.
someecards
user card

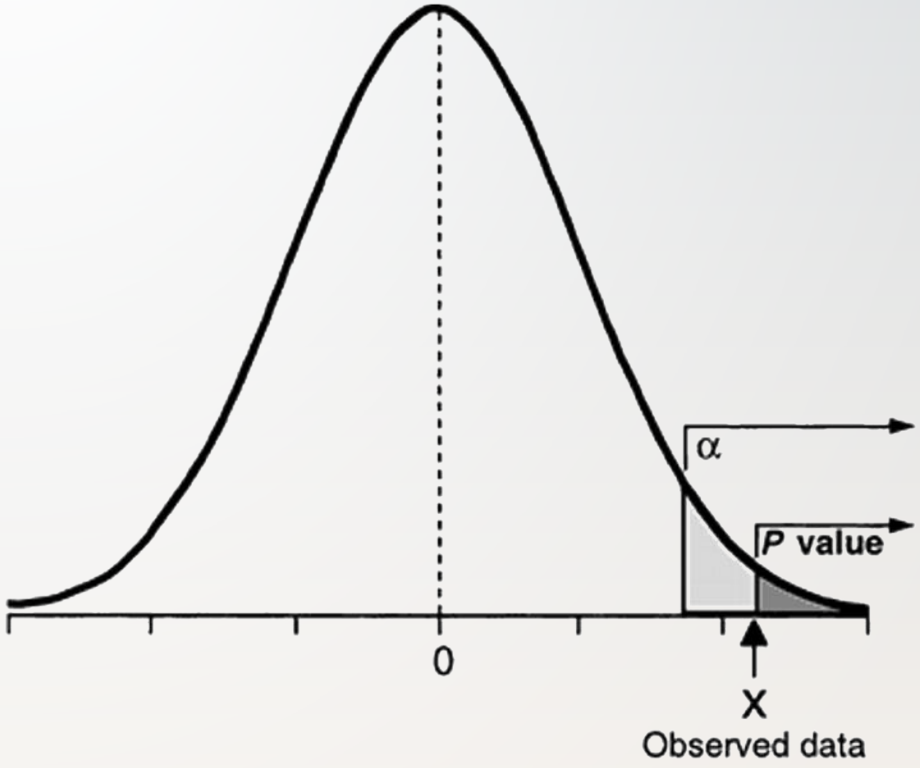
Definition of p-value emerges

- Over the next 100-150 years, an informal definition for a p-value coalesces from Arbuthnot's and other's subsequent work:

The probability, under a specified statistical model or hypothesis, that a summary of the data (for example, a difference in sample means) would be equal to or more extreme than its observed value.

What a *p*-value looks like





Timeline of Testing Ideas

1778
Laplace

1920s
RA Fisher

Fisher & Neyman
fight for
the soul of
statistics

1710
Arbuthnot

1900s
K. Pearson

1933
Neyman &
E. Pearson

Today

Spinoza

Significance Testing

- RA Fisher (1925, 1956)
- Specify null hypothesis, compute p-value
- P-values 'measure the evidence against the null hypothesis'

- Said no uniform cut-off exists
- Proposed 0.05 cut-off, but said cut-off could change per context.
- Attempted to clarify; did a poor job

Hypothesis Testing

No concept of statistical evidence

- E. Pearson & J. Neyman (1933)
- "The problem of testing statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or take step B..."

- "Thus to accept a hypothesis means only to decide to take action A rather than action B. This does not mean that we necessarily believe that the hypothesis H is true." (Neyman 1950)

Hypothesis Testing

- E. Pearson & J. Neyman (1933)
- Solved the problem of how to optimally choose between two competing hypotheses
- Solution:
  - Pre-specify Type I Error (alpha)
  - Check if p-value < alpha to Reject the null
- Uh-oh. This kinds looks like significance testing
- Attempted to clarify; did a poor job

So what do we have?

- Arbuthnot.
- Two competing frameworks for testing that are conceptually quite different but use nearly the same statistical metrics.
- Neither framework stresses the formulation of the null hypothesis nor consideration of effect sizes.
- As a consequence, we have not progressed, in a scientific sense, much further than Arbuthnot.

Over time...

- Elements of significance testing and hypothesis testing were blended.
- Studies are designed by hypothesis testing, results are reported by significance testing.
- This leads to fundamental paradoxes like those of multiple comparisons and multiple looks at the data.

- This also discourages the reporting of effect sizes that are essential to revealing science

Evidential metrics

Example:
Diagnostic Test

1. **Measure of the strength evidence**
   - → Axiomatic and intuitive justification
   - → Summary statistic, yardstick

Positive Test
Negative Test

2. **Propensity to collect data that will yield a misleading #1**
   - → Error rates
   - → Properties of the study design (!)

Sensitivity
Specificity

3. **Probability that an observed #1 is misleading**
   - → False Discovery rate, False Confirmation rate
   - → Chance that an observed result is mistaken
   - → Properties of the observed data (!)

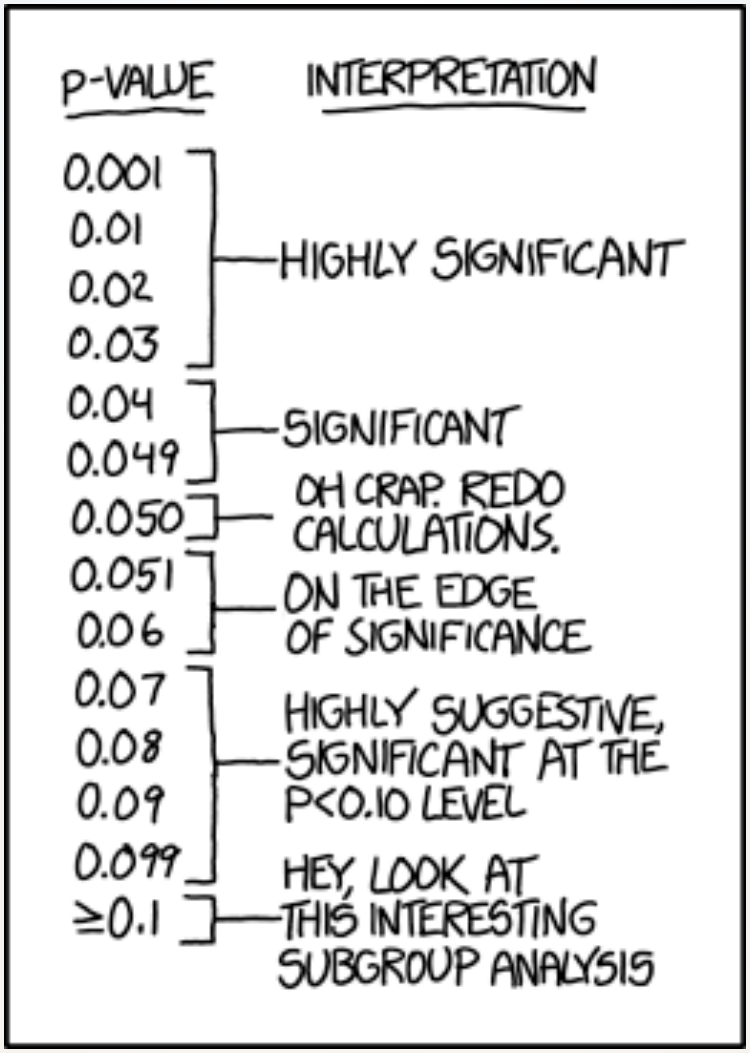
PPV
NPV

This is now

Evidential Metric	What it measures	Hypothesis Testing	Significance Testing
1	strength of the evidence	Absent	**Tail-area probability** (*p*-value)
2	propensity for study to yield misleading evidence	**Tail-area probability** (error rates)	Absent
3	propensity for observed results to be misleading	misinterpret #2	misinterpret #1

- The *tail-area probability* is used to measure *three* distinct metrics

This is also now



Real life descriptions

When $p \sim 0.06$

- almost significant
- almost attained significance
- almost significant tendency
- almost became significant
- almost but not quite significant
- almost statistically significant
- almost reached statistical significance
- just barely below the level of significance
- just beyond significance

Real life descriptions

When $p \sim 0.08$

- a certain trend toward significance
- a definite trend
- a slight tendency toward significance
- a strong trend toward significance
- a trend close to significance
- an expected trend
- approached our criteria of significance
- approaching borderline significance
- approaching, although not reaching, significance

Real life descriptions

Nearly $p \sim 0.05$

- hovered at nearly a significant level (p=0.058)
- hovers on the brink of significance (p=0.055)
- just about significant (p=0.051)
- just above the margin of significance (p=0.053)
- just at the conventional level of significance (p=0.05001)
- just barely statistically significant (p=0.054)
- just borderline significant (p=0.058)
- just escaped significance (p=0.057)
- just failed significance (p=0.057)

"... we only wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05.

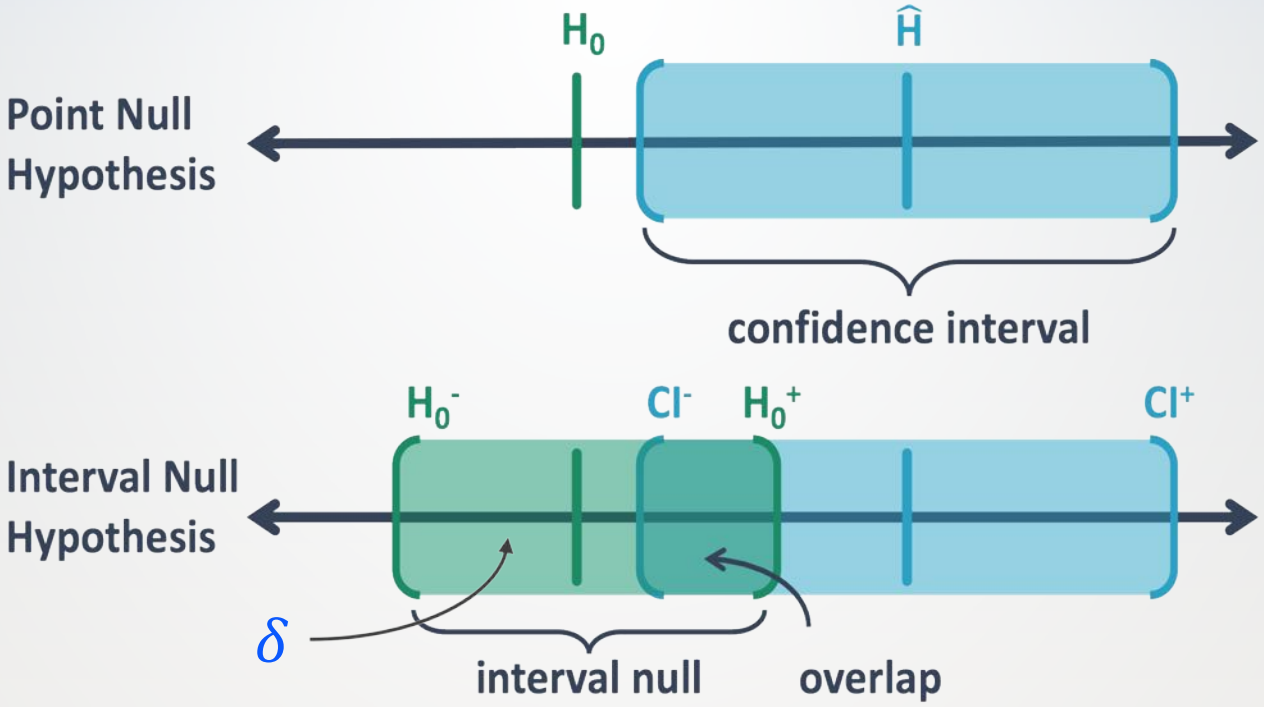
Rosnow, R.L. and Rosenthal, R. 1989. Statistical procedures and the justification of knowledge and psychological science. *American Psychologist* 44: 1276-1284

Thanks to Matthew Hankins and Ron Wasserstein for these quotes.

Back to Basics (?)

- → Collect data
- → Compute a confidence interval (CI)
- → Examine the CI to see if
  - → the null hypothesis is in the CI
  - → other practically null hypotheses are in the CI
  - → Measure how much of the CI is null
- → Re-focuses the statistical analysis on the effect size, the science
  - → this improves error rates!
- → Recommended best practice and SGPV can report this

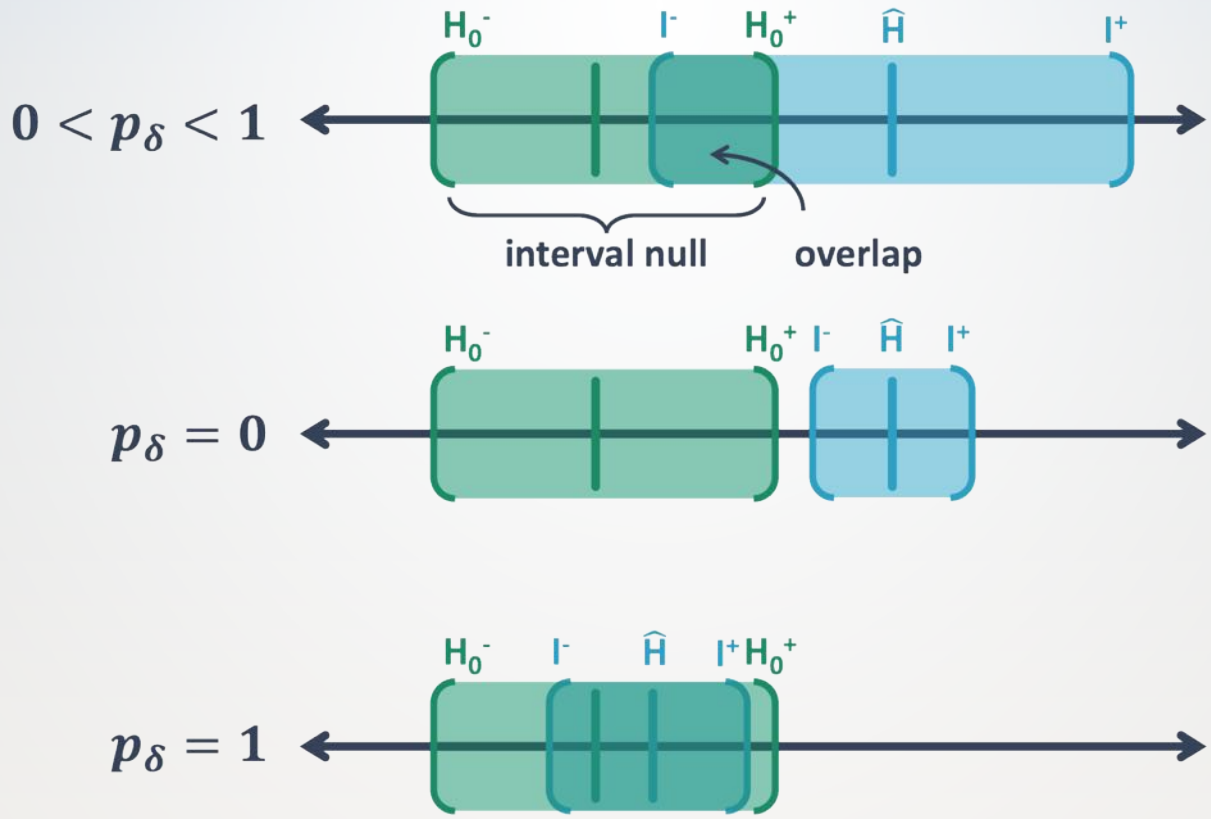
Illustration



Point null hypothesis H_0 and interval null hypothesis $[H_0^-, H_0^+]$

Data-supported hypothesis $\hat{H}$ and confidence interval $[CI^-, CI^+]$

Illustration

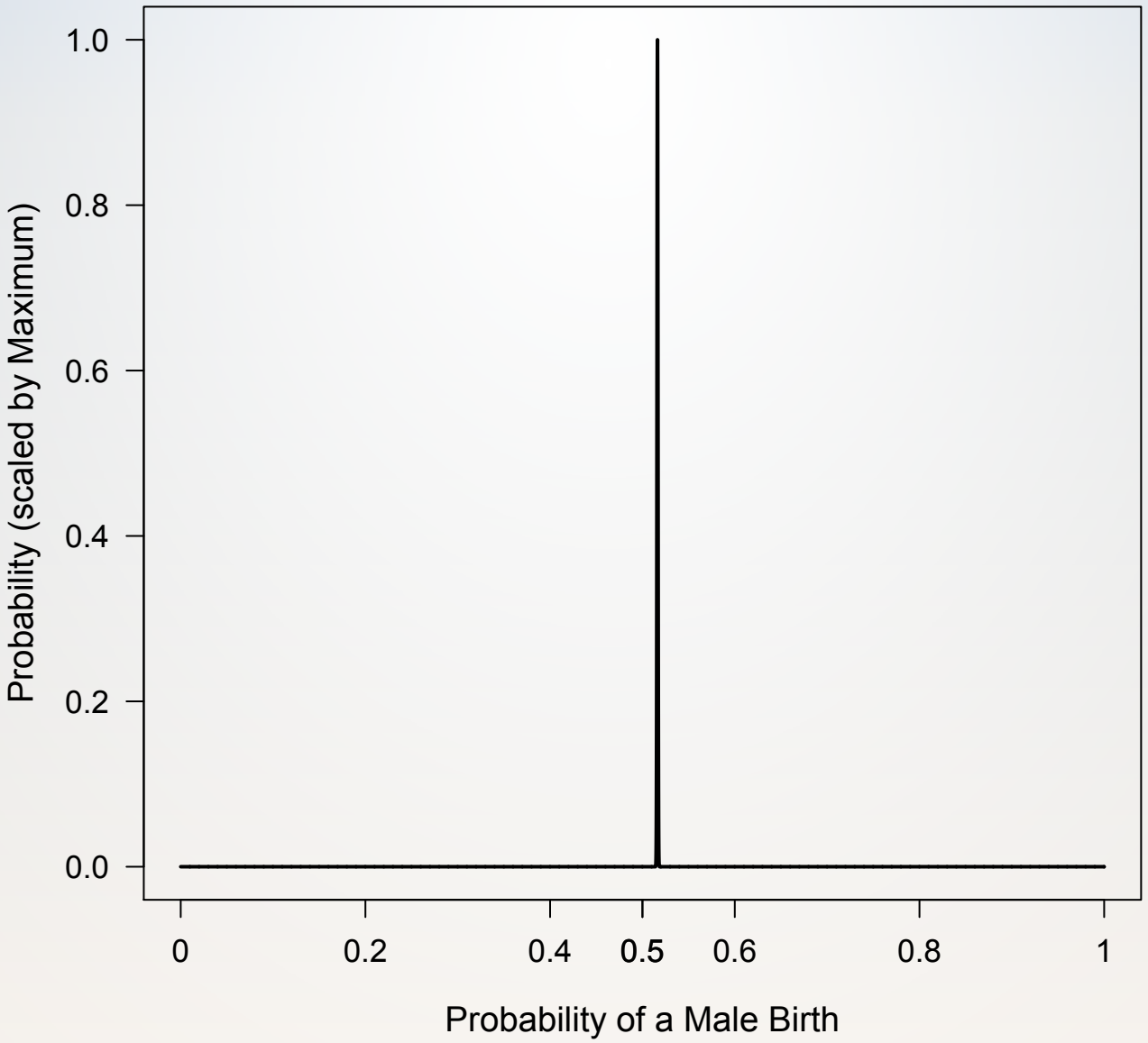


Inferentially Agnostic: Works with confidence, credible, and support intervals

Second-generation p-value

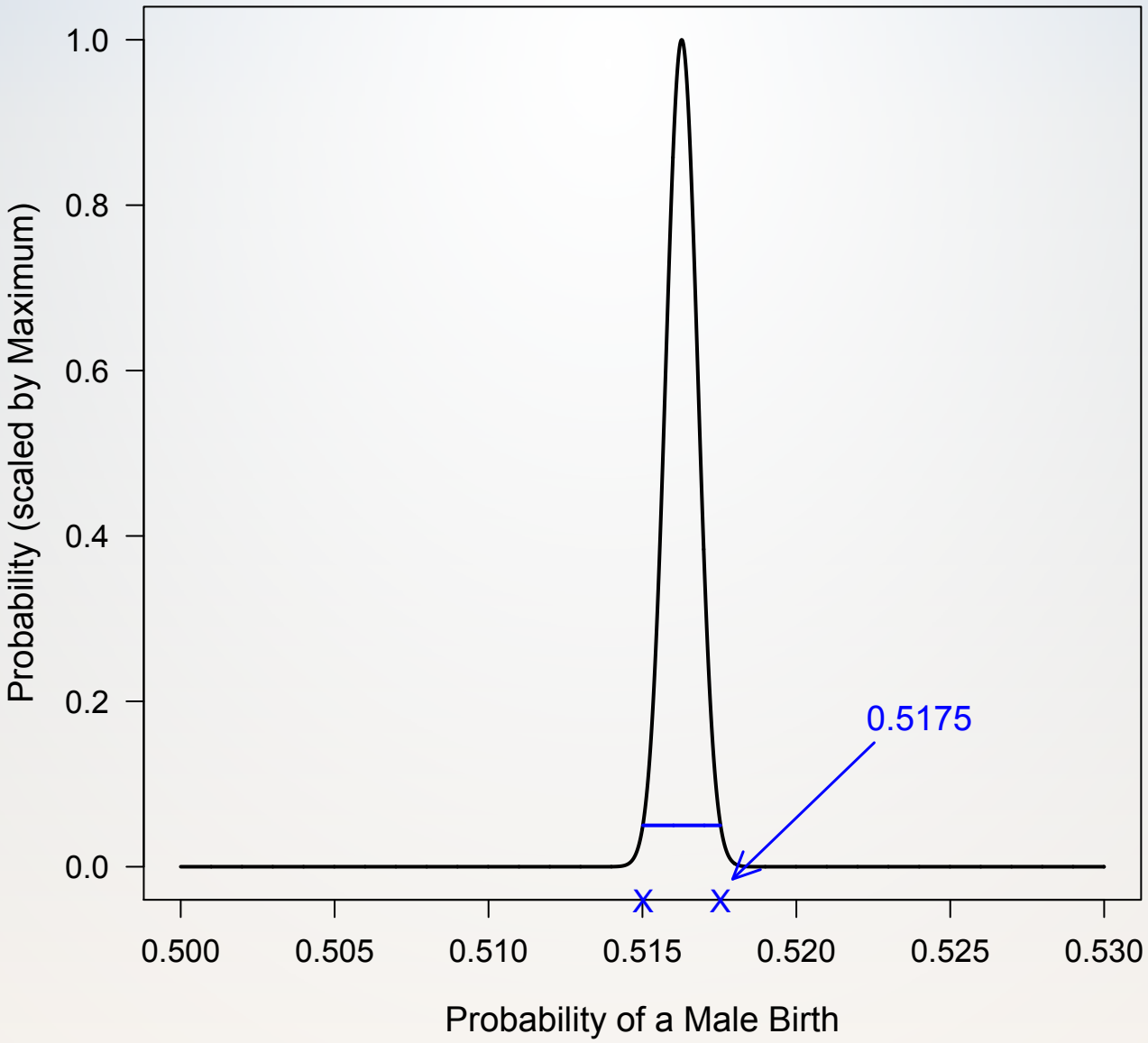
- StatisticalEvidence.com
- Formalizes the use of 'Interval Inference'
- Retains strict error control

Evidential Metric	What it measures	SGPV
1	Summary measure	SGPV (p_δ)
2	Operating characteristics	$P(p_\delta = 0 \mid H_0)$ $P(p_\delta = 1 \mid H_1)$ $P(0 < p_\delta < 1 \mid H)$
3	False discovery rates	$P(H_0 \mid p_\delta = 0)$ $P(H_1 \mid p_\delta = 1)$



Spinoza

Blume, Nov 2019



Spinoza

Blume, Nov 2019

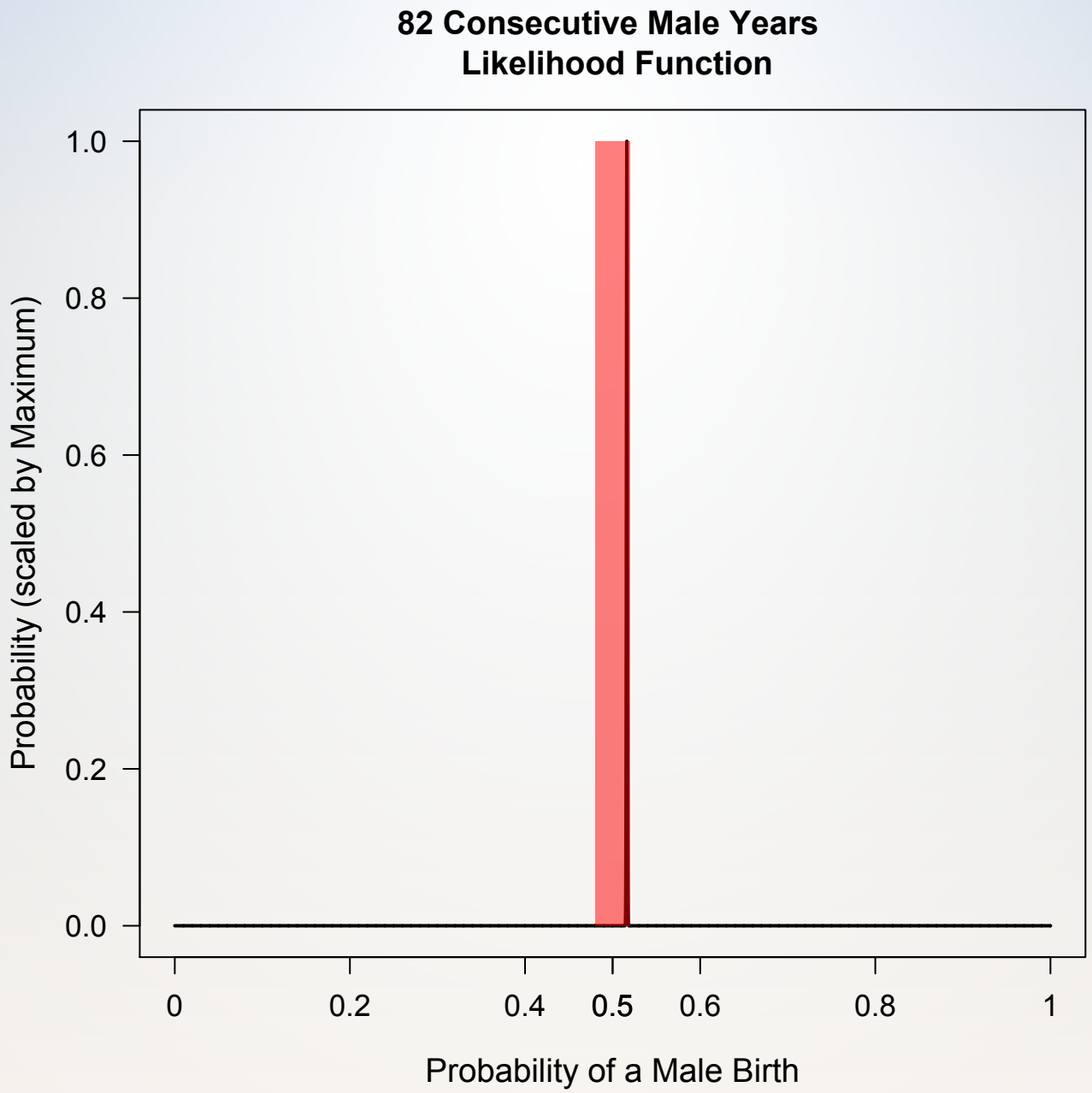
Gamblers Take Note: The Odds in a Coin Flip Aren't Quite 50/50

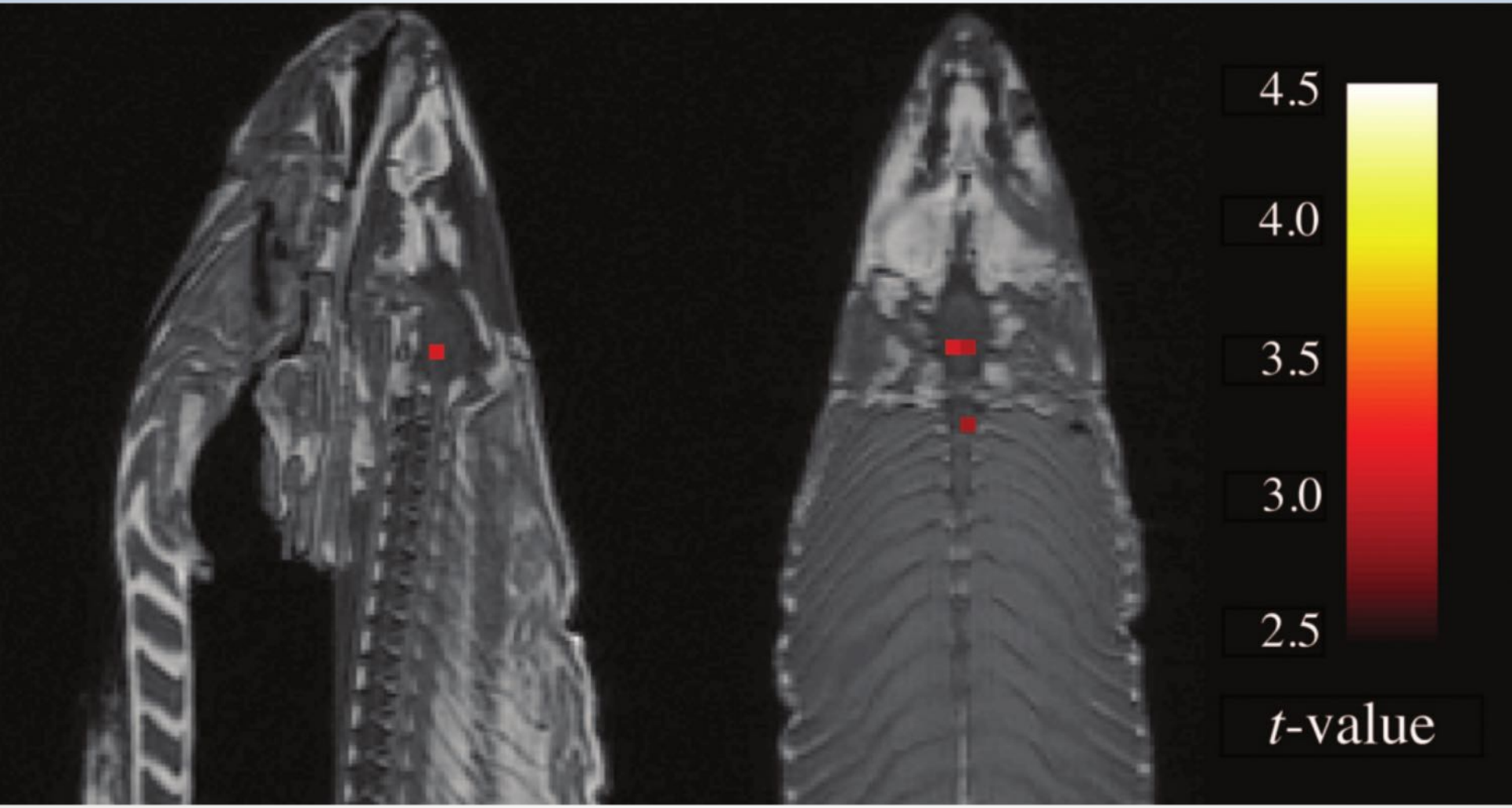
And the odds of spinning a penny are even more skewed in one direction, but which way?

Fair bet? Not if Persi Diaconis is right.

Diaconis is a professor of mathematics and statistics at Stanford University and, formerly, a professional magician. While his claim to fame is determining how many times a deck of cards must be shuffled in order to give a mathematically random result (it's either five or seven, depending on your criteria), he's also dabbled in the world of coin games. What he and his fellow researchers discovered (here's a PDF of their paper) is that most games of chance involving coins aren't as even as you'd think. For example, even the 50/50 coin toss really isn't 50/50 — it's closer to 51/49, biased toward whatever side was up when the coin was thrown into the air.

But more incredibly, as reported by Science News, spinning a penny, in this case one with the Lincoln Memorial on the back, gives even more pronounced odds — the penny will land tails side up roughly 80 percent of the time. The reason: the side with Lincoln's head on it is a bit heavier than the flip side, causing the coin's center of mass to lie slightly toward heads. The spinning coin tends to fall toward the heavier side more often, leading to a pronounced number of extra "tails" results when it finally comes to rest.


82 Consecutive Male Years
Likelihood Function
Probability (scaled by Maximum)
1.0
0.8
0.6
0.4
0.2
0.0
0
0.2
0.4
0.5
0.6
0.8
1
Probability of a Male Birth

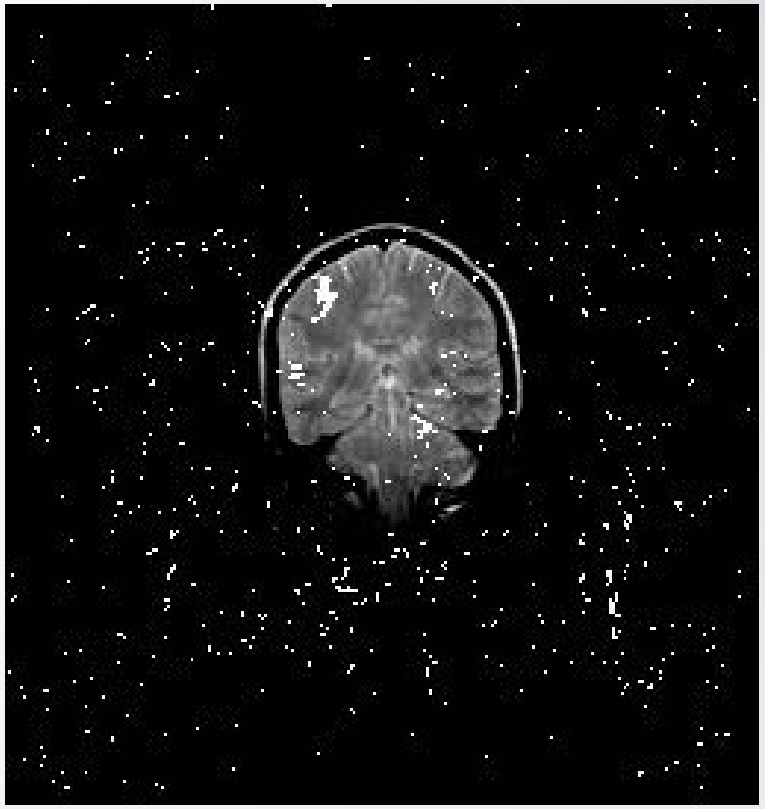


"Sagittal and axial images; t(131) > 3.15, p(uncorrected) < 0.001, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster…in the medial brain cavity and another…in the upper spinal column."

From Bennett et. al., 2010, JSUR 1:1 1-5. ***8064 total voxels; 16 identified.***

Setting interval null

- Before analyzing data (!)
- Measurement error
- Subject matter knowledge
- Impact of findings
- Community standard
- Get creative (fMR example)
- Width not critical, buffer



- Incorporating the scientific scale is the next step in improving testing

Thank you for your attention.

Questions?

www.statisticalevidence.com