Quantitative Aspects of Simpson's Paradox

Branden Fitelson & Vincenzo Crupi
April 21, 2021 (Draft)

Abstract

Simpson's Paradox has received a lot of attention in the contemporary literature. Typically, these presentations focus only on the qualitative structure of the phenomenon, and various explanations of its "paradoxicality" [1, 2]. In this paper, we discuss quantitative aspects of Simpson's Paradox, *via* the use of various Bayesian measures of degree of confirmation. This leads to some interesting new results, both for the general phenomenon of Simpson's Paradox and for Bayesian confirmation theory.

1 Simpson's Paradox & Bayesian Confirmation (Qualitative)

Simpson's Paradox (in its usual, qualitative form) can be stated purely in Bayesian confirmation-theoretic terms [2]. Let H, E, and K be propositions, and let $\Pr(\cdot)$ be some ("prior") probability function that is being used to assess relations of evidential support involving these three propositions (*e.g.*, $\Pr(\cdot)$ might reflect statistical probabilities in some experimental setup). We can define the usual, qualitative, Bayesian confirmation relation as follows.

> **Confirmation** (qualitative). E confirms H, given K iff $\Pr(H \mid E \,\&\, K) > \Pr(H \mid K)$.[1]

If $\Pr(H \mid E \,\&\, K) < \Pr(H \mid K)$, then we say E *dis*confirms H, given K. And, if $\Pr(H \mid E \,\&\, K) = \Pr(H \mid K)$, then we say that E is *confirmationally irrelevant* to H, given K. If K is tautological (*i.e.*, if $K = \top$), then we will drop the "given K" and say simply that "E confirms (or disconfirms or is irrelevant to) H, unconditionally." With these concepts in hand, we can now define (qualitative) Simpson's Paradox, as follows.

> **Simpson's Paradox** (qualitative). Any example in which the following three (qualitative) confirmation relations obtain is an example of (qualitative) Simpson's Paradox.[2]
>
> 1. E confirms H, given K.
> 2. E confirms H, given $\neg K$.
> 3. E disconfirms H, unconditionally (*i.e.*, E disconfirms H, given $\top$).

Here is a toy example of a qualitative Simpson's Paradox.[3] Suppose a graduate school has two departments: K (history) and $\neg K$ (physics) and two genders of applicants: E (female) and $\neg E$ (male). And, let H ($\neg H$) express the proposition than an applicant is *accepted* (*not* accepted) to the graduate school. Table 1 gives all the relevant *acceptance rates* — encoded by the probability function $\Pr(\cdot)$ — regarding the admissions process for the graduate school in question for a given year (which involves a total of 260 applicants, 130 of which applied to history, and 130 of which applied to physics).

[1] Strictly speaking, the confirmation relation is a *four*-place relation, also involving the probability function $\Pr(\cdot)$. For simplicity, we will suppress this relativity to $\Pr(\cdot)$. Since most of our theoretical results will apply to *any* probability function, this simplification will be mostly harmless.

[2] Simpson's Paradox (in its most general sense) can also involve cases in which we go from disconfirmation — given each of K and $\neg K$ — to confirmation, unconditionally; or, cases in which we go from irrelevance — given each of K and $\neg K$ — to relevance, unconditionally; or, cases in which we go from confirmation (or disconfirmation) — given each of K and $\neg K$ — to irrelevance, unconditionally. In this paper, we will focus on Simpson reversals that go from confirmation (conditionally) to disconfirmation (unconditionally). Similar things can be said for these other kinds of "Simpson reversal" (broadly construed). Moreover, Simpson's paradox need not be limited to *dichotomous* random variables. Our results can be generalized to discrete random variables with any finite number of values. For simplicity, we'll focus on the 2×2 case.

[3] This example is based loosely on an example involving the Berkeley graduate school, which was made famous by Nancy Cartwright [3]. That example involved 6 departments instead of 2. Since we are focusing here on the 2×2 case, we have simplified the case considerably.

	E	$\neg E$	Overall
K	$^{40}/_{50} = \Pr(H \mid E \,\&\, K)$	$^{60}/_{80} = \Pr(H \mid \neg E \,\&\, K)$	$^{100}/_{130} = \Pr(H \mid K)$
$\neg K$	$^{20}/_{80} = \Pr(H \mid E \,\&\, \neg K)$	$^{10}/_{50} = \Pr(H \mid \neg E \,\&\, \neg K)$	$^{30}/_{130} = \Pr(H \mid \neg K)$
Overall	$^{60}/_{130} = \Pr(H \mid E)$	$^{70}/_{130} = \Pr(H \mid \neg E)$	$^{130}/_{260} = \Pr(H)$

Table 1: Probabilistic structure of a toy, qualitative Simpson's Paradox

These acceptance rates are to be understood as follows. For instance, $\Pr(H \mid E \,\&\, K) = {}^{40}/_{50}$ means that the probability of acceptance for female applicants to the history department is ${}^{40}/_{50}$ — *i.e.*, ${}^{40}/_{50}$ of the female history applicants were accepted. The crucial thing to notice here is that the following three conditions (constitutive of a qualitative Simpson's Paradox) obtain.

1. E confirms H, given K. That is: $\Pr(H \mid E \,\&\, K) = {}^{40}/_{50} > {}^{100}/_{130} = \Pr(H \mid K)$.

2. E confirms H, given $\neg K$. That is: $\Pr(H \mid E \,\&\, \neg K) = {}^{20}/_{80} > {}^{30}/_{130} = \Pr(H \mid \neg K)$.

3. E disconfirms H, unconditionally. That is: $\Pr(H \mid E) = {}^{60}/_{130} < {}^{130}/_{260} = \Pr(H)$.

In words: being female was positively correlated with acceptance in each department, but negatively correlated with acceptance in the overall population of applicants. The standard explanation of how such a case might arise is that (in the year in question) there happens to be a correlation (but not any causal connection) between being female (E) and applying to the department (physics) with the lower acceptance rate ($\neg K$). Before turning to quantitative generalizations of Simpson's Paradox, it is useful to look at one real-world example of the phenomenon.

Our real-world example comes from a medical study comparing success rates of two treatments for kidney stones [4]. Table 2 shows the success rates and numbers of treatments for cases involving both small (K) and large kidney ($\neg K$) stones, where treatment E includes open surgical procedures and treatment $\neg E$ includes closed surgical procedures. That is, *e.g.*, $\Pr(H \mid E \& K) = {}^{81}/_{87}$ means 81 out of 87 patients who received open surgical procedures for small kidney stones had successful procedures.

	E	$\neg E$	Overall
K	$^{81}/_{87} = \Pr(H \mid E \,\&\, K)$	$^{234}/_{270} = \Pr(H \mid \neg E \,\&\, K)$	$^{315}/_{357} = \Pr(H \mid K)$
$\neg K$	$^{192}/_{263} = \Pr(H \mid E \,\&\, \neg K)$	$^{55}/_{80} = \Pr(H \mid \neg E \,\&\, \neg K)$	$^{247}/_{343} = \Pr(H \mid \neg K)$
Overall	$^{273}/_{350} = \Pr(H \mid E)$	$^{289}/_{350} = \Pr(H \mid \neg E)$	$^{562}/_{700} = \Pr(H)$

Table 2: Probabilistic structure of a real-world, qualitative Simpson's Paradox

In this real-world example, we have the following three facts:

1. E confirms H, given K. That is: $\Pr(H \mid E \,\&\, K) = {}^{81}/_{87} > {}^{315}/_{357} = \Pr(H \mid K)$.

2. E confirms H, given $\neg K$. That is: $\Pr(H \mid E \,\&\, \neg K) = {}^{192}/_{263} > {}^{247}/_{343} = \Pr(H \mid \neg K)$.

3. E disconfirms H, unconditionally. That is: $\Pr(H \mid E) = {}^{273}/_{350} < {}^{562}/_{700} = \Pr(H)$.

In words: open surgical procedures were more effective among both small stone and large stone patients, but closed surgical procedures were more effective in the overall population of patients. The standard explanation of how this actually happened is that the less effective treatment ($\neg E$) is applied more frequently to small stone cases, which makes it appear to be a more effective treatment (*i.e.*, what we have here is suppression of the causal effect of the size of the stones on successful treatment). Presently, we are not concerned with analyses of (qualitative) Simpson's Paradox. Rather, our discussion will focus on quantitative generalizations of Simpson's Paradox, to which we now turn.

2 Simpson's Paradox & Bayesian Confirmation (Quantitative)

Bayesian confirmation theory allows not only for *qualitative* judgments regarding confirmation, disconfirmation, and irrelevance, but also for *quantitative* assessments of *degree of confirmation*. This is achieved *via* the use of Bayesian measures $\mathfrak{c}(H, E \mid K)$ of *the degree to which E confirms H, given K*. Following [5], we will be comparing Bayesian confirmation measures that are defined on a $[-1, 1]$ scale. More precisely, all the Bayesian measures of confirmation we discuss here will satisfy the following desideratum.

> **Confirmation** (quantitative). All Bayesian measures $\mathfrak{c}(H, E \mid K)$ of the degree to which E confirms H, given K should be such that
>
> $$\mathfrak{c}(H, E \mid K) \text{ is } \begin{cases} = +1 & \text{if } E \textit{ maximally confirms } H, \text{ given } K, \text{ according to } \mathfrak{c}. \\ > 0 \text{ (confirmation)} & \text{if } E \text{ confirms } H, \text{ given } K. \\ = 0 \text{ (irrelevance)} & \text{if } E \text{ is confirmationally irrelevant to } H, \text{ given } K. \\ < 0 \text{ (disconfirmation)} & \text{if } E \text{ disconfirms } H, \text{ given } K. \\ = -1 & \text{if } E \textit{ maximally disconfirms } H, \text{ given } K, \text{ according to } \mathfrak{c}. \end{cases}$$

Specifically, we will be comparing and contrasting the following five well-known measures [5, p. 233].

$$d(H, E \mid K) \stackrel{\text{def}}{=} \Pr(H \mid E \,\&\, K) - \Pr(H \mid K)$$

$$r(H, E \mid K) \stackrel{\text{def}}{=} \frac{\Pr(H \mid E \,\&\, K) - \Pr(H \mid K)}{\Pr(H \mid E \,\&\, K) + \Pr(H \mid K)}$$

$$l(H, E \mid K) \stackrel{\text{def}}{=} \frac{\Pr(E \mid H \,\&\, K) - \Pr(E \mid \neg H \,\&\, K)}{\Pr(E \mid H \,\&\, K) + \Pr(E \mid \neg H \,\&\, K)}$$

$$s(H, E \mid K) \stackrel{\text{def}}{=} \Pr(H \mid E \,\&\, K) - \Pr(H \mid \neg E \,\&\, K)$$

$$z(H, E \mid K) \stackrel{\text{def}}{=} \begin{cases} \dfrac{\Pr(H \mid E \,\&\, K) - \Pr(H \mid K)}{1 - \Pr(H \mid K)} & \text{if } \Pr(H \mid E \,\&\, K) \geq \Pr(H \mid K) \\ \dfrac{\Pr(H \mid E \,\&\, K) - \Pr(H \mid K)}{\Pr(H \mid K)} & \text{if } \Pr(H \mid E \,\&\, K) < \Pr(H \mid K) \end{cases}$$

With this confirmation-theoretic background in place, we can now define a *quantitative generalization* of Simpson's Paradox. Here is our proposed generalization.

> **Simpson's Paradox** (quantitative). Any example in which the following three (quantitative) confirmation relations obtain is a **Quantitative Simpson's Paradox of strength** $t > 0$ **for measure** $\mathfrak{c}$ ($\text{QSP}^t_{\mathfrak{c}}$).
>
> 1. $\mathfrak{c}(H, E \mid K) \geq t$.
> 2. $\mathfrak{c}(H, E \mid \neg K) \geq t$.
> 3. $\mathfrak{c}(H, E \mid \top) \leq -t$.

The basic idea behind ($\text{QSP}^t_{\mathfrak{c}}$) is that it involves case in which we have not only a *qualitative* reversal (from conditional confirmation to unconditional disconfirmation), but also a *quantitative* reversal — *of a particular strength* $t > 0$ — according to a confirmation measure $\mathfrak{c}$.

In order to illustrate how ($\text{QSP}^t_{\mathfrak{c}}$) works, we may return to our toy (admissions) example, above. The following three quantitative facts obtain in our toy example — for the d-measure.

1. $d(H,E \mid K) = \Pr(H \mid E \,\&\, K) - \Pr(H \mid K) = {}^{40}/_{50} - {}^{100}/_{130} = {}^{2}/_{65} \approx 0.031$.

2. $d(H,E \mid \neg K) = \Pr(H \mid E \,\&\, \neg K) - \Pr(H \mid \neg K) = {}^{20}/_{80} - {}^{30}/_{130} = {}^{1}/_{52} \approx 0.019$.

3. $d(H,E \mid \top) = \Pr(H \mid E) - \Pr(H) = {}^{60}/_{130} - {}^{130}/_{260} = -{}^{1}/_{26} \approx -0.038$.

Thus, our toy example constitutes an instance of $(\mathrm{QSP}_d^{0.019})$. Similar calculations reveal that our toy example constitutes instances of $(\mathrm{QSP}_r^{0.019})$, $(\mathrm{QSP}_l^{0.052})$, $(\mathrm{QSP}_s^{0.05})$, and $(\mathrm{QSP}_z^{0.025})$, respectively. That is, the reversal in our toy example is *rather small* (from a quantitative point of view), according to all five of our measures. Something similar happens in our real-world example. There, we have:

1. $d(H,E \mid K) = \Pr(H \mid E \,\&\, K) - \Pr(H \mid K) = {}^{81}/_{87} - {}^{315}/_{357} = {}^{24}/_{393} \approx 0.049$.

2. $d(H,E \mid \neg K) = \Pr(H \mid E \,\&\, \neg K) - \Pr(H \mid \neg K) = {}^{192}/_{263} - {}^{247}/_{343} = {}^{895}/_{90209} \approx 0.009$.

3. $d(H,E \mid \top) = \Pr(H \mid E) - \Pr(H) = {}^{273}/_{350} - {}^{562}/_{700} = -{}^{4}/_{175} \approx -0.023$.

Thus, our real-world example constitutes an instance of $(\mathrm{QSP}_d^{0.009})$. Similar calculations reveal that our real-world example constitutes instances of $(\mathrm{QSP}_r^{0.0068})$, $(\mathrm{QSP}_l^{0.024})$, $(\mathrm{QSP}_s^{0.042})$, and $(\mathrm{QSP}_z^{0.028})$, respectively. Interestingly, all of the concrete numerical examples of Simpson's Paradox that we have seen in the literature involve weak/small quantitative reversals (according to each of our five measures).

This raises an interesting theoretical question. Is there an *upper-bound* on *how strong* a quantitative Simpson reversal can be — *in principle* — according to each of our measures? More precisely, for each measure $\mathfrak{c}$, we may ask whether there exists a $\tau \in (0,1]$ such that no cases of $(\mathrm{QSP}_{\mathfrak{c}}^{t})$ are possible, for any $t > \tau$. Trivially, $\tau = 1$ will be one such value (since that's the maximum value possible for each of our measures). If $\tau = 1$ is the smallest such upper bound for a measure $\mathfrak{c}$, then we will say that $\mathfrak{c}$ permits quantitative Simpson's Paradoxes *of arbitrary strength*. Interestingly, while some of our five measures permit QSP's of arbitrary strength, some do not. Specifically, we have the following main result.[4].

Theorem. Regarding maximum possible strength of $(\mathrm{QSP}_{\mathfrak{c}}^{t})$ for each of our five measures, we have:

(1) Measures l and r permit quantitative Simpson's Paradoxes of arbitrary strength (*viz.*, we have $\tau = 1$ for measures l and r). That is, t can be made arbitrarily close to 1 (but must remain less than 1) in cases of (QSP_l^t) and (QSP_r^t).

(2) Measures d, s, and z do *not* permit quantitative Simpson's Paradoxes of arbitrary strength. More precisely, we have the following upper-bounds $t = \tau$ for (QSP_d^t), (QSP_s^t), and (QSP_z^t).

(2.1) For z, we have $\tau = {}^{1}/_{2}$. That is, t can be made arbitrarily close to ${}^{1}/_{2}$ (but must remain less than ${}^{1}/_{2}$) in cases of (QSP_z^t).

(2.2) For d and s, we have $\tau = {}^{1}/_{3}$. That is, t can be made arbitrarily close to ${}^{1}/_{3}$ (but must remain less than ${}^{1}/_{3}$) in cases of (QSP_d^t) and (QSP_s^t).

Proof. To establish (1), it suffices to specify (1_l) a family of probability models such that t can be made arbitrarily close to 1, while maintaining (QSP_l^t); and, (1_r) a family of probability models such that t can be made arbitrarily close to 1, while maintaining (QSP_r^t). See the companion *Mathematica* notebook for this paper (*fn.* 4), which explains how to to use `PrSAT` to construct (1_l) and (1_r).

To establish (2.1), we must show two things. First, we must show that (2.1.1) whenever $z(H,E \mid K) \geq {}^{1}/_{2}$ and $z(H,E \mid \neg K) \geq {}^{1}/_{2}$, it follows that $z(H,E \mid \top) > -{}^{1}/_{2}$. Second, we must specify (2.1.2) a family

[4]We have created a companion *Mathematica* notebook for this paper, which uses the `PrSAT` package [6] to verify all of the technical claims in this paper (including the claims about examples, the theoretical results, and the computer simulations). That notebook can be downloaded from the following URL: `http://fitelson.org/qasp.nb`.

of probability models such that t can be made arbitrarily close to $1/2$, while maintaining (QSP^t_z). See the companion *Mathematica* notebook, which explains how to to use `PrSAT` to construct (2.1.2).

To establish (2.2) for measure d, we must show two things. First, we must show that ($2.2.1_d$) whenever $d(H, E \mid K) \geq 1/3$ and $d(H, E \mid \neg K) \geq 1/3$, it follows that $d(H, E \mid \top) > -1/3$. Second, we must specify ($2.2.2_d$) a family of probability models such that t can be made arbitrarily close to $1/3$, while maintaining (QSP^t_d). See the companion *Mathematica* notebook, which explains how to to use `PrSAT` to construct ($2.2.2_d$).

To establish (2.2) for measure s, we must show two things. First, we must show that ($2.2.1_s$) whenever $s(H, E \mid K) \geq 1/3$ and $s(H, E \mid \neg K) \geq 1/3$, it follows that $s(H, E \mid \top) > -1/3$. Second, we must specify ($2.2.2_s$) a family of probability models such that t can be made arbitrarily close to $1/3$, while maintaining (QSP^t_s). See the companion *Mathematica* notebook, which explains how to to use `PrSAT` to construct ($2.2.2_s$).

Our proofs of the three remaining parts of our **Theorem**: (2.1.1), ($2.2.1_d$), and ($2.2.1_s$) rely on the following Lemma.

Lemma. Measures $\mathfrak{c} \in \{d, s, z\}$ are such that *if* (a) $\mathfrak{c}(x, y \mid u) \geq 0$, *then* (b) $\Pr(x \mid y \,\&\, u) \geq \mathfrak{c}(x, y \mid u)$.

Proof. (d) Suppose, for *reductio*, that (b) $\Pr(x \mid y \,\&\, u) < d(x, y \mid u)$. Then

$$\begin{gathered} \Pr(x \mid y \,\&\, u) < \Pr(x \mid y \,\&\, u) - \Pr(x \mid u) \\ \therefore 0 < -\Pr(x \mid u) \end{gathered}$$

which is impossible since $\Pr(x \mid u) \in [0, 1]$.

(s) Suppose, for *reductio*, that (b) $\Pr(x \mid y \,\&\, u) < s(x, y \mid u)$. Then

$$\begin{gathered} \Pr(x \mid y \,\&\, u) < \Pr(x \mid y \,\&\, u) - \Pr(x \mid \neg y \,\&\, u) \\ \therefore 0 < -\Pr(x \mid \neg y \,\&\, u) \end{gathered}$$

which is impossible since $\Pr(x \mid \neg y \,\&\, u) \in [0, 1]$.

(z) Suppose, for *reductio*, that (a) $z(x, y \mid u) \geq 0$; and, (b) $\Pr(x \mid y \,\&\, u) < z(x, y \mid u)$. Then

$$\begin{aligned} \Pr(x \mid y \,\&\, u) &< \frac{\Pr(x \mid y \,\&\, u) - \Pr(x \mid u)}{1 - \Pr(x \mid u)} \\ \therefore \Pr(x \mid y \,\&\, u) - \Pr(x \mid y \,\&\, u) \cdot \Pr(x \mid u) &< \Pr(x \mid y \,\&\, u) - \Pr(x \mid u) \\ \therefore \Pr(x \mid y \,\&\, u) &< \Pr(x \mid u) \cdot \left[\Pr(x \mid y \,\&\, u) - 1\right] + \Pr(x \mid y \,\&\, u) \\ \therefore 0 &< \Pr(x \mid u) \cdot \left[\Pr(x \mid y \,\&\, u) - 1\right] \end{aligned}$$

which is impossible since $\Pr(x \mid y \,\&\, u), \Pr(x \mid u) \in [0, 1]$. ✦

With our **Lemma** in hand, we may now prove (2.1.1), ($2.2.1_d$), and ($2.2.1_s$).

(2.1.1) Suppose that (i) $z(H, E \mid K) \geq 1/2$ and (ii) $z(H, E \mid \neg K) \geq 1/2$. Then, by our Lemma, $\Pr(H \mid E \,\&\, K) \geq 1/2$ and $\Pr(H \mid E \,\&\, \neg K) \geq 1/2$. If $\Pr(H \mid E \,\&\, K) = \Pr(H \mid E \,\&\, K) = 1/2$, then $\Pr(H \mid E) = 1/2$, which also implies $\Pr(H) < 1$. So in this case either $\Pr(H \mid E) \geq \Pr(H)$ and thus $z(H, E) \geq 0 > -1/2$, or $1 > \frac{Pr(H \mid E)}{\Pr(H)} > 1/2$ and $z(H, E) = \frac{\Pr(H \mid E)}{\Pr(H)} - 1 > 1/2 - 1$, thus $z(H, E) \geq 0 > -1/2$ once again. Otherwise, we have $\Pr(H \mid E \,\&\, K) \geq 1/2$ and $\Pr(H \mid E \,\&\, \neg K) \geq 1/2$ with at least one inequality strict. For $\Pr(\cdot \mid E \,\&\, K)$ and $\Pr(\cdot \mid E)$ to be defined, however, we must also have $\Pr(E \,\&\, K) > 0$ and $\Pr(E) > 0$, so that $\Pr(K \mid E) > 0$, and $\Pr(H \mid E)$ must lie strictly between $\Pr(H \mid E \,\&\, K)$ and $Pr(H \mid E \,\&\, \neg K)$ by the probability calculus. As a consequence, $\Pr(H \mid E) > 1/2$ and $\Pr(\neg H \mid E) < 1/2$. In this case, too, either $\Pr(H \mid E) \geq \Pr(H)$ and

thus $z(H,E) \geq 0 > -1/2$, or $z(\neg H,E) > 0$ but $z(\neg H,E) \leq \Pr(\neg H \mid E) < 1/2$, as implied by our Lemma, so that $z(H,E) = -z(\neg H,E) > -1/2$ once again.

($2.2.1_d$) Suppose that (i) $d(H,E \mid K) = \Pr(H \mid E \,\&\, K) - \Pr(H \mid K) \geq 1/3$ and (ii) $d(H,E \mid \neg K) = \Pr(H \mid E \,\&\, \neg K) - \Pr(H \mid \neg K) \geq 1/3$. If $\Pr(H \mid E \& K) = \Pr(H \mid E \& \neg K) = 1$, then $\Pr(H \mid E) = 1$ too, so $d(H,E) \geq 0$ and, trivially, $d(H,E) > -1/3$. If on the other hand either $\Pr(H \mid E \,\&\, K) < 1$ or $\Pr(H \mid E \,\&\, \neg K) < 1$, then (i) and (ii) imply that $\Pr(H \mid K) \leq 2/3$ and $\Pr(H \mid \neg K) \leq 2/3$ with at least one inequality strict. For $\Pr(\cdot \mid K)$ to be defined, moreover, we must have $\Pr(K) > 0$, and so $\Pr(H)$ must lie strictly between $\Pr(H \mid K)$ and $\Pr(H \mid \neg K)$ by the probability calculus. As a consequence, $\Pr(H) < 2/3$. Also, $\Pr(H \mid E \,\&\, K) \geq 1/3$ and $\Pr(H \mid E \,\&\, \neg K) \geq 1/3$, again from (i) and (ii) via our Lemma, so that $\Pr(H \mid E) \geq 1/3$ too, by the probability calculus. But then $d(H,E) = \Pr(H \mid E) - \Pr(H) > 1/3 - 2/3 = -1/3$.

($2.2.1_s$) Suppose that (i) $s(H,E \mid K) = \Pr(H \mid E \,\&\, K) - \Pr(H \mid \neg E \,\&\, K) \geq 1/3$ and (ii) $s(H,E \mid \neg K) = \Pr(H \mid E \,\&\, \neg K) - \Pr(H \mid \neg E \,\&\, \neg K) \geq 1/3$. If $\Pr(H \mid E \,\&\, K) = \Pr(H \mid E \,\&\, \neg K) = 1$, then $\Pr(H \mid E) = 1$ too, so $s(H,E) \geq 0$ and, trivially, $s(H,E) > -1/3$. If on the other hand either $\Pr(H \mid E \,\&\, K) < 1$ or $\Pr(H \mid E \,\&\, \neg K) < 1$, then (i) and (ii) imply that $\Pr(H \mid \neg E \,\&\, K) \leq 2/3$ and $\Pr(H \mid \neg E \,\&\, \neg K) \leq 2/3$ with at least one inequality strict. For $\Pr(\cdot \mid \neg E \,\&\, K)$ and $\Pr(\cdot \mid \neg E)$ to be defined, moreover, we must have $\Pr(\neg E \,\&\, K) > 0$ and $\Pr(\neg E) > 0$, so that $\Pr(K \mid \neg E) > 0$ and $\Pr(H \mid \neg E)$ must lie strictly between $\Pr(H \mid \neg E \,\&\, K)$ and $Pr(H \mid \neg E \,\&\, \neg K)$ by the probability calculus. As a consequence, $\Pr(H \mid \neg E) < 2/3$. Also, $\Pr(H \mid E \,\&\, K) \geq 1/3$ and $\Pr(H \mid E \,\&\, \neg K) \geq 1/3$, again from (i) and (ii) via our Lemma, so that $\Pr(H \mid E) \geq 1/3$ too, by the probability calculus. But then $s(H,E) = \Pr(H \mid E) - \Pr(H \mid \neg E) > 1/3 - 2/3 = -1/3$. ✦

3 Estimating the Prevalence and Strength of Simpson's Paradoxes

Although some measures *theoretically* allow for arbitrarily strong quantitative Simpson's Paradoxes (while others have strong theoretical upper-bounds on possible reversal strength), actual cases of Simpson's Paradox that are observed "in the wild" tend to be very weak, according to each of our measures. This raises questions about the prevalence and strength distribution of Simpson's Paradoxes. We have performed computer simulations that shed light on both of these questions.

We sampled 10 million probability distributions over H, E, K at random (*i.e.*, assuming a uniform distribution over possible probability distributions). Our simulations revealed some very interesting patterns.

First, we approximated the probability of obtaining a qualitative Simpson's Paradox (assuming a uniform distribution over possible distributions). It turns out that this probability is approximately 0.0083 (approximately 83,000 out of our 10 million sampled distributions exhibited Simpson's Paradox). That is, approximately 1 out of every 125 probability distributions can be expected to exhibit a qualitative Simpson's Paradox. This suggests that Simpson's Paradox is somewhat rare, but not astronomically so (which partly explains why so many empirical cases have been observed, historically).

Second, we computed the average strength of the Simpson reversals, according to each of the five measures. These averages were approximately 0.013 for d, 0.04 for l, 0.019 for r, 0.035 for s, and 0.025 for z. Note that the none of the measures records an average strength of Simpson reversal exceeding 0.04.

Finally, we computed *strength histograms* for each of our five measures. That is, we calculated the *distribution* of strengths t, for each of the (approximately 83,000) cases of Simpson's Paradox (QSP^t_c) that were observed in our 10 million randomly sampled distributions. It is clear from these histograms that the distribution of strengths t in cases of (QSP^t_c) is (roughly) *exponential* in nature — for each of our five measures. The Figure below shows these t-histograms for each of our five measures. It reveals that — even for measures (l and r) that theoretically allow for arbitrarily strong Simpson reversals — very strong reversals are exceedingly rare ("tail events"). We think this goes some way toward explaining why strong Simpson reversals are almost never observed in nature.

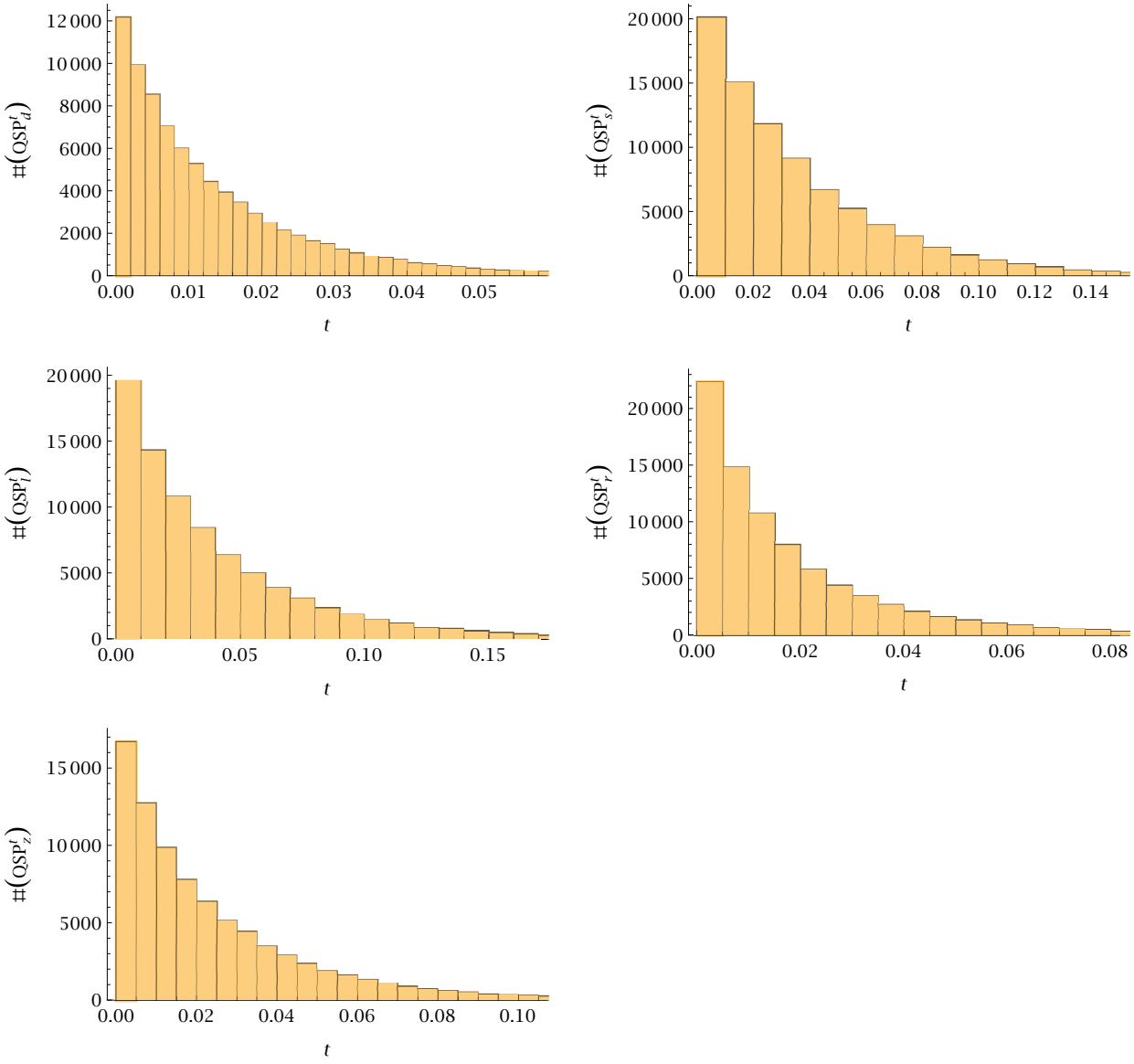

Figure: Histograms of (QSP_c^t) counts (from our ≈83,000 randomly sampled Simpson's Paradoxes)

References

[1] Judea Pearl. Comment: understanding simpson's paradox. *The American Statistician*, 68(1):8–13, 2014.

[2] Branden Fitelson. Confirmation, causation, and simpson's paradox. *Episteme*, 14(3):297–309, 2017.

[3] Nancy Cartwright. Causal laws and effective strategies. *Noûs*, pages 419–437, 1979.

[4] Steven A Julious and Mark A Mullee. Confounding and simpson's paradox. *British Medical Journal*, 309(6967):1480–1481, 1994.

[5] Vincenzo Crupi, Katya Tentori, and Michel Gonzalez. On bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, 74(2):229–252, 2007.

[6] Branden Fitelson. A decision procedure for probability calculus with applications. *Review of Symbolic Logic*, 1(1):111–125, 2008.