Journal of the Economic Science Association (2019) 5:255–268
https://doi.org/10.1007/s40881-019-00069-x

METHODOLOGY PAPER

True lies

Comment on Garbarino, Slonim and Villeval (2018)

David Hugh-Jones[1]

Received: 8 November 2018 / Revised: 30 May 2019 / Accepted: 3 June 2019 / Published online: 13 June 2019

Abstract
Garbarino et al. (J Econ Sci Assoc. https://doi.org/10.1007/s40881-018-0055-4, 2018) describe a new method to calculate the probability distribution of the proportion of lies told in "coin flip" style experiments. I show that their estimates and confidence intervals are flawed. I demonstrate two better ways to estimate the probability distribution of what we really care about—the proportion of liars—and I provide R software to do this.

Keywords Lying · Experiment · Estimation

JEL Classification C91 · C81 · D03

Some people are honest, while others are likely to lie whenever it benefits them. We would like to understand the prevalence of lying, because dishonesty may be economically and socially harmful. Since we cannot simply ask people if they are liars, one way to estimate the proportion of liars in a group is to ask them to report the result of a coin flip or other random device, offering them a payment if they report heads. Liars do not always lie: they only lie when it benefits them. So, they always report heads irrespective of the true coin flip.[1] If there are many more heads than we would expect by chance, we can assume many people are lying. But how many?

A naïve estimate would be that if, e.g., 80 people of 100 report heads, then on average 50 really saw heads and 60% (30/50) of the remainder are lying. More generally, from R reports of a good outcome in a sample of size N, where the bad

[1] Garbarino et al. (2018) maintain this assumption and so shall I.

✉ David Hugh-Jones
D.Hugh-Jones@uea.ac.uk

[1] University of East Anglia, Norwich, UK

Table 1 Parameter values

Parameter	Values
Sample size (N)	10, 50, 100, 500
Probability of lying (λ)	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
Probability of bad random outcome (P)	0.2, 0.5, 0.8

Table 2 Coverage levels for GSV and alternative methods

Method	CI 90%	CI 95%	CI 99%
GSV	65.2%	71.0%	78.9%
Frequentist	91.4%	94.4%	96.5%
Bayesian	91.3%	95.5%	99.1%

outcome happens with probability P, we can estimate that the following proportion are lying (Abeler et al. 2016):

$$\frac{R/N - (1-P)}{P}. \tag{1}$$

The problem with this approach is that the number of heads is not fixed. If we see 1 out of 3 people reporting heads, this method estimates there are less than zero liars. But it is still possible that everyone saw heads and 1 person lied.

Garbarino et al. (2018)—GSV from here on—point out this problem and introduce an alternative method. They claim that their method corrects for this problem and can estimate the full distribution of lying outcomes, and they recommend using it for confidence intervals, hypothesis testing and power calculations.

I ran simulations to check the overall performance of the GSV confidence intervals. Simulations parameters are shown in Table 1. λ is the probability that an individual in the sample lies and report heads when they observe tails:

$$\lambda = \frac{1}{N}\sum_{i=1}^{N} \text{Prob}(i \text{ reports heads} | i \text{ saw tails}). \tag{2}$$

For each parameter combination, I ran 1000 simulations, drawing random coin flips and reports given P, λ and N.

For each simulation and confidence level, I computed whether the GSV confidence interval contained the true value of λ. The first row of Table 2 shows the results.

By definition, 95% of 95% confidence intervals ought to contain the true value, on average. This is called "achieving nominal coverage". GSV confidence intervals are too narrow.[2]

To deal with this problem, I test two alternative methods for calculating confidence intervals on my simulated data. The first ("Frequentist") is the standard method of deriving confidence intervals from a binomial test. The second is a Bayesian method.

[2] This problem holds across all simulated probabilities of the low outcome, confidence levels, and sample sizes. See the "Appendix".

To understand the statistics, start with the probability of getting R reports of heads in total, given λ. Since individuals report heads either if they see heads with probability $1 - P$, or if they see tails but lie, with probability λP, this is just:

$$\Pr(R|\lambda; N, P) = \text{binom}(R, N, (1 - P) + \lambda P). \tag{3}$$

That immediately suggests the "Frequentist" method, which is to estimate the parameter of this distribution, $(1 - P) + \lambda P$, from the proportion of heads reported in the sample, then back out λ. This is the conventional method of, e.g., Abeler et al. (2016). It is justified if the sample is large, because this will lessen sampling variation in the proportion of actual heads observed. Similarly, if the sample is large enough, we can generate hypothesis tests for a value of λ—e.g., zero—using the tails of the binomial distribution. And we can back out confidence bounds for λ from confidence bounds for the population proportion of heads reported in the same way. As GSV point out, in small samples, this method runs the risk that the sample proportion of high outcomes will be different from its expected value.[3] We will see whether this matters.

There are numerous ways to calculate confidence intervals in a test of proportions. See, e.g., Agresti and Coull (1998). Here, I use the binomial exact test of Clopper and Pearson (1934), which is known to be conservative.

The second method uses Bayes' rule. Start with a prior probability density function over λ, $\varphi(\lambda)$. The posterior probability is then:

$$\varphi(\lambda|R; N, P) = \frac{\Pr(R|\lambda; N, P)\varphi(\lambda)}{\int_0^1 \Pr(R|\lambda'; N, P)\varphi(\lambda')\, \mathrm{d}\lambda'}. \tag{4}$$

From this, one can derive confidence intervals and expected values in the usual way. Technically, they are Bayesian "credible" intervals. I used highest posterior density intervals (Hyndman 1996), rather than the central confidence interval. This allows the intervals to include endpoints of the distribution, which is important when, e.g., testing for $\lambda = 0$.

The Bayesian method requires a prior. Here, I used a uniform prior, $\varphi(\lambda) = 1$ on $[0, 1]$.

Results in Table 2 show that both frequentist and Bayesian methods mostly achieve the nominal confidence level, with more than 90/95/99% of intervals containing the true value of λ. The exception is the frequentist 99% confidence interval, which is too narrow.

Frequentist confidence intervals could be less accurate when N is low, since that leads to more sampling variation in the number of true heads. Table 3 checks this by looking separately at simulations with $N = 10$ and $N = 50$. Frequentist 99% confidence intervals indeed appear slightly too narrow for this range. Bayesian confidence intervals are fine.

[3] In particular, this method can estimate confidence bounds for λ lower than 0. If so, we can set them to 0.

Table 3 Confidence interval coverage by sample size

Method	N	CI 90%	CI 95%	CI 99%
GSV	10	61.4%	65.7%	69.8%
	50	67.9%	73.3%	81.1%
Frequentist	10	94.1%	95.7%	96.8%
	50	91.6%	94.5%	96.6%
Bayesian	10	91.3%	95.3%	99.1%
	50	90.9%	95.5%	99.1%

1 Understanding the GSV approach

Why does the GSV method produce narrow confidence intervals? We can get a clue by running the GSV method when there are 10 reports of "heads" out of 10 for a fair coin flip ($R = N = 10, P = 0.5$). The resulting point estimate is that 100% of subjects lied. The upper and lower 99% confidence intervals are also 100%.

This is calculated as follows. First, given R reports of heads, the probability that a total of T "true" heads were observed is calculated as:

$$\text{Prob}(T \text{ heads}|R; N, P) = \frac{\text{binom}(T, N, 1-P)}{\sum_{k=0}^{R} \text{binom}(k, N, 1-P)}. \tag{5}$$

This is the binomial distribution, truncated at R because by assumption, nobody "lies downward" and reports tails when they really saw heads.

Next, from T the number of lies told is calculated as $R - T$; and the proportion of lies told is:

$$\text{Lies} = \frac{R-T}{N-T}, \tag{6}$$

because $N - T$ people saw the low outcome and had the chance to lie. Combining this with the truncated binomial gives a cumulative distribution function of Lies. This is then used to estimate means and confidence intervals.

Putting these together, for $R = N = 10$, the estimated distribution of Lies is calculated as follows:

- With probability $\frac{1}{1024}$, there were really 10 heads. Nobody lied in the sample.[4]
- Otherwise, 1 or more people saw tails, and they all lied. The proportion of liars is 100%.

Hence, the lower and upper confidence intervals are all 100%.

There are two problems with this approach: one statistical, and one conceptual.

[4] But the proportion of people who lied out of those who saw tails is undefined, because no one saw tails. The GSV software seems to resolve this by fixing the proportion of lies to 100%.

First, if many heads are reported, you should learn two things. On the one hand, there are probably many liars in your sample. On the other hand, probably a lot of coins really landed heads. The probability distribution in Eq. (5) does not take account of this.

For example, suppose we are certain that everyone in the sample is a liar who always reports heads. In this case, observing $R = N = 10$ gives us no information about the true number of heads. The posterior probability that $T = 10$ is then indeed 1/1024, the same as the prior. Now, suppose we know that nobody in the sample is a liar. Then on observing $R = 10$, we are sure that there were truly 10 heads: the posterior that $T = 10$ is 1. If exactly 5 out of 10 subjects are liars, then observing $R = 10$ means that all 5 truth-tellers really saw heads. The posterior probability that $T = 10$ is then $1/32$, the chance that all 5 liars saw heads, and so on.

When we are uncertain about the number of liars, our posterior that $T = 10$ will be some weighted combination of these beliefs. Unless we are certain everyone in the sample is a liar, the probability that $T = 10$ will be greater than 1 in 1024. Equation (5) is, therefore, not correct. In this case, it is equivalent to assume that everybody in the sample is a liar, whose report is uninformative about the true number of heads. One then uses the prior distribution of heads to estimate the proportion of those who actually saw tails and lied.

Indeed, in the simulations with $P = 0.5$ and across all values of λ, the overall probability that there were 10 true heads, conditional on $R = N = 10$, was about 1 in 161, not 1 in 1024. Fixing $\lambda = 0.2$, it was about 1 in 4.

This problem means that the GSV estimator of Lies is biased. In the "Appendix", I show that the GSV estimator can have substantial bias, and performs worse than the naïve estimator from Eq. (1), $\frac{R/N-(1-P)}{P}$. Also, the GSV confidence intervals do not always achieve nominal coverage of Lies. When the number of heads reported is either high or low, the percentage of confidence intervals containing Lies may fall below the nominal value.

There is a second, more important problem. The GSV approach attempts to estimate Lies in Eq. (6). This is the proportion of lies actually told, among the subsample of people who saw tails. But we are not usually interested in the proportion of lies actually told. We care about the probability that a subject in the sample would lie if they saw tails—λ in Eq. (2). This λ can be interpreted in different ways. Maybe on seeing a tail, each person in the sample lies with probability λ. Or maybe the sample is drawn from a population of whom λ are (always) liars, and $1 - \lambda$ are truth-tellers. Lies has no interpretation in the population, because the rest of the population has no chance to tell a lie in the experiment.

Lies can be treated as an estimate of λ. It is unbiased: it estimates λ from the random, and randomly sized, sample of $N - T$ people who saw tails. But it can be a very noisy estimate. Again, suppose 10 heads out of 10 are reported, and 9 heads were really observed. Lies is 100%. But it is 100% of just one person.

This means that even the correct confidence intervals for Lies would not be correct for λ. For example, if 3 out of 3 subjects report heads, the GSV software reports a lower bound of 100% for any confidence interval. Indeed, since anyone who had the opportunity to lie clearly did so, this is the correct lower bound (if we arbitrarily define Lies = 1 when $T = N$). But it makes no sense as a confidence interval for λ:

Table 4 GSV confidence interval coverage by proportion of heads reported (*R*/*N*)

R/N	Percentage of simulations	CI 90%	CI 95%	CI 99%
[0.00,0.25)	3.8	84.3%	87.9%	91.5%
[0.25,0.50)	10.6	76.3%	82.4%	89.2%
[0.50,0.75)	25.0	68.2%	75.7%	84.1%
[0.75,1.00]	60.6	60.8%	66.1%	74.1%

Table 5 Mean bias by method and *N*

Method	*N*: 10	*N*: 50	*N*: 100	*N*: 500
Bayesian	0.0025	0.00417	0.00438	0.00275
Frequentist	0.0354	0.01	0.0071	0.0025
GSV	0.048	0.016	0.0109	0.00419

Table 6 Mean squared errors by method and *N*

Method	*N*: 10	*N*: 50	*N*: 100	*N*: 500
Bayesian	0.0409	0.0136	0.00789	0.00184
Frequentist	0.0661	0.0159	0.00852	0.00184
GSV	0.0571	0.0142	0.00799	0.00184

we clearly cannot rule out that one or two subjects truly saw heads, and would have reported tails if they had seen tails.

Because of this problem, the GSV confidence interval coverage of λ is much worse than its coverage of Lies.The issue is especially serious when there are many reports of heads. In this case there were probably many true heads, so T is high and the true sample size $N - T$ is low, making Lies a noisy estimate of λ. Table 4 shows this. It splits the simulations by the proportion of reported heads, *R/N*. GSV coverage levels fall off sharply as *R/N* increases. Note that for fair coin flips, *R/N* is usually greater than 0.5, both in the simulations and in reality.

2 Point estimation

We can also compare the accuracy of point estimates of λ between GSV, Frequentist and Bayesian methods. Table 5 shows bias (the estimated value minus the true value of λ) for different methods by different N. The Bayesian method is always the least biased until $N = 500$, and the GSV method is the most biased.

Table 6 shows the mean squared error for methods by different N. For low N, the best method is Bayesian and the worst is Frequentist, with GSV in between. When N gets large, all methods give about the same estimates and are equally accurate.

The Bayesian method might have an advantage here, since it assumes a uniform prior and the simulations indeed used a uniform distribution of the proportion of

liars *L/N*. In fact, further analysis reveals that the Bayesian method is best across all specific values of *L/N* up to 80%.[5] So, the Bayesian method is likely to be best, unless one is sure that the true *L/N* is rather high.

3 Comparing different groups

Bayesian estimates are accurate, but rely on a choice of prior. A non-informative prior is a reasonable choice. Alternatively one might use information from previous meta-analyses such as Abeler et al. (2016). If the sample size is large enough, the choice of prior should not matter much.

When comparing the dishonesty rates of different groups, an interesting approach is to use the "empirical Bayes" method (Casella 1985). This piece of statistical jiu-jitsu involves estimating a common prior from the pooled data, before updating the prior for each individual group.

We can also test hypotheses using the Bayesian approach. If two samples are independent, then the probability that, e.g., the true proportion of liars in sample 1 is smaller than in sample 2 can be calculated from the posterior distributions for each sample:

$$\int_0^1 \int_0^{\lambda_1} \varphi(\lambda_1)\varphi(\lambda_2)\, \mathrm{d}\lambda_2 \mathrm{d}\lambda_1. \tag{7}$$

4 Applications

Benndorf et al. (2017) use the GSV method to calculate confidence intervals for the proportion of liars in a lying task with a die roll ($P = 5/6$). From 57 reports of the best outcome, out of 98 subjects, they calculate a lying rate of 49.68%, with a 95% CI of (45.3%, 53.95%). Using the Bayesian method with a uniform prior, the confidence interval becomes (38.0%, 61.1%), about twice as big.

Banerjee et al. (2018) use the GSV method to estimate confidence intervals for proportion of liars in a die roll task. They estimate the proportion of liars who report a die roll above 3 ($P = 0.5$), for several treatments. Table 8 in the "Appendix" shows GSV confidence intervals, along with recalculated Bayesian confidence intervals (from a uniform prior), and confidence intervals for the difference between lying to the "Same" and "Other" caste. The Bayesian confidence intervals are much larger than GSV confidence intervals. Only a couple of significant results survive. (Note that significance tests in the original paper were done with standard frequentist techniques, not the GSV method.) More importantly, the N is rather low to make useful inferences about the differences between groups. For example, for the T2-winners-GC group in the "aligned payoffs" treatment, differences in lying could be as much as 40% in either direction.

[5] When *L/N = 1*, all subjects deterministically report heads, and both Frequentist and GSV point estimates are exactly correct.

Hugh-Jones (2016) estimates the dishonesty rates of 15 nations using a coin flip experiment. I use empirical Bayes to check these results. For my prior over λ, I fit a beta distribution using the 15 observations of $2R/N - 1$. I then updated this prior separately for each country to find new confidence intervals and point estimates of the means.[6] There is some "shrinkage" towards the pooled mean from the naïve per-country estimates found by calculating $2R/N - 1$ separately for each country. One of the strengths of empirical Bayes, as Casella (1985) points out, is that it "anticipates regression to the mean". Using Eq. (7), I calculated the probability of different λ values for each pair of countries in the data. Reassuringly, there were still significant differences between countries.

5 Software

The Bayesian methods described here are implemented in R code, available at https://github.com/hughjonesd/GSV-comment. In this section, I give some simple examples of how to use it. More details are available at the website.

To load the code, download the file "bayesian-heads-cts.R" from github, and source it in the R command line:

```
source("bayesian-heads-cts.R")
```

Suppose 33 people report heads out of an N of 50, where the probability of the bad outcome is 0.5. To create a posterior distribution over λ, we use the `update_prior()` function:

```
updated <- update_prior(heads = 33, N = 50, P = 0.5, prior = dunif)
```

Here, we have started with a uniform prior, using R's built in `dunif()` function.

To calculate the point estimate of lambda, call the `dist_mean()` function on the updated posterior:

```
dist_mean(updated)

## [1] 0.3120336
```

To calculate the 95% confidence interval (the highest density region), use `dist_hdr()`:

[6] Results available on request.

```
dist_hdr(updated, 0.95)

## [1] 0.06123949 0.55000000
```

Lastly, we can run power tests by simulating multiple experiments. GSV argue that existing sample sizes may be too small to reject "no lying" ($\lambda = 0$). With a uniform prior and an N of 100, the Bayesian method has 80.6% power to detect λ of 25% and 21.4% power to detect λ of 10%. So, this paper confirms that important point. To run power calculations, use `power_calc()`. Here, we calculate the power to detect $\lambda = 0.1$ in a sample of 300, where the probability of the bad outcome is 0.5, with an alpha level of 0.05 and a uniform prior:

```
power_calc(N = 300, P = 0.5, lambda = 0.1, prior = dunif, alpha = 0.05)

## [1] 0.375
```

6 Conclusion

These results suggest some recommendations when designing and analysing a coin flip style experiment.

1. Use power tests to ensure that your N is big enough.
2. If your N is reasonably large, say at least 100, you can safely use standard frequentist confidence intervals and tests.
3. If your N is small, consider Bayesian estimates and confidence intervals. To estimate differences between subgroups, consider empirical Bayes with a prior derived from the pooled sample.

Appendix

R code to reproduce this comment is available at https://github.com/hughjonesd/GSV-comment, along with code to find Bayesian posterior distributions of λ.

The errors in GSV confidence intervals could be due to a programming error rather than to the algorithm. I could reproduce GSV's expected value to 4 or 5 significant figures by following their method, but I could not reproduce their confidence intervals. For example, when $R = 3, N = 6, P = 0.5$, GSV's Java program

Table 7 Proportion of true results within confidence interval, recalculated GSV method

CI 90%	CI 95%	CI 99%
62.3%	69.3%	77.4%

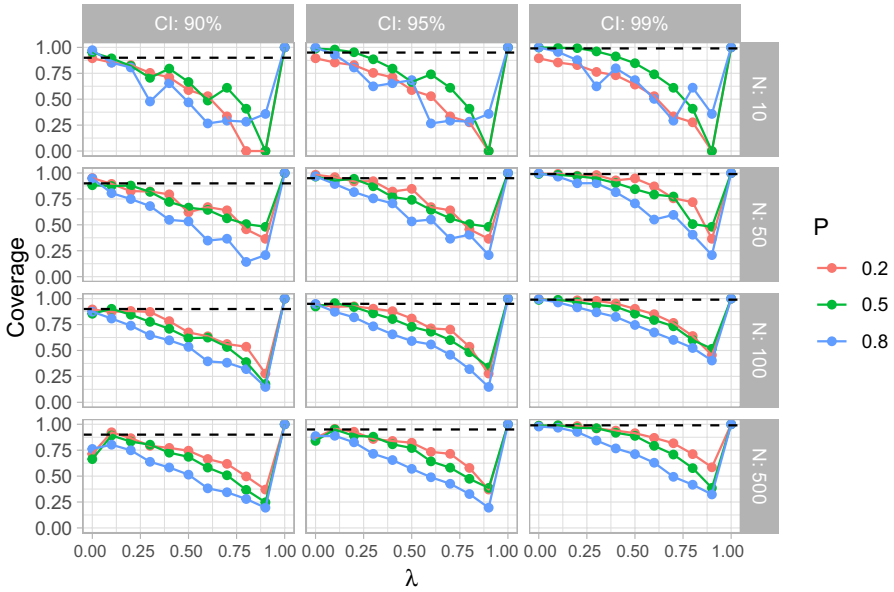

Fig. 1 GSV confidence interval coverage by confidence level, N, P and λ

gives the upper bound of the 95% confidence interval for the proportion of liars as 49.91%. But the possible proportions of lies when there are $T = 0, 1, 2, 3$ true heads, are $(R - T)/(N - T) = 50\%, 40\%, 25\%, 0\%$. It may be that some linear interpolation is being done.

Nevertheless, if I use GSV's method as stated, rather than their program, confidence intervals remain too small, as shown in Table 7.

Figure 1 shows the proportions of true values within the confidence interval for the GSV method, split by N, P, confidence level and λ. Dashed lines show the nominal confidence level. This makes the pattern clear: coverage gets worse as λ increases. (At 100%, coverage jumps back up since results become deterministic). Also, coverage does not get better as N increases.

Figure 2 shows the average bias of expected values by different methods. At $N = 500$, all methods perform reasonably well. For lower values, there is a clear pattern: Bayesian methods are least biased, GSV method is most biased, and the frequentist method is in between.

Figure 3 shows mean squared error by estimation method and λ, for N of 10 and 50. The Bayesian method is best for all values of λ up to 80%.

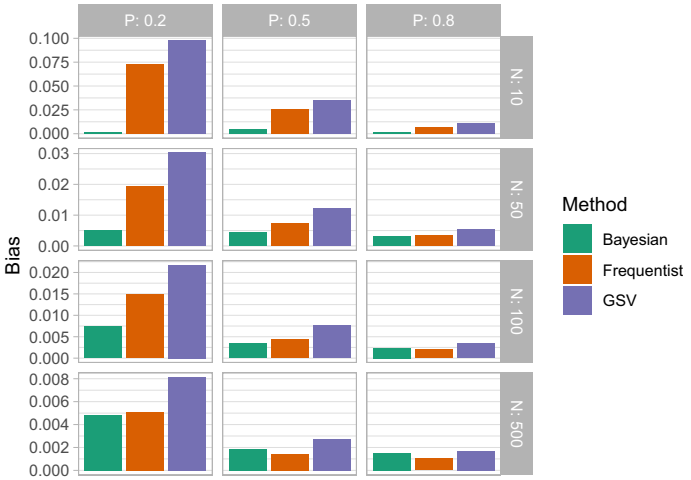

Fig. 2 Bias by method and λ

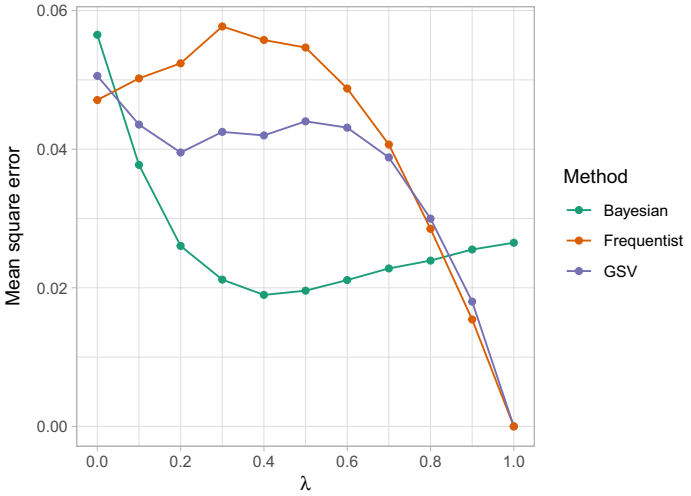

Fig. 3 Errors by method and λ, $N = 10$ and 50

GSV as an estimate of Lies

GSV argue that their method provides a good estimate of Lies, as opposed to λ. Here, I check whether that is true.

I ran 2000 simulations for each of the parameter combinations. I calculated Lies as $(R - T)/(N - T)$, and ignored simulations where $N = T$. I estimated confidence intervals and expected value using the GSV method. For a comparison, I also estimated the expected value of Lies using the "naïve" estimator $\max(0, \frac{R-(1-P))}{P})$.

Figure 4 shows average bias by sample size, P and true Lies. For low sample sizes, both methods perform about the same. For higher sample sizes and

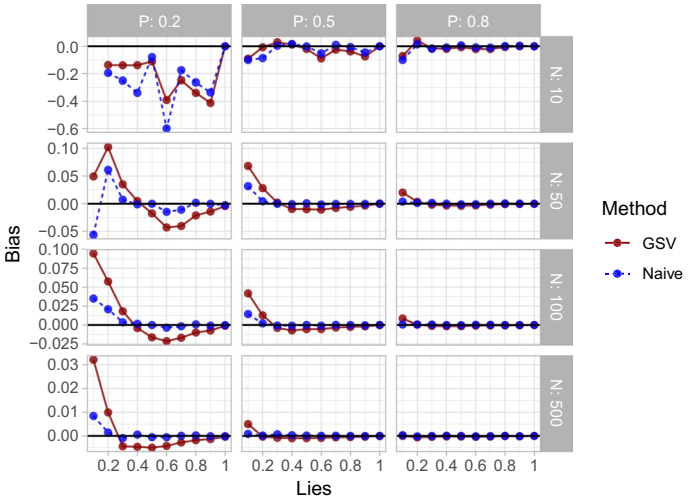

Fig. 4 Bias of GSV method for 'Lies'

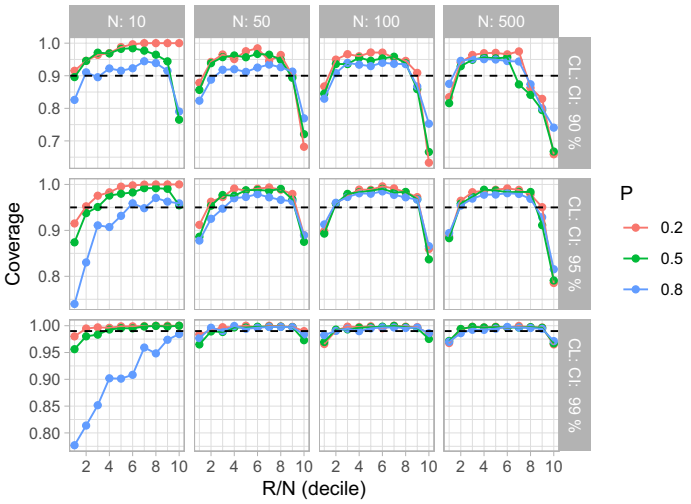

Fig. 5 Coverage of GSV method for 'Lies'

low values of Lies, however, the GSV method is clearly dominated by the naïve method, and shows a lot of upward bias—more than 5 percentage points even when $N = 100$. This is especially problematic for testing whether anyone lied in the sample.

Figure 5 shows the proportion of confidence intervals that contain the true value of Lies. Coverage is shown by confidence level, N, P and the decile (within these groups) of proportion of heads reported (R/N). Overall results are quite

solid, but when the proportion of heads reported is low or high, coverage drops below the nominal intervals. This is probably because, at these extremes, the difference between the true posterior and Eq. 5 becomes large. Interestingly, this problem gets worse as N increases.

Table 8 Confidence intervals from Banerjee et al. 2018, original and recalculated

Treatment	Payoffs	Target	Pct > 3	N	GSV 95% CI	Bayes 95% CI	Diff. 95% CI
T0-GC	Aligned	Same	77.4	84	42–62	35–71	
T0-SC	Aligned	Same	77.4	84	42–62	35–71	
T1-GC	Aligned	Same	83.7	43	53–74	42–85	−69 to 0
		Other	65.1		6–44	4–54	
T1-SC	Aligned	Same	80.5	41	43–69	34–81	−50 to 22
		Other	73.2		21–58	18–69	
T2-GC	Aligned	Same	75.6	41	29–62	23–74	−65 to −1*
		Other	53.7		0–27	0–34	
T2-SC	Aligned	Same	88.4	43	67–81	54–91	−44 to 17
		Other	81.4		47–70	36–82	
T2-winners-GC	Aligned	Same	58.3	12	0–38	0–60	−50 to 37
		Other	50		0–33	0–50	
T2-winners-SC	Aligned	Same	70.6	17	0–58	2–72	−63 to 26
		Other	52.9		0–33	0–47	
T2-losers-GC	Aligned	Same	87.9	33	64–81	49–92	−87 to −32*
		Other	45.5		0–18	0–27	
T2-losers-SC	Aligned	Same	68.2	22	0–53	2–65	−39 to 53
		Other	72.7		14–60	8–74	
T0-GC	Unaligned	Same	68.9	90	22–47	19–55	
T0-SC	Unaligned	Same	66.7	78	13–45	11–53	
T1-GC	Unaligned	Same	57.1	42	0–33	0–39	−7 to 60
		Other	73.8		27–58	20–70	
T1-SC	Unaligned	Same	57.1	42	0–33	0–38	−20 to 47
		Other	66.7		7–46	6–57	
T2-GC	Unaligned	Same	60	40	0–36	0–45	−5 to 64
		Other	77.5		36–64	26–77	
T2-SC	Unaligned	Same	56.8	44	0–32	0–38	−29 to 35
		Other	59.1		0–36	0–41	
T2-winners-GC	Unaligned	Same	33.3	12	0–11	0–36	−30 to 38
		Other	41.7		0–22	0–41	
T2-winners-SC	Unaligned	Same	40	15	0–18	0–36	−13 to 71
		Other	73.3		0–60	5–78	
T2-losers-GC	Unaligned	Same	53.1	32	0–29	0–36	37 to 93*
		Other	93.8		80–90	64–98	
T2-losers-SC	Unaligned	Same	60.7	28	0–39	0–49	−18 to 62
		Other	75		22–61	16–76	

*Significant at 95%, i.e. the confidence interval does not contain 0

Banerjee et al. results

Table 8 recalculates confidence intervals from Banerjee et al. (2018) using data in their Tables 6A and 6B.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2016). Preferences for truth-telling. CESIFO working paper.

Agresti, A., & Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–26.

Banerjee, R., Gupta, N. D., & Villeval, M. C. (2018). The spillover effects of affirmative action on competitiveness and unethical behavior. *European Economic Review*, *101*, 567–604.

Benndorf, V., Moellers, C., & Normann, H.-T. (2017). Experienced vs. inexperienced participants in the lab: Do they behave differently. *Journal of the Economic Science Association*, *3*(1), 12–25.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or Fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–13.

Garbarino, E., Slonim, R., & Villeval, M. C. (2018). A method to estimate mean lying rates and their full distribution. *Journal of the Economic Science Association*,. https://doi.org/10.1007/s40881-018-0055-4.

Hugh-Jones, D. (2016). Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization*, *127*, 99–114.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50*(2), 120–26.