True Lies

Comment on Garbarino, Slonim and Villeval (2018)

David Hugh-Jones

2018-11-08

Abstract: Garbarino, Slonim and Villeval (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 are really interested in: the proportion of liars.

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 don't 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* outcome happens with probability P, we can estimate that the following proportion are lying (Abeler, Nosenzo, and Raymond 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, Slonim, and Villeval (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

[1] GSV maintain this assumption and so shall I.

full distribution of lying outcomes, and they recommend using it for confidence intervals, hypothesis testing and power calculations.

Table 3 of GSV reports Monte Carlo simulations, which show that the probability of a Type II error can be up to 50 per cent. GSV define a Type II error as the 95% confidence interval not containing the true value. A 95% confidence interval which only contains the true value 50% of the time is unusual. What is going on?

I ran simulations to check the overall performance of the GSV confidence intervals. Simulations parameters are shown in Table 1. For each parameter combination, I ran 1000 simulations. λ is the probability that an individual in the sample lies and report heads when they observe tails:

$$\lambda = \text{Prob}(i \text{ reports heads}|i \text{ saw tails}). \tag{2}$$

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

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: Proportion of true results within confidence interval.

	CI 90%	**CI 95%**	**CI 99%**
GSV	65.2%	71.0%	78.9%

By definition, 95% of 95% confidence intervals ought to contain the true value, on average. This is called "achieving nominal coverage". GSV intervals are far from this. The confidence intervals are too narrow. This problem holds across all simulated probabilities of the low outcome, confidence levels, and sample sizes.[2]

Why does this happen? 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%. Naïvely, this might seem wrong. Can we rule out that 8 subjects lied, and the two honest subjects both saw heads?

[2] See the appendix.

In fact, the GSV method works like this. First, given R reports of heads, the probability that a total of T "true" heads were observed is calculated as:

$$\text{Prob}(T\,\text{heads}|R; N, P) = \frac{\text{binom}(T, N, 1-P)}{\sum_{k=0}^{R} \text{binom}(k, N, 1-P)} \tag{3}$$

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:

$$Lies = \frac{R-T}{N-T} \tag{4}$$

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 the proportion of liars. This is then used to estimate means and confidence intervals.

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

- With probability $\frac{1}{1024}$, there were really 10 heads. Nobody lied in the sample.[3]
- 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 is statistical, one is conceptual.

First, if 10 out of 10 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 equation (3) does not take account of this. In particular, for $R = N$, it reduces to the binomial distribution. This is not correct. For example, in my simulations with $P = 0.5$, the overall probability that there were 10 heads when $R = N = 10$ was about 1 in 161, not 1 in 1024. When $\lambda = 0.2$, it was about 1 in 4. The same is true at the other end of the scale: when R is low, we indeed learn that $T \leq R$ but also that the overall distribution is not shaped like the prior.

Second, more importantly, the GSV approach estimates *Lies*, the distribution of 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 people in the sample would lie if they saw tails – λ in equation (2). This λ can be interpreted in different ways. It could be that on seeing

[3] But the proportion of people who lied *out of those who saw tails* is undefined, because noone saw tails. The GSV software seems to resolve this by fixing the proportion of lies to 100%.

a tail, each person in the sample lies with probability λ. Or it could be that the sample is drawn from a population of whom λ are (always) liars, and $1 - \lambda$ are truth-tellers. *Lies*, on the other hand, has no interpretation in the population, because the rest of the population has no chance to tell a lie in the experiment. Indeed, when GSV run Monte Carlo simulations, they themselves check their confidence intervals against λ (which they call Q).

The fact that some people saw heads, and had no reason to lie, is just a nuisance for estimation. Similarly, we are not interested, per se, in whether more lies were actually told in one group than another. Suppose that 9 out of 10 people saw heads. Only one person had a reason to lie, so by definition, *Lies* will be either 0 or 1. This will almost certainly be different from the value of *Lies* in another group where there were fewer heads. But that tells us very little, either about the samples as a whole, or about the population they come from.

Lies can be treated as an estimate of λ. It is unbiased: it is an estimate of λ in the random 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.

The first of these two problems means that the GSV distribution of *Lies* is wrong. In the appendix I show that their estimator can have substantial bias. In fact, it performs worse than the naïve estimator from equation (1), $\frac{R-(1-P)}{P}$. Since the true value of *Lies* is an unbiased estimator of λ, the GSV estimator is also a biased estimator of λ, as we will see. 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.

The second problem means that the GSV confidence intervals are not correct for λ. We can tell that the second problem matters, because the confidence interval coverage of λ is much worse than coverage of *Lies*.

The problem is especially bad 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. Table 3 shows this. It splits the simulations by the proportion of reported heads, *R/N*. Coverage levels fall off sharply as *R/N* increases. For fair coin flips, *R/N* is typically greater than 0.5, in the simulations and in reality.

Table 3: GSV confidence interval coverage by proportion of heads reported (R/N)

R/N:	**Percentage of cases**	**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%

Alternative methods

Can we do better? 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:

$$\Pr(R|\lambda; N, P) = \text{binom}(R, N, (1 - P) + \lambda P) \tag{5}$$

That immediately suggests the first 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, Nosenzo, and Raymond (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 – by 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.[4] We'll 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. I call this the "frequentist" method.

The second method is to use Bayes' rule. Start with a prior probability density function over the probability of liars, $\varphi(\lambda)$. The posterior probability is then:

$$\varphi(\lambda|R; N, P) = \frac{\Pr(R|\lambda; N, P)\varphi(\lambda)}{\int_0^1 \Pr(R|\lambda'; N, P)\varphi(\lambda)\, \mathrm{d}\lambda} \tag{6}$$

From this, one can derive confidence intervals and expected values in the usual way.[5] Technically, they are Bayesian "credible" intervals. I use 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 $[0, 1]$.

I ran these estimation methods on my simulated data. Results are shown in Table 4. Both frequentist and Bayesian methods mostly achieve the nominal confidence level, with more than 90/95/99% of

[4] In particular, this method can estimate confidence bounds for λ lower than 0. If so, we can set them to 0.
[5] R code to do this is available at https://github.com/hughjonesd/GSV-comment.

intervals containing the true value of *Liars*. The exception is the frequentist 99% confidence interval, which is too narrow.

Table 4: 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%

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 5 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. Both alternatives perform much better than the GSV approach.

Table 5: 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%

Benndorf, Moellers, and Normann (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. Obviously, this matters for e.g. sample size calculations.

Point estimation

We can also compare the accuracy of point estimates of *Liars*. Table 6 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 7 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%.[6] So, the Bayesian method is likely to be best unless one is sure that the true L/N is rather high.

Table 6: 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 7: 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

Power tests

GSV argue that existing sample sizes may be too small to reject "no lying" ($\lambda = 0$). We can check this with the Bayesian method by simulations. With a uniform prior and an N of 100, the Bayesian method has 83.0% power to detect λ of 25% and 20.0% power to detect λ of 10%. So, on this we agree.

Empirical Bayes

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, Nosenzo, and Raymond (2016). If the sample size is large enough, the choice of prior should not matter much.

[6] When $L/N = 1$, all subjects deterministically report heads, and both Frequentist and Bayesian point estimates are exactly correct.

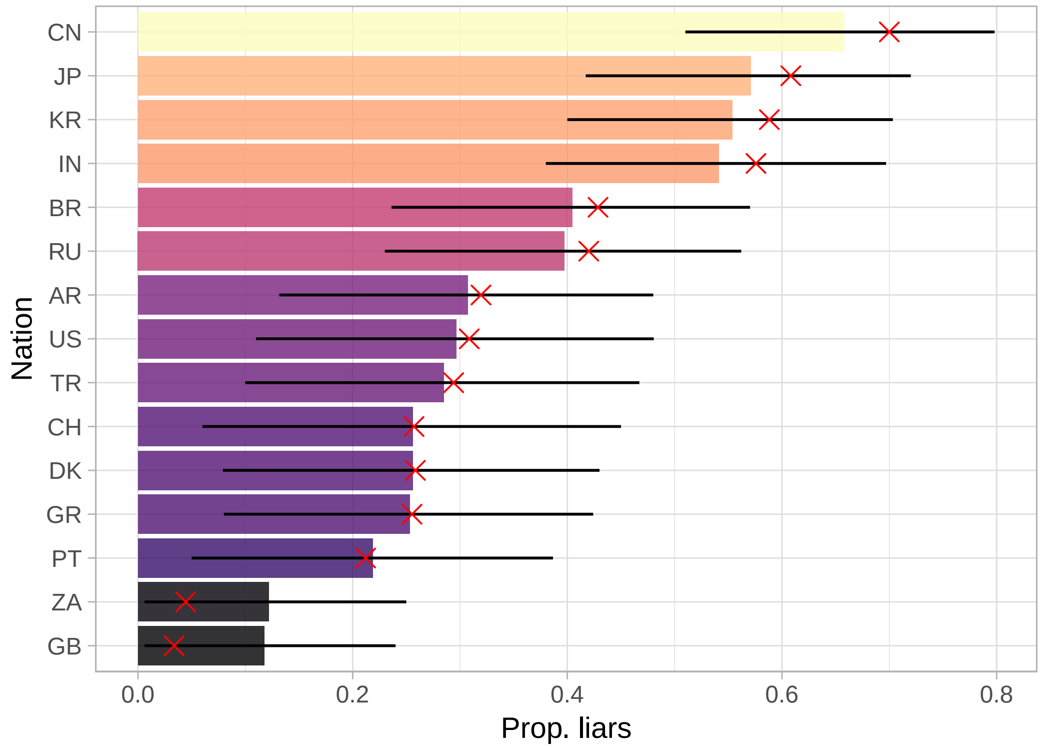

Figure 1: Posterior means and confidence intervals for proportion dishonest in 15 countries. Crosses are naïve estimates.

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.

For example, Hugh-Jones (2016) estimates the dishonesty rates of 15 nations using a coin flip experiment. We can fit a beta distribution using the 15 observations of *2R/N - 1*. This will be our prior.

Results are shown in Figure 1, with means and 95% confidence intervals. 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". That is, as there is noise in the data, the most and least honest groups in the data probably had some extreme error terms. So the British people in the sample may be less honest, and the Chinese people more honest, than the per-group data suggest.

Bayesian hypothesis testing

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 marginal distributions in each sample:

$$\int_0^1 \int_0^{\lambda_1} \varphi(\lambda_1)\varphi(\lambda_2)\ \mathrm{d}\lambda_2 \mathrm{d}\lambda_1$$

Table 8 shows Bayesian comparisons for selected countries. Even though estimates have "shrunk" towards the pooled mean from the naive estimates, differences remain significant, partly because the pooled prior has made posterior confidence intervals narrower. Similarly, we can calculate the full distribution of the differences between countries.

Table 8: Bayesian country comparisons. Cells give the probability that the row country has higher proportion of liars.

	GB	**ZA**	**GR**	**BR**	**CN**
GB	–	0.483	0.115	0.007	0.000
ZA	0.517	–	0.123	0.008	0.000
GR	0.885	0.877	–	0.115	0.000
BR	0.993	0.992	0.885	–	0.014
CN	1.000	1.000	1.000	0.986	–

Banerjee, Gupta, and Villeval (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 9 shows GSV confidence intervals, along with recalculated Bayesian confidence intervals, and confidence intervals for the difference between lying to the "Same" and "Other" caste. Here, I used a non-informative prior rather than a pooled prior, letting the algorithm maintain an "open mind" about treatment effects.

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, it is clear that 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.

Conclusion

These results suggest some recommendations when you are analysing a coin-flip style experiment.

1. Do not use the GSV method. Confidence intervals are always too narrow, risking false inferences.
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.

Lastly, here is a simple experimental design to avoid these statistical shenanigans. Ask a group of N experimental subjects, not to flip a coin, but to draw a black or white ball, without replacement, from an urn containing N balls, of which T are white. Pay them if they report a white ball. You will then know – not estimate – that the proportion of lies told is $(R-T)/(N-T)$, and this will also be the proportion of liars, in a sample of exactly $N-T$ subjects. You can compare different samples, or test for no lying, using a straightforward chi-squared test. So, you can simultaneously estimate the proportion of lies, and of the lying liars who tell them.

Table 9: 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	-68 – -1 *
		Other	65.1		6 – 44	4 – 54	
T1-SC	Aligned	Same	80.5	41	43 – 69	34 – 81	-50 – 22
		Other	73.2		21 – 58	17 – 70	
T2-GC	Aligned	Same	75.6	41	29 – 62	22 – 74	-66 – 0
		Other	53.7		0 – 27	0 – 34	
T2-SC	Aligned	Same	88.4	43	67 – 81	53 – 91	-44 – 17
		Other	81.4		47 – 70	36 – 82	
T2-winners-GC	Aligned	Same	58.3	12	0 – 38	0 – 58	-52 – 39
		Other	50		0 – 33	0 – 49	
T2-winners-SC	Aligned	Same	70.6	17	0 – 58	2 – 72	-63 – 25
		Other	52.9		0 – 33	0 – 46	
T2-losers-GC	Aligned	Same	87.9	33	64 – 81	49 – 92	-87 – -32 *
		Other	45.5		0 – 18	0 – 28	
T2-losers-SC	Aligned	Same	68.2	22	0 – 53	2 – 65	-39 – 53
		Other	72.7		14 – 60	9 – 74	
T0-GC	Unaligned	Same	68.9	90	22 – 47	17 – 56	
T0-SC	Unaligned	Same	66.7	78	13 – 45	11 – 53	
T1-GC	Unaligned	Same	57.1	42	0 – 33	0 – 38	-7 – 60
		Other	73.8		27 – 58	19 – 71	
T1-SC	Unaligned	Same	57.1	42	0 – 33	0 – 39	-20 – 47
		Other	66.7		7 – 46	5 – 58	
T2-GC	Unaligned	Same	60	40	0 – 36	0 – 44	-5 – 64
		Other	77.5		36 – 64	27 – 77	
T2-SC	Unaligned	Same	56.8	44	0 – 32	0 – 37	-30 – 35
		Other	59.1		0 – 36	0 – 41	
T2-winners-GC	Unaligned	Same	33.3	12	0 – 11	0 – 35	-31 – 39
		Other	41.7		0 – 22	0 – 43	
T2-winners-SC	Unaligned	Same	40	15	0 – 18	0 – 36	-13 – 71
		Other	73.3		0 – 60	6 – 77	
T2-losers-GC	Unaligned	Same	53.1	32	0 – 29	0 – 37	38 – 92 *
		Other	93.8		80 – 90	64 – 98	
T2-losers-SC	Unaligned	Same	60.7	28	0 – 39	0 – 50	-18 – 62
		Other	75		22 – 61	15 – 76	

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 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 10.

Table 10: Proportion of true results within confidence interval, recalculated GSV method

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

Figure 2 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 3 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 4 shows mean squared error by estimation method and λ, for *N* of 10 and 50. The Bayesian method is best for all values of *Liars* up to 80%.

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 $\max(0, \frac{R-(1-P))}{P})$.

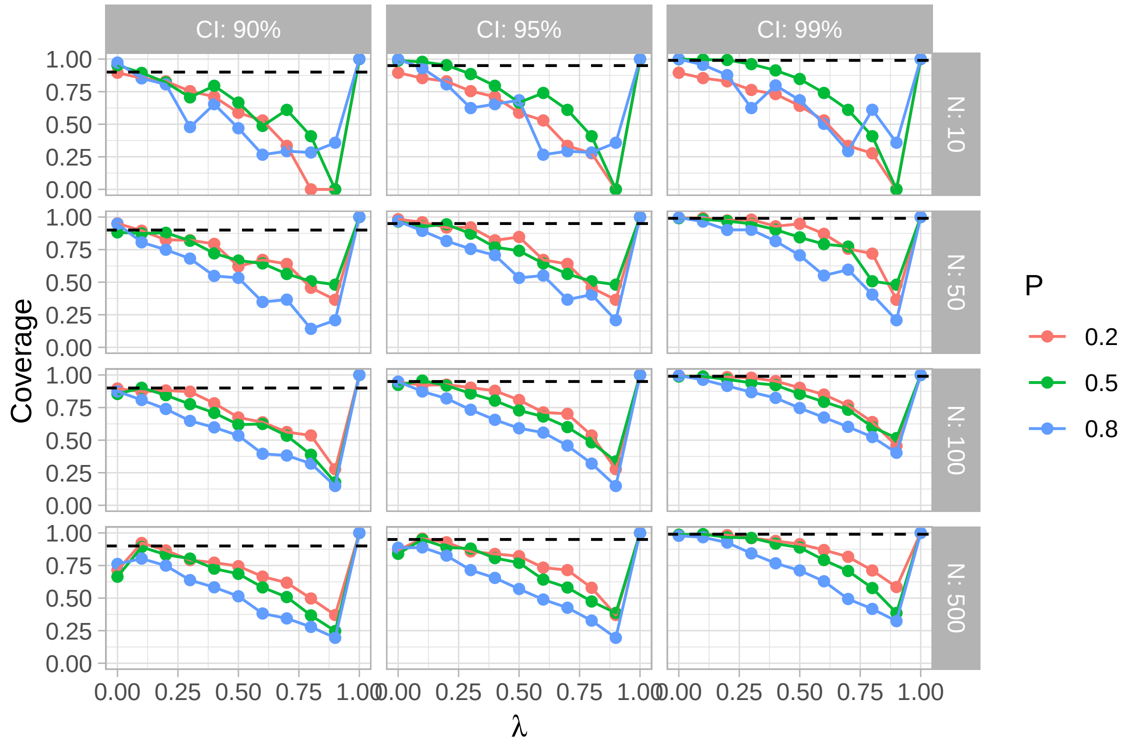

Figure 2: GSV confidence interval coverage by confidence level, N, P and λ

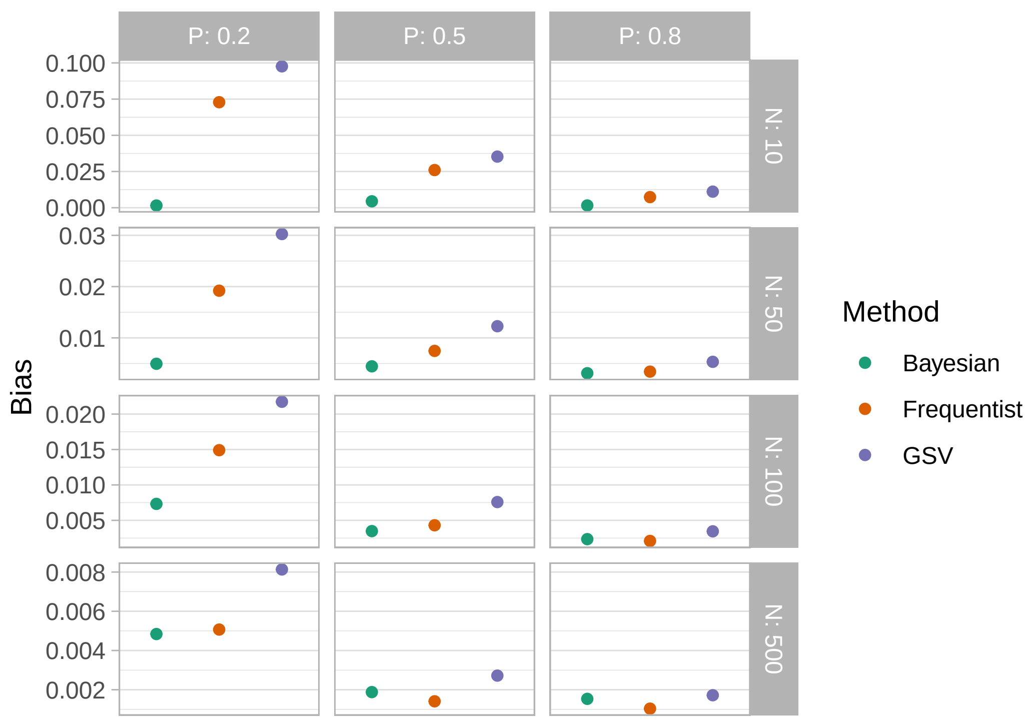

Figure 3: Bias by method and λ

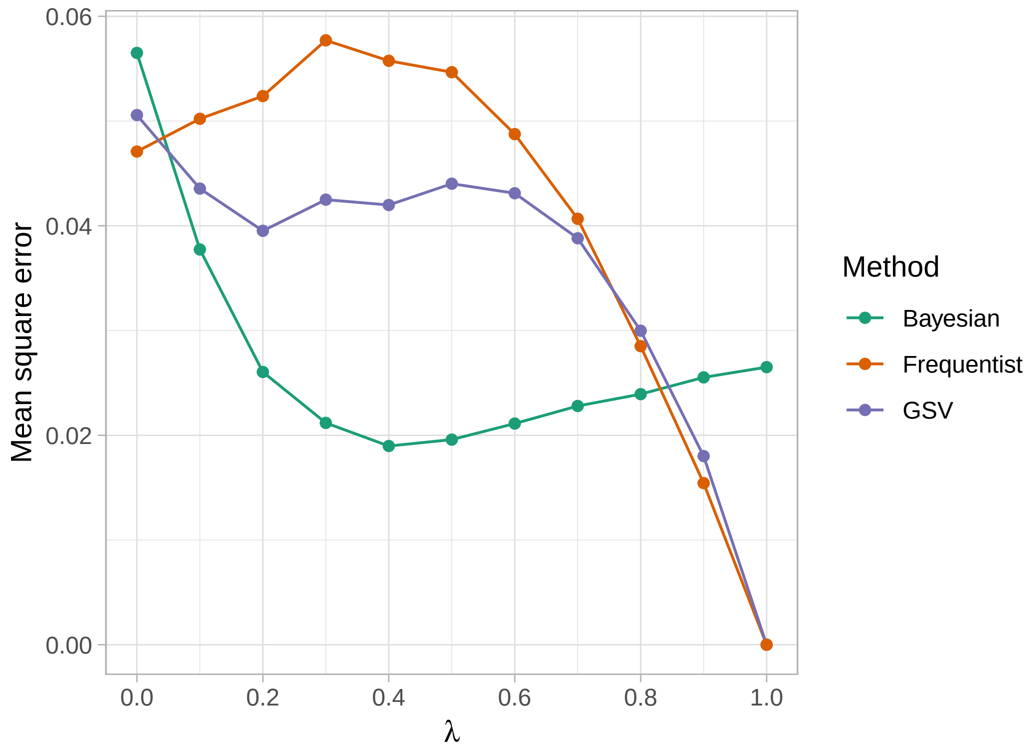

Figure 4: Errors by method and λ, N = 10 and 50

Figure 5 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 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 6 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 equation 3 becomes large. Interestingly, this problem gets worse as N increases.

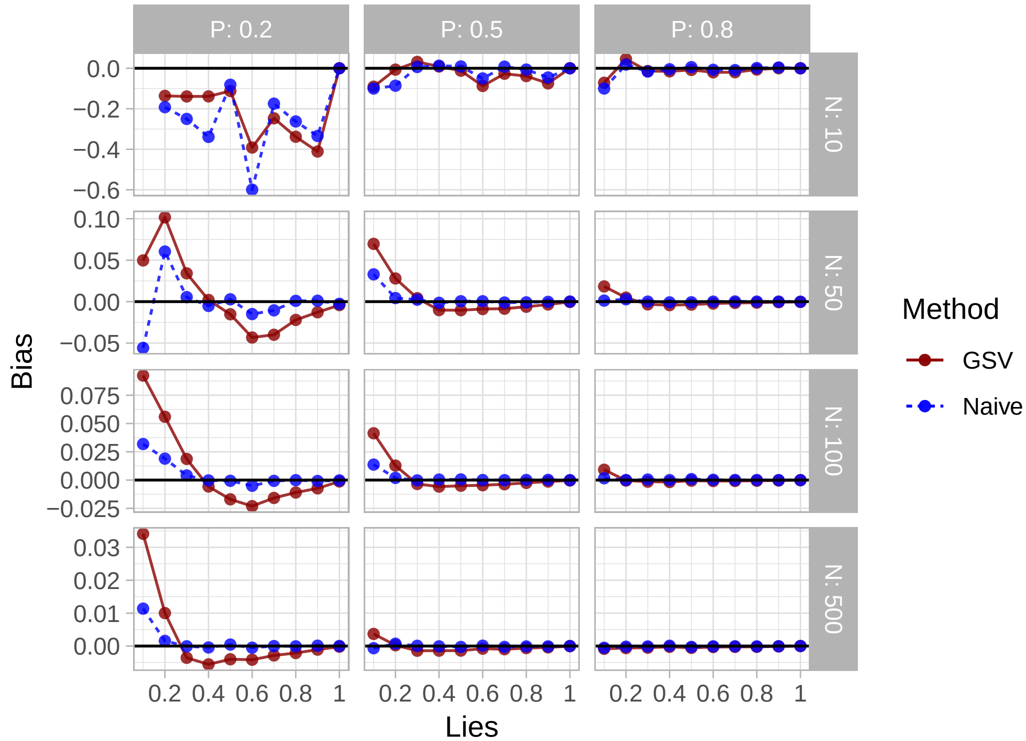

Figure 5: Bias of GSV method for 'Lies'.

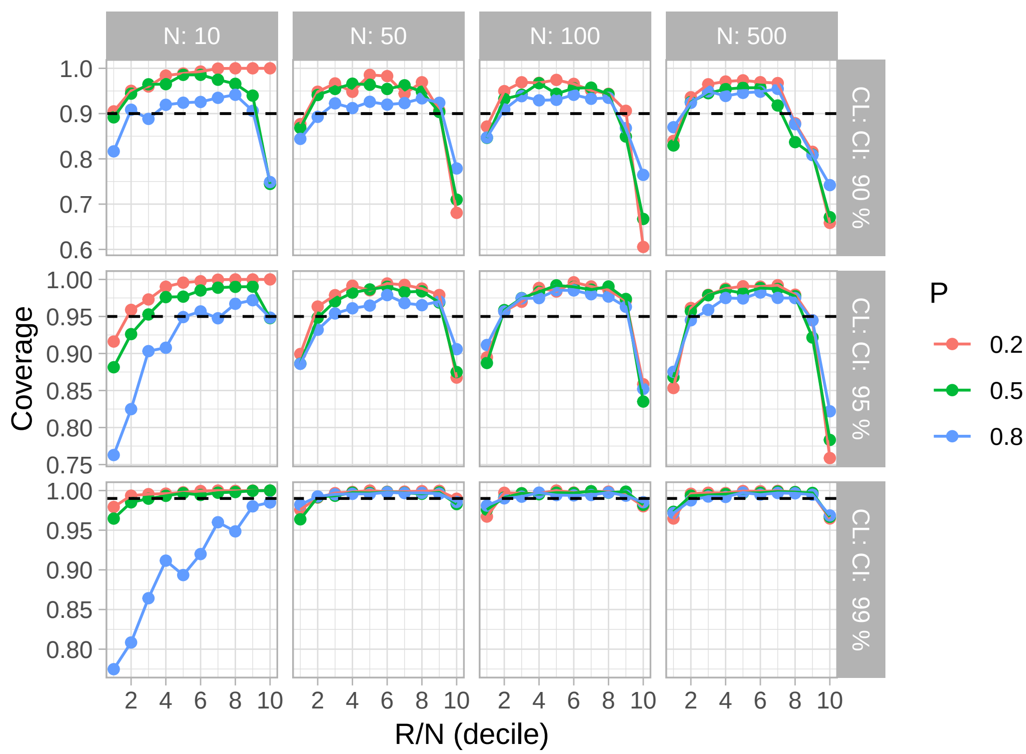

Figure 6: Coverage of GSV method for 'Lies'.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2016. "Preferences for Truth-Telling."

Agresti, Alan, and Brent A Coull. 1998. "Approximate Is Better Than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician* 52 (2): 119–26.

Banerjee, Ritwik, Nabanita Datta Gupta, and Marie Claire Villeval. 2018. "The Spillover Effects of Affirmative Action on Competitiveness and Unethical Behavior." *European Economic Review* 101: 567–604.

Benndorf, Volker, Claudia Moellers, and Hans-Theo Normann. 2017. "Experienced Vs. Inexperienced Participants in the Lab: Do They Behave Differently?" *Journal of the Economic Science Association* 3 (1): 12–25.

Casella, George. 1985. "An Introduction to Empirical Bayes Data Analysis." *The American Statistician* 39 (2): 83–87.

Clopper, Charles J, and Egon S Pearson. 1934. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial." *Biometrika* 26 (4): 404–13.

Garbarino, Ellen, Robert Slonim, and Marie Claire Villeval. 2018. "A Method to Estimate Mean Lying Rates and Their Full Distribution." *Journal of the Economic Science Association*, October. https://doi.org/10.1007/s40881-018-0055-4.

Hugh-Jones, David. 2016. "Honesty, Beliefs About Honesty, and Economic Growth in 15 Countries." *Journal of Economic Behavior & Organization* 127: 99–114.

Hyndman, Rob J. 1996. "Computing and Graphing Highest Density Regions." *The American Statistician* 50 (2): 120–26.