"Do You Like Brussels Sprouts?" How to Implement Population Surveys with Embarrassing Questions*

Keith Hellman†

Colorado Council of Teachers of Mathematics
Denver Conference, September 29, 2006

1 Introduction

Have you ever wondered how studies of "embarrassing questions" are performed? Would you enter a study with questions about drug-use, ethical, or moral behavior and then answer the questions honestly? Even if study participants are assured their answers are anonymous, many people *will not* answer all the questions truthfully, out of fear of recrimination or exposure.

Yet, studies such as these are constantly published. How many high school aged students have already experimented with drugs? How many college students have plagiarized? How many college students have been drunk more than four times in a school year? It's not hard to hear these statistics in the public media — should we believe the results? How are these studies performed without the fear of erroneous conclusions?

One way is to guarantee the study participants *plausible-deniability*. Which means that even if they have answered a question in an unflattering way and later confronted with the fact,[1] they can still claim their answer was a lie. Moreover, this claim must be an entirely reasonable, or plausible, assertion.

These studies are quite simple and can certainly be understood by many middle-schoolers, indeed anyone with a basic understanding of probability and algebra — and this is *important*. If surveyors could not convince the studied population of full plausible-deniability, the study would be left with results of low confidence.

Embarrassing surveys present a useful application of probability and algebra. You might even ask your students to come up with their own questions for a class-wide survey, then conduct the survey and analyze the results.

2 Study Algorithm

The key to the study design is twofold:

1. only yes-no questions are asked, with the "embarrassing" answer always being "yes"
2. an individual yes answer cannot be trusted. Only *all* of the yes answers, grouped together, are trusted.

Let's see how this works. Each study participant is given a printed survey, pencil, and a fair coin (even better if they use their own!). They are led to a reasonably private location, **where the results of coin-flips cannot be seen by onlookers**, and they agree to follow a specific algorithm (Figure 1 on page 3) when answering each question.

1. For every question, they begin by flipping the coin.

Y is a yes answer on the survey (which may be a lie).

N is a no answer on the survey.

Y is a true yes from the studied population.

N is a true no from the studied population.

*Supported through a grant from the Colorado Department of Education and the NCLB

†Colorado School of Mines, Golden Colorado. khellman@mines.edu http://alamode.mines.edu/~khellman/

[1] meaning their anonymity has been compromised

2. Next, they read the survey question, and if their *truthful* answer is Y, they answer **Y** on the survey form.
3. If their *truthful* answer is N, they consult their coin flip. If it is heads, *they lie*, and answer **Y** on the survey form. If the coin flip is tails, then they answer **N** on the survey form.

A study participant has one way of answering **N**, if his truthful answer is N. Every time a participant writes a **Y** on the questionnaire, it may be because his coin-flip was heads. Thus, every participant has plausible-deniability for each **Y** answer provided.

It might be useful to raise some questions about why this algorithm is so important, and how it protects the survey participant from exposure.

- *Why must the coin be flipped for each question? Isn't it a waste of time if the answer is already yes?* Suppose you're taking the test in the far corner of the room, far enough such that your coin flip results cannot be seen by anyone. I'm sitting on the other side of the room with a copy of the survey — if you don't flip your coin for each question, I will know what your answers are. How? First, I count the number of times you write answers on your survey, in this way, I know which question you are on. Then I simply keep track of which questions you don't flip the coin for, since you would only not flip the coin for questions you would answer **Y** to, I know your Y questions.[2]
- *Why must the survey questions be posed as the embarrassing answer always being yes?* It's those words again: *plausible-deniability*. Run through the correct answer-logic algorithm with a poorly formed question which has an embarrassing N answer (perhaps "Do you hate Brussels sprouts?"). If your answer is truly N, and your coin flip is tails, you must answer **N**. Now imagine your anonymity is compromised, since the only way through the answer-logic which arrives at **N** is if your true answer is N, you cannot fall back onto plausible-deniability and claim "the coin flip made me do it." For completeness, consider a properly formed question, such as "Do you like Brussels sprouts?" You really do, so you answer the survey honestly with a **Y**. Now, again, suppose your anonymity is compromised and you are brought before congressional hearings. When asked about your purported love of Brussels sprouts you answer "Brussels sprouts! Echk! I don't like Brussels sprouts, I answered **Y** because my coin flip was heads!"
- *Why can't the alternative answering logic shown in Figure 1 be used?* This answer is the same as the answer above. Run through the logic and you'll see that the alternative, incorrect, algorithm has **Y** answerable only if Y is a person's true answer. The incorrect algorithm does not provide plausible-deniability to a study participant.

3 Survey Analysis

You are hopefully convinced that the described study methods provide participants with the plausible-deniability needed to receive honest survey answers. But how can these results be used when an unknown fraction of the **Y** answers are lies?

Let's look at the answer to just one question, and what happens to the truthful fraction of Y and N answers inside the *answer logic* (Figure 2 on page 4). In this figure, we know the fraction of Ys and Ns existing in the population, and based on this, we can predict the count of **Y**s and **N**s measured by the study. Is it possible to go backwards? If we know the counts of **Y**s and **N**s from a survey question, can we work backwards to an estimate of the true fraction of Ys and Ns in the population? Sure we can![3]

Let p be the percent, or probability, of a survey question having a truthful Y in the population. Then the fraction of Ns will be $1-p$. If each survey participant follows the "answer logic" instructions,[4] there are two sources of **Y** answers: The first source is the percent of truthful Ys in the population, or pn, where n is the number of participants in the survey. The second source is from the participants that would truthfully answer N ($(1-p)n$), since these participants flip fair coins, we expect this number to be about $\frac{1}{2}(1-p)n$. Now we can write an equation for y, the survey's expected number of **Y** answers.

p is the percent, or probability, of a survey question having a truthful Y in the population

n is the number of participants in the survey

y is our survey's expected number of **Y** answers

[2] Granted, this type of distance eavesdropping may be difficult to pull off, but it is plausible enough to explain why it is best to flip the coin for each single question.

[3] I'll present a simplified approach later (Section 4 on page 5), one that is perhaps better suited for middle-school math.

[4] And we expect them to, because this provides them the plausible-deniability if their anonymity is compromised.

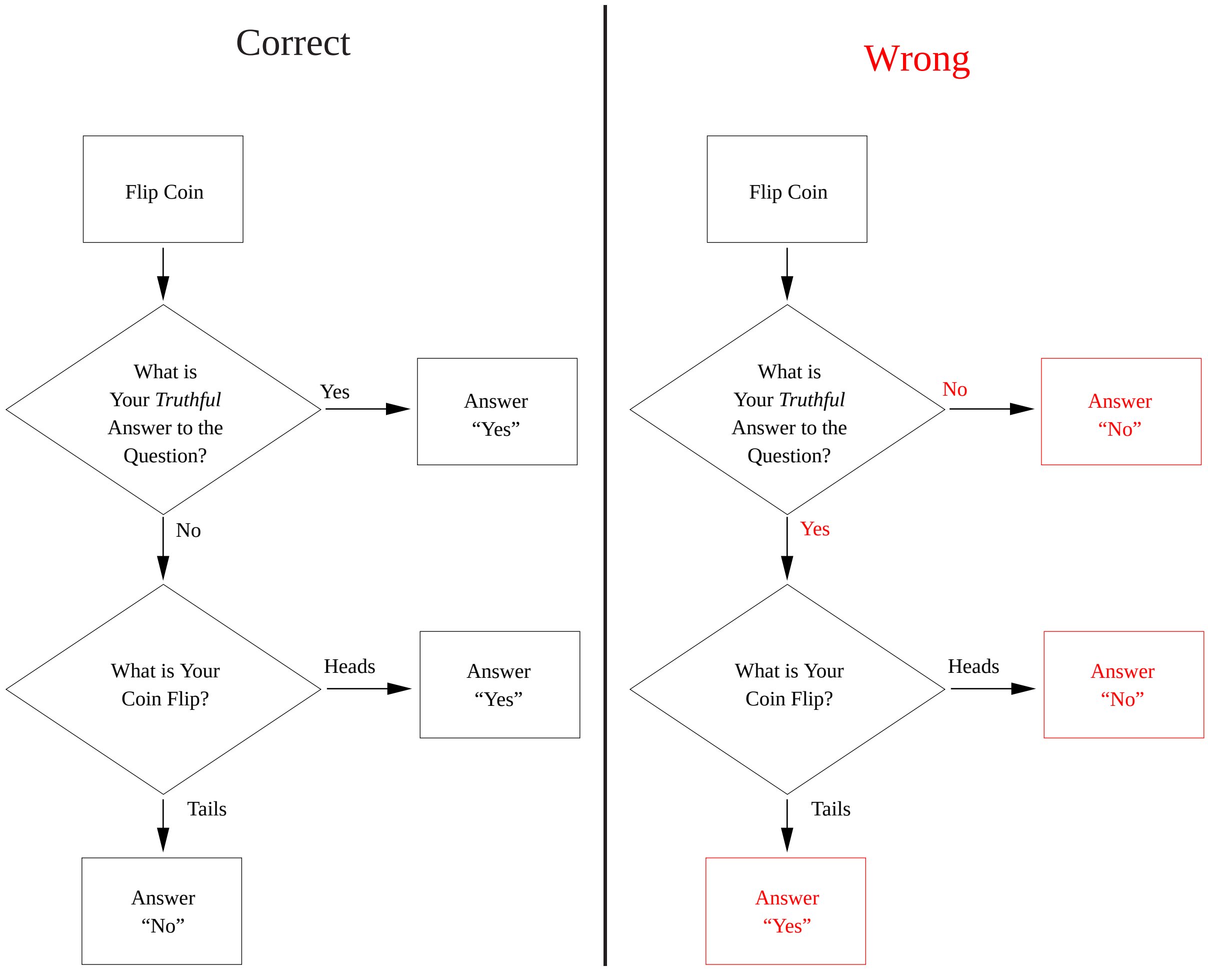


Figure 1: Answer Logic. The left-hand algorithm is correct. The right-hand algorithm (differences highlighted in red) does not offer study participants *plausible-deniability*, why not?

$$
\begin{aligned}
y &= \overbrace{pn}^{\text{truthful } \mathbf{Y}} + \underbrace{\frac{1}{2}\overbrace{(1-p)n}^{\text{truthful } \mathbf{N}}}_{\mathbf{Y} \text{ from fair coin flips}} \\
& \quad \text{solving for } p \ldots \\
\frac{y}{n} &= p + \frac{1}{2}(1-p) \\
\frac{y}{n} &= p + \frac{1}{2} - \frac{1}{2}p \\
\frac{y}{n} - \frac{1}{2} &= p - \frac{1}{2}p \\
\frac{y}{n} - \frac{1}{2} &= \frac{1}{2}p \\
2\left(\frac{y}{n} - \frac{1}{2}\right) &= p \\
p &= \frac{2y}{n} - 1 \qquad (1)
\end{aligned}
$$

Let's examine Equation 1 more closely. If we ask a question for which every survey participant would answer **Y**, then we'll have a total of n **Y**s, or $y = n$, and our calculation of p would be:

$$p = \frac{2n}{n} - 1 = 2 - 1 = 1$$

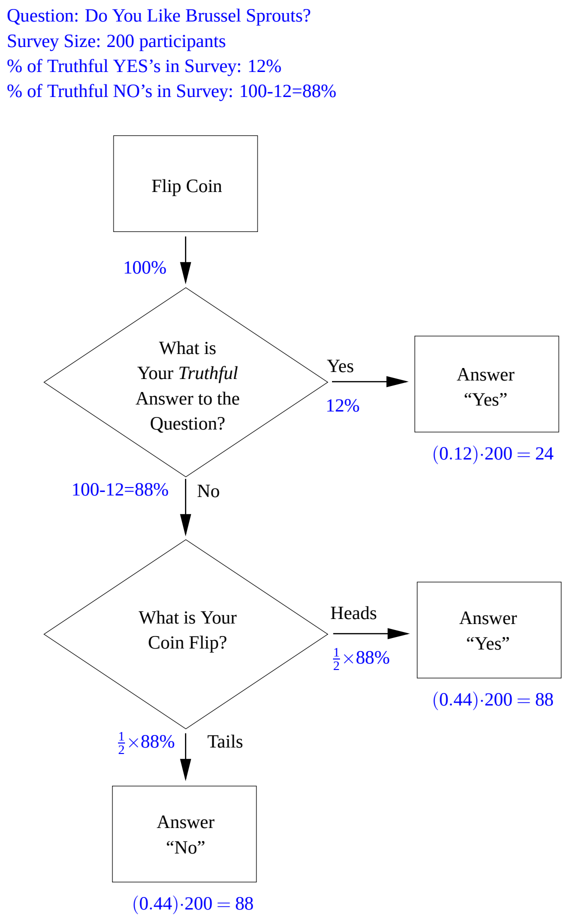


Figure 2: The Yes-No Flow. Survey results for a hypothetical question with known answer probabilities through the answer logic.

100% (or 1.0 in decimal) is certainly the expected result. What if we ask a question for which every participant would answer **N**? In this case, we expect $y = \frac{n}{2}$ (everyone answers according to the coin flip, and we expect half the coin flips to be heads):

$$p = \frac{2\frac{n}{2}}{n} - 1 = \frac{n}{n} - 1 = 1 - 1 = 0$$

Again, we arrive at a reassuring answer.

Figure 3 on the following page shows the relationship between the number of **Y** answers and the estimated true population fraction (p) — but how can p be negative? Let's look at Equation 1 on the previous page again. When will it emit negative values for p?

$$\begin{aligned} \frac{2y}{n} - 1 &< 0 \\ \frac{2y}{n} &< 1 \\ 2y &< n \\ y &< \frac{n}{2} \end{aligned}$$

Equation 1 on the preceding page emits negative values for p when $y < \frac{n}{2}$. We have already discovered this case! When everyone in the study has a truthful N answer, about half of them will write down **Y** due to the coin flip. We don't expect the heads-to-tails ratio of *all* the coin flips to be *precisely* 1:1, so there may be a question where the **Y** answer count is less than $\frac{n}{2}$ — this simply indicates that nobody in the study had a Y answer to the

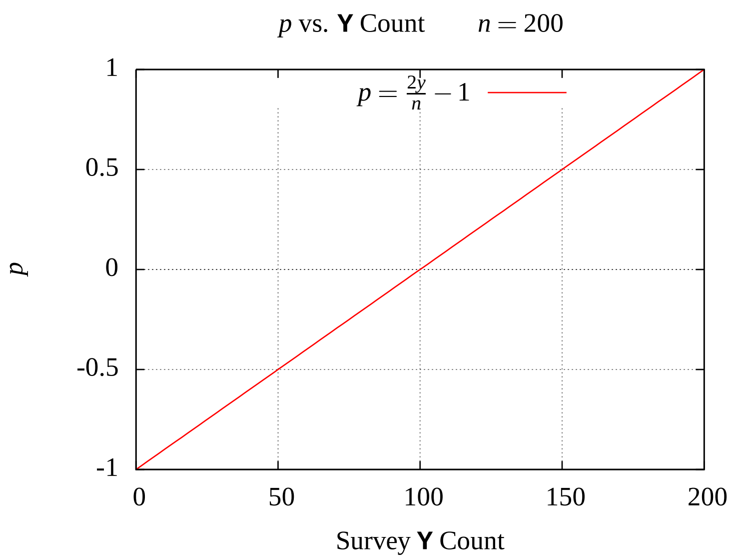


Figure 3: A plot of the study estimated Y fraction (p) against the number of **Y** answers. The graph is for a study of size $n = 200$. The red line shows that a count of about 125 **Y**s suggests a population Y fraction of 25%.

question.[5]

If you are concerned about obtaining usable results from a classroom survey,[6] consider the graphs in Appendix A on the next page and the confidence intervals you would like for the results. You may need to have multiple classes take the same survey, and use questions with a high anticipated p to improve accuracy (they don't need to be embarrassing questions). Keep in mind that this type of study is usually implemented with ns on the order of one or two thousand participants.

4 Simplified Equations

Equation 1 does not have to be derived by a middle-school classroom. Instead, you may wish to keep the analysis grounded in the straightforward logic of the algorithm and the mathematical computations that flow from it:

$$y = \overbrace{pn}^{\text{truthful } \mathbf{Y}} + \underbrace{\frac{1}{2}\overbrace{(1-p)n}^{\text{truthful } \mathbf{N}}}_{\mathbf{Y} \text{ from fair coin flips}}$$

After your survey has been tallied, you will know n, suppose it is $n = 150$.

$$y = 150p + \frac{1}{2}(1-p)150 \qquad (2)$$

Have them prepare a table of y versus p and graph the relationship in Equation 2. Then find p for each survey question graphically.

For each question, y is known. For instance, if $n = 150$ and $y = 102$, then

$$102 = 150p + \frac{1}{2}(1-p)150 \qquad (3)$$

[5]There is a subtlety in this derivation that must be mentioned. We are treating y and **Y** as the same number, but they are truly different. y is the *expected number of* **Y***s given* p, whereas **Y** is simply the number of **Y**s for a survey question. y is a theoretical value, **Y** is a real, tangible number. It turns out that **Y** represents a very good *estimator* of y, I believe it would be the maximum likelihood estimate (although I haven't bothered with the derivation), so we use the two terms interchangeably.

[6]Or in other words, knowing the inherent uncertainty in your survey results.

Have them use calculators and attack the problem numerically. Beginning at a number for p near 0.5 and successively trying better guesses until Equation 3 on the previous page is satisfied.

I think the two best simplifications have been suggested by teachers at CCTM. The first is to simply subtract the count of question **N**s from the survey **Y**s to determine the true number of Ys. Then determining p is straightforward:

$$p = \frac{\textbf{Y}\text{s} - \textbf{N}\text{s}}{n}$$

The second embraces the "let the algorithm guide them" notion to the hilt. Let the students figure out their own technique for determining p from the count of **Y**s and **N**s. Some may find it via guess and check, some may find a linear relationship, and some may subtract the **N**s.

A Study Confidence Intervals

Since coin flips are used,[7] there is always a little bit of error in the measurement of p. The N branch of any study question is more likely to have an uneven-even heads-to-tail split than a perfect 1:1 result. Despite this "wiggle" or uncertainty, statistics is able to place bounds on the confidence of these study results.

The number of **Y** answers is given by

$$y = \overbrace{pn}^{\text{truthful } \textbf{Y}} + \underbrace{\frac{1}{2}\overbrace{(1-p)n}^{\text{truthful } \textbf{N}}}_{\textbf{Y} \text{ from fair coin flips}}$$

The $\frac{1}{2}$ in this equation is a simplification, because a perfect heads-to-tail split from the coin flips is unlikely. This number is actually a random variable representing the number of heads from $(1-p)n$ fair coin tosses, which is a binomial distribution with $(1-p)n$ trials and a success probability of $\frac{1}{2}$. The mean and standard deviation of a binomial distribution is (b subscripts are used to distinguish from similar variable names elsewhere in the document):

$$\mu = n_b p_b$$
$$\sigma = \sqrt{n_b p_b (1-p_b)}$$

In this case, $n_b = (1-p)n$ and $p_b = \frac{1}{2}$,

$$\mu = \frac{1}{2}(1-p)n$$
$$\sigma = \sqrt{(1-p)n\frac{1}{2}(1-\frac{1}{2})} = \frac{1}{2}\sqrt{(1-p)n}$$

A 95% confidence interval ($z = 1.96$) is given by

$$\mu \pm z\sigma$$

$$\frac{1}{2}(1-p)n \pm \frac{z}{2}\sqrt{(1-p)n}$$

And we may rewrite our expression for y as

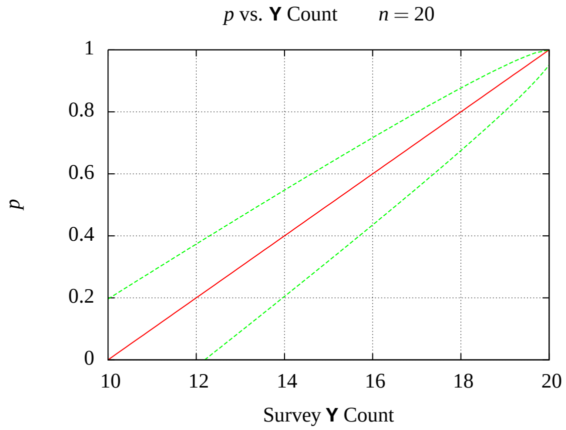
$$\begin{aligned} y &= pn + \left[\frac{1}{2}(1-p)n \pm \frac{z}{2}\sqrt{(1-p)n}\right] \\ &= pn + \frac{n}{2} - \frac{pn}{2} \pm \frac{z}{2}\sqrt{(1-p)n} \\ &= \frac{pn}{2} + \frac{n}{2} \pm \frac{z}{2}\sqrt{(1-p)n} \end{aligned}$$

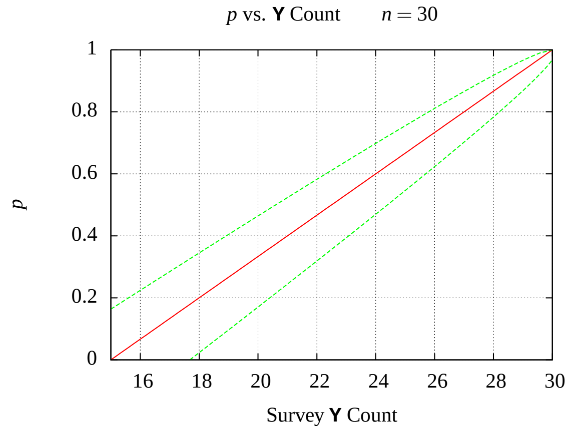
[7] You could also use TI calculators' random number generators, but you will still have this problem.

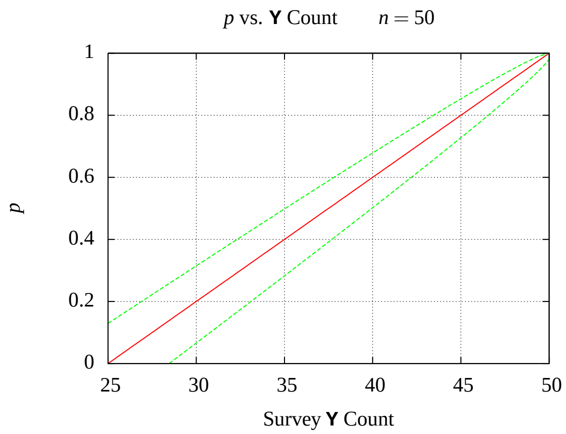
The following graphs show the original y vs p relationship (red line) with 95% confidence intervals (green lines) for varying study sizes (n). Note how both the study size (n) and the fraction of the population answering Y affect the confidence intervals.

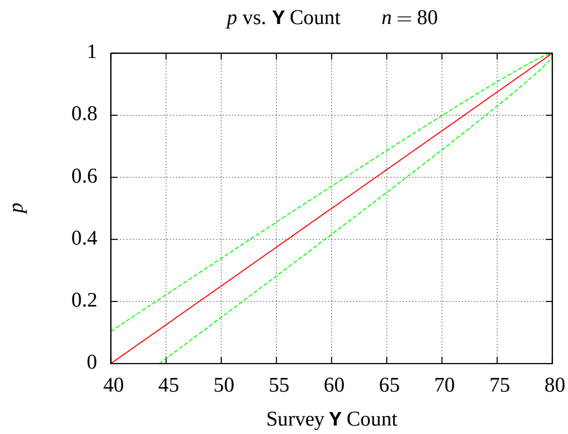
If a question in a study with 80 participants ($n = 80$) had 55 **Y**s, the interpretation would be:

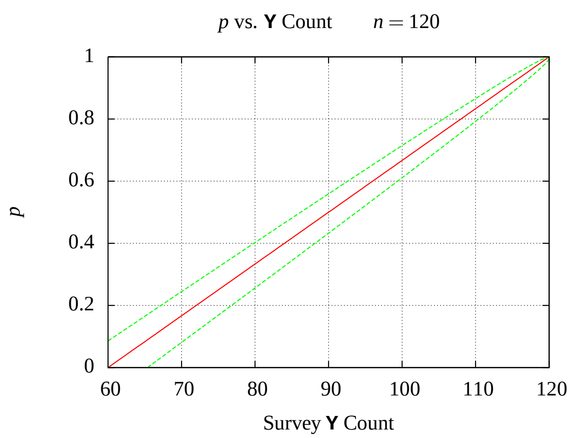
- p (for this question) is near 0.38
- if the study was repeated 100 times, we would expect other estimated values of p to fall within the interval $0.28 < p < 0.46$ for 95 of the 100 studies.

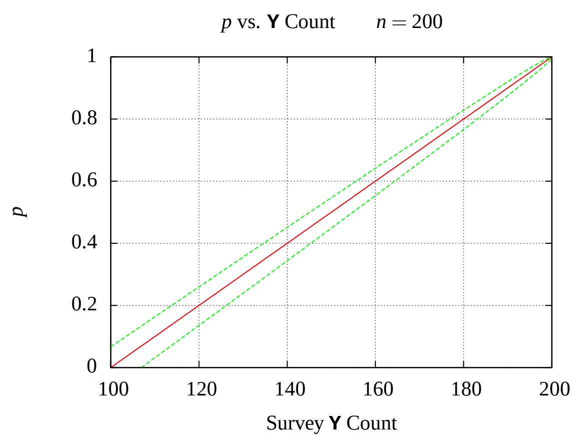

p vs. Y Count n = 20
1
0.8
0.6
0.4
0.2
0
p
10 12 14 16 18 20
Survey Y Count

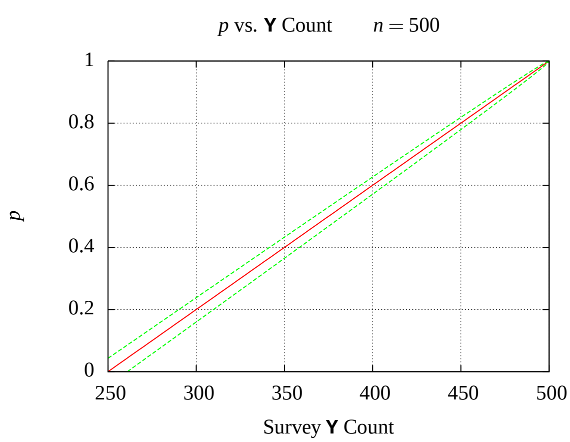

p vs. Y Count n = 30
1
0.8
0.6
0.4
0.2
0
p
16 18 20 22 24 26 28 30
Survey Y Count

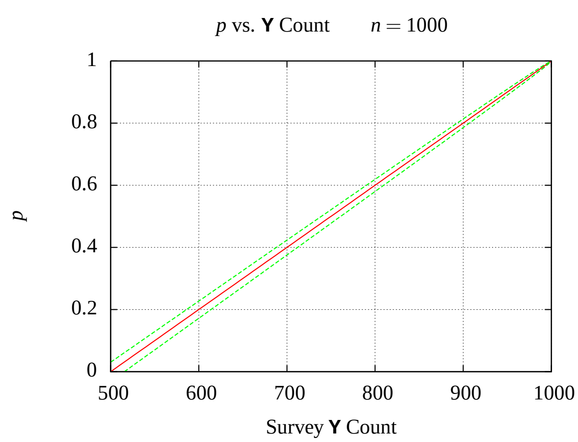

p vs. Y Count n = 50
1
0.8
0.6
0.4
0.2
0
p
25 30 35 40 45 50
Survey Y Count


p vs. Y Count n = 80
1
0.8
0.6
0.4
0.2
0
p
40 45 50 55 60 65 70 75 80
Survey Y Count


p vs. Y Count n = 120
1
0.8
0.6
0.4
0.2
0
p
60 70 80 90 100 110 120
Survey Y Count


p vs. Y Count n = 200
1
0.8
0.6
0.4
0.2
0
p
100 120 140 160 180 200
Survey Y Count


p vs. Y Count n = 500
1
0.8
0.6
0.4
0.2
0
p
250 300 350 400 450 500
Survey Y Count


p vs. Y Count n = 1000
1
0.8
0.6
0.4
0.2
0
p
500 600 700 800 900 1000
Survey Y Count