



Marcello Pagano

# [JOTTER 8 SURVEY SAMPLING]

Stratified sampling, cluster sampling, biases, randomized response.

1996 Presidential Election				
Poll	N	Clinton	Dole	Perot
Harris	1339	51	39	10
ABC-Wash Pst	703	51	39	10
CBS-NYT	1519	53	35	12
NBC-WStJ	1020	49	37	14
Gall-CNN-USA	1200	52	41	7
Reuters/Zogby	1200	44	37	19
Pew Research	1211	49	36	15
Hotline/Battlg	1000	45	36	19
Actual	96m	49.3	40.7	10

Possibly the most popular use of surveys, especially during presidential campaigns, is to make predictions about the outcome of the election. In the political realm, such surveys are called polls.

If we look at the 1996 presidential elections, when President Clinton was running against Senator Dole with Perot as a third candidate, we have here eight polls, each represents the last poll taken by that company before the actual election.

In the actual election, 96 million people voted, with Clinton getting 49.3% of the vote, Dole getting 40.7% of the vote, and Perot getting 10% of the vote.

When we compare the poll results to the actual numbers who voted for each candidate, given in the bottom line, we see a good agreement.

But the fascinating thing is evident when we look at the Harris poll, for example. The Harris poll was based on asking 1,339 people to give us their opinion of how they would vote. Now this is 1,339 versus the 96 million who actually voted. How can 1,339 people predict what 96 million people are going to do? Indeed, the mean sample size for these eight polls is 1,149. So based on what just over a thousand people say, these pollsters predict how 96 million people will vote, at some future time.

This is the magic. How well did the pollsters do in subsequent elections?

<sup>1</sup> <http://www.ncpp.org/files/1936-2000.pdf>

2000 Presidential Election				
Poll	N	Gore	Bush	Nader
Harris (phone)	1348	47	47	5
ABC-Wash Pst	826	45	48	3
CBS	1091	45	44	4
NBC-WStJ	1026	44	47	3
Gall-CNN-USA	2350	46	48	4
Reuters/Zogby	1200	48	46	5
Pew Research	1301	47	49	4
Battleground	1000	45	50	4
Actual	96m	48	48	3

When we look at the results for the 2000 elections, once again 96 million had their votes counted. The reported results had them at a virtual tie; 48% each.

And once again, on the basis of an average of 1,268 responders per poll, the predictions were rather good.

2004 Presidential Election		
Projector	Bush	Kerry
Harris	49	48
ABC-Wash Pst	49	48
CBS	49	47
NBC-WStJ	48	47
USA/Today/Gall	49	47
Zogby	49.4	49.1
Pew Research	51	48
Battleground	51.2	47.8
Actual	50.75	48.3

<sup>2</sup> <http://www.pollingreport.com/wh2gen1.htm>

<sup>3</sup> <http://www.pollingreport.com/2004.htm>

Once again in 2004, where I have not included the number polled, but they remain at roughly what they were in the previous years.

CANDIDATE ESTIMATE ERROR -Preliminary Report		Start Date	End Date	Voter Sample	MoE + / -	Obama	McCain	U n
UNOFFICIAL RESULT - 11/21/08						52.7%	46.0%	
<b>Projections-Uncideds Allocated</b>								
1	GWU/Battleground-Tarrance(R)	2-Nov	3-Nov	800	3.5%	50%	48%	
2	GWU/Battleground-Lake(D)	2-Nov	3-Nov	800	3.5%	52%	47%	
3	Rasmussen	1-Nov	3-Nov	3,000	1.8%	52%	46%	
4	Investors Business Daily/TIPP	1-Nov	3-Nov	981	3.1%	52%	44%	
5	Harris Interactive (Internet)	30-Oct	3-Nov	3,946	1.6%	52%	44%	
6	Gallup /USA Today	31-Oct	2-Nov	2,472	2.0%	55%	44%	
7	McClatchy/Ipsos	30-Oct	2-Nov	760	3.6%	53%	46%	
8	Democracy Corps, GQR (Dem)	30-Oct	2-Nov	1,000	3.1%	53%	44%	
9	Pew Research Center	29-Oct	1-Nov	2,587	1.9%	52%	46%	
<b>Final Trial Heats-No allocation</b>								
10	Marist College	3-Nov	3-Nov	804	3.5%	52%	43%	
11	ABC News/Washington Post	31-Oct	3-Nov	2,904	1.8%	53%	44%	
12	Daily Kos (D)/Research 2000	1-Nov	3-Nov	1,100	3.0%	51%	46%	
13	American Research Group	1-Nov	3-Nov	1,200	2.8%	53%	45%	
14	NBC News/Wall St Journal	1-Nov	2-Nov	1,011	3.1%	51%	43%	
15	Zogby - Reuters	31-Oct	3-Nov	1,226	2.8%	54%	43%	
16	FOX News/Opinion Dynamics	1-Nov	2-Nov	971	3.1%	50%	43%	
17	CBS News	31-Oct	2-Nov	952	3.2%	51%	42%	
18	Hotline-Diageo/FD	31-Oct	2-Nov	887	3.3%	50%	45%	
19	CNN/Opinion Research Corp	30-Oct	1-Nov	714	3.7%	53%	46%	
ESTIMATE AVERAGE/POLL ERROR						52%	44%	4

The 2008 election had results very similar to the previous elections. Roughly the same poll sizes and roughly the same accuracy of the predictions when compared to what actually took place.

How is it that 1,000, or so, people can predict what 100 million will vote?

<sup>4</sup> [http://www.ncpp.org/files/08FNLncppNatIPolls\\_010809.pdf](http://www.ncpp.org/files/08FNLncppNatIPolls_010809.pdf)

Mark Blumenthal

**2012 Poll Accuracy: After Obama, Models And Survey Science Won The Day**

Posted: 11/07/2012 8:04 am EST


**Pollster Model Correctly Predicts Outcome in All States**  
Results as of Nov. 7, 2012, 5:01 a.m. ET

State	Electoral Vote		Pollster Estimate			Unofficial Election Result			
	State	Cumul.	Obama	Romney	Margin	% in	Obama	Romney	Margin
Pennsylvania	20	237	50.1	44.2	+5.8 D	99%	51.9	46.8	+5.1 D
Wisconsin	10	247	50.4	45.8	+4.7 D	99%	52.8	46.1	+6.7 D
Nevada	6	253	50.0	46.5	+3.6 D	98%	52.3	45.7	+6.6 D
Ohio	18	271	49.2	45.8	+3.4 D	99%	50.1	48.2	+1.9 D
Iowa	6	277	48.6	46.0	+2.6 D	99%	52.1	46.5	+5.6 D
New Hampshire	4	281	49.2	46.8	+2.4 D	90%	52.0	46.7	+5.3 D
Virginia	13	294	48.7	46.8	+1.9 D	99%	50.8	47.8	+3.0 D
Colorado	9	303	48.6	46.8	+1.7 D	71%	50.5	47.3	+3.2 D
Florida	29	332	48.4	47.9	+0.5 D	100%	49.8	49.3	+0.5 D
North Carolina	15	206	47.3	48.8	+1.6 R	100%	48.4	50.6	+2.2 R

This last election (2012) the pollsters also predicted down to the State level and got all fifty results correct. Indeed, this is the second presidential election when one pollster, Nate Silver<sup>6</sup>, achieved this feat. Here are the ten States most difficult to predict. These are the results given for the “modelers” who averaged individual pollster’s results. We know from the Central Limit Theorem (standard deviation versus standard error) that they should do better, but it is still amazing that they actually got all fifty correct.

<sup>5</sup> [http://www.huffingtonpost.com/2012/11/07/2012-poll-accuracy-obama-models-survey\\_n\\_2087117.html](http://www.huffingtonpost.com/2012/11/07/2012-poll-accuracy-obama-models-survey_n_2087117.html)

<sup>6</sup> <http://fivethirtyeight.blogs.nytimes.com/>



**NCHS**


The National Center for Health Statistics

The Vital Statistics Program,

The National Health Survey Program

<http://www.cdc.gov/nchs/>

The National Center for Health Statistics<sup>7</sup>, who maintain the Vital Statistics program for the US, have also taken, in the last half-century a number of surveys to measure and monitor the health of the population



**NCHS**

The first three of these national health examination surveys were conducted in the 1960s:

1. 1960-62—National Health Examination Survey I (NHES I);
2. 1963-65—National Health Examination Survey II (NHES II); and
3. 1966-70—National Health Examination Survey III (NHES III).


All 3 surveys had an approximate sample size of 7,500 individuals.

1. 1971-75—National Health and Nutrition Examination Survey I (NHANES I);
2. 1976-80—National Health and Nutrition Examination Survey II (NHANES II);
3. 1982-84—Hispanic Health and Nutrition Examination Survey (HHANES); and
4. 1988-94—National Health and Nutrition Examination Survey (NHANES III).

++

They started out early 1960s with the National Health Examination Survey. They repeated that survey twice more in that decade. Then they expanded to the National Health and Nutrition Examination surveys, starting in 1971 and extending into the mid-90s. These very important health surveys provided invaluable information, and the center continues to provide a wealth of statistical information. It is a wonderful resource, and well worth a visit.

<sup>7</sup> <http://www.cdc.gov/nchs/>




## DHS

Demographic and Health Surveys (USAID)

<http://www.measuredhs.com/>

Another important source of health related survey data—although this is much more global—is one funded by USAID, and it is the collection of Demographic and Health Surveys (DHS)<sup>8</sup>. Go visit them. Lauren has a presentation about this topic this week.



## Accuracy and Precision

- **Accuracy** – how close to reality the measure represents
  - Depends on how much bias
- **Precision** – how confident are we in the reproducibility of the results.
  - Depends on variability

As with all measuring instruments, and certainly surveys fall into that category, we need to know how trustworthy the results are. With inferential tools we like to consider two aspects, and we differentiate between them here: accuracy and precision.

---


<sup>8</sup> <http://www.measuredhs.com/>

Accuracy attempts to quantify how close to the truth the measure is. Another way of explaining this is to look at the opposite view and see whether we have bias in our measurement. For example, if we only survey men, then it is not really telling us about the population. At best the survey will only contain information about half the population. So it is biased in that respect.

Does the bias matter? It may or may not depending on what it is we are measuring; is there a sex difference in what it is we are measuring? The answer depends on the context, but typically we do not wish to take the chance that it does matter.

The precision associated with the survey tells us how confident we are in the reproducibility of the results. This very much depends on how variable the population is that we are attempting to measure. If everyone in the population is equal to each other, then we need a sample of size one, and it is very precise. So the precision depends upon the population variability and the size of the survey.

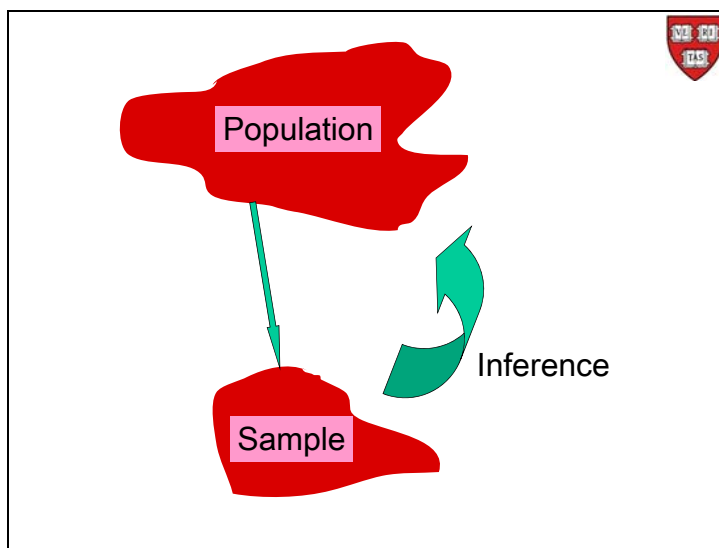
### Precision



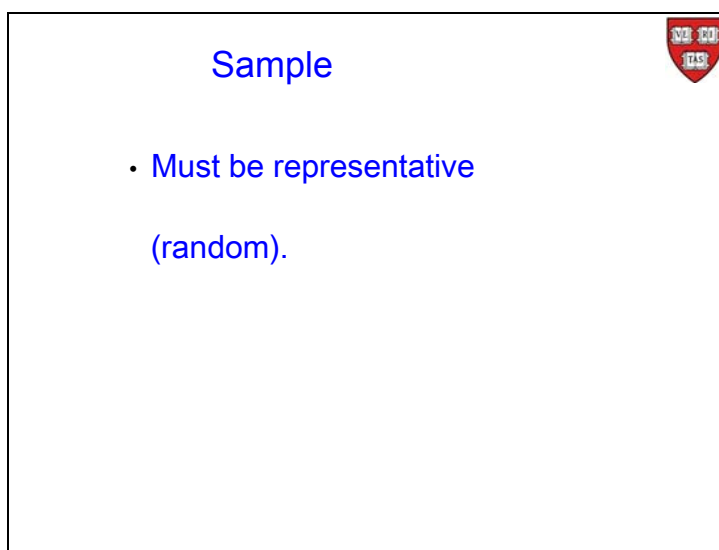
- Function of...
  - Sample size
  - Way that sample was selected
- Some designs yield more precise estimates than others.

Precision is a function of the sample size and, as it turns out, the way that sample was selected. Some designs yield more precision than others, and yet retain representativeness by incorporating external information we may have about the population under study. We elaborate on this point as we go along.





Returning to our introduction to inference, we said that we have a population from which we take a sample, and on the basis of this sample make inference about the population.



Up to now, we have taken a simple random sample from the population—everyone in the population has an equal chance of being chosen to be in the sample. The reason for this choice is that we decided that a representative sample was unobtainable in general, so we resort to random samples to obtain representativeness in aggregate.

## Sampling Theory

Up to now we have assumed:

1. Population infinite
2. Simple random sample

Suppose we take a sample of size  $n$  from a population of size  $N$  with mean  $\mu$  and standard deviation  $\sigma$

So far, we have not made any mention of the population size. We have been going along with what is called the infinite population assumption; namely that the population is huge. Let us investigate that assumption a little.

To this end, introduce some notation: suppose we are to take a sample of size  $n$  from a population of size  $N$  that has a mean  $\mu$  and a standard deviation  $\sigma$ .

Does the population size make any difference to our inference? And the answer is, yes, the population size is small, or the sample is big relative to the population size.

### Sampling fraction

Sampling without replacement  
(no duplication)

Sampling fraction:  $f = \frac{n}{N}$

Simple random sample:

Each individual has an equal chance,  $f$  of being included in the sample.

(More about this equiprobability later.)


What is important, when quantifying our inference, is the sampling fraction,  $f = n/N$ .

We now have the notation to tell us what we mean by simple random sample. It means that every person in the population has the same probability of being in the sample, and that probability is  $f$ . So let us say that we are going to take a sample of 100 people from our population of 1,000. Then  $f$  is equal to  $1/10$ , and that is the probability each person has of being in the sample. Technically, we are talking about *sampling without replacement*, which means that once a person has been chosen to be in the sample then that person is removed from the prospective pool of people to be subsequently chosen for the sample.

This chance of  $f$  is for simple random sample. With some sampling designs we now consider, this may not hold, even if the sampling is random at some level, different individuals will have different probabilities of being chosen. These probabilities are sometimes called weights.

The number  $f$  is called the *sampling fraction*.

### Finite population correction factor



Central limit still holds, except:

$$s.e. = s.d.(\bar{X}) = \frac{\sigma}{\sqrt{n}} \sqrt{1-f}$$

$$\approx \frac{\sigma}{\sqrt{n}} \text{ if } f \approx 0$$

So, if population is huge,  $f \approx 0$   
and population size is not important.

That means that  $n$  and not  $f$  determines the precision.

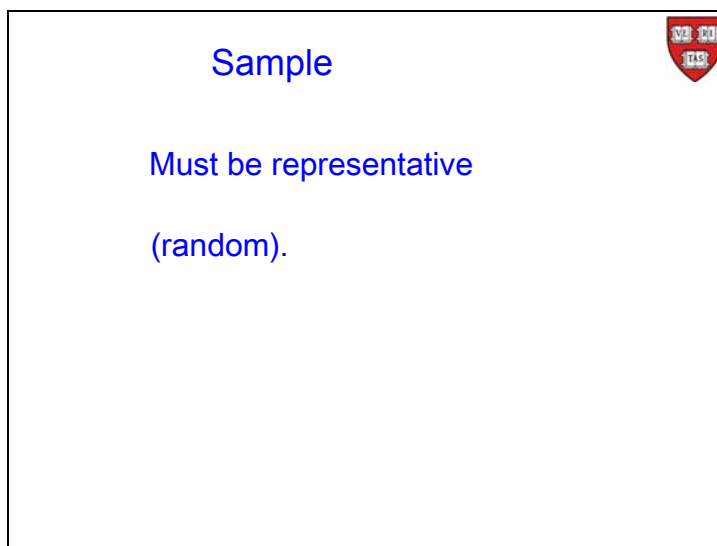
The central limit theorem still holds even when sampling from a finite population. The difference is that we must make the population size explicit. This manifests itself in the formula for the standard deviation of  $\bar{X}$ —namely the standard error—is actually as shown above.

We recognize the  $\sigma/\sqrt{n}$ ; that is what we had before. But now we have what is called the *finite population correction factor*:  $\sqrt{1-f}$ . If  $f \approx 0$ , then we are back to the formula we have become accustomed to, and there is nothing new to be learnt. So if the population is very large and the sample does not represent a sizable fraction of the population, then we can assume that  $f$  is approximately zero and ignore it.



In summary, if we have a huge population—120 million voters in the US, say—and we are going to choose a sample of 1,000 people and ask them about how they are going to vote, then that gives us an  $f$  that is 1 over 120,000. So  $\sqrt{1-f} \approx 1$ . Thus it makes no practical difference to the value of the standard error, if we ignore this factor, or not.

The important point to notice is that the population size  $N$  only enters into the argument via  $f$ , so if we ignore  $f$  as being too small, then it is the sample size  $n$  that is important in the standard error formula and not the population size  $N$ . This is something that must seem counterintuitive, judging by the number of people who do not believe this at first—namely, if I am going to take a sample of size 1,000, it does not matter if I take this 1,000 from the City of New York whose population is approximately 10 million, or if I take that 1,000 from the US, whose population is 300 million or so, or if I take that 1,000 from the world, which has a population of 7 billion. It does not matter as far as the standard error of my estimate is concerned.

This is true as long as we are selecting a truly simple random sample. Remember that that means that everyone in the population has an equal chance of being selected, so that may be the cause if the reticence felt by some to the statement that it is the sample size that matters, and not the population size. Possibly, one is more willing to believe this assumption is satisfied with a smaller rather than with a larger population, such as everyone in the world, but I am speculating now.




Let us return to thinking about the requirements we have of the sample; namely, that it be representative. Intuitively, this makes sense since we are going to base our inference about the population on what we find out in the sample. What does representative really mean, since it feels like it should fit the bill perfectly?


## Representative Sample

Filed Under » [Statistics](#)



### Definition of 'Representative Sample'

A subset of a statistical population that accurately reflects the members of the entire population. A representative sample should be an unbiased indication of what the population is like. In a classroom of 30 students in which half the students are male and half are female, a representative sample might include six students: three males and three females.



### Investopedia explains 'Representative Sample'

When a sample is not representative, the result is known as a sampling error. Using the classroom example again, a sample that included six students, all of whom were male, would not be a representative sample. Whatever conclusions were drawn from studying the six male students would not be likely to translate to the entire group since no female students were studied.

<http://www.investopedia.com/terms/r/representative-sample.asp#axzz2APWlHYEF>

I went onto the web and I found a location called, Investopedia<sup>9</sup>. They define what they mean by a representative sample, and I quote, "it is a subset of a statistical population." Now I don't know what is meant by a statistical population, as opposed to just a population, but to continue, "that accurately reflects the members of the entire population."


The example they give is, in a classroom of 30 students in which half the students are male and half are female, a representative sample might include six students, three males and three females. I do not know what they mean by "might include." I suspect they do not mean that it might not include them either. I suspect they mean it should include three males and three females because that would then be representative of the population of 30 students; because half are male and half are female. It seems like to them a representative sample is a miniature version of the population.

They go on to say, "when a sample is not representative, the result is known as a sampling error." Do not believe everything you read.

They state, "Using the classroom example again, a sample that includes six students, all of whom are male, would not be a representative sample. Whatever conclusions were drawn from studying the six male students would not be likely to translate to the entire group since no female students were studied."

There are some things wrong with that statement. First, what they seem to be saying is that a representative sample must be 50% female and 50% male. This works fine with a single factor, such as sex—assuming we do not take an odd numbered sample!

<sup>9</sup> <http://www.investopedia.com/terms/r/representative-sample.asp#axzz2DoRjoP00>



	Female	Male
Number	15	15
Proportion	50%	50%

	Female	Male
Young	9 (60%)	6 (40%)
Older	6 (40%)	9 (60%)

Other factors, e.g. height, weight, .....

But why stop at a single factor. Suppose that in this classroom, we judge students as being young for that class or older for the class. And suppose the breakdown is as shown above.

So not only do we need to have a 50-50 sex split, but we also must have that amongst the females, we have that 60% are young and 40% are older, and amongst the males, we want that 40% are young and 60% are older, if the sample is to be representative.

So you can imagine what is going to happen once we start looking at additional factors: height, for example. Some are tall, some are short. Do we have to have the correct representation of height amongst the young females, the older females, the young males and the older males? How about weight? Should we have correct representation of weight distribution amongst the tall, young, females; amongst the short, young, females; etc..


Very quickly, we are going to be unable to satisfy these requirements.

There are two things wrong with this approach: One is the fact that once we start looking at more and more and more factors, it becomes more and more complex as we define more and more precise cells to divide up the population, and the sample, and we are not going to be able to find people to fit into each cell. A trivial example, if we stop at just two factors we need to fill four cells (young-female, older-female, young-male and older-male), but if we only need a sample of size three, we cannot proceed. In general if there are  $k$  factors and we assume, in order to simplify the argument, that each factor is only at two levels. Then we would need a sample of size exactly  $2^k$  to claim we have a representative sample.

Problem number two is how is it that we know this much about your population? To be truly representative we need to know how all these factors, assuming we have all the factors accounted for, interact with each other. For example, as above, how the age distribution varies with sex. Knowing this much about the population then raises the issue about the need for sampling?

The profession had this argument some 100 years ago, and consensus was reached that rather than rely on the “experts” to decree what is and what is not representative, we turn to randomization to provide a sample. This then has the advantage that we can quantify the precision and accuracy with which we make our inference. That is not to say that we cannot incorporate important information into our sampling methods, as we show below.

Further, we must be careful in how we actually communicate with individuals. For example, if we carry out a phone survey, then, obviously, only people who own phones can be sampled. Second, we know that the ownership of cell phones and/or land lines are very much dependent on the age, and other demographic factors, of the owner<sup>10</sup>. So if we ignore this fact—for example, if we ignore people reachable only through cell phones—then the resultant survey might well be biased towards one group or another.



If you have knowledge to bring to the table, by all means use it—for example, if you know you have 50% female and 50% male, and sex will impact the outcome of interest, then by all means ensure that your sample is 50% female and 50% male.

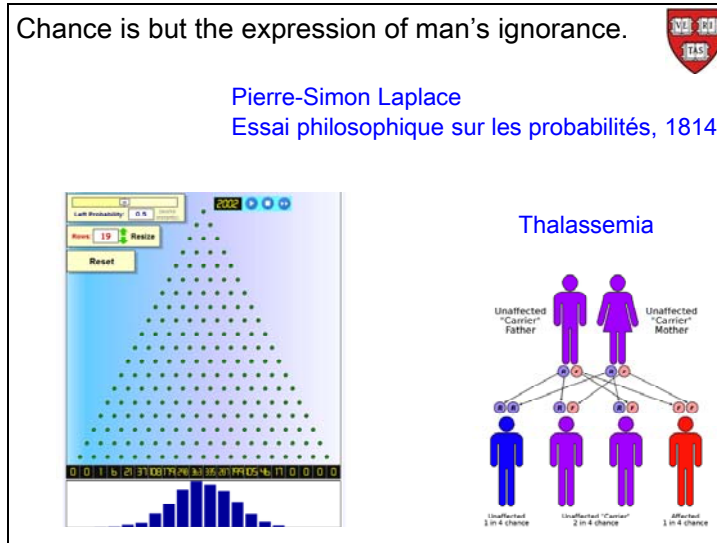
Once you have satisfied your marginal constraints, we turn to,

Simple Random Samples

The point of all this is if you have knowledge to bring to the table, by all means, bring it. For example, if you have 50% female and 50% male, and sex will impact your outcome of interest, then make sure that your sample is 50% female and 50% male. Do not leave it up to chance because chance is not going to give you a 50-50 split every time.

But be careful, as argued above, do not impose too many of these constraints, because you quickly run out of degrees of freedom. Further, once you impose a constraint you lose the ability to estimate the prevalences of those factors. For example, if you impose a 50-50 sex ratio in your, you can no longer estimate the sex ratio in the population.

<sup>10</sup> <http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201112.htm#differences>



Laplace said, "chance is but the expression of man's ignorance." And here we are aiming for a simple random sample and rely on chance as a foundation for our inference!

We know that probability is not very helpful, except to provide uncertainty, in the short run, and the other reason why it is attractive is that it provides us with predictable, steady behavior in the long run. We saw this repeatedly with the Quincunx, and in real life we looked at the long run behavior of families with Thalassemia.

We bank our inference on this long run stability.

### Sampling Frame

In order to ensure that everyone has an equal chance of being sampled, we need what is called a sampling frame; or an itemization of every person in the population.

If the two are not the same (population and sampling frame) then we are actually taking a random sample from the sampling frame.

To obtain a random sample we need to ensure that everyone has an equal chance of being chosen. This is achievable if we have a list of everyone in the population. That list is called the



*sampling frame*. In a sense the two are inextricably related; the population ideally defines the sampling frame, but operationally the sampling frame defines the population from which we can sample.


For example, if you wish to make inference about everyone who lives in a city, your list may consist of all the home addresses in the city. That means you do not have homeless people represented in your sample, and as a result you do not have a sample of everyone in the city.

Another way of stating that is that you may have a random sample of people who have homes in the city.

If your sampling frame consists of everybody who voted at the last election, you are not getting everybody in your town. If you are a jury clerk who chooses potential jurors, if you use such a list then not everyone in the city has an equal chance of being on a jury, which might be a requirement by law.

We also need a random device that can operate on the sampling frame and yield a random sample. Let us assume we have such. The government has rules that should be obeyed by random samples. Let us assume we obey those rules.<sup>11</sup>

### WEIRD Samples



Psychologists rely on Western, Educated, Industrial, Rich, and Democratic subjects.

“The findings suggest that members of WEIRD societies, including young children, are among the least representative populations one could find for generalizing about humans.”

J Henrich, SJ Heine, & A Norenzayan, The weirdest people in the world? *Behavioral and Brain Sciences* (2010) 33, 61–135

Sometimes the samples are not quite as random as we would like to believe. Psychologists speak about WEIRD samples<sup>12</sup>. What they mean is—WEIRD is an acronym for Western, Educated, Industrial, Rich, and Democratic subjects—that their subjects may not truly be as representative as they thought and as a result the theories that they have promulgated on the basis of their studies may be less generalizable than they thought.

<sup>11</sup> [http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards\\_stat\\_surveys.pdf](http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf)

<sup>12</sup> [Henrich J, Heine SJ, Norenzayan A](#). The weirdest people in the world? [Behav Brain Sci](#). 2010 Jun;33(2-3):61-83; discussion 83-135. Epub 2010 Jun 15.

Now we do something similar in our health studies. If in need of a hospital set of patients, our students typically go to teaching hospitals to find them. Most medical studies are also done at teaching hospitals. For example, if we need a sample of children, there is a children's hospital just down the street. It is a wonderful hospital. The research it performs is terrific. But are those children at that hospital representative of all children everywhere in the world?

Possibly health studies are not going to be as badly influenced as the social scientists who are mainly concerned with cultural events. There are a number of health studies, of course, that are very heavily dependent on cultural mores, but it will be interesting to find out as time evolves how dependent our medical studies suffer from this phenomenon.

### Stratified Sampling

Suppose we have information about our population that we wish to incorporate into our survey design and analysis. For example, the country we are surveying has provinces and we want to make inference both at the provincial level and at the national level. Or there may be structural information: for example, we may know that the population is half female and half male. Making use of this external information should prove beneficial in our inference about the population in improving both precision and accuracy. Further, when calculating information about the whole, it is more informative to along the way also be able to obtain information about subgroups that make up the whole.

For example, obtaining information for each province and then combining those provincial estimates to obtain an estimate about the country yields information about how the country aggregate is distributed amongst the provinces. This is an example of *stratified* sampling.

For example, if we have 6 provinces, and suppose we want a sample of size 60 for our country, then taking 10 from each province will give me information about each province. Whereas, had I taken a simple random sample of 60 from the country, it is possible (more than 1 in a thousand) that I get zero, or one person from one of the provinces. So stratified sampling seems more informative than simple random sampling.

Further, there is no dictate that says we need to take the same number of individuals from each province. Intuitively, we may wish to spend more effort (larger sample) in a province that is more variable than one that is more homogeneous.

These are the some of the optimizations we can exercise with stratified sampling.

Note that all the Xs within a group need not be equal.

$$\begin{aligned}
 \bar{X} &= \frac{1}{6}(1+2+3+4+6+8) \\
 &= \frac{1}{6}(\{1+2+3\} + \{4+6\} + 8) \\
 &= \frac{1}{6}\left(3\frac{\{1+2+3\}}{3} + 2\frac{\{4+6\}}{2} + 1\frac{8}{1}\right) \\
 &= \frac{1}{6}(3 \times 2 + 2 \times 5 + 1 \times 8) \\
 &= .5 \times 2 + .33 \times 5 + .17 \times 8 \\
 &= \sum_{i=1}^3 p_i \bar{X}_i = 4
 \end{aligned}$$

A central role in stratification is the construct that breaks the whole into parts and then recombines the results from each to get the result for the whole. Our initial focus is with the overall mean, or the total prevalence, so recall the composition formula that explicitly shows how to combine group means to obtain the overall mean. This formula plays a central role when dealing with stratified sampling.

### HIV Antenatal Clinic Surveillance (Sentinel or Convenient Sample)

ANC HIV surveillance has been carried out among women attending antenatal clinics in more than 115 countries worldwide. (2006 – 600 sites in sub-Saharan Africa)

Annually or bi-annually and they provide ready and easy access to a cross-section of sexually active pregnant women from the general population.

Used to “assess trends” in the epidemic over time.

In generalised epidemics, HIV prevalence among pregnant women has been considered a good approximation of prevalence among sexually active men and women aged 15–49 years.

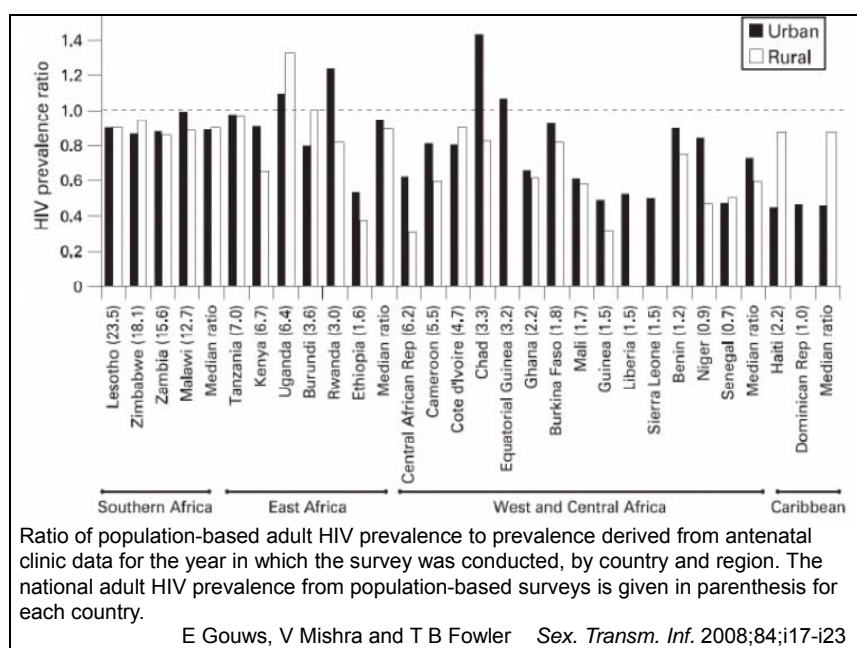
Considering the whole as made up of groups, or parts, is often quite informative because when we look at the composition formula we can see the effect of ignoring any groups. We need everybody to be represented. Consider a case in point, the estimation of HIV prevalence, and incidence, at a global level. A common practice in disease surveillance is to make use of

*sentinel systems*. A sentinel system concentrates on some pre-chosen hospitals, let us say, to use them as sentinels, or proxies, for what is going on in the population.

One that is very widely used, especially in sub-Saharan Africa, to monitor the HIV epidemic, makes use of information available from Antenatal Clinics (ANC). Women go to ANCs to get prenatal care, and part of prenatal care is to collect and store blood samples from the patients. These blood samples can be tested for the presence of the HIV and thus one can obtain an excellent measure of the prevalence, and incidence, of HIV infection amongst this group of women.

The difficulty arises when one attempts to use these measures to infer values for the population as a whole. A simple argument that this group of women represents the whole population is problematic in a number of ways: these are young, sexually active women who can access the ANC. So neither are men, nor are women in general represented—for example, women not represented include those not in their child bearing years, those who are not sexually active (that is, of course, a critical consideration when dealing with a sexually transmitted disease), and women who do not live near an ANC, or afford themselves of any ANC services.

Are these just theoretical considerations, or does it make a difference to our estimators?



E Gouws, V Mishra and T B Fowler *Sex. Transm. Inf.* 2008;84;i17-i23


<sup>13</sup> E Gouws, V Mishra and T B Fowler, Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data *Sex. Transm. Inf.* 2008;84;i17-i23

In this study the authors compare the results of the ANC Surveillance, to results from a DHS done at approximately the same time. The graphic shows the ratio of the population based prevalence (DHS) to that obtained from the ANC. The dotted line represents when the ratio is unity, and thus is the national average—reported parenthetically next to the country's name.

The solid blocks are for urban areas, and the open blocks represent rural areas. A few of these bars cross the unit line, but the great majority of them fall beneath that line. That means that the prevalences at the Antenatal Clinic are consistently, except for the four bars in the center, higher than the national averages, as we expect.

Some argue that, these ANC averages may not provide good estimates of the prevalence, but they do track the epidemic, and provide a good idea of how things are moving; in other words, a good estimate of the incidence.<sup>14</sup> For this to be true we would have to argue that the same mechanism that caused the prevalence to go up or down in sexually active women of child bearing age who have access to ANCs will have the same effect on the rest of the population. In order for this to be true, it would have to be a perfect storm that needs to remain in place throughout the life of the sentinel system.

Bias



UNAIDS/WHO estimates of national adult HIV prevalence have been based on prevalence data collected over time from pregnant women attending antenatal clinics.

But without an element of randomness, we cannot claim unbiasedness, and we cannot measure the bias.

Modest proposal to retain the knowledge in the convenient sample but introduce unbiasedness.

Hedt & Pagano Statistics in Medicine 2010

What we have is a biased system since we are only measuring a subset of the population. Can this bias be removed? And the answer is yes, but it requires some sampling, preferable random, of the rest of the population. We can take advantage of the information from the ANC, especially since the economics of the situation make large samples readily available from the ANC. The methods are covered in this report<sup>15</sup>.

<sup>14</sup> This hope that two (prevalence) curves that are not equal will have the same derivatives (incidences) associated with them, can be labeled wishful thinking.

<sup>15</sup> [Hedt BL, Pagano M](#). Health indicators: eliminating bias from convenience sampling estimators. [Stat Med](#). 2011 Feb 28; 30(5): 560-8.

## Stratified Sampling



Aim: Increase precision for same cost.

Suppose the population is made up of  $g$  groups, with group

Sizes:	$N_1$	$N_2$	...	$N_g$
Means:	$\mu_1$	$\mu_2$	...	$\mu_g$
Std. Devs.	$\sigma_1$	$\sigma_2$	...	$\sigma_g$
samples	$n_1$	$n_2$	...	$n_g$
Sample means	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_g$

Consider stratified sampling in general. Suppose our population is composed of  $g$  groups, or strata. And suppose that, for each stratum, say the  $i^{\text{th}}$ , we know the size of the stratum,  $N_i$ , but we do not know its mean,  $\mu_i$ , or its standard deviation  $\sigma_i$ . We are interested in estimating these parameters by taking random samples of size  $n_i$  from each of the  $g$  strata. Let us also assume that the population size is  $N$  and the total sample size is  $n$ .

## Stratified Sampling



$$\mu = \sum_{i=1}^g \frac{N_i}{N} \mu_i$$

So, estimate by

$$\hat{x} = \sum_{i=1}^g \frac{N_i}{N} \bar{x}_i$$

whose variance is

$$\sum_{i=1}^g \left( \frac{N_i}{N} \right)^2 \frac{\sigma_i^2}{n_i} \left( 1 - \frac{n_i}{N_i} \right)$$


which is minimized by choosing.

We are interested in the overall population mean, which from our composition formula we know we can express as above. So it makes sense that we estimate this overall mean by the weighted average,  $\hat{x}$ , we show above. It is a weighted average of the individual strata means.

We see that the weights are proportional to the size of the population within each stratum. Thus if we are to use this estimator, we need to know the stratum sizes.

We can also calculate the variance of this estimator, whose square root is the standard error. We have control of the sample sizes from each stratum, so if we want to choose a design to minimize this standard error, we can modify the individual sample sizes—a straightforward mathematical problem.

**Stratified Sampling**



$$n_i = n \frac{N_i \sigma_i}{\sum_{i=1}^g N_i \sigma_i}$$

Thus, to maximize precision, the sampling fraction in each stratum should be proportional to the standard deviation in that stratum, i.e.

$$\frac{n_i}{N_i} = \sigma_i \times \text{constant}$$

If the cost per observation is  $c_i$ , then to maximize the precision for a fixed cost:

$$\frac{n_i}{N_i} = \frac{\sigma_i}{\sqrt{c_i}} \times \text{constant}$$

Here is the solution: to minimize the standard error of the estimator of the population mean, choose the sampling fraction in each stratum to be proportional to the standard deviation of the population in that stratum. If the sampling cost varies between strata, then minimizing the standard error within a fixed cost leads to the solution shown above.

Intuitively, cost aside, this answer makes sense. What it says is spend your effort where the variability is greatest. At an extreme, for example, if everyone within a stratum were equal to each other, we need only take a single observation from that stratum. This is the statistical version of the squeaky wheel getting the attention.

## Sampling Weights

### Sample Weights



If the probability that a person is sampled is not  $f$ , then this must be reflected in the analysis.

e.g. Suppose we have 2 strata, the first is of size  $2n$  and the second is of size  $n$ . If we take a single sample from each stratum ( $x_1$  and  $x_2$ ), then each can be said to “represent” their stratum ( $2n$  and  $n$ ). Any subsequent estimator using both observations must reflect this. For example, if we want to estimate the overall mean, then we would give  $x_1$  weight  $2n$  and  $x_2$  weight  $n$ . So the estimator would be,

$$\bar{x} = \frac{2n x_1 + n x_2}{2n + n}$$

In summary then, let us talk about sample weights. We stated that for simple random samples,  $f$ , the sampling fraction, is the (same) probability that anyone in the population is in the sample. This may vary for other sampling schemes. We deviate from this equality when we use other sampling designs. For example, with a stratified scheme, the probability of each individual being in the sample is not constant, but varies across individuals.

This variability in the probability of being in the sample matters when it comes to the calculation of the estimators of the population parameters. The probability with which a person is chosen to be in the sample should be reflected in the formula for the estimator, just as we saw it was in the calculation of  $\hat{x}$ , above in the case of stratified sampling. This allowed us to ensure that the mean of the sampling distribution of  $\hat{x}$  is  $\mu$ , the population mean—we label this property unbiasedness.

Here is a simple example. If we have two strata, and suppose that the size of the first stratum is  $2n$ , and the size of the second stratum is  $n$ ; so one stratum is twice as big as the other. Now, take a single sample from each of the strata.

Now the observation from the first stratum can be thought of as representing  $2n$  people, whereas the one from the second stratum only represents  $n$  people. So intuitively the one from the first stratum should carry twice as much weight as the one from the second stratum in the estimation of the population mean.

This is the rationale behind sampling weights. So, when analyzing sample surveys, sometimes, such as in DHS surveys, each observation comes with its own sampling weight. In a sense that weight represents how much information this one observation has relative to the others.



### How to determine strata?



To **maximize precision** of estimation, construct strata so that:

1. Their averages are as different as possible.
2. The standard deviations within a stratum are as small as possible.

If we have any latitude in determining the strata, then how should we choose them? If we want to maximize the precision of the estimation, then choose the strata to be as homogeneous as possible within a stratum, and as different, or heterogeneous as possible across strata.

### Cluster Sampling

#### Cluster Sampling




<http://phil.cdc.gov/phil/details.asp?pid=7224>

Another common sampling strategy is to use cluster sampling. The idea of cluster sampling is similar, and at the same time dissimilar, to stratified sampling. The idea is roughly to create

clusters that resemble the population, and then sample a few of them to measure, and then measure each one thoroughly. This typically turns out to be much cheaper than stratified sampling, and not as accurate.

One of the biggest cluster sampling designs in the 20<sup>th</sup> century was used to overcome one of the biggest scourges we had in the 20<sup>th</sup> century, but which declined precipitously in the second half of the 20<sup>th</sup> century, and that is polio. The challenge was how to test the polio vaccine, once it had been introduced? They decided to use a cluster design to test 1.8 million children (440,000 received the vaccine, 210,000 received a placebo, and 1.2 million served as controls) the biggest clinical trial in history<sup>16</sup>.

**Cluster Sampling**



Substantial loss in precision but cheaper.

1. Divide the population into clusters (just like strata).
2. Choose which clusters to measure (total or sample).

For maximum precision form clusters so that individuals within a cluster vary as much as possible.  
(Two-stage, ... )

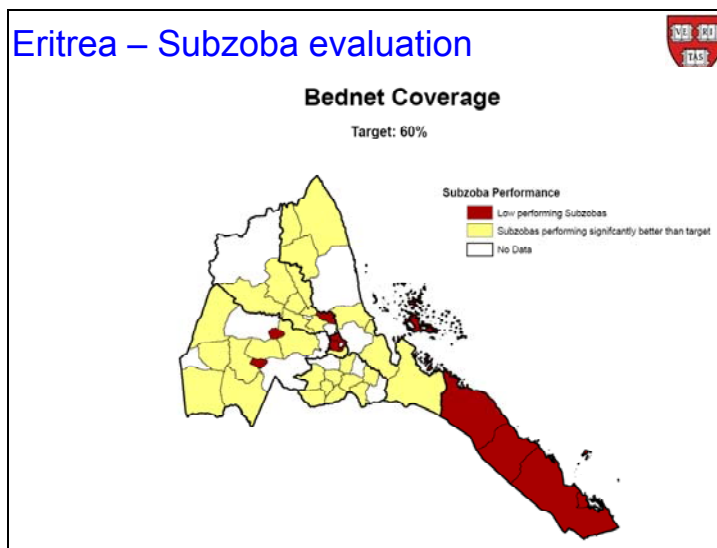
There is a substantial loss in precision in a cluster sample, but it is much cheaper than either simple random sampling or stratified sampling. One starts by dividing the population into clusters, just as we did for stratified sampling. Now, instead of choosing every single one of those clusters, you choose which clusters to measure. So you might choose one, two, three clusters. And then you measure those clusters.

After some thought, if you are only going to measure a few clusters, then if you want a good representation of the population, you would want each of the chosen clusters to be as representative of the population as possible. So you want to maximize variability within a cluster; in contrast to stratified sampling where we sought homogeneity within a stratum.

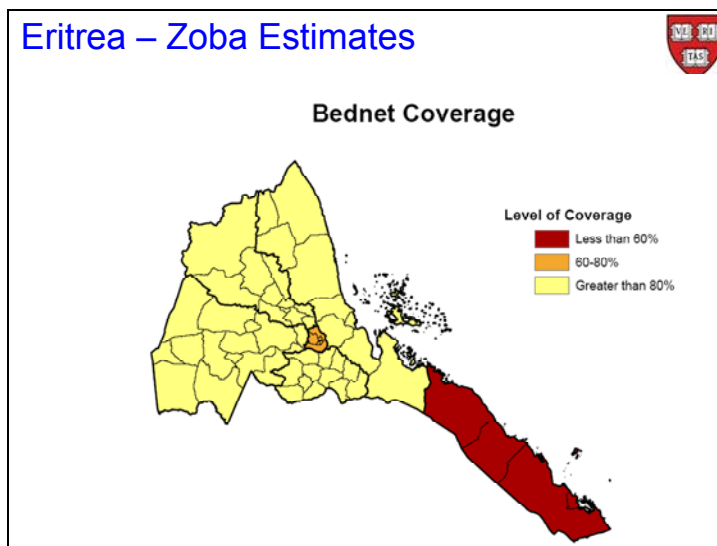
One can also do what is called two-stage cluster sampling. There you choose some clusters, and then within those clusters you might take a simple random sample and not sample everyone within a cluster.

Indeed, you can mix and match as many layers of clustering and stratification and simple random sampling as you wish.

<sup>16</sup> American Journal of Public Health, Volume 45, Issue 5\_Pt\_2 (May 1955)



Here is an example of cluster sampling. In the struggle against malaria in Eritrea, they needed to evaluate the use of bed nets countrywide. The target they set was 60% coverage, and they wanted to classify each locale as having reached the target or not. This map is at the subzoba level. The subzobas are classified into one of three colours: red, low performance ( $<60\%$ ) ; yellow, high performance ( $\geq 60\%$ ); and white, no data. These classifications were based on a simple random sample in each of the subzobas chosen.



The subzobas can then be combined to form zobas, and the sampled subzobas provide estimates for the zobas.



People in the same cluster are more likely to be similar to one another than to people in a different cluster.

So choosing another person from the same cluster will not be as informative as choosing someone from another cluster.

e.g. In a malnutrition study, choosing 3 children in the same household is not as informative as 3 children at random in a village, which in turn is not as informative as 3 children at random in a province, etc.

The **design effect** (DEFF) is the ratio of the variance under the design used, to the variance with simple random sampling, e.g.  $DEFF = 2$  means you need twice as big a sample to get the same variance as you would get with a simple random sample.

One of the problems with cluster sampling is that people in the same cluster are more likely to be more similar to one another than they are to people in different clusters.

So for example, one (cluster) design might consider each household a cluster and we sample every child between the ages of two and five in that household. In contrast, we might consider the household as the sampling unit and only possibly (assuming such a child exists in the household) sample one child between the ages of two and five within each household. It is not unreasonable to believe that kids within a household are more likely to share common characteristics than kids across different households. For example, when doing a vaccination study, if the parents are responsible enough to have one child vaccinated, chances are they will get all the kids vaccinated. Or kids within the same village, when doing a malnutrition study, are more likely to look similar than kids in other villages. So there is a correlation, if you will, between people within the same cluster.

So here is the problem, suppose you have sampled one kid. If you now choose the second kid in the same household or the same village, it is going to be much cheaper, because you do not have to go across town, or to another village to choose another kid. But because of the shared genes or environment, it is not going to be as informative as choosing that second kid across town or in another village.


So how much do you lose? Well, it all depends on how close the kids are to each other; how related they are to each other. If you have a cluster of physicians in a clinic who were all trained at the same school by the same teachers, chances are they are all going to act the same way when faced with the same patient.

To quantify this effect statisticians have come up with is what is called the DEFF; called the *design effect*. The design effect is the ratio of the variance under the design used to the variance with simple random sampling. For example, cluster sampling is going to give you a bigger variance because you don't have as good representation, so the DEFF is going to be greater than one.

For example, if the design effect is two, so the ratio of the variances is two, that means you will need twice as big a sample with your cluster design than you would with a simple random sample to get the same precision. So this is the basis for differentiating between designs when considering costs and accuracy.

Design effects of one-and-a-half to two, or three are usually what one experiences in the field when you have a well-designed study. But big DEFFs have been documented too, all the way up to 30, and you can understand why, because of the relationship between people within the same cluster.

### Sources of error



So far we have been talking about biases & imprecision caused by *sampling* variability.


Other sources of error:

1. Selection
2. Non-response
3. Recall
4. Lying

The generic survey involves asking selected people questions and recording their answers. As a result, different types of errors can crop into surveys, beyond what we have learnt to expect due to sampling variability.

Let us consider just four types of possible error: selection bias, non-response bias, recall bias, and the simplest one that occurs when the person surveyed lies.

## Selection



Literary Digest and Presidential Election of 1936

Landon	1,293,669
Roosevelt	972,897

Final Returns in The Digest's Poll of Ten Million Voters

Good reputation predicting 1920 - 1932 elections.

(Majority of respondents had voted for Hoover!)

See autopsy results.


Selection bias is as the name implies due to the fact that we were biased in who we chose to be in the sample. We have already mentioned the WEIRD sample phenomenon as well as the ANC HIV surveillance. Another, classic example of selection bias is something that happened in 1936. The Literary Digest, a very reputable publication, predicted the outcome of the presidential election of 1936, and they had Landon, who got 1.3 million votes in their survey, to beat Roosevelt, who got only 972,000 votes in their survey. Of course, we know that they were wrong.

As a result of their wrong call, the Literary Digest actually went out of business. They had a terrific track record in predicting correctly every four years from 1916 to 1932, but with this one bad prediction they lost their credibility.

The problem is that this was quite predictable. You see, this was the Depression. They had sent out 10 million cards and about 2 million were returned. First, how did they choose the 10 million to whom they sent the cards and secondly which 2 million responded? Remember this was the depression, people tended to spend their money wisely. Further, they also asked how voters had voted in 1932, so they could see how well the responders represented the voting public.

And this is exactly the problem we had with Ray Pearl when he looked at autopsy results. Who gets autopsy? Do not say dead people. Not everybody who dies has an equal chance of being autopsied. That is what you need to ask yourself with any survey, are we getting a random sample of the population we think we are surveying.

### Non-response



Survey of sexual abuse of patients by US psychiatrists.

Surveyed	5,574
Responders	1,442 (hostile 19)
Response rate	26%

Admit having sex with patients:

Male	Female
1,057	257
7.1%	3.1%

``Psychiatrist-patient sexual contact: results of a national survey, I: Prevalence"  
 N. Gatrell et al. A.J. Psy. 143 (1986) 1126-31

The non-response bias is another important bias. Just as we saw with the Literary Digest, 10 million cards were sent out and only 2 million responded. Are those 2 million who did respond representative of the 10 million? Of the 8 million who did not respond? This is sometimes called the high school reunion effect. Who shows up at a high school reunion? Typically, it is the people who have been successful in life, and want to show off to their high school friends. They typically are not representative of those who do not show up.

Here is another example of this bias. A survey was carried out by Nancy Gatrell to determine the pervasiveness of the problem of psychiatrists sexually abusing their patients. So she sent out 5,574 questionnaires, and 1,442 responded, and that included 19 who just scrawled cuss words on the responses. So the generous response rate was just 26%.

Amongst the responders, 7.1% of the males admitted to having had sexual relations with their patients, and 3.1% of the females admitted to having had sexual relations with their patients.

Is it reasonable to think that these 7.1% and 3.1% rates can be used as estimators of the practice in general? What about the non-responders? Is it possible that one of the reasons why the 74% did not respond might be related to possibly self-admission of something that is expressly prohibited in the Hippocratic oath? In other words, might there be a relation between not responding and whether they had sexually abused their patients, or not?


If the 26% who responded acted in pretty much the same way as the 74% who responded, then there is no bias and the result of the low response rate is that the standard error will be based on 1,442 responders and not 5,574. If, on the other hand, there is some relationship between not responding and the outcome we are measuring, then the estimators based on the 26% are biased estimators of what we are seeking to measure.

### Benjamin Franklin 1759 :



As the practice of Inoculation always divided people into parties, some contending warmly for it, and others as strongly against it; the latter asserting that the advantages pretended were imaginary, and that the Surgeons, from view of interest conceal'd or diminish'd the true numbers of deaths occasion'd by Inoculation, and magnify'd the number of those who died of the Small-pox in the common way: It was resolved by the Magistrates of the town, to cause a strict and impartial enquiry to be made by the Constables of each ward, who were to give in their returns upon oath; and that the enquiry might be made more strictly and impartially, some of the partisans for and against the practice were join'd as assistants to the officers, and accompany'd them in their progress through the wards from house to house. Their several returns being receiv'd and summ'd up together, the numbers turn'd out as follows,

The trustworthiness of surveys is apparently an old problem. Benjamin Franklin<sup>17</sup> reports in 1759 on a survey he did to gauge the effectiveness of smallpox inoculation. The idea of whether or not to inoculate was a rather heated topic at the time, as it is again today! So they decided to do a study. The solution was, "So it was resolved by the magistrates of the town to cause a strict and impartial inquiry to be made by the constables of each ward. So they sent the constables out to do the survey. "



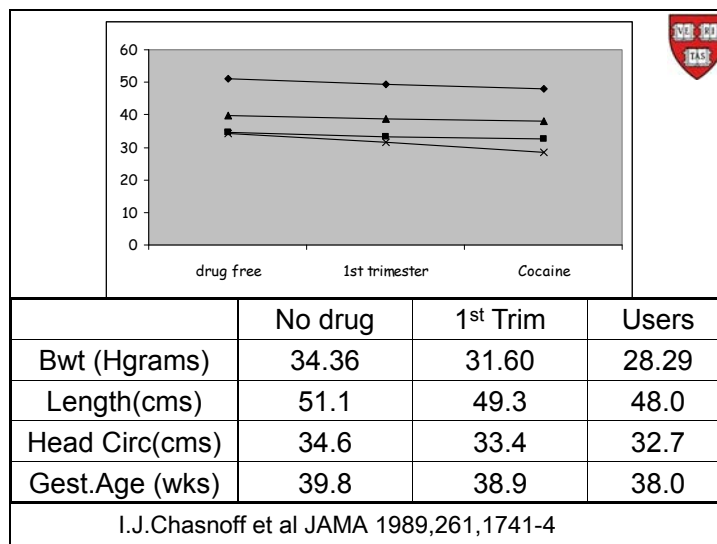
Had the Small-pox in the common way		Of these died		Received the distemper by Inoculation		Of these died	
Whites	Blacks	Whites	Black	Whites	Black	Whites	Black
5059	485	452	62	1974	139	23	7
[9.3%]				[1.4%]			

Survey following Smallpox outbreak in Boston in 1753-4, as reported by Benjamin Franklin; percentages added.

<sup>17</sup> Benjamin Franklin, Some account of the success of inoculation for the smallpox in England and America, London W.Strahan, 1759



Here are the results of the study, and they seem convincingly in favor of inoculation. And we know that the constables took care of this data. So we can trust it.



Here is another study where lying plays a role, although the rationale for lying is elusive.

The study was done on pregnant women to try and determine the impact on birth of the use of illicit drugs (such as marijuana and cocaine) during the pregnancy, and here are the results for cocaine.

The authors obtained all the right permissions to make the study ethical including to have the mothers tested for cocaine every time they came in for an antenatal visit.

Simultaneously, as they were doing the testing, they also asked the mothers about their drug use. As a result they had two answers to the drug usage question: one verbally provided by the mother (who had also given permission to be tested), and the other from the biological test.

The analysis of pregnancy outcomes was performed as a comparison of three groups: (i) those mothers who partook of no illicit drugs, (ii) those who quit illicit drug usage before the first trimester of their pregnancy, and (iii) those who used illicit drugs beyond the first trimester of their pregnancy.

They looked at these four pregnancy outcomes: (i) birth weight of the baby, (ii) length of the baby, (iii) head circumference of the baby, and (iv) the baby's gestational age. Above we see the four results, and they all show a significant ( $\alpha = 0.05$ ) trend, when the classifications in to the three groups is done on the basis of the biological tests.

Yet when they redid this analysis, but using this time basing the classification into the three groups on the basis of the mothers' verbal responses, the results lost their significance. This is rather surprising that we reach qualitatively different conclusions because of lying, even when there really was not much incentive to lie.

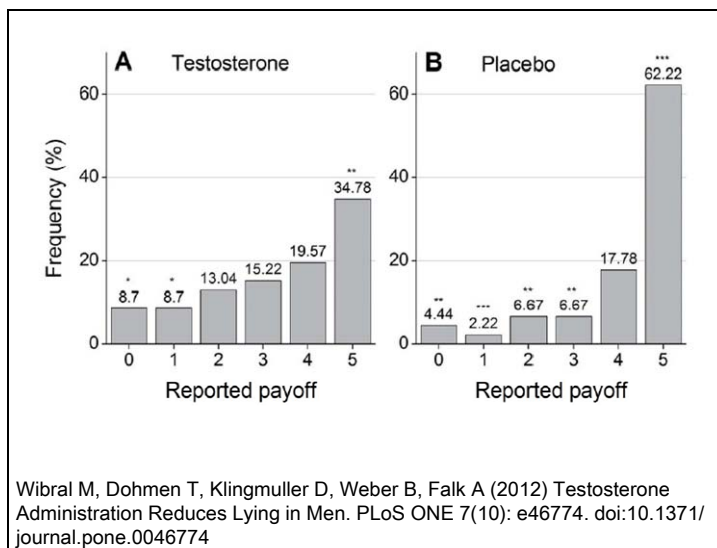
### Be a man: tell the truth



Higher levels of [testosterone](#) have been implicated in some *negative* qualities associated with men, such as impaired empathy. Recent research, though, is showing some upsides of the same hormone; now, a new study finds that [testosterone inhibits lying](#). Men were given either testosterone or a placebo and were later asked to roll a six-sided die in private, and whatever number the men reported rolling would determine their payoff for the experiment. Men with higher testosterone reported they'd scored payoffs far closer to the rate of probability. *While men who were administered the placebo reported rolling the highest payoff 62% of the time, those who were administered testosterone reported rolling the highest payoff just 35% of the time.*

Wibral, M. et al., "Testosterone Administration Reduces Lying in Men," *PLoS ONE* (October 2012).

AN article that just appeared led to the above headline, "Be a man," it says. "Tell the truth." The study<sup>18</sup> was of 91 men where roughly half the men got a shot of testosterone, and the other half were administered a placebo. They then had the men roll a die to get a payoff whose size depended on the role of the die. The catch was that no one watched the role of the die and each man reported the value of the die when claiming the payoff.



Here is the distribution of the payoffs. They were given fair die, so the expected value was that all these bars should be approximately the same size. The contention of the paper is that there

<sup>18</sup> Wibral M, Dohmen T, Klingmu N Iler D, Weber B, Falk A (2012) Testosterone Administration Reduces Lying in Men. *PLoS ONE* 7(10): e46774. doi:10.1371/journal.pone.0046774

is a significant difference between these two bar graphs, with the placebo group showing a greater tendency to lie, as measured by too large a bar on the right hand side of the graphic.

Column = die face reported

. tabi 4 4 6 7 9 16 \2 1 3 3 8 28 , row chi

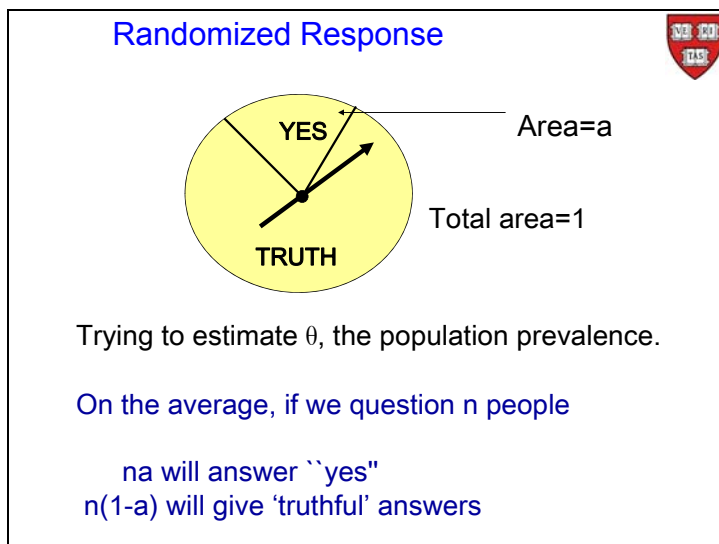
row	1	2	3	4	5	6	Total
1	4 8.70	4 8.70	6 13.04	7 15.22	9 19.57	16 34.78	46 100.00
2	2 4.44	1 2.22	3 6.67	3 6.67	8 17.78	28 62.22	45 100.00
Total	6 6.59	5 5.49	9 9.89	10 10.99	17 18.68	44 48.35	91 100.00

Pearson chi2(5) = 8.3882 Pr = 0.136

1 = Testosterone group  
2 = Placebo group

Their analysis is somewhat suspect, given that they perform all sorts of subtable analyses possibly in order to get significance? Anyway, when I fed the above table to Stata you can see that the Chi-squared analysis shows the differences between the two groups to be insignificant.

### Randomized Response



So is there anything we can do to overcome lying? Possibly not.

Ex post adjustments, for example take 80% of the ANC prevalence estimates and use those as population estimates, are unsatisfactory ad hoc solutions, especially if the truth can go in either direction. For example, in some nutritional studies, people who think they are too thin will tell you that they eat more than they actually have, and at the other end of the spectrum, people who think they are overweight will under-report how much they have eaten.

There is one design, introduced in the 1950s, that utilizes probability to mask the individual response, and thus elicit more truthful behaviour. The idea is simple, before asking a question a randomization device, hidden from the interviewer, is introduced and this device tells the person being interviewed how to respond. Since the device is hidden from the interviewer, the individual response cannot be linked to the individual being interviewed with any certainty. This thus protects the privacy of the individual, yet at the same time we can gauge aggregate behavior, which is our original intent in taking a survey, anyway!

For example, suppose you are carrying out a survey and you ask the person being interviewed to spin a pointer, which is hidden from your view, before responding to a question. The spinner can come to rest in one of two areas: if in the one, then the person being interviewed is instructed to answer, yes, whatever the question; and if the pointer falls in the other area, then the person should answer the question truthfully.

That way, if the area of the “yes” part is  $a$ , if the area of the whole is 1, and if we can consider this to be an unbiased spinner, then the probability that the person being interviewed is instructed to say “yes” by the pointer, is  $a$ . We can think of this  $a$  as the obfuscation factor; the larger we make  $a$ , the more uncertainty we introduce into the study. This would argue for a small  $a$ , but too small an  $a$  will defeat the purpose of the obfuscation, which was to convince the person being interviewed that we cannot link them to the truthful situation, unless it is a “no”. Presumably the “no” is the non-controversial label.

This should remind you of an imperfect diagnostic test. We have introduced a specificity of  $a$  by making asking that proportion of those who should answer “no”, to answer “yes”. Yet at the same time we hope that this will increase the sensitivity of this device; ideally to one. We are introducing some imprecision, hoping to gain some veracity.

## Randomized Response



So, how many will answer “Yes”? ( $m$ )

On average:

$na$  because of the dial

plus

$n(1-a)\theta$  because answering truth

In a particular sample:

$$m = na + n(1-a)\hat{\theta}$$

$$\hat{\theta} = \frac{\{m/n\} - a}{1-a}$$

The modification we need to make to our answers are the same as those we made to accommodate an imperfect diagnostic test.

## Appraisal



Advantage :

Lessens “lying” fraction

Disadvantages:

1. Increases variability of estimator (versus idealized “everyone telling truth”)
2. Costly --- difficult to use in mail survey

The advantage, we hope, is that this device lessens the lying fraction. This disadvantage is that it increases the variability but with respect to what? With respect to the idealized 1, which does not exist, which is when everybody tells you the truth.

It is a costly solution. It is difficult to use in a mail survey, for example. It requires explaining, and that may introduce errors. It has been used, but I am surprised it has not been used more often.

## Survey



Phone survey to determine illicit marijuana use --- 1986.

Estimates of prevalence:

Direct questioning estimate 40%

Randomized response estimate 64%

This design was used in a phone survey in 1986, where the person being interviewed was told to get three pennies and spin them. If the coins landed as all heads, then the person interviewed was instructed to say, yes, whatever the question. If the coins landed all tails, the answer was to be, no. The person interviewed was instructed to answer the question truthfully for any other configuration of heads and tails. This design is a little bit different than the one above;  $1/8$  of the people will say "yes,"  $1/8$  of the people will say "no," and the remainder will (hopefully) tell the truth. So we not only have a specificity of  $7/8$ , but now we also have a sensitivity of  $7/8$ , by design, assuming everyone tells the truth.

They used this design to survey the usage of marijuana (which was illegal then). They first carried out the survey without the coins and 40% of those interviewed said they used marijuana. They then redid the survey, but this time they incorporated the flipping of the coins. With this randomized response design, the estimate of marijuana use increased from 40% to 64%, a 60% increase.