PAI 705
McPeak
Lecture 7

Sampling

- The process of selecting observations is called sampling.
- Our goal is to generalize to a larger population from a sample.

Overview of the history of sampling.

- It has developed alongside political polling.
- Literary Digest, 1890 to 1938.
- Mailed postcards to readers in six states to ask who they were planning on voting for, Woodrow Wilson or Charles Evans Hughes.
- Also relied on names selected from telephone directories and automobile registration lists.
- Correctly predicted the presidential election in 1916, 1920, 1924, 1928, and 1932.
- 1936, sent out 10 million cards, got back a little over 2 million (22% response rate).
- Alf Landon predicted to win by 57% to 43% over incumbent Franklin Roosevelt.
- Roosevelt won 61% of the vote
    - Landon got 8 electoral votes compared to Roosevelt's 523.

  - Problem with the sampling frame, especially coming out of the great depression in 1936. Who was likely to have a phone and a car to be in a phone book and a vehicle registration list?
  - The poll may have predicted the voting intentions of phone owners and car owners, but they are not a representative sample of the voting population.
  - Lower income households are not well represented in this sampling frame and members of lower income households generally voted for Roosevelt.
  - Beyond that, the 22% response rate places real limits on making statements about the voting intentions of even those in the sample.

Quota Sampling (to be returned to later, but for now fits in the history of polling and sampling)

- Gallup in 1936 developed his poll with the idea of quota sampling.
- You need to have information about the population, and base your sample on matching sample size population proportion.
- Gallup used income stratification for his polls with the American Institute of Public Opinion. Correctly picked winner in 1936, 1940, and 1944.
- He got it wrong in 1948, predicting Dewey would beat Truman. Truman won 50% to 45% [the States Rights (Thurmond) and Progressives (Wallace) each got over 2%]

- What happened?
- Gallup used 1940 census data as a framework to identify population proportions. Why is 1940 profile not capturing essential characteristics of 1948's population profile?

Nonprobability Sampling.

- Convenience sampling. Haphazard sampling. "person on the street" interviews.
    - It gives you a representative sample of people passing by where you are, and that is assuming you have an equal probability of getting people to stop and answer the question.
    - Does our sample of reactions to the commercials of people in this classroom give us much ability to extrapolate to the larger US population?
        - In what ways are you systematically NOT representative of the US population.
- Purposive / judgmental sampling.
    - Pretest survey on the widest variety of people as possible to test the ability of your questions to handle a wide variety of responses / not highly likely but not impossible situations.
    - I want to compare left leaning students and right leaning students. I might interview members of left leaning student groups and right leaning student groups.

    - A subset of the population of interest, but maybe good enough for the purposes you have in mind.
  - I want to get a sense of women's perspective on the proposed dry season garden.
    - I can have a focus group meeting with the members of the local women's group.
  - Guided by convenience.
  - Also might want to select intentionally on deviance from the norm to better understand the norm.
  - Might find out things along the way that lead you to add people you need to interview; the student groups talk to these faculty, so you want to add the faculty to understand more fully the student groups.
    - "Theoretical sampling" because your evolving theoretical understanding drives who you need to speak to in order to understand the subject.
- Snowball sampling.
  - When you are finding your way, and where coming up with a list of the population or the location of the population is hard to get.
    - Homeless
    - Nomadic herders
    - Undocumented farmworkers
    - People involved in illegal activities
  - Good for exploratory, not great for representativeness.

  - Good for tracing networks. Good for uncovering “the dynamics of natural and organic social networks”.

- Quota sampling.
  - You need to know the characteristics of the population for this kind of sampling.
  - What are the relevant characteristics you need to use to stratify your population?
    - Gender
    - Age
    - Income
    - Social status
    - Ethnicity
    - Clan
    - Livelihood category from typology
    - Educational attainment
  - Develop a matrix of these with the share of the population at each intersection in the matrix. These proportions define the quota frame.
    - Interview people having the characteristics of each cell in the matrix.
      - Note that the selection of people in each cell may not be representative of people in that cell unless you set it up to be random at that stage.

        - Systematically picking people to fill the quota, but doing it in a way that might not make them representative of the whole population in question; those in the mall, those having land lines, those waiting for a Centro bus.

Selecting informants.

- At the most basic level, somebody from the group who can speak about it with some informed knowledge.
    - Often are in a bit of a bind, as the person you can work with shares a common language with you, thus making them to some degree not an insider in the way you are trying to understand.
    - Their willingness to work with you may make them not typical; the female who will sit and answer your questions.
    - Their caste or clan status might make them have a particular view of things.
- Sampling list. The list of elements from which a sample is drawn. It is (hopefully) the list of the population that is relevant to the study.
- What kinds of lists are out there to be used?
- University directories, school rosters, voter registration, vehicle registration, tax rosters, humanitarian aid rosters, white pages, licensed professionals….

- Think about white pages (what are those?) and phone sampling. As cell phone technology advances, those with land lines listed in the white pages become less and less representative of the overall population.
- Cell phones and the option not to be listed make this more difficult to use as a sampling frame.
- Random digit dialing (RDD).
- But Telephone Consumer Protection Act of 1991 puts limits on the use of phones, aimed at telemarketing and robocalls, but this also impacts survey researchers.
- Address Based Sampling (ABS) based on US postal service lists of postal addresses.
- Spread of cell phones in developing countries provides new opportunities for phone sampling that did not exist before.
- Still issues of non-representativeness.
- Do respondents get charged for the minutes.
- Bias in terms of who is in range of the network.
- Also issues of online surveys.
- Self-selection in.
- Non response.
- Multiple response.

PROBABILITY SAMPLING

- Figure 7-3. Researcher is standing in bottom right corner. Talks to the 10 people closest to her. 10% of the population. But a biased sample.

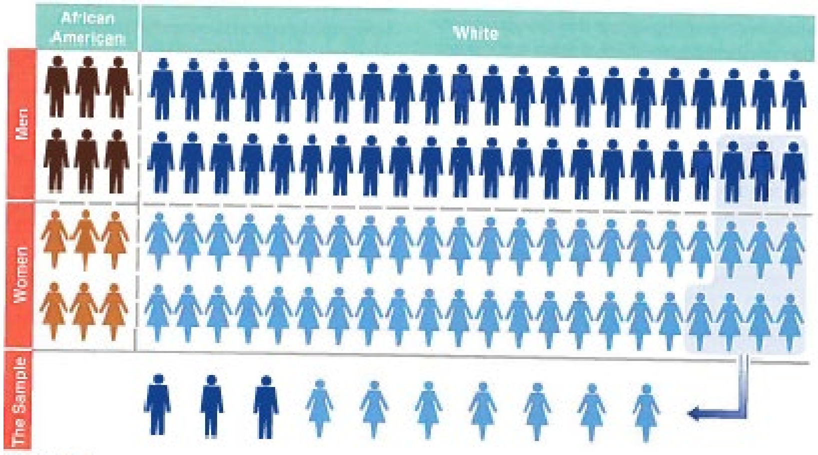


FIGURE 7-3

A Sample of Convenience: Easy, but Not Representative. Simply selecting and observing those people who are most readily at hand is the simplest method, perhaps, but it's unlikely to provide a sample that accurately reflects the total population.

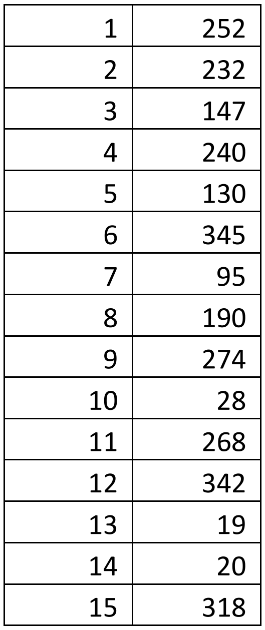
© Cengage Learning®

- Representativeness. The aggregate characteristics of the sample closely approximate the aggregate characteristics of the population.
    - Not necessarily in all characteristics, but in all relevant characteristics where relevance is context dependent.
    - All sampled units should be drawn in such a way that each member of the population has an equal chance of being selected into the sample.
        - EPSEM. Equal probability of selection method.
            - It needed a vowel so SE from selection.
    - An element is the unit about which information is being collected. The unit of analysis, with the nuance that the element describes the unit in reference to

sample selection and the latter that same unit in reference to data analysis.

- A population is the theoretical aggregation of our elements to a larger group that were interested in generalizing about.
    - This requires some specification about what we mean specifically.
    - All households resident in the village when the survey is to be conducted.
- The study population is the list that we use that is the practical representation of the population.
    - The list of all households in the village that is in the office of the mayor.

- Random selection.
    - Simple Random Sample.
        - Excel:



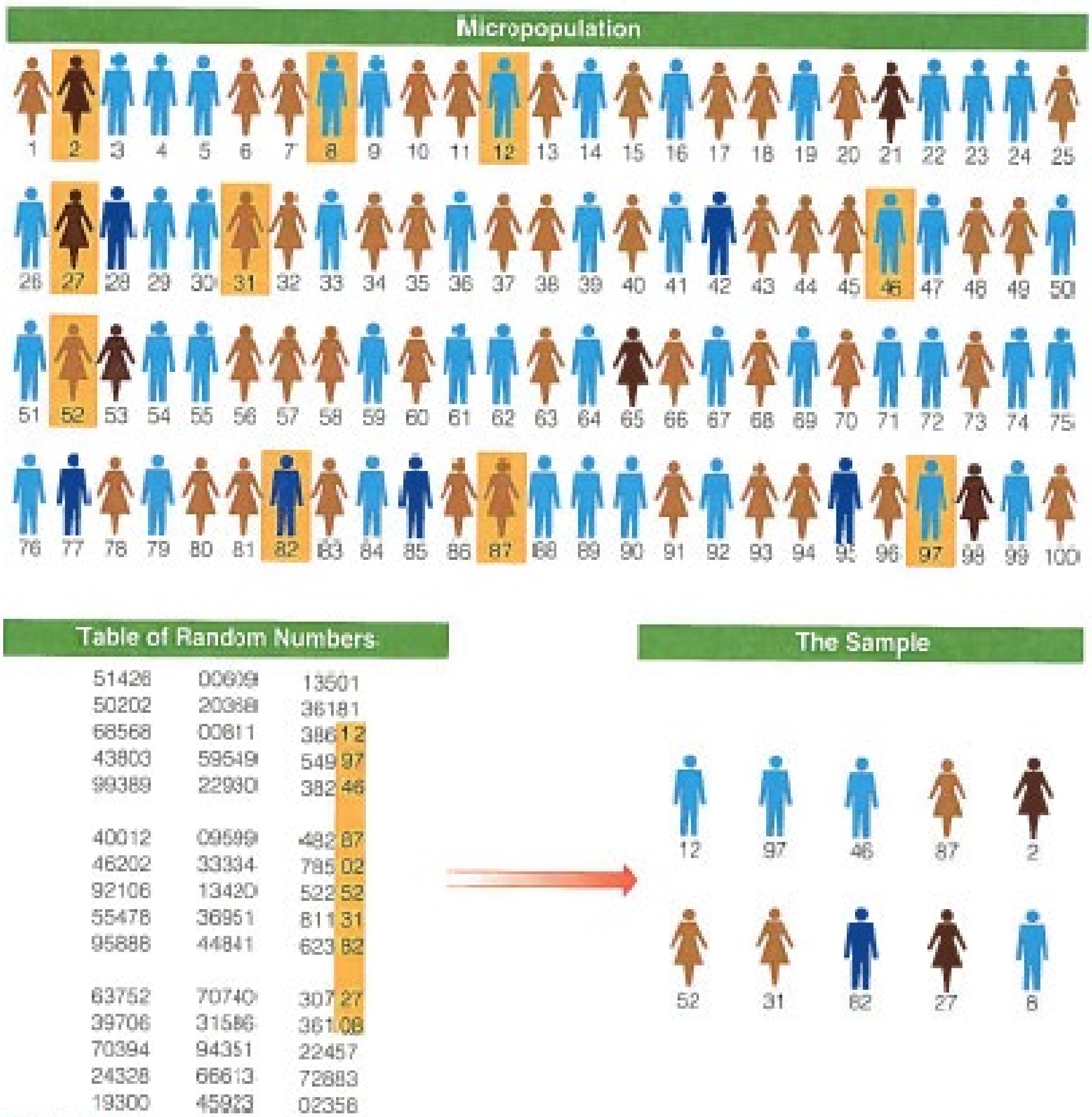
1	252
2	232
3	147
4	240
5	130
6	345
7	95
8	190
9	274
10	28
11	268
12	342
13	19
14	20
15	318

-
    - Random number table (p 517)
    - Pseudo random number generators.
        - =INT*(N*(RAND())) in excel.
    - Systematic Sampling.
    - Flip coin 10 times, the number of heads is the first number from which to start sampling.
    - Sampling interval, the N/target sample size rounded to an integer.
    - Sampling ratio = sample size / population size.

- 358 households in the village, I want 15 in my sample for whatever reason.
    - 358 divided by 15 rounds to 24. Flip coin ten times, get 4 heads, start with household 4 on the list.
        - 1-4,
        - 2-28 (4+24),
        - 3-52, (4+48)
        - 4-76,
        - 5-100,
        - 6- 124,
        - 7- 148,
        - 8-172,
        - 9-196,
        - 10-220,
        - 11-244,
        - 12 – 268,
        - 13- 292,
        - 14- 316,
        - 15-340.

No odd numbers.

Figure 7-11.



51426	00609	13501
50202	20368	36181
68568	00811	38612
43803	59549	54997
99389	22930	38246
40012	09599	48287
46202	33334	78502
92106	13420	52252
55478	36951	61131
95888	44841	62382
63752	70740	30727
39706	31586	36108
70394	94351	22457
24328	66613	72883
19300	45923	02356

FIGURE 7-11

A Simple Random Sample. Having numbered everyone in the population, we can use a table of random numbers to select a representative sample from the overall population. Anyone whose number is chosen from the table is in the sample.

© Cengage Learning®

- Potential problem; is the order of elements on the list arranged in some kind of pattern.
    - Book example; study of soldiers, arranged according to rosters. Took every tenth name on the roster. The rosters were organized by sergeants first, then corporals and privates. Each squad had ten members. Every tenth person was the squad sergeant.
    - My research; household names in one community were organized by clan. Clan rivalry ran high. "His computer is biased against the Sale clan". Stratify by size of the clan in the overall population to weight what share of the 30 households came from each clan.
- Stratified sample.
    - Ensuring that appropriate numbers of elements are drawn from distinct sub-groups in a heterogeneous population.
    - The sub group is homogeneous with regard to the selected strata.
    - You can select the share of the total sample based on the share of the population that that sub-group represents.
    - For example, figure 7-12. Line them up and sample from the sample frame.

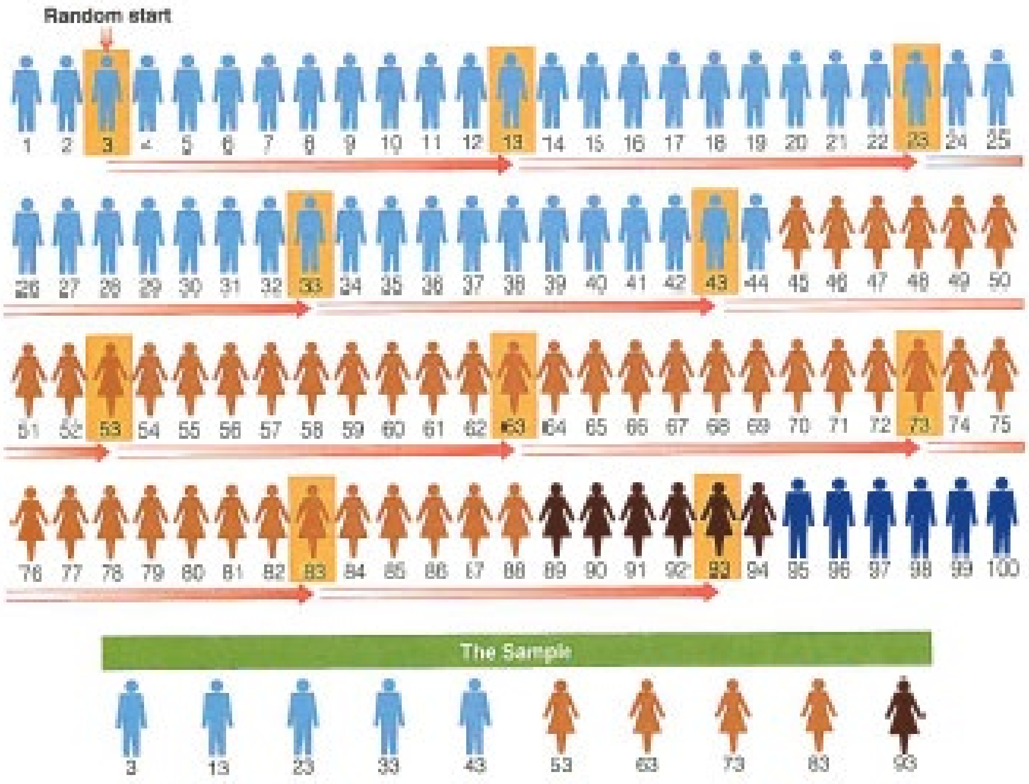


FIGURE 7-12

A Stratified, Systematic Sample with a Random Start. A stratified, systematic sample involves two stages. First the members of the population are gathered into homogeneous strata; this simple example merely uses gender as a stratification variable, but more could be used. Then every *k*th (in this case, every 10th) person in the stratified arrangement is selected into the sample.

© Cengage Learning®

        - Note dark brown females and dark blue males make up 6% of the population; for the females that translates into 10% of the sample for the males that translates into 0% of the sample.

- You want to pay attention to any order that the list may have:
    - Example in the book of University of Hawaii students ordered by class (1st year, 2nd year, 3rd year, 4th year) but within class by student ID.
        - The subtle issue was that the student ID numbers started with a code reflecting the

students Social Security Number, and the codes are distinct representing the geographical area you got your card issued.

Cluster sampling.

- Initial sampling of groups of elements, then selecting elements of groups within the cluster.
- It is impractical, not possible, not feasible to get a list of all the possible elements.

A two stage approach: first, pick the sample frame of a subset of the population that represents a category of the population.

Survey sites

The sites were chosen to capture variation in agricultural potential, market access, livestock mobility and ethnic diversity. The following table summarizes the basic characteristics of the survey sites located in southern Ethiopia and northern Kenya.

Code	Name	Country	Market Access	Ethnic Majority	Agricultural Potential	Annual Rainfall
DG	Dirib Gumbo	Kenya	Medium	Boran	High	650
KA	Kargi	Kenya	Low	Rendille	Low	200
LL	Logologo	Kenya	Medium	Ariaal	Medium-Low	250
NG	Ng'ambo	Kenya	High	Il Chamus	High	650
NH	North Horr	Kenya	Low	Gabra	Low	150
SM	Sugata Marmar	Kenya	High	Samburu	Medium	500
DH	Dida Hara	Ethiopia	Medium	Boran	Medium	500
DI	Dillo	Ethiopia	Low	Boran	Low	400
FI	Finchawa	Ethiopia	High	Guji	High	650
QO	Qorate	Ethiopia	Low	Boran	Low	450
WA	Wachille	Ethiopia	Medium	Boran	Medium	550

More complex, capturing elements of treatment and control that we will be moving towards. Index Based Livestock Insurance in Ethiopia.

Encouragement Design

Researchers came up with 3 ways to encourage households to purchase IBLI. These 3 approaches were designed to stimulate interest in purchasing insurance for their animals.

1. **Discount Coupons:**

These coupons ranged from 10% to 80% and were issued to 80% of sampled households for the purchase of up to 15 TLU (Tropical Livestock Units). The remaining 20% got no discount coupon but for those households participating in the Herd Migration Study, 50% of them received no discount coupon and the other 50% received 100% worth discount coupons. This was randomized across all study sites.

2. **Poet Skit Tape:**

A skit was prepared by the underwriter (Oromia Insurance Company) during the launch of IBLI and the skit taped. Development Agents were asked to play the skit to some of the sampled households.

3. **Cartoon:**

OIC came up with a caricature representation of the product to be used as an extension tool. The contents are first read out to the sampled households and they are latter allowed to read/look through again as many times as they want.

We came up with some criteria that ensured community level observable variance was kept at a minimum thus collapsing the study sites into 3 clusters based on:

1. **Market Participation**: This contained sites located closer to major livestock markets in Dubluq and Haro Bake. These markets would offer pastoralists an opportunity for timely offtake of livestock in the event of a drought, therefore sites within a 10km[1] radius from these markets were lumped into one cluster. This grouped Iddi Halle, Harboro, Dambala Saden and Dambala Dhibayu into one cluster.
2. **Rainfall distribution**: Some sites suffered severe lack of rainfall and were thus most affected by drought, suppressing livestock population growth. Consequently, we anticipate that sites under this criterion would have similar demand for IBLI. In this cluster we had Qancharo, Magole, Baha, Magole and Gofa (Low-Medium TLU).
3. **High TLU**: These households are remotely located and far from functioning livestock markets. They depend almost entirely on their livestock for subsistence and are composed of Ella Dima, Dibe Gaya, Web, Wachile, Saba, Gorile, Sarite and Hirimaye.

Two treatments (Encouragement tools) were assigned to each of the 3 clusters at random and at least one site acted as a control i.e. no treatment. The randomized treatment effect yielded the matrix:

Cluster	Study Site	Cartoon	Poet Tape	No Treatment
Market participation	Iddi Halle	✓		
	Dambala Saden		✓	
	Harboro			✓
	Dambala Dhibayu	✓		
Sparse rainfall	Magole		✓	
	Qancharo			✓
	Magado	✓		
	Gofa		✓	
	Baha			✓
Large TLU	Web	✓		
	Wachile		✓	
	Dibe Gaya			✓
	Sarite	✓		
	Saba		✓	
	Gorile			✓
	Hirmaye	✓		
	Ella Dima		✓	

[1] We assume a 10km walk would not hinder pastoralists from market participation

- Since this is about livestock insurance, an oversampling of the places with lots of livestock since that is the population of interest.
- In contrast, for the land certification project in the same area, we had the following approach:

At the time of the impact evaluation design, some communities were selected as places where certification would go forward and others where it was not scheduled to take place in the initial phase of the project. For our purposes, we do not draw a distinction between treatment (i.e., sites designated for future certification) and control sites, as there is no treatment yet in place. It is worth noting that the desire to have paired treatment and control sites led the survey of the Borana sites to tilt towards more agro-pastoral zones than would be the case if the sample was purely random selection from the Borana Zone. This meant drier, more pastoral areas are underrepresented in the survey sample, although interviews with key informants and focus group discussions were held in some of these more arid locations. For example, the mean herd size revealed in our Borana sample is 43% of the mean herd size of the Index Based Livestock Insurance (IBLI) monitoring study mean herd size collected in the same year in communities in the same zone[2].

- The price of the efficiency and feasibility of the cluster sample approach is that it is a less accurate sample of the larger population.
    - One source of error, initial sample of clusters represents the range of clusters with some sampling error.
        - How well do these four sites reflect all the sites that are associated with this concept of relatively easy market participation?

Market participation	Iddi Halle, Dambala Saden, Harboro, Dambala Dhibayu

[2] The IBLI sample is more heavily weighted towards areas more oriented to livestock production as the aim of the project is to extend insurance to livestock keepers (Ikegami and Sheahan, 2010). We assume the overall Borena zonal mean household herd size in 2015 lies somewhere between the 8.3 TLU per household found in the LAND dataset and the 19.4 TLU per household found in the IBLI dataset.

    - Second source of error, the elements selected from the sampling frame chosen to represent the larger population of the selected cluster with sampling error.
        - How well do the households selected in Iddi Halle represent the population of Iddi Halle?
- The clusters (study villages) chosen to represent the category that define the cluster (all villages with higher market access) will best represent the larger category's population if a large number of clusters are selected and if all clusters within the category are very much alike.
- A sample of elements will best represent all elements in a given cluster (village) if a large number of elements (households) are selected from a cluster and the elements are alike within the clusters.
- Here is the tension; often you have a total sample size that is supportable by your research budget.
- I can increase my number of villages studied per cluster, but that decreases the number of households per village. I can increase my number of households per village, but that decreases the number of villages I can cover.
- Total sample size = number of categories for the clusters*number of villages per cluster*number of households per village.

- If the households within the village are generally homogenous, and the variation in the population is mostly generated by cross village heterogeneity, you would want to increase the number of villages while decreasing the sample size per village.
- One, you don’t really know whether the heterogeneity is more pronounced within or between going in (that is why you have to do the research!).
- Two, increasing the number of villages is generally costlier than increasing the number of households within a village. Travel time, logistics of getting there, gathering people together….

Probability Proportionate to Size sampling.

- Remember we want each item in the population to have an equal chance of being selected to ensure the sample is representative.
- Two stages, chance of being selected as the village representative of that cluster, then chance of being selected as the household within that village.

Violating this rule:

We have committed to a sample size of 600.

- 400 from Mali
- 200 from Senegal.

We have identified three Cercles in Mali (Motpi, Koro, Douentza) in the Region of Mopti and four Departements in Senegal (Koungheul, Birkelane, Maleme Hodar, and Kaffrine) in the Region of Kaffrine.

- 400 from Mali by three cercles gives ~135 households per cercle.
- 200 from Senegal from four departments gives 50 households per departement.

To get some representative variation of the different administrative units, we will sample purposively when selecting the villages where the survey will be conducted. Proposed strata:

- Agroecological to contain some areas that are riverine and some that are not.
- Production system to include some communities that are primarily cultivators and some that are predominantly livestock producers.
- Market access / tarmac access with some having a larger weekly market within a 10 kilometer radius and some not.

Three dimensions gives eight possible combinations

Riverine, cultivator, high market access	Not riverine, cultivator, high market access
Riverine, cultivator, low market access	Not riverine, cultivator, low market access
Riverine, livestock, high market access	Not riverine, livestock, high market access
Riverine, livestock, low market access	Not riverine, livestock, low market access

In Senegal, 8 different communities / villages could be selected with 25 households in them to cover this variation and arrive at a sample size of 200. It would seem that 2 villages per departement would be one way to do this.

In Mali, 16 different communities / villages could be selected with 25 households in them to cover this variation with two sites per cell in the table above.

In each village where the household survey is being conducted, obtain the list of residents from the chief of the village. Count the total number of households listed as resident in the village. Divide this number by 25. Round to the nearest integer. This tells you what number you add to the household number you have selected to interview to identify the next household to interview.

- This violates the equal probability of selection if the number of households differs across villages.
  - In a community with 100 households the chance of being selected is 25% (25/100).

    - In a community of 500 the chance of being selected is 5% (25/500).
- If we know the populations per village afterwards, you have an ability to clean up afterwards by assigning weights.
    - Each element that was selected can have their observations given a weight equal to the inverse of their likelihood of selection.
    - If each element has the same probability of being selected we don’t need to do this, and we have a self-weighting sample.
- We also may intentionally sample in a disproportionate way to ensure we get adequate observations on populations of particular interest.
    - Female headed households
    - Poor households.
    - Households from a particular caste

Probability Theory, Sampling Distributions, and Estimates of Sampling Error.

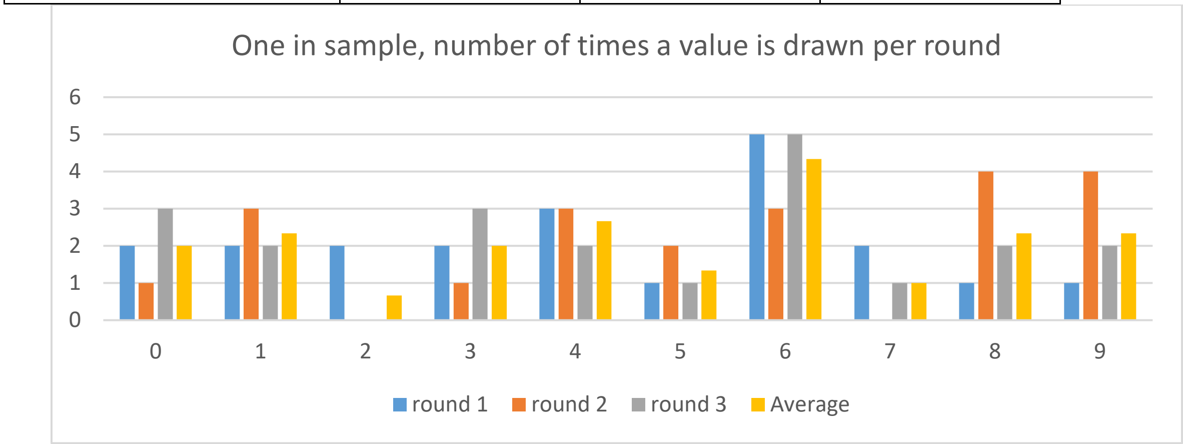
- A parameter is the summary description of a given variable in the population.
- To illustrate the 'law of large numbers' let us look at a population of 10 with the following distribution of income.

John	Renee	Josh	Mike	Bob	Stuart	Sabina	Mary	Jeb	Catherine
$0	$1	$2	$3	$4	$5	$6	$7	$8	$9

- Sum is $45 and there are 10 people so $4.50 is the computed average – the true value we are estimating by sampling.

Take a sample size of 1, 20 draws, 3 rounds with =INT(10*RAND()) per draw per round and replacement after drawing.

Draw number	round 1	round 2	round 3
0	1	4	1
1	6	9	7
2	0	6	8
3	3	1	0
4	4	6	6
5	8	4	1
6	7	1	6
7	6	9	8
8	5	9	9
9	3	1	3
10	6	4	5
11	7	6	3
12	0	9	0
13	2	8	9
14	4	5	0
15	1	8	6
16	4	3	4
17	2	8	6
18	9	5	6
19	6	0	3
20	6	8	4



This should tend to be equal for each value as more rounds are conducted (in this case, tend towards 3 since 3 rounds).

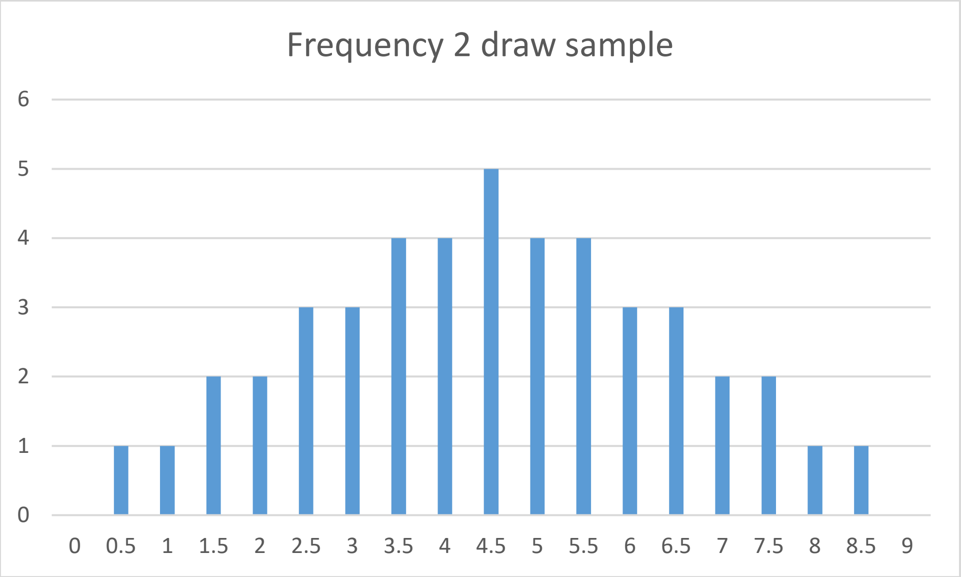
Then allow the sample size to be 2 draws at a time, and after each draw we calculate the average of the 2 values we have selected. We then replace and draw again.

If we do that, the table below shows there are 45 possible combinations that lead to different average values. [n!/{(n-r)!*r!}]= [10!/{(10-2)!*2!}]=45

	0	1	2	3	4	5	6	7	8	9
0										
1	0.5									
2	1	1.5								
3	1.5	2	2.5							
4	2	2.5	3	3.5						
5	2.5	3	3.5	4	4.5					
6	3	3.5	4	4.5	5	5.5				
7	3.5	4	4.5	5	5.5	6	6.5			
8	4	4.5	5	5.5	6	6.5	7	7.5		
9	4.5	5	5.5	6	6.5	7	7.5	8	8.5	

Carry on to now look at figure 7-7.

With samples of 2, how many ways are there to get each of the averages labeled in the following frequency graph?



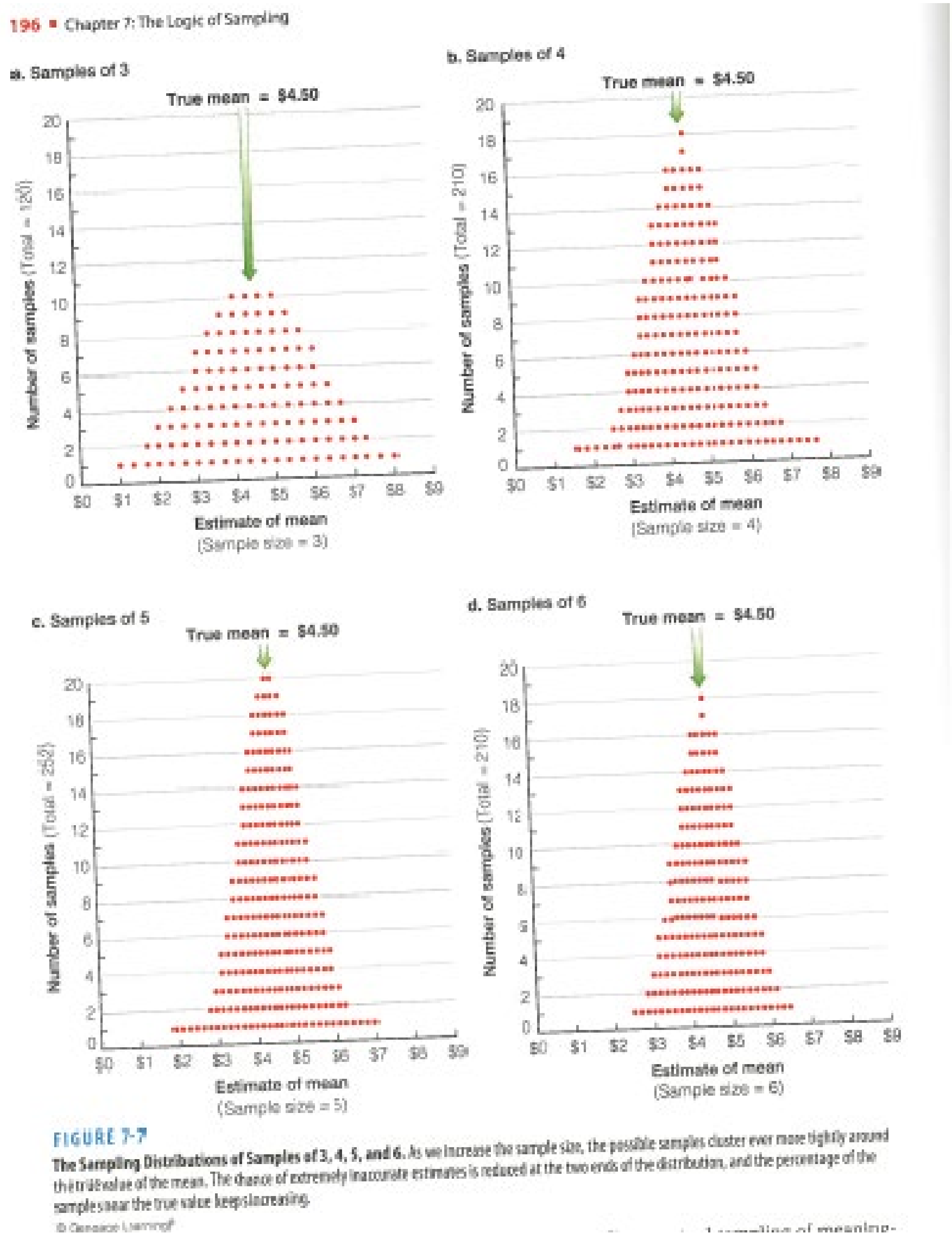


FIGURE 7-7

The Sampling Distributions of Samples of 3, 4, 5, and 6. As we increase the sample size, the possible samples cluster ever more tightly around the true value of the mean. The chance of extremely inaccurate estimates is reduced at the two ends of the distribution, and the percentage of the samples near the true value keeps increasing.

© Cengage Learning®

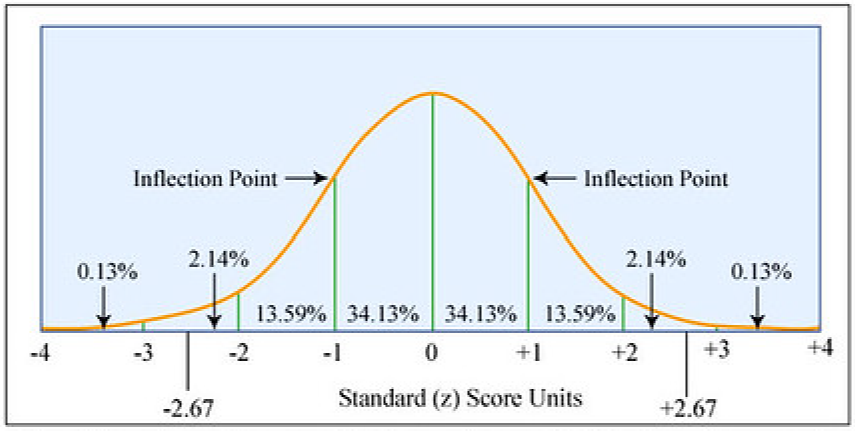
1-10, 2-45, 3-120, 4-210, 5-252, 6-210,….

- We get more and more observations piled up on the true value, and less dispersion about the true value.

- 4.5 emerges as that statistic, the summary description of the income variable in question, used to estimate the population parameter of average income.
- Normal curve emerges; the bell curve.
- To be specific, a random variable x is distributed normally with parameters μ (mean) and σ^2 (variance) if the density of x is given by:

$$f(x) = \frac{1}{\sqrt{2*\pi}*\sigma} e^{\frac{-(x-\mu)^2}{2*\sigma^2}}, -\infty<x<\infty$$

- So to say something is normally distributed has a specific mathematical meaning.
    - Standard normal



Proportions of scores in specific areas under the normal curve. The inflection points are one standard deviation from the mean.

We can calculate the sampling error / sampling deviation. Given the two outcomes P and Q and n outcomes, the sampling error is:

$$s = \sqrt{\frac{P * Q}{n}}$$

- In the example in the book, P = .5, Q = .5, n = 100. Let P be the share in support, Q be the share opposed. Our interest is in P.
- s= sqrt((.5*.5)/100) = .05. 5 %. We can call 5% one standard deviation.
- In a normal distribution, around 34% of estimates will fall within one standard deviation below the parameter P, another 34% will fall within one standard deviation above.
- If the true parameter is P=.5, there is a 68% chance our sample will give a value between .45 and .55.
- Given that about 95% of samples fall within two standard deviations above and below the true parameter, there is a 95% chance our sample will give us a value between .40 and .60.
- Given that about 99.9 % of samples will give a value above or below three standard deviations from the true parameter of .5, 99.9% of samples will give us values between .35 and .65 as an estimate of P (which is .5)

- Returning to the formula for s, we can see that s is an increasing function of the product of P and Q. They reach their maximum (.25) at even splits of .5, .5.

.5*.5	.6*.4	.7*.3	.8*.2	.9*.1	.99*.01
.25	.24	.21	.16	.09	.0099

  - So if the true values for P and Q are P=90% and Q=10% we have s = sqrt((.9*.1)/100)=.03. So we have a 68% chance our sample of 100 will give values between .87 and .93, a 95% chance our sample will give values between .84 and .96,…

- Going back to the P=.50 and Q=.50 example, we can also see s is a decreasing function of n.
  - If we go to n= 400, we have sqrt(.25/400)=2.5%.
  - If we go to n=2500 we get to sqrt(.25/2500)=1%.
- Now all of this proceeded with the idea that we knew P and Q.
- However, that is not the case in social science.
- We would not need to do the research if we knew the parameter value already!
- This is to think through the logic about how the statistic we can calculate from our sample is related to the (unknown) true population parameter.

- Confidence level – the probability that the value of the population parameter lies within a specific interval.
- Confidence interval. The upper and lower boundaries and the range between them that are associated with the confidence level.
- When we are determining sample size and confidence intervals, it is worth keeping in mind that the size of the population is not generally of direct concern.
- The background assumption is that population is so large that the size of our sample is 'small' in comparison.
- If a sample is large enough (say 5% or more of the population) we can start to adjust the confidence intervals by scaling them down to be a smaller range at a given probability level.
- The idea is that the larger our sample is as a share of the total population, we are moving towards what would happen if we sample the whole population so the average is a math calculation with no standard deviation.
- Finite population correction =$\sqrt{\frac{N-n}{N-1}}$
- If my population is 12,000 and my sample size is 1,200, this is $\sqrt{\frac{12{,}000-1{,}200}{12{,}000-1}}=\sqrt{\frac{10{,}800}{11{,}999}}=\sqrt{0.9}=0.95.$
- Returning to the example of when the true parameter is P=.5 and there was a 68% chance our sample will give a value between .45 and .55.

- Now take the .95 correction and multiply it by .05. This gives you 0.0475.
- With the finite population correction calculated above for 1,200 I could now say there was a 68% chance the value is between [(.5-.0475)=.4525 and (.5+.0475= .5475)].
- If I had 6000 out of 12,000 I would narrow it down to 68% chance the value is between [.465 and .535]
- Take it to the extreme, as n tends towards N this correction tends toward zero, so as you sample the whole population your confidence interval compresses towards no sampling error – since you are getting close to not sampling but enumerating the whole population!

Technical explanation of sampling and variance with regard to Minimum Detectible Effect.

When we have existing data sets that we think might provide responses in the range of what we will find in Senegal and Mali, we conduct Minimum Detectable Effect size calculations.

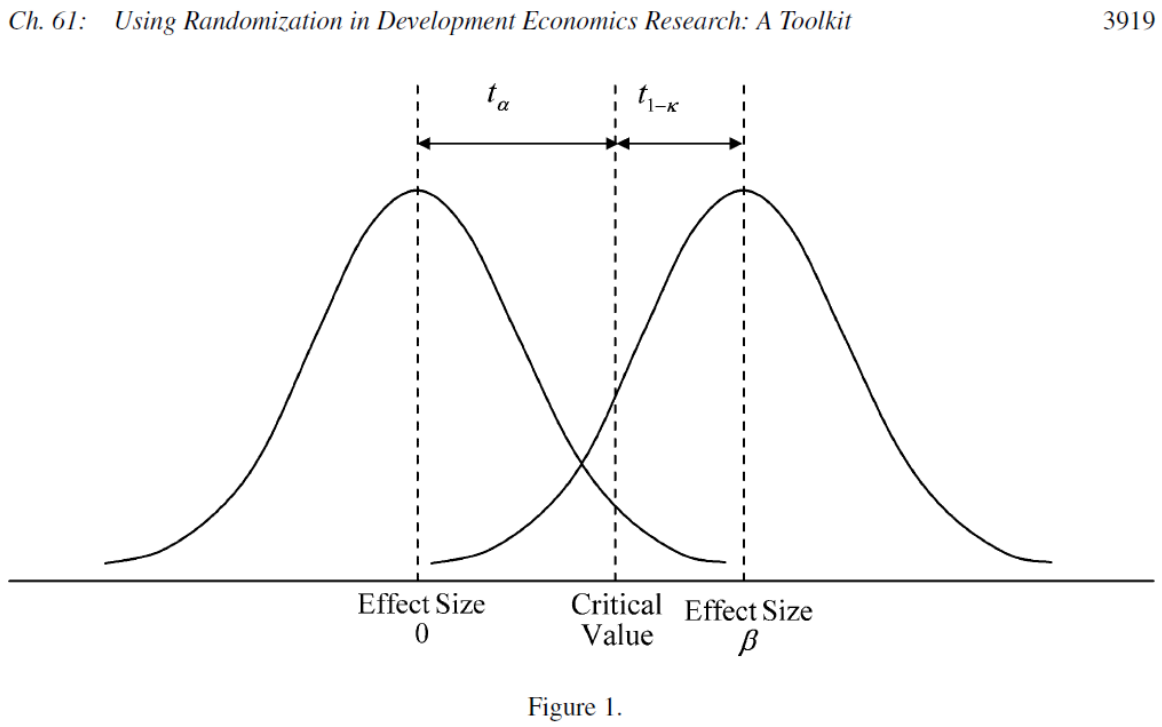
MDE is based on the idea that we are interested in assessing the impact of a given intervention.

The null hypothesis is that the intervention has no impact (called H_0).

The significance level of a test is the chance that we reject the hypothesis of no effect (by saying there is an effect) when the null hypothesis is in fact true (a type I error where we think there is an impact where there really is not.) – we get it wrong.

The power of a test is the probability we reject H_0 (the null of no impact) when it is false (and there really an impact) – we get it right.

Following the example in Duflo et al. (2008) we choose a significance level of 10% and a power of 80%. MDE is also a function of the percent of the sample anticipated to be beneficiaries (P) and those not anticipated to be beneficiaries (1-P) of the intervention.



Ch. 61: Using Randomization in Development Economics Research: A Toolkit 3919

Figure 1.

$$MDE = (t_{(1-\kappa)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \tag{7}$$

Indicator 1 Assets.

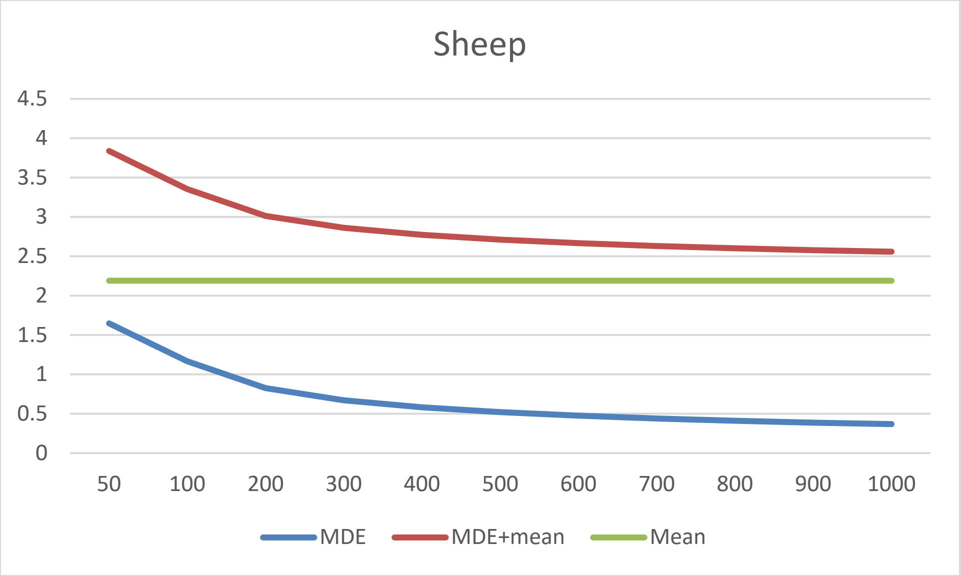
In this context we consider livestock as a key asset. In a survey conducted in 2012-2013 in Matam, Bakel, and Kidira areas of Senegal, the following question was asked (share of the population in each category in each cell):

Herd Size.

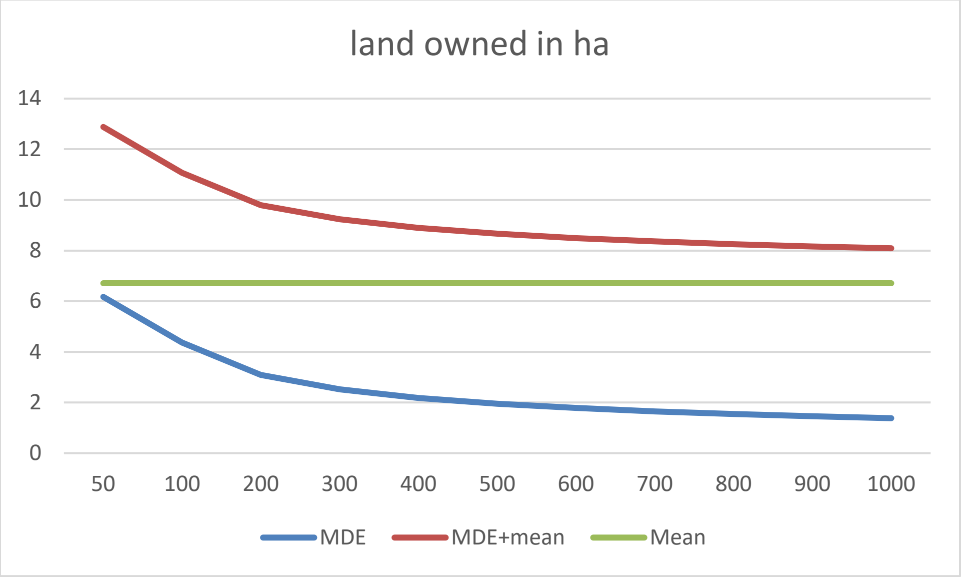
Category	Range	Cattle	Sheep	Goats	Camels	Donkeys	Traction animal
1	0	38%	32%	39%	80%	50%	24%
2	1-10	26%	30%	27%	18%	36%	76%
3	10-25	11%	17%	17%	1%	4%	0%
4	25-50	11%	13%	11%	1%	4%	0%
5	50-100	8%	6%	5%	0%	3%	0%
6	>100	5%	2%	1%	0%	1%	0%
	Mean	2.10	2.19	1.88	0.78	1.61	1.64
	Variance	3.27	2.44	2.59	0.92	3.69	0.46

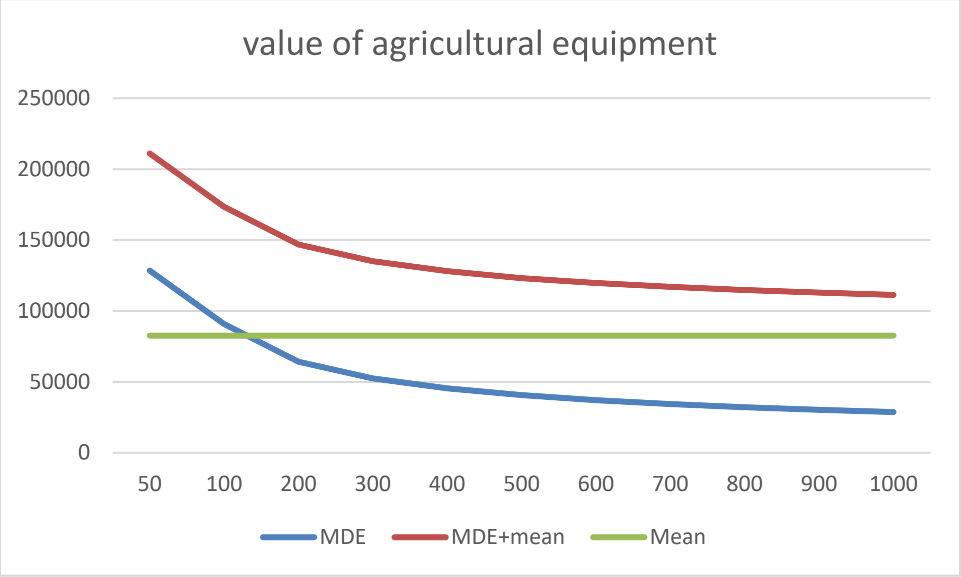
Take for example sheep. Sheep may be a key species to consider as sheep are more likely, though not necessarily, owned by women compared to cattle. Sheep are also likely to be owned by the poorer population, note in the table above the second lowest % of zeros is for sheep (only animal traction animal is lower).

In figure 1, the y axis is the mean of the category numbers from table 1. Thus the mean of 2.2 means that the average household is above the 1-10 range and in the 10-25 range, but probably closer to the 10 than the 25 head. The way to read the red line is to say how much the mean category score would have to increase before we felt confident that there was a statistically significant increase in this variable in the sense developed above. The x axis is the sample size that is drawn from the population. Focusing on the 600 sample size proposed above, this implies that the mean of the category of 2.19 would need to increase to 2.67 or more before we felt confident in there being an effect. We also see that increasing the sample over the range of 600 to 1000 does not give us all that much benefit in terms of our ability to be confident that we could identify an impact in this domain.



If we turn to land and the value of agricultural material owned by the household as a measure of assets that apply more to the cultivation domain, we have the following results.



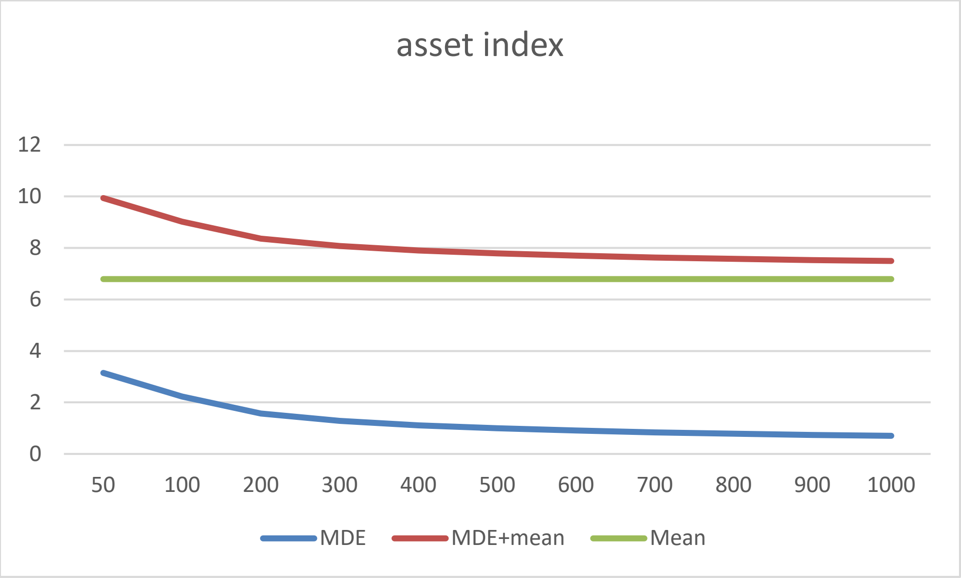


These we can see are going to be much harder to use to detect a difference (assuming this area has means and variance like that seen in the sample used to generate the mean and variance of course). At a sample size of 600, land ownership would have to increase by more than 1.8 ha and the value of agricultural equipment by more than 37,000 CFA to be detected. A strategy in this case may be to focus on a key type of agricultural equipment as an indicator, as we did above on sheep.

Finally, we might look at household belongings as a form of assets. In the survey used above, people were asked about the following home possessions.

Goods and equipment	OWNS?	
Cell Phone	Yes	No
Bike	Yes	No
Radio	Yes	No
Télévision	Yes	No
Flashlight	Yes	No
Electric Lamps	Yes	No
Animal drawn cart	Yes	No
Traction animal	Yes	No
Shop	Yes	No
House or land in a bigger town:	Yes	No
Vehicle	Yes	No
Battery for light in the house	Yes	No
Solar Panel	Yes	No
Motorcycle	Yes	No

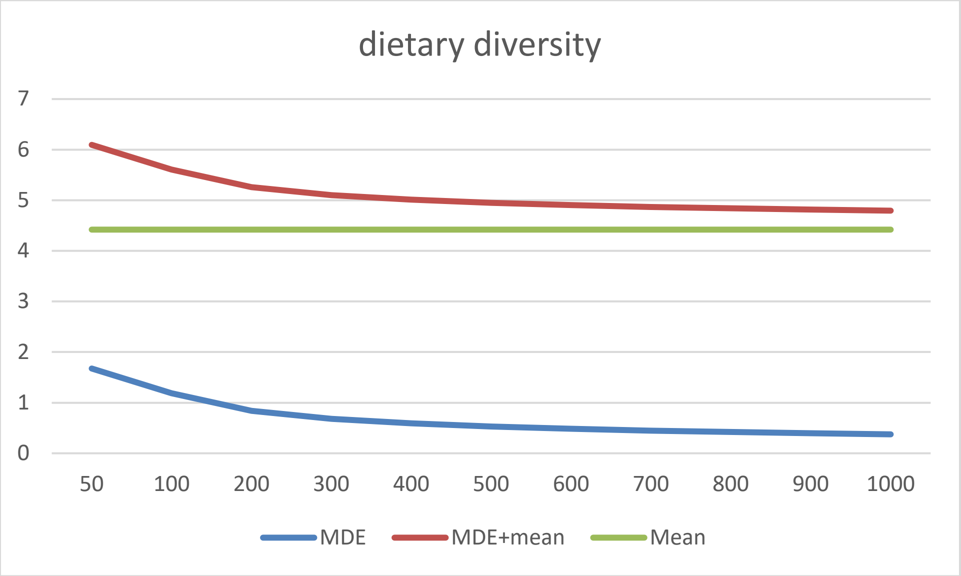
As a way of constructing a household level index, we can assign a score of 1 for each yes and a 0 for each no. When this is done, the mean score is 6.8 and the variance is 8.9. This leads us to the following graph:



At our proposed sample size of 600, we would be able to pick up a change in the mean of this measure by anything larger than .91. The index as currently constructed is not weighted by value, so you get the same credit for a flashlight and a refrigerator so there is some crudeness involved, but still this would suggest there is some prospect for detecting impact in something like this measure.

Dietary diversity.

Dietary diversity has been proposed as a relatively easy measure to gather that is closely correlated with nutrition (Hoddinott, IFPRI). The basic logic is that people who eat relatively few different things over the course of a day (milky tea, milky tea, milky tea =1) are likely to be at a lower plane of nutrition than someone who has different things (millet porridge, maize meal and kale, maize meal and meat = 4). Drawing on some recent data from Agropastoral Ethiopia, we have the following result.



At the proposed sample size of 600 we would need the mean of this measure to increase by .48 for us to detect the improvement.