Sampling Strategy

The Quantitative Way

Bharat Bhushan Verma

Agenda

1. History of Sampling
2. Let us get terminology right
3. Why sampling works
4. How to choose a right sample size

History of Sampling

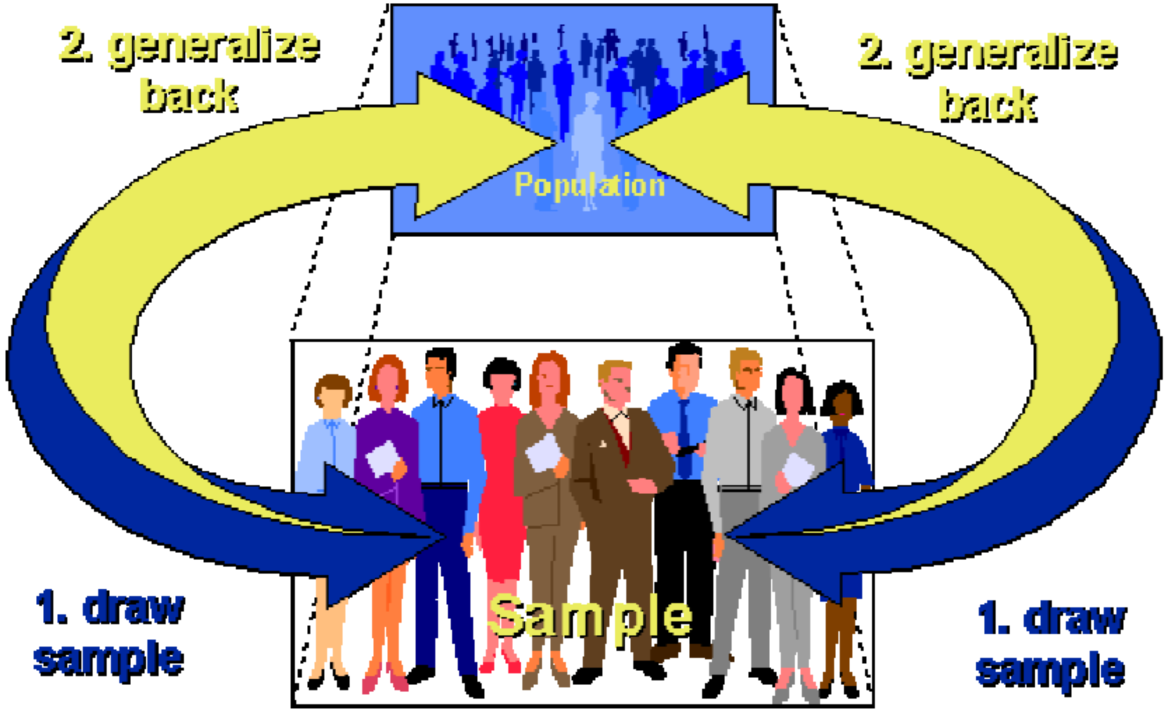
- In *The Histories* (440 BC) by Herodotus, a Persian King estimated his troops size using sampling. http://classics.mit.edu/Herodotus/history.html
- Neyman (1934) laid conceptual foundation of probability sampling. http://www.stat.cmu.edu/~brian/905-2008/papers/neyman-1934-jrss.pdf
- Godambe (1955) in unified theory of sampling established the joint optimality of Horvitz-Thompson estimates. http://www.stat.cmu.edu/~brian/905-2008/papers/Godambe-1955.pdf
- Primarily, two schools of thoughts viz. Model based (Heckman 2008) and Design based (Holland) approach are foundation of modern sampling methods. https://www.youtube.com/watch?v=S6xSEiB6E2s

What is Sampling?

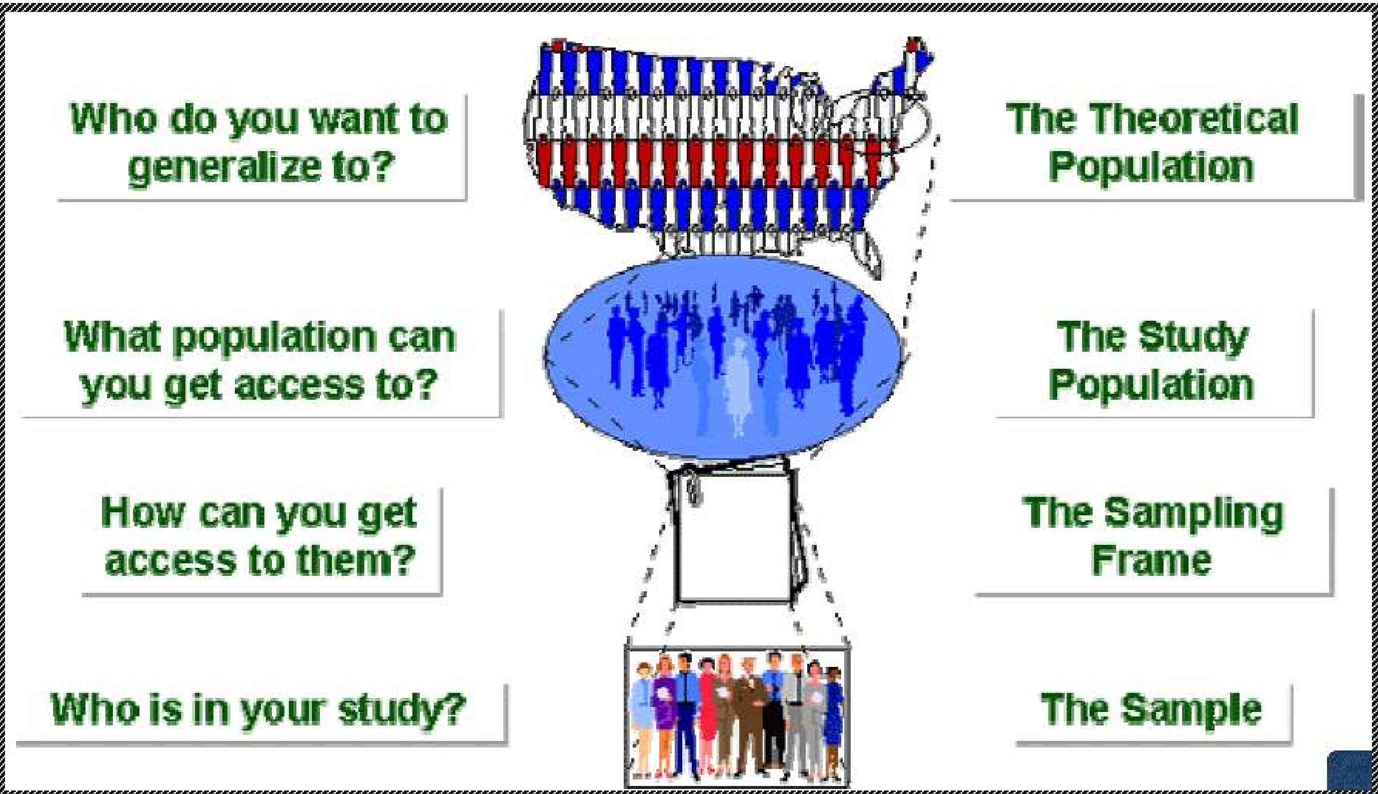
- Sampling is the process of selecting units (people, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen.

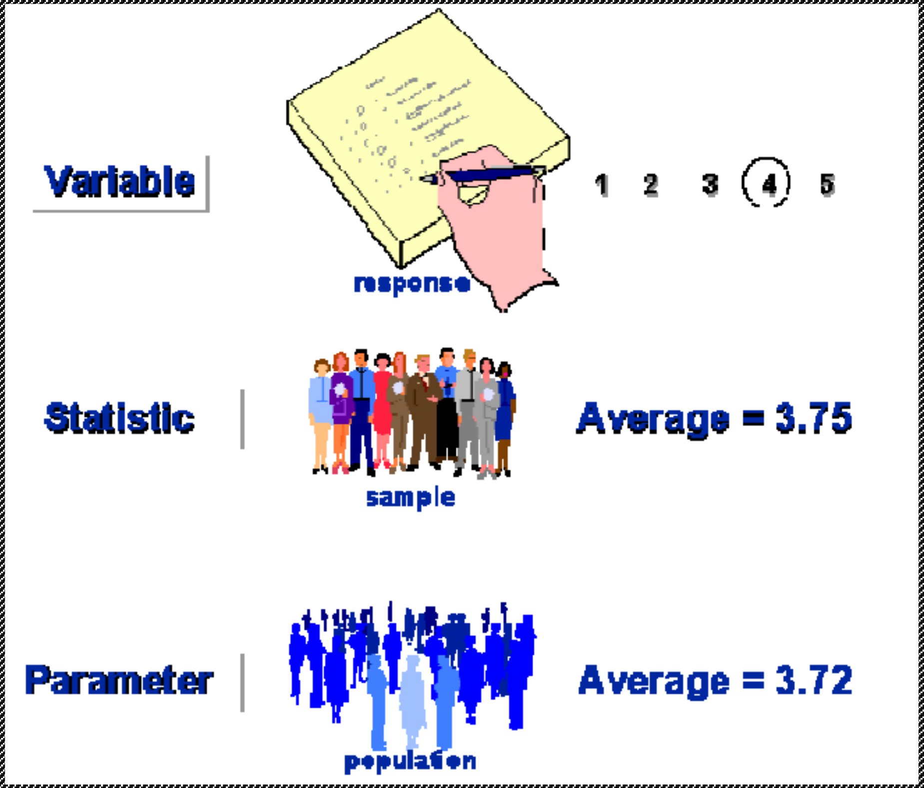
External Validity

- Assuming that there is a causal relationship between the constructs of the cause and the effect, can we generalize this effect to other persons, places or times?

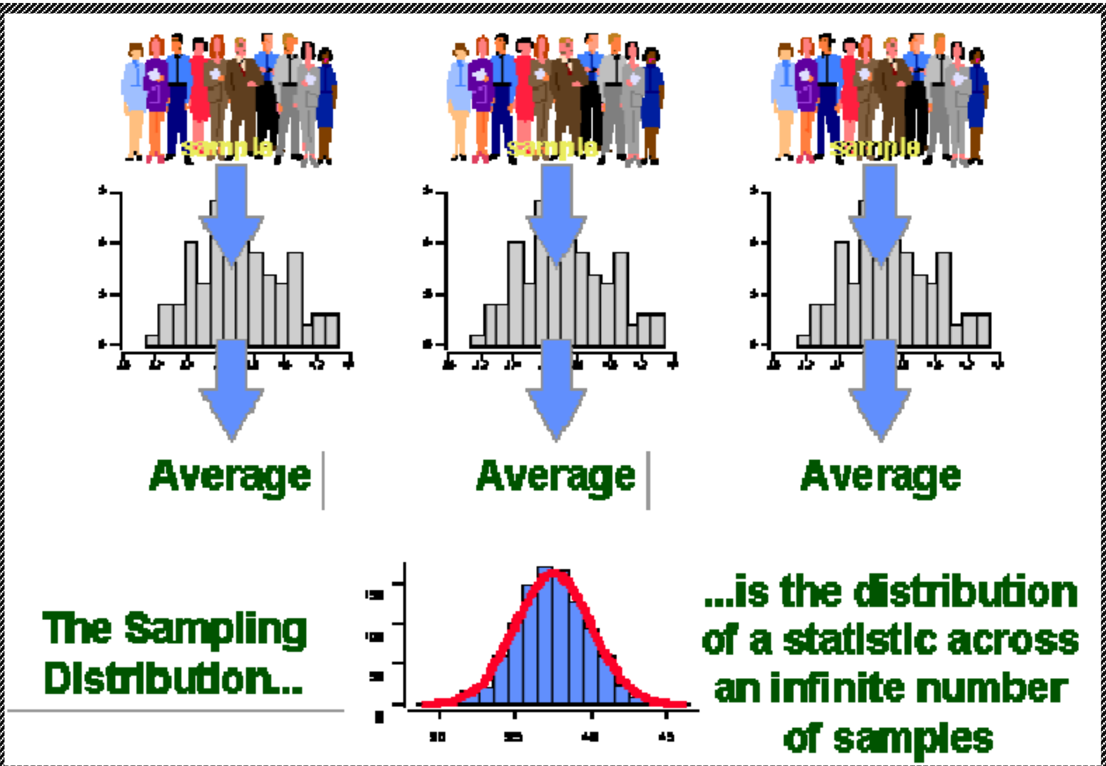


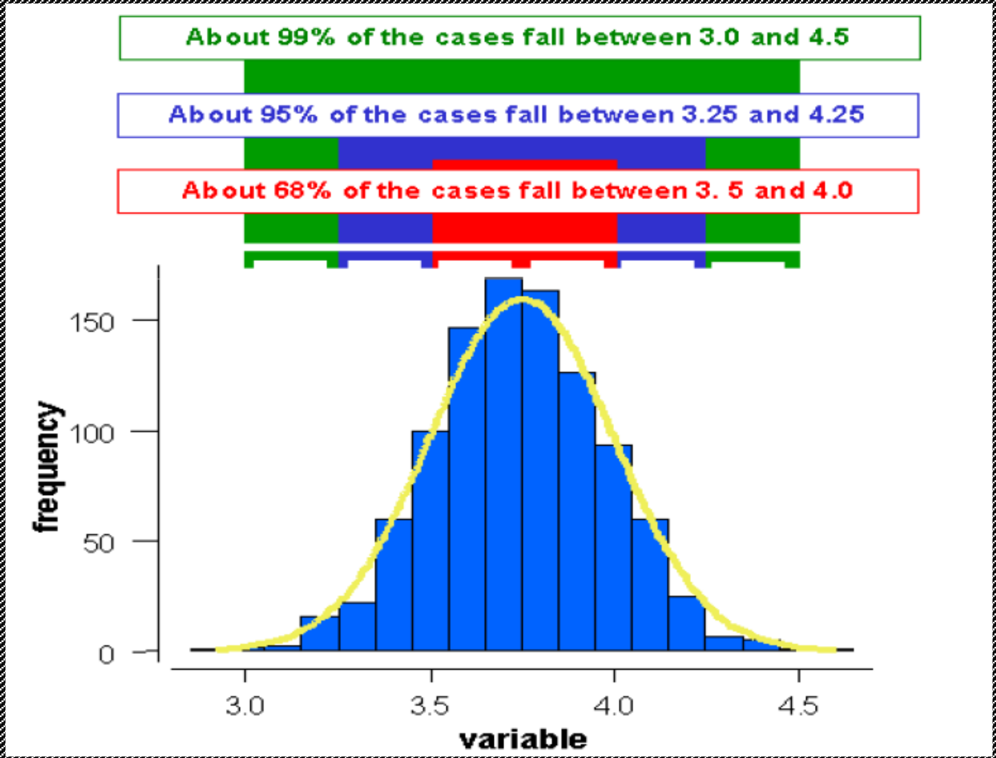
Terminology



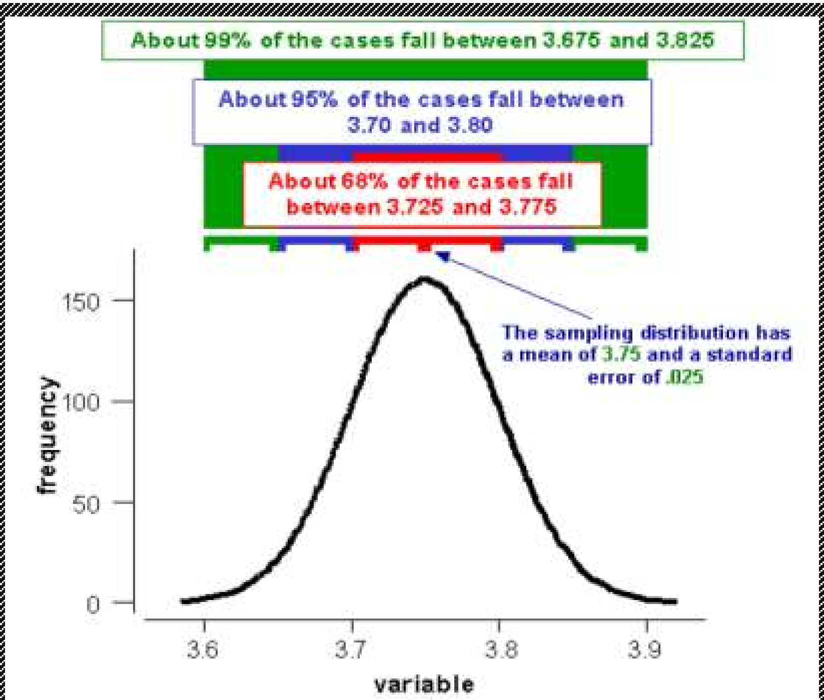


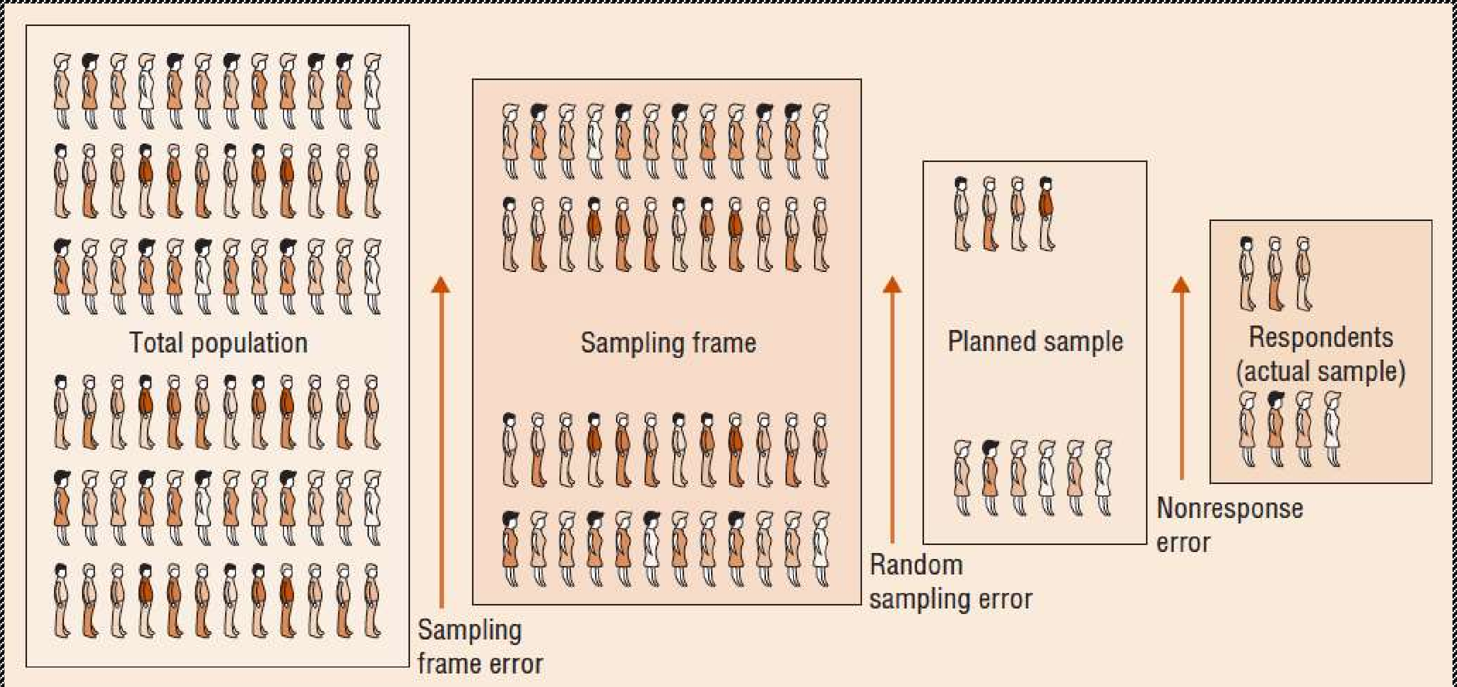
Sampling Distribution





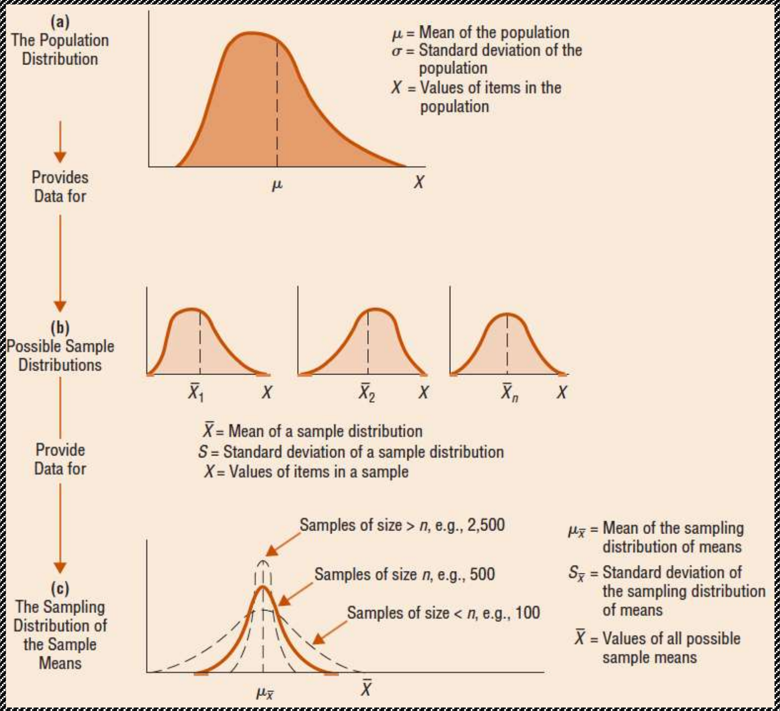
Sampling Error



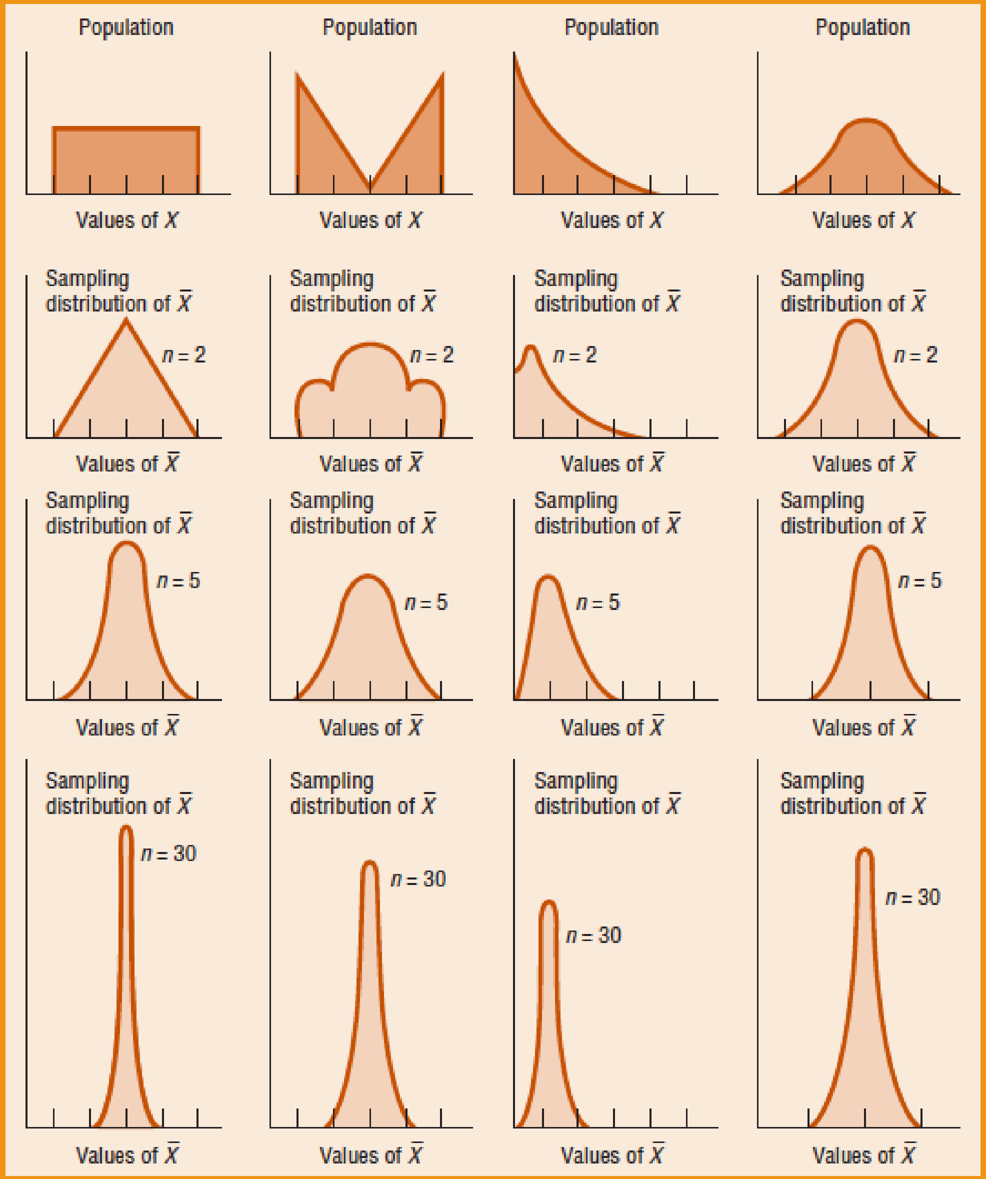


Cox, Keith K. and Ben M. Enis, The Marketing Research Process (Pacific Palisades, CA: Goodyear, 1972)
Bellenger, Danny N. and Barnet A.Greenberg, Marketing Research: A Management Information Approach (Homewood, IL: Richard D. Irwin, 1978

Why Sampling Works?

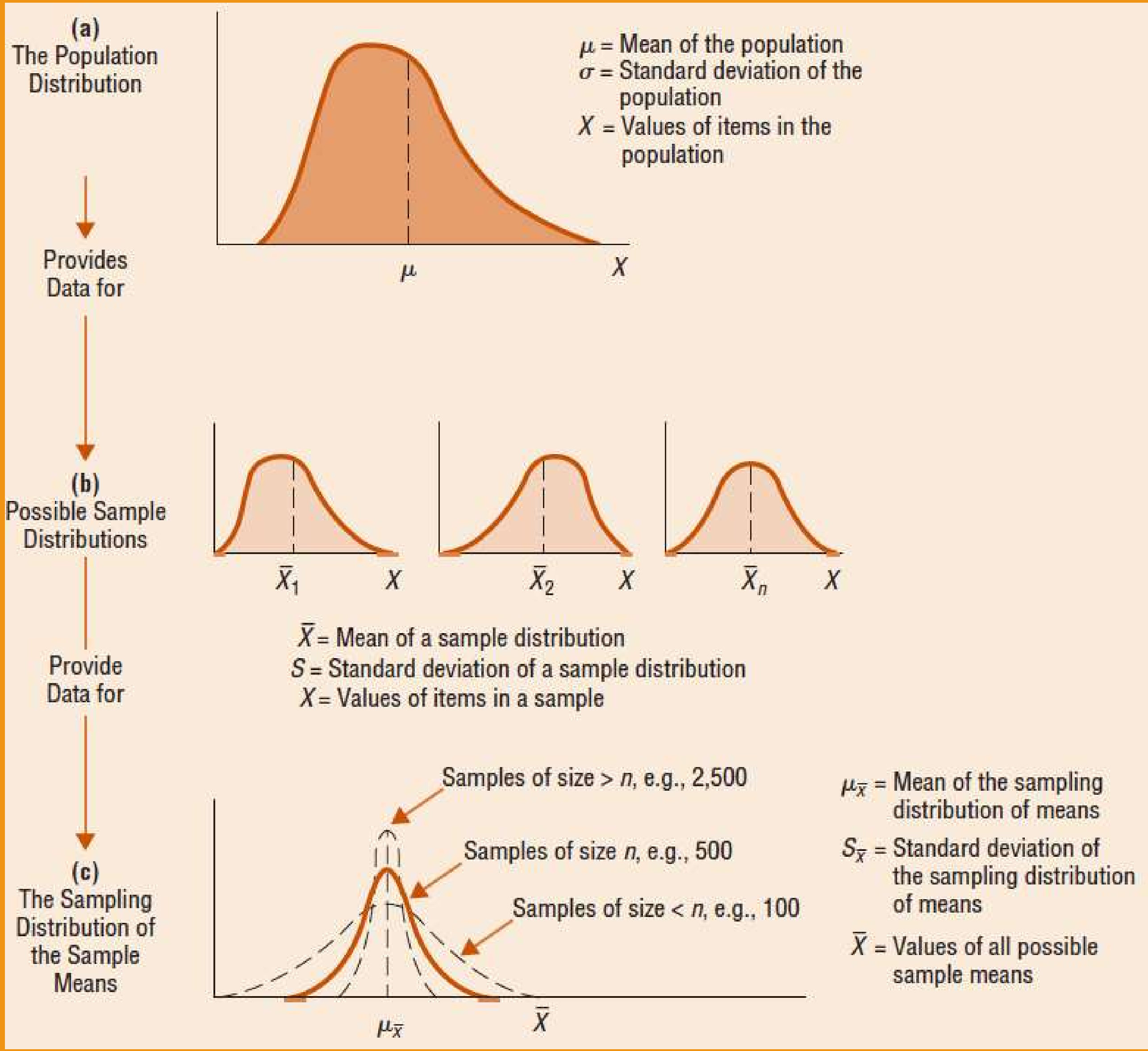


Sanders, D. H., A. F. Murphy, and R. J. Eng, Statistics: A Fresh Approach (New York: McGraw-Hill, 1980)



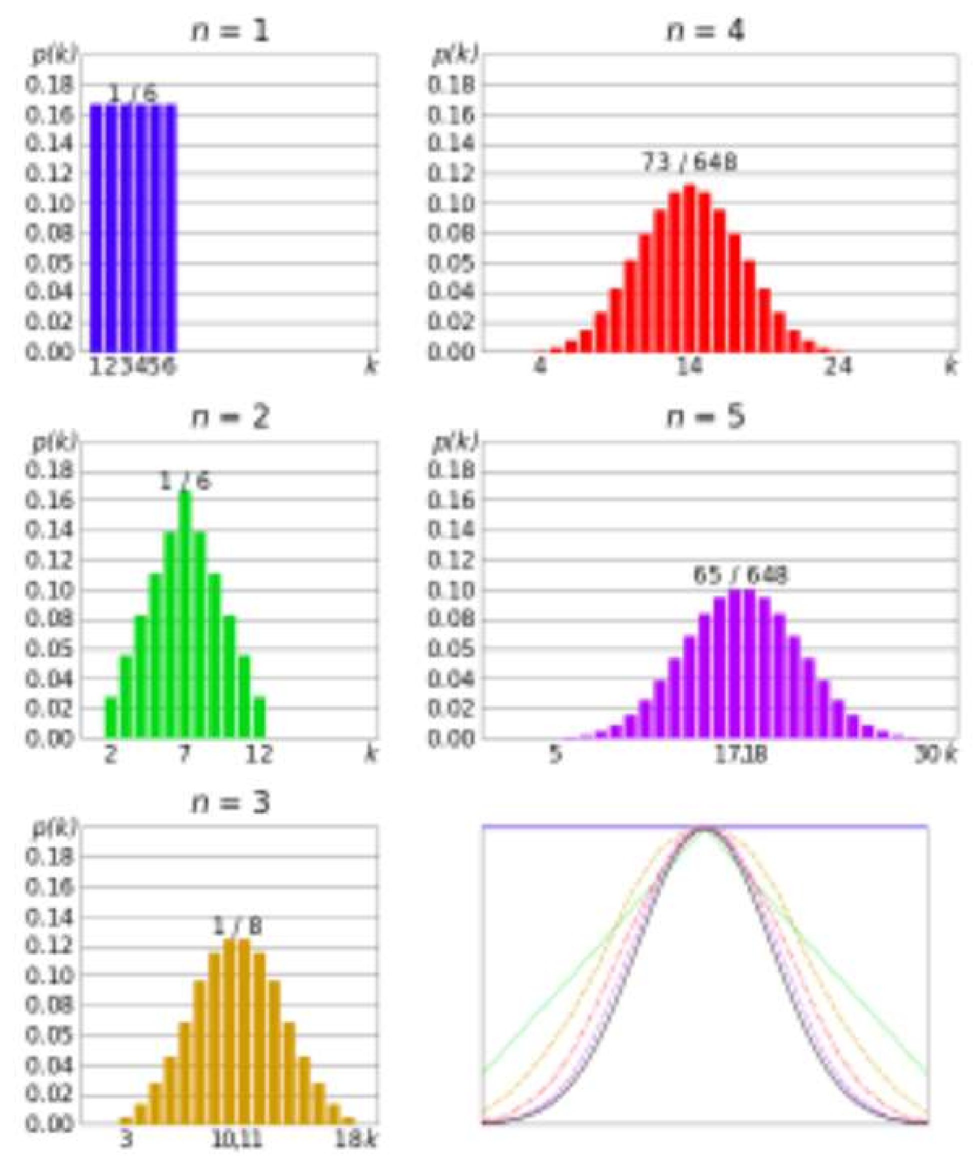
Central Limit Theorem

- The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger — *no matter what the shape of the population distribution*. This fact holds especially true for sample sizes over 30.



Example of CLT

- The **more times you roll the die**, the more likely the shape of the distribution of the means tends to look like a **normal distribution graph**.



Error in Sampling

- In 1936 Literary Digest magazine conducted a survey and predicted that Republican Alf Landon would win over Democrat Franklin D. Roosevelt by a landslide in that year’s presidential election. This prediction was wrong—and the error was due to sample selection. The post-mortems showed that Literary Digest had sampled its readers and names drawn from telephone books and auto registrations. In 1936, not everyone had a telephone or a car; thus the sample was biased toward people with means. In reality, Roosevelt received over 60 percent of the popular vote.
- In 2004, early exit polls led many to believe that John Kerry would win the U.S. presidential election.13 The exit polls were performed early on election day and done mostly in highly urban areas in the Northeast, areas that are predominantly Democratic. The resulting sample of voters responding to the early exit polls did not represent the entire U.S. population, and Kerry lost to Bush by over 3 million votes, or about 3 percent of all votes cast.

Choosing Right Sample Size

- Homogeneity of Population
- Size of Population
- Desired Margin of Error
- Desired Confidence Level
- Magic number 384 is valid for random samples.
- Compute sample size for each Strata.
- Ensure each variable has 5 - 6 non zero entries.
- Respondents = N(Variables) * 10
- Maximum 10% unless it exceeds 1000
- Read The Survey Research Handbook by Pamela Alreck and Robert Settle
- Use good deal of common sense and pragmatism
- Relax and stop worrying about the formulas

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

Issues with large samples

- In the Current era of large sample sizes
  - Little power of Goodness of Fit tests
    - Bickel, Ritov, Stoker (2001). Tailor-made tests for goodness of fit for semi parametric hypotheses.
  - Residual Analysis fails to uncover lack of fit beyond 4 – 5 dimensions.
    - Cleveland, Grousse (1991). Computational methods for local regression.
  - False Positive significance levels (p-values)
    - Galit Shmueli (2012). Too big to fail: Large samples and p-value problem

Some Quotes

- Robert Geary (1947). Normality is a myth; there never was, and never will be, a normal distribution.
- Edwards Deming (1942). Data are not taken for museum purposes; they are taken as a basis for doing something.
- If all a man has is a hammer, then every problem looks like a nail.
- Tim Hartford (2014). Statisticians have spent the past 200 years figuring out what traps lie in wait when we try to understand the world through data. The data are bigger, faster and cheaper these days - but we must not pretend that the traps have all been made safe. They have not.
- Spiegelhalter (2014). There are a lot of small data problems that occur in big data. They do not disappear because you have got lots of stuff. They get worse.
- Nate Silver (2012). The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. Before we demand more of our data, we need to demand more of ourselves.
- Tukey (1980). Neither exploratory nor confirmatory is sufficient alone. To try to replace either by other is madness. We need them both.