Lecture 6
Sampling
DHX_MET1 Methodology 1

Stanislav Ježek

Faculty of Social Studies MU

SAMPLING

- Sampling strategies and representativeness
- Sample size determination

REPRESENTATIVENESS

- Inability to collect and/or analyze ALL data – need for sampling
- Samples of units
  - people, organizations, economies, events…
- Sampling within units (in time)
  - behavior – states
- Samples represent population (~ALL data) – representativeness
- **Representativeness**
  - ideal - sample differs from population only in size (or irrelevant characteristics)
  - achievable only probabilistically
  - R in relevant characteristics – can we list them all?

STATISTICS & PARAMETERS

Statistics (*M, r, SD, B*…) computed on our sample are estimates of their population values – **parameters** (μ, ρ, σ, β…)

- $\hat{b}$ *is an estimate of* b

The difference between their values is **sampling error**

SAMPLING ERRORS

Random

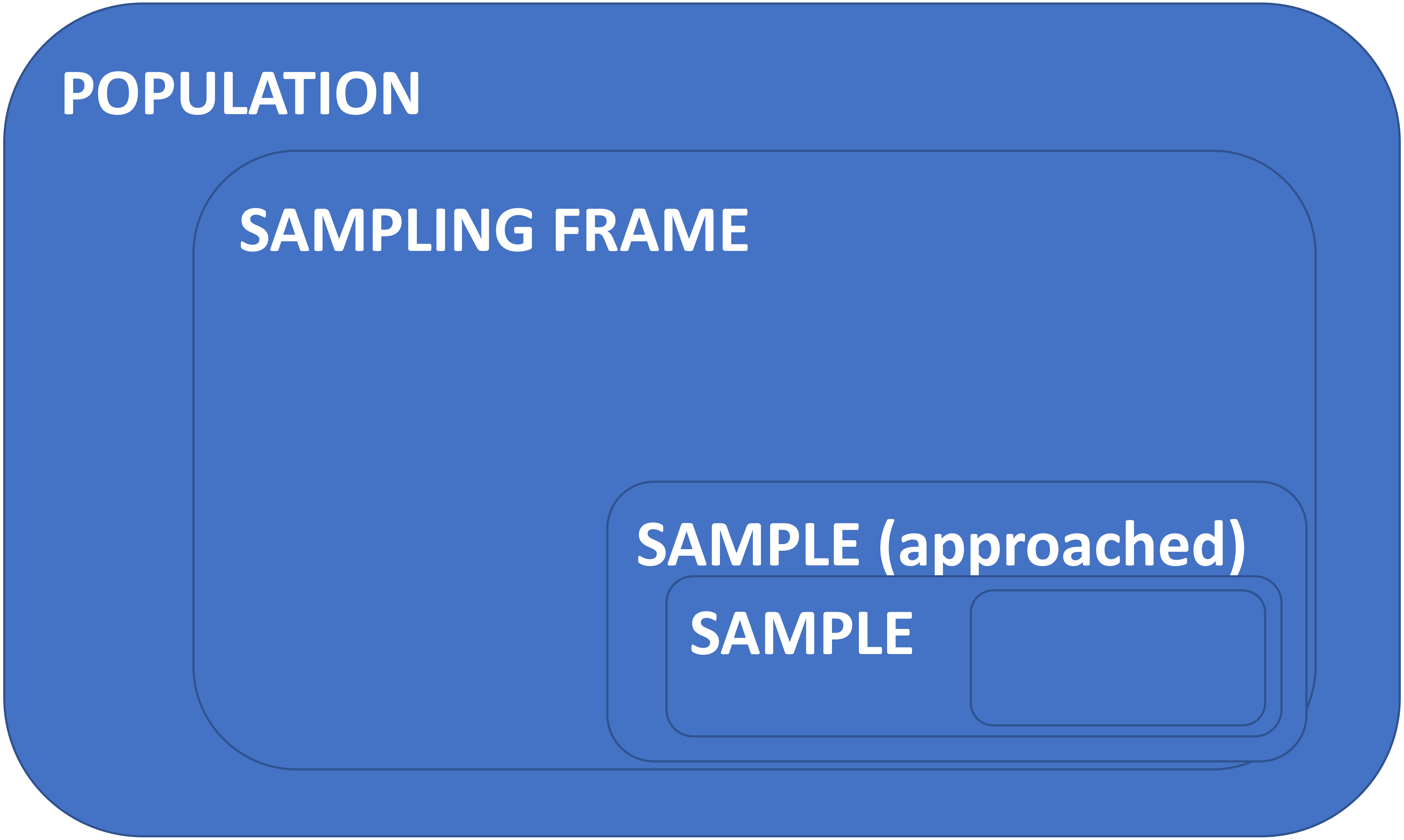
- quantifiable, estimable from probability theory
- in the long run does not bias estimates of researched characteristics

Systematic

- hard to control for unless we know exactly the process (variable) creating the error
- selection bias
- response bias

SAMPLING

1. Define the **population**.
2. Choose the **sample(-ing) frame**.
3. Decide the **sampling design/strateg**y.
4. Estimate the appropriate **sample size**.
5. **Execute** the sampling process.


POPULATION
SAMPLING FRAME
SAMPLE (approached)
SAMPLE

POPULATION

- NOT necessarily population in demographic sense
- POPULATION
  - a set of all units to which I want to generalize
  - a set of all units I want to have a sample of
- Widely/vaguely defined populations hard to sample
- Better to have a representative sample of a narrowly defined population than a biased sample of a wide one.

SAMPLING FRAMES

- LISTS, SETS of (all) units in a population from which we **can** select
  - SETS of approachable units in some communication channel, place…
  - Registries of all kind

- Function of sampling frames
  - Allow sampling, allow for a level of control over sampling stratégy
  - Allow reasoning about external validity - generalization

- First, come up with a frame, second, consider its limitations

SAMPLING STRATEGIES

NON-STRATEGIES

- Convenience samples
- Self-selected samples
- Naive snow-ball

PURPOSIVE, NON-PROBABILISTIC STRATEGIES

- Careful creation of a sample making it representative in relevant variables
- Quota sample

PROBABILISTIC STRATEGIES

- Strategies based on random selection

SAMPLING – **NON-STRATEGIES** (convenience sampling)

We have little to no control (or knowledge) over the processes leading to including a particular unit in or sample

- Difficult to argue about bias
- Difficult to argue the processes are the same as in other studies
- Difficult to apply statistical inference

- „Heterogeneity“ nor „homogeneity“ are not the solution if not considered systematically
- Making the sample bigger makes it worse – false confidence

- If it *must* be used, strive for maximum randomness

SAMPLING – **NON-PROBABILISTIC**

QUOTA SAMPLING

- building the sample so that it is representative in particular characteristics
- typically demographics – settlement, age, race, gender…
- quota = proportion of units in each category found in **population**
- unless the quota variables are super-relevant it may not be worth the effort

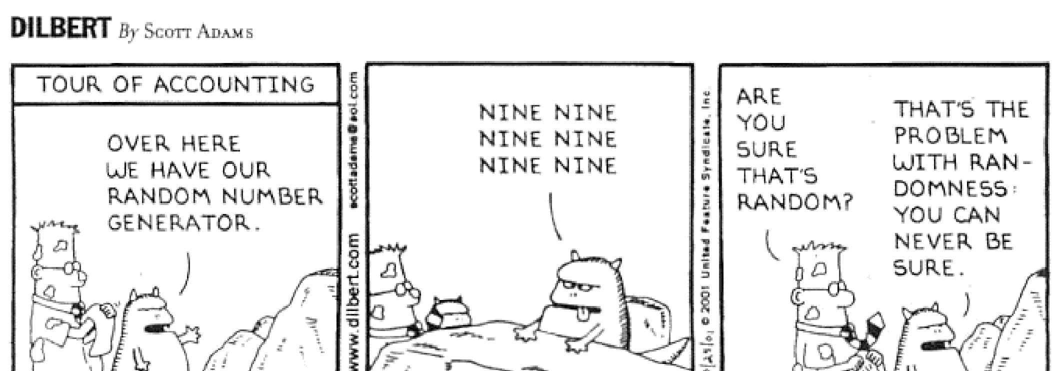
PURPOSIVE SAMPLING, THEORETICAL SAMPLING

- selection of individual units based on current needs of a (qualitative) study
- to compare, contrast…
- Emmel, N. (2013). *Sampling and choosing cases in qualitative research. A realist approach.* Sage.

SAMPLING – **PROBABILISTIC STRATEGIES**

Probabilistically unbiased estimates of parameters

- SIMPLE RANDOM, SYSTEMATIC
- STRATIFIED – let‘s assist probability; need for sub-population parameters
  - PROPORTIONAL (proportionate)
  - NON-PROPORTIONAL (disproportionate), e.g. oversampling rare subgroups
- CLUSTER (MULTISTAGE) – let‘s make it more practical
  - Hierarchical sampling procedure – higher-order units, lower-order units, individuals
  - At all levels we need sufficient numbers of units
- SNOW-BALL (probabilistic)
  - network sampling, link-tracing



Normality of Distributi[illegible]

Attributes or characteristics of the po[illegible]n are generally no[illegible]istributed. For instance, when attributes such as height and weight are considere[illegible]eople will [illegible]ered around the mean, leaving only a small number at the extremes who are either very [illegible]ry heavy or very light, and so on, as indicated in Figure 13.3. If we are to estimate the population [illegible]ics from those represented in a sample with reasonable accuracy, the sample has to be chosen so t[illegible]ibution of the characteristics of interest follows the same pattern of normal distribution in the sa[illegible] population. From the central limit theorem, we know that the sampling distribution of th[illegible]e mean is [illegible] distributed. As the **sample size** n increases, the means of the random samples ta[illegible]n practically any p[illegible]on approach a normal distribution with

SAMPLING – WHAT SAMPLE SIZE DO WE NEED?

- Large enough to make sure that relevant observed properties of the sample are unlikely to be due to sampling error
  - High „signal-to-noise ratio“
- QUAN - Power analysis, precision analysis
- QUAL – Saturation

Often seems difficult to determine beforehands – rules of thumb, e.g.

- high tens of participants in each group for a between-subject experiment
- hundreds of participants for regression models
- low tens for a within-subject experiment
- 3-5 cases for IPA
- about 10 for a GT

Rules of thumb (like p.264) **should be avoided – world is just not that simple.**

PRECISION

At a preferred level of confidence we want our (confidence) interval estimates to be as narrow as possible.

- For mean $S.E. = SD/\sqrt{N}$ and $95\%CI = M \mp 1.96S.E.$
  - The higher SD, the wider the interval estimater
  - The higher N, the narrower – precise the interval estimate
- **How large a sample do I need to have a certain level of precision and confidence?**
  - $N = \frac{2 \times 1.96SD^2}{width\ of\ CI}$ where 1,96 can be replaced by an appropriate quantile of normal or t distribution

- With other statistics and more complicated sampling desings such estimate is more difficult to compute

STATISTICAL POWER ($1-\beta$)

- In the context of statistical hypothesis testing – the probability that an effect will be found statistically significant (provided it exists)
  - $P(p<\alpha|H_0)$
  - Probability of avoiding Type II error
- Factors affecting power
  - (signal) – effect size, measurement reliability
  - (noise) – standard error … sample size, design
  - (required certainty) - alpha
- [Simulation: https://rpsychologist.com/d3/NHST/](https://rpsychologist.com/d3/NHST/)

CONSEQUENCES…

- …OF INSUFFICEINT POWER (eg. 60%)
- even if you are right about your hypothesis & have designed your study well, the data may not support your hypothesis
- gambling with funding money & participants time -> ethics
- effect size inflation
  - in confirmatory studies due to publication bias
  - in exploratory studied due to publication bias, fishing and insuffficent correction of p-values for multiple tests
- … OF EXTREMELY HIGH POWER (eg. > 95%)
  - may be just an inefficient use of research budget
  - combined with fishing and other metodological *sins* (QRPs) allows to identify very small significant efects - artefacts

Practically - 2 big questions:

1. What is the expected effect size?
   - Many standardized measures of effect size
     - distance based – Cohen‘s *d*, Hedges‘ *g*….
     - based on explained variance – R^2, r, η^2, ω^2…
   - It is safer to consider published effect sizes inflated, unless they come from meta-analysis
2. How to do power analysis for more complicated analyses than a *t*-test?
   - G*Power: http://www.gpower.hhu.de/
   - Dattalo, P. (2008). *Determining sample size: balancing power, precision, and practicality*. OUP.

SAMPLING IN QUALITATIVE RESEARCH

- Representativeness – what is to be represented?
- If there are relevant phenomena in the the studied population we want to be fairly confident they could have been encountered during study.
- Purposive sampling
  - careful selection of each case based on accummulated knowledge and immediate needs
  - Some selected before analysis, some after analyses of first cases
  - Often return to cases
- (Theoretical) Saturation – subjective belief that adding further cases would not improve the theory enough to be justifiable
- Both the reasons for selecting each case and reasons behind saturation are reported/discussed in the research report.
- Again, rules of thumb should be avoided

Sampling from finite populations

- Finite and fairly small - when it makes sense to think about the percentage of population in the sample
- *fpc* - finite population correction to standard error
- $SE_{fpc} = SE\sqrt{\frac{N-n}{N-1}}$ where *N* is population size and *n* sample size
- Rarely used bc the population is often thought to be more general than we initially see
- Eg. Current employees are only a sampling frame from possible employees or employees over some time period