Methodological issues in naturalistic driving designs

Jacques J.F. Commandeur (SWOV Leidschendam, VU Amsterdam)

November 23, 2012

DaCoTA

Overview

- Population and sample
- Simple random sampling
- How many units to sample?
- More than one item
- Other types of sampling
  - Stratified random sampling
  - Two-stage sampling
- Other estimators
- Non-sampling errors
- Conclusions for naturalistic driving designs

Population and sample

- One of the aims of naturalistic driving studies is to obtain estimates of exposure (e.g., total number of kilometers traveled) and safety performance indicators (e.g., seat belt use, speed) of all car drivers in a country. All car drivers in a country is then the *population*.
- Question: how many car drivers should then be selected in a country? How large should the sample be?
- I only consider *probabilistic* sampling techniques.
- These have in common that the probability of each car driver in the population ending up in the sample is *known*.
- Only then will the estimates of population characteristics like the mean or total from the sample be *unbiased*.

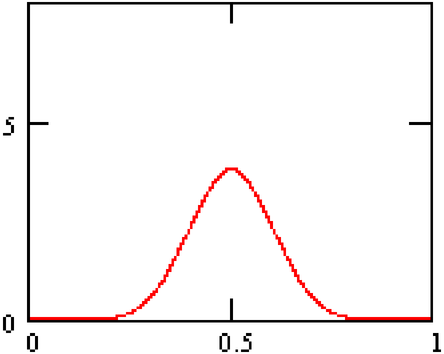
Simple random sampling

- In simple random sampling all car drivers in a country have the *same probability* of being sampled.
- If there are N car drivers, then that probability is $1/N$.
- In simple random sampling it is assumed that there is a *sampling frame*: a centralized numbered list of all members of the population.
- The n members of the sample are drawn one by one at random from this list.
- The difference between the mean or total or percentage in the sample and that in the population is called the *sampling error*.

How many units to sample?

- The secret to the answer to this question is the remarkable *central limit theorem*:
- When we repeatedly draw random samples of size *n* from a population, and we increase *n*, then the means of all these samples *more and more approximate a normal distribution, even if the variable of interest in the population does not have a normal distribution at all*.

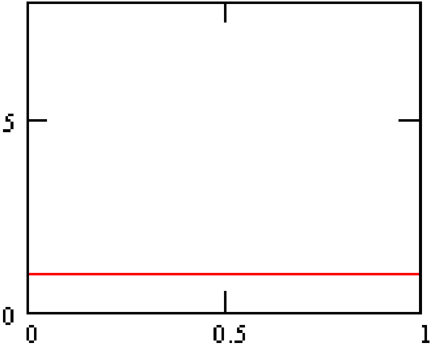
Example of a normal distribution:



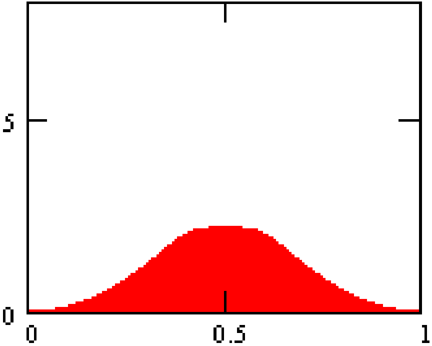
- Symmetrical and bell-shaped.

Distribution of a variable y

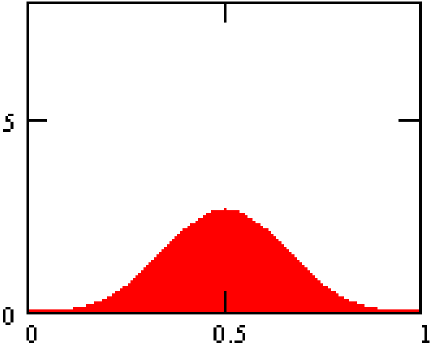
- Suppose a variable y in the population has the following distribution:



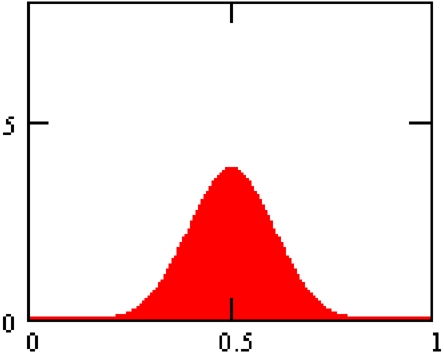
Distribution of means $\bar{y}$ for samples of size 3



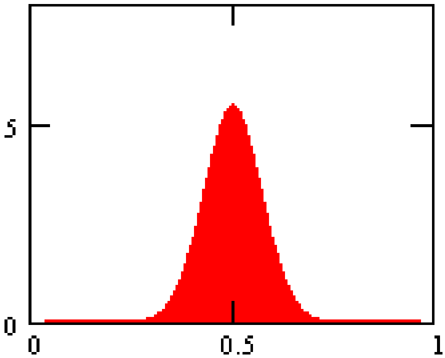
Distribution of means $\bar{y}$ for samples of size 4



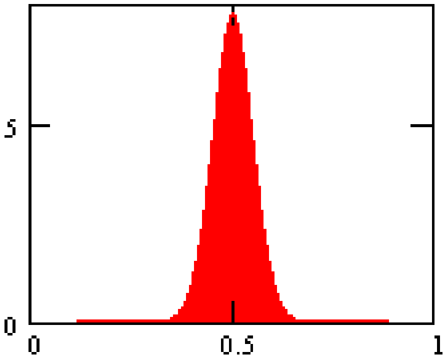
Distribution of means $\bar{y}$ for samples of size 8



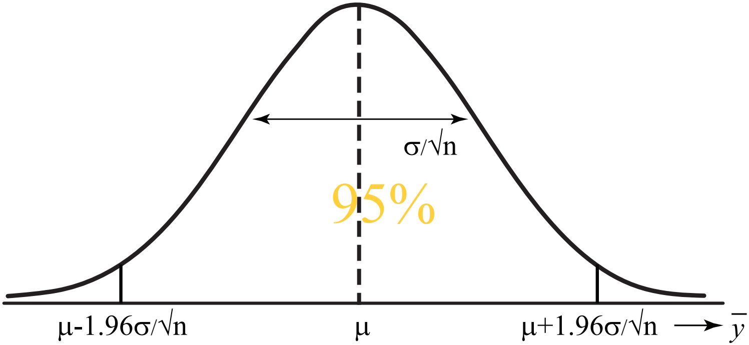
Distribution of means $\bar{y}$ for samples of size 16



Distribution of means $\bar{y}$ for samples of size 32



Sampling distribution of the mean $\bar{y}$



- In this normal distribution, where μ is the mean of y in the population and σ is its standard deviation in the population, we have that

$$\mu - 1.96\sigma/\sqrt{n} < \bar{y} < \mu + 1.96\sigma/\sqrt{n}$$

with a 95% probability.

What sample size?

- After some algebra it can be shown that we should sample

  $$n = \frac{1.96^2\sigma^2}{d^2}$$

  units from the population, where $d = |\bar{y} - \mu|$ is the *precision* we want.
- The required sample size thus gets larger as
  - we increase the precision d,
  - we increase the probability (1.96),
  - the variance σ^2 of the variable y in the population is larger.

Application

- Sample sizes required for the estimation of total number of vehicle kilometers traveled by cars in a country with precision levels of ±10%, ±5%, and ±1%, population standard deviations of $\sigma = 5,000$ and $\sigma = 15,000$, and a probability of 95%:

$\sigma = 5,000$			$\sigma = 15,000$		
±10%	±5%	±1%	±10%	±5%	±1%
43	171	4,269	385	1,537	38,416

More than one item

- When multiple variables like the number of vehicle kilometers traveled, speed, and seat belt use are involved, the minimal sample size required to achieve a certain precision should be estimated for each of these variables separately.
- The largest sample size estimate is then used to guarantee the required precision for all variables simultaneously.

Stratified random sampling

- There are several ways to improve precision – and thus to reduce sample size.
- One way is to divide the population of size N into a number of mutually exclusive sub-populations of sizes $N_1, N_2, \ldots, N_L$ such that $N = N_1 + N_2 + \cdots + N_L$, and then to apply simple random sampling to each of these L sub-populations separately. These mutually exclusive sub-populations (e.g., males and females) are called *strata*. This sampling procedure is called *stratified random sampling*.

Stratified random sampling

- With stratified random sampling considerable gains in precision of the estimates or considerable reduction in costs can be obtained when the population variance of the variable of interest is (much) smaller within each stratum than in the total population.
- This sampling strategy requires knowledge (or estimates) of the variance in each stratum of the total population.

Other estimators

- A second way to improve precision – and thus to reduce sample size – is to use other estimators like
  - the ratio estimator
  - the regression estimator
- Both estimators require knowledge of the mean or the total in the population of an auxiliary variable that is highly correlated with the variable of interest.
- In naturalistic driving a natural application of this method would be to use the data from the previous year (or quarter or month) as auxiliary variable.

Two-stage sampling

- When no centralized sampling frame is available to randomly select sample units from, two-stage sampling is used to reduce costs.
- In travel surveys in Great Britain, for example, first a random sample of 684 postcode sectors is selected in the first stage, and then 22 random addresses are selected within each randomly sampled postcode sector in the second stage, resulting in a total sample of $684 \times 33 = 15,048$ addresses.
- When estimating sample size for this type of sampling, knowledge of the variances of the units in the first stage, and of the variances of the units in the second stage is required.

Non-sampling errors

- These are
  - Measurement errors resulting in unreliable data. This is where naturalistic driving studies should really shine.
  - Selection bias as the result of non-response.

Conclusions for sample size estimation in naturalistic driving studies

- Use a probabilistic sampling design in order to obtain unbiased estimators
- Decide upon an a priori specified degree of precision with an a priori specified probability.
- For the corresponding sample size estimation, some knowledge is required about
  - the population variance(s) of the variable(s) of interest in simple random sampling,
  - the population variances in the different strata in stratified random sampling,
  - the variances of the primary and secondary units in two-stage sampling.
- When sample size is estimated for proportions or percentages, the situation is easier because a conservative estimate can always be obtained by assuming the population proportion to be equal to 0.5.

Conclusions for naturalistic driving studies

- When the objective is to measure *change* in the population over time, as is the case in naturalistic driving studies, the required precision should be established by considering the minimal difference in estimates between consecutive time points that we want to detect with certainty,.
- When information on auxiliary variables in the population is available that are highly correlated with the variable of interest this opens up the possibility of improving the precision of the estimates obtained with simple random sampling by using stratified random sampling. For kilometers traveled by cars, this could be year of construction of the car, fuel type (petrol versus diesel), ownership (private versus business), for example.

Conclusions for naturalistic driving studies

- When several items need to be estimated, estimate sample size for each of these items separately.
- If costs are not an issue, the largest sample size should be used in order to guarantee the required precision for all items.
- The continuous nature of the measurements obtained in a naturalistic driving study implies that the ratio and/or regression estimators are natural and well-suited candidates for statistically improving the precision of the population parameter estimates.

Conclusions for naturalistic driving studies

- When estimates for sub-populations of the total passenger car population in a country are required, it is recommended to use these sub-populations as strata in a stratified random sampling design because this yields more precise estimates than when the sub-populations cut through the strata.
- The estimation of the required sample size for a pre-specified precision should always take the problem of non-response into account, and the estimated sample size should be increased accordingly.

Conclusions for naturalistic driving studies

- In some countries at least, it should be possible to get information on the characteristics of the non-respondents by using the double sampling for non-response approach. This can be applied in two ways: either by obtaining a random sub-sample of the non-respondents and then make sure that they participate in the study after all, or by obtaining a random sub-sample of the non-respondents and then consulting a second sampling frame also containing (estimates of) the required information.
- Finally, if distributions of other variables in the total populations are known, such as demographic variables for car drivers, or technical characteristics for passenger cars, then these can be used to *calibrate* the sample to conform with the population distributions, thus partly correcting for non-response.