STA304

Neil Montgomery
2016-01-14

general sampling concepts

basic definitions

A motivating example: *a natural gas distribution company wishes to determine the proportion of a certain valve that is in a failed state.*

- **element** is a unit on which a measurement it taken: *a valve*.
- **population** is a collection of elements: *all the valves*.
  - **target population** is the *intended* population of interest: *all the valves*.
  - **sampling population** is the population effectively sampled: *all the valves they have records of*.

more basic definitions

- A collection of **sampling units** is a *partition* of the population
- A partition of a set is mathematical jargon - it's a collection of subsets that satisfy these properties:
    - every element of the set is in one of the subsets
    - but only in *one* of the subsets
- The simplest partition of $\{e_1, e_2, \ldots, e_N\}$ is just $\{\{e_1\}, \{e_2\}, \ldots, \{e_N\}\}$.
- But other collections of sampling units also occur naturally, as we shall see.
- **frame** is a list of sampling units: *the list of valves in the database*.
- **sample** is a collection of sampling units
- *Technically a sample is actually the union of elements from a collection of sampling units.*

conveniences and conventions

Notice how N has crept in as *population size*.

In theoretical discussions we'll just say **population** (not worrying about target vs. sampling)

An element is not the same thing as a value measured on the element. A population is the set of elements themselves, say

$$\{e_1, e_2, \ldots, e_N\}$$

with corresponding values

$$\{y_1, y_2, \ldots, y_N\},$$

and we will lazily refer to the latter set as the **population** when there is no confusion (defined as "when the instructor is not confused").

more conveniences and conventions

Sample size is n. The notation for the sample values perhaps ought to be:

$$\{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}, \text{ where } \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\},$$

but the convention is to flagrantly abuse the index notation and just use:

$$\{y_1, y_2, \ldots, y_n\}$$

Thanks a lot William G. Cochrane!

"replacement"

- *with replacement*: a sampling unit can appear more than once in the sample.
- *without replacement*: a sampling unit can appear only once in the sample.

statistical concepts, revisited for sampling

random variable

- A tricky concept from probability theory: "a real valued function of a sample space". (The full theory is even more involved.)
- The fundamental property of a random variable is its *distribution*: the possible outcomes and their probabilities.
- This course happens to be about *discrete* random variables, whose distributions are simply represented by the so-called *probability (mass) function* (pmf) expressed as (for random variable X):

$$p(x) = P(X = x).$$

- Sadly, the convention in sampling is to use y as the "generic random variable". We'll call its generic pmf $p(y)$.

expected value and friends

$$E(y) = \sum_y y\, p(y)$$

$$V(y) = E\big((y - E(y))^2\big) = \sum_y (y - E(y))^2\, p(y)$$

The book prematurely aliases these to μ and σ^2, but I'm going to wait.

$$E(ay + b) = aE(y) + b$$

$$E(x + y) = E(x) + E(y)$$

$$V(ay + b) = a^2 V(y)$$

covariance

$$\begin{aligned} Cov(x, y) &= E((x - E(x))(y - E(y))) \\ &= E(xy - xE(y) - E(x)y + E(x)E(y)) \\ &= E(xy) - E(x)E(y) \end{aligned}$$

When x and y are *independent random variables* $E(xy) = E(x)E(y)$ and $Cov(x, y) = 0$.

$$V(ax + by) = a^2 V(x) + b^2 V(y) + 2abCov(x, y)$$

$$Cov(y, y) = V(y)$$

source of randomness *for this course*

- In this course the population values $\{y_1, y_2, \dots, y_N\}$ are considered to be a fixed list of numbers. Randomness comes from picking values at random only: *design-based sampling*.
- There is also *model-based sampling* in which the population values are considered to be a sequence of random variables, giving a second source of randomness.

simple random sampling

the *definition* of three population parameters

- **population total:** $\tau = \sum_{i=1}^{N} y_i$ $(= N\mu)$
- **population mean:** $\mu = \frac{1}{N} \sum_{i=1}^{N} y_i = \tau/N$
- **population variance:** $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2$
- These last two numbers are *analogous* to what we understand as "mean" and "variance" as a properties of a random variable.
- We don't know what these values are for the population, so we will gather a sample $\{y_1, y_2, \ldots, y_n\}$ in order to *infer* their values.

"… so we will gather a sample …"

- The sampling technique will determine how to estimate the population parameter(s)
- One technique to select a sample of size n in a way that all samples of size n have the same probability of being selected. This is called **simple random sampling.**
- It is *necessary* that each unit has the same probability of being selected, but not *sufficient*

how to simply random sample?

- If you have a sampling frame, use a computer to randomly select units from the frame.
- The computer will use its own internal list of (pseudo-)random numbers between 0 and 1 to do this for you.
- You could build a time machine and go back to the 1950's and use a printed table of random digits.
  - Index the population from 1 to N.
  - Figure out the d such that $10^{d-1} < N \leq 10^d$.
  - Use successive groups of d digits from the table to select units by their index, discarding any random group of digits if it is larger than N or it has appeared already.
- What if you don't have a frame?

properties of a simple random sample (without replacement)

- The sample $\{y_1, y_2, \ldots, y_n\}$ is a list of random variables.
- Each of the y_i have the *same* distribution.
  - Name: *discrete uniform distribution over* $\{y_1, y_2, \ldots, y_N\}$.
  - $E(y_i) = \sum_{i=1} y_i \frac{1}{N} = \mu$
  - $V(y_i) = \sum_{i=1} (y_i - \mu)^2 \frac{1}{N} = \sigma^2$
- But they are *not* independent random variables. In particular:

$$Cov(y_i, y_j) = -\frac{1}{N-1}\sigma^2$$