Definition of Samples

James J. Heckman
Center for the Economics of Human Development
University of Chicago

Econ 312, Spring 2022

- General definitions of
  1. Random Sampling
  2. Censored Sampling
  3. Truncated Sample
  4. Choice Based Sample
  5. Also consider truncated and censored random variables.

(1) Random sampling: (Really *simple* random sampling)

- iid. random variables with density $f(X)$. Random sampling in general is derivation of a sample by a *calculatable* rule.

$$\text{Prob. of sample } = f(X_1)f(X_2)f(X_3)\ ...\ f(X_N)$$

- Problem of getting an X is $f(X)$. Thus in a population, the probability of getting into the sample is $f(X)$. This is simple random sampling.

(2) Truncated Sample

$$\begin{aligned} f(X) &: \quad \text{density of a random variable} \\ a &< X < b : b, a \text{ may be infinite} \end{aligned}$$

- We observe X if $X < R$ (right truncation)
- or if if $X > L$ (left truncation).
- Key property: latent variable X in the population — we know

$$X^* = X \qquad \text{when } L < X < R$$

- (Assume simple random sampling of a larger population). We only observe X^* and we do not know the number of observations in (larger) random sample for which X is outside the interval. We only know the reduced sample if density in population (untruncated) is $f(X)$, then density of X^* is

$$\frac{f(X^*)}{\int_L^R f(z)dz} \qquad L \le X^* \le R$$

- (Note further that there are an infinity of underidentified distributions consistent with the truncated one.)
   3. Censored Sample: We observe X^* as before but *we know the number of observations outside interval.* We encounter two types of censoring:
      a. Type one censoring : we only observe a variable if it lies in a range, number of values of Y outside the range is known.
      b. Type Two Censoring: *Fixed proportion* of the sample is censored in advance (*e.g.* stop observing light bulb burnout when we have a proportion - say m).

(4) If we have that in both (3) and (2), X is a *truncated random variable* (the range of the random variable is truncated).

(5) New term: coined in recent econometric work — *censored random variable*. It is inherently a bivariate concept. Joint *pdf* $-f(Y_1, Y_2)$. Then we have that we observe Y_1 only if Y_2 exceeds some value or lies in some range, *e.g.*

$$L < Y_2 < R \tag{1}$$

Prob. of this event is

$$\int_L^R f_2(Y_2)dY_2$$

- The random variable Y_1 is *not truncated*. We observe Y_1 only if the condition on Y_2 is satisfied.

- The *sample* may or may not be truncated. Thus, it is the case that if we observe Y_1, given selection criterion (*), but we do not know the number of observations in the larger random sample variable for which the Y_2 restriction is violated, we have a truncated sample and a censored random variable. Now clearly we may put a restriction on Y_1 *e.g.* we observe Y_1 only if $L_2 < Y_2 < R_2$ and $L_1 < Y_1 < R_1$. Thus define $Y_1^* = Y_1$ for $L_1 < Y_1 < R_1$.

$$g(Y_1^*) = \frac{\displaystyle\int_{L_2}^{R_2} f(Y_1^*, Y_2) dY_2}{\displaystyle\int_{L_1}^{R_1} \int_{L_2}^{R_2} f(Y_1, Y_2) dY_1 dY_2}$$

(6) New term in discrete choice literature – *choice based sampling*. Consider the random variable Y to be discrete. Z are exogenous explanatory variables. The theory produces a $g(Y \mid Z, \theta)$: discrete choice model $h(Z)$ in the distribution of the population exogenous variables.

$$Y_j \subset \{1, \ldots, J\}$$

elements of choice set.

Exogenous Sampling: we pick Z, then observe Y. Sample Z according to the density $k(Z)$ and observe the value of Y, the choice. Likelihood of an observation (Y, Z) is

$$g(Y \mid Z, \theta)k(Z)$$

when $k(Z) = h(Z)$, we have random sampling. Otherwise we have *stratified* sampling.

Choice Based Samples

- Pick Y first (*e.g.* travel mode). Probability of selecting Y is $C(Y)$.
- $f(Y, Z)$ is the joint density of Y and Z in the population.

$$\begin{aligned} f(Y, Z \mid \theta) &= g(Y \mid Z, \theta) h(Z) = \varphi(Z \mid Y) f(Y \mid \theta) \\ f(Y \mid \theta) &= \int g(Y \mid Z, \theta) h(Z) dZ \end{aligned}$$

- Given Y we observe Z (the implicit assumption is that we are sampling only on Y, not on Y and Z). Probability of *sampled* Z, Y is $\varphi(Z \mid Y) C(Y)$.
- A fact we use later is

$$\begin{aligned}\varphi(Z \mid Y)C(Y) &= \left\{ \frac{g(Y \mid Z)h(Z)}{f(Y)} \right\} C(Y) \\ &= \frac{g(Y \mid Z)h(Z)C(Y)}{\left[\int g(Y \mid Z)h(Z)dZ\right]}.\end{aligned}$$

When $C(Y) = f(Y) = \int g(Y \mid Z)h(Z)dZ$, choice based sampling is random sampling.

- Note, the likelihood function in an exogenous sampling scheme is

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{I} f(Y_i, Z_i) = \prod_{i=1}^{I} f(Y_i \mid Z_i, \theta) h(Z_i) \\ \ln \mathcal{L} &= \sum_{i=1}^{I} \ln f(Y_i \mid Z_i) + \sum \ln h(Z_i). \end{aligned}$$

- By exogeneity, we get the lack of dependence of distribution of Z on θ.

- Likelihood function for a choice-based sampling scheme is

$$\ln \mathcal{L} = \sum_{i=1}^{I} \left[\ln g(Y_i \mid Z_i) + \ln h(Z_i) - \ln f(Y_i) + \ln C(Y_i) \right].$$

- In several, $f(Y)$ depends on parameters θ. $\therefore$ Max with θ.

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = \sum_{i=1}^{I} \frac{\partial \ln g(Y_i \mid Z_i)}{\partial \theta} - \sum_{i=1}^{I} \frac{\partial \ln f(Y_i)}{\partial \theta}.$$

- We neglect the second term in forming the usual estimators using only the first term. That is the source of the inconsistency.

Choice Based Sample:

- An example in discrete choice.
- (c) Draw d by $\varphi(d)$.
- (d) Draw X by $f(X \mid d = 1)$.
- Joint density of data:

$$\begin{aligned} & \varphi(d = 1) f(X \mid d = 1, \theta_0) \\ = \quad & \varphi(d = 1) \left[\frac{\Pr(d = 1 \mid X, \theta_0) f(X)}{\Pr(d = 1 \mid \theta_0)} \right] \end{aligned}$$

- Now in a choice-based sample

$$\Pr^*(d = 1 \mid X) = \frac{f(X \mid d = 1, \theta_0)\varphi(d = 1)}{g^*(X)}$$

where $g^*(X)$ is the sampled X data. Joint density of *data* X is given by:

$$g^*(X) = f(X \mid d = 1, \theta)\varphi(d = 1) + f(X \mid d = 0, \theta)\varphi(d = 1)$$

and

$$\Pr(d = 1 \mid X) = \frac{f(X \mid d = 1)\Pr(d = 1)}{f(X)}$$

- Assume $f(X) > 0$. Using Bayes' theorem for Y write:

- $\Pr^*(d = 1 \mid X) =$

$$\frac{\dfrac{\Pr(d=1 \mid X,\theta) f(X)}{\Pr(d=1 \mid \theta)}\varphi(d=1)}{\dfrac{\Pr(d=1 \mid X,\theta) f(X)}{\Pr(d=1 \mid \theta)}\varphi(d=1) + \dfrac{\Pr(d=0 \mid X,\theta) f(X)}{\Pr(d=0 \mid \theta)}\varphi(d=0)}$$

$$= \frac{\Pr(d=1 \mid X,\theta)\varphi(d=1)/\Pr(d=1 \mid \theta)}{\Pr(d=1 \mid X,\theta)\dfrac{\varphi(d=1)}{\Pr(d=1 \mid \theta)} + \Pr(d=0 \mid X,\theta)\dfrac{\varphi(d=0)}{\Pr(d=0 \mid \theta)}}.$$

- Now we missample the population with density $f(X \mid d = 1)$ in a choice based sample:

$$
\begin{aligned}
\Pr{}^*(d &= 1 \,|\, X) = \frac{f(X \mid d=1,\theta_0)\varphi(d=1)}{f(X \mid d=1,\theta)\varphi(d=1) + f(X \mid d=0,\theta_0)\varphi(d=0)} \\
&= \frac{\dfrac{f(X)\Pr(d=1 \mid X)}{\Pr(d=1)}\varphi(d=1)}{\dfrac{f(X)\Pr(d=1 \mid X)}{\Pr(d=1)}\varphi(d=1) + \dfrac{f(X)\Pr(d=0 \mid X)}{\Pr(d=0)}\varphi(d=0)} \\
&= \frac{\Pr(d=1 \mid X)}{\Pr(d=1 \mid X) + \Pr(d=0 \mid X)\dfrac{\varphi(d=0)}{\varphi(d=1)} \cdot \dfrac{\Pr(d=1)}{\Pr(d=0)}} \\
&= \frac{1}{1 + \left[\dfrac{\Pr(d=0 \mid X)}{\Pr(d=1 \mid X)}\right] \cdot \dfrac{\varphi(d=0)}{\varphi(d=1)} \cdot \dfrac{\Pr(d=1)}{\Pr(d=0)}}
\end{aligned}
$$

- With logit we get

$$\Pr^*(d=1 \mid X) = \frac{1}{1 + e^{-(\alpha_0 + X\beta) + \ln\left[\frac{\varphi(d=0)}{\varphi(d=1)} \cdot \frac{\Pr(d=1)}{\Pr(d=0)}\right]}}.$$

This goes into an intercept term:

$$= \frac{e^{\alpha^* + X\beta}}{1 + e^{\alpha^* + X\beta}}$$

$$\alpha^* = \alpha_0 - \ln\left[\frac{\varphi(d=0)}{\varphi(d=1)} \cdot \frac{\Pr(d=1)}{\Pr(d=0)}\right].$$

- How to solve problem: Reweight data by relative frequency in population.
- (Idea due to C.R. Rao, 1965, 1986.)
- Joint density of the data is

$$f(X \mid d = 1)\varphi(d = 1).$$

Use Bayes' rule to obtain

$$\frac{P(d = 1 \mid X)f(X)}{P(d = 1)}\varphi(d = 1).$$

- Now weight by

$$\frac{P(d = 1)}{\varphi(d = 1)}.$$

- Solution: Reweight the data to form the following weighted likelihood:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\Pr(d_i=1)}{\varphi(d_i=1)}(d_i^*)\ln\Pr(d_i=1\mid X,\theta)+\frac{\Pr(d_i=0)}{\varphi(d_i=0)}(1-d_i^*)\ln\Pr(d_i$$

$$P\int\{[\Pr(d=1\mid X,\theta_0)f(X\mid\theta_0)]\ln\Pr(d=1\mid X,\theta)+$$

$$\int[\Pr(d=0\mid X,\theta_0)f(X\mid\theta_0)]\ln\Pr(d=0\mid X,\theta)\}\,f(X\mid d)dX$$

- This step uses the result that reweighting the data gives us the true density.
- Better way to see what is giving on:

$$\frac{f(X \mid d=1)\varphi(d=1)}{g^*(X)} = \frac{\Pr(d=1 \mid X)f(X)}{g^*(X)} \frac{\varphi(d=1)}{\Pr(d=1)}.$$

- Reweight the data: when we reweight the data, g^* is restored to f.

$$f(X) = f(X \mid d=1)\varphi(d=1)\left[\frac{P(d=1)}{\varphi(d=1)}\right] + f(X \mid d=0)\varphi(d=0)\frac{\Pr(d=0)}{\varphi(d=0)}$$