Sample Selection: a central idea for last 35 years
See an observed relationship in sample data. **Causal?**
Think hard about what is and is not measured or observed.
Data generating process: loosely what we observe.

Involves two distinct processes: sampling rule by survey organization. For example, sample poor households at twice their population occurrence.

The other dimension of the data generating process is the decisions made by the agents whose behavior we study.

When considering household surveys, there is a potential third process: responding process, whether a sampled individual or household participates in the survey (unit) or answer a particular question (item).
Let $Y$ denote a vector of outcomes of interest. Let $X$ denote control or explanatory variables. Population distribution function of $(Y, X)$ is $F(y, x)$ with density $f(y, x)$.

Will use $f$ to denote population quantities (marginal, joint, cdf) etc, and $g$ to denote sample quantities.

Any sampling rule can be interpreted as producing a non–negative weighting function $\omega(y, x)$ that alters the population density.

$$g(y, x) = \tilde{\omega}(y, x)f(y, x)$$
Fundamental Density Equation

\[ g(y, x) = \frac{\omega(y, x)f(y, x)}{\int \omega(y, x)f(y, x) \, dy \, dx} \]

\[ \tilde{\omega}(y, x) = \frac{\omega(y, x)}{\int \omega(y, x)f(y, x) \, dy \, dx} \]

\[ g(y, x) = \tilde{\omega}(y, x)f(y, x) \]
Schemes with $\omega(y, x) = 0$ for some values of $(y, x)$ create special problems.

Entertain notion that not all values of $(Y, X)$ are observed (sampled).

Let $i(y, x) = 0$ if a potential observation at $(y, x)$ can not be sampled. And equal to 1 otherwise.

Let $\delta = 1$ record that a potential observation is sampled; the value of $(Y, X)$ is observed. And $\delta = 0$, if it is not.
In the population the proportion sampled equals

\[ \Pr(\delta = 1) = \int i(y, x) f(y, x) \, dy \, dx \]

while,

\[ \Pr(\delta = 0) = 1 - \Pr(\delta = 1). \]
Truncated and Censored Samples

Samples in which $\omega(y, x) = 0$ nor non-negligible proportion of the population, $\Pr(\delta = 0) > 0$,

Truncated Sample is one in which $\Pr(\delta = 1)$ is not known or can not be consistently estimated from sample.

The density $g(y, x)$ is of all sampled values of $(Y, X)$.

Censored Sample is one in which $\Pr(\delta = 1)$ is known or can be consistently estimated from the sample.

It is known where $i(y, x) = 0$ for all values of $y, x$. But do not know values $Y, X$ when $\omega(y, x) = 0$.

Notationally convenient to define $(Y^*, X^*) = (0, 0)$ for $\omega(y, x) = 0 = i(y, x)$ if no point–mass at $(0, 0)$. 
Simple Random Sample

Textbook sampling process: simple random sample.

Definition or Key feature: ?

Ans. Every element in population has an equal chance of being included in the sample. Or \( \omega(y, x) = \omega = 1. \)

Thus, can treat the sample as if it were the population.
As one might expect: $\omega(y, x) = \omega(x)$.

Recognize that we can integrate out $y$ from the denominator. Then sample density is:

$$g(y, x) = \frac{\omega(x)f(y, x)}{\int \omega(x)f(x) \, dx}$$

$$= f(y|x) \left( \frac{w(x)f(x)}{\int \omega(x)f(x) \, dx} \right)$$

$$= f(y|x) \cdot g(x)$$
Exogenous Sampling

Just to be clear:

\[ g(x) = \frac{\omega(x)f(x)}{\int \omega(x)f(x) \, dx}. \]

So that

\[ g(y|x) = \frac{g(y, x)}{g(x)} = f(y|x). \]

Thus, for analyses in which conditional distribution \( g(y|x) \) is the object of interest, then exogenous sampling process need not be recognized.
While conditional distribution in the sample equals the conditional distribution in the population, with exogenous sampling

\[ g(x) \neq f(x) \]

This implies that \( \bar{X} \neq \mu_x \) And \( \bar{Y} \neq \mu_y \).

Use sample weights to make correct inference on characteristics of marginal distribution.

May want to use weights in regression analysis to reduce size of standard errors.
General Stratified Sampling

Sampling rule depends on both $Y$ and $X$.

Endogenous sampling, when sampling rule depends on $Y$.

Endogenous sampling frequently called choice–based sampling when endogenous variable is discrete.
Form sample data it is not possible to recover $f(y, x)$ without knowledge of weighting rule.

If:

- $\omega(y, x)$ is known;
- Sample density $g(y, x)$ is known.
- The support of $(y, x)$ is known;
- $\omega(y, x) > 0$, $\forall(y, x)$

can recover $f(y, x)$. 
Divide Fundamental sample equation by weighting function, to yield,

$$\frac{g(y, x)}{\omega(y, x)} = \frac{f(y, x)}{\int \omega(y, x) f(y, x) \, dy \, dx}$$

By hypothesis, LHS numerator and denominator are known. And since

$$\int \frac{g(y, x)}{\omega(y, x)} \, dy \, dx = \int \left( \frac{f(y, x)}{\int \omega(y, x) f(y, x) \, dy \, dx} \right) \, dy \, dx$$

$$= \frac{1}{\kappa} \Rightarrow$$

$$f(y, x) = \kappa \frac{g(y, x)}{\omega(y, x)}$$
Assumptions Not Innocuous

Critical to see that assumptions 3 and 4 are not innocuous.

The support of \((y, x)\) is known. And

\[ \omega(y, x) > 0 \forall (y, x) \] allow sample information to recover population density.

\[ \omega(y, x) = 0 \] in many applications. And it is impossible to recover \(f\) without invoking further assumptions to determine population distribution of \((Y, X)\) at values for which \(\omega = 0\). If support is not known, can not determine if \(\omega = 0\) because \(f=0\) at those points or because of the sampling plan.
Observe income $Y$ if $Y > c$. Hence, the weighting function is:

$$
\omega(y) = \begin{cases} 
1 & \text{if } Y > c, \\
0 & \text{if } Y \leq c.
\end{cases}
$$

Because $\omega = 0$ for some values of $Y$, knowledge of the sampling rule does not suffice to recover the population density.

From a random sample, can consistently estimate:
- sample distribution of $Y > c$,
- Proportion of the original sample with $Y \leq c$, $F(c)$.
- But we do not observe values of $Y$ below $c$.

Observed income is a truncated (at $c$) random variable.
Example 1: Component Probabilities

The sample of observed income is censored if the proportion of the original random sample with income below $c$ can be consistently estimated.

Let $Y_o = Y$, $Y > c$, and $Y_o = \min(Y_o) - 1$, $Y < c$.

First determine, $\Pr(\delta = 1)$,

Find $G(y_o|\delta)$ the distribution function for the sample

$$G(y_o|\delta) = \begin{cases} 
G(y|\delta = 1) = \frac{F(y_o)}{1-F(c)}, & \delta = 1 \\
G(0|\delta = 0) = 1, & \delta = 0 
\end{cases}$$
Joint Distribution

Find the Joint distribution of \((Y_o, \delta)\) as

\[
G(y, \delta) = G(y|\delta) \Pr(\delta)
\]

\[
= \left( \frac{F(y_o)}{1 - F(c)} (1 - F(c)) \right)^\delta (1 \cdot F(c))^{1-\delta}
\]

\[
= F(y_o)^\delta F(c)^{1-\delta}
\]

If \(Y \sim N(\mu, \sigma^2)\) then \(G(y, \delta)\) is a Tobit model.
The National Longitudinal Survey of Older Men was one of the first large nationally represented longitudinal studies. It’s purpose was to study the labor market activity and well-being of elderly men. (Asst. Secretary of Labor, Patrick Moynihan, lobbied for the study.)

To be included in the sample the male respondent had to be age 45 to 59 as of April 1, 1966. Thus respondents were born in 1907 to 1921. The group served to represent the civilian non-institutionalized population of men in the same age group residing in the United States at the time the sample was selected.
Age is truncated. No respondent is younger than 45 years old. \( T \) is respondent’s age at interview, \( T \geq 45 \).

Sample is truncated. As there is no information in the sample to determine what fraction of the original birth cohorts (i.e., men born in 1907 to 1921) were at risk to be included in the sample.

A regression of \( T \) on \( X \) recovers the effect of the covariates after age \( \tau \).
Survivor Function

Let $S(\tau)$ be the survivor function $\Pr(T > \tau)$. The hazard function is,

$$h(t) = -\frac{d}{dt} \ln(S(t)) = \frac{f(t)}{S(t)}$$

Hence, the density of the survival time $T$ in the population is $f$

$$f(t) = h(t)S(t)$$

$$= h(t) \exp \left( - \int_0^t h(u) \, du \right).$$

Using the representation $S(t) = \exp \left( - \int_0^t h(u) \, du \right)$. 
The sample density of the survival waiting time for NLS Older Men Cohort is

\[ g(t|T \geq \tau) = \frac{f(t)}{S(\tau)} = \frac{h(t) \exp \left( - \int_0^t h(u) \, du \right)}{\exp \left( - \int_0^\tau h(u) \, du \right)} = h(t) \exp \left( - \int_\tau^t h(u) \, du \right) \]

Add the covariate vector, \( X = x \)

\[ g(t|T \geq \tau, X = x) = h(t|X = x) \exp \left( - \int_\tau^t h(u|X = x) \, du \right). \]
Example 3: Size Biased Sampling

Example of sampling via consumer side of the market to find informal (unregulated) child care providers.

Problem: Concern without additional assumptions will identify larger providers.