STAT 332 - Probability Sampling

- Notation
- Probability Sampling Protocols
- Toy Example
- Sampling Estimators
- Course Notes Coverage: Beginning of Chapter 5.

Probability Sampling

Probability sampling: selecting units for the sample based on a probability model.

- Major advantage:

- If we have a statistical model for how we've sampled units,

In reality, our sample will (almost certainly) differ from the target population, so we will have uncertainty about the population parameter we're interested in.

Probability Sampling

Here are a few probability sampling techniques, some of which we'll cover in detail:

- Simple random sampling without replacement.
- Simple random sampling with replacement.
- Systematic sampling.
- Cluster sampling.
- Stratified sampling.
- Ratio and regression estimation.

Notation:

-
-
-
-
-

Sampling Protocol:

A **sampling protocol** or **sampling design** is the mechanism by which we choose our samples.

-
-
-
-

Design Space Example

e.g. if $N = 3$, we have $U = \{1, 2, 3\}$ and consider the set of distinct samples:

-
-
-

Design Example

Some probability sampling protocols or sampling designs based on the previous set of samples are

-
-

Step by step Example

In practice a design is described by a step-by-step procedure and induces the sampling design $P(s)$

-
-

Inclusion probabilities:

The sampling design **inclusion probabilities** will help us derive the expectation and variance of the estimators.

-
-
-

Example 1 Inclusion Probabilities

s	s_1	s_2	s_3	s_{12}	s_{13}	s_{23}	s_{123}
Sample	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$P(s)$	0	0	0	1/3	1/3	1/3	0

Can compute p_i for each unit.

What about p_{ij}?

Example 2 Inclusion Probabilities

s	s_1	s_2	s_3	s_{12}	s_{13}	s_{23}	s_{123}
Sample	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$P(s)$	$1/4$	0	0	$1/4$	$1/4$	0	$1/4$

Can compute p_i for each unit.

What about p_{ij}?

Example Inclusion Probabilities

Intuitively, we might think of phrasing a sampling protocol as "each unit has the same chance of being sampled". **This is not correct!** Consider the following two designs

s	s_1	s_2	s_3	s_{12}	s_{13}	s_{23}	s_{123}
Sample	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$P_1(s)$	0	0	0	$1/4$	$1/4$	$1/4$	$1/4$
$P_2(s)$	0	0	0	$1/3$	$1/3$	$1/3$	0

Compute p_i for each unit for design one

Compute p_i for each unit for design two

Probability Sampling

An **important** note:

- It's common to see abbreviations, especially of long phrases like 'simple random sampling without replacement'.
- However, this can lead to confusion!
- e.g. 'SRS' could be 'simple random sampling' or 'stratified random sampling'.

In **all** contexts (assignments, tests, final exam) you must only use abbreviations if they are defined (either in the question, or by you in your answer). **Never** assume 'they'll know what I mean'!

(This is a problem throughout the scientific literature: it's easy to assume other people will know/use the same abbreviations you do!)

Probability Sampling

Recall the population total, average and variance.

$$\tau = \sum_{i \in U} y_i, \quad \mu = \frac{1}{N} \sum_{i \in U} y_i \quad \text{and} \quad \sigma^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \mu)^2$$

Using sample estimates of these population quantities are

-
-
-

Note: we use $\hat{h}\hat{a}\hat{t}\hat{s}$ for sample estimates - this is an important distinction!

Probability Sampling

s	s_1	s_2	s_3	s_{12}	s_{13}	s_{23}	s_{123}
Sample	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$P(s)$	0	0	0	$1/4$	$1/4$	$1/4$	$1/4$

We can now compute the sample average and variance for each of these samples.

- e.g. for $s = s_{12}$

- e.g. for $s = s_{23}$

Estimators

Subtly different to an estimate is an **estimator**:

-
-

Using this we can construct the distribution of the estimator and compute its expected value.

Note:

Example Probability Sampling

The probability mass function for the estimator $\tilde{\mu}$ is

s	s_{12}	s_{13}	s_{23}	s_{123}
Sample	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$\overline{y}(s)$	$(y_1+y_2)/2$	$(y_1+y_3)/2$	$(y_2+y_3)/2$	$(y_1+y_2+y_3)/3$
$P(s)$	$1/4$	$1/4$	$1/4$	$1/4$

-
-

The expectation of the estimator is

Example Probability Sampling Cont.

$$
\begin{aligned}
&= E[\tilde{\mu}] \\
&= y(s_{12})\,P(S = s_{12}) + y(s_{13})\,P(S = s_{13}) + y(s_{23})\,P(S = s_{23}) + y(s_{123})\,P(S = s_{123}) \\
&= \left(\frac{y_1 + y_2}{2}\right) \times \frac{1}{4} + \left(\frac{y_1 + y_3}{2}\right) \times \frac{1}{4} + \left(\frac{y_2 + y_3}{2}\right) \times \frac{1}{4} + \left(\frac{y_1 + y_2 + y_3}{3}\right) \times \frac{1}{4} \\
&= \frac{1}{3}[y_1 + y_2 + y_3] \\
&= \mu
\end{aligned}
$$

Note:

Example Estimators

Suppose we have 6 people and want to estimate their average age.

Unit	1	2	3	4	5
Age (y)	10	20	30	40	50

The population average is $\mu = 30$.

Suppose we'll pick one person at random.

For a specific sample, the *sample estimate* of μ is denoted $\hat{\mu}$.
e.g., if we picked Unit 2, $\hat{\mu} = 20$.

Estimators

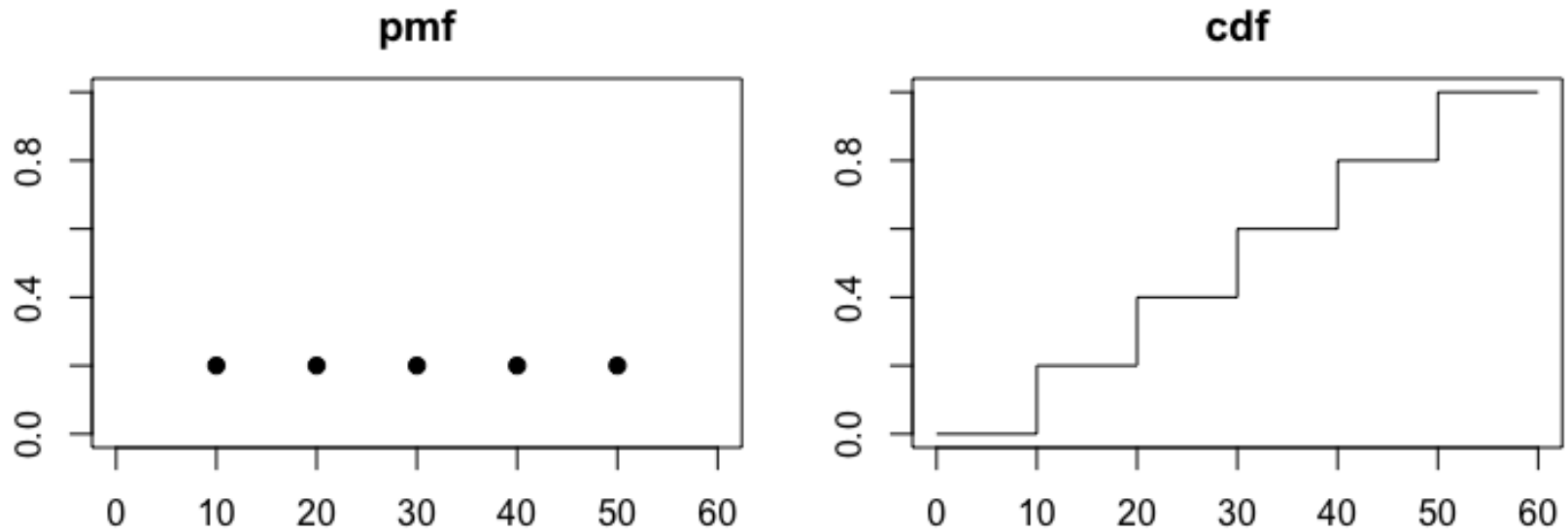
It's important to remember that μ and $\hat{\mu}$ are **fixed**. They are not random variables, so (e.g.) their variance is zero.

In contrast, the corresponding **estimator**, $\tilde{\mu}$, takes the randomness of the sampling protocol into account.

For this example

Estimators

We can examine the cumulative probability function (cdf) and probability mass function (pmf) for this estimator when we pick one person at random.



We can describe a random variable by its distribution or its expectation and variance.

Summary

-
-
-
-
-