# Introduction and Review

**Introduction to Survey Sampling**

**Math review for discrete random variables**

**Simple random sampling**

# Intro to survey sampling

## Basic Notation

$\mathcal{U} = \{ y_1, y_2, \ldots, y_N \}$ is a finite population

It contains $N$ units where $N$ is known.

$i$ is the label identifying a unit.

The $y_i$'s are the values of the characteristic of interest and are unknown.

We may select (how?) $n$ units from the population and use them to estimate the population mean.

## A simple example

$\mathcal{U}$ is a class of 45 students and $y_i$ is the year the mother of the $i$th student was born.

Let $smp$ be the labels of the $n$ students selected in the sample. Then the

sample mean $= \bar{y}_{smp} = \sum_{i \in smp} y_i / n$

could be a good guess for the

population mean $= \sum_{i=1}^{N} y_i / N$.

## Real problems are more complicated

**Target population** $= \mathcal{U} =$ population of interest.

**Sampled population** is the collection of all the possible units we could see in our sample.

Unfortunately these two populations are usually not identical.

Not every member of the target population belongs to the sampled population. (**Undercoverage**)

The sampled population contains members which do not belong to the target population.

Not all members in the intersection will be reachable and even if they are they may choose not to respond or be incapable of responding. (**Nonresponse**)

**Sampling unit** is what we can actually sample. We may be interested in individuals but only have a list of households.

**Sampling frame** is the complete list of sampling units.

**Measurement bias** is when we cannot observe $y_i$ directly but we see $y_i$ plus some random error.

**Questionnaire design** is important. How you ask the questions is important. Keep it simple and clear.

## Why sample at all?

Collecting a sample can be cheaper, quicker and even more accurate that taking a census.

The design of the sample is much more important than the size.

Important to balance Sampling vs Nonsampling error.

*Introduction to Survey Sampling* by Graham Kalton is a 1983 SAGE publication with lots of good practical information.

## Math review for Discrete random variables

### A single random variable

$X$ is a discrete rv with possible values $x_1 < x_2 < \cdots < x_K$

$\Pr\{ X = x_i \} = p(x_i)$.

$E(X) = \sum_i x_i p(x_i) = \mu$

$V(X) = \sum_i (x_i - \mu)^2 p(x_i) = \sigma^2$

Recall

$$V(X) = E(X - \mu)^2 = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2$$

If $0 < x_1$ and $\lambda > 0$ then

$$E(X) = \sum_i x_i p(x_i) = \sum_{x_i < \lambda} x_i p(x_i) + \sum_{x_i \geq \lambda} x_i p(x_i) \geq \lambda \Pr\{ X \geq \lambda \}$$

Or

$$\Pr\{ X \geq \lambda \} \leq E(X)/\lambda$$

which is *Chebyshev's Inequality*.

From this we have that for an $\epsilon > 0$

$$\Pr\{ |X - \mu| \geq \epsilon \} = \Pr\{ |X - \mu|^2 \geq \epsilon^2 \} \leq \sigma^2/\epsilon^2$$

## A pair of random variables

### The joint distribution

Consider the experiment where we randomly place 3 distinguishable balls into 3 distinguishable urns.

$X =$ the number of balls in urn 1

$Y =$ the number of occupied urns

We want to find the joint probability distribution of $X$ and $Y$, $\Pr\{ X = x, Y = y \} = p(x, y)$

| | | $Y$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | $p_1(x)$ |
| | 0 | 2/27 | 6/27 | 0 | 8/27 |
| | 1 | 0 | 6/27 | 6/27 | 12/27 |
| $X$ | 2 | 0 | 6/27 | 0 | 6/27 |
| | 3 | 1/27 | 0 | 0 | 1/27 |
| $p_2(y)$ | | 3/27 | 18/27 | 6/27 | |

$X$ and $Y$ are independent if $p(x, y) = p_1(x)p_2(y)$ for all $x$ and $y$.

Knowledge of $p(x, y)$ always yields knowledge of $p_1(x)$ and $p_2(y)$ but the converse is true only when $X$ and $Y$ are independent.

## Finding moments

$E(X) = \sum_x x p_1(x) = 0 \times 8/27 + 1 \times 12/27 + 2 \times 6/27 + 3 \times 1/27 = 1$

$E(Y) = \sum_y y p_2(y) = 1 \times 3/27 + 2 \times 18/27 + 3 \times 6/27 = 19/9$

For any function $\phi(X, Y)$ we have

$$E(\phi(X, Y)) = \sum_{(x,y)} \phi(x, y)\, p(x, y)$$

It is easy to check that

$$E(X + Y) = \sum_{(x,y)} (x + y) p(x, y) = 28/9 = E(X) + E(Y)$$

$$\begin{aligned} V(X+Y) &= E(X+Y-E(X)-E(Y))^2 \\ &= E(X-E(X))^2 + E(Y-E(Y))^2 \\ &\qquad + 2E\Big[(X-E(X))(Y-E(Y))\Big] \\ &= V(X) + V(Y) + 2Cov(X,Y) \end{aligned}$$

Note

$$Cov(X,Y) = E\Big[(X-E(X))(Y-E(Y))\Big]$$

measure how $X$ and $Y$ vary together.

If $Cov(X,Y) = 0$ then $V(X+Y) = V(X) + V(Y)$.

Note for any real numbers $a, b, c$ and $d$

$$Cov(aX + b, cY + d) = a\,c\,Cov(X, Y)$$

and

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

the correlation coefficient of $(X, Y)$ is a unit free measure of association between $X$ and $Y$.

## An important fact

If $X$ and $Y$ are independent then $E(XY) = E(X)E(Y)$. Why?

$$E(XY) = \sum_{(x,y)} x\,y\,p(x,y) \overset{\text{ind}}{=} \sum_{(x,y)} x\,y\,p_1(x)\,p_2(y)$$
$$\Big[\sum_x x\,p_1(x)\Big]\Big[\sum_y y\,p_2(y)\Big] = E(X)E(Y)$$

So if $X$ and $Y$ are independent then $V(X+Y) = V(X) + V(Y)$. Why?

$$E\Big[(X - E(X))(Y - E(Y))\Big] = E\Big[X\,Y - E(Y)X - E(X)Y + E(X)E(Y)\Big]$$
$$= E(XY) - E(X)E(Y)$$

## Simple random sampling

Suppose an urn contains $N$ distinct objects.

We choose an item from the urn at random and set it aside. From the remaining $N-1$ items we select another at random and set it aside. We continue until we have select $n < N$ items. This is simple random sampling without replacement which we denote by SRS.

If at each stage we replace the selected item before drawing again this is simple random sampling with replacement, SRSWR.

Let $A_i(j)$ be the event that the $i$th item appears on the $j$th draw.

Let $A_{i_0,i_1}(j_0, i_1)$ be the event that the $i_0$th and $i_1$ items appear on $j_0$ and $j_1$ draw.

Under SRSWR

$$\Pr\{\, A_i(j) \,\} = 1/N$$

$$\Pr\{\, A_{i_0,i_1}(j_0, i_1) \,\} = (1/N)(1/N)$$

Under SRS

$$\Pr\{ A_9(3) \} = \frac{(N-1)(N-2)1(N-3)\cdots(N-n+1)}{N(N-1)(N-2)(N-3)\cdots(N-n+1)}$$
$$= 1/N = \Pr\{ A_i(j) \}$$

$$\Pr\{ A_{4,6(2,4)} \} = \frac{(N-2)1(N-3)1(N-4)\cdots(N-n+1)}{N(N-1)(N-2)(N-3)\cdots(N-n+1)}$$
$$= (1/N)(1/(N-1)) = \Pr\{ A_{i_0,i_1}(j_0, i_1) \}$$

Let $y_i$ equal either 0 or 1. Let $D = \sum_{i=1}^{N} y_i$ be the population total. For $i = 1, 2, \ldots, n$ let

$$\begin{aligned} X_i &= 1 \text{ if } i\text{th item selected has the value 1} \\ &= 0 \text{ if } i\text{th item selected has the value 0} \end{aligned}$$

then

$$X = X_1 + X_2 + \cdots X_n$$

is the sample total.

Under SRSWR $X$ is binomial$(n, D/N)$ and

$$E(X) = n(D/N) \quad \text{and} \quad V(X) = n(D/N)(1 - D/N)$$

Under SRS $X$ is hypergeometric$(n, D, N)$ and

$$E(X) = n(D/N) \quad \text{and} \quad V(X) = n(D/N)(1-D/N)((N-n)/(N-1))$$

In both cases $E(X) = n(D/N)$ since $\Pr\{ X_i = 1 \} = D/N$.

Note under SRS we have

$$E(X_i X_j) = 1 \times \frac{D(D-1)}{N(N-1)} + 0 \times (1 - \frac{D(D-1)}{N(N-1)}) = \frac{D(D-1)}{N(N-1)}$$

Under SRS

$$
\begin{aligned}
V(X) &= V(X_1 + X_2 + \cdots + X_n) \\
&= nV(X_3) + 2\binom{n}{2} Cov(X_4, X_7) \\
&= n\frac{D}{N}(1 - \frac{D}{N}) + n(n-1)\left(\frac{D(D-1)}{N(N-1)} - \left(\frac{D}{N}\right)^2\right) \\
&= n\frac{D}{N}\left(1 - \frac{D}{N} + (n-1)\left(\frac{D-1}{N-1} - \frac{D}{N}\right)\right) \\
&= n\frac{D}{N}\left(1 - \frac{D}{N} + (n-1)\left(-\frac{N-D}{N(N-1)}\right)\right) \\
&= n\frac{D}{N}(1 - \frac{D}{N})(1 - \frac{n-1}{N-1}) \\
&= n\frac{D}{N}(1 - \frac{D}{N})\frac{N-n}{N-1}
\end{aligned}
$$