

Lecture 5: Random Sampling, Sampling Distribution

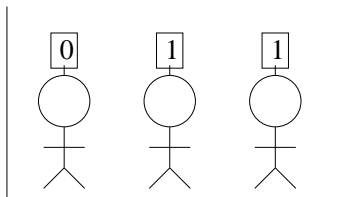
References: Nolan/Speed - Chapter 2, Pitman - Section 2.5

Sampling Survey: estimate numerical features of a population from a sample. To minimize sampling bias, it is impartial and objective to use probability methods to sample from a population.

e.g. Gallup polls before elections, telephone surveys about commercial products

Sampling is needed because of limited resources (time, money, etc).

A USEFUL WAY TO THINK ABOUT THE SAMPLING PROCESS is through the Box Model



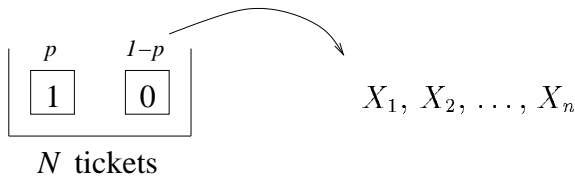
0 = Vote for Gore
 1 = Vote for Bush
 Interest: π = proportion of 1's in the box

$N = \# \text{ tickets} = \# \text{ voters}$

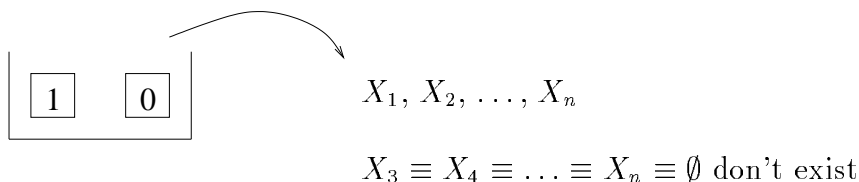
How to sample? or How to draw tickets out?

Simple Random Sampling (SRS) - the most basic ...

- SRS = sampling without replacement



$N = 2, p = 1/2$



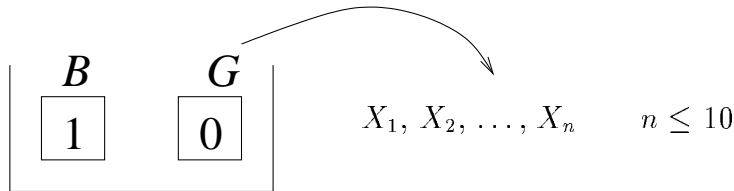
X_1, X_2 very dependent (correlation negative) and have the same marginal distribution $(1/2, 1/2)$.

		X_2
		0 1
	0	0 $1/2$
X_1	1	$1/2$ 0

$N = 1000, \pi = 1/2$ and $n = 2$. X_1, X_2 not that dependent (correlation negative, small) and have the same marginal distribution $(1/2, 1/2)$.

		X_2
		0 1
	0	$500/1000 \times 499/999$ $500/1000 \times 500/999$
X_1	1	$500/1000 \times 500/999$ $500/1000 \times 499/999$

• Product Inspection



$G = \#$ “good”,
 $B = \#$ “bad”,
 $N = \#G + \#B = 10$

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ sample mean
 $\bar{Y}_n = n\bar{X}$ sample sum

Let $n = 2, G = 8, B = 2$:

$$\text{pr}(X_1 = 1, X_2 = 0) = \text{pr}(X_1 = 1)\text{pr}(X_2 = 0|X_1 = 1) = \frac{2}{10} \times \frac{8}{9}$$

$$\text{pr}(X_1 = 1, X_2 = 1) = \frac{2}{10} \times \frac{1}{9}$$

$$\text{pr}(X_1 = 0, X_2 = 1) = \frac{8}{10} \times \frac{2}{9}$$

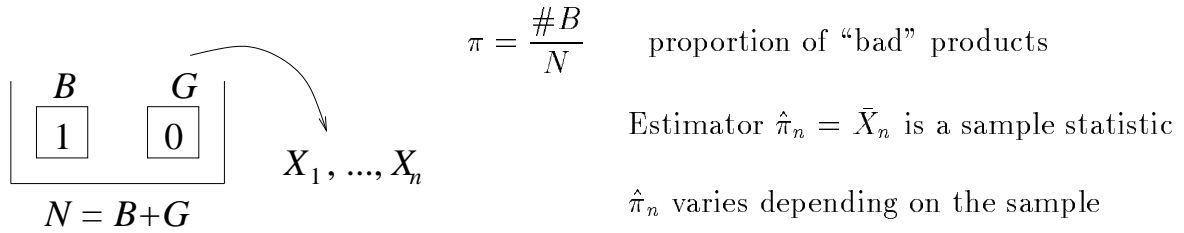
$$\text{pr}(X_1 = 0, X_2 = 0) = \frac{8}{10} \times \frac{7}{8}$$

$$\text{pr}(Y_2 = 0) = \frac{8}{10} \times \frac{7}{9}$$

$$\begin{aligned} \text{pr}(Y_2 = 1) &= \frac{2}{10} \times \frac{8}{9} + \frac{8}{10} \times \frac{2}{9} \\ \text{pr}(Y_2 = 2) &= \frac{2}{10} \times \frac{1}{9} \end{aligned}$$

Note: $Y_2 = X_1 + X_2$

Population Parameter (of interest)



What do we know about $\hat{\pi}_n$? or equivalently, \bar{X}_n or Y_n ? They are random variables coming from sampling. Their distributions are called “sampling distributions”.

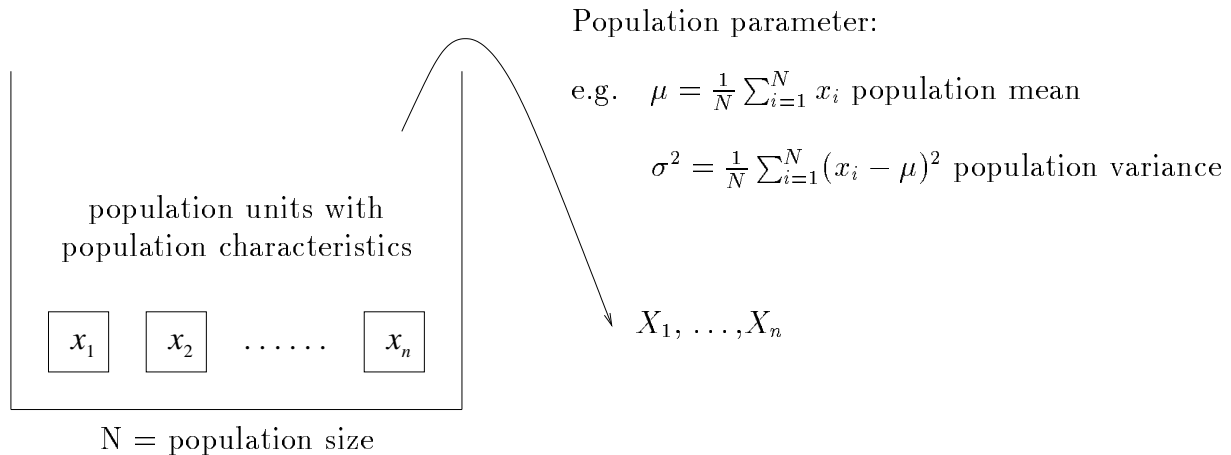
In particular,

$$\begin{aligned} E(\hat{\pi}_n) &= \pi \\ \text{Var}(\hat{\pi}_n) &= ? \end{aligned}$$

In general,

$$\begin{aligned} E(X) &= \sum x_i \text{pr}(X = x_i) \\ \text{Var}(X) &= E[X - E(X)]^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

Population: “box”



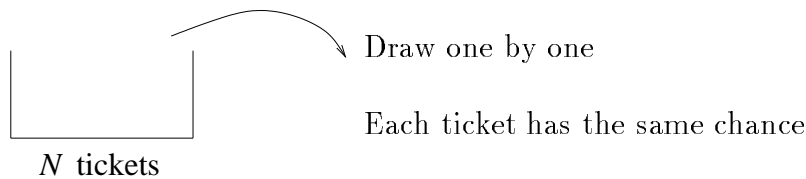
For the sample,

- sample units
- sample size n
- sample statistic: e.g. \bar{X}_n, X_n, \dots

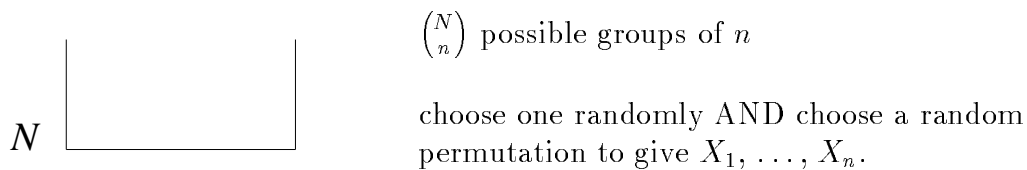
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad \text{sample variance}$$

Two equivalent ways to carry out SRS:

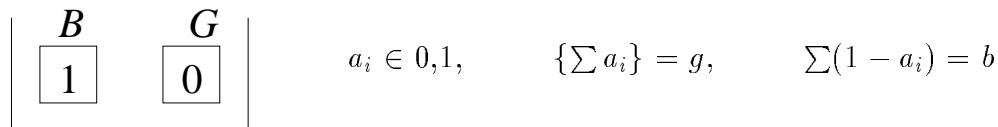
1.



2.



e.g.



1.

$$\begin{aligned} \text{pr}(X_1 = a_1, \dots, X_n = a_n) &= \frac{G \dots (G - g + 1) \times B \dots (B - b + 1)}{N \dots (N - n + 1)} \\ &= \frac{(N - n)!}{N!} \frac{G!}{(G - g)!} \frac{B!}{(B - b)!} \end{aligned}$$

2.

$$\begin{aligned} \text{pr}(X_1 = a_1, \dots, X_n = a_n) &= \frac{\binom{B}{b} \binom{G}{g}}{\binom{N}{n}} \frac{1}{\binom{n}{g}} \\ &= \frac{(N - n)!}{N!} \times \frac{G!}{(G - g)!} \times \frac{B!}{(B - b)!} \end{aligned}$$

Theorem: X_1, \dots, X_n - sampling without replacement

(i) $E(\bar{X}_n) = \mu$

(ii) $\text{Var}(\bar{X}_n) = \frac{1}{n}\sigma^2 + \frac{n-1}{n}\text{Cov}(X_1, X_2)$

Proof: Use the second way to carry out SRS that is symmetric. Then

(i) X_1, \dots, X_n have the same distribution as X_1

$$E(\bar{X}_n) = E(X_1) = \frac{1}{N} \sum_{i=1}^N x_i = \mu$$

(ii) $(X_j, X_k), j \neq k$, have the same distribution.

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \frac{1}{n^2} \sum \text{var}(X_i) + \frac{1}{n^2} \sum_{j \neq k} \text{Cov}(X_j, X_k) \\ &= \frac{1}{n}\sigma^2 + \frac{n-1}{n}\text{Cov}(X_1, X_2) \end{aligned}$$

Moreover, $\text{Cov}(X_1, X_2) = -\frac{\sigma^2}{N-1}$. Hence

$$E(\bar{X}_n) = \mu$$

$$\text{Var}(\bar{X}_n) = \frac{1}{n}\sigma^2 \frac{N-n}{N-1}$$

$$\text{SD}(\bar{X}_n) = \frac{1}{\sqrt{n}}\sigma \sqrt{\frac{N-n}{N-1}}$$

$\sqrt{\frac{N-n}{N-1}}$ is known as the **correction factor**.

