Design based inference

Week 2 (2.2)

Stat 260, St. Clair

Content:

- Small example
- Comparing estimators
- Notation

Small example

Suppose we have $N = 3$ lakes in a town. We are asked to estimate the total number residents who like around the lakes.

Sampling Design: We are going to compare three different designs where

- Population = three lakes
- Sampling unit = Observation unit = lake
- Measurement = y_i is the number of people who reside around lake i
- **Parameter** of interest = $\boxed{t = \sum_{i=1}^{3} y_i}$ total number of residents around all three lakes

??

population total

Small example

We want to estimate t using a sample of size $n = 2$ lakes. Why does it make sense to use

$$\boxed{\hat{t} = 3\bar{y}}$$

as an estimator of t where $\bar{y}$ is our average response in our sample?

$\bar{y}$ = avg. # residents per lake in sample

3 = # lakes in pop.

$3 \times \bar{y}$ = estimate of total # residents in all 3 lakes

Design-based idea

What is the **sampling distribution** of $\hat{t}$ **under a particular sampling design?**

1. List out all possible samples of n sampling units from a population of size N.

2. For each set of sampled units $\mathcal{S}$:

    - Compute the probability of each sample **based on your chosen sampling design**

    - Compute the estimator value $\hat{t}$

Design-based idea

Step (2) generates the **sampling distribution** for $\hat{t}$.

Properties:

- **Expected value**: $E(\hat{t})$ (on avg., what value of $\hat{t}$) ?
- **Bias**: $bias(\hat{t})$ systematically over/under estimating t
- **SD**: $SD(\hat{t})$ variability of $\hat{t}$ around $\boxed{E(\hat{t})}$
- **Mean Square Error**: $MSE(\hat{t})$ variability $\hat{t}$ around $\boxed{t}$

Different sampling designs will yield different sampling distributions *and* different estimator properties.

Small example

Design 1: Roll a 6-sided die twice.

- A 1 or 2 samples unit 1
- a 3 or 4 samples unit 2
- a 5 or 6 samples unit 3

Use this same scheme to sample two lakes (with replacement so repeats are possible!)

What samples are possible?

1 = lake 1 2 = lake 2 3 = lake 3

S = unorder list of lakes

$s_1 = \{1,1\}$ $s_2 = \{1,2\}$ $\{1,3\}$

$\{2,2\}$ $\{2,3\}$ $\{3,3\}$

unit #

Small example

Design 1: Roll a 6-sided die twice.

- A 1 or 2 samples unit 1 $= 2/6$
- a 3 or 4 samples unit 2 $= 2/6$
- a 5 or 6 samples unit 3 $= 2/6$

What is the probability of each sample?

$\boxed{S_1 = \{1, 1\}}$

$\underbrace{\text{1st roll}} = \text{label 1} = \text{rolled die 1/2} = \frac{2}{6}$

$\text{2nd roll} = \text{label 1} = \text{die 1/2} = 2/6$

$P(S_1) = P(\{1,1\}) = P(\text{die 1/2 and die 1/2})$

$= \frac{2}{6} \times \frac{2}{6} = \boxed{\frac{1}{9}} = P(\{2,2\})$

$= P(\{3,3\})$

$S_2 = \{1, 2\}$

$P(\{1,2\}) =$

$P\Big(\underbrace{\text{label 1}}_{\text{1st roll}}, \underbrace{\text{label 2}}_{\text{2nd}}$ OR

$\underbrace{\text{label 2}}_{\text{1st}}, \underbrace{\text{label 1}}_{\text{2nd}}\Big)$

$= \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3}$

$= \frac{1}{9} + \frac{1}{9} = \boxed{\frac{2}{9}}$

$= P(\{1,3\})$

$= P(\{2,3\})$

Small example

No.	**Sample**	**D1:** $P(\mathcal{S}_i)$	**Data**	$\bar{y}_s$	$\hat{t} = 3\bar{y}_s$
1	$\mathcal{S}_1 = \{1, 1\}$	1/9	?	?	?
2	$\mathcal{S}_2 = \{1, 2\}$	2/9	?	?	?
3	$\mathcal{S}_3 = \{1, 3\}$	2/9	?	?	?
4	$\mathcal{S}_4 = \{2, 2\}$	1/9	?	?	?
5	$\mathcal{S}_5 = \{2, 3\}$	2/9	?	?	?
6	$\mathcal{S}_6 = \{3, 3\}$	1/9	?	?	?

We can't fill in the table (with numbers) unless we know what the lake responses y_i are.

Small example

pop. total

$t = 10 + 8 + 12 = \underline{\underline{30}}$

Suppose our population looks like:

lake i	1	2	3
y_i	10	8	12

→ data in population

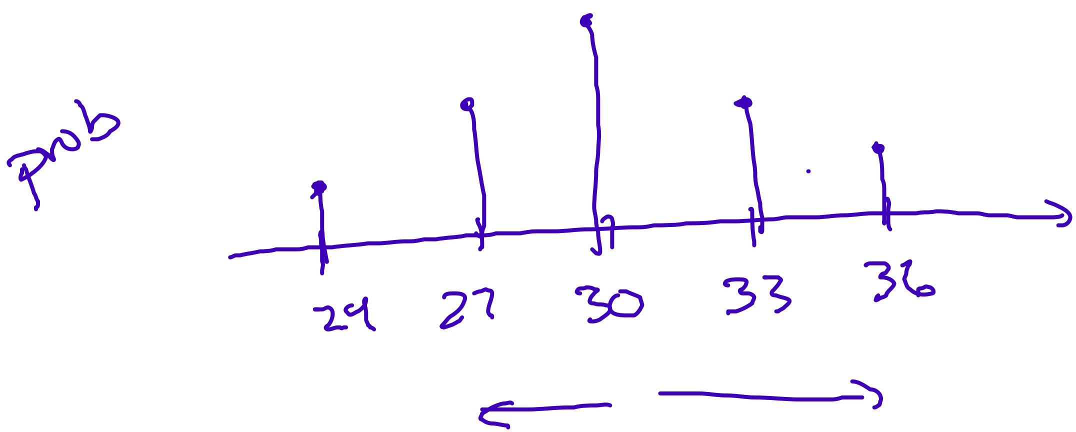
then:

No.	Sample	D1: $P(\mathcal{S}_i)$	Data	$\bar{y}_s$	$\boxed{\hat{t} = 3\bar{y}_s}$
1	$\mathcal{S}_1 = \{1,1\}$	1/9	10,10	10	30
2	$\mathcal{S}_2 = \{1,2\}$	2/9	10,8	9	27
3	$\mathcal{S}_3 = \{1,3\}$	2/9	10,12	11	33
4	$\mathcal{S}_4 = \{2,2\}$	1/9	8,8	8	24
5	$\mathcal{S}_5 = \{2,3\}$	2/9	8,12	10	30
6	$\mathcal{S}_6 = \{3,3\}$	1/9	12,12	12	36

Small example

Under **design 1**, the sampling distribution of $\hat{t}$ is the following **discrete probability model**:

$\hat{t}$	24	27	30	33	36
Prob.	$1/9$	$2/9$	$3/9$	$2/9$	$1/9$



$\hat{t} = 3\bar{y}$

Small example

Under **design 1**, the (estimator) bias of $\hat{t}$ is

- Expected value

$$\underline{E(\hat{t})} = \sum_{\text{all values } \hat{t}} \hat{t}\, P(\hat{t}) = 24\left(\frac{1}{9}\right) + 27\left(\frac{2}{9}\right) + 30\left(\frac{3}{9}\right) + \dots$$

$$= \underline{30}$$

- $\text{Bias}(\hat{t}) = E(\hat{t}) - t$

$$= 30 - 30 = \underline{\underline{0}}$$

No bias

Under design 1 $\Rightarrow$ $\hat{t} = 3\bar{y}$ is unbiased

Small example

Under **design 1**, the SD of $\hat{t}$ is

$$SD(\hat{t}) = \sqrt{Var(\hat{t})} = \sqrt{\sum_{\text{all } \hat{t}} (\hat{t} - E(\hat{t}))^2 P(\hat{t})}$$

$$= \sqrt{(24-30)^2\left(\frac{1}{9}\right) + (27-30)^2\left(\frac{2}{9}\right) + \cdots\cdots}$$

$$= \sqrt{12} \approx \boxed{3.5}$$

Small example

Under **design 1**, the MSE of $\hat{t}$ is

$$MSE(\hat{t}) = \underbrace{Var(\hat{t})}_{SD(\hat{t})^2} + [Bias(\hat{t})]^2$$

$$= 12 + 0 = \boxed{12}$$

Small example

Design 2: Toss two darts at a map of the town where

- lakes 1 and 2 are the same area and
- lake 3 is three times the area of 1 and 2.

Use this same scheme to sample two lakes (with replacement so repeats are possible!)

No.	**Sample**	**Data**	$\bar{y}_s$	$\hat{t} = 3\bar{y}_s$	**D1:** $P(\mathcal{S}_i)$	**D2:** $P(\mathcal{S}_i)$
1	$\mathcal{S}_1 = \{1, 1\}$	10,10	10	30	1/9	?
2	$\mathcal{S}_2 = \{1, 2\}$	10,8	9	27	2/9	?
3	$\mathcal{S}_3 = \{1, 3\}$	10,12	11	33	2/9	?
4	$\mathcal{S}_4 = \{2, 2\}$	8,8	8	24	1/9	?
5	$\mathcal{S}_5 = \{2, 3\}$	8,12	10	30	2/9	?
6	$\mathcal{S}_6 = \{3, 3\}$	12,12	12	36	1/9	?

Small example

Design 2: Toss two darts at a map of the town where

- lakes 1 and 2 are the same area and $\Rightarrow 1/5$
- lake 3 is three times the area of 1 and 2. $\Rightarrow 3/5$

$p = P(\text{lake } 1) = P(\text{lake } 2)$

$P(\text{lake } 3) = 3p$

$L1 \quad L2 \quad L3$

$1 = p + p + 3p$

$p = \frac{1}{5}$

Small example

Design 2: Toss two darts at a map of the town where

- lakes 1 and 2 are the same area and
- lake 3 is three times the area of 1 and 2.

No.	**Sample**	**Data**	$\bar{y}_s$	$\hat{t} = 3\bar{y}_s$	**D1:** $P(\mathcal{S}_i)$	**D2:** $P(\mathcal{S}_i)$
1	$\mathcal{S}_1 = \{1, 1\}$	10,10	10	30	1/9	1/25
2	$\mathcal{S}_2 = \{1, 2\}$	10,8	9	27	2/9	2/25
3	$\mathcal{S}_3 = \{1, 3\}$	10,12	11	33	2/9	6/25
4	$\mathcal{S}_4 = \{2, 2\}$	8,8	8	24	1/9	1/25
5	$\mathcal{S}_5 = \{2, 3\}$	8,12	10	30	2/9	6/25
6	$\mathcal{S}_6 = \{3, 3\}$	12,12	12	36	1/9	9/25

PAUSE: Write down the sampling distribution of $\hat{t}$ under design 2 and computed the expected value, bias, SD and MSE.

Small example

Design 2: Toss two darts at a map of the town where

- lakes 1 and 2 are the same area and
- lake 3 is three times the area of 1 and 2.

Sampling distribution of $\hat{t}$ under **design 2**

$\hat{t}$	**24**	**27**	**30**	**33**	**36**
Probability	1/25	2/25	7/25	6/25	9/25

- Expected value: $E(\hat{t}\,) = 32.4$
- Bias: $Bias(\hat{t}\,) = 2.4$
- SD: $SD(\hat{t}\,) \approx 3.39$
- MSE: $MSE(\hat{t}\,) \approx 17.28$

Small example

Design 3: Put three pieces of paper in a hat labeled 1-3, draw 2 pieces at random, without replacement.

No.	Sample	Data	$\bar{y}_s$	$\hat{t} = 3\bar{y}_s$	D1: $P(\mathcal{S}_i)$	D2: $P(\mathcal{S}_i)$	D3: $P(\mathcal{S}_i)$ *
1	$\mathcal{S}_1 = \{1,1\}$	10,10	10	30	1/9	1/25	? 0
2	$\mathcal{S}_2 = \{1,2\}$	10,8	9	27	2/9	2/25	? 1/3
3	$\mathcal{S}_3 = \{1,3\}$	10,12	11	33	2/9	6/25	? 1/3
4	$\mathcal{S}_4 = \{2,2\}$	8,8	8	24	1/9	1/25	? 0
5	$\mathcal{S}_5 = \{2,3\}$	8,12	10	30	2/9	6/25	? 1/3
6	$\mathcal{S}_6 = \{3,3\}$	12,12	12	36	1/9	9/25	? 0

Small example

Design 3: Put three pieces of paper in a hat labeled 1-3, draw 2 pieces at random, without replacement.

Small example

Design 3: Put three pieces of paper in a hat labeled 1-3, draw 2 pieces at random, without replacement.

No.	Sample	Data	$\bar{y}_s$	$\hat{t} = 3\bar{y}_s$	D1: $P(\mathcal{S}_i)$	D2: $P(\mathcal{S}_i)$	D3: $P(\mathcal{S}_i)$
1	$\mathcal{S}_1 = \{1, 1\}$	10,10	10	30	1/9	1/25	0
2	$\mathcal{S}_2 = \{1, 2\}$	10,8	9	27	2/9	2/25	1/3
3	$\mathcal{S}_3 = \{1, 3\}$	10,12	11	33	2/9	6/25	1/3
4	$\mathcal{S}_4 = \{2, 2\}$	8,8	8	24	1/9	1/25	0
5	$\mathcal{S}_5 = \{2, 3\}$	8,12	10	30	2/9	6/25	1/3
6	$\mathcal{S}_6 = \{3, 3\}$	12,12	12	36	1/9	9/25	0

PAUSE: Write down the sampling distribution of $\hat{t}$ under design 3 and computed the expected value, bias, SD and MSE.

Small example

Which design is "better"?

$\hat{t} = 3\bar{y}$

design	Bias	SD	MSE
1 (die)	$\boxed{0}$	3.5	12
2 (darts	2.4	3.39	17.28
3 (srs)	$\boxed{0}$	$\boxed{2.45}$	6

with replacement

Bias = 0

small