

TEKNIK SAMPLING DAN DATA WRANGLING

Stratified Random Sampling (STRA)

Keuntungan **Stratified**

Keuntungan:

- Meningkatkan efisiensi desain/presisi estimasi karakteristik populasi.

Dengan Prinsip:

1. Unit/element yang karakteristiknya hampir sama dikelompokkan dalam satu strata Unit-unit dalam strata (within stratum) --> homogen
 2. Perbedaan rata-rata karakteristik antar strata dibuat sebesar mungkin Unit-unit antar strata (between stratum) --> heterogen
-
- Masing-masing strata bisa dianggap sebagai populasi tersendiri sehingga bisa diterapkan desain sampling yang berbeda.
 - Estimasi bisa dilakukan untuk penyajian sampai level strata
 - Untuk kemudahan administratif

Estimasi Rata-Rata

Estimasi	Domain	
	Strata ke-h	Populasi
Rata-rata	$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$	$\bar{y}_{st} = \sum_{h=1}^L \frac{N_h}{N} \cdot \bar{y}_h = \sum_{h=1}^L W_h \cdot \bar{y}_h$
Varians rata-rata	$v(\bar{y}_h) = (1 - f_h) \frac{s_h^2}{n_h}$	$v(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \cdot v(\bar{y}_h)$
Standar Error	$se(\bar{y}_h) = \sqrt{v(\bar{y}_h)}$	$se(\bar{y}_{st}) = \sqrt{v(\bar{y}_{st})}$
RSE	$RSE(\bar{y}_h) = \frac{se(\bar{y}_h)}{\bar{y}_h} \times 100\%$	$RSE(\bar{y}_{st}) = \frac{se(\bar{y}_{st})}{\bar{y}_{st}} \times 100\%$
(1 - α)% Confidence interval	$\left[\bar{y}_h - Z_{\frac{\alpha}{2}} \cdot se(\bar{y}_h); \bar{y}_h + Z_{\frac{\alpha}{2}} \cdot se(\bar{y}_h) \right]$	$\left[\bar{y}_{st} - Z_{\frac{\alpha}{2}} \cdot se(\bar{y}_{st}); \bar{y}_{st} + Z_{\frac{\alpha}{2}} \cdot se(\bar{y}_{st}) \right]$

Estimasi Total

Estimasi	Domain	
	Strata ke-h	Populasi
Total	$\hat{Y}_h = N_h \cdot \bar{y}_h$	$\hat{Y}_{st} = N \cdot \bar{y}_{st} = \sum_{h=1}^L \hat{Y}_h$
Varians Total	$v(\hat{Y}_h) = N_h^2 \cdot v(\bar{y}_h)$	$v(\hat{Y}_{st}) = N^2 \cdot v(\bar{y}_{st}) = \sum_{h=1}^L v(\hat{Y}_h)$
Standar Error	$se(\hat{Y}_h) = \sqrt{v(\hat{Y}_h)}$	$se(\hat{Y}_{st}) = \sqrt{v(\hat{Y}_{st})}$
RSE	$RSE(\hat{Y}_h) = \frac{se(\hat{Y}_h)}{\hat{Y}_h} \times 100\%$	$RSE(\hat{Y}_{st}) = \frac{se(\hat{Y}_{st})}{\hat{Y}_{st}} \times 100\%$
(1 - α)% Confidence interval	$\left[\hat{Y}_h - Z_{\frac{\alpha}{2}} \cdot se(\hat{Y}_h); \hat{Y}_h + Z_{\frac{\alpha}{2}} \cdot se(\hat{Y}_h) \right]$	$\left[\hat{Y}_{st} - Z_{\frac{\alpha}{2}} \cdot se(\hat{Y}_{st}); \hat{Y}_{st} + Z_{\frac{\alpha}{2}} \cdot se(\hat{Y}_{st}) \right]$

Estimasi Proporsi

Estimasi	Domain	
	Strata ke-h	Populasi
Proporsi	$p_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{a_h}{n_h}$	$p_{st} = \sum_{h=1}^L \frac{N_h}{N} \cdot p_h = \sum_{h=1}^L W_h \cdot p_h$
Varians proporsi	$v(p_h) = (1 - f_h) \frac{p_h q_h}{n_h - 1}$	$v(p_{st}) = \sum_{h=1}^L W_h^2 \cdot v(p_h)$
Standar Error	$se(p_h) = \sqrt{v(p_h)}$	$se(p_{st}) = \sqrt{v(p_{st})}$
RSE	$RSE(p_h) = \frac{se(p_h)}{p_h} \times 100\%$	$RSE(p_{st}) = \frac{se(p_{st})}{p_{st}} \times 100\%$
(1 - α)% Confidence interval	$\left[p_h - Z_{\frac{\alpha}{2}} \cdot se(p_h); p_h + Z_{\frac{\alpha}{2}} \cdot se(p_h) \right]$	$\left[p_{st} - Z_{\frac{\alpha}{2}} \cdot se(p_{st}); p_{st} + Z_{\frac{\alpha}{2}} \cdot se(p_{st}) \right]$

Alokasi Sampel

- Alokasi Sembarang → jarang dipakai
- Alokasi Equal
- Alokasi Proporsional
- Alokasi Neyman
- Alokasi Optimum

Alokasi Equal (Sama)

- Alokasi ini sering digunakan jika varians strata (S_h^2) hampir sama.
- Jumlah sampel untuk setiap strata sama.
- Ukuran sampel untuk strata ke-h

$$n_h = \frac{n}{L}$$

- n : jumlah sampel
- n_h : jumlah sampel di strata ke-h
- L : jumlah strata
- Ukuran sampel keseluruhan:

$$n = \frac{L \sum_{h=1}^L N_h^2 \cdot S_h^2}{N^2 D^2 + \sum_{h=1}^L N_h \cdot S_h^2}$$

- Keterangan:

$$D = \frac{d}{Z_{\alpha/2}}$$

Alokasi Equal (Sama)

- y_{hi} : nilai karakteristik unit ke-i strata ke-h
- N : jumlah populasi
- N_h : jumlah populasi di strata ke-h

$$\sum_{h=1}^L N_h = N$$

- W_h : penimbang strata ke-h (stratum weight)

$$W_h = \frac{N_h}{N}$$

- n : jumlah sampel
- n_h : jumlah sampel di strata ke-h

$$\sum_{h=1}^L n_h = n$$

- f_h : fraksi sampling strata ke-h

$$f_h = \frac{n_h}{N_h}$$

Alokasi Equal (Sama)

Suatu survei dilakukan untuk mengetahui karakteristik wanita usia subur (WUS) di suatu kecamatan dengan menggunakan desa sebagai strata. Dari survei terdahulu, diperoleh rata-rata WUS beserta standar deviasinya:

No	Strata	Populasi rumah tangga (N_h)	Rata-rata jumlah WUS per rumah tangga (\bar{y}_h)	Standar deviasi (s_h)
1	Desa A	1200	1.5	0.5
2	Desa B	800	1.25	0.4
3	Desa C	600	1.2	0.32
4	Desa D	400	0.8	0.18

Jika alokasi sampel untuk survei di atas dilakukan secara **equal allocation**, berapakah **ukuran sampel (n)** dan **ukuran sampel tiap desa (n_h)**? Diketahui tingkat kepercayaan **95%** dan persentase **margin of error 5%** dari nilai rata-ratanya.

Alokasi Equal (Sama)

$$Z_{\alpha/2} = 1.96, \quad d' = 5\% = 0.05$$

No	Strata	N_h	\bar{y}_h	s_h	$W_h \cdot \bar{y}_h$	$N_h s_h^2$	$N_h^2 s_h^2$
1	Desa A	1200	1.5	0.5	0,600	300	360000
2	Desa B	800	1.25	0.4	0,333	128	102400
3	Desa C	600	1.2	0.32	0,240	61,44	36864
4	Desa D	400	0.8	0.18	0,107	12,96	5184
Jumlah		3000			1,28	502,4	504448

$$D = \frac{d}{1.96} = \frac{d' \cdot \bar{y}}{1.96} = \frac{0.05 \cdot \bar{y}}{1.96} = \frac{0.05 \cdot 1.28}{1.96} = 0.03265$$

$$n = \frac{L \sum_{h=1}^L N_h^2 \cdot S_h^2}{N^2 D^2 + \sum_{h=1}^L N_h \cdot S_h^2} = \frac{4 \cdot 504448}{3000^2 \cdot 0.03265^2 + 502.4} = 199.813 \approx 200$$

$$n_h = \frac{n}{L} = \frac{200}{4} = 50$$

Alokasi Proporsional (Sebanding)

- Alokasi ini sering digunakan jika varians strata (S_h^2) tidak berbeda signifikan antara strata yang satu dengan strata yang lainnya.
- Jumlah sampel untuk setiap strata sebanding dengan ukuran populasi di strata tsb.
- Ukuran sampel strata ke-h

$$n_h = \frac{N_h}{N} \cdot n$$

- n : jumlah sampel
- n_h : jumlah sampel di strata ke-h
- N : jumlah populasi
- N_h : jumlah populasi di strata ke-h
- Ukuran sampel keseluruhan:

$$n = \frac{N \sum_{h=1}^L N_h \cdot S_h^2}{N^2 D^2 + \sum_{h=1}^L N_h \cdot S_h^2}$$

Alokasi Proporsional (Sebanding)

- Fraksi sampling sama untuk setiap strata

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

- Dengan menggunakan alokasi proportional, akan membentuk selfweighting design (desain yang tertimbang otomatis), hal ini dibuktikan:

$$\bar{y}_{st} = \sum_{h=1}^L W_h \cdot \bar{y}_h = \sum_{h=1}^L \frac{N_h}{N} \cdot \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{1}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi}$$

$$\begin{aligned} v(\bar{y}_{st}) &= \sum_{h=1}^L W_h^2 \cdot \frac{1 - f_h}{n_h} \cdot s_h^2 \\ &= \sum_{h=1}^L \frac{N_h^2}{N^2} \cdot \frac{1 - f}{n_h} \cdot s_h^2 = \frac{1 - f}{n} \sum_{h=1}^L W_h \cdot s_h^2 \end{aligned}$$

Alokasi Proporsional (Sebanding)

Suatu survei dilakukan untuk mengetahui karakteristik wanita usia subur (WUS) di suatu kecamatan dengan menggunakan desa sebagai strata. Dari survei terdahulu, diperoleh rata-rata WUS beserta standar deviasinya:

No	Strata	Populasi rumah tangga (N_h)	Rata-rata jumlah WUS per rumah tangga (\bar{y}_h)	Standar deviasi (s_h)
1	Desa A	1200	1.5	0.5
2	Desa B	800	1.25	0.4
3	Desa C	600	1.2	0.32
4	Desa D	400	0.8	0.18

Jika alokasi sampel untuk survei di atas dilakukan secara **proporsional allocation**, berapakah **ukuran sampel (n)** dan **ukuran sampel tiap desa (n_h)**? Diketahui tingkat kepercayaan **95%** dan persentase **margin of error 5%** dari nilai rata-ratanya.

Alokasi Proporsional (Sebanding)

$$Z_{\alpha/2} = 1.96, \quad d' = 5\% = 0.05$$

No	Strata	N_h	\bar{y}_h	s_h	$W_h \cdot \bar{y}_h$	$N_h s_h^2$
1	Desa A	1200	1.5	0.5	0,600	300
2	Desa B	800	1.25	0.4	0,333	128
3	Desa C	600	1.2	0.32	0,240	61,44
4	Desa D	400	0.8	0.18	0,107	12,96
Jumlah		3000			1,28	502,4

$$D = \frac{d}{1.96} = \frac{d' \cdot \bar{y}}{1.96} = \frac{0.05 \cdot \bar{y}}{1.96} = \frac{0.05 \cdot 1.28}{1.96} = 0.03265$$

$$n = \frac{N \sum_{h=1}^L N_h \cdot S_h^2}{N^2 D^2 + \sum_{h=1}^L N_h \cdot S_h^2} = \frac{3000 \cdot 502.4}{3000^2 \cdot 0.03265^2 + 502.4} = 149.2513 \approx 150$$

$$n_1 = \frac{N_1}{N} \cdot n = \frac{1200}{3000} \cdot 150 = 60, \quad n_3 = \frac{N_3}{N} \cdot n = \frac{600}{3000} \cdot 150 = 40,$$

$$n_2 = \frac{N_2}{N} \cdot n = \frac{800}{3000} \cdot 150 = 30, \quad n_4 = \frac{N_4}{N} \cdot n = \frac{400}{3000} \cdot 150 = 20$$

Alokasi Neyman

- Jika ada variabel pendukung yang bisa digunakan untuk mengetahui nilai varians strata (S_h^2 atau s_h^2), maka alokasi Neyman akan meningkatkan presisi dari metode sampling.
- Biaya setiap strata diasumsikan sama.
- Dalam metode ini sampel dialokasikan ke dalam setiap strata agar diperoleh standar error sekecil mungkin dan dengan memperhatikan besarnya varians.
- Makin besar varians strata, maka jumlah sampel yang dialokasikan ke dalam strata tsb akan semakin besar.
- Ukuran sampel strata ke-h:

$$n_h = \frac{N_h \cdot S_h}{\sum_{h=1}^L N_h \cdot S_h} \cdot n$$

- Ukuran sampel keseluruhan:

$$n = \frac{\left(\sum_{h=1}^L N_h \cdot S_h \right)^2}{N^2 D^2 + \sum_{h=1}^L N_h \cdot S_h^2}$$

Alokasi Neyman

Suatu survei dilakukan untuk mengetahui karakteristik wanita usia subur (WUS) di suatu kecamatan dengan menggunakan desa sebagai strata. Dari survei terdahulu, diperoleh rata-rata WUS beserta standar deviasinya:

No	Strata	Populasi rumah tangga (N_h)	Rata-rata jumlah WUS per rumah tangga (\bar{y}_h)	Standar deviasi (s_h)
1	Desa A	1200	1.5	0.5
2	Desa B	800	1.25	0.4
3	Desa C	600	1.2	0.32
4	Desa D	400	0.8	0.18

Jika alokasi sampel untuk survei di atas dilakukan secara **Neyman allocation**, berapakah **ukuran sampel (n)** dan **ukuran sampel tiap desa (n_h)**? Diketahui tingkat kepercayaan **95%** dan persentase **margin of error 5%** dari nilai rata-ratanya.

Alokasi Neyman

$$Z_{\alpha/2} = 1.96, \quad d' = 5\% = 0.05$$

No	Strata	N_h	\bar{y}_h	s_h	$W_h \cdot \bar{y}_h$	$N_h s_h^2$
1	Desa A	1200	1.5	0.5	0,600	300
2	Desa B	800	1.25	0.4	0,333	128
3	Desa C	600	1.2	0.32	0,240	61,44
4	Desa D	400	0.8	0.18	0,107	12,96
Jumlah		3000			1,28	502,4

$$D = \frac{d}{1.96} = \frac{d' \cdot \bar{y}}{1.96} = \frac{0.05 \cdot \bar{y}}{1.96} = \frac{0.05 \cdot 1.28}{1.96} = 0.03265$$

$$n = \frac{\left(\sum_{h=1}^L N_h \cdot S_h \right)^2}{N^2 D^2 + \sum_{h=1}^L N_h \cdot S_h^2} = \frac{1184^2}{3000^2 \cdot 0.03265^2 + 502.4} = 138.8196 \approx 139$$

$$n_1 = \frac{N_1 S_1}{\sum_{h=1}^4 N_h S_h} \cdot n = \frac{600}{1184} \cdot 139 = 70$$

$$n_2 = \frac{N_2 S_2}{\sum_{h=1}^4 N_h S_h} \cdot n = \frac{320}{1184} \cdot 139 = 38$$

$$n_3 = \frac{N_3 S_3}{\sum_{h=1}^4 N_h S_h} \cdot n = \frac{192}{1184} \cdot 139 = 23$$

$$n_4 = \frac{N_4 S_4}{\sum_{h=1}^4 N_h S_h} \cdot n = \frac{72}{1184} \cdot 139 = 8$$

Alokasi Optimum

- Sampel yang berukuran n dialokasikan ke dalam setiap strata sedemikian rupa sehingga diperoleh varians sekecil mungkin dengan biaya yang tersedia atau meminimumkan biaya dengan varians tertentu.
- Fungsi biaya:

$$C = C_0 + \sum_{h=1}^L c_h n_h$$

- Ukuran sampel strata ke- h : [Rumus]

$$n_h = \frac{\frac{N_h \cdot S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \left(\frac{N_h \cdot S_h}{\sqrt{c_h}} \right)} \cdot n$$

- Keterangan:
 - C = Total Biaya
 - C_0 = biaya tidak dipengaruhi desain dan metode sampling
 - c_h = biaya per elemen untuk strata ke- h

Alokasi Optimum

- Ukuran sampel keseluruhan:

1. Meminimumkan biaya dengan varians tertentu

$$n = \frac{\left(\sum_{h=1}^L N_h S_h \sqrt{c_h} \right) \left(\sum_{h=1}^L \frac{N_h S_h}{\sqrt{c_h}} \right)}{N^2 D^2 + \sum_{h=1}^L N_h S_h^2}$$

2. Meminimumkan varians dengan biaya tertentu

$$n = \frac{(C - C_0) \sum_{h=1}^L \frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L N_h S_h \sqrt{c_h}}$$

Alokasi Optimum

Suatu survei dilakukan untuk mengetahui karakteristik wanita usia subur (WUS) di suatu kecamatan dengan menggunakan desa sebagai strata. Dari survei terdahulu, diperoleh rata-rata WUS beserta standar deviasinya:

No	Strata	Populasi rumah tangga (N_h)	Rata-rata jumlah WUS per rumah tangga (\bar{y}_h)	Standar deviasi (s_h)	Biaya pencacahan per rumah tangga (c_h)
1	Desa A	1200	1.5	0.5	Rp 40.000,00
2	Desa B	800	1.25	0.4	Rp 22.500,00
3	Desa C	600	1.2	0.32	Rp 22.500,00
4	Desa D	400	0.8	0.18	Rp 62.500,00

Jika alokasi sampel untuk survei di atas dilakukan secara **Neyman allocation**, tingkat kepercayaan 95%, berapakah **ukuran sampel (n)** dan **ukuran sampel tiap desa (n_h)** jika

- Persentase *margin of error* yang diinginkan adalah 5% (dari nilai rata-ratanya)
- Biaya yang disediakan 20 juta, biaya tetapnya 4 juta

Alokasi Optimum

a. Varians ditentukan, meminimumkan biaya

No	Strata	N_h	\bar{y}_h	S_h	c_h	$W_h \cdot \bar{y}_h$	$N_h S_h$	$N_h S_h^2$	$N_h S_h \sqrt{c_h}$	$N_h S_h / \sqrt{c_h}$
1	Desa A	1200	1.5	0.5	40000	0,600	600	300	120000	3,000
2	Desa B	800	1.25	0.4	22500	0,333	320	128	48000	2,133
3	Desa C	600	1.2	0.32	22500	0,240	192	61,44	28800	1,280
4	Desa D	400	0.8	0.18	62500	0,107	72	12,96	18000	0,288
Jumlah		3000				1,28	1184	502,4	214800	6,701

$$D = \frac{d}{1.96} = \frac{d' \cdot \bar{y}}{1.96} = \frac{0.05 \cdot \bar{y}}{1.96} = \frac{0.05 \cdot 1.28}{1.96} = 0.03265$$

$$n = \frac{\left(\sum_{h=1}^L N_h S_h \sqrt{c_h} \right) \left(\sum_{h=1}^L \frac{N_h S_h}{\sqrt{c_h}} \right)}{N^2 D^2 + \sum_{h=1}^L N_h S_h^2} = \frac{214800 \cdot 6.701}{3000^2 \cdot 0.03265^2 + 502.4} = 143$$

$$n_1 = \frac{\frac{N_1 S_1}{\sqrt{c_1}}}{\sum_{h=1}^4 \frac{N_h S_h}{\sqrt{c_h}}} \cdot n = \frac{3.000}{6.701} \cdot 143 = 64$$

$$n_2 = \frac{\frac{N_2 S_2}{\sqrt{c_2}}}{\sum_{h=1}^4 \frac{N_h S_h}{\sqrt{c_h}}} \cdot n = \frac{2.133}{6.701} \cdot 143 = 46$$

$$n_3 = \frac{\frac{N_3 S_3}{\sqrt{c_3}}}{\sum_{h=1}^4 \frac{N_h S_h}{\sqrt{c_h}}} \cdot n = \frac{1.280}{6.701} \cdot 143 = 27$$

$$n_4 = \frac{\frac{N_4 S_4}{\sqrt{c_4}}}{\sum_{h=1}^4 \frac{N_h S_h}{\sqrt{c_h}}} \cdot n = \frac{0.288}{6.701} \cdot 143 = 6$$

Conclusion

Ukuran sampel pada suatu strata akan **lebih besar** dari strata lainnya jika:

- **Jumlah populasi** di strata tsb lebih besar
- **Varians** strata lebih besar
- **Biaya lebih murah** pada strata yang bersangkutan

Cluster Sampling

Cluster Sampling

Cluster Sampling

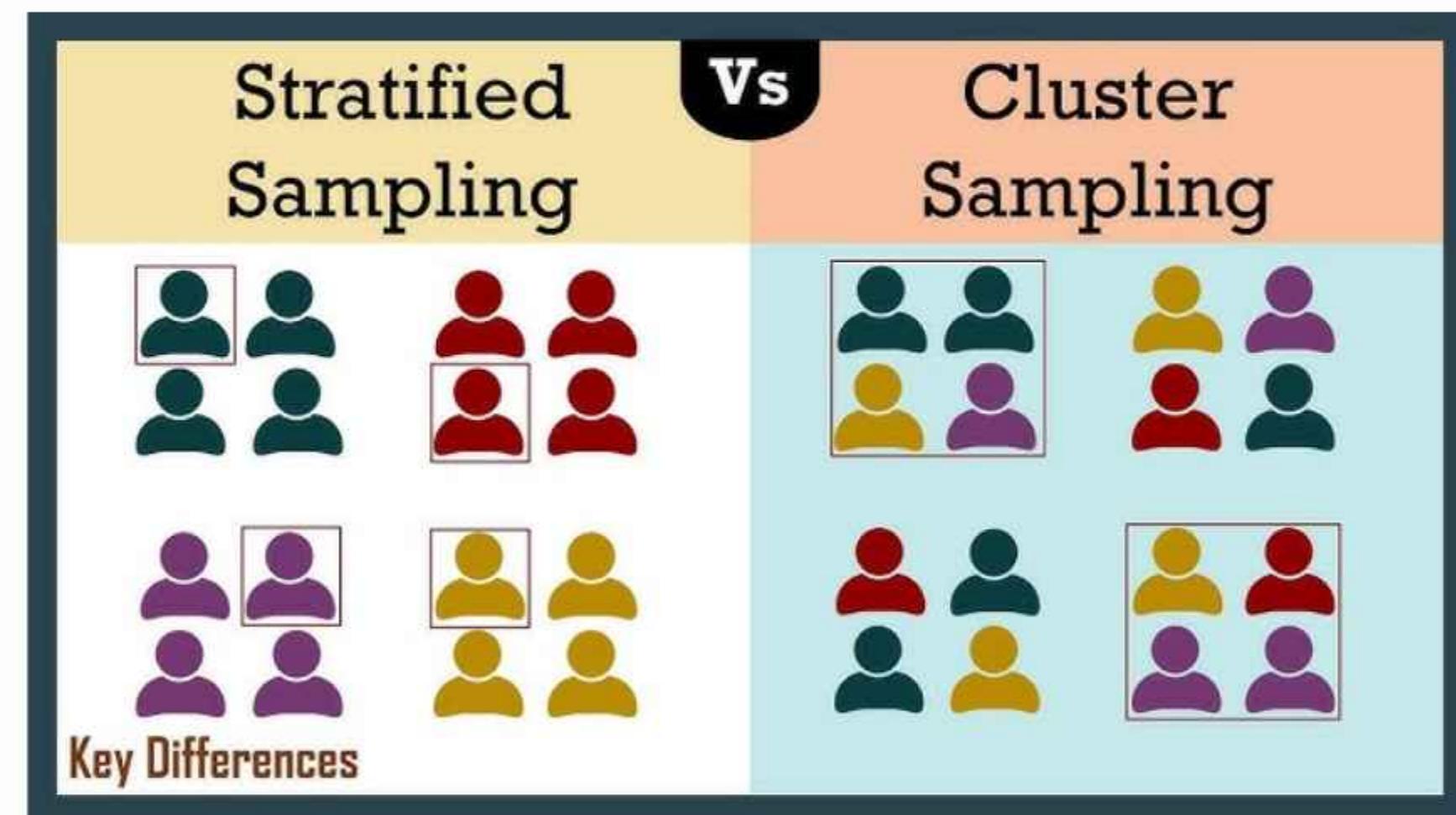
Cluster sample adalah **sampel probabilitas** (sampel acak) **di mana unit sampling** yang dipilih bukan individu langsung, melainkan **sekumpulan individu** (cluster).

Unit Sampling

Kelompok atau kelas yang berisi anggota dengan karakteristik sama

Cluster Sampling

Cluster vs Stratified



Primary dan Secondary Unit Sampling

Primary Unit Sampling

Unit pertama yang diacak

Secondary Unit Sampling

Unit yang diacak di dalam kelompok terpilih.

Cluster Sampling

Notasi pada Cluster Sampling

N = the number of clusters in the population

n = the number of clusters selected in a simple random sample

m_i = the number of elements in cluster i , $i = 1, \dots, N$

$\bar{m} = \frac{1}{n} \sum_{i=1}^n m_i$ = the average cluster size for the sample

$M = \sum_{i=1}^N m_i$ = the number of elements in the population

$\bar{M} = M/N$ = the average cluster size for the population

y_i = the total of all observations in the i th cluster

source : Sampling by Madenball (book)

Cluster Sampling

Notasi pada Cluster Sampling

- N = jumlah klaster dalam populasi.
 - Contoh: Jumlah total blok di sebuah kota.
- n = jumlah klaster yang dipilih dalam sampel acak sederhana.
 - Contoh: Jumlah blok yang dipilih untuk disurvei.
- m_i = jumlah elemen dalam klaster i .
 - Contoh: Jumlah rumah tangga di blok ke- i .
- \bar{m} = ukuran klaster rata-rata untuk sampel.
 - Ini adalah rata-rata sederhana dari ukuran-ukuran klaster yang terpilih dalam sampel. Jika Anda memilih 3 klaster ($n=3$) dengan ukuran masing-masing 10, 15, dan 20 elemen (m_1, m_2, m_3), maka ukuran klaster rata-rata sampelnya adalah $(10 + 15 + 20) / 3 = 15$.
- M = jumlah total elemen dalam populasi.
 - Contoh: Jumlah total semua rumah tangga di seluruh kota.
- \bar{M} = ukuran klaster rata-rata untuk populasi.
 - Ini adalah ukuran klaster rata-rata yang sebenarnya di seluruh populasi.
- y_i = total dari semua observasi dalam klaster ke- i yang terpilih.
 - Contoh: Total pendapatan dari semua rumah tangga di blok ke- i yang disurvei.

Cluster Sampling One Stage

Estimator untuk Rata-rata

Rata-rata sampel (ratio estimator)

$$\hat{\mu} = \frac{\sum_{j=1}^n y_j}{\sum_{j=1}^n m_j}$$

Varians rata-rata

$$\hat{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{s_r^2}{nM^2} \text{ dengan } s_r^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \hat{\mu}m_j)^2$$

Standard Error (SE)

$$SE(\hat{\mu}) = \sqrt{\hat{V}(\hat{\mu})}$$

RSE (Relative Standard Error)

$$RSE(\hat{\mu}) = \frac{SE(\hat{\mu})}{\hat{\mu}} \times 100\%$$

Confidence Interval (1- α)%

$$\hat{\mu} \pm Z_{\alpha/2} \cdot SE(\hat{\mu})$$

Cluster Sampling One Stage

Estimator untuk Total Populasi

Estimator total populasi

$$\hat{\tau} = N \cdot \bar{y}_t, \quad \bar{y}_t = \frac{1}{n} \sum_{j=1}^n y_j$$

Jika ukuran cluster rata-rata M diketahui

$$\hat{\tau} = M \cdot \sum_{j=1}^n \bar{y}_j$$

Varians total

$$\hat{V}(\hat{\tau}) = N^2 \cdot \hat{V}(\bar{y}_t)$$

Standard Error (SE)

$$SE(\hat{\tau}) = \sqrt{\hat{V}(\hat{\tau})}$$

RSE (Relative Standard Error)

$$RSE(\hat{\tau}) = \frac{SE(\hat{\tau})}{\hat{\tau}} \times 100\%$$

Confidence Interval $(1-\alpha)\%$

$$\hat{\tau} \pm Z_{\alpha/2} \cdot SE(\hat{\tau})$$

$$\hat{V}(\bar{y}_t) = \left(1 - \frac{n}{N}\right) \frac{s_c^2}{n}, \quad s_c^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y}_t)^2$$

Cluster Sampling One Stage

Estimator untuk Proporsi Populasi

Estimator proporsi

$$\hat{p} = \frac{\sum_{j=1}^n a_j}{\sum_{j=1}^n m_j}$$

Varians proporsi

$$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{s_r^2}{nM^2}$$

Definisi s_r^2

$$s_r^2 = \frac{1}{n-1} \sum_{j=1}^n (a_j - \hat{p}m_j)^2$$

Standard Error (SE)

$$SE(\hat{p}) = \sqrt{\hat{V}(\hat{p})}$$

RSE (Relative Standard Error)

$$RSE(\hat{p}) = \frac{SE(\hat{p})}{\hat{p}} \times 100\%$$

Confidence Interval $(1-\alpha)\%$

$$\hat{p} \pm Z_{\alpha/2} \cdot SE(\hat{p})$$

Cluster Sampling One Stage

Contoh 8.2 Madenhall page 279

Seorang produsen gergaji pita ingin memperkirakan biaya perbaikan rata-rata per bulan untuk gergaji yang ia jual ke beberapa industri. Ia tidak dapat memperoleh biaya perbaikan untuk setiap gergaji secara individual, tetapi ia dapat memperoleh jumlah total yang dikeluarkan untuk perbaikan gergaji dan jumlah gergaji yang dimiliki oleh setiap industri.

Karena itu, ia memutuskan untuk menggunakan cluster sampling, dengan setiap industri dianggap sebagai satu klaster. Produsen memilih sampel acak sederhana sebanyak $n = 20$ dari $N = 96$ industri yang ia layani. Data tentang total biaya perbaikan per industri dan jumlah gergaji per industri diberikan dalam tabel berikut. **Perkirakan** biaya perbaikan **rata-rata per gergaji** untuk bulan lalu dan **tentukan batas kesalahan estimasi**.

Industry	Number of saws	Total repair cost for past month (dollars)
1	3	50
2	7	110
3	11	230
4	9	140
5	2	60
6	12	280
7	14	240
8	3	45
9	5	60
10	9	230
11	8	140
12	6	130
13	3	70
14	2	50
15	1	10
16	4	60
17	12	280
18	6	150
19	5	110
20	8	120

Cluster Sampling One Stage

Jawab 8.2 Madenball page 279

anggap Number of Saw Yi dan Total Repair Mi:

1. Hitung miu hat : $2565/130 = 19.7308$
2. Hitung varians cluster (s^2) : 845.5607
3. Hitung varians rata rata : 0.792192
4. Hitung Se : 0.89005

Cluster Sampling One Stage

Contoh Lain 8.3 & 8.4 MadenHall Page 279

- 8.3 For the data in Exercise 8.2, estimate the total amount spent by the 96 industries on band saw repairs. Place a bound on the error of estimation.

Translate : Untuk data pada Latihan 8.2, estimasikan total pengeluaran dari 96 industri untuk perbaikan gergaji pita. Berikan batas galat (margin of error) dari estimasinya.

- 8.4 After checking his sales records, the manufacturer in Exercise 8.2 finds that he sold a total of 710 band saws to these industries. Using this additional information, estimate the total amount spent on saw repairs by these industries and place a bound on the error of estimation.

Translate : Setelah memeriksa catatan penjualannya, produsen di Latihan 8.2 menemukan bahwa ia telah menjual total 710 gergaji pita ke industri-industri ini. Menggunakan informasi tambahan ini, perkirakan total jumlah yang dihabiskan untuk perbaikan gergaji oleh industri-industri ini dan berikan batas kesalahan dari perkiraan tersebut.

Latihan Soal

Simple Random Sampling

1. Sebuah survei oleh dinas pendidikan dilakukan di sebuah kabupaten untuk menaksir proporsi siswa SMA yang menggunakan transportasi umum untuk berangkat ke sekolah. Kabupaten tersebut memiliki total 850 siswa SMA yang tercatat di data pokok pendidikan. Sampel acak sederhana sebanyak $n = 75$ siswa dipilih dari daftar tersebut. Setelah survei lapangan selesai, dari 75 siswa yang dijadikan sampel, ditemukan bahwa 24 siswa menggunakan transportasi umum. Perkirakan proporsi populasi sebenarnya p dan buat batasan pada kesalahan estimasi.

2. Sebuah perpustakaan universitas ingin mengestimasi total jumlah halaman dari seluruh buku fiksi ilmiah yang dimilikinya. Berdasarkan katalog, terdapat total $N = 1.200$ buku dalam kategori fiksi ilmiah. Pustakawan mengambil sampel acak sederhana sebanyak $n = 50$ buku. Dari sampel tersebut, dihitung rata-rata jumlah halaman per buku adalah $\bar{y} = 320$ halaman, dengan varians sampel $s^2 = 2.500$.

Pertanyaan:

- Buatlah taksiran untuk **total jumlah halaman (τ)** dari **seluruh** koleksi buku fiksi ilmiah di perpustakaan tersebut.
- Tentukan **batas kesalahan estimasi (Margin of Error)** untuk taksiran total tersebut.

Simple Random Sampling

3. Pihak rektorat Institut Teknologi Sepuluh Nopember dengan total $N = 8.000$ mahasiswa aktif ingin melakukan survei untuk mengetahui proporsi mahasiswa yang setuju dengan kebijakan baru mengenai jam malam di asrama. Rektorat ingin hasil taksiran memiliki batas kesalahan (B) tidak lebih dari 0.03 (3%) dengan tingkat kepercayaan 95%.

Pertanyaan:

- Jika **tidak ada informasi awal** mengenai proporsi kepuasan mahasiswa, berapa ukuran sampel minimum (n) yang harus diambil?
- Bagaimana jika survei serupa tahun lalu menunjukkan bahwa sekitar **70% mahasiswa setuju** dengan kebijakan tersebut? Berapa ukuran sampel yang dibutuhkan sekarang?

4. Departemen Sains Data Stanford University ingin menaksir rata-rata Indeks Prestasi Kumulatif (IPK) seluruh mahasiswa angkatan pertama. Total populasi mahasiswa angkatan pertama adalah $N = 2.500$ orang. Departemen ingin hasil taksiran memiliki batas kesalahan (B) tidak lebih dari 0.05 poin IPK dengan tingkat kepercayaan 95%. Karena tidak ada data varians dari tahun sebelumnya, asumsikan rentang nilai IPK adalah dari 0.0 hingga 4.0.

Pertanyaan:

- Tentukan **ukuran sampel minimum (n)** yang harus diambil untuk mencapai presisi yang diinginkan.

Stratified Random Sampling

1. Sebuah perusahaan e-commerce ingin mengestimasi rata-rata nilai transaksi (dalam ribuan Rupiah) yang dihabiskan oleh penggunanya setiap bulan. Pengguna dikelompokkan berdasarkan tingkat keanggotaan mereka: Silver, Gold, dan Platinum. Jumlah pengguna dalam setiap kelompok adalah sebagai berikut.

Strata (Tingkat Keanggotaan)	Ukuran Sub-populasi (N_h)	Ukuran Sampel (n_h)	Rata-rata / bulan (ribu Rp) (\bar{y}_h)	Deviasi Standar (ribu Rp) (s_h)
Silver	4	150	350	80
Gold	2,5	100	700	120
Platinum	1	50	1,5	250

Pertanyaan:

- Tentukan taksiran dari **rata-rata nilai transaksi per bulan** untuk **seluruh** pengguna.
- Tentukan taksiran **rata-rata nilai transaksi per bulan** untuk pengguna dengan tingkat keanggotaan **Platinum**.
- Tentukan **batas kesalahan estimasi (Margin of Error)** dari jawaban a) dan b).

Stratified Random Sampling

2. Sebuah distributor resmi produk Apple dan Samsung ingin mengestimasi total unit iPhone dan Samsung Galaxy yang terjual dalam sebulan di seluruh toko resminya. Distributor tersebut memiliki total 150 toko yang dikelompokkan menjadi tiga strata berdasarkan tipe toko: Premium Store di mal besar, Urban Store di jalan raya utama, dan Neighborhood Store di area ruko. Berikut adalah data populasi dan hasil dari sampel acak berstrata yang diambil:

Strata (Tipe Toko)	Ukuran Populasi (N_i)	Ukuran Sampel (n_i)	Rata-rata Unit Terjual (\bar{y}_i)	Varians Sampel (s_i^2)
1. Premium Store	20	5	1.500	40.000
2. Urban Store	50	10	800	14.400
3. Neighborhood Store	80	15	300	2.500
Total	$N = 150$	$n = 30$		

Pertanyaan:

- Taksirlah **total unit smartphone (τ)** yang terjual di seluruh 150 toko resmi tersebut.
- Tentukan **batas kesalahan estimasi (Margin of Error)** untuk taksiran total tersebut.

Stratified Random Sampling

3. Sebuah survei dilakukan untuk menaksir proporsi total mahasiswa ITS yang menggunakan layanan konseling psikologi dalam setahun terakhir. Populasi mahasiswa distratifikasi berdasarkan tiga rumpun ilmu: Fakultas Sains dan Analitika Data (FSAD), Fakultas Teknologi Industri dan Rekayasa Sistem (FTIRS), dan Fakultas Teknologi Elektro dan Informatika Cerdas (FTEIC). Berikut adalah data populasi dan hasil dari sampel yang diambil:

Strata (Rumpun Ilmu)	Ukuran Populasi (N_i)	Ukuran Sampel (n_i)	Jumlah Sampel yang Menggunakan Layanan (a_i)
1. FSAD	4.000	100	25
2. FTIRS	6.000	150	30
3. FTEIC	3.000	75	9
Total	$N = 13.000$	$n = 325$	

Pertanyaan:

- Taksirlah **proporsi gabungan (p_st)** mahasiswa yang menggunakan layanan konseling.
- Tentukan **batas kesalahan estimasi (Margin of Error)** untuk taksiran proporsi gabungan tersebut.

Stratified Random Sampling

4. Kota Surabaya dengan total $N = 40.000$ rumah tangga ingin menaksir proporsi rumah tangga yang memiliki akses internet fiber optik. Populasi dibagi menjadi dua strata: Area Perkotaan dan Area Pinggiran. Diketahui N_1 (Perkotaan) = 25.000 dan N_2 (Pinggiran) = 15.000. Target batas kesalahan (B) adalah 0.04 (4%) dengan tingkat kepercayaan 95%. Tidak ada informasi awal mengenai proporsi di setiap area.

Pertanyaan:

- a. Tentukan **ukuran sampel total (n)** dan alokasi sampel untuk setiap area (n_1 dan n_2) menggunakan **Alokasi Proporsional**.
- b. Tentukan **ukuran sampel total (n)** dan alokasi sampel untuk setiap area (n_1 dan n_2) menggunakan **Alokasi Sama (Equal Allocation)**.

Stratified Random Sampling

5. Dinas Pertanian kabupaten ngawi ingin melakukan survei untuk menaksir rata-rata hasil panen padi per hektar di seluruh wilayahnya. Total lahan sawah di kabupaten tersebut dibagi menjadi tiga strata berdasarkan sistem irigasi, karena sistem irigasi sangat mempengaruhi produktivitas dan variabilitas hasil panen. Dinas Pertanian ingin hasil taksiran memiliki batas kesalahan (B) tidak lebih dari 0.1 ton per hektar. Data awal dari survei pendahuluan dan data luas lahan adalah sebagai berikut:

Strata (Sistem Irigasi)	Luas Lahan (N_i) (dalam hektar)	Simpangan Baku (σ_i) (ton/hektar)	Biaya Survei per Hektar (c_i) (dalam Rp)
1. Irigasi Teknis	10.000	0.8	100.000
2. Irigasi Semi-Teknis	6.000	1.0	120.000
3. Tadah Hujan	4.000	1.5	150.000

Pertanyaan:

- Tentukan ukuran **sampel total (n)** dan alokasi sampel untuk setiap jenis irigasi (n_i) menggunakan **Alokasi Neyman** (asumsikan biaya survei sama untuk semua strata).
- Tentukan ukuran **sampel total (n)** dan alokasi sampel untuk setiap jenis irigasi (n_i) menggunakan **Alokasi Optimum** (menggunakan biaya survei yang berbeda seperti yang tertera di tabel).

Perbandingan SRS vs STRA

1. Sebuah perusahaan kecil memiliki total $N = 100$ karyawan yang terbagi dalam dua departemen (strata): Penjualan dan Produksi. Manajemen ingin menaksir rata-rata gaji bulanan karyawan. Data populasi lengkap adalah sebagai berikut:

Departemen	Ukuran Populasi (N_i)	Rata-rata Gaji (μ_i) (juta Rp)	Varians Gaji (σ_i^2)
1. Penjualan	40	12	4
2. Produksi	60	8	2.25

Diketahui rata-rata gaji keseluruhan populasi adalah $\mu = 9.6$ juta dan varians gaji keseluruhan populasi adalah $\sigma^2 = 5.44$.

Pertanyaan:

Jika akan mengambil sampel sebanyak $n = 20$ karyawan, hitung dan bandingkan varians teoritis dari taksiran rata-rata (\bar{y}) untuk kedua desain sampling berikut:

a. **Simple Random Sampling (SRS).**

b. **Stratified Random Sampling** dengan Alokasi **Proporsional**.

Ialu interpretasikan

Thank You