Estimating Prevalence Correctly

Complex Sampling in National Surveys

Dr Mohd Azmi Bin Suliman

azmi.suliman@moh.gov.my

Pusat Penyelidikan Penyakit Tak Berjangkit, Institut Kesihatan Umum

Sunday, 16 November 2025

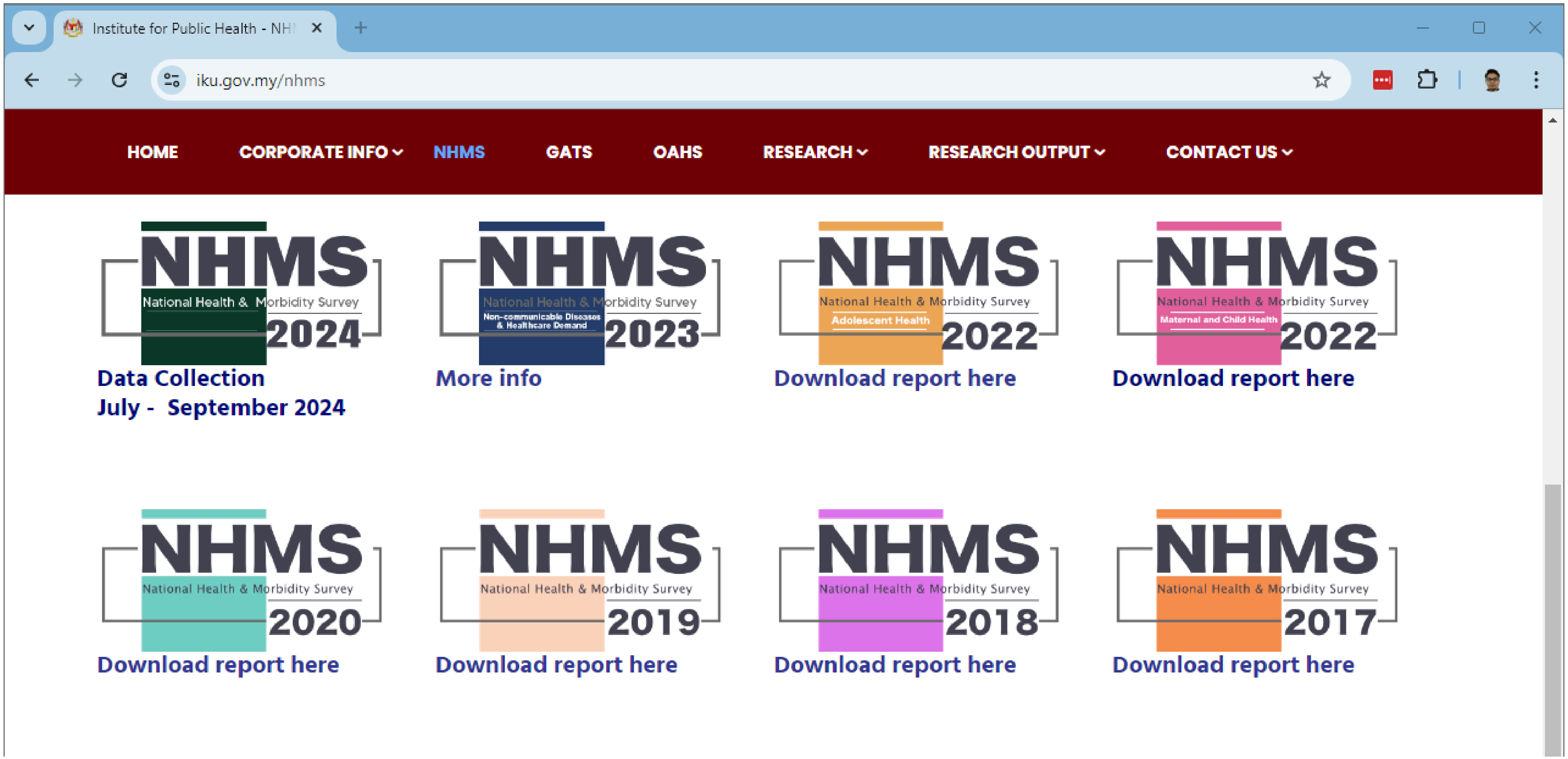
Institut Kesihatan Umum (IKU)

Who are we?

- **National Health Surveys**
- **Public Health Research**
- **Policy Support**

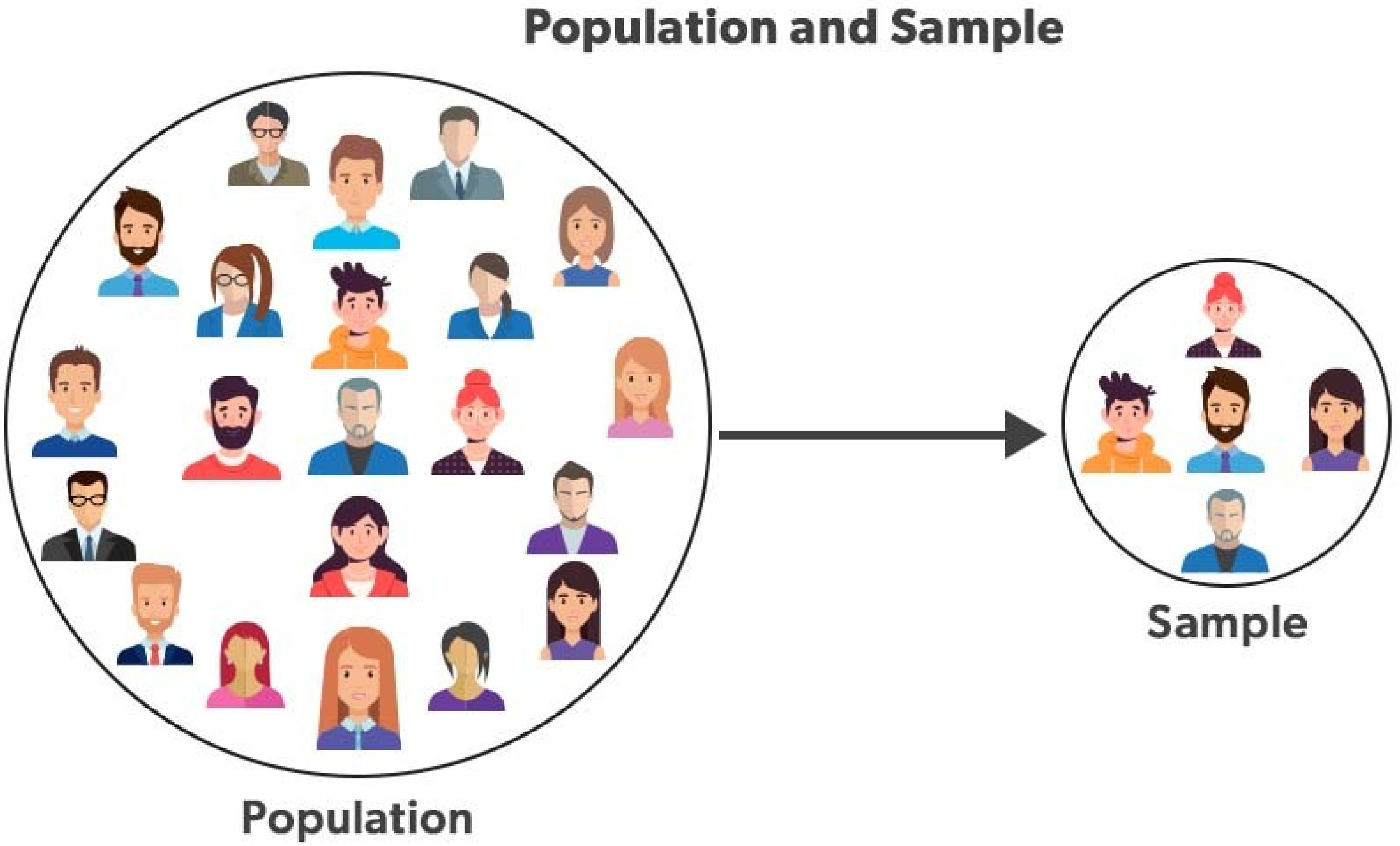

NHMS Reports

https://iku.nih.gov.my/nhms



From Population to Sample

The Sampling Problem

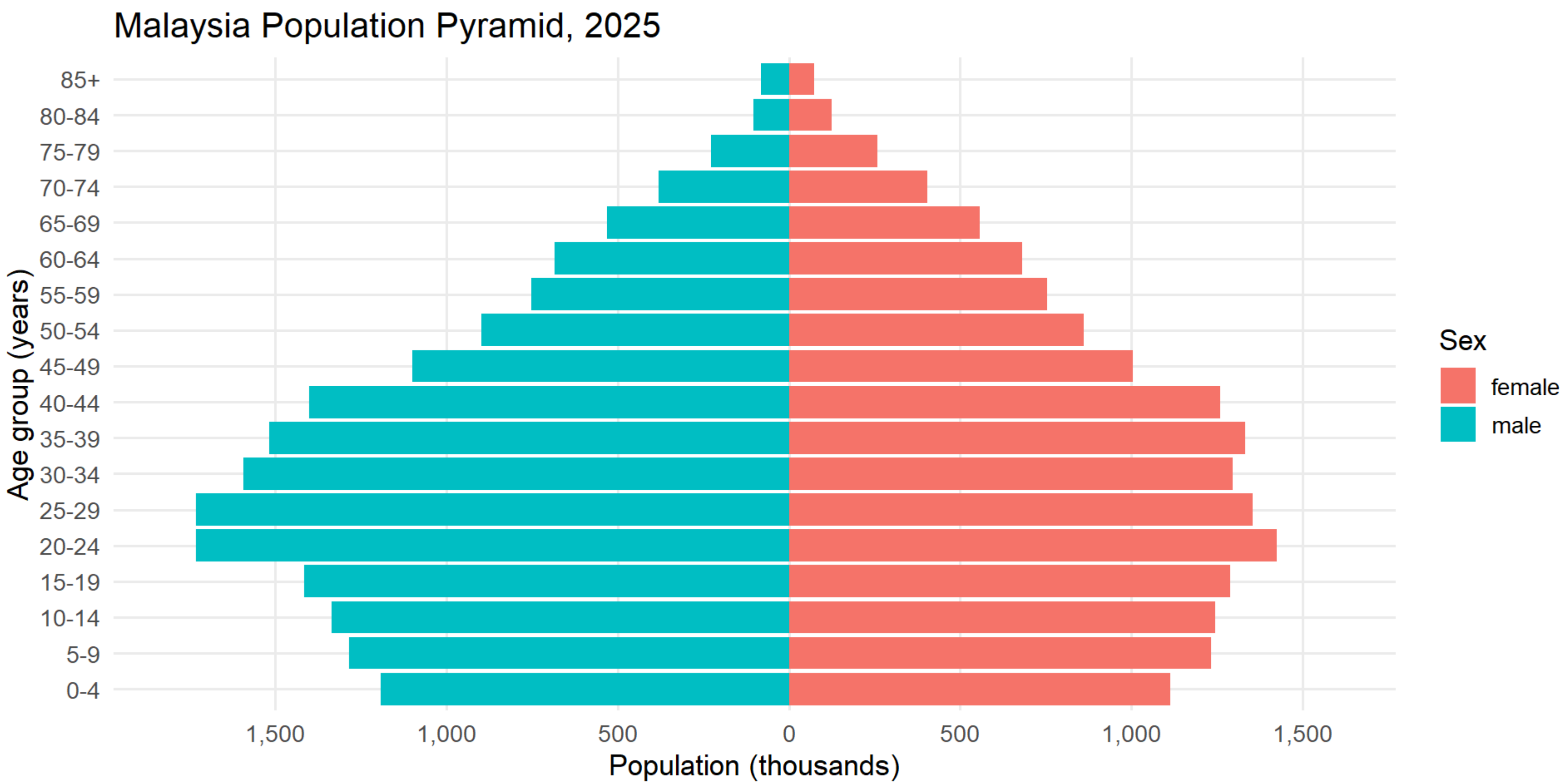


The Sampling Problem

- In describing a population, we often use a handful of **samples** rather than the **whole population**.
- Unfortunately, sample distribution may **differ** from the population - gender, ethnicity, age.
- Small studies typically limit their sample; clearly **define the target population** using **inclusive** and **exclusive criteria**.
- But national surveys, including health surveys, require the sample to **represent the general population** (e.g., adult population, older person population, maternal and child population).

Malaysian Population, 2025

- This is the Malaysian population pyramid.
- Source: Open DOSM Data Dashboard



Speaker notes

- This is Malaysia’s official population pyramid (as of 2025).
- Note the large base of working-age adults and a growing older population.

The codes

```r
pacman::p_load(tidyverse, arrow)

pyr_df <- read_parquet("https://storage.dosm.gov.my/population/population_malaysia.parquet") %>%
  filter(date == as.Date("2025-01-01"), sex %in% c("male", "female"),
         age != "overall", ethnicity == "overall") %>%
  mutate(pop_k = population, pop = if_else(sex == "male", -pop_k, pop_k),
         age0 = readr::parse_number(age), age = fct_reorder(age, age0))

my_pyr_plot <- ggplot(pyr_df, aes(x = age, y = pop, fill = sex)) +
  geom_col(width = 0.9) + coord_flip() +
  scale_y_continuous(limits = c(-2000, 2000), breaks = seq(-2000, 2000, 500),
                     labels = function(x) scales::comma(abs(x)),
                     expand = expansion(mult = c(0.02, 0.02))) +
  labs(title = "Malaysia Population Pyramid, 2025", x = "Age group (years)",
       y = "Population (thousands)", fill = "Sex") +
  theme_minimal(base_size = 13) + theme(panel.grid.minor = element_blank())

my_pyr_plot
```

Complex Sampling

Why Complex Sampling?

- **Sampling**: We use a sample to estimate the population efficiently, saving time, cost, and resources while still capturing key characteristics.
- **Stratification**: Stratifying (by gender, ethnicity) ensures all important subgroups are represented and improves precision of estimates.
- **Clustering**: Clustering respondents by area makes data collection logistically practical and cost-efficient.

Speaker notes

- Complex designs make national health surveys operationally feasible and statistically robust, balancing representativeness and cost.

What is Complex Sampling?

- **Structured selection** – Instead of simple random sampling, respondents are chosen through stratified and clustered sampling to ensure representation across diverse groups.
- **Unequal probabilities** – Some groups are oversampled (e.g., small states, older adults) to obtain reliable estimates, necessitating the use of sampling weights to correct for these differences.
- **Design-based inference** – Analysis must account for the survey's design, including strata, clusters, and weights,so that standard errors and prevalence estimates accurately reflect the true population.

Speaker notes

- Complex sampling combines stratification and clustering to achieve efficient, representative national surveys.
- Since some groups are over- or under-sampled, weighting is needed to correct their contribution to population estimates.

Example (NHMS 2023) – Diabetes Prevalence

Category	Overall %	95% CI	Male %	95% CI	Female %	95% CI
Malaysia	15.6	14.4–16.9	15.0	13.6–16.5	16.2	14.7–18.0
Age Group						
18–29	3.2	2.2–4.6	3.7	2.2–6.1	2.6	1.7–4.1
30–39	6.5	5.2–8.1	6.9	5.0–9.3	6.0	4.5–7.9
40–49	15.2	13.2–17.4	13.7	11.1–16.8	16.8	14.2–19.8
50–59	28.8	25.0–33.0	28.4	24.2–33.0	29.3	24.4–34.7
60+	38.0	35.4–40.7	37.7	34.0–41.5	38.4	35.0–41.8
Ethnicity						
Malay	16.2	15.1–17.4	15.5	14.1–17.1	16.9	15.4–18.4
Chinese	15.1	11.6–19.5	14.8	11.2–19.3	15.5	11.0–21.3
Indian	26.4	22.1–31.2	28.4	22.1–35.7	24.5	19.4–30.4
B. Sabah	9.3	7.3–11.8	9.5	6.8–13.0	9.1	6.5–12.6
B. Sarawak	17.2	13.0–22.3	14.9	10.4–21.0	19.3	14.3–25.6

Speaker notes

- Diabetes prevalence rises sharply with age.
- Diabetes also common among the Indians.
- These national estimates come from a complex survey design that accounts for stratification, clustering, and weighting.

Complex Sampling Demonstration

Simulation

- We try to mimic typical survey at field.
  - → 1,100 synthetic respondents.
  - → Age range: 18 to 100 years old
  - → Sex ratio 40 % male / 60 % female.
  - → Ethnicity: 65 % Malay, 20 % Chinese, 15 % Indian.
  - → Out of 1,100 respondents, 242 have DM
- The simulated dataset is available on GitHub: https://github.com/MohdAzmiSuliman/MyRUG_ComplexSamplingNHMS

Speaker notes

- To simplify the demonstration, a synthetic dataset was generated instead of using the original NHMS data.
- The simulated dataset is available on GitHub: https://github.com/MohdAzmiSuliman/MyRUG_ComplexSamplingNHMS
- This simulation contains 1,100 respondents, mimicking a typical survey at the field:
  - → Age distribution: 200 each for 18–29, 30–39, 40–49, 50–59, and 300 for 60+ years
  - → Sex ratio: 40% male, 60% female
  - → Ethnicity ratio: 65% Malay, 20% Chinese, 15% Indian
- Diabetes status (DM): 242 respondents (22%) simulated as having diabetes

The codes

```
tibble(age_group = c("18-29","30-39","40-49","50-59","60+"), n_total = c(200, 200, 200, 200, 300)) %>%
  mutate(male = as.integer(round(.4*n_total)), female = n_total - male) %>%
  pivot_longer(male:female, names_to = "gender", values_to = "n_gender") %>%
  mutate(malay = as.integer(round(.65*n_gender)), chinese = as.integer(round(.2*n_gender)),
         indian = n_gender - malay - chinese) %>%
  pivot_longer(malay:indian, names_to = "ethnicity", values_to = "n_ethnic") %>%
  uncount(n_ethnic) %>% select(-starts_with("n_")) %>% group_by(age_group) %>%
  mutate(age = case_when(age_group == "18-29" ~ sample(18:29, n(), replace = T),
                         age_group == "30-39" ~ sample(30:39, n(), replace = T),
                         age_group == "40-49" ~ sample(40:49, n(), replace = T),
                         age_group == "50-59" ~ sample(50:59, n(), replace = T),
                         .default = sample(60:90, n(), replace = T))) %>% ungroup() %>%
  mutate(dm = c(rep(0, 50), rep(1, 2), rep(0, 15), rep(1, 1), rep(0, 11), rep(1, 1), rep(0, 76),
              rep(1, 2), rep(0, 23), rep(1, 1), rep(0, 17), rep(1, 1), rep(0, 48), rep(1, 4), rep(0,
              rep(1, 1), rep(0, 11), rep(1, 1), rep(0, 73), rep(1, 5), rep(0, 23), rep(1, 1), rep(0,
              rep(1, 2), rep(0, 45), rep(1, 7), rep(0, 14), rep(1, 2), rep(0, 9), rep(1, 3), rep(0, 6
              rep(1, 13), rep(0, 20), rep(1, 4), rep(0, 13), rep(1, 5), rep(0, 37), rep(1, 15),
              rep(0, 12), rep(1, 4), rep(0, 6), rep(1, 6), rep(0, 55), rep(1, 23), rep(0, 18),
              rep(1, 6), rep(0, 9), rep(1, 9), rep(0, 49), rep(1, 29), rep(0, 16), rep(1, 8), rep(0,
```

Crude Proportion

- In epidemiology, prevalence refers to the proportion of a population that has a specific condition at a given time.
- Here, it reflects the proportion of individuals with diabetes mellitus (DM) in our simulated data.
- We know that 242 out of 1,100 respondents have DM — so the crude prevalence should be 22.0%, right?

Characteristic	**N = 1,100**[1]
dm	
No DM	858 (78.0%)
DM	242 (22.0%)

[1] n (%)

Speaker notes

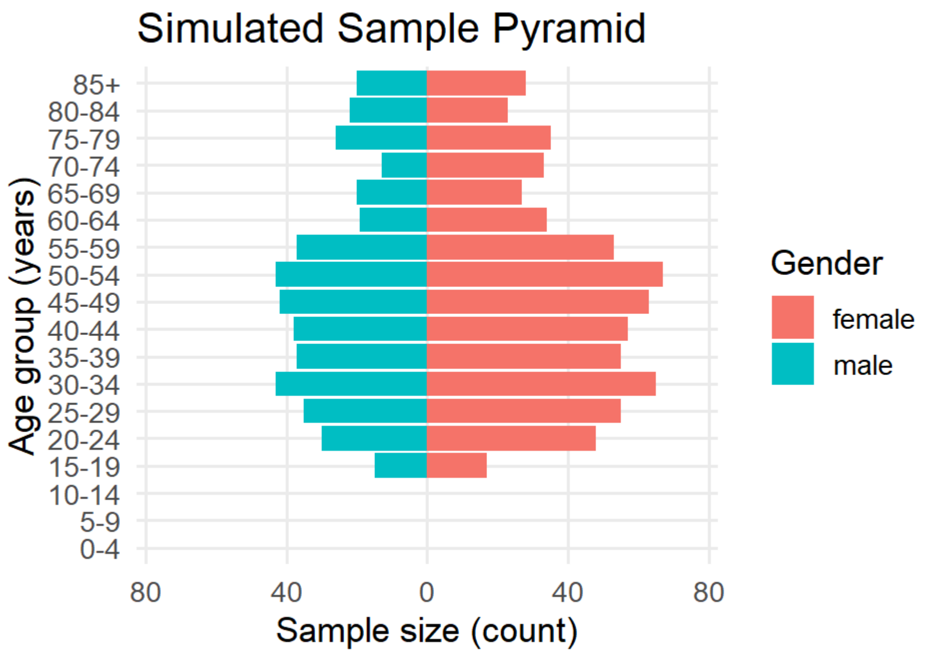
- Crude prevalence simply divides positive cases by total respondents.
- If the sample is unbalanced by age or sex, this crude figure will be biased.

Respondent vs Target Population

- But do our respondents actually reflect our target population?
- Lets compare our respondent and the Malaysian population distribution.

Respondent vs Target Population

- But do our respondents actually reflect our target population?
- These our simulated respondents.

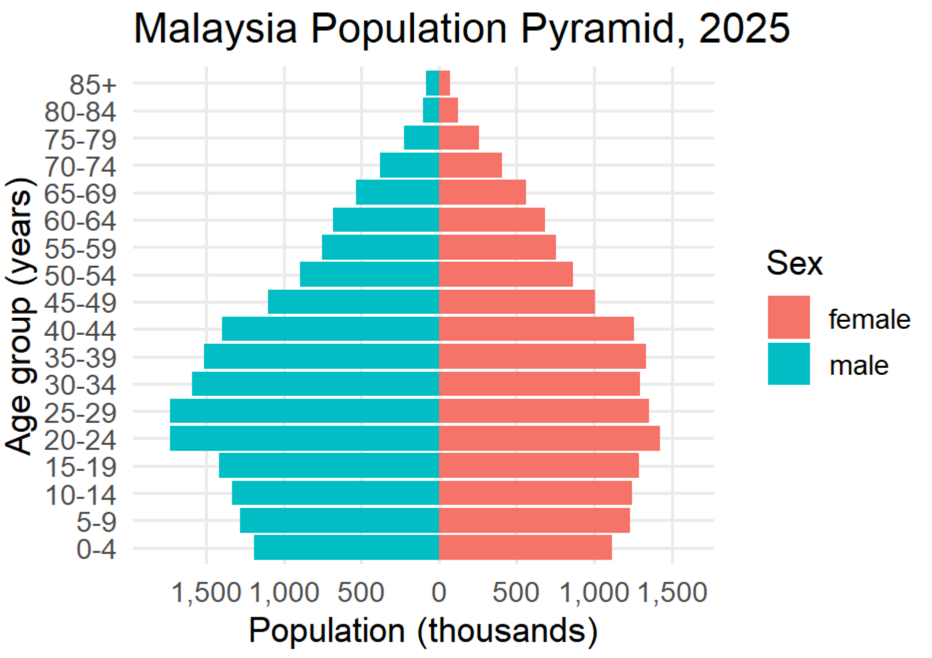


- 1,100 synthetic respondents.
- Age groups: 18–29 to 60+., 200 for each group.
- Sex ratio 40 % male / 60 % female.
- Ethnicity: 65 % Malay, 20 % Chinese, 15 % Indian.

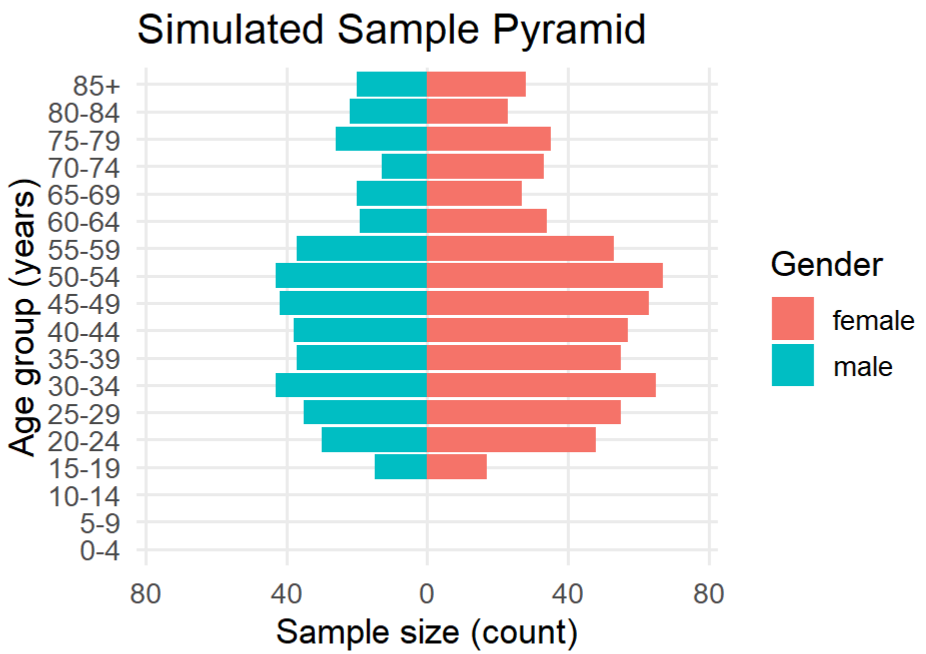
Respondent vs Target Population

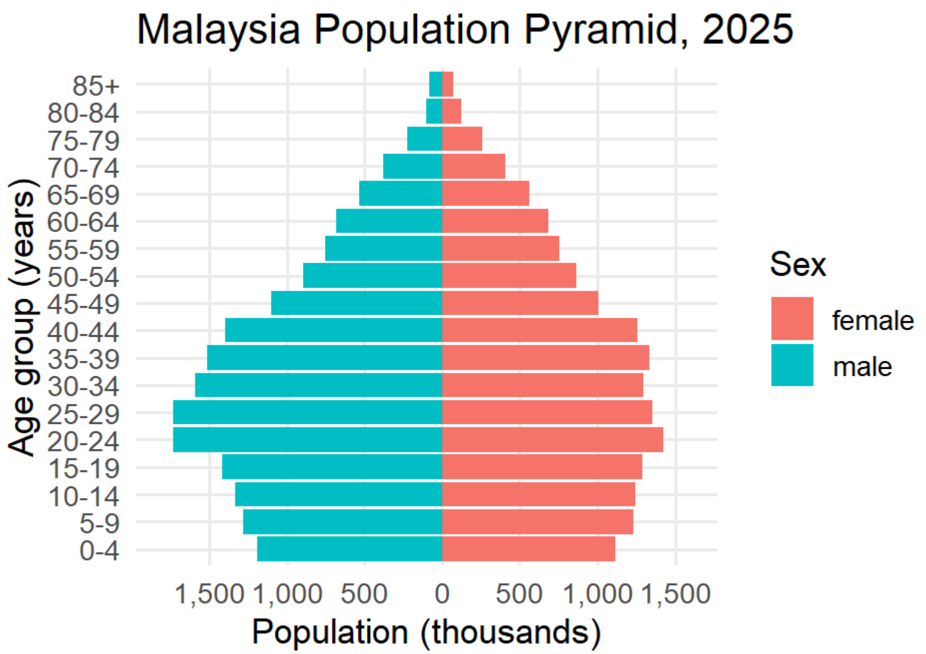
- But do our respondents actually reflect our target population?
- Let look back at our Malaysian population distribution

- Large base of working-age adults.
- Male slightly more than femal.

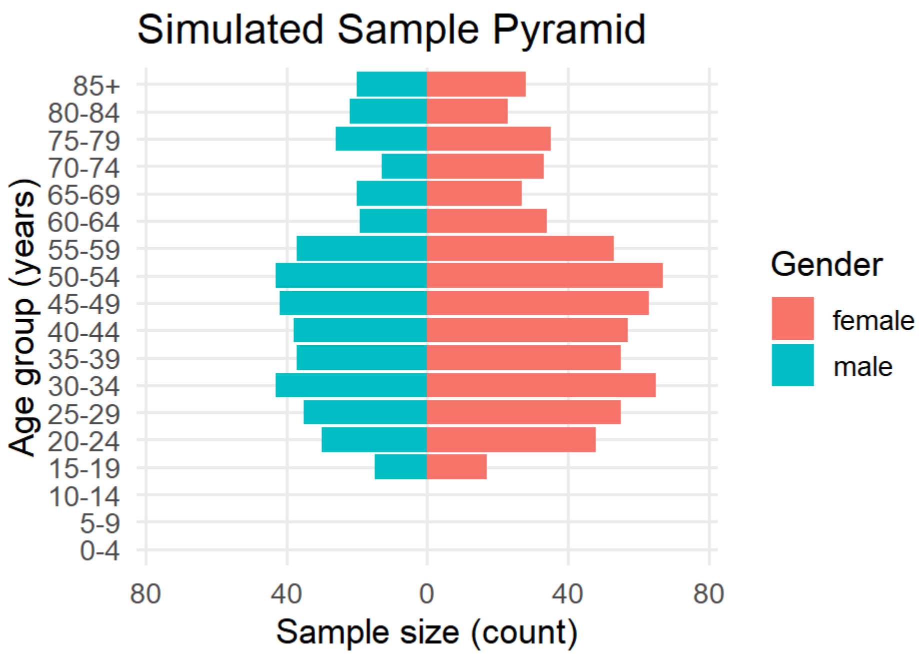


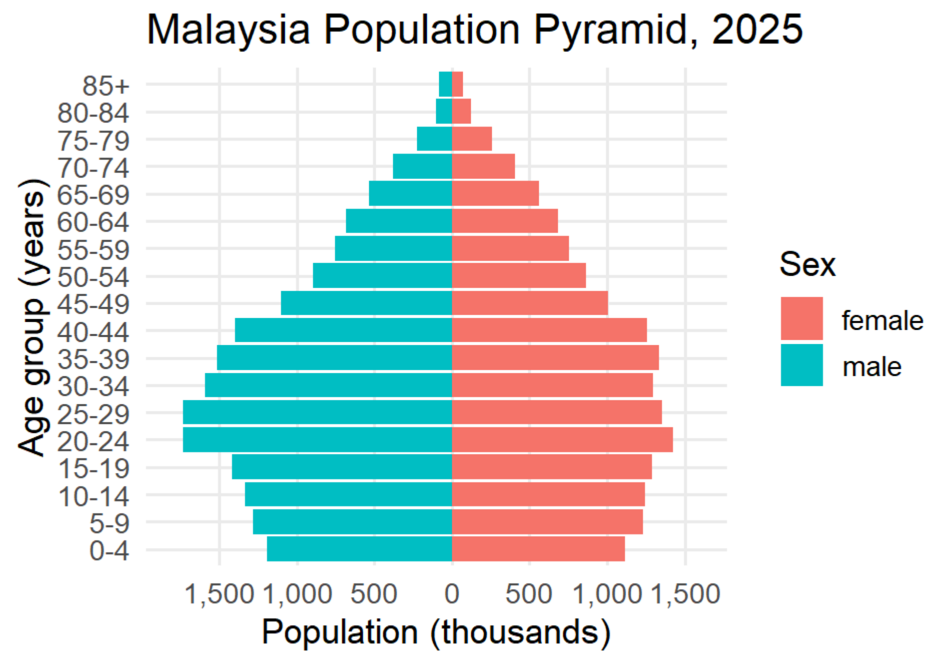
Respondent vs Target Population





The Calibration - Post-stratification





- To match sample structure to true population totals.
- Ensures estimates represent Malaysia accurately.

Speaker notes

- Because our sample's demographic structure differs from the Malaysian population, we need to adjust the sample weights so that our estimates correctly represent the national population.
- This adjustment process is called calibration in general survey methodology.
- After calibration (i.e., the post-stratification), estimates such as prevalence will better reflect the true population distribution, not just the sample composition.

The Calibration - Post-stratification

- Aligns weights by **age, sex, ethnicity**.
- Focuses on respondent count, national population by strata, and adjustment factor.

Age Group	Sex	Ethnicity	Sample Count (n)	Init. Est. Pop.	Malaysia Population ('000)	Post-strat Factor
18-29	male	malay	52	1040	1910.68	1.8371923
18-29	male	indian	12	240	201.46	0.8394167
18-29	female	malay	78	1560	1790.28	1.1476154
18-29	female	indian	18	360	188.38	0.5232778
40-49	male	malay	52	1040	1232.30	1.1849038
40-49	male	indian	12	240	161.50	0.6729167
40-49	female	malay	78	1560	1203.60	0.7715385
40-49	female	indian	18	360	155.50	0.4319444
60+	male	malay	78	1560	982.90	0.6300641
60+	male	indian	18	360	129.20	0.3588889
60+	female	malay	117	2340	1064.90	0.4550855

Post-Strat Effect – Age

- Younger adults up-weighted → weight ↑.
- Older adults down-weighted → weight ↓.

Age Group	Sex	Ethnicity	Sample Count (n)	Init. Est. Pop.	Malaysia Population ('000)	Post-strat. Factor
18-29	male	malay	52	1040	1910.68	1.8371923
30-39	male	malay	52	1040	1419.40	1.3648077
40-49	male	malay	52	1040	1232.30	1.1849038
50-59	male	malay	52	1040	814.80	0.7834615
60+	male	malay	78	1560	982.90	0.6300641
18-29	female	malay	78	1560	1790.28	1.1476154
30-39	female	malay	78	1560	1419.10	0.9096795
40-49	female	malay	78	1560	1203.60	0.7715385
50-59	female	malay	78	1560	828.40	0.5310256
60+	female	malay	117	2340	1064.90	0.4550855

Speaker notes

- The youngest age groups were under-represented in our sample, so their weights are inflated.
- This ensures their contribution matches the national age distribution.

Post-Strat Effect – Gender

- Males under-sampled → weight ↑.
- Females over-sampled → weight ↓.

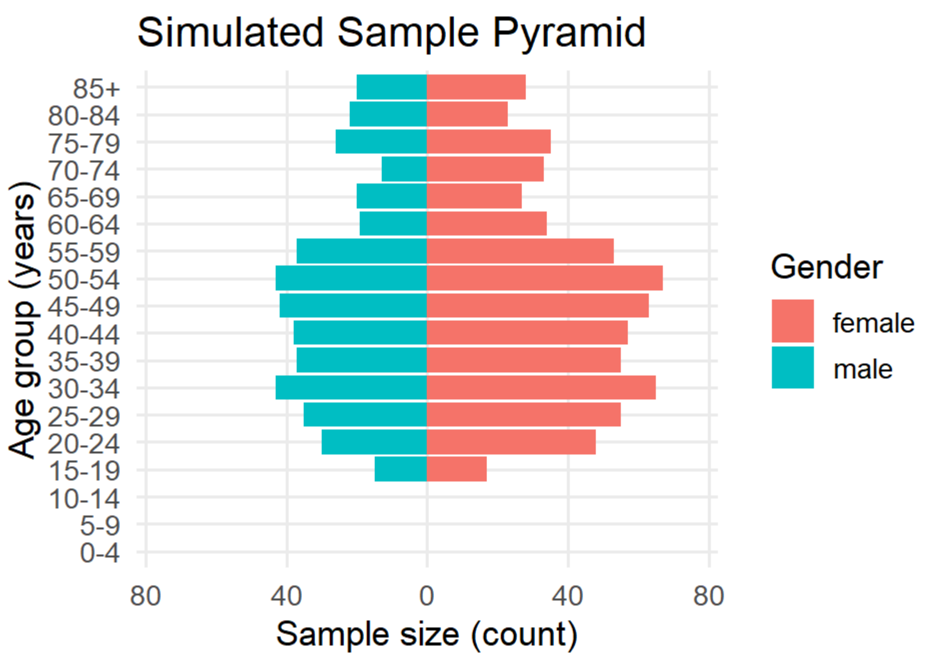
Sex	Ethnicity	Age Group	Sample Count (n)	Init. Est. Pop.	Malaysia Population ('000)	Post-strat. Factor
male	malay	18-29	52	1040	1910.68	1.8371923
female	malay	18-29	78	1560	1790.28	1.1476154
male	malay	40-49	52	1040	1232.30	1.1849038
female	malay	40-49	78	1560	1203.60	0.7715385
male	malay	60+	78	1560	982.90	0.6300641
female	malay	60+	117	2340	1064.90	0.4550855

Speaker notes

- Male respondents were fewer than expected.
- Post-stratification increases male weights and reduces female weights so that the total mirrors the real population ratio.

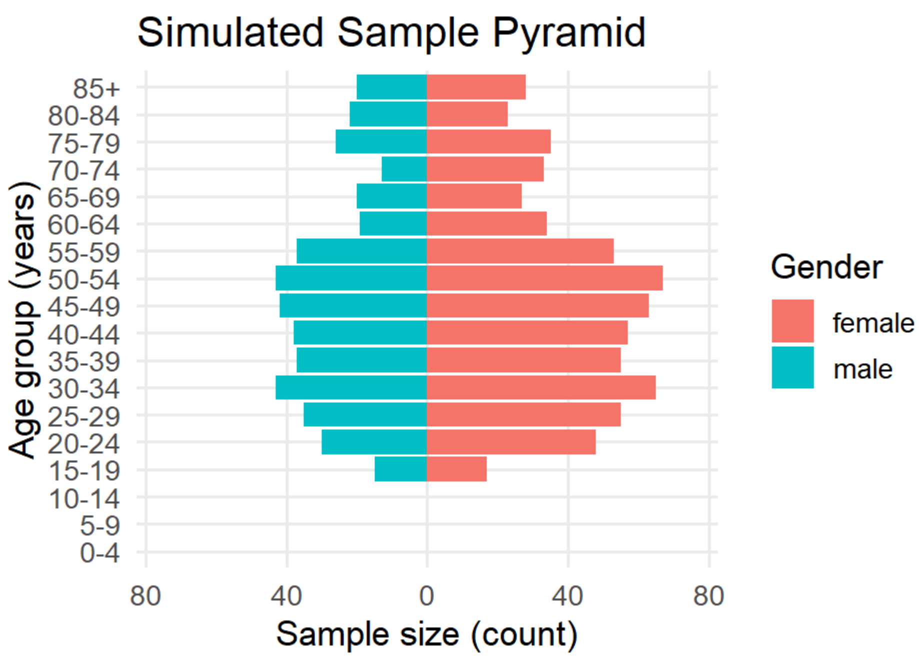
Before and After Weighting

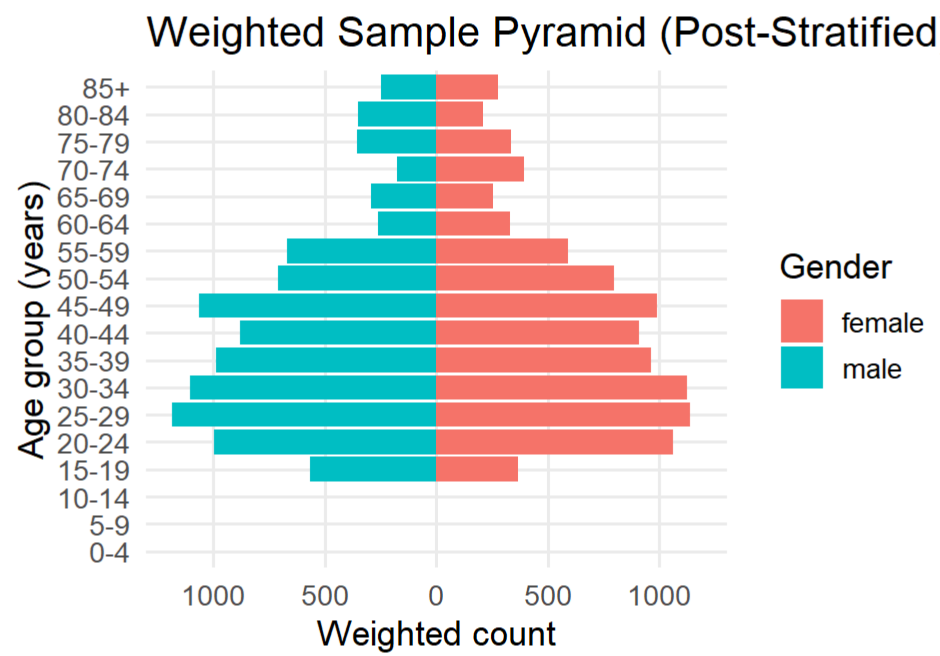
- Weighting restores population structure.



Before and After Weighting

- Weighting restores population structure.





Speaker notes

- Post-stratification weighting adjusts the sample to align with Malaysia’s actual age–sex distribution.
- The weighted pyramid (right) now mirrors Malaysia’s actual age–sex pattern, showing the success of post-stratification in correcting sample imbalance.

Corrected Prevalence

- After post-stratification, our sample now reflects Malaysia's true age–sex–ethnicity structure.
- We can now apply complex sampling analysis to obtain correct prevalence estimates.
- In R, we use the `survey` package.
- The key step is to convert the dataset into a survey design object (svydesign) by specifying:
  - → Cluster
  - → Strata
  - → Sampling weight

Speaker notes

- The idea on this slide: post-stratification fixes the mismatch between the sample and the population. Once the weights are calibrated, we can run proper survey analysis. In national surveys, we must tell R about the sampling structure—clusters, strata, and weights—because these influence both the estimate and the standard error.
- The `survey` package is the base package for design-based inference in R.
- `svydesign()` constructs the design object. After that, all analysis is done through functions like `svymean`, `svyglm`, `svyciprop`, etc.
- This step is essential to get nationally representative numbers with correct confidence intervals.

Corrected Prevalence

```r
dmsi_des <- svydesign(id = ~1, # cluster (PSU)
                      strata  = NULL, # define if applicable
                      weights = ~final_wt, # sampling weight (ADW × PSF)
                      data    = dmsi_ds_final)

svymean(~dm, design = dmsi_des) # weighted prevalence
```

```
              mean     SE
dmNo DM 0.83299 0.0112
dmDM    0.16701 0.0112
```

```r
mean(dmsi_ds_final$dm == "DM") # crude prevalence (unweighted)
```

```
[1] 0.22
```

Speaker notes

- This slide shows the practical difference between weighted and unweighted prevalence. svymean(~dm) produces the prevalence after applying the design and post-stratification weights. This is what we use for national reporting.
- The crude mean uses equal weight for every respondent, which is not valid when the sampling probabilities differ.
- This comparison helps emphasise why we cannot rely on raw proportions from complex survey data.

Corrected Prevalence

Comparison of Crude and Weighted Estimates of Diabetes Prevalence

Characteristic	Crude (Unweighted) DM N = 242[1]	Weighted (Post-Stratified) DM N = 3,269[1]
my		
Overall	242 (22.0%)	3,269 (16.7%)
gender		
female	147 (22.3%)	1,672 (17.2%)
male	95 (21.6%)	1,597 (16.2%)
age_group		
18-29	8 (4.0%)	211 (4.0%)
30-39	14 (7.0%)	284 (6.8%)
40-49	34 (17.0%)	609 (15.8%)
50-59	63 (31.5%)	816 (29.5%)
60+	123 (41.0%)	1,349 (38.8%)

[1] n (%)

Speaker notes

- This table highlights the key message of the session: complex sampling adjusts both the estimate and the precision.
- Crude proportions ignore true sampling probabilities and population structure, often leading to biased estimates.
- Weighted, design-based estimates correct these issues and give results that can be generalised to the national population. The table also demonstrates how different age–sex patterns shift once weights are applied.

Caveats in Complex Sampling

- Needs a **known sampling frame**.
- Requires larger sample to offset design effect.
- **Intra-cluster correlation** reduces precision.
- Standard tests without weights → invalid results.

Speaker notes

- Complex sampling assumes known selection probabilities (e.g., list of houses from DOSM).
- Design effects increase variance, so larger samples are needed.
- Ignoring weights leads to underestimated standard errors and misleading confidence intervals.

Summary

- Complex sampling improves representativeness.
- Weighting corrects for unequal selection.
- Post-stratification aligns sample to population.
- Corrected estimates are valid and comparable nationally.

Speaker notes

- In national surveys like NHMS, complex sampling and calibration ensure data accurately represent Malaysia’s population.
- Understanding weighting and design effects is key to valid public health inference.