Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

About the dataset:
Context:
According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
Each row in the data provides relavant information about the patient.

Relevant Attribute Information:

1) id: unique identifier
2) age: age of the patient
3) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

4) hypertension: 0 if the patient doesn't have any hypertension, 1 if the patient has a hypertension
5) avg_glucose_level: average glucose level in blood
6) bmi: body mass index
7) smoking_status: "formerly smoked", "never smoked", "smokes"
8) stroke: 1 if the patient had a stroke or 0 if not

Objectives:

To develop and evaluate efficient sampling strategies for estimating stroke risk factors in diverse populations, specifically focusing on: a) The relationship between BMI and glucose levels across different age groups b) The association between smoking status and hypertension prevalence while comparing the statistical efficiency of various estimation methods to inform future large-scale health screening protocols.

Why Is This Meaningful?

Health Significance:

- Stroke disease is the leading cause of death globally
- Early detection of risk factors (BMI, hypertension) is crucial for prevention
- Understanding age-specific BMI patterns can inform targeted interventions strategies
- The relationship between smoking and hypertension affects screening strategies

Methodological Value:

- Compares efficiency of different sampling designs in health surveillance
- Tests whether auxiliary information (glucose levels) improves estimation accuracy
- Evaluates cost-effectiveness of stratification strategies
- Provides guidance for optimal resource allocation in health surveys

Methods:

Estimate the Mean BMI using both a regression and a ratio estimator in SRS, with glucose level serving as the helper variable in the regression estimator and the binarized version of glucose (high/low glucose) – proportion of high glucose level serving as the auxiliary variable in the ratio estimate. For stratified sampling, stratify by age groups.

Estimate mean BMI:

- SRS:
    - Regression estimator: using average glucose level as helper var
    - Ratio estimator: using average glucose level as auxiliary var
    - Ratio estimator: using binarized version of glucose level as auxiliary var
- Stratified Sampling: stratify by age groups [10 years/level, merge 0-10 and 10-19 because of small sample size in 0-10 group (2 in sample, 9 in population), also align with literature].
    - Regression estimator: using average glucose level as helper var
    - Ratio estimator: using average glucose level as auxiliary var
    - Ratio estimator: using binarized version of glucose level as auxiliary var

References:

- Zierle-Ghosh A, Jan A. Physiology, Body Mass Index. [Updated 2023 Nov 5]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535456/
- Mathew TK, Zubair M, Tadi P. Blood Glucose Monitoring. [Updated 2023 Apr 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK555976/
- Alexander, C. M. "The influence of age and body mass index on the metabolic syndrome and its components." *Diabetes, obesity & metabolism*, vol. 10, no. 3, 03/2008, pp. 246-250, , doi:10.1111/j.1463-1326.2006.00695.x.

Estimate the Proportion of Hypertension using SRS with both vanilla estimate and a ratio estimator, again using glucose level as the auxiliary variable. For stratified sampling, smoking status (“formerly smoked,” “never smoked,” “smokes”) will be used as the stratifying variable. (FYI: regression estimation for binary variable is omitted because the model with only one helper and binary variable is really unstable)

- SRS:
    - Vanilla estimator
    - Ratio estimator: using average glucose level as auxiliary var
    - Ratio estimator: using binarized version of glucose level as auxiliary var
- Stratified Sampling: stratify by smoking status.
    - Vanilla estimator
    - Ratio estimator: using average glucose level as auxiliary var

- Ratio estimator: using binarized version of glucose level as auxiliary var

References:

- Virdis A, Giannarelli C, Neves MF, Taddei S, Ghiadoni L. Cigarette smoking and hypertension. Curr Pharm Des. 2010;16(23):2518-25. doi: 10.2174/138161210792062920. PMID: 20550499.
- Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nat Rev Nephrol. 2020 Apr;16(4):223-237. doi: 10.1038/s41581-019-0244-2. Epub 2020 Feb 5. PMID: 32024986; PMCID: PMC7998524.
- Mathur RK. Role of diabetes, hypertension, and cigarette smoking on atherosclerosis. J Cardiovasc Dis Res. 2010 Apr;1(2):64-8. doi: 10.4103/0975-3583.64436. PMID: 20877688; PMCID: PMC2945206.

Results:

Mean of BMI estimates:

True Population Mean BMI: 30.29

SRS:

1. Regression Estimation Results:

Estimated Mean BMI: 30.571

Standard Error: 0.269

95% CI: [30.043 , 31.098]

2. Ratio Estimator with Continuous Glucose Level:

Estimate: 30.462

Standard Error: 0.522

95% CI: [29.439 , 31.485]

3. Ratio Estimator with Binary Glucose Level:

Estimate: 31.403

Standard Error: 1.349

95% CI: [28.759 , 34.047]

Results by Stratum, the estimates are quite different in each age span, which aligns with the literature:

stratum	N_h	n_h	regression	ratio_cont	ratio_bin
0-19	229	44	26.249	28.084	30.903
20-29	422	81	29.310	28.257	22.137
30-39	500	96	30.733	31.217	29.904
40-49	560	108	30.260	30.314	29.909

50-59	660	127	31.311	30.545	29.496
60-69	485	94	30.863	32.416	37.970
70-79	420	81	29.570	30.606	31.109
80+	149	29	26.160	25.768	24.493

Overall estimates with stratified sampling:

	Method	Estimate	SE	CI_Lower	CI_Upper
1	Regression	29.969	0.239	29.500	30.437
2	Ratio (Continuous)	30.223	0.473	29.297	31.150
3	Ratio (Binary)	29.991	1.324	27.395	32.586

Proportion of Hypertension estimates:
True Population Proportion: 0.1191

SRS:
1. Vanilla Estimator:
Estimated Proportion: 0.1173
Standard Error: 0.0049
95% CI: [0.1077 , 0.127]

2. Ratio Estimator (Continuous Glucose):
Estimated Proportion: 0.1176
Standard Error: 0.0049
95% CI: [0.108 , 0.1271]

3. Ratio Estimator (Binary Glucose):
Estimated Proportion: 0.1159
Standard Error: 0.0052
95% CI: [0.1057 , 0.1261]

Results by stratum, which aligns with what I saw in the literature (smoking people having the most proportion of getting hypertension).

	stratum	N_h	n_h	vanilla	ratio_cont	ratio_bin
1	formerly smoked	836	464	0.1185	0.1186	0.1206
2	never smoked	1852	1028	0.1177	0.1171	0.1158
3	smokes	737	409	0.1296	0.1298	0.1304

Overall estimates with stratified sampling:

	Method	Estimate	SE	CI_Lower	CI_Upper
1	Vanilla	0.1205	0.0050	0.1107	0.1302
2	Ratio (Continuous)	0.1202	0.0049	0.1106	0.1298

3	Ratio (Binary)	0.1201	0.0053	0.1098	0.1305