The need for people of any age to be covered by some health insurance is becoming increasingly popular in today's world.

Health insurance is a way of paying for one's medical bills and health care costs however, depending on the type of health insurance plan an individual is signed up for, there may be excess charges that they would need to pay for - out of their own pockets - at the end of any given year.

Sampling Survey Analysis on the Health Insurance Costs of US citizens

Done by:

- Yasthika Ramgobin (214517334)
- Siphesihle Khumalo (219002362)
- Ayanda Ndlovu (217001859)

Aim of Study:

Our aim is to estimate those average annual health insurance costs (excess charges) of a beneficiary residing in the US, as well as to estimate the proportion of beneficiaries whose health insurance costs amount to more than $10,000 in a particular year, using four different sampling techniques to determine which sampling method gives us the most accurate results.

Discussion of the dataset and Variables of interest:

There are numerous factors that may influence the health insurance costs (charges) of US citizens, some of which include: the beneficiary's age, their body mass index (BMI), the number of dependents covered by their health insurance plan, whether they are a smoker or not, and their location (region).

In preparing our chosen dataset for analysis, we found that it contained one duplicate observation which was removed, resulting in a total population size of 1337 individuals / healthcare insurance beneficiaries.

Our main variable of interest in this study was charges, which represents the amount on medical costs (in US dollars) that the health insurance does not cover and that the primary beneficiary is required to pay at the end of the year. Another variable of interest was smoker, which indicates whether the beneficiary is a smoker or not. It is well-known that smoking is the main cause of lung cancer and can lead to many other health problems such as heart disease and stroke. Hence, we might expect beneficiaries who smoke to have higher charges than those who do not. Region, which represents the area in the US in which the beneficiary resides, was also a variable of interest in this study as we would anticipate each of the given areas to comprise of some people whose medical charges are relatively low and others whose medical charges are higher, thereby reflecting the diversity of the population of interest.

Sampling types:

We used the following probability sampling methods to obtain results for our estimates of the average charges and the proportion of individuals whose charges are higher than $10,000:

- *Simple Random Sampling Without Replacement (SRS) and Simple Random Sampling With Replacement (SRS WR):*
  For SRS, a sample of 225 individuals was randomly selected from the population of 1337 individuals *without* replacement. The sample size of 225 was calculated by specifying a margin of error of no more than $1,500 and then conducting SRS on a small (pilot) sample to obtain its variance. For SRS WR, a sample of 225 individuals was randomly selected from the population of 1337 individuals *with* replacement, which differs from SRS in that there could have been repeated observations in the chosen sample.
- *Stratified SRS:* The variable smoker (categorised as either yes or no) was used as the stratification variable, hence the population of 1337 was divided into two groups (strata), i.e., non-smoker (1063) and smoker (274). Thereafter, SRS was conducted independently within each stratum, which resulted in 179 non-smokers and 46 smokers being randomly selected for the sample, giving us a total sample size of 225.

- *Cluster SRS:* The variable region (categorised as either northeast, southeast, southwest or northwest) was used as the cluster variable, hence the population of 1337 was divided into four groups (clusters), where each cluster was not of equal size. Thereafter, SRS was conducted and randomly selected 2 out of the 4 clusters to form the sample, giving us a total sample size of 688.

Results:

	Sampling Type:			
	SRS	**SRS WR**	**Stratified SRS**	**Cluster SRS**
Estimated Average	12845	13371	13116	14110
Estimated Standard Error of Average	706.728317	785.775838	491.155990	468.293677
95% Confidence Interval (C.I.) for Average	(11452.4887; 14237.8618)	(11822.0848; 14920.2095)	(12148.2866; 14084.0885)	(8159.29715; 20059.7678)
C.I. width = 2x error	2785.3731	3098.1247	1935.8019	11900.47065
	SRS	**SRS WR**	**Stratified SRS**	**Cluster SRS**
Estimated Proportion	0.466667	0.452381	0.466996	0.485465
Estimated Standard Error of Proportion	0.030399	0.031609	0.025546	0.011736
95% Confidence Interval (C.I.) for Proportion	(0.40676122; 0.52657212)	(0.39006708; 0.51469483)	(0.41665495; 0.51733802)	(0.33634870; 0.63458153)
C.I. width = 2x error	0.1198109	0.12462775	0.10068307	0.29823283

Conclusion:

From the results obtained above, we observe that cluster sampling had the smallest standard error for the estimates of both parameters of interest – the average charges and the proportion of individuals whose charges are more than $10,000. **However**, the seemingly lower variance found for Cluster Sampling may be attributed to the fact that it used a sample size of 688, which was more than 3x bigger than the specified sample size of 225 used in the other three sampling methods. Using a bigger sample size will always reduce the variance of the estimate we obtain, but we also observe that Cluster SRS gave us the widest 95% C.I. for both of our estimates, which suggests that we would get much less accuracy from using those results. We can therefore conclude that stratified sampling gave us the most reliable estimates overall since it had a much lower standard error and produced the narrowest 95% C.I. for both the average and proportion parameters of interest compared to SRS and SRS WR. Thus, the estimated average annual health insurance costs of a beneficiary residing in the US is $13,116 and we are 95% sure that the true average of those costs is between $12,148.29 and $14,084.09. Similarly, the estimated proportion of beneficiaries whose health insurance costs amount to more than $10,000 in a particular year is 46.7% and we are 95% sure that the true proportion of those beneficiaries (amongst the 1337) is between 41.7% and 51.7%.