Diabetes Dataset

The aim of this analysis was to explore how sampling and bootstrap resampling influence the estimation of statistics such as means, maximum values, standard deviations, and percentiles. By comparing these estimates with the values from the full dataset (treated as the population), we can understand how sample size affects accuracy and stability.

1. Random Sample of 25: Comparison of Glucose Mean and Maximum

Statistic	Sample Value	Population Value
Mean Glucose	130.36	120.8945
Max Glucose	197	199

What I Did

- I selected 25 observations at random from the full dataset using a fixed seed for reproducibility.
- For this sample, I computed the mean and maximum Glucose levels.
- Then I calculated the same two statistics using the entire dataset.
- Both sets of values were displayed side-by-side using visual comparisons such as bar charts.

Interpretation

- The sample estimates differed noticeably from the population values, which is expected when drawing small samples.
- A sample of only 25 individuals does not capture the full variation in Glucose levels present in the population.
- The differences highlight natural sampling variability — small samples may either overestimate or underestimate the true mean or maximum.
- The charts visually emphasize how much the sample deviated from the actual population statistic

2. 98th Percentile of BMI: Sample vs Population

Statistic	Sample Value	Population Value
98th Percentile (BMI)	45.264	47.526

What I Did

- Using the same sample of 25 individuals, I computed the 98th percentile of BMI.
- I also calculated the 98th percentile using all 768 observations.

- A bar chart was used to compare the two percentile values and visualize the difference.

Interpretation

- The percentile calculated from the small sample didn't match the population's 98th percentile closely.
- This is expected because extreme percentiles depend heavily on having a large number of observations.
- With only 25 data points, the upper tail of the BMI distribution isn't well represented, leading to unstable percentile estimates.
- This reinforces the idea that percentiles, especially high ones, are more reliable when calculated from larger datasets.

3. Bootstrap Analysis (500 Samples of Size 150)

Using BloodPressure Variable

Statistic	Bootstrap Average	Population Value
Mean	69.1546	69.1055
Standard Deviation	19.2049	19.356
98th Percentile	98.0235	98.0

What I Did

- I created 500 bootstrap samples, each containing 150 observations drawn with replacement.
- For each bootstrap sample, I calculated the mean, standard deviation, and 98th percentile of BloodPressure.
- The results were stored and visualized using histograms and summary plots.
- I compared the average across the 500 bootstrap statistics with the actual population values.

Interpretation

- The bootstrap averages were very close to the population statistics, showing that a sample size of 150 provides stable and reliable estimates.
- Unlike the earlier small sample of 25, the bootstrap sampling distribution centered tightly around the true population numbers.
- The variability across the 500 bootstrap samples was small, indicating low uncertainty.

- The percentile estimate was also much more accurate due to the larger sample size used in each bootstrap resample.

- Overall, the bootstrap technique effectively illustrated how sample-based estimates converge toward the population values with sufficient sample size.

Conclusion:

- Small samples (like n = 25) can give rough estimates but show substantial variability, especially for extreme statistics like maxima and high percentiles.
- Bootstrap resampling helps us see how much our estimates might change from one sample to another.
- Increasing the sample size (to n = 150) narrows the bootstrap distributions, bringing estimates much closer to the true population values.
- Together, these experiments emphasize the importance of sample size and resampling methods in statistical inference.
-