Medical Insurance Cost Prediction using Regression Analysis

# Introduction

Among the available sources of links, I chose the “Private industry workers by occupational and industry group” data set for my project. The data set provides a comprehensive overview of the distribution of workers across various occupational categories and industry sectors within the private sector. By categorizing workers based on their occupations and industries, the table offers valuable insights into the composition of the private industry workforce.  
  
***Source: -*** United States Department of Labor, Bureau of Labor Statistics. (n.d.). Table 4. Average hourly and weekly earnings of all employees on private nonfarm payrolls by industry sector, seasonally adjusted. Retrieved from   
<https://www.bls.gov/news.release/ecec.t04.htm>

***Modified dataset* (sample size = 141 data points)**

[**BDP Mid Project Dataset**](https://docs.google.com/spreadsheets/d/1Fhb6AglGEb-ap1KUYMIDiTkpKPE31DbakjWrSj6djXc/edit#gid=0)[*https://docs.google.com/spreadsheets/d/1Fhb6AglGEb-ap1KUYMIDiTkpKPE31DbakjWrSj6djXc/edit#gid=0*](https://docs.google.com/spreadsheets/d/1Fhb6AglGEb-ap1KUYMIDiTkpKPE31DbakjWrSj6djXc/edit#gid=0)

# Module 1: Probability with Z-score

In the Management, business, and financial operations occupations, the mean number of private industry workers is 5000 with a standard deviation of 1000. The probability of randomly selecting a company that employs more than 6000 workers is calculated using the *Z-score:*Z = (6000 - 5000) / 1000 = 1  
P(Z ≥ 1) = 0.8413 (from Z-table)  
Therefore, the probability is approximately 84.13%.

# Module 2: Confidence Interval Estimation

Sample of 100 workers from Sales and Related Occupations with mean = 350, std = 30.  
Using a 95% confidence level and a t-value of 1.984:  
CI = 350 ± 1.984 \* (30/√100) = [344.048, 355.952]

# Module 3: Hypothesis Testing

Null Hypothesis (Ho): µ = 400  
Alternative Hypothesis (Ha): µ ≠ 400  
Sample mean = 395, std = 45, n = 100  
Z = (395 - 400) / (45/√100) = -1.111  
P-value = 0.266  
Conclusion: Fail to reject Ho. No significant difference from 400.

# Regression Analysis: Insurance Charges

Dataset Source: Kaggle - https://www.kaggle.com/datasets/mirichoi0218/insurance  
Modified Dataset: https://docs.google.com/spreadsheets/d/15EztHgS-q6WrVakS3vCKHJMplBOIKySzAzvafDB6RJw  
  
Independent Variables:  
- Age  
- BMI  
- Smoker (0 or 1)  
- Sex (0 = male, 1 = female)  
- Children (integer)  
Dependent Variable:  
- Charges  
  
Regression Equation:  
Estimated charges = -4260.94 + 387.67 × age + 23609.39 × smoker + 308.58 × sex + 599.60 × children

## Interpretation of Regression Output

- R = 0.813, suggesting a strong correlation.  
- R² = 0.6616: ~66.2% of variation in charges is explained by the model.  
- BMI: every unit increase increases charges by ~$887.  
- Smoker: smokers pay ~$23,609 more.  
- Sex: effect is minor and statistically insignificant.  
- Children: Each child adds ~$599 to charges.

# Summary

This project explored statistical tools (Z-score, Confidence Intervals, Hypothesis Testing) and applied Regression Analysis on a simulated medical insurance dataset. Findings show that smoking status, BMI, and number of children significantly impact medical costs. The Tableau visualizations created for this project further demonstrate these insights in a graphical format.