

ELEMENTS OF STATISTICS AND PROBABILITY R EXERCISE

Submission Deadline: Tuesday 28 September 2021

The salary data frame contains information about 474 employees hired by a Midwestern bank between 1969 and 1971. It was created for an Equal Employment Opportunity (EEO) court case involving wage discrimination. The file contains beginning salary (SALBEG), salary now (SALNOW), age of respondent (AGE), seniority (TIME), gender (SEX coded 1 = female, 0 = male) among other variables.

1. Load the data from the SPSS file.
2. Run a frequency table for “gender”. Report how many men are in the sample. Which type of graph would best display data for a variable like gender?
3. Run a frequency table for “jobcat”. Report the most common job category and its percentage.
4. Make a pie chart for “jobcat”
5. Report the percentage of people who are either clerical workers or security officers
6. Run a frequency distribution for “salnow”
7. Report the highest salary and how many people make that salary
8. Report the mean and median salary
9. What is the salary of the 10% highest paid employees in this bank?
10. Create a histogram of education level (edlevel). Describe the shape of the distribution. Run descriptives for education level (edlevel). Write a short description/report.
11. Report the mean, standard deviation and range of all quantitative variables.
12. Perform a Z-transformation on all quantitative variables. Report the mean, standard deviation and range of the transformed scores
13. Using z-transformed variables, select females only. What is their average standardized score for education level?

14. Reselect all cases. Calculate a new variable called “raise” which is the difference between current salary and beginning salary. Report the mean, median, and standard deviation for “raise”.
15. Which person (report case ID) had the greatest increase in salary from beginning until now? Report the value of the increase.
16. If you have many different values for a variable, it may be more meaningful to group scores together. Recode “salbeg” into a new variable called “salbeg2”.
0 = \$0-\$4,999
1 = \$5,000 - \$9,999
2 = \$10,000 - \$14,999
3 = \$15,000 - \$19,999
4 = \$20,000 - \$24,999
17. How many people (and their %) have beginning salaries between \$10,000 and \$14,999?
18. Prepare a scatterplot showing the relationship between education level (edlevel) and current salary (salnow). Put education on the x-axis and current salary on the y-axis.
19. How does the relationship between education and current salary appear overall? Linear or non-linear?
20. Add the regression line to the graph.

Deliverables

- A short report with comments on your findings
- Your script in an R file