Homework 1

2024-10-03

Problem Description

Our objective is to study the difference in weight between babies born to mothers who smoked during pregnancy and those who did not. We aim to determine whether this difference is significant for the health of the baby.

You are provided with a Child Health and Development Studies (CHDS) dataset, which includes all pregnancies that occurred between 1960 and 1967 among women in the **Kaiser Health Plan in Oakland, California**. (Dataset is in module section - babies.txt)

Data is comprised of 1236 babies:

1. all the same gender: boys
2. single births: no twins
3. all lived at least 28 days

Variables measured:

1. bwt: birthweight, in ounces
2. gestation: length of gestation, in days
3. parity: binary indicator for a first pregnancy (0 = first pregnancy)
4. age: mother's age in years
5. height: mother's height in inches
6. weight: mother's weight in pounds
7. smoke: binary indicator for whether the mother smokes (0 = no)

Questions

1. The first step of any report is understanding the data
    a. Examine all your variables. Check variable types, summarize all values, plot histograms (save description as *variable_name_description* for example - *bwt_description*)
    b. Catch inconsistencies- NA values, and outliers (save final dataframe as *cleaned_df*)
    c. Evaluate whether the dataset qualifies as a simple random sample, and discuss the implications for its generalizability.
2. Summarize **numerically** the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy. Find the following values to summarize.
    a. Minimum and maximum values (save as *smoker_min_bwt*, *smoker_max_bwt*, *nonsmoker_min_bwt*, *nonsmoker_max_bwt*)
    b. Mean (save as *smoker_mean_bwt*, *nonsmoker_mean_bwt*)
    c. Median (save as *smoker_median_bwt*, *nonsmoker_median_bwt*)

   d. What can be inferred about skewness when comparing the mean and median values?
   e. Quartiles (save as *smoker_q1_bwt*, *smoker_q2_bwt*, *smoker_q3_bwt*, *nonsmoker_q1_bwt*, *nonsmoker_q2_bwt*,*nonsmoker_q3_bwt*)
   f. Standard deviation (save as *smoker_std_bwt*, *nonsmoker_std_bwt*)

3. Summarize the two distributions **graphically**

   a. Draw at least two graphs to compare the two birth weight distributions (of women who smoked and who did not).
   b. Write your learnings about the distribution of the data from the graph and compare the two groups.

4. Compare the **incidence** (frequency) of low-birth-weight babies (those weighing under 100 ounces) between the two groups.

   a. What percentage of the babies weigh under 100 ounces to women who used to smoke? (save as *low_bwt_smoker*)
   b. What percentage of the babies weigh under 100 ounces to women who did not smoke? (save as *low_bwt_nonsmoker*)
   c. How would the incidence of low birth weight change if we changed the threshold so that more or fewer babies were classified as low birth weight.
   d. How would the change in part c affect the comparison between the two groups.

5. Evaluate how reliable the three types of comparisons—numerical, graphical, and incidence—are, based on the variability you found in your analysis. Highlight the strengths and weaknesses of each type.

Submission Guidelines

Please review the **HW Submission Guidelines** document and **HW reporting format** available in the Canvas modules. The guidelines are consistent for all assignments.