Mairead Dillon

Longitudinal Analysis HW 7

1. For each outcome (stroke, kcal, crp, alcohol, sysbp) and the scientific questions for each outcome, please provide the following descriptive statistics:
   * Sample sizes
   * Arithmetic means and variances (use medians and IQR’s for crp)
   * Skewness, kurtosis

Please provide these summaries for the full dataset and stratified summaries for the comparison groups (e.g., sex, age group) and mean differences (medians for crp). Please note relative sample sizes and relative variances between the comparison groups for each question.

1. For each outcome, briefly comment on whether the populations are normally or approximately normally distributed. You can use descriptive summaries from Problems 1 and/or other descriptive summaries you choose (e.g., histograms, Q-Q plots, etc.).

Physical Activity (kcal)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample Size | Mean | Variance | Skewness | Kurtosis | Relative Sample Size | Relative Variance |
| All | 5868 | 1708.44 | 4110950 | 2.2864 | 9.6827 |  |  |
| < 70 years | 5074 | 1784.575 | 4216202 | 2.1823 | 9.0236 | 6.3904 | 1.3305 |
| 70 years | 794 | 1221.905 | 3168709 | 3.3192 | 17.8116 |
| Females | 3385 | 1551.233 | 3736498 | 2.5047 | 11.2771 | 1.3633 | 0.8224 |
| Males | 2483 | 1922.754 | 4543486 | 2.0510 | 8.1714 |

Chart, histogram

Description automatically generated

Based on the histogram, the population is not normally distributed.

Alcohol Consumption (alcoh)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample Size | Mean | Variance | Skewness | Kurtosis | Relative Sample Size | Relative Variance |
| All | 5865 | 2.4664 | 45.1811 | 5.2202 | 43.1752 |  |  |
| < 70 years | 5073 | 2.5540 | 47.6218 | 5.1310 | 41.4727 | 6.4053 | 1.6295 |
| 70 years | 792 | 1.9054 | 29.2241 | 5.7631 | 56.8495 |
| Current Smokers | 696 | 3.8394 | 107.3366 | 4.4645 | 27.6304 | 0.1348 | 2.9362 |
| Former/Non-Smokers | 5163 | 2.2806 | 36.5558 | 4.9459 | 40.6224 |

Chart, histogram

Description automatically generated

Based on the histogram, the population is not normally distributed.

Systolic Blood Pressure (sysbp)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample Size | Mean | Variance | Skewness | Kurtosis | Relative Sample Size | Relative Variance |
| All | 5879 | 136.5742 | 477.0281 | 0.5826 | 3.4835 |  |  |
| < 70 years | 5084 | 135.4165 | 455.0954 | 0.5877 | 3.5526 | 6.3950 | 0.8206 |
| 70 years | 795 | 143.9779 | 554.5707 | 0.4455 | 3.0803 |
| Females | 3387 | 137.2079 | 495.8720 | 0.5782 | 3.5090 | 1.3591 | 1.1012 |
| Males | 2492 | 135.7129 | 450.3171 | 0.5787 | 3.4097 |

Chart, histogram

Description automatically generated

Based on the histogram, the population is approximately normally distributed.

C-Reactive Protein (crp)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample Size | Mean | Variance | Median | Skewness | Kurtosis | Relative Sample Size | Relative Variance |
| All | 5806 | 4.9494 | 69.4532 | 2.6809 | 6.9119 | 78.1199 |  |  |
| Females | 3341 | 4.8142 | 57.6953 | 2.6809 | 6.9154 | 88.0337 | 1.3554 | 0.6759 |
| Males | 2465 | 5.1327 | 85.3609 | 2.6809 | 6.7241 | 67.1150 |

Chart, histogram

Description automatically generated

Incidence Stroke (stroke)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample Size | Mean | Variance | Skewness | Kurtosis | Relative Sample Size | Relative Variance |
| All | 5888 | 0.0421 | 0.0404 | 4.5592 | 21.7859 |  |  |
| Current Smokers | 700 | 0.0414 | 0.0398 | 4.6023 | 22.1812 | 0.1351 | 0.9824 |
| Former/Non-Smokers | 5182 | 0.0423 | 0.0405 | 4.5504 | 21.7062 |
| Females | 3393 | 0.0309 | 0.0300 | 5.4172 | 30.3462 | 1.3599 | 0.5550 |
| Males | 2495 | 0.0573 | 0.0541 | 3.8090 | 15.5084 |

Chart, histogram

Description automatically generated

Based on the histogram, the population is not normally distributed.

**INVESTIGATION OF THE VALIDITY OF STATISTICAL INFERENCE**

*For the problems below, consider the homework dataset in Problems 1 and 3 to be the relevant study “population”.*

For each of the SCIENTIFIC QUESTIONS, use Monte Carlo (MC) simulations to determine what sample sizes you feel are sufficient to trust the validity of your inference.

* You can use samples increments of 20 (e.g., n = 60, 80, 100, ...). You are welcome to use smaller increments, but beware that the smaller the increments, the longer it will take your sample size evaluations to execute.
* Draw “simple random samples” from the study “populations” (i.e., the full homework dataset) for your Monte Carlo simulations.
* For each sample size under consideration, perform a reasonably large number of Monte Carlo (MC) iterations. At minimum, use at least 500 for each sample size considered. You can use more iterations (e.g., in the thousands) to minimize your simulation error, but be aware that the more iterations performed, the longer it may take for your simulations to execute. You might begin strategically using a smaller number of MC iterations (e.g., B=200) to get an idea of the behavior of your estimators, and then increase the MC iterations after you’ve narrowed the scope of your simulations (e.g., range of sample sizes) for your “final” simulations.

**Sample size determination: Please evaluate**

* 1. Coverage: Estimate 95% confidence intervals for each simulation/iteration. Because you have the true population parameters for each scientific question, you know the “true” differences to compare your 95% confidence intervals against.
  2. Bias: For each MC iteration, you can compare your estimate versus the population parameters. You can compare the mean of your MC estimates and the population value (i.e., mean difference). You may (optionally) estimate the mean absolute bias or the mean squared bias.
  3. Optional: You may choose to evaluate whether the estimators behave like normal random variables, at given sample sizes, using graphical summaries (e.g., Q-Q plots), evaluate coverage levels for different coverages (e.g., 99%). These summaries might provide additional insights but are not required.

**ROBUST OR VARIANCE CORRECTED ESTIMATORS**

1. For each of the first four scientific question (a) – (c), determine what sample sizes are sufficiently large for you to have confidence of the validity of the statistical inference. **Please use either linear regression with sandwich variance estimates or Welch’s (unequal variances) t-tests for your work.** Please fill in the values for Table 1 below:

Table 1. Approximate minimal sample sizes for valid inference

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Outcome | Scientific Question | Minimum Sample Size | Nominal Coverage | Mean Bias | Relative N | Relative Variance |
| (a) kcal | (i) |  |  |  |  |  |
|  | (ii) |  |  |  |  |  |
| (b) alcoh | (i) |  |  |  |  |  |
|  | (ii) |  |  |  |  |  |
| (c) sysbp | (i) |  |  |  |  |  |
|  | (ii) |  |  |  |  |  |

1. For each of the three outcomes, were the sample size requirements approximately equal for each of the two scientific questions? If they were not, what might have contributed to the differences, especially for different sample size requirements for the same outcome?
2. kilo-calories outcome –
3. alcohol outcome –
4. systolic blood pressure outcome –

**MODEL BASED OR NOMINAL ESTIMATORS**

1. For the first four scientific question (a) – (c), rerun your simulations to determine sample sizes that are sufficiently large for you to have confidence of the validity of the statistical inference. **Please use either linear regression without sandwich variance estimates or use the classical Student’s t-test (that assume equal variances) for your work.** Please pay attention to the nominal coverage of your confidence intervals and bias. Please fill in the values for Table 2 below.
2. For each of the three outcomes and their two scientific questions, were the results similar between these results and the “robust” results? How were they comparable? If they were different, how were they different? Was one approach superior to the other? Please refer to the summaries in the Tables 1 and 2, and other observations made from your investigation.