# Problem Set 4

*Statistics 104*

*Due March 26, 2020 at 11:59 pm*

**Problem set policies.** *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, be sure to include the code in your solution.*

*Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.*

*We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.*

**Problem 1.**

A company that produces coffee for use in commercial machines monitors the caffeine content in its coffee. The company selects 35 8-oz samples each hour from its production line to analyze. The samples collected one morning between 8:00 - 9:00 am contained on average 96.1 mg of caffeine, with standard deviation 1.2 mg.

a) Compute and interpret a 95% confidence interval for mean caffeine content based on the collected data.

b) According to production standards, the mean amount of caffeine content per 8 ounces should be no more than 95 mg. An overly high caffeine content indicates that the coffee beans have not been roasted long enough.

   Conduct a formal hypothesis test to investigate whether production standards are being met, based on the observed data. Summarize your findings to the CEO using language accessible to someone who has not taken a statistics course and make a recommendation as to whether an adjustment needs to be made to the bean roasting time.

c) A set of samples collected between 10:00 - 11:00 am on the same day has average caffeine content of 95.3 mg, with standard deviation 1.1 mg. Based on observing this data, would you change your recommendation in part b)? Explain your answer.

**Problem 2.**

The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. The survey is conducted face-to-face with an in-person interview of a randomly selected sample of adults. One of the questions on the survey is "After an average workday, about how many hours do you have to relax or pursue activities that you enjoy?" A 95% confidence interval from the 2010 GSS survey for the collected answers is 3.53 to 3.83 hours.

Identify each of the following statements as true or false. Explain your answers.

a) If the researchers wanted to report a confidence interval with a smaller margin of error based on the same sample of 1,154 Americans, the confidence interval would be larger.

b) We can be 95% confident that the interval (3.53, 3.83) hours contains the mean hours that the sampled adults have for leisure time after an average workday.

c) The confidence interval of (3.53, 3.83) hours contains the mean hours that U.S. adults have for leisure time after an average workday.

d) The survey provides statistically significant evidence at the $\alpha = 0.05$ significance level that the mean hours U.S. adults have for leisure time after the average workday is 3.6 hours.

e) There is a 5% chance that the interval (3.53, 3.83) hours does not contain the mean hours that U.S. adults have for leisure time after an average workday.

f) The interval (3.53, 3.83) hours provides evidence at the $\alpha = 0.05$ significance level that U.S. adults, on average, have fewer than 3.9 hours of leisure time after a typical workday.

**Problem 3.**

After graduating, you decide to apply for positions as a statistical consultant. A friend who graduated a few years prior and already works in the field has sent you the following set of scenarios. If she is impressed by your responses, she will recommend your name to the hiring team.

For each part, provide a clear explanation and limit your answer to no more than five sentences.

a) A pharmaceutical company is planning to run an initial trial on a potential new cholesterol lowering drug compound to assess whether they should continue to invest money on its development. They would like your recommendation as to whether the initial trial should be run at the $\alpha = 0.05$ or $\alpha = 0.10$ significance level.

b) Melatonin pills are a popular remedy for providing insomnia relief. A company developed a new formulation of melatonin pill and conducted a trial on a random sample of 2,000 American adults who were already experiencing insomnia and using melatonin pills. A 95% confidence interval for mean increase in sleep hours per typical night was calculated based on data from the trial: (0.10, 0.15) hours. Assess whether the evidence suggests the new formulation is a substantial improvement over melatonin pills currently available on the market.

**Problem 4.**

A hospital administrator has been asked by her supervisor to assess whether waiting time in the emergency room (ER) has changed from last year, when average wait time was 127 minutes. The administrator collects a simple random sample of 64 patients and records the time between when they checked in at the ER until they were first seen by a doctor; the average wait time is 137 minutes, with standard deviation 39 minutes.

a) Compute and interpret a 95% confidence interval for mean ER wait time at the hospital. Based on the interval, is mean ER wait time statistically significantly different from 127 minutes at the $\alpha = 0.05$ level?

b) Would the conclusion in part a) change if the significance level were changed to $\alpha = 0.01$?

c) Suppose that upon seeing the results from part a), the supervisor criticizes the hospital administrator on the basis that ER wait times have increased greatly from last year and the

administrator must be at fault. Present a brief argument in favor of the administrator; be sure to fully explain your reasoning.

**Problem 5.**

The file `nc.csv` contains data on 1,000 randomly sampled births from birth records released by the state of North Carolina in 2004. The following variables are contained in the dataset:

| Variable | Description |
|---|---|
| `fage` | Father's age, in years |
| `mage` | Mother's age, in years |
| `mature` | Whether the mother was considered a `younger mom` or `mature mom` |
| `weeks` | Length of pregnancy, in weeks |
| `premie` | Whether the birth was full-term (`full term`) or premature (`premie`) |
| `visits` | Number of hospital visits during pregnancy |
| `marital` | Whether the mother was married at the time of the birth |
| `gained` | Weight gained by the mother during pregnancy, in pounds |
| `weight` | Weight of the baby at birth, in pounds |
| `lowbirthweight` | Whether the baby was classified as low birthweight or not |
| `gender` | Sex of the baby, female or male |
| `habit` | Whether the mother was a nonsmoker or smoker |
| `whitemom` | Whether the mother was white or nonwhite |

a) From the data, compute and interpret a 95% confidence interval for the mean birthweight of babies born in North Carolina in 2004.

b) From national census figures, the mean birthweight for full-term births is 7.5 pounds. Conduct a formal hypothesis test to assess whether mean birthweight for premature births in North Carolina is different from 7.5 pounds. Summarize your findings; do they cohere with your expectations?

c) Preterm birth occurs disproportionately among communities with lower socieconomic status. Suppose an investigator takes a random sample of 40 birth weights from several teaching hospitals located in an inner-city neighborhood. The results of the analysis will inform whether an income support and maternal health program (providing some medical services free of cost) will be started at the neighborhood community center.

   i. State the hypotheses for comparing mean birth weight for babies born in these teaching hospitals to 7.5 pounds. Justify your choice of alternative hypothesis.

   ii. Discuss whether an $\alpha$ of 0.10 or 0.01 might be more appropriate than the standard value of 0.05.

**Problem 6.**

In each of the following scenarios,

i. Identify whether the data are paired or independent.

ii. Consider whether a design that produces the "other type" of two-sample data might be preferable (for example, if the data are paired, consider whether a design that produces independent data might be preferable). If so, briefly describe the alternative study design. If not, briefly explain why the design presented in the question statement is preferable.

For each scenario, limit your description/explanation to at most five sentences.

a) Does salary differ by gender for full professors at American universities? Draw a random sample of 50 female full professors and 50 male full professors, and compare their salaries.

b) Does Vitamin E increase artery thickness? Measure artery thickness for a group of patients before they start taking Vitamin E regularly for two years. Compare initial artery thickness to artery thickness at the end of the study.

c) Is a Mediterranean diet effective for weight loss? Compare the weights of individuals before and after going on the diet regimen.

d) Do Intel's stock and Southwest Airlines' stock have similar rates of return? Take a random sample of 60 days, and compare Intel's and Southwest's stock on those days.

**Problem 7.**

A possible important environmental determinant of lung function in children is amount of cigarette smoking in the home. Suppose this question is studied by selecting two groups: Group 1 consists of 23 nonsmoking children 5-9 years of age with two parents who smoke, and have mean forced expiratory volume (FEV) of 2.1 L and a standard deviation of 0.7 L; group 2 consists of 20 nonsmoking children of comparable age, with two parents who do not smoke, and have mean FEV of 2.3 L and a standard deviation of 0.4 L.

a) Do the children of two parents who smoke have mean FEV different from the children of two parents who do not smoke? Conduct a formal hypothesis test. Use the $\min(n_1 - 1, n_2 - 1)$ approximation for the degrees of freedom.

b) Summarize your findings in language that someone who has not taken a statistics course would understand.

c) Suppose you have been asked to assess the design of a new study addressing the same question. In the new study, researchers plan to recruit 20 nonsmoking children 5-9 years of age with two parents who smoke and 20 nonsmoking children 5-9 years of age with two parents who do not smoke; children will be matched based on similar household income. Briefly explain whether you believe the new study has more potential to show an association between smoke exposure and lung function than the original study.