Section 6

Statistics 104

Fall 2019

Topics

- Two-sample t-test for paired data
- Two-sample t-test for independent group data
- Statistical power and sample size
- ANOVA

Notes

1. **The Paleo Diet.** The Paleo diet allows only for foods that humans typically consumed prior to the development of agriculture about 10,000 years ago. To study the efficacy of the Paleo diet, researchers recruited 500 adult volunteers interested in losing weight from around the Boston area. Volunteers were randomly assigned to either of two equally sized treatment groups. One group spent six months following the Paleo diet, while the other group received a weekly email about healthy eating practices such as reducing portion size. At the beginning of the study, the average difference in weights between the two groups was about 0. The file `diets.Rdata` contain the change in weight for each volunteer, measured in pounds and calculated as (weight at beginning of study - weight at end of study).

    a) Calculate the 90% confidence interval for the difference between the mean weight loss of the two groups (Paleo - control). Interpret this interval in the context of the data.

b) Based on the confidence interval, do the data provide convincing evidence that the Paleo diet is more effective for weight loss than the weekly email newsletter?

c) Would it be reasonable to believe that the study results generalize to the population of American adults? Explain your answer.

d) Without explicitly performing the hypothesis test, determine whether the results would have indicated a significant difference in population means if the Paleo group had lost 8.25 pounds on average instead of 7.25. Assume that all other numbers remained constant.

e) Suppose a new study will be conducted that specifically recruits twins; of a pair of twins, one individual will be assigned to the Paleo diet group, while the other will be assigned to the email newsletter group. The study team is interested in detecting any average difference of at least 5 pounds, and anticipate that the standard deviation of weight will be about 15 pounds. Calculate the number of twin pairs that should be recruited to achieve a power level of 80%, if the study data will be analyzed at $\alpha = 0.10$.

2. **Indoor Air Quality.** A study was conducted in 1980 to compare the indoor air quality in offices where smoking was permitted with that in offices where smoking was not permitted. Measurements were made of carbon monoxide (CO) at 1:20 p.m. in 40 work areas where smoking was permitted and in 40 work areas where smoking was not permitted; CO levels were measured in parts per million (ppm). The data are in `air_quality.Rdata`. [1]

   Analyze the data using a hypothesis test and confidence interval. Summarize your conclusions in language that a non-statistician would understand.

3. **Global Warming.** Is there strong evidence of global warming? Let's consider a small-scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. The readings are in `us_temps.Rdata`, in units of degrees Fahrenheit. We are interested in determining whether these data provide strong evidence of temperature warming in the continental US.

   According to the worldwide scientific community, emission of greenhouse gases like carbon dioxide are a major cause of global warming; emissions have increased over time due to use of fossil fuels for electricity, transportation, etc. Suppose that our analysis will be used to inform whether existing legislation to limit carbon dioxide emissions should be maintained.

   a) State the hypotheses and α. Consider whether an α different from 0.05 could be reasonable. Justify your choice of alternative hypothesis and α.

   b) Conduct the hypothesis test and interpret your conclusions.

   c) Calculate and interpret a confidence interval that corresponds to the test conducted in part b).

[1] Data simulated from summary statistics in Rosner, *Fundamentals of Biostatistics*, 7^{th} ed., p. 310

4. **Work Hours and Education.** The General Social Survey collects data on demographics, education, and work, among many other characteristics of U.S. residents. The dataset `gss_work.Rdata` contains the simulated responses for 954 participants. The variable `work.hours` contains the number of hours the respondent works in an average week, while `education.level` contains the highest level of education attained.

   Investigate whether the average amount of hours worked per week differs between individuals who have different levels of educational attainment, at significance level $\alpha = 0.05$.

   a) Evaluate the assumptions for ANOVA.

   b) State the null and alternative hypotheses.

   c) Conduct the F-test and summarize the results.

   d) Complete the analysis using pairwise comparisons.

      i. What is the appropriate significance level $\alpha^\star$ for the individual comparisons, as per the Bonferroni correction?

      ii. Conduct pairwise comparisons and summarize the results.

5. **Corn Yield.** A large farm wants to try out a new type of fertilizer to evaluate whether it will improve the farm's corn production. The land is broken into plots that produce an average of 1,215 pounds of corn with a standard deviation of 94 pounds per plot. The owner is interested in detecting any average difference of at least 40 pounds per plot. Assume each plot of land gets treated with either the current fertilizer or the new fertilizer. Let $\alpha = 0.05$.

   a) How many plots of land would be needed for the experiment if the desired power level is 90%?

   b) What would the power of the study be if 160 total plots of land were used for the experiment, with an equal number being treated with the current fertilizer versus new fertilizer?

   c) Would the power of the test increase or decrease if the significance level were increased to $\alpha = 0.10$? Explain your answer.