9.07 Introduction to Statistics for Brain and Cognitive Sciences
Emery N. Brown

Homework Assignment 8
November 28, 2016
Due December 7, 2016 at 5:00 PM

This homework assignment has problems covering the material discussed during the last two weeks of class. You need only hand in solutions to problems 1-3, 5,6 and 9. Please do the other problems. We will hand out solutions to all of these problems with the solution set.

1. (Confidence Intervals and $z-$test)
The gross calorific value is a measure in megajoules per kilogram (MJ/kg) of the heat content of coal. Coal that meets a certain standard receives an International Organization of Standardization (ISO) certificate. If the assessment is carried out by ISO standards the measurements are known to be Gaussian with a standard deviation $\sigma = 0.1$. A sample of gross caloric values from a laboratory are in the file COAL.txt on the course Homework page.

   A. Test the null hypothesis that the true mean value for the lab is 23.73 against the alternative that it could be either greater than 23.73 or less than 23.73 at $\alpha = 0.05$. What do you conclude?

   B. What is the $p-$value?

   C. Construct a 95% confidence interval for the true gross calorific value based on this sample. Do your results agree with the results in A? Why or why now?

   D. Which is more informative, A and B, or C? Why?

2. (Confidence Intervals and $t-$test)

   A. Estimate σ^2 from the data in Problem 1 as

   $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

   B. Use the results in **Lecture 12** pages 10-11 to test the same null and alternative hypotheses in 1A using a $t-$test based on s^2 estimated in 2A. Is the conclusion the same as in 1A? Why or why not?

   C. Compute a 95% confidence interval for the true gross calorific value based on this sample using the estimate of s^2 in 2A and the $t-$distribution. How does it differ from the findings in 1C?

3. In **Lecture 12** we computed for **Example 2.1** the power of rejecting H_0 $p = \dfrac{1}{2}$ under the alternative hypothesis that the true probability of correctly responding is $p = 0.72$.

    A. Find c_α, if $\alpha = 0.01$. (See page 5 of **Lecture 12**.)

    B. Compute the power using c_α in 3A for the alternative hypothesis $H_A : p_A = 0.28$, to indicate impaired learning.

4. Polygraphs that are used in criminal investigations are supposed to indicate whether a person is lying or telling the truth. However the procedure is not infallible, as is illustrated by the following example. An experienced polygraph examiner was asked to make an overall judgment for each of a total of 280 records, of which 140 were from guilty suspects and 140 were from innocent subjects. The results are listed in Table 1. We view each judgment as a problem in hypothesis testing, with the null hypothesis corresponding to "suspect is innocent" and the alternative hypothesis corresponding to "suspect is guilty." Estimate the probabilities of a type I error and of a type II error using these data.

Table I. Results of Polygraph Tests

		Suspect's True Status	
		Innocent	Guilty
Examiner's	Acquitted	131	15
Assessment	Convicted	9	125

5. Power and Sample Size Calculation
The power for a one-sided alternative for testing a hypothesis about the mean of a Gaussian distribution is with known variance is

$$\text{power} = \Phi\!\left(z_\alpha + \frac{n^{\frac{1}{2}}(\mu_A - \mu_0)}{\sigma}\right).$$

    A. If a study is to have a type I error of 0.05, for a sample of 49 subjects, from a Gaussian distribution with a variance of 49 and an expected difference between μ_0 and μ_A of 0.5, what is the power of the study?

    B. What is the type II error?

    C. Find a sample size so that the power for this study is 0.99.

6. A clinical trial is conducted at the gynecology unit of a major hospital to determine the effectiveness of drug A in preventing premature birth and low birth weight babies. In the trial, 30 pregnant women are to be studied, 15 in a treatment group to receive drug A and 15 in a control group to receive a placebo. The patients are to take a fixed dose of each drug on a one-time-only basis between the 24th and 28th weeks of pregnancy. The patients are assigned to groups using computer-generated random numbers, where for every 2 patients eligible for the study, one is assigned randomly to the treatment group and the other to the control group. The presumption is that if the drug is effective then the mothers are more likely to carry their pregnancies to term and be less likely to have a low birth weight child. The weights of the babies from the mothers entered in the study are in the data file Baby_Weights.txt on the course Homework page. Column 1 has the weights of the babies whose mothers were in the treatment group and column 2 has the weights of the babies whose mothers were in the control group.

    A. What is the average birth weight in each of the groups?

    B. What is the sample standard deviation of the weights in each of the groups?

    C. Conduct a two sample $t-$test to determine whether or not the average birth weight in the treatment group differed from the average birth weight in the control group. Use $\alpha = 0.05$ and conduct a two-sided test. (See **Lecture 12**, **Example 12.3**.)

    D. What do you conclude about the effectiveness of the treatment?

7. A study was designed to investigate the effectiveness of epidural steroid injections in alleviating sciatica (pain in the distribution of the sciatic nerve of the leg). The treatment group received an epidural injection of methyl prednisolone whereas the control group received a placebo injection of comparable volume of normal saline. Recall that a placebo is an inert form of the treatment designed to assess whether the process of treating rather than the treatment itself may explain some component of the therapeutic benefit. Straight-leg raising was one of the outcome variables. Improvement in straight leg-raising was scored as a binary variable, i.e. either improvement or no improvement. There were 25 patients assigned to the placebo group and 25 assigned to the treatment group. Ten of the placebo group and 15 of the treatment group showed improvement. Use the Gaussian approximation to the binomial to construct a $z-$test of the hypothesis that there is no difference in the two groups using a two-sided alternative. Recall that the $z-$statistic with continuity correction for the two sample binomial test is

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - (\frac{1}{2n_1} + \frac{1}{2n_2})}{[\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})]^{\frac{1}{2}}}.$$

where $\hat{p} = (n_1 + n_2)^{-1}(k_1 + k_2)$.

8. For comparing two binomial proportions with a two-sided alternative hypothesis the sample size formula for the number of subjects per group

$$n = \frac{\left[\left[\frac{\bar{p}(1-\bar{p})}{2}\right]^{\frac{1}{2}} z_{1-\alpha/2} + [p_1(1-p_1) + p_2(1-p_2)]^{\frac{1}{2}} z_{1-\beta}\right]^2}{\Delta^2}$$

where $\Delta = |p_2 - p_1|$, $\bar{p} = \frac{p_1 + p_2}{2}$ and n is the number of subjects per group. Placebo effects are an extremely important consideration in clinical studies. A placebo effect of 0.30 is not uncommon. That is, a 30% improvement may occur by the application of a placebo. If we take $p_1 = 0.30$, the power, $1-\beta = 0.80$, and $\alpha = 0.05$, how many subjects total should be in the epidural steroid study if the true probability of improvement with the steroid injection is 0.7? How does the sample size required per group change if the true probability of improvement with the steroid injection is 0.8, or 0.9?

9. On the course Homework page is a data set (CALCIUM_RATE.txt) on the relationship between mortality and calcium particulate concentration measured in parts per million. An epidemiologist wishes to study the extent to which mortality can be predicted linearly from calcium concentration.

   A. Perform a regression analysis to analyze the relationship between mortality and calcium concentration. Report the regression coefficients, their standard errors, the R^2 and the analysis of variance table.

   B. Provide 95% confidence intervals for each regression coefficient.

   C. Plot the data with the regression fit superimposed. Does the model describe the data well?

   D. What is the predicted mortality for a calcium concentration of 60 ppm?

10. The simple linear regression analysis can be carried out using a bootstrap procedure to assess uncertainty. The key to this, as we discussed in **Lecture 11**, is to draw bootstrap samples that are at least approximately independent, identically distributed. Here is a bootstrap prescription for a simple linear regression model,

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$$

where the ε_k's are independent, identically distributed Gaussian random variables with mean zero and σ^2. First compute the maximum likelihood (least-squares) estimates. $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$. Then compute $\hat{\varepsilon}_k = y_k - \hat{y}_k$ for $k = 1, \dots, n$. As in **Lecture 11,** let $F_n(\varepsilon)$ be the empirical cdf defined

by putting mass n^{-1} at each $\hat{\varepsilon}_{(k)}$ where the $\hat{\varepsilon}_{(k)}$'s are the ordered $\hat{\varepsilon}_k$'s. The bootstrap prescription is as follows (modified from **Algorithm 11.3**):

Pick B.

i)   Draw a random sample of size n with replacement of the $\hat{\varepsilon}_k$'s from $F_n(\varepsilon)$. Denote this sample as $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)$.

ii)   Construct the bootstrap sample $y^* = (y_1^*, ..., y_n^*)$, where y_k^* is $y_k^* = \hat{\beta}_0 + \hat{\beta}_1 x_k + \varepsilon_k^*$ for $k = 1, ... , n$.

iii)   Estimate $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ from $y^* = (y_1^*, ..., y_n^*)$ by maximum likelihood (least-squares).

iv)   Repeat i), ii), and iii) B times.

v)   Compute the uncertainty in $\hat{\beta}_0$ and $\hat{\beta}_1$ as
        1. the sample standard errors.
        2. the histograms of the $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$.
        3. the 95% confidence interval from the histogram.

where the sample standard error is

$$se_B(\hat{\beta}_0^*) = \left[(B-1)^{-1} \sum_{b=1}^{B} (\hat{\beta}_{0,b}^* - \bar{\beta}_0^*)^2 \right]^{\frac{1}{2}}$$

and where

$$\bar{\beta}_0^* = B^{-1} \sum_{b=1}^{B} \hat{\beta}_{0,b}^*$$

A. Using $B = 100$, conduct a non-parametric bootstrap analysis of the mortality and calcium data and compute

    a. the $B = 100$ $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, and the $se_B(\hat{\beta}_0^*)$ and $se_B(\hat{\beta}_1^*)$.

    b. approximate 95% confidence intervals for β_0 and β_1.

B. How do these bootstrap estimates and standard errors compare with those from the original simple linear regression analysis computed in Problem 8?

11. In **Lecture 16**, we performed a one-way ANOVA on these performance data

<div align="center">Drug Level</div>

Age	Control	Dose One	Dose Two
Young	56,62,57,72	64,34,64,41	33,37,40,16
Old	62,72,61,91	64,48,34,63	17,21,49,54

Here is the partially completed two-way ANOVA table for these data.

Source	Sum of Squares	Degrees of Freedom	MS	F	$p-\text{value}$
Rows	150.00	1			
Columns	4,434.25	2			
Interactions	72.75	2			
Residual	3,495.00	18			
Mean	61,206.00	1			
Total	69,358.00	24			

Use the results on pages 7 and 8 of **Lecture 16** to complete the two-way ANOVA table. Determine if there is: 1) an effect of age on performance; 2) an effect of drug dose level on performance; and 3) if there is an interaction between age and drug dose level that affects performance.

9.07 Statistics for Brain and Cognitive Science
Fall 2016