SUTD ESD

Statistics 2017

Homework Assignment 3

Due: 1pm, Thursday 6 April.

You may submit this assignment on *eDimension*, or in the homework box on level 7, Building 1. Show working.

Question 1. While held as a prisoner of war during WWII, the British mathematician John Kerrich flipped a coin 10000 times and obtained 5067 Heads. Let p be the probability of a Head on a single coin flip.

(a) Set up the hypotheses to check if the coin was fair. Perform a hypothesis test with $\alpha = 0.1$.

(b) Find a 90% confidence interval for p.

Question 2. While trying to find the least square regression line for some data points (x_i, y_i), a drunk statistician used the points (y_i, x_i) instead.

(a) Does the correlation coefficient of the resulting regression line agree with the correct r?

(b) What about the slope of the resulting regression line: is it the same as the correct $\hat{\beta}_1$, or is it $\hat{\beta}_1$ flipped around the line $y = x$, or does something else happen?

Question 3. Refer to the spreadsheet for some triple jump data.

(a) Using the formulas given in class, compute the p-value of $\hat{\beta}_1$, under the hypotheses

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$.

(b) Using the regression line, predict the winning distance for 2020, and provide a 99% two-sided confidence interval for the prediction.

Question 4. The average orbital distances of some planets and dwarf planets are given in the spreadsheet.

(a) Explore the data, and find an appropriate transformation to linearize the relationship between 'planet number' and distance.

(b) Produce a regression line for the linearized data, and hence find an equation relating distance to 'planet number'.

Question 5. We have looked at the distribution of the digits of π in base 10, and the result is consistent with the null hypothesis that the digits of π are random. However, it is plausible for a number to appear random in one base but not in another. The spreadsheet shows the distribution of the digits of π in base 16.

(a) Perform a chi-squared test on the null hypothesis, using $\alpha = 0.05$.

(b) Give an example of a number whose base 10 digits are clearly *not* random, yet the chi-squared test fails to reject the null hypothesis.

Question 6. The Salk polio vaccine trial was the most elaborate program of its kind ever, involving millions of participants. Some of its results are given in the spreadsheet: out of the 200,000+ people given a placebo, 142 developed polio, while out of a similar number of people given the vaccine, 57 did. We would like to test (and hopefully, reject) the null hypothesis that the vaccine is ineffective. To do so, we could rephrase H_0 as: the incidence of polio is *independent* of which group one is in. Pick a reasonable α and perform a chi-squared test for independence to test this.

Question 7. Prove that in simple linear regression,

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1}{s/(s_x\sqrt{n-1})}.$$

Hint: first, simplify the $\sqrt{1-r^2}$ term using $r^2 = 1-\text{SSE}/\text{SST}$.

Question 8. In the spreadsheet you can find some French listening test scores for 20 people, before and after they took a course.

(a) This is an example of a matched pairs design (see Week 5). Let μ_1 be the true mean of the 'after' scores and μ_2 be the true mean of the 'before' scores. Test the hypotheses $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$, using $\alpha = 0.01$.

(b) We can also test for H_0 using a *permutation* test. In this test, each resample is generated by randomly permuting (without replacement) every person's scores. From 1000 resamples, construct a histogram and estimate the p-value of the observed difference in means.

For the histogram, you only need to submit a picture and the data used to construct it; do *not* submit the resamples. However, do *submit your code* for generating the resamples. It is possible to write the code in one line in *Excel*; *R* is not required.

Question 9. In the spreadsheet you will find some data from an early attempt to measure the speed of light, as well as instructions on *bootstrap* resampling (with replacement).

(a) Resample 1000 times, and construct a histogram of the resample means.

(b) Locate the average of the resample means. Provide a 95% two-sided confidence interval using the histogram. Does the true value, 33, lie within the CI?

(c) With this set of data, how many different resamples are possible?

(d) Why is it not wise to assume that the data is normally distributed, and then construct a CI using a t-distribution?

For the histogram, you only need to submit a picture and the data used to construct it; *do not* submit the resamples.