SUTD ESD

Statistics 2017

Assignment 3 Solutions

Question 1. (a) $H_0 : p = 1/2$, $H_1 : p \neq 1/2$. The alternative is two-sided because we do not know beforehand which way the coin may be biased.

Assume H_0 is true, then the z-score for the observed value, $\hat{p} = 5067/10000$, is

$$z = \frac{0.5067 - 0.5}{\sqrt{0.5 \times 0.5/10000}} = 1.34.$$

The p-value is therefore $2\,P(Z > z) = 0.1802 > \alpha = 0.1$, and hence we do not reject H_0.

(b) The 90% confidence interval for p is given by

$$\left[0.5067 - z_{0.95}\sqrt{\frac{0.5067(1-0.5067)}{10000}},\ 0.5067 + z_{0.95}\sqrt{\frac{0.5067(1-0.5067)}{10000}}\right] = [0.4985, 0.5149].$$

(This contains the value 1/2.)

Question 2. (a) Since $r = s_{xy}/(s_x s_y)$ is a symmetric expression in x and y (you can check from the formula that $s_{xy} = s_{yx}$), swapping x and y does not affect r.

(b) The slope is $\hat{\beta}_1 = s_{xy}/s_x^2$, so swapping x and y changes it, the new slope being s_{xy}/s_y^2. Note that the new slope is not the old one flipped around $y = x$, since that would make the slope $1/\hat{\beta}_1$.

Question 3. (a) Using either *Excel* or the explicit formula $\hat{\beta}_1 = s_{xy}/s_x^2$, we find that $\hat{\beta}_1 = -0.009786$.
From the week 10 slides, we know that this is drawn from a normal distribution with mean $\beta_1 = 0$ (assuming H_0 to be true) and variance $\sigma^2/(n-1)/s_x^2$. The variance can be approximated by $s^2/(n-1)/s_x^2 = 7.202 \times 10^{-5}$.
Therefore the t-score for $\hat{\beta}_1$ is $t = (-0.009786 - 0)/\sqrt{7.202 \times 10^{-5}} = -1.153$, and the p-value is $2\,P(T_4 < t) = 0.3131$ (note that $s^2 = \text{MSE}$ has 4 degrees of freedom). This p-value can also be generated by *Excel*.

(b) Using $\hat{\beta}_1$ and $\hat{\beta}_0 = 34.86$, the prediction for 2020 is 15.09m. Using the formula given in the week 10 slides, the confidence interval is [14.48, 15.70].

Question 4. See spreadsheet. (a) There are many ways to see that the distance D is roughly an exponential function of the planet number N (for instance, an exponential trendline fits the original data well). Therefore, we guess $D = \alpha \exp(\beta N)$. Taking the natural log of both sides, we find that $\log(D) = \log(\alpha) + \beta N$.
So the linearization uses $\log(D)$ instead of D.

(b) Simple linear regression gives $\hat{\beta}_0 = \log(\alpha) = 3.097$, $\hat{\beta}_1 = \beta = 0.5223$. Therefore, $D = 22.13\,e^{0.5223N} = 22.13 \times 1.686^N$.

Remark: It is convenient to define $D' = D/(\text{earth's distance})$, which expresses the distance in terms of *astronomical units*. Another plausible model is $D' = \alpha \exp(\beta N) + \gamma$, and it turns out (amazingly) that $D' \approx 0.075 \times 2^N + 0.4$ for $N = 1, 2, \ldots, 8$, an observation known as Bode's Law. However, the 'law' breaks down for $N = 9$ and 10, and is widely believed to be a coincidence.

Question 5. (a) If the digits were random, then each e_i would be $10^{12}/16$; the n_i's are given in the spreadsheet. Therefore, the chi-squared statistic is

$$\chi^2 = \sum_{i=1}^{16} \frac{(n_i - e_i)^2}{e_i} \approx 7.944,$$

while $\chi_{15,0.95} \approx 25.00 > \chi^2$, hence we do not reject the null hypothesis. See spreadsheet for details.

(b) An example would be $0.\dot{1}234567890\dot{} = 0.123456789012345678901234567890\ldots$ (repeating). The digits are not random as they follow a simple pattern. Yet whenever we sample the first N digits, each digit occurs (roughly) the same number of times, the chi-squared test for the distribution of individual digits will fail to reject the null hypothesis. (However, a more sophisticated chi-squared test can detect the non-randomness: e. g. test for the occurrences of the pairs of digits 00, 01, 02, ..., 99.)

Question 6. Refer to spreadsheet for details. We treat the row and column sums as fixed, therefore there's only 1 degree of freedom (specifying the value of 1 entry in the table allows us to determine the other 3 entries). The expected values can be computed as follows: assume H_0, then the incidence of polio is independent of the group, so

$$\frac{\text{expected number of polio cases in vaccine group}}{\text{size of vaccine group}} = \frac{\text{total number of polio cases}}{\text{total size of all groups}}.$$

Thus the expected number of polio cases in the vaccine group is $200745 \times \frac{199}{401974} \approx 99.38$. The other expected numbers can be similarly calculated, leading to $\chi^2 \approx 36.12$, and a p-value of 1.86×10^{-9}. Thus (for any reasonable choice of α) we reject H_0.

Question 7. We first simplify the $\sqrt{1-r^2}$ term using $r^2 = 1-$ SSE/SST, where SSE $= (n-2)s^2$ and SST $= (n-1)s_y^2$. So we obtain

$$\sqrt{1-r^2} = \frac{\sqrt{n-2}}{\sqrt{n-1}} \frac{s}{s_y}.$$

Therefore,

$$\text{LHS} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = r\, s_y \frac{\sqrt{n-1}}{s}.$$

Recall that $r = \frac{s_{xy}}{s_x s_y}$ and $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$, so the last term simplifies to

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

Question 8. (a) Let d_i be 'after' scores − 'before' scores. Then $\bar{d} = 2.5$ and $s_d = 2.893$. Assuming H_0, the t-score is 3.865, leading to a one-sided p-value of 0.0005, thus we reject H_0.

(b) Suppose H_0 is true, then there is no statistical difference between the 'after' and 'before' scores, so let us randomly permute each person's scores and recalculate d_i. That is, for each person, with probability 1/2 we swap the two scores, and with probability 1/2 we do not swap the scores. (It does not make sense here to mix up scores from different people.)

If the scores are swapped then d_i changes sign, otherwise it doesn't. This observation leads to a short piece of code to perform the resampling (see highlighted cell in spreadsheet).

The histogram of the resample means should be roughly symmetric, with mean very close to 0. There should be very few values (0–4) greater than or equal to $\bar{d} = 2.5$, giving a p-value of 0.0 – 0.008.

Question 9. (a) See spreadsheet. The histogram is not symmetrical, and has a longer left tail.

(b) The average of the resample means should be around 26.2 (in agreement with the average of the original data).

The confidence interval can be constructed by *cutting off* 2.5% of the values from each end. Since the histogram is constructed using 1000 values, we cut off 25 values from each end, and use the range of the remaining values as our CI.

The true value does not lie within the CI (which indicates systematic errors in the experiment). The CI should look like $[23.x, 28.y]$ (for example, [23.2, 28.5]).

(c) There are 23 distinct data values, and 66 data entries. In a resample, the 1st entry can be any of the 23 values, the 2nd entry can be any of the 23 values, ..., the 66th entry can also be any of the 23 values. Therefore there are in total $23^{66} \approx 7.48 \times 10^{89}$ different resamples. (Even if you could resample at a rate of 10^{15} per second, it would take 2.37×10^{67} years to generate all the resamples. So in practice we can never come close to doing a complete resample.)

(d) The data do not seem to be normal, mainly because of the two outliers (you can verify this using a Q-Q plot). If we used a t-distribution (which would be inappropriate due to non-normality), then the CI would be [23.57, 28.85], which lies to the right of the correct CI due to these outliers.