STA2002 Probability and Statistics II
Assignment 6
Due date: 11 December 2020 5pm TC414 Assignment dropbox
Please submit an electronic copy of your answers on Blackboard too

Please work on the assignment by yourself only. We follow a strict rule on academic honesty.

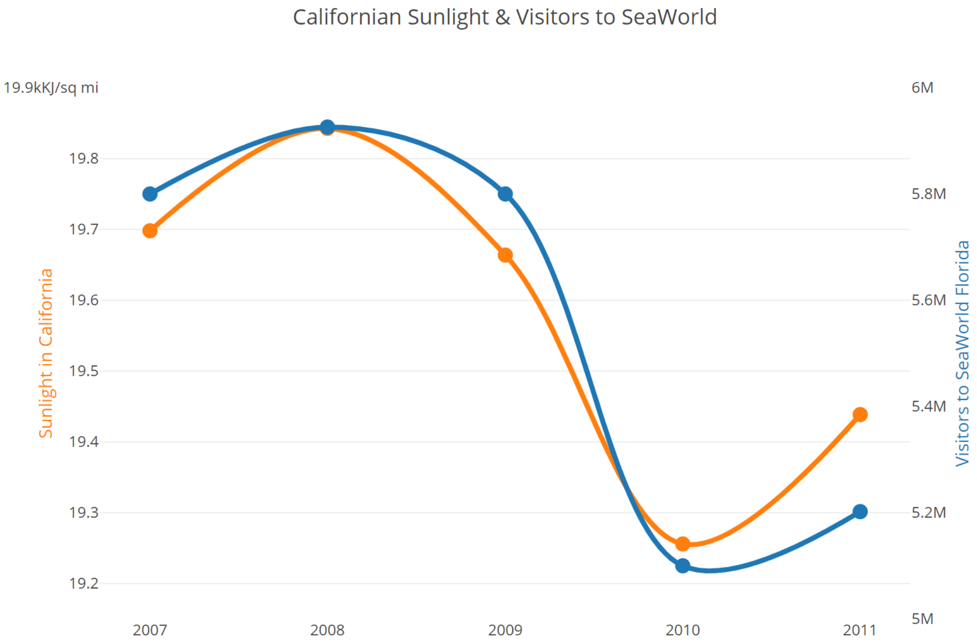
1. (Spurious correlation/Correlation does not imply causation) Sometimes, data can be misleading. When two variables are statistically correlated, it does not mean that one causes the other (causation). This phenomenon is what we call "spurious correlation" or "correlation does not imply causation".

   On the website `http://tylervigen.com/`, you can find plenty interesting examples of real-world spurious correlations.

   In this question, we shall explore one such spurious correlation. For Year 2007 to 2011, let X be the amount of California sunlight (in KJ per square meter) and Y be the number of visitors to SeaWorld Florida.

Year	2007	2008	2009	2010	2011
California Sunlight (X)	19698.04	19842.65	19663.69	19255.82	19438.55
Visitors to SeaWorld Florida (Y)	5800000	5926000	5800000	5100000	5202000

   The plot is the following:



   Obviously these two variables are not related. However, we shall see that they are statistically correlated.

(a) Compute the sample correlation coefficient r between X and Y.

(b) To test the hypothesis

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0$$

at $\alpha = 5\%$, using part (a), the test statistic $|r|$ and Table IX in the book, determine the outcome of this test.

(c) To test the hypothesis

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0$$

at $\alpha = 5\%$, using the test statistic

$$|t| = \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right|,$$

determine the outcome of this test.

(d) Test the hypothesis

$$H_0 : \rho = 0.7, \quad H_1 : \rho > 0.7$$

at $\alpha = 5\%$. Determine the outcome of this test.

2. Let $X_1, \ldots, X_{10}$ be a random sample of $n = 10$ from a normal distribution $N(0, \sigma^2)$.

(a) Find a best critical region of size 0.05 for testing

$$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 = 2.$$

(b) Deduce the power of the test in part (a), that is, compute the power function $K(2)$. Feel free to use any computing language to help you compute the power.

(c) Find a best critical region of size 0.05 for testing

$$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 = 4.$$

(d) Deduce the power of the test in part (c), that is, compute the power function $K(4)$.

(e) Find a best critical region of size 0.05 for testing

$$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 = \sigma_1^2,$$

where $\sigma_1^2 > 1$.

(f) Find a uniformly most powerful test and its critical region of size 0.05 for testing

$$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 > 1.$$

3. Consider two independent normal distributions $N(\mu_1, 400)$ and $N(\mu_2, 225)$. Let $\theta = \mu_1 - \mu_2$. Let $\overline{x}$ and $\overline{y}$ denote the observed means of two independent random samples, each of size n, from these two distributions. To test

$$H_0 : \theta = 0, \quad H_1 : \theta > 0,$$

we use the critical region

$$C = \{\overline{x} - \overline{y} \geq c\}.$$

(a) Express the power function $K(\theta)$ in terms of the standard normal distribution cdf Φ. It depends on n and c.

(b) Find n and c so that the probability of type I error is 0.05, and the power at $\theta = 10$ is 0.9, approximately. Assume that $z_{0.10} = 1.28$.

4. Exercise 9.6-6 in book.

5. Let $X_1, \ldots, X_{10}$ be a random sample of $n = 10$ from a Poisson distribution with mean θ. We want to test

$$H_0 : \theta = 0.1, \quad H_1 : \theta = 0.5.$$

(a) Prove that

$$C = \left\{ \sum_{i=1}^{10} x_i \geq 3 \right\}$$

is a best critical region.

(b) For the test in part (a), using Table III in the book determine the significance level α.

(c) For the test in part (a), using Table III in the book determine the power at $\theta = 0.5$.

6. (Pooled z-test as likelihood ratio test) Let $X_1, \dots, X_m \overset{i.i.d.}{\sim} N(\mu_1, 1)$ and $Y_1, \dots, Y_n \overset{i.i.d.}{\sim} N(\mu_2, 1)$. Suppose that these two samples are independent. We would like to test

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

using the likelihood ratio test.

(a) Prove that the likelihood function can be written as

$$L(\mu_1, \mu_2) = \frac{1}{(2\pi)^{(m+n)/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} (x_i - \mu_1)^2 \right\} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} (y_j - \mu_2)^2 \right\}.$$

(b) Prove that the maximum likelihood estimators of μ_1 and μ_2 are respectively the sample mean of X and Y, that is,

$$\widehat{\mu}_1 = \overline{X} = \frac{1}{m} \sum_{i=1}^{m} X_i, \quad \widehat{\mu}_2 = \overline{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j.$$

(c) Under $H_0 : \mu_1 = \mu_2$, let us write $\mu_1 = \mu_2 = \mu$. Prove that the maximum likelihood estimator of μ, assuming H_0 is true, is the pooled estimator

$$\widehat{\mu} = \frac{\sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j}{m+n} = \frac{m\overline{X} + n\overline{Y}}{m+n}.$$

(d) Using part (a), (b) and (c), prove that the likelihood ratio can be written as

$$\frac{\max_{\mu_1=\mu_2=\mu} L(\mu, \mu)}{\max_{\mu_1,\mu_2} L(\mu_1, \mu_2)} = \exp\left[-\frac{1}{2} \frac{mn}{(m+n)} (\bar{x} - \bar{y})^2 \right].$$

(e) Using part (d), prove that the critical region of the likelihood ratio test with significance level α is of the form

$$C = \left\{ \sqrt{\frac{mn}{m+n}} |\bar{x} - \bar{y}| \geq z_{\alpha/2} \right\}.$$

7. (CO2 and global temperature) It is commonly believed that an increasing amount of Carbon Dioxide (CO2) leads to rising global temperature, thus these two variables should be highly correlated.

On Blackboard, download the dataset "CO2_data.csv", which contains the annual CO2 data compiled by Institute for Atmospheric and Climate Science at ETH Zurich `http://www.iac.ethz.ch/` from Year 1880 to 2014.

On Blackboard, download the dataset "Temp_data.csv", which contains the annual average global surface temperature compiled by National Oceanic and Atmospheric Administration `https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php#anomalies` from Year 1880 to 2014.

For this question, feel free to use any function or package in your choice of programming language to compute the correlation, and attach your code at the end of your assignment.

(a) Compute the sample correlation coefficient r between the CO2 level and the global surface temperature.

(b) Denote by ρ to be the correlation between CO2 and global temperature. To test the hypothesis

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0$$

at $\alpha = 5\%$, using the test statistic

$$|t| = \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right|,$$

determine the outcome of this test. As n is large in this dataset, use the standard normal distribution quantile $z_{0.025}$ instead of $t_{0.025}(n-2)$.

(c) Test the hypothesis

$$H_0 : \rho = 0.8, \quad H_1 : \rho > 0.8$$

at $\alpha = 5\%$. Determine the outcome of this test.