Inference from Scientific Data, 2020 - Worksheet 3

Canvas submission deadline: Wednesday, 16 December
Marks out of a total of 20 are shown in brackets

Question 1: A sensitive probe is used to measure the displacement (in microns) of an optical component from its nominal position. Since there is a suspicion that these deviations are related to thermal effects in the instrument, the ambient temperature (in Celsius) at which each measurement is taken was also recorded. The temperature and displacement data file (called "Tdisp.dat") can be found on Canvas.

1. Test whether these data provide evidence for a relationship between displacement and ambient temperature, by calculating the correlation and the significance of correlation between the two quantities using:

   (a) the Pearson correlation coefficient, **[2]**

   (b) the Spearman rank correlation coefficient. **[2]**

2. Plot the data and comment on your results. Which of the two methods for testing for correlation do you prefer in this case? **[1]**

3. Now perform a simple linear regression on these data, deriving the least squares slope and intercept. Swap the axes and repeat the fit. How similar are your two regression lines? Plot the lines on top of the data. **[2]**

Question 2: The luminosity function (i.e. distribution in luminosity) of galaxies in the field follows a "Schechter function" form,

$$P(L)\mathrm{d}L \propto \left(\frac{L}{L_*}\right)^{\alpha} \exp\left(\frac{-L}{L_*}\right) \frac{\mathrm{d}L}{L_*}\,,$$

where we take $L_* = 1.4 \times 10^{10} L_\odot$ and $\alpha = -0.7$. A survey is sensitive to galaxies with $L > 10^9 L_\odot$.

1. What is the expectation value for galaxy luminosity in the survey? **[2]**

2. A region of space was found to contain galaxies with luminosities:

   $$[1.39, 1.40, 1.29, 5.95, 2.97, 1.63, 1.17, 2.06, 4.69, 2.48] \times 10^9 L \odot\ .$$

   Plot a cumulative distribution of the observed galaxy luminosities, along with the theoretical expectation for this distribution. **[2]**

3. Using an appropriate statistical test, are these data consistent with the theoretical expectation at the 90% level? What do you conclude? **[3]**

Question 3: Suppose we have observations $\{x_1, x_2, \ldots, x_N\}$ where each x_i is an independent random variable drawn from the same distribution $P(x)$.

1. Write down the integral definitions of the population mean μ and variance σ^2 for the distribution $P(x)$. **[1]**

2. Write down the definition of the sample mean, $\hat{\mu}$, given in lectures. Show that $\hat{\mu}$ is an unbiased estimator of the population mean. **[1]**

3. The true variance of the distribution $P(x)$ is given by

$$\sigma^2_{\text{True}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \,.$$

   The following is proposed as an estimator of the population variance,

$$\sigma^2_{\text{Est}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \,.$$

   By considering the expectation of difference $\langle \sigma^2_{\text{True}} - \sigma^2_{\text{Est}} \rangle$, or otherwise, show that σ^2_{Est} is NOT an unbiased estimator of the population variance. **[4]**

In fact using a factor of $N-1$ in the denominator, instead of N, gives an unbiased estimator of the population variance. The correction $(N-1)/N$ is known as Bessel's correction. Of course, in the limit of large N this correction factor tends to unity; the bias is only important for small N.