Math 2a Homework 8 Solutions

Problem 1. For $x = 0, 1, \ldots, 10$, we compute the likelihood ratio

$$f_0(x)/f_1(x) = \frac{\binom{10}{x} \times 0.6^x \times 0.4^{10-x}}{\binom{10}{x} \times 0.7^x \times 0.3^{10-x}} = \left(\frac{4}{3}\right)^{10} \times \left(\frac{9}{14}\right)^x.$$

Numerically, this gives

x	0	1	2	3	4	5	6	7	8	9	10
$f_0(x)/f_1(x)$	17.76	11.42	7.34	4.72	3.03	1.95	1.25	0.81	0.52	0.33	0.21

The ratio decreases as x increases. So the likelihood ratio test rejects for larger values of x. The rejection region is of the form $\{x \geq n\}$. The corresponding significance level is the probability that we reject the null hypothesis when it is true, namely, $P_{0.6}(X \geq n)$ i.e. the probability that a binomial random variable with 10 trials and a probability of success equal to .6 is greater or equal to n. The values are given below (as a function of the cut-off n).

n	significance level
10	0.006046618
9	0.046357402
8	0.167289754

Problem 2. (a) The mean lifetime of a battery is equal to

$$\int_0^\infty t f(t|\lambda)dt = \int_0^\infty \frac{t}{\lambda} e^{-t/\lambda} dt = \lambda \int_0^\infty s e^{-s} ds,$$

where we set $s = t/\lambda$. Since

$$\int_0^\infty s e^{-s} ds = -(s+1)e^{-s}|_0^\infty = 0 - (-1) = 1,$$

we see that the mean lifetime of a battery is λ.

(b) The likelihood ratio is

$$\frac{\prod_1^n f(T_i|1)}{\prod_1^n f(T_i|1.5)} = \frac{\prod_1^n (e^{-T_i/1}/1)}{\prod_1^n (e^{-T_i/1.5}/1.5)}$$
$$= 1.5^n \exp(-\sum_1^n T_i + \sum_1^n T_i/1.5) = 1.5^n \exp(-\sum_1^n T_i/3).$$

The ratio decreases as $\sum T_i = n\bar{T}$ increases. So the likelihood ratio test rejects for larger values of $\bar{T}$, which means that the rejection is of the form $\{\bar{T} \geq C\}$. The corresponding significance level is

$$P_1\left(\sum T_i \geq nC\right) = \int_{\sum t_i \geq nC} \prod f(t_i|1) dt_1 \ldots dt_n$$

$$= \int_{t_i \geq 0, \sum t_i \geq nC} e^{-\sum t_i} dt_1 \dots dt_n.$$

If we set $S = \sum t_i$ and $y_i = t_i/S$, then $dt_1 \dots dt_n = S^{n-1} dS dy_1 \dots dy_{n-1}$. And $(y_1, \dots, y_{n-1})$ runs over the region $y_i \geq 0$, $\sum_1^{n-1} y_i \leq 1$. So the above formula is equal to

$$\int_{nC}^{\infty} \int_{y_i \geq 0, \sum_1^{n-1} y_i \leq 1} S^{n-1} e^{-S} dy_1 \dots dy_{n-1} dS = \int_{nC}^{\infty} \frac{S^{n-1}}{(n-1)!} e^{-S} dS.$$

By integration by parts, we compute the significance level to be equal to $e^{-nC} \sum_{k=0}^{n-1} \frac{(nC)^k}{k!}$.

Problem 3. (a) For $i = 1, \dots, 16$, let x_i denote the number of cans of beer the i-th student drank, and y_i be the corresponding BAC number. We compute $\overline{x} = \sum_1^{16} x_i/16 = 4.8125$ and $\overline{y} = \sum_1^{16} y_i/16 = 0.07375$. Thus,

$$s_x = \sqrt{\sum_1^{16} (x_i - \overline{x})^2/(16-1)} = 2.197536;$$

$$s_y = \sqrt{\sum_1^{16} (y_i - \overline{y})^2/(16-1)} = 0.04414;$$

$$r = \frac{\sum_1^{16} (x_i - \overline{x}) \times (y_i - \overline{y})}{(16-1) \times s_x \times s_y} = 0.894338.$$

It then follows that the slope of the regression line is

$$b_1 = r \frac{s_y}{s_x} = 0.017964$$

and the intercept is

$$b_0 = \overline{y} - b_1 \overline{x} = -0.0127.$$

Finally, $r^2 = 0.799841$ and the equation of the regression line is given by

$$y = -0.0127 + 0.017964x.$$

(b) Let ρ denote the correlation between x and y. We will test

$$H_0 : \rho = 0 \text{ versus } H_a : \rho > 0.$$

Compute the t-statistic:

$$t = \frac{r\sqrt{16-2}}{\sqrt{1-r^2}} = 7.479592.$$

In terms of a random variable T having $t(16-2)$ distribution, the P-value for the test is

$$P(T \geq t) = 1.48 \times 10^{-6},$$

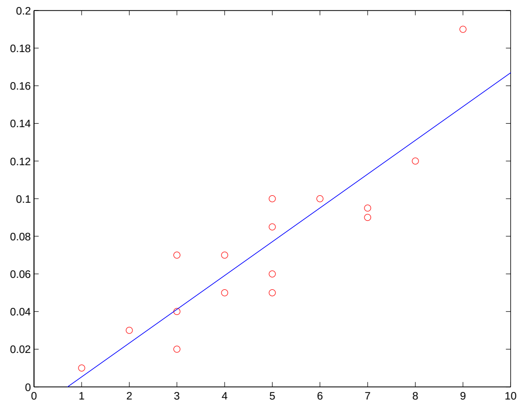

Figure 1: Regression plot

which is very small. The conclusion is that there is very strong evidence that drinking more beers increases blood alcohol.

Problem 4. (a) Let $y =$ interval until the next eruption and $x =$ duration of an eruption, both in minutes, for all eruptions of Old Faithful Geyser over the 8-day period. The general linearity of the scatter plot of time-between(y) versus duration (x) suggests use of a simple linear regression model,

$$y = \beta_0 + \beta_1 x + \text{ error}$$

The fitted model is

$$y = b_0 + b_1 x,$$

where

$$b_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = 10.3582,$$
$$b_0 = \bar{y} - b_1\bar{x} = 33.9668.$$

For reference, $\bar{x} = 3.57613$ and $\bar{y} = 71.0090$.

(b) We have made $n = 222$ observations. Suppose we have just observed $x = d$, we want to predict y. Note that the **mean value** of the interval until the next eruption is $\mathbf{E}y = \beta_0 + \beta_1 d$, so we use $b_0 + b_1 d$ as a (least square) estimator of $\mathbf{E}y$. In class, we showed that

$$t = \frac{b_0 + b_1 d - (\beta_0 + \beta_1 d)}{s\sqrt{\frac{1}{n} + \frac{(d-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}$$

is a student's t distribution with $n - 2 = 220$ degrees of freedom; here,

$$s = \frac{RSS}{n-2} = \frac{\sum(y_i - b_0 - b_1 x_i)^2}{n-2}$$

is our estimate of the std deviation of the errors. Therefore, a 95% confidence interval for $\mathbf{E}y$ is

$$b_0 + b_1 d \pm s\sqrt{\frac{1}{n} + \frac{(d-\bar{x})^2}{\sum(x_i - \bar{x})^2}} \cdot 1.9708,$$

where 1.9708 is such that $P(|T| \leq 1.9708) = 95\%$, where T has $t(220)$ distribution.
As an aside, recall the key intermediate results

$$\mathbf{E}(b_0 + b_1 d) = \beta_0 + \beta_1 d, \quad \text{Var}(b_0 + b_1 d) = \sigma^2 \Big[\frac{1}{n} + \frac{(d - \bar{x})^2}{\sum(x_i - \bar{x})^2}\Big].$$

(c) Let $y = \beta_0 + \beta_1 d + \epsilon$ be a new observation. In this case,

$$y - (b_0 + b_1 d) = \epsilon + (\beta_0 + \beta_1 d) - (b_0 + b_1 d)$$

Thus $y - \Big(b_0 + b_1 d\Big)$ is a normal random variable with mean 0 and variance $\sigma^2 \Big[1 + \frac{1}{n} + \frac{(d-\bar{x})^2}{\sum(x_i - \bar{x})^2}\Big]$.
As in part (b), we have

$$\frac{y - \Big(b_0 + b_1 d\Big)}{s\sqrt{1 + \frac{1}{n} + \frac{(d-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}$$

is a student's t distribution with 220 degrees of freedom. Therefore, a 95% confidence **prediction interval** for y is

$$b_0 + b_1 d \pm s\sqrt{1 + \frac{1}{n} + \frac{(d - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \cdot 1.9708$$

(d) We compute $s = 6.15853$. For the case $d = 4.0$, $b_0 + b_1 d = 75.39958$, and

$$s\sqrt{\frac{1}{n} + \frac{(d - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \cdot 1.9708 = 0.874932,$$

$$s\sqrt{1 + \frac{1}{n} + \frac{(d - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \cdot 1.9708 = 12.16902.$$

So the numerical answers to (b) and (c) are $[74.52465, 76.27451]$ and $[63.23056, 87.56860]$, respectively.

Appendix. There are a total of 222 observations in total in the dataset. Some of them are not in the 8-day period. If we used the whole dataset, we would obtain the following results:
The fitted line is

$$y = 33.96676 + 10.35821x.$$

When $d = 4$, a 95% confidence interval for the mean is

$$75.39958 \pm 0.87493 = [74.52465, 76.27451],$$

and a 95% prediction interval for y is

$$75.39958 \pm 12.16877 = [63.23081, 87.56835].$$

Some intermediate results are: $\bar{x} = 3.57613$, $\bar{y} = 71.00901$, $s = 6.15853$, and the C such that $P(|T| \leq C) = 95\%)$ for a t distribution with 220 degrees of freedom is $C = 1.97081$.