Probability and Statistics for JMC
Solutions 6 — Estimation

1. If $(X_1, \ldots, X_n)$ are a random sample from an exponential distribution with rate parameter λ, find the maximum likelihood estimate for λ.

   The pdf for a single observation X_i is $f(x) = \lambda e^{-\lambda x}$ so the log-likelihood function is

$$\begin{aligned}\ell(\lambda \,|\, x_1, \ldots, x_n) &= \log\left(\prod_{i=1}^{n} f(x_i)\right) \\ &= \sum_{i=1}^{n} (\log(\lambda) - \lambda x_i) \\ &= n\log(\lambda) - n\lambda\bar{x},\end{aligned}$$

   where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean. Taking the derivative of $\ell(\lambda)$ wrt λ gives $d\ell/d\lambda = \frac{n}{\lambda} - n\bar{x}$. Setting it to zero gives the MLE $\hat{\lambda} = 1/\bar{x}$.
   To check it's a maximum we need the second derivative of $\ell(\lambda)$: $d^2\ell/d\lambda^2 = -n\lambda^{-2}$, which is negative when $\lambda = \hat{\lambda}$.

2. Derive the maximum likelihood estimate for λ for n independent samples from Poisson(λ).

   The pmf for a single observation X_i is $p(x) = e^{-\lambda}\lambda^x/x!$ so the log-likelihood is

$$\begin{aligned}\ell(\lambda \,|\, x_1, \ldots, x_n) &= \log\left(\prod_{i=1}^{n} p(x_i)\right) \\ &= \sum_{i=1}^{n} (-\lambda + x_i\log(\lambda) - \log(x_i!)) \\ &= -n\lambda + n\bar{x}\log(\lambda) + \text{const.}\end{aligned}$$

   The derivative wrt to λ is $d\ell/d\lambda = -n + n\bar{x}/\lambda$. Setting this to zero gives the MLE $\hat{\lambda} = \bar{x}$.
   To check it's a maximum take the second deriv of the log-likelihood: $d^2\ell/d\lambda^2 = -n\bar{x}\lambda^{-2}$, which is negative for $\lambda = \hat{\lambda}$.
   For the case where all the x_i are zero, we have $\ell(\lambda) = -n\lambda$, which clearly is a maximum at $\lambda = 0 = \bar{x}$ (restricting to $\lambda > 0$, which it must be as the rate param of a Poisson distribution). So the result holds in this case as well.

3. In a study of traffic congestion, data were collected on the number of occupants in private cars on a certain road. These data, collected for 1469 cars, are given below

Count	1	2	3	4	5	≥ 6
Frequency	902	403	106	38	16	4

One theory suggests that these data may have arisen from a modified geometric distribution, in which the probability that there are x occupants in a car is

$$p(x) = p(1-p)^{x-1},\ x = 1, 2, \ldots$$

(a) Find the maximum likelihood estimate of the parameter p of the geometric distribution for these data. (Note that $\mathrm{P}(X \geq x) = (1-p)^{x-1}$.)

We need the probability that out of 1469 observations, $n_1 = 902$ of them are $x = 1$, $n_2 = 403$ of them are $x = 2$, ..., $n_5 = 16$ of them are $x = 5$ and $n_{\geq 6} = 4$ of them are greater than or equal to 6. The probability of this is

$$L(p \,|\, n_1, n_2, n_3, n_4, n_5, n_{\geq 6}) \propto p(1)^{n_1}\, p(2)^{n_2}\, p(3)^{n_3}\, p(4)^{n_4}\, p(5)^{n_5}\, \left((1-p)^{6-1}\right)^{n_{\geq 6}},$$

where the constant of proportionality is some multinomial coefficient with a bunch of factorials to account for all the ways of choosing 1469 observations such that 902 of them are 1, 403 of them are 2, etc. But the point is that this constant of proportionality doesn't depend on p so it does not affect our MLE calculation.
The log-likelihood is therefore

$$\begin{aligned}\ell(p \,|\, \{n_i\}) &= \log\left[\left(\prod_{i=1}^{5} p(i)^{n_i}\right)\left((1-p)^{5n_{\geq 6}}\right)\right] \\ &= 5n_{\geq 6}\log(1-p) + \sum_{i=1}^{5} n_i\left(\log(p) + (i-1)\log(1-p)\right).\end{aligned}$$

Taking the derivative wrt to p gives

$$\begin{aligned}\frac{d\ell}{dp} &= -\frac{5n_{\geq 6}}{1-p} + \sum_{i=1}^{5} n_i\left(\frac{1}{p} - \frac{i-1}{1-p}\right) \\ &= -\frac{5n_{\geq 6}}{1-p} + \frac{\sum_{i=1}^{5} n_i}{p} - \frac{\sum_{i=1}^{5} n_i(i-1)}{1-p} \\ &= \frac{\sum_{i=1}^{5} n_i}{p} - \frac{\sum_{i=1}^{6} n_i(i-1)}{1-p}.\end{aligned}$$

Setting this to zero and solving for p gives

$$(1-\hat{p})(n_1 + n_2 + n_3 + n_4 + n_5) = \hat{p}(n_2 + 2n_3 + 3n_4 + 4n_5 + 5n_{\geq 6})$$

or

$$\hat{p} = \frac{n_1 + n_2 + n_3 + n_4 + n_5}{n_1 + 2n_2 + 3n_3 + 4n_4 + 5n_5 + 5n_{\geq 6}}.$$

Substituting in the values from the table (i.e. $n_1 = 902$, $n_2 = 403$, etc) gives $\hat{p} = 0.643$.
To check whether $\hat{p}$ is a maximum of $\ell(p)$ take another derivative wrt p to get

$$\frac{d^2\ell}{dp^2} = -\frac{\sum_{i=1}^{5} n_i}{p^2} - \frac{\sum_{i=1}^{6} n_i(i-1)}{(1-p)^2} < 0.$$

(b) *[To be attempted after the lectures on hypothesis testing]* Describe how a hypothesis test could be carried out, at the 1% level, to see if these data do come from a geometric distribution.

Consider the data as a histogram of 6 bins with heights $n_1, n_2, \ldots n_{\geq 6}$. If the MLE $\hat{p} = 0.643$ were the true value of p then the expected heights of

the bins would be $E_i = 1469p_i$, where $p_i = \hat{p}(1-\hat{p})^{i-1}$ for $i = 1, \ldots, 5$ and $p_{\geq 6} = (1-\hat{p})^5$.

Form the χ^2 test statistic $\chi^2 = \sum_{\text{bins } i} \frac{(O_i - E_i)^2}{E_i}$, where $O_i = n_i$, the observed counts in each bin. If $\hat{p}$ were the true value of p then χ^2 should be approximately be distributed as a χ^2 RV with $6 - 1 - 1 = 4$ degrees of freedom (subtract 1 because of the fit to the parameter p and the usual additional 1 because the histogram bins are not independent since they must add to 1469). The 0.99-quantile of the $\chi^2(4)$ distribution, is at $\chi^2 = 13.28$. Thus if the observed χ^2 is greater than 13.28 we reject the hypothesis that the data came from a geometric distribution with $p = \hat{p}$ because if the hypothesis were true there is less than a 1% chance of observing χ^2 to be so large.

4. (a) For a random sample of size n from a normal distribution with unknown mean μ and known variance σ^2, what is the confidence level for for each of the following confidence intervals for μ?

   i. $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

   ii. $\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}$

   iii. $\bar{x} \pm 2.575\frac{\sigma}{\sqrt{n}}$

   iv. $\bar{x} \pm 0.99\frac{\sigma}{\sqrt{n}}$

   An α-confidence level interval for μ is $\bar{x} \pm \epsilon_\alpha \frac{\sigma}{\sqrt{n}}$, where the area under a standard normal pdf between $x = -\epsilon_\alpha$ and $x = +\epsilon_\alpha$ is α. In other words, $\alpha = 1 - 2\Phi(-\epsilon_\alpha)$. E.g. for $\epsilon_\alpha = 1.96$, $1 - 2\Phi(-1.96) = 0.95$ so $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is a 95% confidence interval for μ. Likewise, $\epsilon_\alpha = 1.645$ corresponds to a 90% CI, 2.575 to 99%, and 0.99 to 68%.

   (b) A random sample of 64 observations from a population produced the following summary statistics:

$$\sum_i x_i = 700 \qquad \sum_i (x_i - \bar{x})^2 = 4238.$$

   Find a 95% confidence interval for μ, and interpret this interval.

   The sample mean is $\bar{x} = \frac{700}{64} = 10.94$ and an estimate of the population standard deviation σ is $s = \sqrt{\frac{4238}{64}} = 8.14$. Therefore, the sample mean statistic is approximately normally distributed with a mean of μ and a standard deviation of $s/\sqrt{64}$. An approximate 95% confidence interval for μ is then $\bar{x} \pm 1.96\frac{s}{\sqrt{64}} \approx 10.9 \pm 2$.

5. Compute confidence intervals at the 95% level for the means of the distributions from which the following sample values were obtained:

   (a) $n = 100, \quad \sum_i x_i = 250, \quad \sum_i x_i^2 = 725000$

   This is a situation where we don't know the true variance so in principle we need the Student-t distribution. The confidence interval for the mean is $\overline{x} \pm t_{n-1,1-\alpha/2} \frac{s_{n-1}}{\sqrt{n}}$, where $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student-t distribution with $n-1$ degrees of freedom. However, with $n = 100$ samples the Student-t will be extremely close to a standard normal so we'll calculate quantiles using N(0,1). In our case $\alpha = 0.05$, i.e. we need to find the value

a such that there's a 95% probability a standard normal variable is between $-a$ and $+a$. Using our standard normal tables this is 1.96.

We need the sample mean $\overline{x} = \frac{1}{n}\sum x_i = 250/100 = 2.5$, as well as the unbiased estimate of the variance $s^2_{n-1} = \frac{1}{n-1}\sum(x_i - \overline{x})^2$. To get s^2_{n-1} from what we are given we need to compute it as $s^2_{n-1} = \dfrac{1}{n-1}\sum x_i^2 - \dfrac{n}{n-1}\overline{x}^2$ (get this either by expanding out the square or by noticing that $s^2_{n-1} = \frac{n}{n-1}s^2$, where s^2 is the biased estimate of the variance with the n in the denominator, along with $s^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2$). This gives $s_{n-1} = \sqrt{\dfrac{725000}{99} - \dfrac{100}{99}(2.5)^2} = 85.54$.

The resulting CI is then $2.5 \pm 1.96\dfrac{85.54}{\sqrt{100}} = 2.5 \pm 16.77$.

(b) $n = 100, \quad \bar{x} = 83.2, \quad s_{n-1} = 6.4$

Exactly the same as above. The CI for the mean is,

$$\overline{x} \pm 1.96\frac{s_{n-1}}{\sqrt{n}} = 83.2 \pm 1.96\frac{6.4}{\sqrt{100}} = 83.2 \pm 1.25$$

.

6. The following random sample was selected from a normal distribution:

7.53, 4.35, 7.66, 7.54, 5.83, 1.92, 3.14, 4.41

(a) Construct a 90% confidence interval for the population mean.

We do not know the variance so we must estimate it from the data as s^2_{n-1} and use a Student-t quantile to calculate the confidence interval. The sample mean is $\bar{x} = 5.2975$ and the unbiased estimator of the variance is $s^2_{n-1} = 4.80365$. We need the quantile from a Student-t distribution with $\nu = n - 1 = 7$ degrees of freedom. For a 90% CI we need the quantile such that there is a probability of 5% in each of the two tails, i.e. we want the 0.95-quantile. The table gives us $t_{7,0.95} = 1.89$.

The 90% CI for the mean is,

$$5.30 \pm 1.89\frac{\sqrt{4.80365}}{\sqrt{8}} = 5.30 \pm 1.46,$$

or [3.83, 6.76].

(b) Construct a 99% confidence interval for the population mean.

The only thing we need to change is the quantile of the Student(7) distribution. We need $t_{7,0.995} = 3.50$. Therefore, the 99% CI for the mean is,

$$5.30 \pm 3.50\frac{\sqrt{4.80365}}{\sqrt{8}} = 5.30 \pm 2.71,$$

or [2.59, 8.01].