

Solutions to Assignment #4

1.

Solution: The data $s = (r, p, w)$ of size $n = r + p + w$ is assumed to be multinomial with a single parameter $\theta \in \Omega = (0, 1)$ that specifies all of the category probabilities, θ^2 , $2\theta(1 - \theta)$, and $(1 - \theta)^2$, that individual flowers are red, pink, or white, respectively. We observe $s = (35, 156, 209)$.

(a) To find the MLE, we can follow the standard approach.

$$\begin{aligned}
 L(\theta | s) &= \binom{n}{r, p, w} \theta^{2r} (2\theta(1 - \theta))^p (1 - \theta)^{2w} \\
 \ell(\theta | s) = \log L(\theta | s) &= \log \binom{n}{r, p, w} + 2r \log(\theta) + p \log(2\theta(1 - \theta)) + 2w \log(1 - \theta) \\
 &= \log \binom{n}{r, p, w} + (2r + p) \log(\theta) + p \log(2) + (2w + p) \log(1 - \theta) \\
 \frac{\partial \ell(\theta | s)}{\partial \theta} &= \frac{2r + p}{\theta} - \frac{2w + p}{1 - \theta} = 0 \\
 \hat{\theta} &= \frac{2r + p}{2(r + p + w)} = \frac{2r + p}{2n}
 \end{aligned}$$

For the observed data

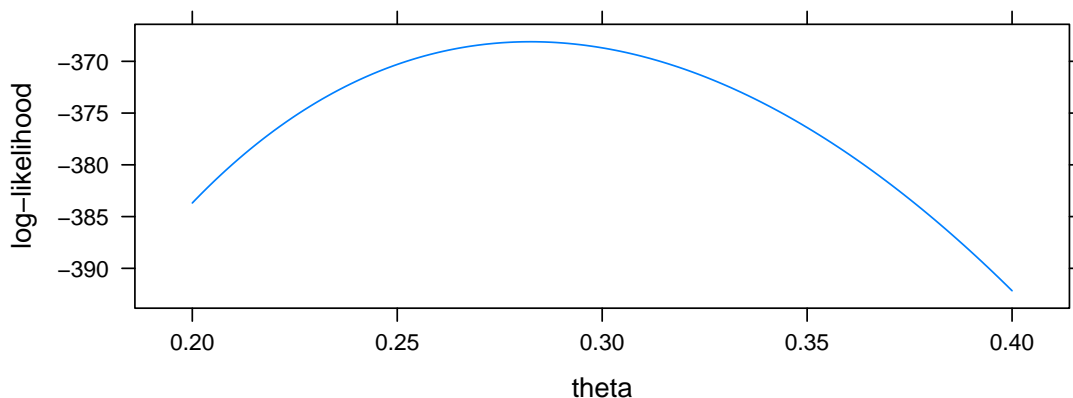
$$\hat{\theta} = \frac{2 \cdot 35 + 156}{2(35 + 156 + 209)} = \frac{226}{800} \doteq 0.2825.$$

A plot of the loglikelihood verifies this point estimate.

```

> library(lattice)
> source("../R/stat310.R")
> likelihoodPlot(c(0.2, 0.4), c(35, 156, 209), hardy.weinberg.logl)

```



Here is one possible specification of `hardy.weinberg.logl()`.

```

hardy.weinberg.logl = function(theta, s, log.like=T, ...) {
  n = sum(s) # = r + p + w
  r = s[1] # = #red flowers in sample
  p = s[2] # = #pink flowers in sample

```

```

w = s[3] # = #white flowers in sample
logl = r*2*log(theta) + p*log(2*theta*(1-theta)) + w*2*log(1-theta)
if(log.like)
  return(logl)
else
  return(exp(logl))
}

```

(b) Repeat the previous problem where now the likelihood is a function of two parameters, $\theta = (\theta_1, \theta_2)$ with category probabilities θ_1 , θ_2 , and $1 - \theta_1 - \theta_2$.

The set $\Omega \subset \mathbb{R}^2$ is the triangle bounded by the lines $\theta_1 = 0$, $\theta_2 = 0$, and $\theta_1 + \theta_2 = 1$.

The likelihood is

$$L(\theta | s) = \binom{n}{r, p, w} \theta_1^r \theta_2^p (1 - \theta_1 - \theta_2)^w .$$

To find the MLE, we take partial derivatives of $\log L(\theta | s)$ with respect to both θ_1 and θ_2 , set both equations equal to 0, and solve.

$$\begin{aligned} \ell(\theta | s) &= \log \binom{n}{r, p, w} + r \log(\theta_1) + p \log(\theta_2) + w \log(1 - \theta_1 - \theta_2) \\ \frac{\partial \ell(\theta | s)}{\partial \theta_1} &= \frac{r}{\theta_1} - \frac{w}{1 - \theta_1 - \theta_2} = 0 \\ \frac{\partial \ell(\theta | s)}{\partial \theta_2} &= \frac{p}{\theta_2} - \frac{w}{1 - \theta_1 - \theta_2} = 0 \end{aligned}$$

The difference of these equations and rearranging gives

$$r\theta_2 = p\theta_1$$

while summing them and rearranging results in

$$\frac{r}{\theta_1} + \frac{p}{\theta_2} = \frac{2w}{1 - \theta_1 - \theta_2} .$$

The first equation implies $\theta_2 = p\theta_1/r$ (provided $r > 0$ — if so, then $\hat{\theta}_1 = 0$), and substitution of this into the second equation gives

$$\frac{r}{\theta_1} + \frac{p}{p\theta_1/r} = \frac{2w}{1 - \theta_1 - p\theta_1/r}$$

which simplifies to

$$\begin{aligned} \frac{r}{\theta_1} &= \frac{w}{1 - \theta_1(r+p)/r} \\ r - (r+p)\theta_1 &= w\theta_1 \\ \hat{\theta}_1 &= \frac{r}{n} \end{aligned}$$

where $n = r + p + w$. Substitution into an earlier expression shows that $\hat{\theta}_2 = p/n$.

Numerically, we have the solutions $\hat{\theta}_1 \doteq 0.0875$ and $\hat{\theta}_2 \doteq 0.39$.

It is interesting to compare the two models probabilities for flower color.

Model	Red	Pink	White
Hardy-Weinberg	0.0798	0.4054	0.5148
General	0.0875	0.39	0.5225

2. Measurements in centimeters are assumed to be an i.i.d. normal sample. The sample ($n = 10$) is 4.7, 5.5, 4.4, 3.3, 4.6, 5.3, 5.2, 4.8, 5.7, and 5.3.

Solution: Begin by computing the sample mean and sample standard deviation. I will use R.

```
> hw4.3 = c(4.7, 5.5, 4.4, 3.3, 4.6, 5.3, 5.2, 4.8, 5.7, 5.3)
> x.bar = mean(hw4.3)
> s = sd(hw4.3)
> se = s/sqrt(10)
> c(x.bar, s, se)
```

```
[1] 4.8800000 0.6957011 0.2200000
```

- (a) Assuming the base distribution is $N(\mu, \sigma_0^2)$ where $\sigma_0^2 = 0.5$, a 90% confidence interval for μ is

$$4.88 \pm 1.6449 \times \frac{\sqrt{(0.5)}}{\sqrt{10}} \quad \text{or} \quad 4.88 \pm 0.37 \quad \text{or} \quad (4.51, 5.25)$$

We are 90% confident that the mean of the underlying normal distribution is between 4.51 and 5.25 cm. The z quantile is the 0.95 quantile of the standard normal distribution.

```
> qnorm(0.95)
```

```
[1] 1.644854
```

- (b) If we assume that the distribution for a single measurement is $N(\mu, \sigma^2)$ where σ^2 is unknown, find a 90% confidence interval for μ .

Under these assumptions, the multiplier comes from a t distribution and we use s and not $\sqrt{0.5}$ in the expression for the SE.

$$4.88 \pm 1.8331 \times 0.22 \quad \text{or} \quad 4.88 \pm 0.4 \quad \text{or} \quad (4.48, 5.28)$$

We are 90% confident that the mean of the underlying normal distribution is between 4.48 and 5.28 cm. The t quantile is the 0.95 quantile of the t_9 distribution.

```
> qt(0.95, 9)
```

```
[1] 1.833113
```

Note that it is generally good practice to round estimates (such as the mean) to one more digit of accuracy than the original data and to round the margin of error to the same accuracy. Another good rule of thumb is to round the margin of error to two significant digits and then round the estimate to the same accuracy.

3. Suppose that $X_1, \dots, X_n \sim \text{Exponential}(\lambda)$.

- (a) Find formulas for the MLE and the standard error of the MLE.

Solution: This follows the derivations in lecture.

$$\begin{aligned} L(\lambda | s) &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \\ \ell(\lambda | s) = \log L(\lambda | s) &= n \log \lambda - \lambda \sum_{i=1}^n x_i \\ \frac{\partial \ell(\lambda | s)}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^n x_i} = 1/\bar{x} \end{aligned}$$

The standard error is the square root of the mean square error. In lecture, I derived the MSE by finding the bias and the variance. Here, I will derive it directly. The derivation depends on the fact that $\bar{X} \sim \text{Gamma}(n, n\lambda)$. Rather than computing integrals directly, the approach is to multiply by constants to obtain known gamma densities which integrate to one.

$$\begin{aligned}
 \text{MSE}(\hat{\lambda}) &= \int_0^\infty \left(\frac{1}{t} - \lambda\right)^2 \frac{(n\lambda)^n}{\Gamma(n)} t^{n-1} e^{-n\lambda t} dt \\
 &= \int_0^\infty (t^{-2} - 2\lambda t^{-1} + \lambda^2) \frac{(n\lambda)^n}{\Gamma(n)} t^{n-1} e^{-n\lambda t} dt \\
 &= \frac{(n\lambda)^n}{\Gamma(n)} \left(\int_0^\infty t^{(n-2)-1} e^{-n\lambda t} dt - 2\lambda \int_0^\infty t^{(n-1)-1} e^{-n\lambda t} dt + \lambda^2 \int_0^\infty t^{n-1} e^{-n\lambda t} dt \right) \\
 &= \frac{(n\lambda)^n}{\Gamma(n)} \left(\frac{\Gamma(n-2)}{(n\lambda)^{n-2}} - 2\lambda \frac{\Gamma(n-1)}{(n\lambda)^{n-1}} + \lambda^2 \frac{\Gamma(n)}{(n\lambda)^n} \right) \\
 &= \frac{(n\lambda)^2}{(n-1)(n-2)} - 2\lambda \frac{n\lambda}{n-1} + \lambda^2 \\
 &= \frac{n^2\lambda^2 - 2n(n-2)\lambda^2 + (n-1)(n-2)\lambda^2}{(n-1)(n-2)} \\
 &= \frac{(n^2 - 2n^2 + 4n + n^2 - 3n + 2)\lambda^2}{(n-1)(n-2)} \\
 &= \frac{(n+2)\lambda^2}{(n-1)(n-2)}
 \end{aligned}$$

Therefore, the standard error is $\lambda\sqrt{(n+2)/((n-1)(n-2))}$.

- (b) Let $s = (3.96, 9.74, 3.50, 4.14, 1.73, 1.88, 18.49, 5.12)$ be modeled as the realization of an i.i.d. random sample from an Exponential distribution. Use the large sample MLE approximation $\hat{\lambda} \pm z\text{SE}(\hat{\lambda})$ where z is the quantile from a normal curve to construct an approximate 95% confidence interval for λ .

Solution: For this sample, $\bar{x} = 6.07$ so that $\hat{\lambda} \doteq 0.165$. The normal quantile is $z = 1.96$ and the standard error is estimated to be $0.165\sqrt{10/72} \doteq 0.08$. An approximate 95% confidence interval is 0.165 ± 0.158 . We are 95% confident that $0.007 < \lambda < 0.322$.

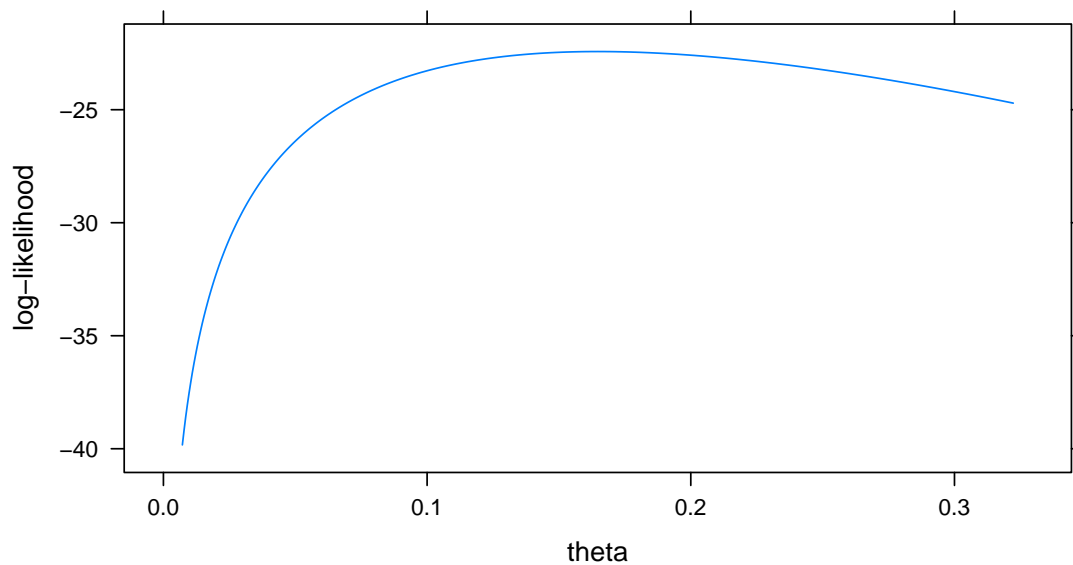
- (c) For the data in the previous part, graph the log likelihood versus lambda and mark the points on the curve at the boundaries of the confidence interval. Is the confidence interval of the form $\{\lambda \in \Omega : \log L(\lambda | s) \geq c\}$? Briefly explain.

Solution:

```

> s = c(3.96, 9.74, 3.5, 4.14, 1.73, 1.88, 18.49, 5.12)
> n = length(s)
> s.mean = mean(s)
> lambda.hat = 1/s.mean
> z = qnorm(0.975)
> se = lambda.hat * sqrt((n + 2)/((n - 1) * (n - 2)))
> a = lambda.hat - z * se
> b = lambda.hat + z * se
> likelihoodPlot(c(a, b), s, exponential.logl)

```



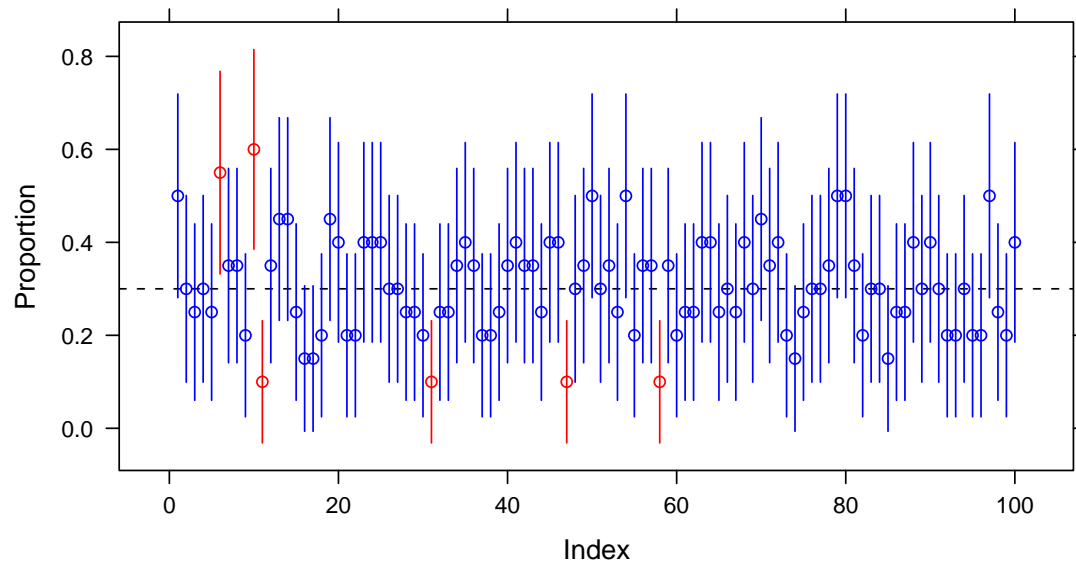
The confidence interval does not contain all λ with likelihood above some constant since the log likelihood evaluated the endpoints of the confidence interval are not equal. This is not too surprising since the sample size $n = 8$ is not large, and the sampling distribution of $\hat{\lambda}$ may not be close to normal. A future homework assignment will explore this in more detail.

4. Use the `binomialSimulation()` in `stat310.R` to simulate 100 95% confidence intervals for θ from a `Binomial(20, 0.3)` model and plot the intervals. How many contain the true value of θ ?

Solution:

```
> set.seed(14234)
> binomialSimulation(100, 20, 0.3)
```

```
[1] "94 of 100 confidence intervals contain theta = 0.3"
```



Your answer may differ.

5. Repeat the previous problem for a sample of 10,000, but do not plot. How many intervals contain the true value of θ ?

Solution:

```
> binomialSimulation(10000, 20, 0.3, plot = F)
```

```
[1] "9473 of 10000 confidence intervals contain theta = 0.3"
```