

Tutorial 2 Solutions

Question 1

(9.78) Let Y_1, Y_2, \dots, Y_n be a random sample from a power family distribution with parameters α and $\theta = 3$. Then for $\alpha > 0$,

$$f(y|\alpha) = \begin{cases} \frac{\alpha y^{\alpha-1}}{3^\alpha}, & 0 \leq y \leq 3 \\ 0, & \text{otherwise} \end{cases}$$

(a) Show that $\mathbf{E}(Y_1) = 3\alpha/(\alpha + 1)$.

$$\begin{aligned} \mathbf{E}(Y_1) &= \int_0^3 y \cdot \frac{\alpha y^{\alpha-1}}{3^\alpha} dy \\ &= \frac{\alpha}{3^\alpha} \int_0^3 y^\alpha dy \\ &= \frac{\alpha}{3^\alpha} \cdot \frac{1}{\alpha + 1} y^{\alpha+1} \Big|_{y=0}^{y=3} \\ &= \frac{\alpha}{\alpha + 1} \cdot \frac{3^{\alpha+1}}{3^\alpha} \\ &= \frac{3\alpha}{\alpha + 1} \end{aligned}$$

(b) Derive the method of moments estimator for α .

To find the method of moments estimator for α , we set the first population moment equal to the first sample moment and solve for $\hat{\alpha}$. The first sample moment is given by:

$$\frac{1}{n} \sum_{i=1}^n Y_i,$$

which is the sample average. As such, for the sake of readability, I will replace the above quantity with \bar{Y} in the second step below.

$$\mathbf{E}(Y_1) = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$\frac{3\hat{\alpha}}{\hat{\alpha} + 1} = \bar{Y}$$

$$3\hat{\alpha} = \bar{Y}(\hat{\alpha} + 1)$$

$$3\hat{\alpha} = \bar{Y}\hat{\alpha} + \bar{Y}$$

$$3\hat{\alpha} - \bar{Y}\hat{\alpha} = \bar{Y}$$

$$(3 - \bar{Y})\hat{\alpha} = \bar{Y}$$

$$\hat{\alpha} = \frac{\bar{Y}}{3 - \bar{Y}}$$

Question 2

(9.84) A certain type of electronic component has a lifetime Y (in hours) with probability density function given by:

$$f(y|\theta) = \begin{cases} \frac{1}{\theta^2} ye^{-y/\theta}, & y > 0 \\ 0, & \text{otherwise} \end{cases}$$

That is, Y has a gamma distribution with parameters $\alpha = 2$ and θ . Let $\hat{\theta}$ denote the MLE of θ . Suppose that three such components, tested independently, had lifetimes (in hours):

$$120 \quad 130 \quad 128.$$

- (a) Find the MLE of θ and provide an estimate using the given data.

The likelihood function is given by:

$$\begin{aligned} \mathcal{L}(\theta|\mathbf{y}) &= \prod_{i=1}^n f(y_i|\theta) \\ &= \frac{1}{\theta^{2n}} \left(\prod_{i=1}^n y_i \right) \exp \left\{ -\sum_{i=1}^n y_i/\theta \right\} \end{aligned}$$

The log-likelihood function is given by:

$$\begin{aligned} \ell(\theta|\mathbf{y}) &= \log(\mathcal{L}(\theta|\mathbf{y})) \\ &= -2n \log(\theta) + \sum_{i=1}^n \log(y_i) - \frac{1}{\theta} \sum_{i=1}^n y_i \end{aligned}$$

The first and second derivatives of the log-likelihood function are:

$$\begin{aligned} \frac{d}{d\theta} \ell(\theta|\mathbf{y}) &= -\frac{2n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n y_i \\ \frac{d^2}{d\theta^2} \ell(\theta|\mathbf{y}) &= \frac{2n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n y_i \end{aligned}$$

From first year calculus, recall that the extrema of a function occur when its first derivative equals zero, i.e.

$$\begin{aligned}\frac{d}{d\theta}\ell(\theta|\mathbf{y})\Big|_{\theta=\hat{\theta}} &= 0 \\ -\frac{2n}{\hat{\theta}} + \frac{1}{\hat{\theta}^2}\sum_{i=1}^n y_i &= 0 \\ -2n\hat{\theta} + \sum_{i=1}^n y_i &= 0 \\ \hat{\theta} &= \frac{\sum_{i=1}^n y_i}{2n}\end{aligned}$$

We should verify that $\ell(\hat{\theta}|\mathbf{y})$ is indeed a maximum.

$$\begin{aligned}\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{y})\Big|_{\theta=\hat{\theta}} &= \frac{2n}{\left(\frac{\sum_{i=1}^n y_i}{2n}\right)^2} - \frac{2}{\left(\frac{\sum_{i=1}^n y_i}{2n}\right)^3}\sum_{i=1}^n y_i \\ &= \frac{(2n)^3}{\left(\sum_{i=1}^n y_i\right)^2} - \frac{2(2n)^3}{\left(\sum_{i=1}^n y_i\right)^2} \\ &= -\frac{(2n)^3}{\left(\sum_{i=1}^n y_i\right)^2}\end{aligned}$$

The above quantity is less than zero, so $\hat{\theta}$ is indeed a maximum. Therefore,

$$\hat{\theta} = \frac{\sum_{i=1}^n Y_i}{2n} = \frac{\bar{Y}}{2}$$

is a MLE for θ .

Using the provided data, an estimate of θ is:

$$(120 + 130 + 128)/(2 \cdot 3) = 63.$$

- (b) Find $\mathbf{E}(\hat{\theta})$ and $\mathbf{Var}(\hat{\theta})$. Is $\hat{\theta}$ an unbiased estimator of θ ?

Recall that if $Y_i \sim \text{Gamma}(\alpha = 2, \theta)$, then

$$\mathbf{E}(Y_i) = 2\theta, \quad \mathbf{Var}(Y_i) = 2\theta^2.$$

$$\begin{aligned}\mathbf{E}(\hat{\theta}) &= \mathbf{E}\left(\frac{\sum_{i=1}^n Y_i}{2n}\right) \\ &= \frac{1}{2n}\mathbf{E}\left(\sum_{i=1}^n Y_i\right) \\ &= \frac{1}{2n}\sum_{i=1}^n \mathbf{E}(Y_i)\end{aligned}$$

$$= \frac{1}{2n} \cdot n \cdot 2\theta$$

$$= \theta$$

Since $\mathbf{E}(\hat{\theta}) = \theta$, $\hat{\theta}$ is an unbiased estimator for θ .

$$\begin{aligned}\mathbf{Var}(\hat{\theta}) &= \mathbf{Var}\left(\frac{\sum_{i=1}^n Y_i}{2n}\right) \\ &= \frac{1}{(2n)^2} \mathbf{Var}\left(\sum_{i=1}^n Y_i\right) \\ &= \frac{1}{(2n)^2} \sum_{i=1}^n \mathbf{Var}(Y_i) \\ &= \frac{1}{(2n)^2} \cdot n \cdot 2\theta^2 \\ &= \frac{\theta^2}{2n}\end{aligned}$$

For a sample of size 3, we can simplify this to $\theta^2/6$.

- (c) What is the MLE for the variance of Y ?


As Y has a $\text{Gamma}(\alpha = 2, \theta)$ distribution, its variance is given by $2\theta^2$, which is a function of θ . By the invariance property of maximum likelihood estimators, if $\hat{\theta}$ is a maximum likelihood estimator for θ , then $g(\hat{\theta})$ is a maximum likelihood estimator for $g(\theta)$, where $g(\cdot)$ is a one-to-one function of θ .

Thus, a maximum likelihood estimator for the variance of Y is given by:

$$g(\hat{\theta}) = 2\hat{\theta}^2 = 2\left(\frac{\bar{Y}}{2}\right)^2 = \frac{\bar{Y}^2}{2}.$$

Question 3

Generate an observation from the binomial distribution with $p = 0.4$ and $n = 40$.

- (a)  Find a 90% confidence interval for p , assuming that you did not know the true value of p .

As usual, we should set our seed before beginning our simulation study. Once again, choose any number you like.

```
set.seed(25)
```

I will store my observation in a variable called `sample_values`.

```
sample_values <- rbinom(n=1, size=40, prob=0.4)
```

```
sample_values
```

```
## [1] 15
```

Note that the above gives the number of successes in a binomial experiment consisting of 40 trials. If we want a proportion, we should divide it by the number of trials. This will be our \hat{p} .

```
sample_values <- sample_values / 40
```

If we seek a $100(1 - \alpha)\%$ confidence interval for p and do not know the true value of p , we can use the formulation given on page 414 of the textbook, which uses \hat{p} instead:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}},$$

where $\alpha/2$ is the area to the right of corresponding quantile.

Using our value of \hat{p} stored in the `sample_values` variable above, our confidence interval is computed as:


```
alpha <- 0.10
zval <- qnorm(alpha/2, lower.tail=FALSE)

se_phat <- sqrt(sample_values * (1 - sample_values) / 40)

sample_values + c(-1, 1) * zval * se_phat

## [1] 0.2490921 0.5009079
```

We are 90% confident that the true value of p lies between 0.25 and 0.5. In this instance, our interval does in fact contain our true value of $p = 0.4$.

- (b)  Consider the interval from (a). Construct 200 such intervals based on a random sample of size $m = 200$ from the Binomial($n = 40$, $p = 0.4$) distribution.

We begin by re-setting our seed.

```
set.seed(69420)
```

We will take an approach similar to what was shown in question 3d of Tutorial 1. I first initialise a data frame with a column called `props` to keep track of the generated proportions, obtained by generating random deviates from the binomial distribution and scaling them by the number of trials.

```
q3b <- data.frame(props = rbinom(n=200, size=40, prob=0.4) / 40)

head(q3b)
```

```
## props
## 1 0.425
## 2 0.300
## 3 0.400
## 4 0.350
## 5 0.275
## 6 0.375
```

Note that the value of `alpha` and `zval` remain the same and can be reused from (a). The values that will change are the ones stored in `se_phat` since these depend on the newly generated values of \hat{p} . As in question 3d of Tutorial 1, I will make use of the pipe (requires **R** 4.1+) and the `transform()` function.

Something to note is that the variable `se_phat` created in (a) exists in our global environment. In the next step, we will be creating another `se_phat`. However, this one will be confined to the scope of its

containing data frame.

```
q3b <- q3b |>
  transform(se_phat = sqrt(props * (1 - props) / 40))

head(q3b)
```

```
## props se_phat
## 1 0.425 0.07816249
## 2 0.300 0.07245688
## 3 0.400 0.07745967
## 4 0.350 0.07541552
## 5 0.275 0.07060011
## 6 0.375 0.07654655
```

I briefly mentioned in the last tutorial that the `transform()` function has a limitation where we cannot create new columns that depend on columns created in the same call. This means that since I have just created the column `se_phat` in the above call to `transform()`, I cannot immediately construct the lower and upper bounds of the corresponding confidence intervals in the same call since these values require the values of `se_phat`. As such, we must make multiple calls to `transform()` to achieve our desired result.


There are ways to get around this limitation through the use of other functions, but we likely won't get to them in this course.

Now that we have the values of `se_phat` in our data frame, we can call `transform()` to create our lower and upper bounds, and once again to check which intervals contain our true value of p .

```
q3b <- q3b |>
  transform(
    lwr = props - zval * se_phat,
    upr = props + zval * se_phat
  ) |>
  transform(contained = (0.4 >= lwr) & (0.4 <= upr))

head(q3b)
```

```
## props se_phat lwr upr contained
## 1 0.425 0.07816249 0.2964341 0.5535659 TRUE
## 2 0.300 0.07245688 0.1808190 0.4191810 TRUE
## 3 0.400 0.07745967 0.2725902 0.5274098 TRUE
## 4 0.350 0.07541552 0.2259525 0.4740475 TRUE
## 5 0.275 0.07060011 0.1588732 0.3911268 FALSE
## 6 0.375 0.07654655 0.2490921 0.5009079 TRUE
```

- (c)  How many of your intervals contained the true value of p ? Was this expected or unexpected?

The number of intervals that contained the true value of $p = 0.4$ is:

```
sum(q3b$contained)
```

```
## [1] 177
```

We can also obtain the percentage of intervals that contained the true value of $p = 0.4$.

```
mean(q3b$contained) * 100
```

[1] 88.5

88.5% of our intervals contained the true value of $p = 0.4$. This value is reasonable. Recall that to say we are 90% confident that our confidence interval (e.g. from (a)) contains the true value of p means that if we were to simulate new samples a sufficiently large number of times under identical conditions and compute confidence intervals with these samples, 90% of these intervals would contain the true value of p .

However, as this is a simulation and there is randomness involved, we shouldn't expect to achieve *exactly* 90%. Any value reasonably close to 90% (whether above or below) is fine. For example, a value of 30% would *not* be fine.