Tutorial 1 Solutions

Question 1

In the 1986 issue of *Consumer Reports*, some data on the calorie content of beef hot dogs is given. Here are the numbers of calories in 20 different hot dog brands:

186,	181,	176,	149,	184,	190,	158,	139,	175,	148
152,	111,	141,	153,	190,	157,	131,	149,	135,	132

Assume that these numbers are the observed values from a random sample of twenty independent normal random variables with mean μ and variance σ^2, both unknown. Find a 90% two-sided confidence interval for the mean number of calories μ.

As the data are assumed to come from a normal distribution, the sample size is small ($n < 30$), and we are estimating σ^2 using S^2, a $100(1-\alpha)\%$ confidence interval for μ is given by:

$$\overline{X} \pm t_{n-1,\,\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

We begin the calculation of this interval by first entering our data into **R**.

```
calories <- c(186, 181, 176, 149, 184, 190, 158, 139, 175, 148,
              152, 111, 141, 153, 190, 157, 131, 149, 135, 132)
```

Next, we calculate some intermediate values that we will store as variables.

```
xbar <- mean(calories)
s <- sd(calories)

n <- length(calories)

alpha <- 0.10
tval <- qt(alpha/2, df=n-1, lower.tail=FALSE)
```

Note that the `lower.tail=FALSE` means that we seek a value q whose area to the **right** is $\alpha/2$.

We will take advantage of the fact that basic mathematical operations in **R** are vectorized. This means that when they are applied to a vector, they are applied element-wise. For example, the following are equivalent:

```
3 + c(-1, 1) * 2
```

```
## [1] 1 5
```

```
c(3 + (-1 * 2), 3 + (1 * 2))
```

```
## [1] 1 5
```

Calculating the confidence interval:

```
xbar + c(-1, 1) * tval * s / sqrt(n)
```

```
## [1] 148.0956 165.6044
```

We are 90% confident that the true mean caloric content of these hot dogs is between 148.096 and 165.604.

Question 2

Suppose that Y is normally distributed with mean 0 and unknown variance σ^2. Then Y^2/σ^2 has a chi-square distribution on 1 degree of freedom. Use the pivotal quantity Y^2/σ^2 to find:

(a) A 95% confidence interval for σ^2 and σ.

A $100(1-\alpha)\%$ confidence interval for σ^2 is derived as follows:

$$\mathbf{P}\left(\chi^2_{1,1-\frac{\alpha}{2}} \leq \frac{Y^2}{\sigma^2} \leq \chi^2_{1,\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\mathbf{P}\left(\frac{\chi^2_{1,1-\frac{\alpha}{2}}}{Y^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{1,\frac{\alpha}{2}}}{Y^2}\right) = 1-\alpha$$

$$\mathbf{P}\left(\frac{Y^2}{\chi^2_{1,1-\frac{\alpha}{2}}} \geq \sigma^2 \geq \frac{Y^2}{\chi^2_{1,\frac{\alpha}{2}}}\right) = 1-\alpha$$

$$\mathbf{P}\left(\frac{Y^2}{\chi^2_{1,\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{Y^2}{\chi^2_{1,1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

A 95% confidence interval ($\alpha = 0.05$) for σ^2 is given by:

$$\left[\frac{Y^2}{\chi^2_{1,0.025}}, \frac{Y^2}{\chi^2_{1,0.975}}\right],$$

where 0.025 and 0.975 are the areas to the right of their respective quantiles. Taking the square root of the endpoints of the above interval gives a 95% confidence interval for σ.

(b) A 95% upper confidence bound for σ^2 and σ.

To obtain the 95% upper confidence bound for σ^2, we simply take the upper bound of the interval from (a) and replace all instances of $\alpha/2$ with α. This gives the interval:

$$\left(0, \frac{Y^2}{\chi^2_{1,0.95}}\right],$$

where 0.95 is the area to the right of the respective quantile. Note that the lower bound of this interval is zero since σ^2 is the variance of a normal distribution and must be greater than zero! Taking the square root of the endpoints of the above interval gives a 95% confidence interval for σ.

(c) A 95% lower confidence bound for σ^2 and σ.

To obtain the 95% lower confidence bound for σ^2, we simply take the lower bound of the interval from (a) and replace all instances of $\alpha/2$ with α. This gives the interval:

$$\left[\frac{Y^2}{\chi^2_{1,\,0.05}},\ \infty\right)$$

where 0.05 is the area to the right of the respective quantile. Taking the square root of the lower endpoint of the above interval gives a 95% confidence interval for σ.

Question 3

Assume that $X_1, \ldots, X_n$ is a random sample of size n from a gamma distribution with $\alpha = 2$ and unknown β.

(a) Use the method of moment generating functions to show that $2\sum_{i=1}^{n} X_i/\beta$ is a pivotal quantity and has a chi-square distribution with $4n$ degrees of freedom.

Each $X_i \sim \text{Gamma}(\alpha = 2,\ \beta)$. The MGF for each X_i is given by:

$$M_X(t) = (1-\beta t)^{-2}, \quad t < \frac{1}{\beta}.$$

Let $U = 2\sum_{i=1}^{n} X_i/\beta$. By the method of moment generating functions:

$$\begin{aligned}
M_U(t) &= \mathbf{E}\left(\exp\{tU\}\right) \\
&= \mathbf{E}\left(\exp\left\{\frac{2t}{\beta}\left(X_1 + X_2 + \ldots + X_n\right)\right\}\right) \\
&= \mathbf{E}\left(\exp\left\{\frac{2t}{\beta}X_1\right\} \cdot \exp\left\{\frac{2t}{\beta}X_2\right\} \cdot \cdots \cdot \exp\left\{\frac{2t}{\beta}X_n\right\}\right) \\
&= \mathbf{E}\left(\exp\left\{\frac{2t}{\beta}X_1\right\}\right) \cdot \mathbf{E}\left(\exp\left\{\frac{2t}{\beta}X_2\right\}\right) \cdot \cdots \cdot \mathbf{E}\left(\exp\left\{\frac{2t}{\beta}X_n\right\}\right) && \text{(Independence)} \\
&= \left[M_X\left(\frac{2t}{\beta}\right)\right]^n && \text{(Identically distributed)} \\
&= \left[\left(1-\beta\left(\frac{2t}{\beta}\right)\right)^{-2}\right]^n \\
&= (1-2t)^{-2n}, \quad t < \frac{1}{2}
\end{aligned}$$

The MGF of U is that of the chi-square distribution on $4n$ degrees of freedom. It follows that U has a chi-square distribution on $4n$ degrees of freedom. Note that although U depends on the unknown β, **its distribution does not.** Since the distribution of U does not depend on the unknown β, $U = 2\sum_{i=1}^{n} X_i/\beta$ is a pivotal quantity.

(b) Use the pivotal quantity $2\sum_{i=1}^{n} X_i/\beta$ to derive a 95% two-sided confidence interval for β.

A $100(1-\alpha)\%$ two-sided confidence interval for β is obtained as follows:

$$\mathbf{P}\left(\chi^2_{4n,\,1-\frac{\alpha}{2}} \leq \frac{2}{\beta}\sum_{i=1}^{n} X_i \leq \chi^2_{4n,\,\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\mathbf{P}\left(\frac{\chi^2_{4n,\,1-\frac{\alpha}{2}}}{2\sum_{i=1}^{n} X_i} \leq \frac{1}{\beta} \leq \frac{\chi^2_{4n,\,\frac{\alpha}{2}}}{2\sum_{i=1}^{n} X_i}\right) = 1-\alpha$$

$$\mathbf{P}\left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{4n,\,1-\frac{\alpha}{2}}} \geq \beta \geq \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{4n,\,\frac{\alpha}{2}}}\right) = 1-\alpha$$

$$\mathbf{P}\left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{4n,\,\frac{\alpha}{2}}} \leq \beta \leq \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{4n,\,\frac{1-\alpha}{2}}}\right) = 1-\alpha$$

Thus, a 95% confidence interval ($\alpha = 0.05$) for β is given by:

$$\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{4n,\,0.025}},\ \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{4n,\,0.975}}\right],$$

where 0.025 and 0.975 are the areas to the right of their respective quantiles.

(c) Generate a sample of $n = 30$ observations from a gamma distribution with $\alpha = 2$ and $\beta = 5$. Use the result of part (b) to find a 95% two-sided confidence interval for β.

Before generating random samples, we should set a seed for reproducibility. Set it to whatever number you wish.

```
set.seed(120)
```

Next, we initialise some intermediate variables, for convenience. Note that we do not actually care about the individual values of our generated sample, only its sum.

```
n <- 30

numerator <- 2 * sum(rgamma(n=n, shape=2, scale=5))

alpha <- 0.05
chi_lower <- qchisq(alpha/2, df=4*n, lower.tail=FALSE)
chi_upper <- qchisq(1-alpha/2, df=4*n, lower.tail=FALSE)
```

Putting it all together:

```
numerator / c(chi_lower, chi_upper)
```

```
## [1] 4.379201 7.279077
```

(d) Consider the interval in part (c). Construct 100 such intervals based on 100 independent samples of size $n = 30$ from a gamma distribution with $\alpha = 2$ and $\beta = 5$. How many of these intervals contain the true β?

Again, for reproducibility, we should re-set our seed. Once again, you can set it to whatever number you wish.

```r
set.seed(99)
```

Since we will repeat this procedure 100 times, we will store our results in a data frame. This way, we can once again take advantage of vectorization of basic mathematical operations in **R** by operating on the columns of our data frame.

We begin by initialising our $2\sum_{i=1}^{n} X_i$ values, storing them in a column called `double_sum`. Again, we do not actually care about the individual values in our samples, only their sum.

```r
q3d <- data.frame(
  double_sum = replicate(n=100, 2 * sum(rgamma(n=n, shape=2, scale=5)))
)

head(q3d)
```

```
##   double_sum
## 1   678.0035
## 2   519.8549
## 3   630.0145
## 4   773.5394
## 5   625.2225
## 6   699.4331
```

Next, we create columns to keep track of the lower and upper bounds of the resulting confidence intervals. We make use of the base-**R** pipe (requires **R** 4.1+), `|>`, which takes the value on the left and passes it to the first argument of the function on the right. It just so happens that the first argument of `transform()` is the data frame that we wish to operate on!

```r
q3d <- q3d |>
  transform(
    lwr = double_sum / chi_lower,
    upr = double_sum / chi_upper
  )

head(q3d)
```

```
##   double_sum      lwr      upr
## 1   678.0035 4.454354 7.403996
## 2   519.8549 3.415348 5.676968
## 3   630.0145 4.139075 6.879942
## 4   773.5394 5.082007 8.447276
## 5   625.2225 4.107593 6.827612
## 6   699.4331 4.595142 7.638013
```

Now that we have the lower and upper bounds of our confidence intervals, we can check which intervals contain the true value of $\beta = 5$. We will first create a column of logicals to keep track of which intervals contained the true value of $\beta = 5$. Here, we take advantage of the fact that logical comparisons are also vectorized!

```r
q3d <- q3d |>
  transform(contained = (5 >= lwr) & (5 <= upr))

head(q3d)
```

```
##   double_sum      lwr      upr contained
```

```
## 1   678.0035 4.454354 7.403996      TRUE
## 2   519.8549 3.415348 5.676968      TRUE
## 3   630.0145 4.139075 6.879942      TRUE
## 4   773.5394 5.082007 8.447276     FALSE
## 5   625.2225 4.107593 6.827612      TRUE
## 6   699.4331 4.595142 7.638013      TRUE
```

Recall that we can perform mathematical operations with `TRUE`s (mapped to 1) and `FALSE`s (mapped to 0). The number of intervals that contained the true value of $\beta = 5$ is:

```
sum(q3d$contained)
```

```
## [1] 98
```

The proportion of intervals that contained the true value of $\beta = 5$ is:

```
mean(q3d$contained)
```

```
## [1] 0.98
```

The main takeaways of this computational exercise is to highlight the fact that you don't always need to write a `for` loop and that you should take advantage of vectorization wherever possible. This will allow you to write code that is functional **and** readable!

Question 4

The number of traps (defects of a certain kind) in a particular type of metal oxide semiconductor transistor has a Poisson distribution with (unknown) mean λ. A sample of $n = 40$ metal oxide semiconductor transistors were randomly selected from a large lot and the number of traps in each transistor in the sample was recorded. The following data was obtained:

$$1,\ 3,\ 2,\ 3,\ 2,\ 1,\ 6,\ 3,\ 3,\ 4,\ 5,\ 4,\ 3,\ 5,\ 2,\ 4,\ 4,\ 3,\ 6,\ 1$$
$$1,\ 1,\ 4,\ 6,\ 2,\ 2,\ 2,\ 3,\ 4,\ 1,\ 7,\ 1,\ 3,\ 3,\ 1,\ 3,\ 2,\ 3,\ 7,\ 2$$

(a) Find a large-sample 98% two-sided confidence interval, $[\hat{\lambda}_L, \hat{\lambda}_U]$, for the true average number of traps, λ.

From lecture, it was shown that for a sufficiently large ($n \geq 30$) random sample from a Poisson distribution, a $100(1-\alpha)\%$ large-sample two-sided confidence interval is given by:

$$\overline{X} \pm z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}}{n}}$$

We begin the computation of this interval by first reading the data into **R**.

```
traps <- c(1, 3, 2, 3, 2, 1, 6, 3, 3, 4,
           5, 4, 3, 5, 2, 4, 4, 3, 6, 1,
           1, 1, 4, 6, 2, 2, 2, 3, 4, 1,
           7, 1, 3, 3, 1, 3, 2, 3, 7, 2)
```

As before, we compute the intermediate values and store them as variables, for convenience.

```r
xbar <- mean(traps)

n <- length(traps)

alpha <- 0.02
zval <- qnorm(alpha/2, lower.tail=FALSE)
```

The 98% two-sided confidence interval is computed as:

```r
xbar + c(-1, 1) * zval * sqrt(xbar / n)
```

```
## [1] 2.429989 3.720011
```

(b) Find a large-sample 98% one-sided confidence interval, $[\hat{\lambda}_L, \infty)$, for the true average number of traps, λ.

This is a one-sided lower confidence bound, so we simply take the lower end of the formula from (a) and swap out $z_{\alpha/2}$ for z_α. The lower bound of our one-sided confidence interval is given by:

$$\overline{X} - z_\alpha \cdot \sqrt{\frac{\overline{X}}{n}}.$$

To construct this interval in **R**, we can reuse most of the values from (a). The only value that needs to be updated is `zval` (though `alpha` will remain the same).

```r
zval <- qnorm(alpha, lower.tail=FALSE)
```

For clarity, we append an `Inf` (infinity) to the right side of our interval. The 98% one-sided lower confidence interval is computed as:

```r
c(xbar - zval * sqrt(xbar / n), Inf)
```

```
## [1] 2.505571      Inf
```

(c) Interpret the results obtained in (a) and (b).

From (a), we can be 98% confident that the true average number of traps is between 2.43 and 3.72, i.e.

$$2.43 \leq \lambda \leq 3.72.$$

From (b), we can be 98% confident that the true average number of traps is greater than or equal to 2.506, i.e.

$$\lambda \geq 2.506.$$

Note: The data used in this question were actually generated from a Poisson distribution with $\lambda = 3$.