# W203 HW7

*Suhas Gupta*

*10/22/2018*

**Question 1: The Meat**

**(a)** No, we don't expect the average consumption of ground beef per capita per month to be normally distributed. The random variable in question here is a function (average) of the actual beef consumption per capita per month. This is the full population distribution and can be skewed based on the tendency of beef consumption by Americans. Since this is not a sample statistic, we cannot assume the distribution to be normal.

**(b)** If the distribution of average beef consumption by American does not have a exterme skew, then we expect the distribution of sample mean from sample size of 100 to be normally distributed. This follows from the central limit theorem that states that for sample size greater than 30, the sample statistic can be assumed to be normally distributed.

**(c)** The 95% confidence interval for a sample size > 40 is defined based on the z-distribution as:

$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

The 95% confidence interval is thus defined by the interval

$$\bar{X} \pm 1.96\sigma\sqrt{n}$$

$$\text{Upper Limit, } ul = 2.45 + 1.96 \cdot \frac{2}{\sqrt{100}} = 2.45 + 0.392 = 2.842$$

and

$$\text{Lower Limit, } ll = 2.45 - 1.96 \cdot \frac{2}{\sqrt{100}} = 2.45 - 0.392 = 2.058$$

Thus, the 95% confidence interval for berkeley students is **(2.058,2.842)**

**Question 2 GRE Scores**

Given, the formula used to compute 95% confidence interval assuming large sample size:

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

For a sample size < 40, we cannot assume the sample statistic is normally distributed. Thus, for n = 10 we need to use the student's t-distribution to compute the real confidence interval.

The CI of a t-distribution for small sample size (n<40) is given by:

$$\left(\bar{X} - t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}\right)$$

For **n = 10, df = 10 - 1 = 9**

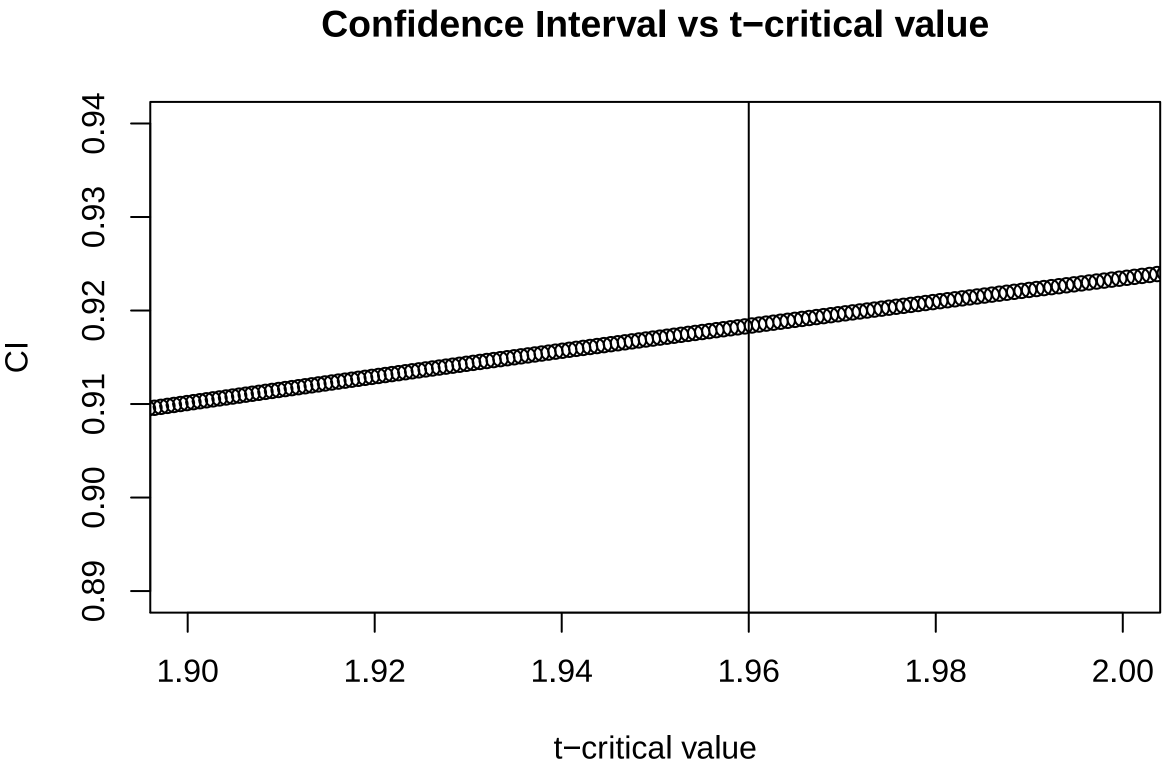
Comparing the two expressions, we get

$$t_{\alpha/2,n-1} = 1.96$$

Thus we need to find the value of $\alpha$ which corresponds to the real confidence interval.

From the t-distribution tables we conclude that the value of $100(1-\alpha)$ % will be between 90% and 95%. Now we can use the R function **qt** to sweep over values of $\alpha$ to get the real CI that results in the above critical value.

```
n = 10
alpha_values = seq(0.05,0.1,0.0001)
ci = 1 - alpha_values
t_values = qt((1-alpha_values/2), df = n-1)
plot(t_values,ci,xlim=c(1.9,2.0),ylim=c(0.91,0.92),asp=1,
     xlab="t-critical value", ylab="CI",
     main="Confidence Interval vs t-critical value"
     )
abline(v=1.96)
```



From above, we determine the real confidence interval for n = 10 is **between 91% and 92%**

When n = 200 (>40), we can approximate the t-distribution with a z-distribution. In this case, the confidence interval for critical value of 1.96 is **95%**

**Question 3 Maximim Likelihood Estimation for an Exponential Distribution**

Given the exponential distribution funtion

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

**(a)** Likelihood function,

$$L(\lambda) = f(x_1, x_2, x_3, ..., x_n|\lambda)$$

Since the arrival times are independent, the above expression can be written as:

$$L(\lambda) = f(x_1|\lambda)f(x_2|\lambda)...f(x_n|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

### (b)

Taking the log of likelihood,

$$ln[L(\lambda)] = ln\Big(\prod_{i=1}^{n} \lambda e^{-\lambda x_i}\Big)$$

$$ln[L(\lambda)] = \sum_{i=1}^{n} ln(\lambda e^{-\lambda x_i})$$

$$ln[L(\lambda)] = n \cdot ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

### (c)

Equating the derivative of log of likelihood w.r.t. $\lambda$ to 0 will let us solve for the MLE for the observed data:

$$\frac{\partial}{\partial\lambda} ln(L(\lambda)) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\implies \hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i}$$

From the above expression we can see that MLE for paramter $\lambda$ is the reciprocal of the mean time between arrivals $\mu$

### (d)

```
library("kableExtra")
times = c(2.65871285, 8.34273228, 5.09845548,
          7.15064545, 0.39974647, 0.77206050,
          5.43415199, 0.36422211, 3.30789126,
          0.07621921, 2.13375997, 0.06577856,
          1.73557740, 0.16524304, 0.27652044)
inverse_of_mean = 1/mean(times)

# Define the likelihood function (negative product for maximum optimization)
lik.exp <- function(x, lambda){
    -prod(lambda * exp(-lambda * x))
}

# Optimize the likelihood function to find the max
lambda_mle_optim = optim(par = 2, lik.exp, x = times,
                         method="Brent", lower=0,upper=1)$par

df = data.frame(inverse_of_mean,lambda_mle_optim,
                100*abs(inverse_of_mean-lambda_mle_optim)/inverse_of_mean)
kable(df, format= "pandoc", booktable = T, longtable = T,
      align="c", position="center",digits = 5,
      caption="Comparison of algebraic and R optimized MLE
      for a sample from exponential distribution",
      col.names=(c("Algebraic MLE","Simulated MLE","Percent Difference")))
```

Table 1: Comparison of algebraic and R optimized MLE for a sample from exponential distribution

| Algebraic MLE | Simulated MLE | Percent Difference |
|---|---|---|
| 0.39493 | 0.39493 | 0 |