Statistics 2 Unit 1

Team 8

April 5, 2018

Contents

1 Task37:

```
bodytemp <- read.csv("bodytemp.txt")
males <- bodytemp$temp[bodytemp$gender == 1]
females <- bodytemp$temp[bodytemp$gender == 2]
n <- length(males)
meanm <- mean(males)
meanf <- mean(females)
sdm <- sd(males)
sdf <- sd(females)
```

We have the following results:

```
meanm
## [1] 98.10462
sdm
## [1] 0.6987558
```

```
meanf
## [1] 98.39385
sdf
## [1] 0.7434878
```

From lecture notes, we know that for normal distribution:
$\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\overline{\sigma}} \sim t_{n-1} \Rightarrow \left(\overline{X}-\mu\right) \sim \frac{t_{n-1}\overline{\sigma}}{\sqrt{n}}$

```
diffm <- qt(0.975, df= n - 1)*sdm/sqrt(n)

difff <- qt(0.975, df= n - 1)*sdf/sqrt(n)

m95 <- c(meanm - diffm, meanm + diffm)

f95 <- c(meanf - difff, meanf + difff)
```

From which we get the following results:

```
diffm

## [1] 0.1731432

difff

## [1] 0.1842272

m95

## [1] 97.93147 98.27776

f95

## [1] 98.20962 98.57807
```

It seems that the standard folklor is working not that accurately.

2 Task38:

As X_i are i.i.d and uniformly distributed on [0,θ] than

$$F(x) = \begin{cases} 1, & \text{if } x > \theta \\ \frac{x}{\theta}, & \text{if } 0 \leq x \leq \theta \\ 0 & \text{if } x < 0 \end{cases}$$

a)We need to show that if the distribution of $\frac{X_n}{\theta}$ is independent of θ, then it is a pivot.

Substitute $\frac{X_n}{\theta}$ for y

$$\begin{aligned}
\mathrm{F}_Y(y) &= P(Y \leq y) \\
&= P(\frac{max(X_1, \ldots, X_n)}{\theta} \leq y) \\
&= P(max(X_1, \ldots, X_n) \leq \theta y) \\
&= F_{max(X)}(\theta y) \\
&= P(X_1 \leq x_n, \ldots, X_n \leq x_n) \\
&= [F(x_{(n)})]^n \\
f(x_{(n)}) &= nF(x_{(n)})^{n-1} \\
f(x_{(n)}) &= \begin{cases} \frac{n}{\theta^n} x_{(n)}^{n-1} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

b)$P(X_{(n)} \leq \theta \leq \alpha^{-1/n} X_{(n)})$ has to be equal to 1 – α, if this is a real CI.

$$\begin{aligned}
P(X_{(n)} \leq \theta \leq \alpha^{-1/n} X_{(n)}) &= P(X_{(n)} \alpha^{1/n} \geq \theta \alpha^{-1/n} \geq X_{(n)}) \\
&= P(\alpha^{1/n} \geq X_{(n)} \theta \alpha^{-1/n} \geq 1) \\
&= P(\theta \geq X_{(n)} \geq \theta \alpha^{-1/n}) \\
&= P(\theta \alpha^{-1/n} \leq X_{(n)} \leq \theta) \\
&= \int_{\theta \alpha^{-1/n}}^{\theta} f_{X_{(n)}} dX_{(n)} \\
&= \int_{\theta \alpha^{-1/n}}^{\theta} \frac{nx}{\theta^n} dX_{(n)} \\
&= \frac{n}{\theta^n} \int_{\theta \alpha^{-1/n}}^{\theta} x dX_{(n)} \\
&= \frac{n}{\theta^n} \left[\frac{\theta^n}{n} - \frac{(\alpha^{1/n} \theta)^n}{n} \right] \\
&= \frac{n}{\theta^n} \frac{\theta^n (1 - \alpha)}{n} \\
&= 1 - \alpha
\end{aligned}$$

3 Task39:

For the RVs $X_1, ..., X_n$, which are i.i.d from $N(\mu, a^2\mu^2)$ the pivot will be random variable with Student distribution:

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S},$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ According to MLE estimation of Normal distribution parameters:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now we need to find unbiased estimates of μ and σ^2.
$\hat{\mu} = \bar{X}$, shown above. However $\hat{\sigma}^2 = a^2\hat{\mu}^2$ will be biased. In order to find unbiased estimate S^2 we have to transform $\hat{\sigma}^2$ by multiplication by n:

$n\hat{\sigma}^2 = \frac{n}{n}\sum_{i=1}^{n}(X_i - \bar{X})^{=} na^2\hat{\mu}^2$

Now, plug this result in the equation of S:

$$S^2 = \frac{n}{n-1}a^2\hat{\mu}^2$$

Finally, we obtain our pivot:

$\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\sqrt{n}(\hat{\mu}-\mu)}{\sqrt{\frac{n}{n-1}a^2\hat{\mu}^2}} = \frac{\sqrt{n-1}(\hat{\mu}-\mu)}{a\hat{\mu}} = \frac{\sqrt{n-1}\hat{\mu}}{a\hat{\mu}} - \frac{\sqrt{n-1}\mu}{a\hat{\mu}} = \frac{\sqrt{n-1}}{a}(1 - \frac{\mu}{\hat{\mu}}) \sim t_{n-1},$
where a and n are known and $n-1$ is degrees of freedom.

Now we need to construct a $100(1-\alpha)$ percent confidence interval for μ. Confidence interval is based on calculated pivot:

$$P(Q_{t_{n-1}}(\alpha/2) \leq \frac{\sqrt{n-1}}{a}(1 - \frac{\mu}{\hat{\mu}}) \leq Q_{t_{n-1}}(1-\alpha/2)) = 1 - \alpha$$

From this we can derive:

$$P(Q_{t_{n-1}}(\alpha/2)\frac{a\hat{\mu}}{\sqrt{n-1}} - \hat{\mu} \leq -\mu \leq Q_{t_{n-1}}(1-\alpha/2)\frac{a\hat{\mu}}{\sqrt{n-1}} - \hat{\mu}) = 1 - \alpha$$

$$P(\hat{\mu} - Q_{t_{n-1}}(1-\alpha/2)\frac{a\hat{\mu}}{\sqrt{n-1}} \leq \mu \leq \hat{\mu} - Q_{t_{n-1}}(\alpha/2)\frac{a\hat{\mu}}{\sqrt{n-1}}) = 1-\alpha$$

Since Student distribution is symmetric
$\Rightarrow -Q_{t_{n-1}}(\alpha/2) = Q_{t_{n-1}}(1-\alpha/2)$

$\Rightarrow$

$$P(\hat{\mu} - Q_{t_{n-1}}(1-\alpha/2)\frac{a\hat{\mu}}{\sqrt{n-1}} \leq \mu \leq \hat{\mu} + Q_{t_{n-1}}(1-\alpha/2)\frac{a\hat{\mu}}{\sqrt{n-1}}) = 1-\alpha$$

4 Task40:

4.1 a)

A confidence interval for the mean of a normally distributed sample (where the standard deviation is known) has the form

$$100(1-\alpha) = \left(\bar{X} - z_{1-\lambda_1}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\lambda_2}\frac{\sigma}{\sqrt{n}}\right).$$

where $\lambda_1 + \lambda_2 = \alpha$.

Since in the exercise $\lambda_1 = \alpha - \beta$ and $\lambda_2 = \beta$, what just sums up to α, the expression is a $100(1-\alpha)$ confidence interval.

4.2 b)

The length of the interval is
$\left(\bar{X} - Z_{\beta}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - Z_{\alpha-\beta}\frac{\sigma}{\sqrt{n}}\right) = (Z_{\alpha-\beta} - Z_{\beta})\frac{\sigma}{\sqrt{n}}$.
Because when $\beta \vee \alpha$ is varied $\frac{\sigma}{\sqrt{n}}$ remains constant, this is minimized provided $Z_{\alpha-\beta} - Z_{\beta}$ is minimized.

Intuitively, the density of the normal distribution is symmetric about μ, and steadily decreasing for positive z. So if we want area say 0.95, the shortest interval is obtained if we are using the "fat" part of the density function.

5 Task 41

If X_i is lognormally distributed, then $Y_i = ln(X_i) \sim N(\mu, \sigma)$. We also know that for n independently normally distributed variables Y_i and its μ

estimation it holds that

$$\frac{(\bar{Y}-\mu)}{S/\sqrt{n}} \sim t_{n-1}$$

By calculating sample mean $\bar{Y}$ and sample variance S^2:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} ln(X_i)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(ln(X_i)-\bar{Y})^2$$

and using previous statement we could construct confidence interval for μ:

$$P(t_{n-1}(\alpha/2) \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq t_{n-1}(1-\alpha/2)) = 1-\alpha$$

since t-distribution is symmetric, $t_{n-1}(\alpha/2) = -t_{n-1}(1-\alpha/2)$. Therefore:

$$P(\bar{Y} - \frac{S}{\sqrt{n}}t_{n-1}(1-\alpha/2) \leq \mu \leq \bar{Y} + \frac{S}{\sqrt{n}}t_{n-1}(1-\alpha/2)) = 1-\alpha$$

$$P(\frac{1}{n}\sum_{i=1}^{n} ln(X_i) - \frac{S}{\sqrt{n}}t_{n-1}(1-\alpha/2) \leq \mu \leq \frac{1}{n}\sum_{i=1}^{n} ln(X_i) + \frac{S}{\sqrt{n}}t_{n-1}(1-\alpha/2)) = 1-\alpha$$

where in our task $\alpha = 0.05$.
So with the targeted parameter μ the scheme of constructing of confidence interval on practice is the following:
a) Generate random samples $X_1^j ... X_n^j$, where $j = 1, ..., m$
b) Compute the confidence interval C_j for the j^{th} sample
c) Compute $y_j = I(\mu \in C_j)$ for the j^{th} sample
d) Compute the empirical confidence level $\bar{y} = \frac{\sum_{j=1}^{m} y_j}{m}$
Now combining all parts together we could compute an empirical estimate of confidence interval:

```
ex41 <- function (mu, sigma, n_sim, sample_size) {
  alpha <- 0.05
  x <- rep(0, n_sim)
  for (i in (1:n_sim)) {
    smpl <- rlnorm(sample_size,mu,sigma)
    lsmpl <- log(smpl)
    logmean <- mean(lsmpl)
```

```
    logvar <- sum((lsmpl - logmean)^2)/(sample_size -
       1)
    low <- logmean - qt(1 - alpha/2, df = sample_size
       - 1) * sqrt(logvar)/sqrt(sample_size)
    up <- logmean + qt(1 - alpha/2, df = sample_size -
        1) * sqrt(logvar)/sqrt(sample_size)
    x[i] <- low < mu & mu < up
  }
  sum(x/n_sim)
}

ex41(2,5,1000,1000)

## [1] 0.929
```

6 Task 42

```
ex42 <- function(){

   alpha <- 0.05
   n <- 20
   m <- 1000

   upper <- numeric(m)
   lower <- numeric(m)

       for(i in 1:m){
   #x is chi-sqr distributed with the sample size 20
      and 2 degrees of freedom
   x <- rchisq(n, 2)
   # to estimate the coverage probability, we need to
      find the bounds
   upper[i] <- mean(x) + sd(x)/sqrt(n)*qt(1 - alpha /
      2, df=n-1)
   lower[i] <- mean(x) - sd(x)/sqrt(n)*qt(1 - alpha /
      2, df=n-1)
       }
```

```
   # True mean of chi-squared distribution with 2
      degrees of freedom is known to be 2.
   # Therefore we need to take the average, which
      satisfies both of the conditions of lower bound
      to be less than 2
   # and upper bound to be more than 2. This would be
      a MC coverage probability estimation
   return(mean(lower < 2 & upper > 2))

   }

ex42()

## [1] 0.916
```

Thus the estimated mean's t-interval is covered by the one, for which data is assumed to be chi-sqr distributed, in over 90% of times.

7 Task 43

From the lecture notes we know that Fisher Information is:

$$I(\theta) = E((\frac{\partial \log(f(x|\theta)}{\partial \theta})^2)$$

Using this formula let's calculate Fisher Information for univariate normal distribution:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$l(x|\theta) = -\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2$$

$$\frac{\partial(x|\theta)}{\partial\mu} = \frac{2(x-\mu)}{2\sigma^2} = \frac{(x-\mu)}{\sigma^2}$$

Now we need to take a derivative with respect to σ^2.

$$\frac{\partial(x|\theta)}{\partial\sigma^2} = \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}$$

In order to get our Fisher Information matrix, we need to multiply column vector of 1st partial derivatives by its row vector.

$$\begin{bmatrix} \frac{(x-\mu)}{\sigma^2} \\ \\ \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \end{bmatrix} \times \begin{bmatrix} \frac{(x-\mu)}{\sigma^2} & \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{(x-\mu)^2}{\sigma^4} & -\frac{(x-\mu)}{2\sigma^4} + \frac{(x-\mu)^3}{2\sigma^6} \\ -\frac{(x-\mu)}{2\sigma^4} + \frac{(x-\mu)^3}{2\sigma^6} & \frac{1}{4\sigma^4} - \frac{(x-\mu)^2}{2\sigma^6} + \frac{(x-\mu)^4}{4\sigma^8} \end{bmatrix}$$

Now we need to calculate the expected value of every element of this matrix. Note that the matrix is symmetric. So let's start:

$E(\frac{(x-\mu)^2}{\sigma^4}) = \frac{1}{\sigma^4}(E(x^2) - 2E(x)\mu + \mu^2) = \frac{1}{\sigma^4}(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2) = \frac{1}{\sigma^2}$

Remark: $E(x^2) = \sigma^2 + \mu^2$

$E(-\frac{(x-\mu)}{2\sigma^4} + \frac{(x-\mu)^3}{2\sigma^6}) = \frac{1}{2\sigma^4}E(x-\mu) + \frac{1}{2\sigma^6}(E(x^3) - 3\mu E(x^2) + 3\mu^2 E(x) - \mu^3) =$
$= \frac{1}{2\sigma^6}(\mu^3 + 3\mu\sigma^2 - 3\mu(\mu^2 + \sigma^2) + 3\mu^3 - \mu^3) = 0$

Remark: $E(x^3) = \mu^3 + 3\mu\sigma^2$

And we also know that 3rd central moment $E(x-\mu)^3$ as well as the 1st central moment for the standard normal distribution is equal to zero.

$E(\frac{1}{4\sigma^4} - \frac{(x-\mu)^2}{2\sigma^6} + \frac{(x-\mu)^4}{4\sigma^8}) = \frac{1}{4\sigma^4} - \frac{\sigma^2}{2\sigma^6} + \frac{3\sigma^4}{4\sigma^8} = \frac{1}{4\sigma^4} - \frac{1}{2\sigma^4} + \frac{3}{4\sigma^4} = \frac{1}{2\sigma^4}$

Remark: the 4th central moment of standard normal distribution is
$E(x-\mu)^4 = 3\sigma^4$.

So, our final result of Fisher Information matrix of univariate normal distribution is:

$$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Small check: Fisher Information matrix should be symmetric and positive semidefinite.
Also from lecture notes we know the theorem that when sample size tends to infinity, the distribution of maximum likelihood estimate is $N(\theta_0, \frac{1}{nI(\theta_0)})$.
So, let's consider 2 cases:
1) We do not know μ;
2) We do not know σ.
1st Case:
Let's assume that the true value of μ is μ_0, and the MLE of μ is $\hat{\mu}$. Since we estimate μ, Fisher Information $I = \frac{1}{\sigma^2}$. Therefore, the asymptotic distribution of $\hat{\mu} \sim N(\mu_0, \frac{\sigma^2}{n})$. If we substract mean and divide by σ, we will

get that random variable $\frac{(\hat{\mu}-\mu_0)\sqrt{n}}{\sigma} \sim N(0,1)$. In general, the theorem says $\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \sim N(0,1)$.
2nd Case:
Now let's assume that the true value of σ^2 is σ_0^2, and the MLE of σ^2 is $\hat{\sigma^2}$. Since we estimate σ^2, Fisher Information $I = \frac{1}{2\sigma^4}$. Therefore, the asymptotic distribution of $\hat{\sigma^2} \sim N(\sigma_0^2, \frac{2\sigma^4}{n})$.So, random variable $\frac{(\hat{\sigma^2}-\sigma_0)\sqrt{n}}{\sqrt{2}\sigma^2} \sim N(0,1)$.

8 Task 44

The pmf of the multinomial distribution is given. By using the hint from the exercise we can write the likelihood function as

$$f(x_1,...,x_k|p_1,...,p_k) = \frac{n!}{\prod_{j=1}^{k} x_j!} \prod_{j=1}^{k-1} p_j^{x_j} \ (1-\sum_{j=1}^{k-1} p_j)^{1-\sum_{j=1}^{k-1} x_j}.$$

The log-likelihood is then given by

$$log(f(x|p)) = log(n!) - \sum_{j=1}^{k} log(x_j!) + \sum_{j=1}^{k-1} x_j log(p_j) + (1-\sum_{j=1}^{k-1} x_j) log(1-\sum_{j=1}^{k-1} p_j).$$

The first partial derivative wrt to p_i is just

$$\frac{\partial log f}{\partial p_i} = \frac{x_i}{p_i} - \frac{1-\sum_{j=1}^{k-1} x_j}{1-\sum_{j=1}^{k-1} p_j}.$$

The second derivatives are given by

$$\frac{\partial^2 log f}{\partial p_i^2} = -\frac{x_i}{p_i^2} - \frac{1-\sum_{j=1}^{k-1} x_j}{(1-\sum_{j=1}^{k-1} p_j)^2}$$

(which will be the entries on the diagonal) and

$$\frac{\partial^2 log f}{\partial p_i p_l} = -\frac{1-\sum_{j=1}^{k-1} x_j}{(1-\sum_{j=1}^{k-1} p_j)^2}, i \neq l.$$

The marginal distribution of a multinomial distributed random variable is Bernoulli distributed. Note that the expectation of a Bernoulli-distributed random variable is just p.
Since we know from the hint in the exercises that $p_k = 1-(p_1+...+p_{k-1})$ the expectation of the second partial derivatives on the diagonal are given

by $-\frac{1}{p_i} - \frac{1}{p_k}$ and of the off-diagonal entries by $-\frac{1}{(p_k)}$.
So we end up with a Fisher-information matrix like (note that the information matrix of the second derivatives has to be multiplied with -1)

$$I(\Theta) = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_k} & \frac{1}{p_k} & \dots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2} + \frac{1}{p_k} & \frac{1}{p_k} & \dots \\ \dots & \frac{1}{p_k} & \dots & \dots \\ \frac{1}{p_k} & \dots & \dots & \frac{1}{p_{k-1}} + \frac{1}{p_k} \end{pmatrix}$$

This is just the expression from the exercise sheet.